The Randomized Dependence Coefficient

David Lopez-Paz[1,2], Philipp Hennig[1], Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems; [2]University of Cambridge

What defines a good measure of dependence?

In 1959 [4], Alfréd Rényi argued that a measure of dependence $\rho^* : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ between random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ should satisfy seven fundamental properties:

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 $(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\rho^*(X, Y) = |\rho(X, Y)|$, where ρ is the correlation coefficient.

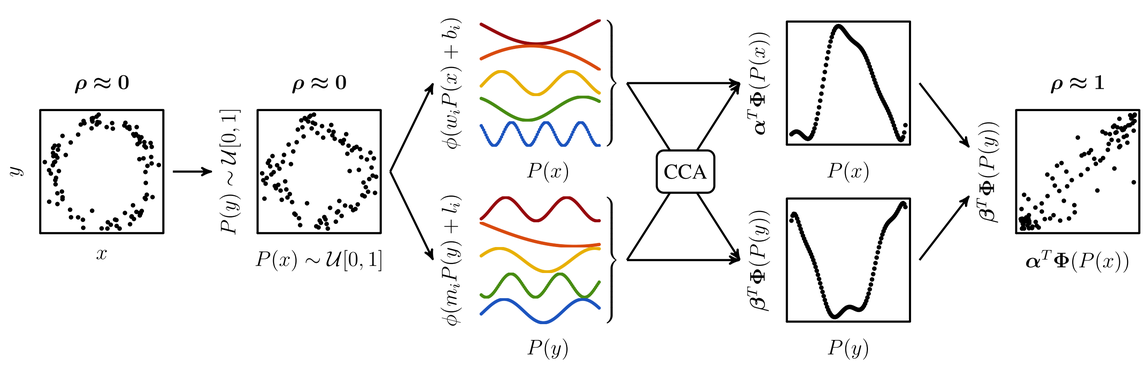
Rényi also showed the *Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient* (HGR) [4] to satisfy all these properties. HGR was defined by Gebelein in 1941 as the supremum of Pearson's correlation coefficient ρ over all Borel-measurable functions f, g of finite variance:

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

Since this supremum is over an infinite-dimensional space, HGR is not computable. It is an abstract concept, not a practical dependence measure. **In the following we propose a scalable estimator with the same structure as HGR: the Randomized Dependence Coefficient.**

Building the RDC Statistic

The *Randomized Dependence Coefficient* (RDC) measures the dependence between random samples $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{q \times n}$ as the largest canonical correlation of k non-linear random projections of their copulas:



Step 1: Estimation of the Copulas

To achieve invariance with respect to transformations on marginal distributions (such as shifts or rescalings), we operate on the empirical copula transformation of the data [2]. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with continuous marginal cumulative distribution functions (cdfs) P_i, $1 \leq i \leq d$. Then the vector $\boldsymbol{U} = (U_1, \ldots, U_d) := \boldsymbol{P}(\boldsymbol{X}) = (P_1(X_1), \ldots, P_d(X_d))$, known as the copula transformation. In our experiments, we compute empirical cdfs to estimate the copula. This has a cost of $O(dn \log n)$.

Step 2: Generation of Non-Linear Random Projections

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

$$\boldsymbol{\Phi}(\boldsymbol{X}; k, s) := \begin{pmatrix} \phi(\boldsymbol{w}_1^T \boldsymbol{x}_1 + b_1) & \cdots & \phi(\boldsymbol{w}_k^T \boldsymbol{x}_1 + b_k) \\ \vdots & \vdots & \vdots \\ \phi(\boldsymbol{w}_1^T \boldsymbol{x}_n + b_1) & \cdots & \phi(\boldsymbol{w}_k^T \boldsymbol{x}_n + b_k) \end{pmatrix}^T, \quad \boldsymbol{w}_i, b_i \sim \mathcal{N}(0, s), \quad \phi(x) = \sin(x). \tag{1}$$

the $k-$th order non-linear random projection of $\boldsymbol{X}$. Rahimi and Recht [3] proved that these projections have large non-linear modeling power. Computing them has a cost of $O(kn \log(d))$ [1].

Step 3: Computation of Canonical Correlations

The final step of RDC is to search for the linear combination of the augmented empirical copula transformations that has maximal correlation. This is the largest canonical correlation, and gives birth to RDC:

$$\mathrm{rdc}(\boldsymbol{X}, \boldsymbol{Y}; k, s) := \sup_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \rho \left(\boldsymbol{\alpha}^T \boldsymbol{\Phi}(\boldsymbol{P}(\boldsymbol{X}); k, s), \boldsymbol{\beta}^T \boldsymbol{\Phi}(\boldsymbol{P}(\boldsymbol{Y}); k, s) \right).$$

When $k \gg n$, the cost of CCA is $O(k^2 n)$. Hence, we achieve a cost in terms of the sample size of $O(n \log n)$.

