The Randomized Dependence Coefficient

David Lopez-Paz & Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
dlopez@tue.mpg.de

Abstract

We introduce the Randomized Dependence Coefficient (RDC), a novel measure of non-linear dependence between random variables of arbitrary dimension. Throughout theoretical analysis and empirical evaluation, it is shown that RDC is a highly-scalable (running over millions of samples in seconds), very easy to implement (five lines of R code) estimator of the well-known Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient.

1 Introduction

Measuring dependence between random variables is a fundamental task in statistics. However, given the exponential amount of possible association patterns and the curse of dimensionality induced by finite sample sizes, there is *no free lunch* when tackling this challenging problem. Historically, the most commonly used measures of statistical dependence are Pearson's rho, Spearman's rank or Kendall's tau. Although computationally efficient and theoretically well understood, these coefficients only take into account a very limited class of possible association patterns, i.e., linear or monotonically increasing functions.

In recent years, there has been an increase in interest to develop more general non-linear statistical dependence measures. Some examples are the Alternating Conditional Expectations (ACE) [3], the Hilbert-Schmidt Independence Criterion (HSIC) [5], the Distance or Brownian Correlation Coefficient (dCor) [15, 16], the Maximal Information Coefficient (MIC) [11], the Copula Maximum Mean Discrepancy (Copula-MMD) [8] or the Mutual Information Dimension Coefficient (MID) [14]. However, these methods have very demanding computational complexities (quadratic in the number of samples for ACE, HSIC, dCor, Copula-MMD or MIC), are limited to measure dependencies between scalar random variables (ACE, MIC, MID), show poor performance under the existence of additive noise (MIC, MID) or are very difficult to implement (ACE, MIC).

In an effort to address this problems, we present the *Randomized Dependence Coefficient* (RDC), an estimator of the Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient (HGR). RDC defines the dependence between two random variables as the largest canonical correlation between k random non-linear proyections of their respective marginal observation ranks. Furthermore, RDC is capable of measuring dependence between random variables of arbitrary dimension within a computational cost of $O(k^2 n)$, and it is extremely easy to implement: just five lines of R code, that are included in our Appendix A.

The rest of this article is organized as follows. In Section 2 we review the classical work of Alfréd Rényi [10], in which seven desirable fundamental properties of dependence measures are proposed and the Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient (HGR) is proved to satisfy. In Section 3 we introduce the Randomized Dependence Coefficient, an efficient estimator of HGR. Properties of RDC are studied in Section 4. Finally, in Section 6, a series of numerical experiments are conducted to verify the performance of the proposed estimator when compared to other state-of-the-art non-linear dependence measures.

2 Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient

In 1959, Alfréd Rényi introduced seven fundamental properties that any measure of dependence $\rho^* : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ should satisfy [10]:

1. $\rho^*(X, Y)$ is defined for any pair of non-constant random variables X and Y.
2. $\rho^*(X, Y) = \rho^*(Y, X)$
3. $0 \leq \rho^*(X, Y) \leq 1$
4. $\rho^*(X, Y) = 0$ iff X and Y are statistically independent.
5. For bijective Borel-measurable functions $f, g : \mathbb{R} \to \mathbb{R}$, $\rho^*(X, Y) = \rho^*(f(X), g(Y))$.
6. $\rho^*(X, Y) = 1$ if for Borel-measurable functions f or g, $Y = f(X)$ or $X = g(Y)$.
7. If $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$, then $\rho^*(X, Y) = |\rho(X, Y)|$, where ρ is the correlation coefficient.

In this same work it is proved that the *Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient* (HGR) [4] satisfies all of the seven proposed fundamental properties. Such statistic was defined by Gebelein in 1941 [4] as:

$$\mathrm{hgr}(X, Y) = \sup_{f,g} \rho(f(X), g(Y)), \tag{1}$$

where f and g run over all Borel-measurable functions with finite variance. Given the infinite-dimensional search space defined over the supremum, its computation becomes infeasible. In the following section we propose a novel, scalable and very easy to implement estimator of the HGR correlation: the Randomized Dependence Coefficient.

3 Randomized Dependence Coefficient

The *Randomized Dependence Coefficient* (RDC) measures the strength of dependence between the random variables $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ as the maximum canonical correlation between k random non-linear proyections of their respective marginal observation ranks.

In the following we describe each of the necessary steps to construct the RDC statistic which are (i) obtain the observation marginal ranks (ii) project them randomly in a non-linear way and (iii) compute the maximal canonical correlation between the two sets of non-linear random projections.

