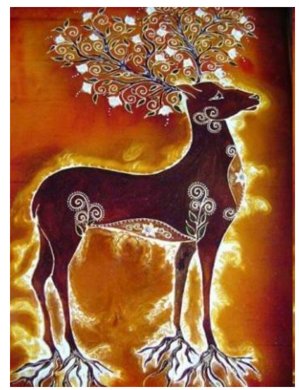
REGO: **R**ank-based estimation of Rényi information using **E**uclidean **G**raph **O**ptimization

Barnabás Póczos (University of Alberta)
Sergey Kirshner (Purdue University)
Csaba Szepesvári (University of Alberta)



AISTATS 2010

May 14, 2010

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

RL & AI

UNIVERSITY OF ALBERTA
QUAECUMQUE VERA

Outline

GOAL: dependence estimation using **mutual information**

- Importance in machine learning
- Background on information and entropy

Graph optimization methods
TSP, MST, kNN

Copula transformation

Information estimation

Who cares about dependence?

- **Unsupervised learning**
  - Which observations are dependent / independent

- **Supervised learning**
  - Is there dependence between inputs and outputs?

- **Analysis of stock markets, physical, biological, chemical systems**
  - Dependence between observations?

- **Other applications**

  Feature selection, Boosting, Clustering

  Information theory, Channel capacity, Information geometry,

  Optimal experiment design, active learning,

  Prediction of protein structure, Drug design, fMRI data processing,

  Microarray data processing, Image registration, ICA/ISA… etc

What is dependence δ?

$(A1)$ $0 \le \delta(\mathbf{X}) \le \gamma$, $\mathbf{X} = (X_1, \ldots, X_d)$, γ can be ∞

$(A2)$ $\delta(\mathbf{X}) = 0 \Leftrightarrow (X_1, \ldots, X_d)$ independent

$(A3)$ $\delta(\mathbf{X}) = \gamma \Leftrightarrow$ deterministic relation in $(X_1, \ldots, X_d)$

$(A4)$ invariance for 1-to-1 transform., permutation

$(A5)$ consistent with $|corr|$ for normal distribution

$(A6)$ superadditivity: $\mathbf{X} = (\mathbf{Y}, \mathbf{Z}) \Rightarrow \delta(\mathbf{X}) \ge \delta(\mathbf{Y}) + \delta(\mathbf{Z})$

Alfréd Rényi

Rényi’s information

$$I_\alpha(X_1, \ldots, X_d) = \frac{1}{\alpha - 1} \log \int \left(\prod_{i=1}^{d} f(x_i) \right)^{1-\alpha} f^\alpha(x_1, \ldots, x_d) dx_1 \cdots dx_d$$

Rényi’s entropy

$$H_\alpha(X_1, \ldots, X_d) = \frac{1}{1 - \alpha} \log \int f^\alpha(x_1, \ldots, x_d) dx_1 \cdots dx_d$$

$$\lim_{\alpha \to 1} I_\alpha = I, \quad \lim_{\alpha \to 1} H_\alpha = H$$

Shannon mutual information

$$I(\boldsymbol{X}) = \int f(x_1,\ldots,x_d)\log\frac{f(x_1,\ldots,x_d)}{f(x_1)\cdots f(x_d)}dx_1\cdots dx_d$$
$$= \sum_{i=1}^{d} H(X_i) - H(X_1,\ldots,X_d)$$

Measuring uncertainty (Shannon entropy)

$$H(X_1,\ldots,X_d) = -\int f(\mathbf{x})\log f(\mathbf{x})\,d\mathbf{x}$$

Claude Shannon

The estimation problem:

Let $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^n \in \mathbb{R}^d$ be i.i.d. variables.

Estimate $I_\alpha(\mathbf{X})$ given the sample $\mathbf{X}^{1:n} = [\mathbf{X}^1, \ldots, \mathbf{X}^n] \in \mathbb{R}^{d\times n}$.

$$I_\alpha(\mathbf{X}) = \frac{1}{\alpha-1}\log\int_{\mathcal{X}}(f_{\mathbf{X}}(\mathbf{x}))^{\alpha}\left(\prod_{j=1}^{d} f_{X_j}(x_j)\right)^{1-\alpha} d\mathbf{x}$$

$$I_\alpha(\mathbf{X}) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} f_{\mathbf{X}}^{\alpha}(\mathbf{x}) \left(\prod_{j=1}^{d} f_{X_j}(x_j) \right)^{1-\alpha} d\mathbf{x},$$

$$I_1(\mathbf{X}) = \int_{\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) \log \frac{f_{\mathbf{X}}(\mathbf{x})}{\prod_{j=1}^{d} f_{X_j}(x_j)} d\mathbf{x} = -H(\mathbf{X}) + \sum_{j=1}^{d} H(X_j)$$

How can we estimate them?

