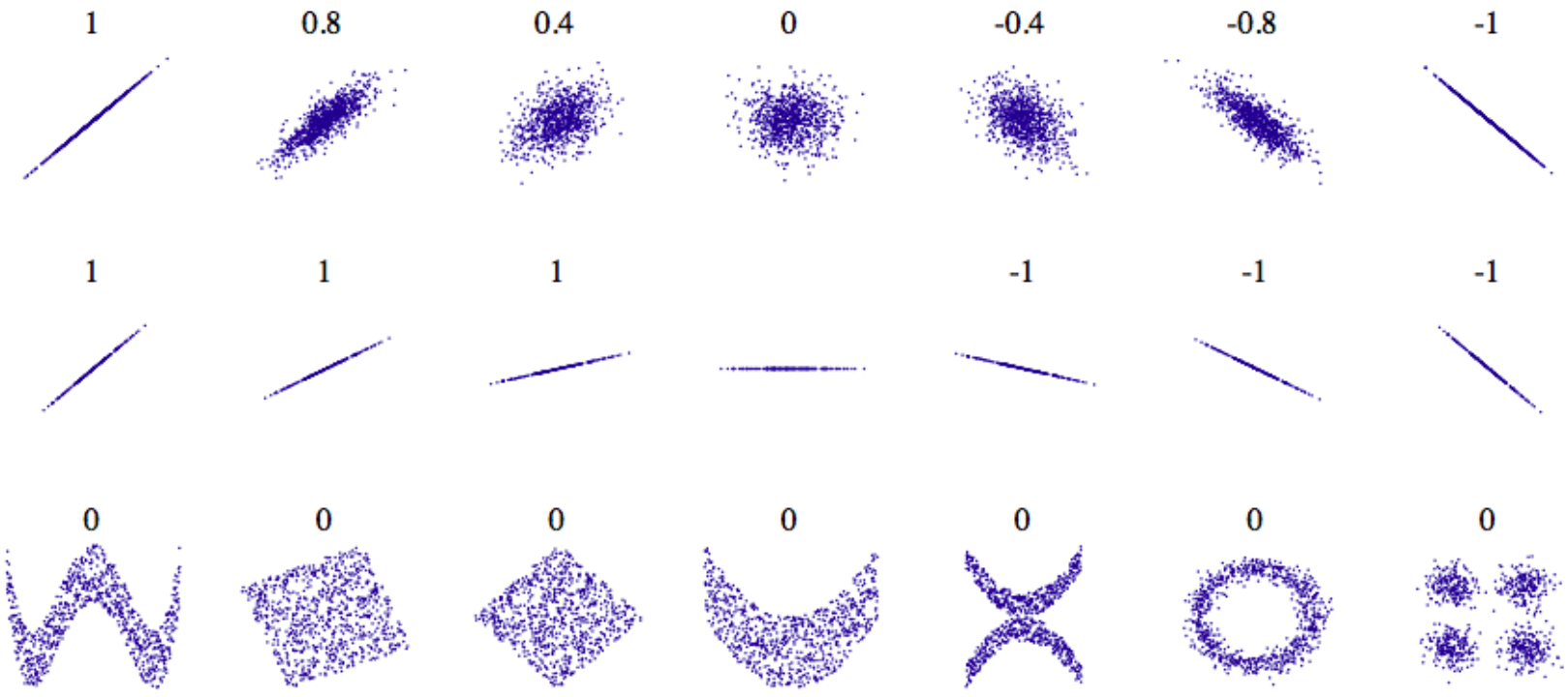
The Randomized Dependence Coefficient

Lopez–Paz, Hennig & Schölkopf

December 2013

Correlation versus Dependence



Alfred Rényi's Properties (1959)

1. $\rho^*(X, Y)$ defined for any non-constant rr.vv. X and Y.
2. $\rho^*(X, Y) = \rho^*(Y, X)$
3. $0 \leq \rho^*(X, Y) \leq 1$
4. $\rho^*(X, Y) = 0 \iff X \perp Y$.
5. For bijective $f, g : \mathbb{R} \to \mathbb{R}$, $\rho^*(X, Y) = \rho^*(f(X), g(Y))$.
6. $\rho^*(X, Y) = 1$ if for f or g, $Y = f(X)$ or $X = g(Y)$.
7. If $(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\rho^*(X, Y) = |\rho(X, Y)|$.

The *Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient*

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

satisfies all these properties.

Alfred Rényi's Properties (1959)

1. $\rho^*(X, Y)$ defined for any non-constant rr.vv. X and Y.
2. $\rho^*(X, Y) = \rho^*(Y, X)$
3. $0 \leq \rho^*(X, Y) \leq 1$
4. $\rho^*(X, Y) = 0 \iff X \perp Y$.
5. For bijective $f, g : \mathbb{R} \to \mathbb{R}$, $\rho^*(X, Y) = \rho^*(f(X), g(Y))$.
6. $\rho^*(X, Y) = 1$ if for f or g, $Y = f(X)$ or $X = g(Y)$.
7. If $(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\rho^*(X, Y) = |\rho(X, Y)|$.

The *Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient*

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

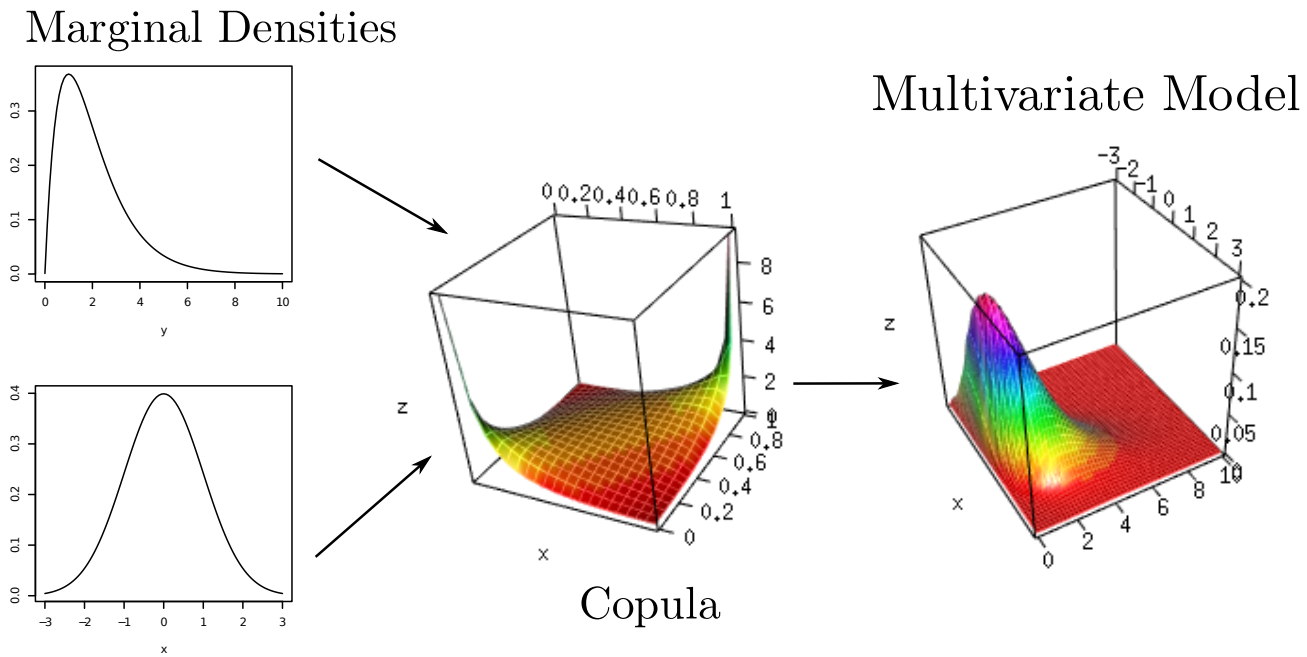
satisfies all these properties. But is **intractable**. We propose *The Randomized Dependence Coefficient*. It has **3 ingredients**.