3.1 Estimation of Marginal Observation Ranks

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with continuous marginal cumulative distribution functions (cdfs) P_i, $1 \leq i \leq d$. The following theorem assures that the transformed random vector $\boldsymbol{U} = (U_1, \ldots, U_d) := (P_1(X_1), \ldots, P_d(X_d))$ has uniform marginals:

Theorem 1. *(Probability Integral Transform [7]) Given a random variable X with cdf P, the random variable $Y := P(X)$ is uniformly distributed on $[0, 1]$:*

The random variables $U_1, \ldots, U_d$ are also known as the marginal observation ranks of $X_1, \ldots, X_d$. Interestingly, $\boldsymbol{U}$ preserves the dependence structure of the original random vector $\boldsymbol{X}$, but ignores each of the d marginal forms. This property is desirable when measuring dependence, since working with observation ranks makes our statistic invariant respect to changes in marginal distributions (i.e., respect to arbitrary monotonic transformations). The joint distribution of $\boldsymbol{U}$ is known as the copula of $\boldsymbol{X}$:

Theorem 2. *(Sklar's [13]) Let the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ have continuous marginal cumulative distribution functions P_i, $1 \leq i \leq d$. Then, the joint cumulative distribution of $\boldsymbol{X}$ can be uniquely expressed as:*

$$P(X_1, \ldots, X_d) = C(P_1(X_1), \ldots, P_d(X_d)), \tag{2}$$

where the distribution C is known as the copula of $\boldsymbol{X}$.

In practice, the estimation of univariate cdfs is easily done given a few hundred of observations. Moreover, non-parametric empirical cumulative distribution functions estimates (ecdfs) converge uniformly to the true distribution along all the whole domain of the random variable, result formalized by the Glivenko-Cantelli theorem:

Theorem 3. *(Glivenko-Cantelli [17]) Let $X_1, \ldots, X_n$ be iid random variables with common cumulative distribution function P. Then, the empirical cumulative distribution function, defined as*

$$P_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \tag{3}$$

converges uniformly to P:

$$\|P_n - P\|_\infty = \sup_{x \in \mathbb{R}} |P_n(x) - P(x)| \xrightarrow{a.s.} 0 \tag{4}$$

Computing ecdfs involves sorting and ranking each of the marginal observations. Therefore, to obtain the marginal observation ranks of two random variables $\boldsymbol{X} \in \mathbb{R}^p$ $\boldsymbol{Y} \in \mathbb{R}^q$ we need to perform $O(2(p+q)n\log(n))$ operations.

3.2 Generation of Random Non-Linear Proyections

In a very elegant result, Rahimi and Brecht [9] proved that random, non-linear proyections of data features can generate enough expressiviness to obtain high-performance regressors when linearly combined:

Theorem 4. *(Rahimi-Brecht) Let p be a distribution on Ω and $|\phi(\boldsymbol{x}; \boldsymbol{w})| \leq 1$. Fix $f^* \in \mathcal{F} = \left\{ f(\boldsymbol{x}) = \int_\Omega \alpha(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w}) \mathrm{d}\boldsymbol{w} \middle| |\alpha(\boldsymbol{w})| \leq Cp(\boldsymbol{w}) \right\}$. Draw $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ iid from p. Then with probability at least $1 - 2\delta$, with $\delta > 0$, and respect to some $L-$Lipschitz loss function $\boldsymbol{R}$ and training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, there exist $\alpha_1, \ldots, \alpha_K$ such that $\hat{f}_K(x) = \sum_{i=1}^K \alpha_i \phi(\boldsymbol{x}; \boldsymbol{w}_i)$ satisfies:*

$$\boldsymbol{R}[\hat{f}_K] - \min_{f \in \mathcal{F}} \boldsymbol{R}[f] \leq O\left(\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{K}} \right) LC \sqrt{\log \frac{1}{\delta}} \right) \tag{5}$$

In other words, Theorem 4 states that one can achieve close to state-of-the-art regressors of the form $\sum_{i=1}^K \alpha_i \phi(\boldsymbol{x}; \boldsymbol{w}_i)$ when we randomize over the weights $\boldsymbol{w}_i$ inside the non-linearities ϕ but optimize the linear mixing coefficients $\boldsymbol{\alpha}$. In the rest of this article we will use sigmoids $\phi(x) = 1/(1 + \exp(-x))$ as non-linearities and random weights $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. When ϕ is sigmoidal, Barron showed that any function whose derivative was an absolutely integrable in the Fourier domain could be approximated with L_2 error below $O(1/\sqrt{K})$ terms [2].