- **Plug-in estimators**: estimate the densities $f_{\mathbf{X}}, f_{X_1}, \ldots, f_{X_d}$
  - density estimators (histograms or kernel density estimators)
  - tuneable parameters, cross validation for model selection
  - density function is a nuisance parameter
- **Direct (not plug-in based) estimators**

History of Graph optimization methods
TSP, MST, kNN

J. Hammersley

- **1959, Beardwood, Halton, Hammersley**
  - Given uniform iid samples on $[0,1]^2$. TSP length=?

$$L(\mathbf{X}^1, \ldots, \mathbf{X}^n)/\sqrt{n} \quad \rightarrow \quad \beta_2 > 0 \quad \text{a.s.}$$

  - Observations with density f on $[0,1]^d$

$$L(\mathbf{X}^1, \ldots, \mathbf{X}^n)/n^{(d-1)/d} \rightarrow \beta_d \int f(x)^{(d-1)/d} dx \quad \text{a.s.}$$

- **1981 –** TSP ⇒ MST, Minimal Matching graphs, conjecture for kNN

Michael Steele

Joseph Yukich

Wansoo Rhee

$$\| \cdot \|_2 \Rightarrow \| \cdot \|_2^p, \qquad 0 < p < d$$

$$L(\mathbf{X}^1, \ldots, \mathbf{X}^n)/n^{(d-p)/d} \rightarrow \beta_{d,p} \int f(x)^{(d-p)/d} dx \quad \text{a.s.}$$

$$H_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log \int f^\alpha(\mathbf{x}) d\mathbf{x}$$

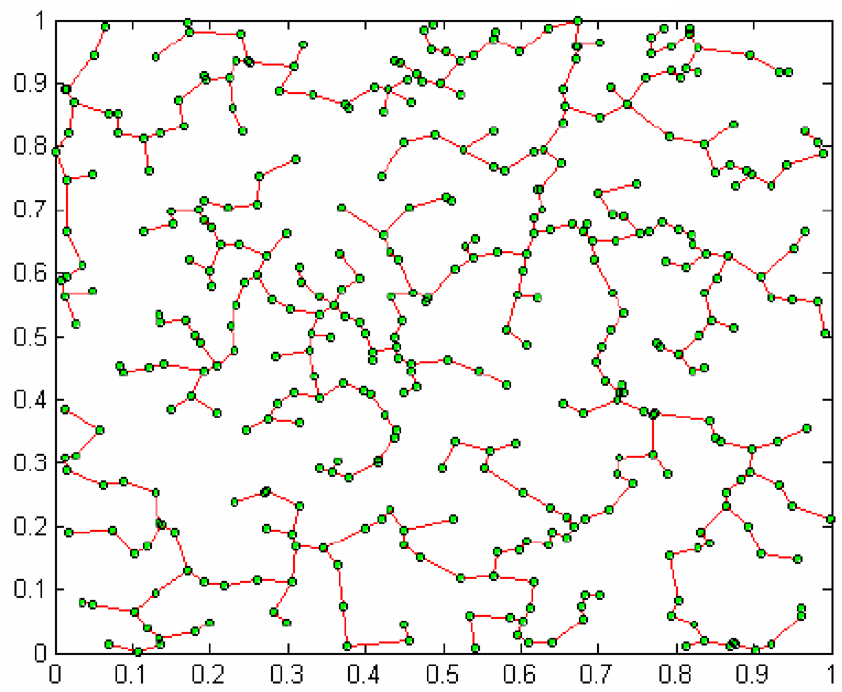
Entropy?

Steele, Yukich theorem for MST, TSP, Minimal Matching

Let $\mathbf{Z}, \mathbf{Z}^1, \ldots, \mathbf{Z}^n$ be i.i.d. on $[0,1]^d$ with density $f_{\mathbf{Z}}$.

$d \geq 2$, $0 < \alpha < 1$. Let $p = d - d\alpha$

MST on 2D uniform:



Define Euclidean functional:

$$L_n(\mathbf{Z}^{1:n}) = \min_{G \in \mathfrak{G}} \sum_{(i,j) \in E(G)} \|\mathbf{Z}_i - \mathbf{Z}_j\|^p$$

The entropy estimator:

$$H_n(\mathbf{Z}^{1:n}) \doteq \frac{1}{1-\alpha} \log \frac{L_n(\mathbf{Z}^{1:n})}{\beta_{d,p}\, n^{\alpha}}$$

$\Rightarrow H_n \to H_\alpha(\mathbf{Z})$ almost surely as $n \to \infty$.

Sensitive to outliers…!