Ingredient 1: Copulas (Sklar, 1959)

A density p on $\mathbf{x} \in \mathbb{R}^d$ can be uniquely factorized as:

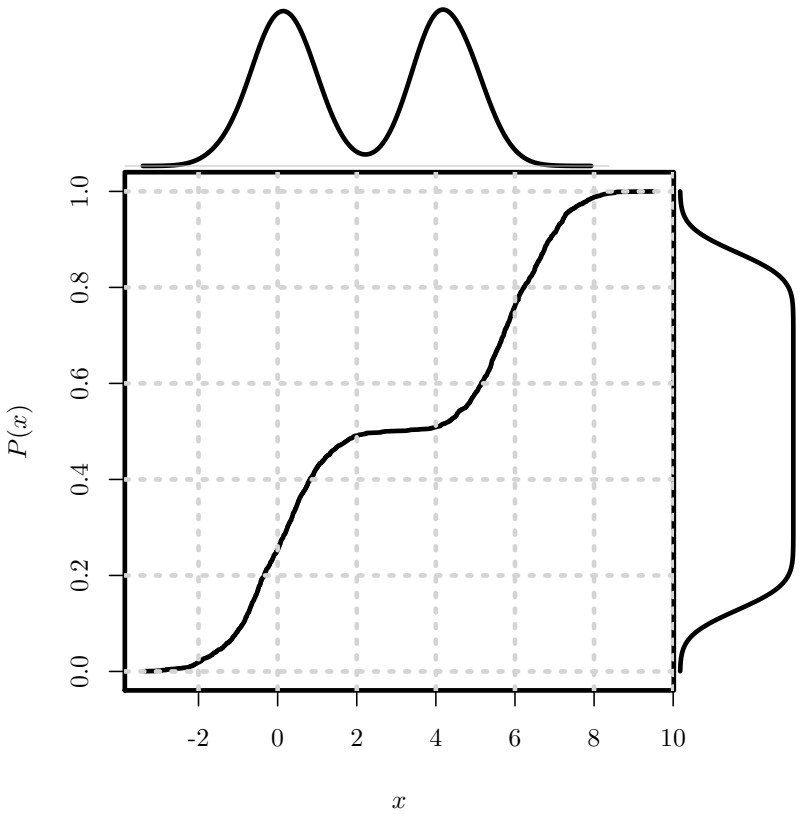
$$p(\mathbf{x}) = \underbrace{c(P_1(x_1), \ldots, P_d(x_d))}_{\text{copula}} \underbrace{\prod_{i=1}^{d} p_i(x_i)}_{\text{marginals}} .$$

Copulas separate the modeling of the marginal densities from the modeling of the dependence structure that links them together.



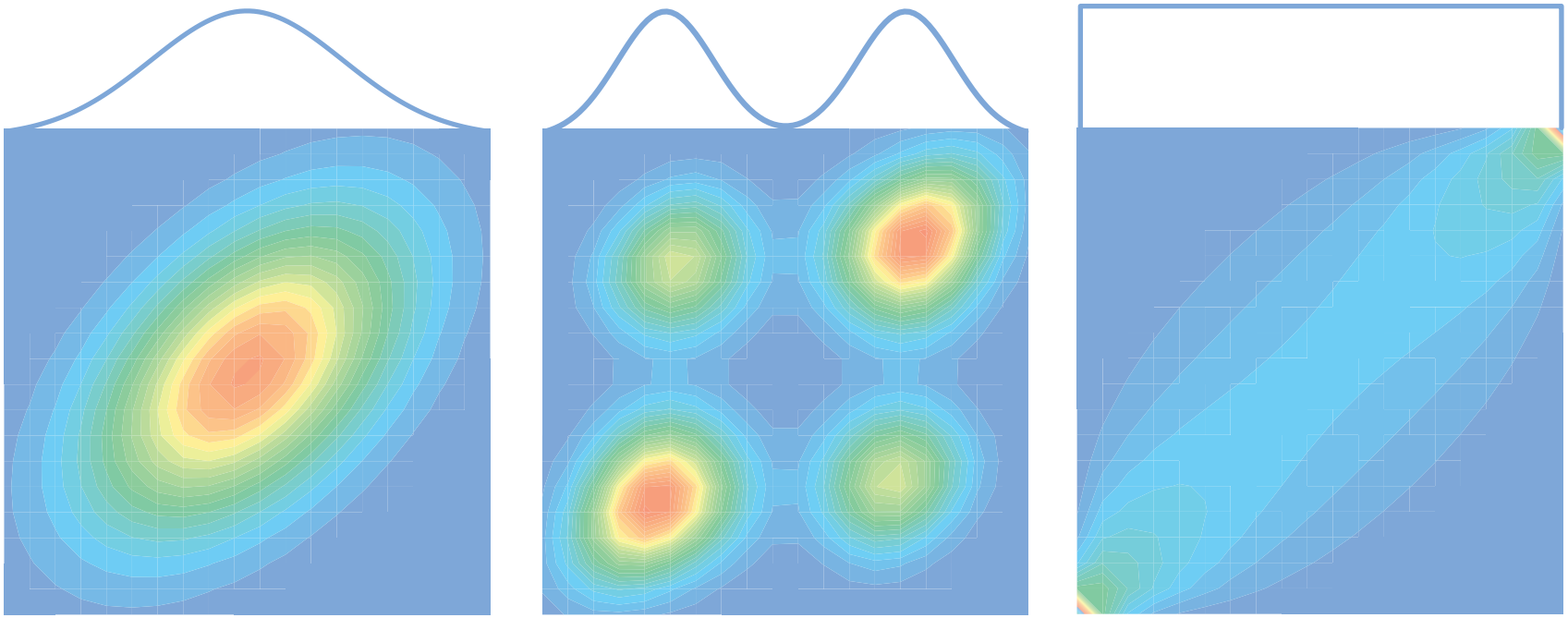
Ingredient 1: Copulas (Sklar, 1959)

$$p(\mathbf{x}) = c(P_1(x_1), \ldots, P_d(x_d)) \prod_{i=1}^{d} p_i(x_i)$$



This is a core concept in *Optimal Transportation*.

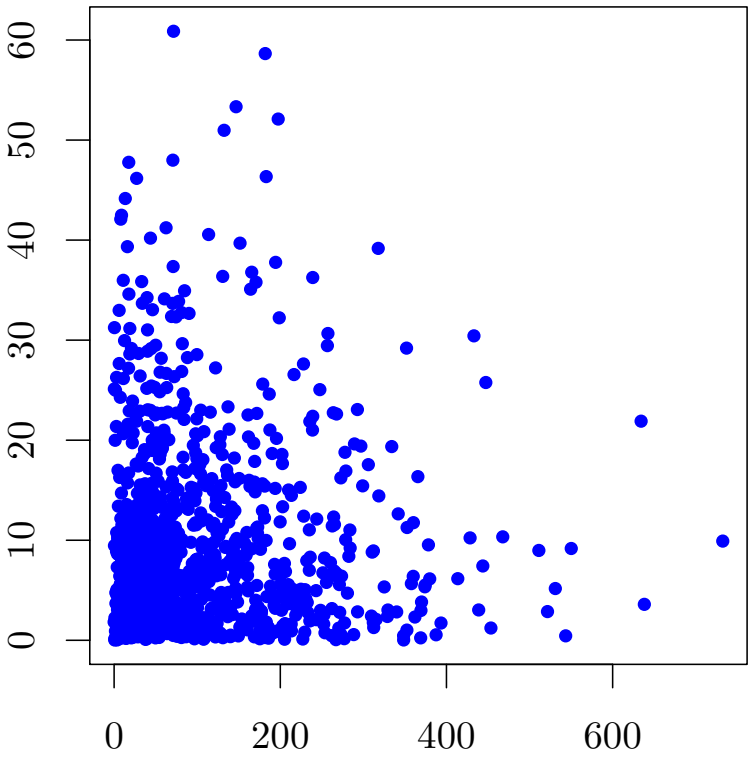
Ingredient 1: Copulas (Sklar, 1959)

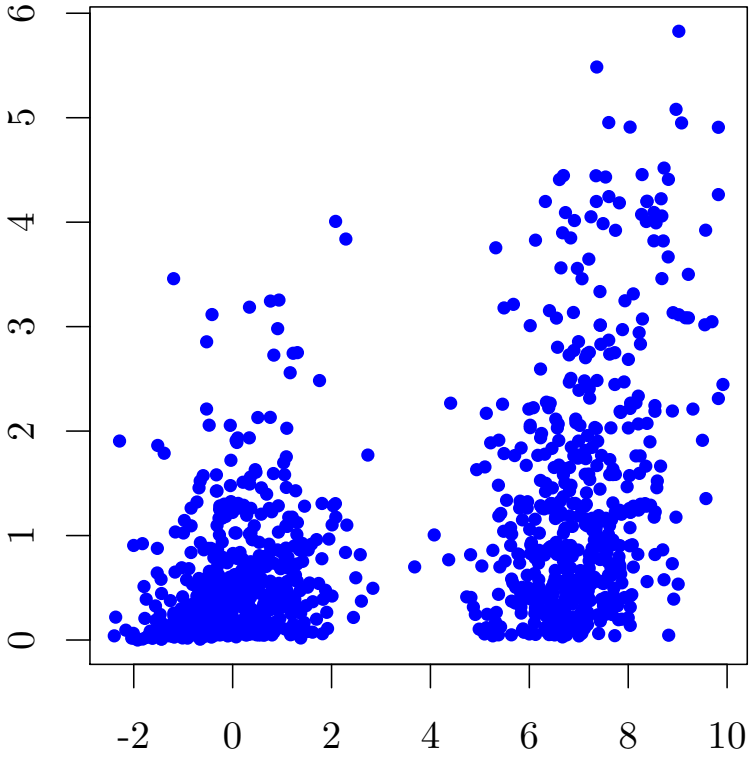


Three different bivariate densities with the same underlying Copula (Gaussian, $\rho = 0.8$)