Given a data collection $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we will denote by:

$$\boldsymbol{\Phi}(\boldsymbol{X}; k, s) = \begin{pmatrix} \phi(\boldsymbol{w}_1^T \boldsymbol{x}_1) & \cdots & \phi(\boldsymbol{w}_k^T \boldsymbol{x}_1) \\ \vdots & \vdots & \vdots \\ \phi(\boldsymbol{w}_1^T \boldsymbol{x}_n) & \cdots & \phi(\boldsymbol{w}_k^T \boldsymbol{x}_n) \end{pmatrix}^T \tag{6}$$

the $k-$th order random non-linear projection from $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ to $\boldsymbol{\Phi}_{\boldsymbol{X}}^k := \boldsymbol{\Phi}(\boldsymbol{X}; k) \in \mathbb{R}^{k \times n}$. Thus, the computational complexity of computing $\boldsymbol{\Phi}_{\boldsymbol{X}}^k$ is $O(kdn)$.

3.3 Obtaining Maximal Canonical Correlations

Canonical Correlation Analysis (CCA, [6]) is the calculation of pairs of basis vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that the proyections $\boldsymbol{\alpha}^T \boldsymbol{X}$ and $\boldsymbol{\beta}^T \boldsymbol{Y}$ of two random variables $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ are maximally correlated. The correlations between the projected (or canonical) random variables are referred to as canonical correlations, and one can compute up to $\max(\mathrm{rank}(\boldsymbol{X}), \mathrm{rank}(\boldsymbol{Y}))$ of them.

Canonical correlations ρ^2 are the eigenvalues from the set of eigenvalue equations:

$$\begin{aligned} \boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{\alpha} &= \rho^2\boldsymbol{\alpha} \\ \boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{\beta} &= \rho^2\boldsymbol{\beta}, \end{aligned} \tag{7}$$

where $\boldsymbol{C}_{xy} = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})$. Therefore, the largest canonical correlation ρ_1 between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be expressed as the supremum over linear projections respect to the correlation coefficient:

$$\rho_1(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \rho(\boldsymbol{\alpha}^T\boldsymbol{X}, \boldsymbol{\beta}^T\boldsymbol{Y}). \tag{8}$$

If we are not interested in the form of the eigenvectors (as it is in our case), the computational complexity of canonical correlation analysis reduces to $O(\max(\mathrm{rank}(\boldsymbol{X}), \mathrm{rank}(\boldsymbol{Y}))n)$.

3.4 Formal Definition or RDC

Given the random variables $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{Y} \in \mathbb{R}^q$ and a number $k \in \mathbb{N}_+$ we define the Randomized Dependence Coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the largest canonical correlation between the $k-$th order random non-linear projections of their respective marginal observation ranks:

$$\mathrm{rdc}(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \rho(\boldsymbol{\alpha}^T\boldsymbol{\Phi}^k_{P(\boldsymbol{X})}, \boldsymbol{\beta}^T\boldsymbol{\Phi}^k_{P(\boldsymbol{Y})}), \tag{9}$$

where $P(\boldsymbol{X}) := (P_1(X_1), \ldots, P_p(X_p))$ and $P(\boldsymbol{Y}) := (P_1(Y_1), \ldots, P_p(Y_q))$.

4 Properties of RDC

Computational Complexity: In the typical setup (very large n, large d, small k) computations are dominated by the QR matrix factorizations involved in the canonical correlation analysis step. Hence, we achieve a very competitive computational cost of $O((p+q)n\log n + kdn + k^2n) \approx O(k^2n)$.

Ease of Implementation: The five lines of code needed to implement our method in R are included in the Appendix A.

Theorem 5. *The Random Dependence Coefficient is an estimator of the Hirschfeld-Gebelein-Rényi's Maximum Correlation Coefficient, hence satisfying the seven fundamental properties introduced in Section 2.*

Proof. It suffices to prove that the transformations $\hat{f} := \boldsymbol{\alpha}^T\boldsymbol{\Phi}^k_{P(\boldsymbol{X})}$ *and* $\hat{g} := \boldsymbol{\beta}^T\boldsymbol{\Phi}^k_{P(\boldsymbol{Y})}$ *from eq. (9) converge to the true optimal transformations* f *and* g *from eq (1) as the sample size and the number of generated random features tend to infinity. To prove so, we assume that the true transformations* f *and* g *belong to the function class* $\mathcal{F}$ *defined in Theorem 4. Since the CCA step performs optimization over the mixing coefficients of the non-linear random projections* $\boldsymbol{\Phi}^k_{P(\boldsymbol{X})}$ *and* $\boldsymbol{\Phi}^k_{P(\boldsymbol{Y})}$*, our procedure obtains the same convergence to the true optimal transformations* f *and* g *as the one derived in Theorem 4.* □