(Fig taken from J. Costa and A. Hero)

How can we get information estimators from entropy estimators?

How can we make the estimators more robust?

⇩

The invariance trick

Information is preserved under monotonic transformations.

Let $\mathbf{Z} = (Z_1, \ldots, Z_d) = (g_1(X_1), \ldots, g_d(X_d)) = g(\mathbf{X})$
where $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, d$, is a monotone function.

$$I_\alpha(\mathbf{Z}) = \frac{1}{\alpha - 1} \log \int_{\mathcal{Z}} \left(\frac{f_{\mathbf{Z}}(\mathbf{z})}{\prod_{j=1}^{d} f_{Z_j}(z_j)} \right)^{\alpha} \left(\prod_{j=1}^{d} f_{Z_j}(z_j) \right) d\mathbf{z} = I_\alpha(\mathbf{X})$$

When the marginals of $\mathbf{Z}$ are uniform, $\Rightarrow I_\alpha(\mathbf{Z}) = -H_\alpha(\mathbf{Z})$, too.

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

RL & AI

UNIVERSITY OF ALBERTA QUAECUMQUE VERA

Transformation to get uniform marginals

Monotone transformation leading to uniform marginals?

Prob theory 101: $X_i \sim F_i$ cont. $\Rightarrow F_i(X_i) \sim U[0,1]$

The transformation **(copula transformation):**

Let $\mathbf{X} = [X_1, \ldots, X_d] \rightarrow [F_1(X_1), \ldots, F_d(X_d)] = [Z_1, \ldots, Z_d] = \mathbf{Z}$

$$\Rightarrow I_\alpha(\mathbf{X}) = I_\alpha(\mathbf{Z}) = -H_\alpha(\mathbf{Z})$$

Monotone transform　Uniform marginals

- *Information estimation problem is reduced for Rényi's entropy estimation*
- *a little problem: we don't know F_i distribution functions*

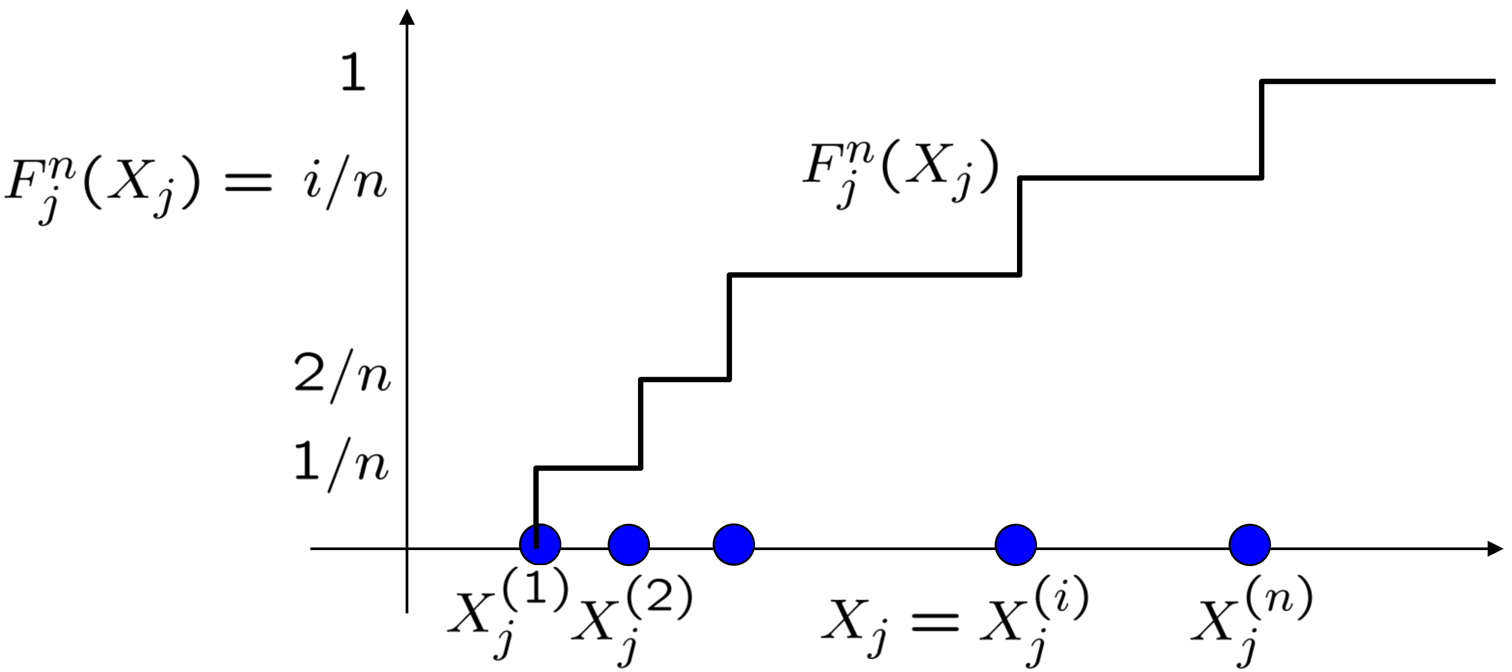
Empirical copula transformation

Solution:

Use empirical distributions F_j^n and empirical copula transform.
We need this in 1D only! ⇒ no curse of dimensionality.