Ingredient 1: Copulas (Sklar, 1959)

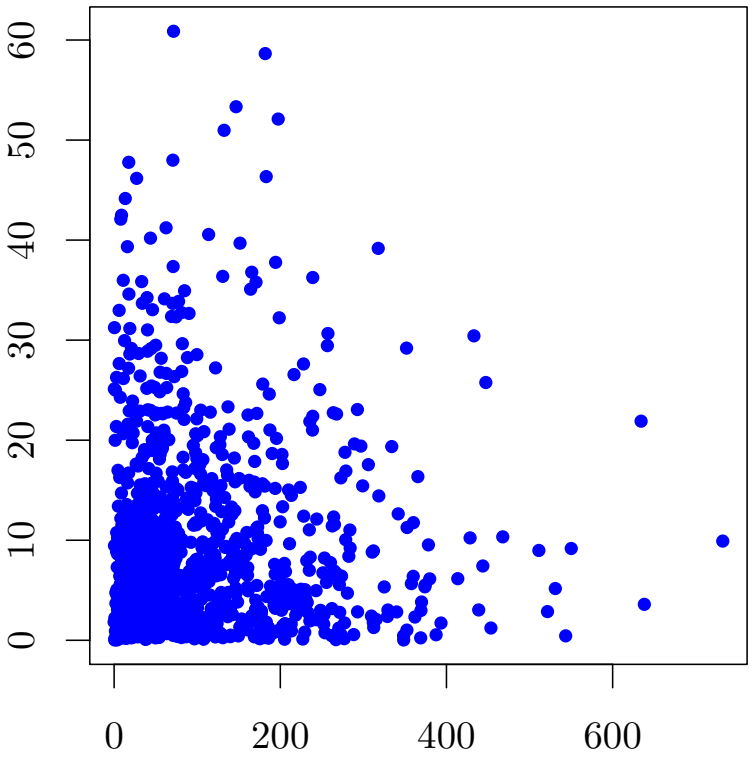
Are X and Y independent?



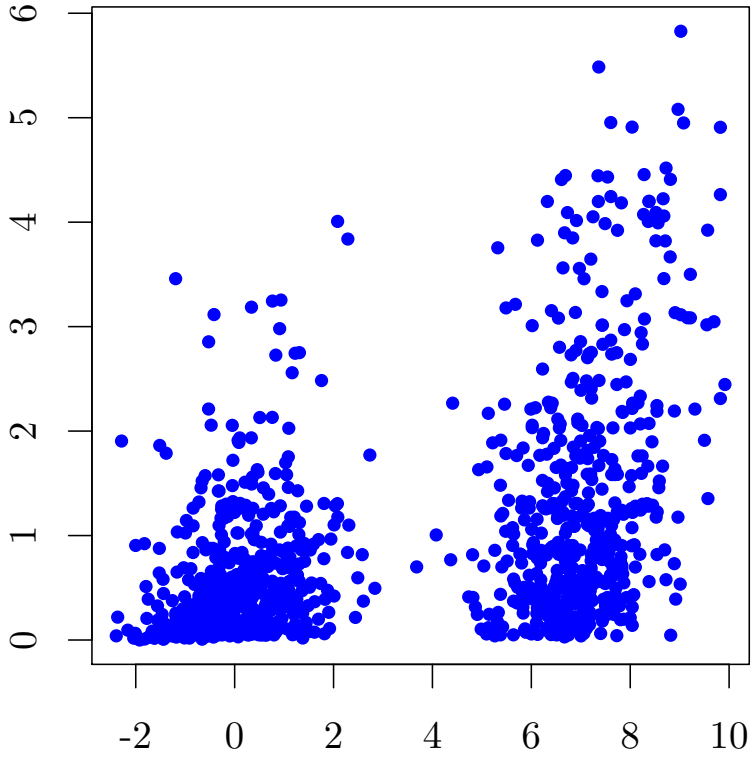


Ingredient 1: Copulas (Sklar, 1959)

Are X and Y independent?



Yes



No

Ingredient 1: Copulas (Sklar, 1959)

Estimate samples from the copula using each empirical CDF $\hat{P}_i$:

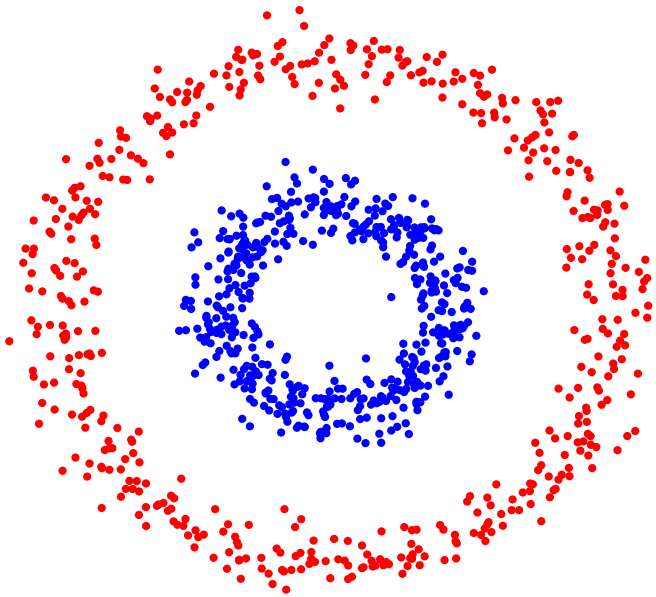
$$P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x),$$
$$\hat{\mathbf{P}}(\mathbf{X}) = [\hat{P}_1(X_1), \ldots, \hat{P}_d(X_d)].$$

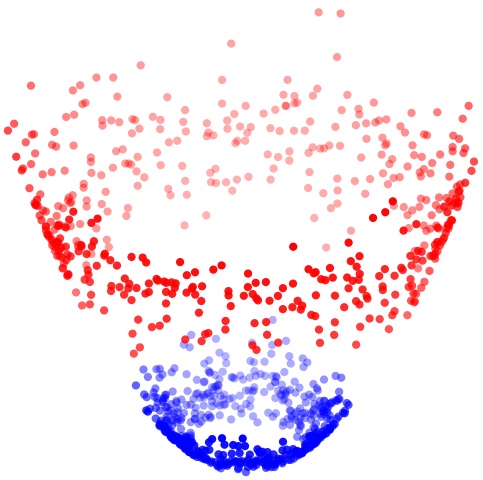
Given that each of the d marginals is transformed independently, we achieve good convergence to the true copula samples[1]:

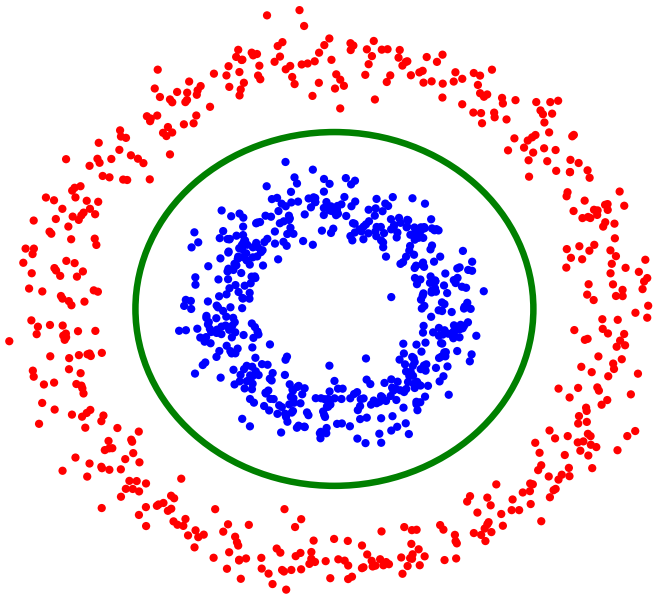
$$\Pr\left(\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{P}(\mathbf{x}) - \hat{\mathbf{P}}(\mathbf{x})\|_2 > \epsilon\right) \leq 2d \exp\left(-\frac{2m\epsilon^2}{d}\right).$$

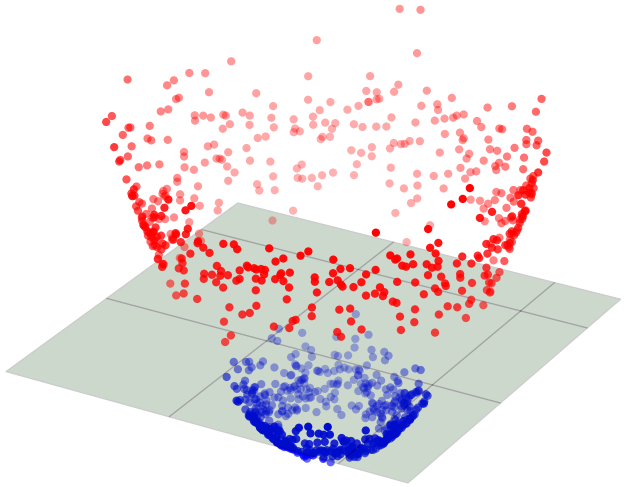
This computations are parallelizable and take $O(dn \log n)$.