Relation to Other Estimators: Table 5 provides a comparison betwen RDC and other dependence measures mentioned in Section 1. In the table it is noted, for each estimator, if they allow general for non-linear dependence estimation, if they handle multidimensional random variables and if they are invariant respect to changes in the marginal distributions (monotonic transformations in each of the components of the input random vectors). The computational complexity of each of the estimators is also included.

5 Selecting k

The major drawback of RDC is the selection of its parameter: the number of random non-linear projections k. This is of special interest under the existence of noise, where a large number of random features can prone the estimator to overfit.

In our numerical experiments, we have found that values of $k \in \{1, 5, 10\}$ work very well for most configurations.

Coefficient	Non-Linear	Vector	Marginal Invariant	Complexity
Pearson's ρ	×	×	×	$O(n)$
Spearman's ρ	×	×	✓	$O(n \log n)$
CCA	×	✓	×	$O(d^2 n)$
ACE [3]	✓	×	×	$O(d^2 n^2)$
MIC [11]	✓	×	×	$O(2^n)$
MID [14]	✓	×	×	$O(n \log n)$
dCor [15]	✓	✓	×	$O(n^2)$
HSIC [5]	✓	✓	×	$O(n^2)$
Copula-MMD [8]	✓	✓	✓	$O(n^2)$
RDC	✓	✓	✓	$O(k^2 n)$

Table 1: Comparison between non-linear dependence measures.

6 Experimental Results

We report the results of two numerical experiments that validate the performance of RDC. In all our experiments, we set the number of random features for RDC to $k = 5$, and the random sampling width to $s = 10^{-2}$. Competing kernel methods make use of a Gaussian Kernel with width hyper-parameter set to the median of the euclidean distances between the samples of each of the input random variables.

We first turn to the issue of estimating the *power* of the RDC estimator. We define the *power* of a dependence measure as the percentage of times that it is able to discern between two samples with equal marginals, but only one of them truly containing dependence. In the spirit of Simon and Tibshirani [12], we conducted experiments to estimate the power of RDC as a measure of non-linear dependence. To do so, we studied 8 different association patterns: linear relationship, parabola, cubic polynomial, sinusoidal wave, variable-amplitude sinusoidal wave, logarithm, circle and Gaussian bell. For each of the 8 association patterns, we generated 500 repetitions of 500 samples, in which the input variable was uniformly distributed on the unit interval. Second, we regenerated the input variable randomly, to generate paired independent versions of each sample with equal marginals. Figure 1 shows the *power* for several of the discussed dependence measures as the variance of some Normal distributed additive noise increases from 1/10 to 3. We believe that RDC shows the most consistent behaviour of all estimators, whilst having much lower running time than its best contenders. One could perform similar experiments for multiple dimensional random variables, in which the only competing algorithms could be HSIC and dCor. Copula-MMD did show the same behaviour as HSIC in our experiments.

Figure 2 shows the sample correlations obtained by the coefficients RDC, ACE, dCor, MIC, Pearson's ρ, Spearman's rank and Kendall's τ for 28 different associations of two scalar random variables. RDC seems able to capture all the proposed dependencies, whilst scoring lower values for the independent configurations (which are the number 4, 21 and 28).