Properties of RDC

Name of Coeff.	Non-Linear	Vector Inputs	Marginal Invariant	Renyi's Properties	Coeff. $\in [0, 1]$	Number Param.	Comp. Cost
Pearson's ρ	×	×	×	×	✓	0	n
Spearman's ρ	×	×	✓	×	✓	0	$n \log n$
Kendall's τ	×	×	✓	×	✓	0	$n \log n$
CCA	×	✓	×	×	✓	0	n
KCCA	✓	✓	×	×	✓	1	n^3
ACE	✓	×	×	✓	✓	1	n
MIC	✓	×	×	×	✓	1	$n^{1.2}$
dCor	✓	✓	×	×	✓	1	n^2
HSIC	✓	✓	×	×	×	1	n^2
CHSIC	✓	✓	✓	×	×	1	n^2
RDC	✓	✓	✓	✓	✓	2	$n \log n$

Rate of Convergence to HGR with $f, g \in \mathcal{F}$

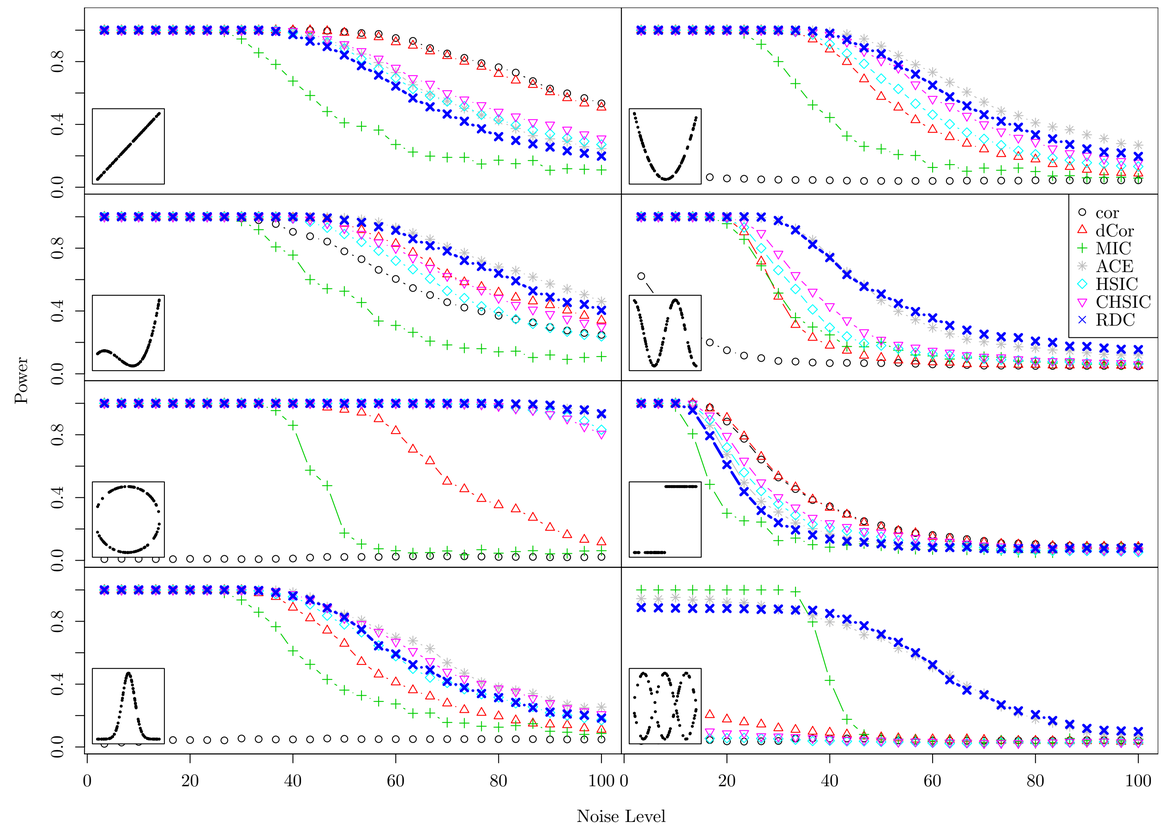
$$\mathrm{hgr}(\boldsymbol{X}, \boldsymbol{Y}; \mathcal{F}) - \mathrm{rdc}(\boldsymbol{X}, \boldsymbol{Y}; k) = O\left(\left(\frac{\|\boldsymbol{m}\|_F}{\sqrt{n}} + \frac{LC}{\sqrt{k}} \right) \sqrt{\log \frac{1}{\delta}} \right), \quad \boldsymbol{m} := \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \boldsymbol{\beta}\boldsymbol{\beta}^T.$$