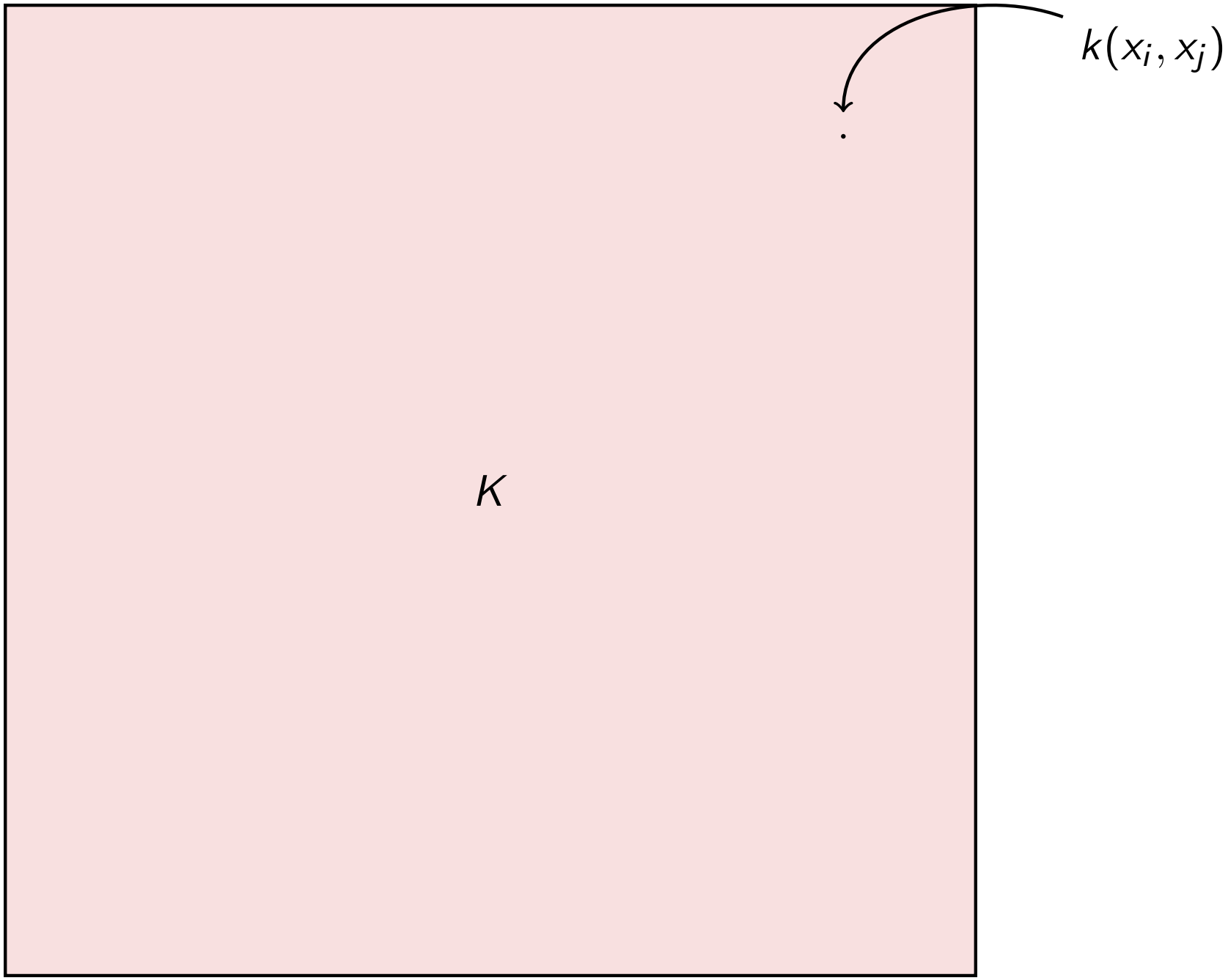
[1] Glivenko-Cantelli (1933) $\rightarrow$ Kiefer-Dvoretzky-Wolfowitz (1956) $\rightarrow$ Massart (1990)










$k(x_i, x_j)$
K

Ingredient 2: Kitchen Sinks (Rahimi & Recht, 2008)

$$\text{Approximate: } f(\mathbf{x}) = \int_{\Omega} \alpha(\mathbf{w})\phi(\mathbf{x};\mathbf{w})\mathrm{d}\mathbf{w}$$

$$\text{Using: } f_T(\mathbf{x}) = \sum_{i=1}^{T} \alpha_i \phi(\mathbf{x};\mathbf{w}_i)$$

Consider the loss $\|f - f_T\|_\mu = \sqrt{\int_{\mathcal{X}} (f_T(x) - f(x))^2 \mu(dx)}$

Greedy Fitting

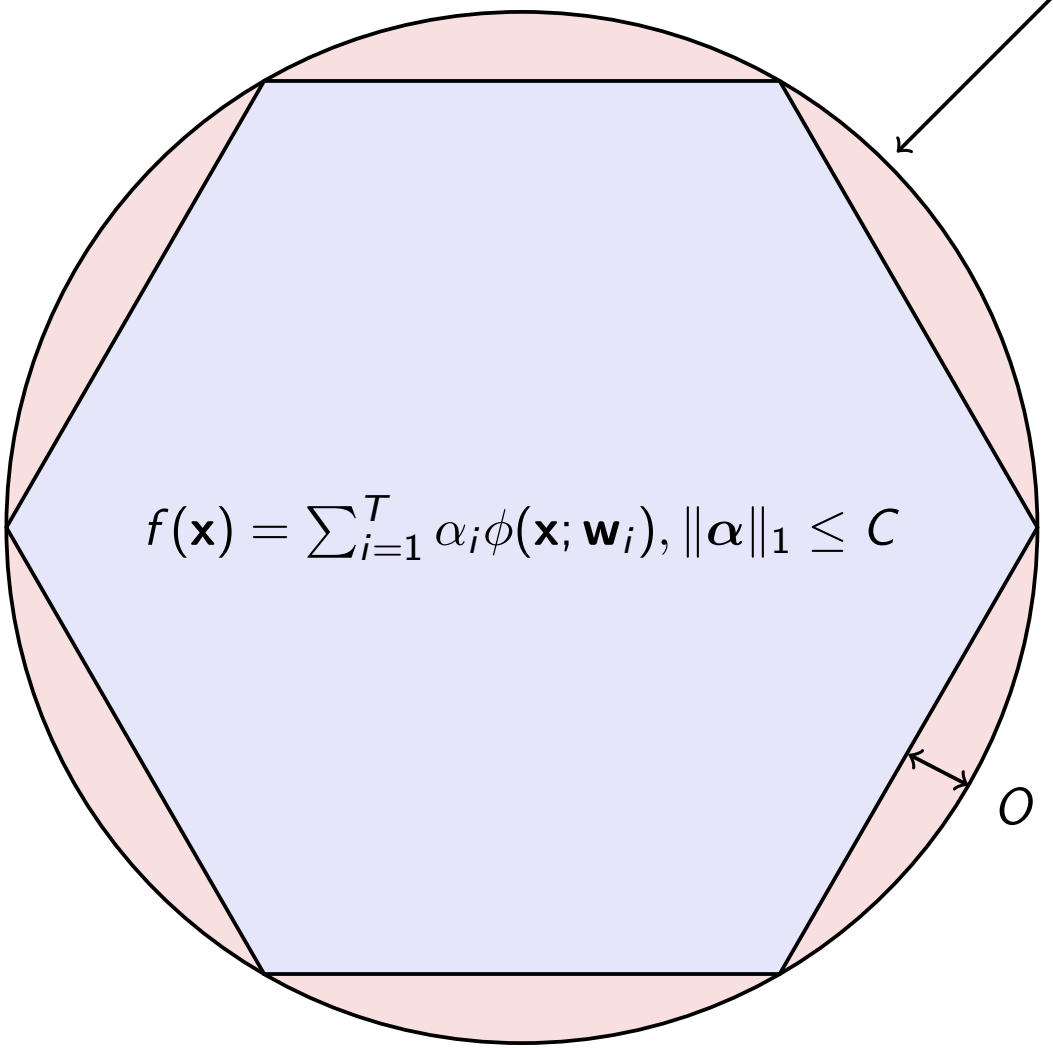
$$(\boldsymbol{\alpha}^\star, \mathbf{W}^\star) = \min_{\boldsymbol{\alpha},\mathbf{W}} \left\| \sum_{i=1}^{T} \alpha_i \phi(;\mathbf{w}_i) - f \right\|_\mu$$