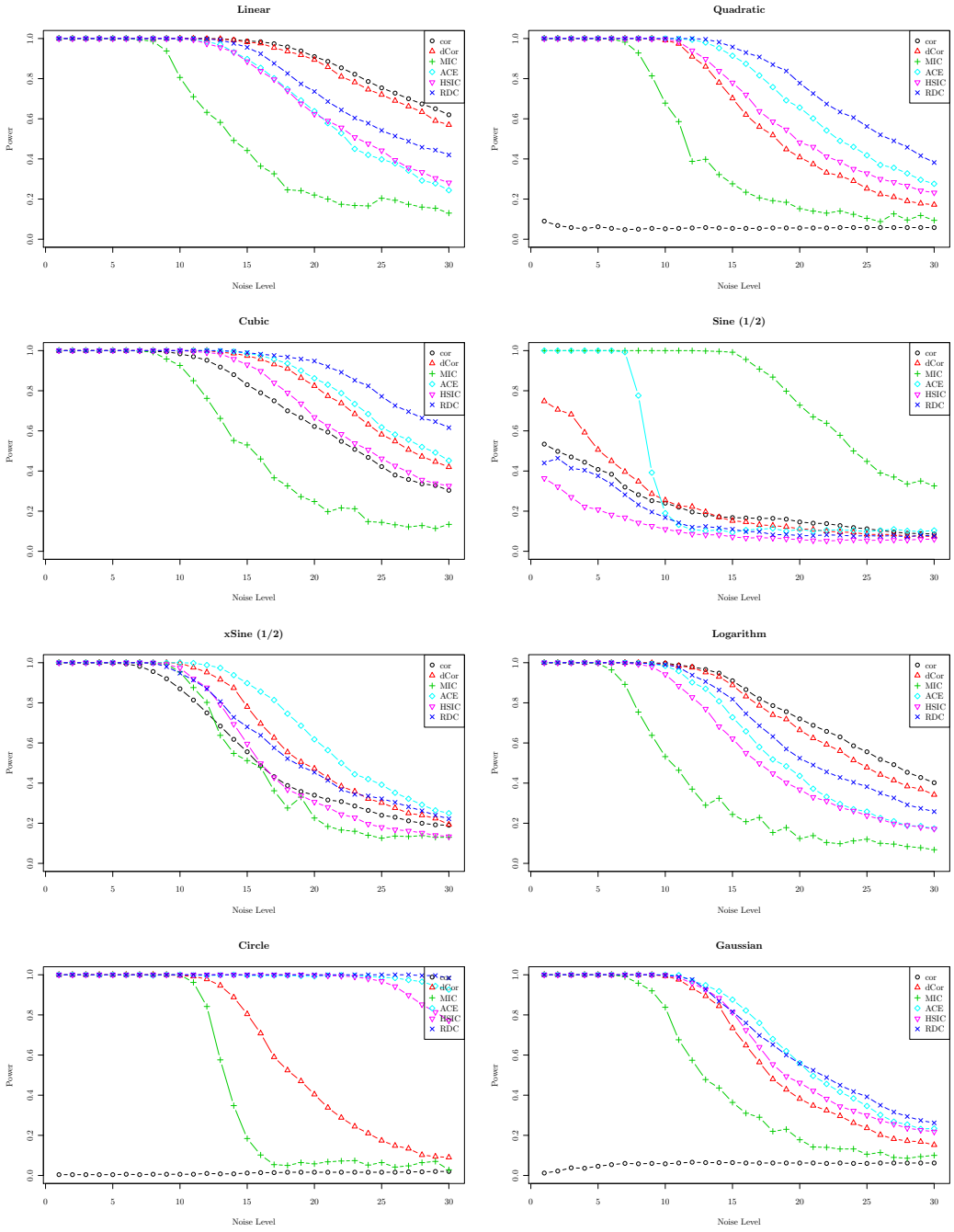

Figure 1: Power of estimators, i.e., percentage of times that they are able to distinguish between dependent and independent samples with equal marginals. Experiments based on [12].