Experimental Results

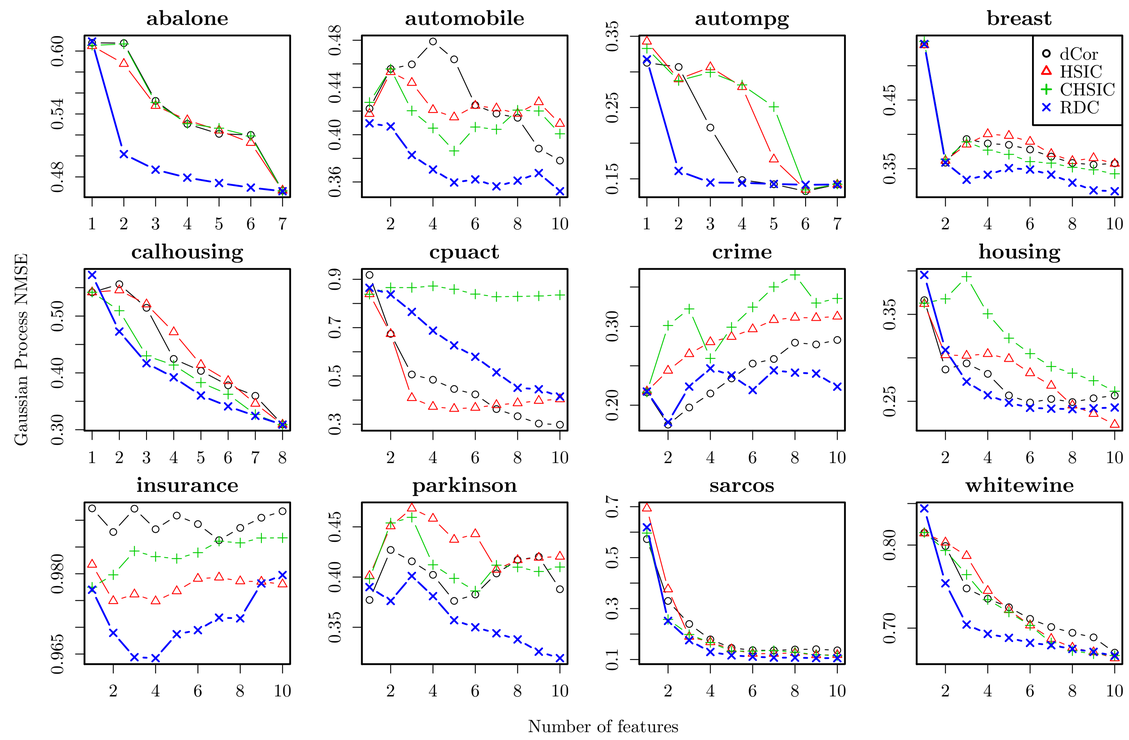
Running Times on Univariate Samples

sample size	Pearson's ρ	RDC	ACE	KCCA	dCor	HSIC	CHSIC	MIC
1,000	0.0001	0.0047	0.0080	0.402	0.3417	0.3103	0.3501	1.0983
10,000	0.0002	0.0557	0.0782	3.247	59.587	27.630	29.522	—
100,000	0.0071	0.3991	0.5101	43.801	—	—	—	—
1,000,000	0.0914	4.6253	5.3830	—	—	—	—	—

Statistical Power with Increasing Additive Gaussian Noise

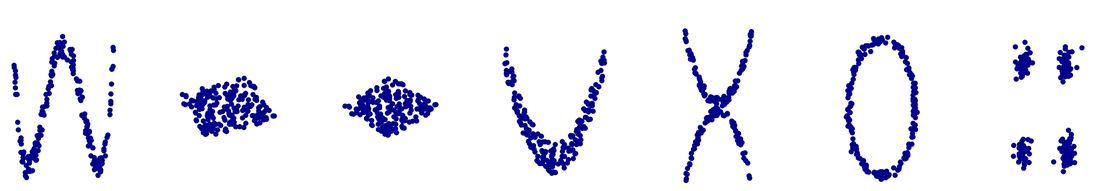


Greedy Feature Selection as Dependence Maximization



Examples of RDC, ACE, dCor, MIC, Pearson, Spearman, Kendall $[0, 1]-$scores

1.0	1.0	0.4	1.0	0.3	0.3	0.1	0.2	0.5	0.5	0.1	0.2	1.0	1.0	0.5	0.9	1.0	1.0	0.3	0.6	1.0	1.0	0.2	0.6	0.1	0.1	0.0	0.1
	0.0	0.0	0.0		0.0	0.0	-0.0		0.0	0.0	0.0		0.0	0.0	0.0		0.1	0.1	0.1		-0.0	-0.0	-0.0		-0.0	-0.0	-0.0



RDC Source Code in R

```
rdc <- function(x,y,k=20,s=1/6,f=sin) {
  x <- cbind(apply(as.matrix(x),2,function(u)rank(u)/length(u)),1)
  y <- cbind(apply(as.matrix(y),2,function(u)rank(u)/length(u)),1)
  x <- s/ncol(x)*x%*%matrix(rnorm(ncol(x)*k),ncol(x))
  y <- s/ncol(y)*y%*%matrix(rnorm(ncol(y)*k),ncol(y))
  cancor(cbind(f(x),1),cbind(f(y),1))$cor[1]
}
```

References

[1] Q. Le, T. Sarlos, and A. Smola. Fastfood – Approximating kernel expansions in loglinear time. In *ICML*, 2013.

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

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

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