Kitchen Sinks

$$\mathbf{w}_i^\star, \ldots, \mathbf{w}_T^\star \sim p(\mathbf{w}), \quad \boldsymbol{\alpha}^\star = \min_{\boldsymbol{\alpha}} \left\| \sum_{i=1}^{T} \alpha_i \phi(;\mathbf{w}_i^\star) - f \right\|_\mu$$

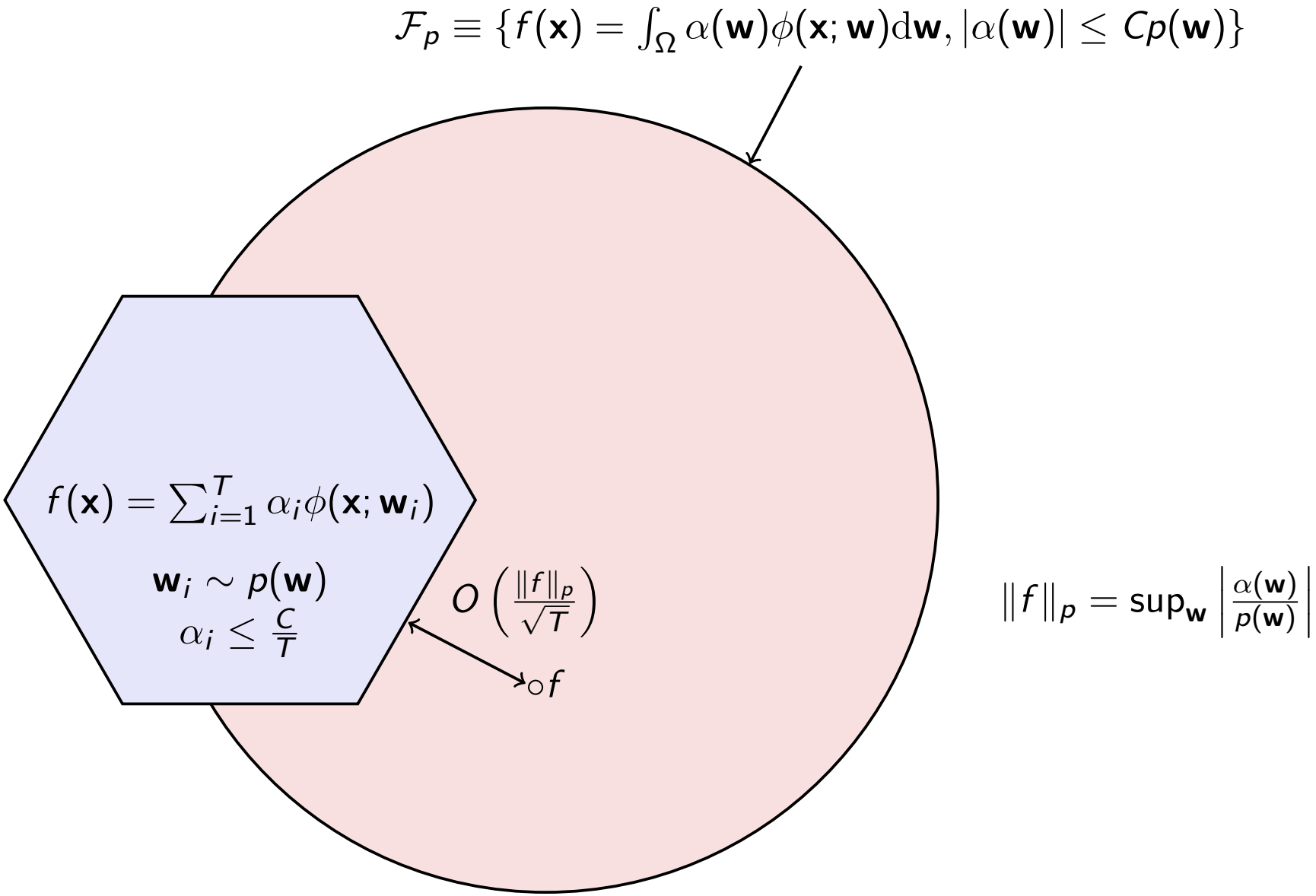
Greedy Approximation of Functions

$$\mathcal{F} \equiv \{f(\mathbf{x}) = \int_\Omega \alpha(\mathbf{w})\phi(\mathbf{x};\mathbf{w})\mathrm{d}\mathbf{w}, |\boldsymbol{\alpha}(\mathbf{w})| \leq C\}$$

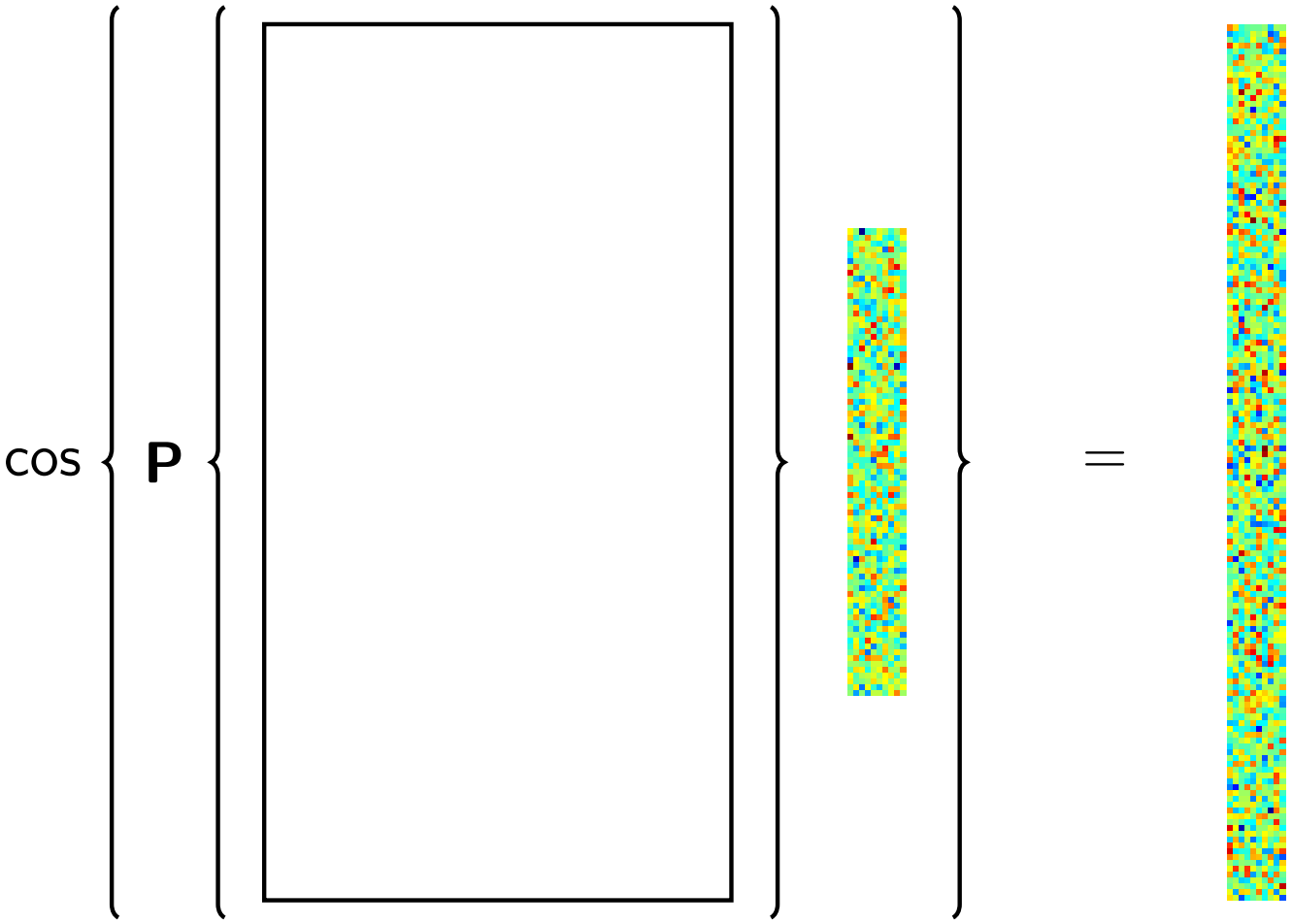


$$\|f_T - f\|_\mu = \sqrt{\int_{\mathcal{X}} (f_T(x) - f(x))^2 \mu(dx)} = O\left(\frac{C}{\sqrt{T}}\right) \text{ (Jones, 1992)}$$