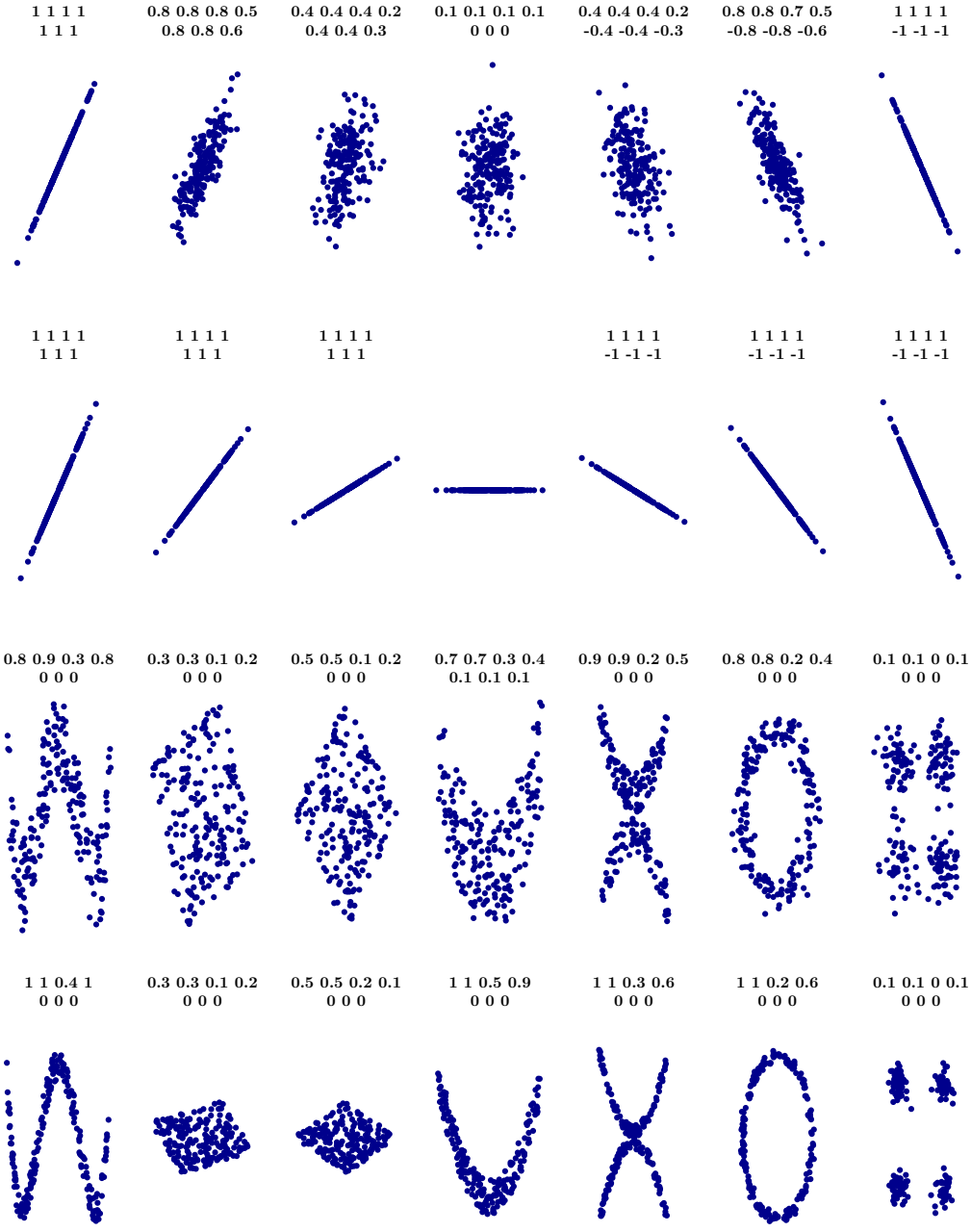

Figure 2: Estimates of RDC ($k = 10$), ACE, dCor, MINE, Pearson's ρ, Spearman's rank and Kendall's τ for different association patterns. Experiments based on [1].

A R Source Code

```
rdc  <- function(x,y,k) {
   x <- apply(as.matrix(x),2,function(u) ecdf(u)(u))
   y <- apply(as.matrix(y),2,function(u) ecdf(u)(u))
   cancor(plogis(x%*%matrix(rnorm(ncol(x)*k),ncol(x),k)),
          plogis(y%*%matrix(rnorm(ncol(y)*k),ncol(y),k)))$cor[1]
}
```

References

[1] Correlation and dependence - Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Correlation_and_dependence`, 2013.

[2] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.

[3] L. Breiman and J. H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[4] H. Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *Zeitschrift fr Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

[5] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th international conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag, 2005.

[6] Wolfgang K. Härdle and Leopold Simar. *Applied Multivariate Statistical Analysis*. Springer, 2nd edition, 2007.

[7] R. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics, 2nd edition, 2006.

[8] B. Poczos, Z. Ghahramani, and J. Schneider. Copula-based kernel dependency measures. In *ICML*, 2012.

[9] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *NIPS*, 2008.

[10] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:441–451, 1959.

[11] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[12] N. Simon and R. Tibshirani. Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et. al (Science Dec. 2011). *`http://www-stat.stanford.edu/~tibs/reshef/comment.pdf`*, 2011.

[13] A. Sklar. Fonctions de repartition à n dimension set leurs marges. *Publ. Inst. Statis. Univ. Paris*, 8(1):229–231, 1959.

[14] M. Sugiyama and K. M. Borgwardt. Measuring statistical dependence via the mutual information dimension. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages XX–XX, Beijing, China, August 2013.

[15] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2007.

[16] Gábor J. Székely and Maria L. Rizzo. Rejoinder: Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1303–1308, 2009.

[17] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.