The Randomized Dependence Coefficient

http://papers.nips.cc/paper/5138-the-randomized-dependence-coefficient.pdf

▼ What did the authors try to accomplish?

- Introduce the Randomized Dependence Coefficient (RDC), a measure of nonlinear dependence between random variables of arbitrary dimension based on the Hirschfeld-Gebelein-Renyi Maximum Correlation Coefficient

▼ What were the key elements of the approach?

- Defines dependence between two random variables as the largest canonical correlation between random non-linear projections of their respective empirical copula-transformations. RDC is invariant to monotonically increasing transformations, operates on random variables of arbitrary dimension, and has computational cost of O(n log n)

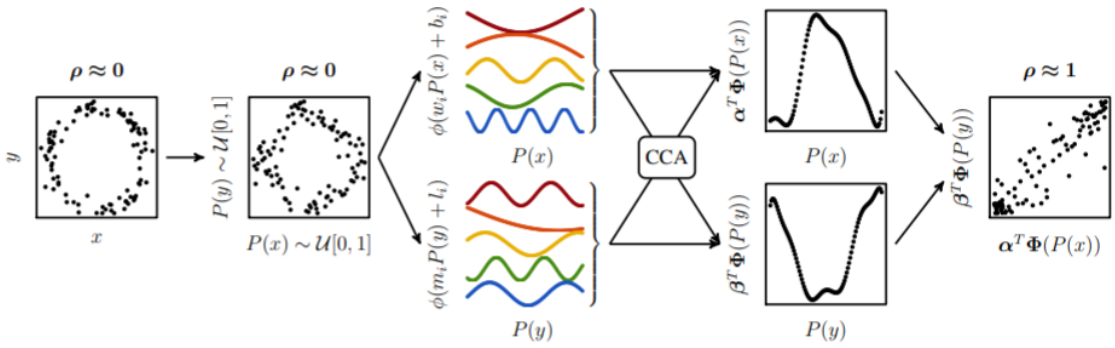

Figure 1: RDC computation for a simple set of samples $\{(x_i, y_i)\}_{i=1}^{100}$ drawn from a noisy circular pattern: The samples are used to estimate the copula, then mapped with randomly drawn non-linear functions. The RDC is the largest canonical correlation between these non-linear projections.

1. Estimation of Copula-Transformations

- 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. Generation of Random Non-Linear Projections

- Augment the empirical copula transformations with non-linear projections, so that linear methods can subsequently be used to capture non-linear dependencies on the original data
- The choice of the non-linearities φ : R → R is the main and unavoidable assumption in RDC
- The only way to favour one such family and distribution over another is to use prior assumptions about which kind of distributions the method will typically have to analyse.
- We use random features instead of the Nystrom method because of their smaller memory and computation requirements. In our experiments, we will use sinusoidal projections, φ(wT x + b) := sin(wT x + b)

3. Computation of Canonical Correlations

- compute the linear combinations of the augmented empirical copula transformations that have maximal correlation. Canonical Correlation Analysis (CCA, [7]) is the calculation of pairs of basis vectors (α, β) such that the projections αT X and β T Y of two random samples X ∈ R^(p×n) and Y ∈ R^(q×n) are maximally correlated

Formal definition of RDC

Given the random samples $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{q\times n}$ and the parameters $k \in \mathbb{N}_+$ and $s \in \mathbb{R}_+$, the Randomized Dependence Coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as:

$$\mathrm{rdc}(\boldsymbol{X}, \boldsymbol{Y}; k, s) := \sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \rho\left(\boldsymbol{\alpha}^T \boldsymbol{\Phi}(\boldsymbol{P}(\boldsymbol{X}); k, s), \boldsymbol{\beta}^T \boldsymbol{\Phi}(\boldsymbol{P}(\boldsymbol{Y}); k, s)\right). \quad (9)$$

▼ What can you use yourself?

- Randomized Dependence Coefficient
- Implementation: https://github.com/garydoranjr/rdc

▼ What other references do you want to follow?

- **Distance Correlation:** G. J. Szekely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. Annals of Statistics, 35(6), 2007.
- **Brownian Correlation:** G. J. Szekely and M. L. Rizzo. Rejoinder: Brownian distance covariance. ´ Annals of Applied

Statistics, 3(4):1303–1308, 2009.