Randomized Approximation of Functions



The picture so far



$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\quad$ $\mathbf{W} \in \mathbb{R}^{d \times k}$ $\quad$ $\mathbf{\Phi}(\mathbf{P}(\mathbf{X}); k, s) \in \mathbb{R}^{n \times k}$

$\mathbf{W}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{s}I)$

When $\mathbf{W}$ is random we can multiply in $O(nk \log d)$! (Sarlos, 2006)

Ingredient 3: CCA (Hotelling, 1936)

CCA finds $\mathbf{A} \in \mathbb{R}^{p \times r}$, $\mathbf{B} \in \mathbb{R}^{q \times r}$ s.t. $r = \max(\mathrm{rank}(\mathbf{X}), \mathrm{rank}(\mathbf{Y}))$,

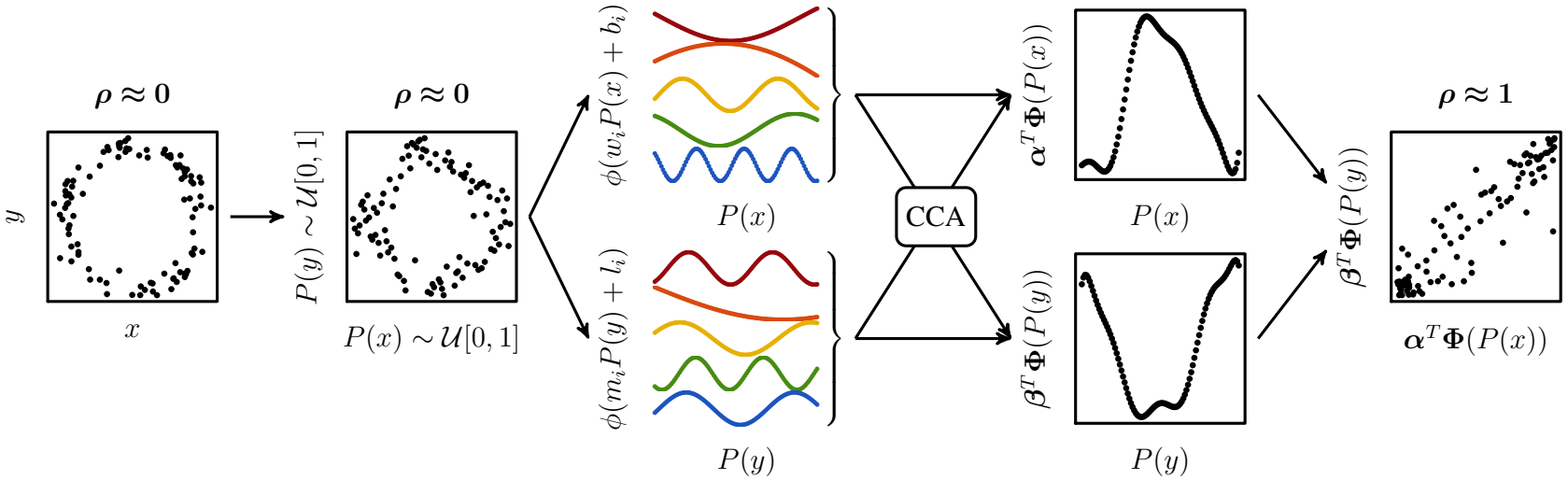
$$\begin{aligned}
&\rho(\mathbf{X}\mathbf{A}_i, \mathbf{Y}\mathbf{B}_i) \text{ is maximized,} \\
&\rho(\mathbf{X}\mathbf{A}_i, \mathbf{Y}\mathbf{B}_j) = 0 \text{ for all } i \neq j, \\
&\rho(\mathbf{X}\mathbf{A}_i, \mathbf{X}\mathbf{A}_j) = 0 \text{ for all } i \neq j \text{ and} \\
&\rho(\mathbf{Y}\mathbf{B}_i, \mathbf{Y}\mathbf{B}_j) = 0 \text{ for all } i \neq j.
\end{aligned}$$

It does so by solving the eigen-problem:

$$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \rho^2 \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}.$$

This involves computing two QR decompositions and one truncated SVD. Complexity of $O(k^2 n)$.

The Randomized Dependence Coefficient



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

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

Closest relative: KCCA (Bach & Jordan, 2002)

Other Properties

Convergence to HGR when restricted to $\mathcal{F}_p$

$$\mathrm{hgr}(\mathbf{X}, \mathbf{Y}; \mathcal{F}_p) - \mathrm{rdc}(\mathbf{X}, \mathbf{Y}; k) = O\left(\left(\frac{\|\mathbf{m}\|_F}{\sqrt{n}} + \frac{LC}{\sqrt{k}}\right)\sqrt{\log\frac{1}{\delta}}\right)$$

Total computational complexity

$$O(\underbrace{(p+q)n\log n}_{\text{copulas}} + \underbrace{kn\log(p+q)}_{\text{kitchen sinks}} + \underbrace{k^2 n}_{\text{CCA}}) \approx \underbrace{O(n\log n)}_{\text{w.r.t. sample size}}$$

Limiting behaviours

As $s \to 0$, converges to Spearman's rank.
As $k \to \infty$, does it converge to "Copularized" kCCA

Implementation

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]
}
```

MATLAB

```
function r = rdc(x,y,k,s)
  n = size(x,1);
  x = [tiedrank(x)/n ones(n,1)];
  y = [tiedrank(y)/n ones(n,1)];
  x = sin(s/size(x,2)*x*randn(size(x,2),k));
  y = sin(s/size(y,2)*y*randn(size(y,2),k));
  [~,~,r] = canoncorr([x ones(n,1)],[y ones(n,1)]);
```

RDC versus Others

Name of Coeff.	Non-Linear	Vector Inputs	Marginal Invariant	Renyi's Properties	Coeff. $\in [0,1]$	# Par.	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
MMD	✓	✓	×	×	×	1	n^2
CMMD	✓	✓	✓	×	×	1	n^2
RDC	✓	✓	✓	✓	✓	2	$n \log n$

Is it fast?

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	—	—	—	—	—

Score Examples

1.0	1.0	0.4	1.0
	0.0	0.0	0.0

0.3	0.3	0.1	0.2
	0.0	0.0	-0.0

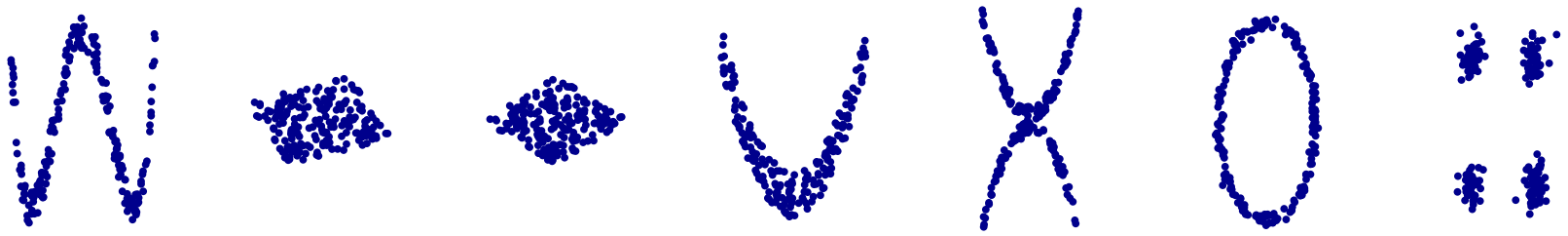
0.5	0.5	0.1	0.2
	0.0	0.0	0.0

1.0	1.0	0.5	0.9
	0.0	0.0	0.0

1.0	1.0	0.3	0.6
	0.1	0.1	0.1

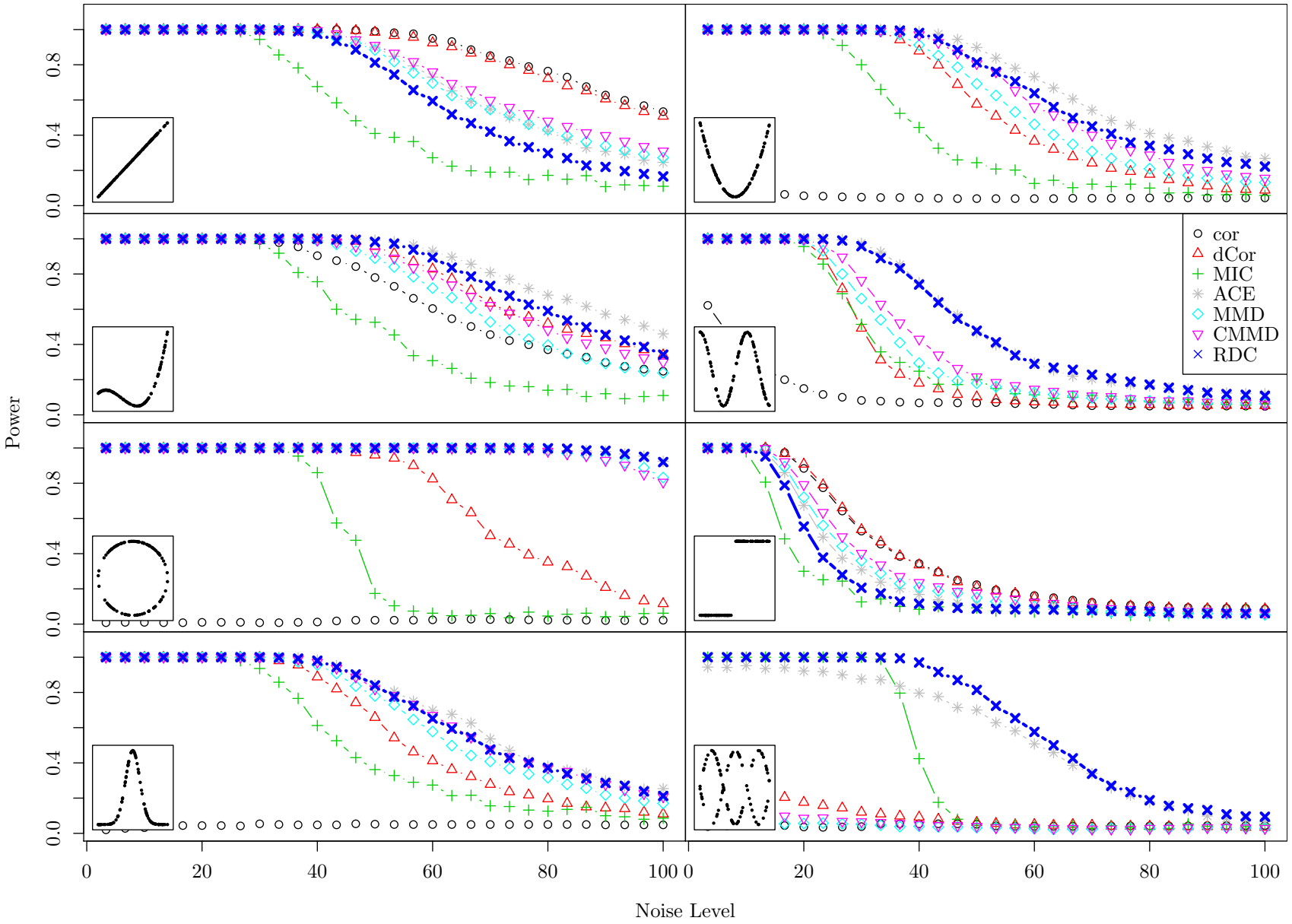
1.0	1.0	0.2	0.6
	-0.0	-0.0	-0.0

0.1	0.1	0.0	0.1
	-0.0	-0.0	-0.0

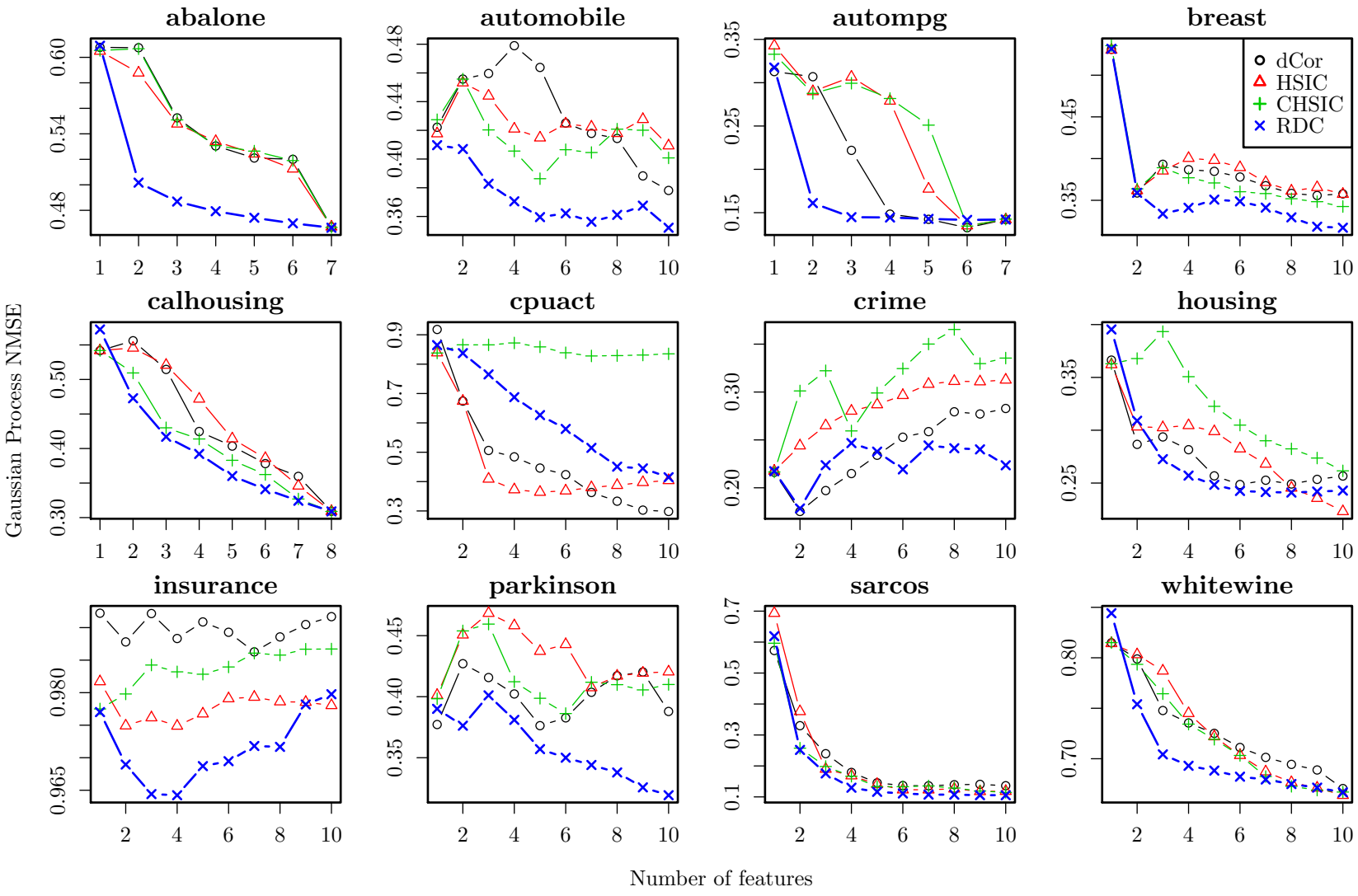


RDC, ACE, dCor, MIC, Pearson's, Spearman's, Kendall's

Power Experiments



Feature Selection Experiments



Mehta Data Experiments

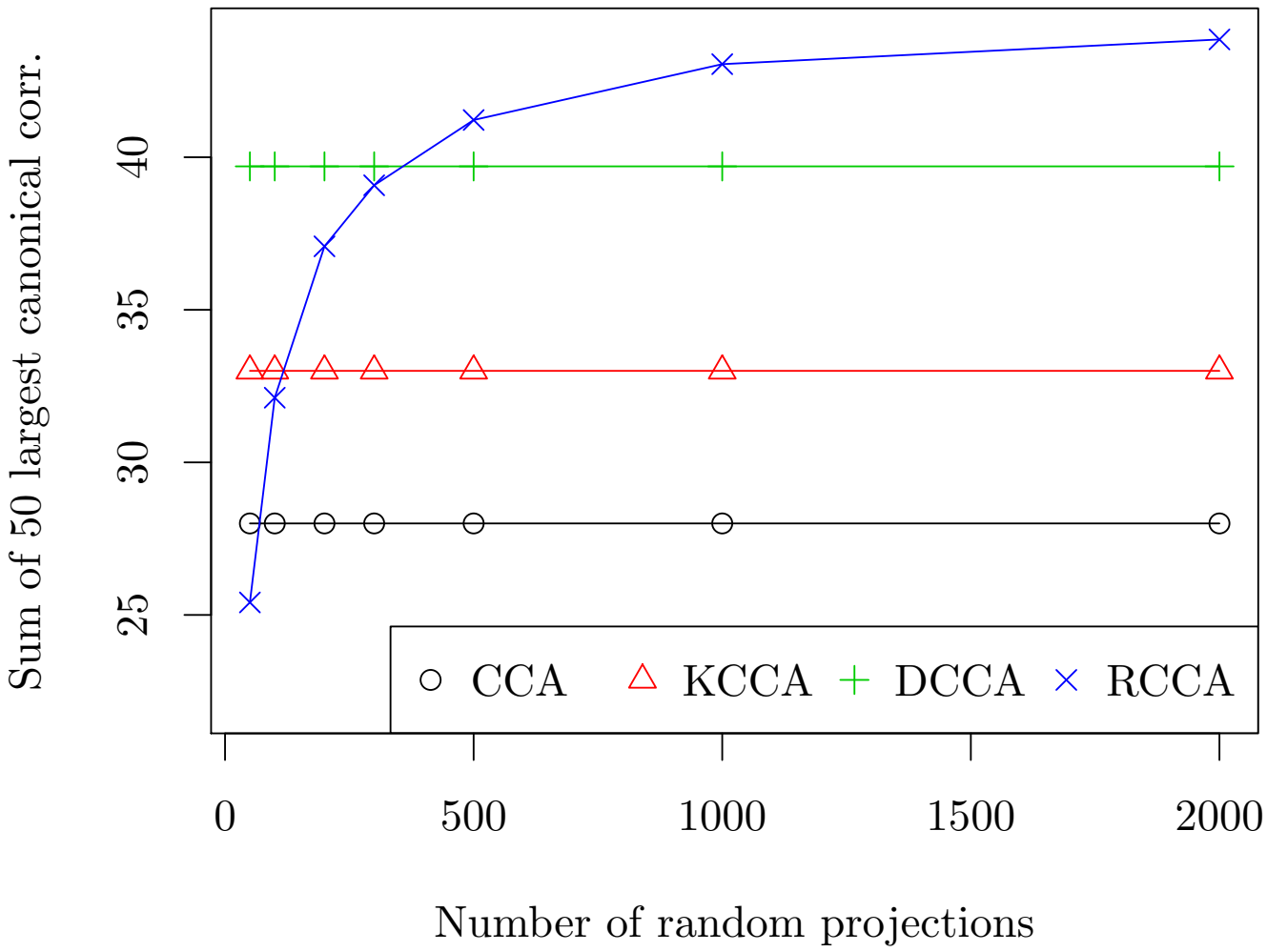
Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder

Divya Mehta[a,1], Torsten Klengel[a], Karen N. Conneely[b], Alicia K. Smith[c], André Altmann[a], Thaddeus W. Pace[c,d], Monika Rex-Haffner[a], Anne Loeschner[a], Mariya Gonik[a], Kristina B. Mercer[e], Bekh Bradley[c,f], Bertram Müller-Myhsok[a], Kerry J. Ressler[c,e,g], and Elisabeth B. Binder[a,c]

[a]Max Planck Institute of Psychiatry, 80804 Munich, Germany; [b]Department of Human Genetics, [c]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322; [d]Winship Cancer Institute, Emory University, Atlanta, GA 30322; [e]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789; [f]Veteran's Affairs Medical Center, Decatur, GA 30033; and [g]Yerkes National Primate Research Center, Atlanta, GA 30322

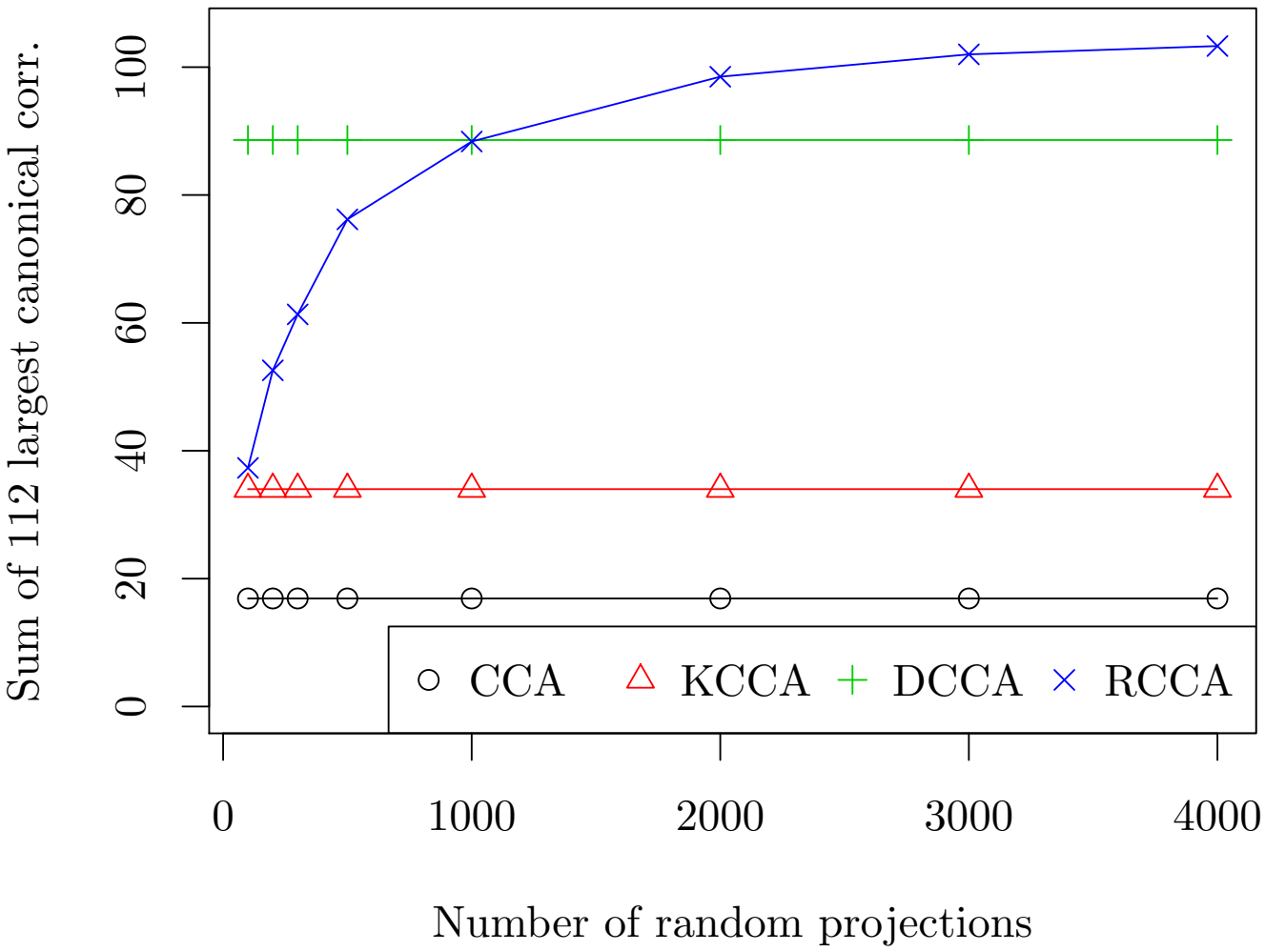
1. Calculate linear model and RDC for $n = 97057278$ (expression, methylation) transcript pairs against the phenotype. Assess significance using 100.000 permutations.
2. Search for transcripts with $R^2 < 0.01$ and $\mathrm{RDC} > 0.4$
3. Several hundreds of hits (under study at the Max Planck Institute of Psychiatry, Munich)

Learning Common Representations with CCA (I)



Learn 50 correlated representations between left and right halves of handwritten digit images (MNIST dataset)

Learning Common Representations with CCA (II)



Learn 112 correlated representations between speech audio and tongue positions (XRMB dataset)

Learning Common Representations with CCA (III)

Dataset of 35.000 images of animals with 85 high-level paired attributes.

Make two pools of 25 animals each. Do binary classification with:

- SURF features: 55.28% accuracy
- CCA features: 62.11% accuracy
- High-level features: 100% accuracy

Experiments on the pipeline

Can CCA learn representations...

- unsupervisedly from data (*autoencoding*)
- from different data modalities (*priviledged information*)
- shared between multiple tasks (*transfer learning*)
- ...?

Thanks!