We don't know $F_1,\ldots,F_d$ distribution functions
⇒ *estimate them with empirical distribution functions*

$$[X_j^{(1)} \leq X_j^{(2)}, \ldots, \leq X_j^{(n)}] \doteq sort\{X_j^1, \ldots, X_j^n\}$$



Empirical copula transformation

"true" copula:

$$\mathbf{X} = [X_1, \ldots, X_d] \to [\underbrace{F_1(X_1)}_{Z_1}, \ldots, \underbrace{F_d(X_d)}_{Z_d}] = [Z_1, \ldots, Z_d] = \mathbf{Z}$$

empirical copula:

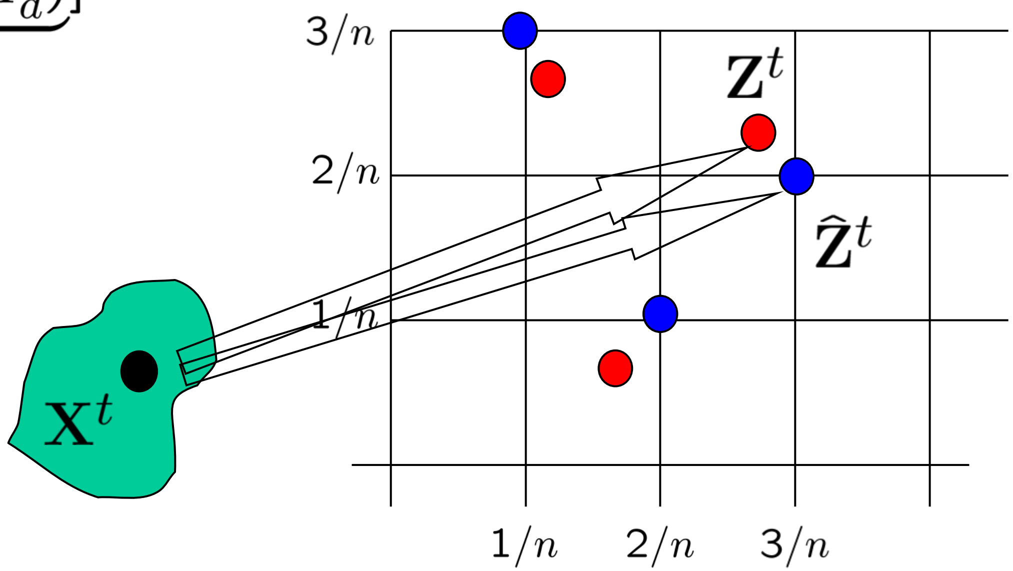
$$\mathbf{X} = [X_1, \ldots, X_d] \to [\underbrace{F_1^n(X_1)}_{\widehat{Z}_1}, \ldots, \underbrace{F_d^n(X_d)}_{\widehat{Z}_d}] = \left[\widehat{Z}_1, \ldots, \widehat{Z}_d\right] = \hat{\mathbf{Z}}$$

empirical copula transformation of the observations:

$$\left[X_1^t, \ldots, X_d^t\right] \to [\underbrace{F_1^n(X_1^t)}_{\widehat{Z}_1^t}, \ldots, \underbrace{F_d^n(X_d^t)}_{\widehat{Z}_d^t}]$$

$$\widehat{Z}_j^t = F_j^n(X_j^t)$$

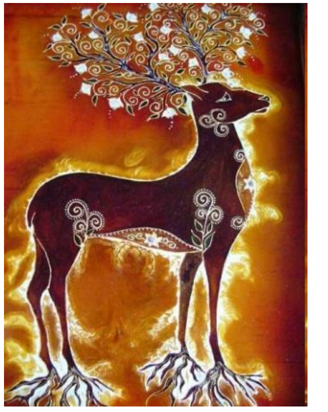
$$\widehat{\mathbf{Z}}^{1:n} = [\widehat{\mathbf{Z}}^1, \ldots, \widehat{\mathbf{Z}}^n] \in \mathbb{R}^{d \times n}$$



🔴 True copula transform

🔵 Empirical copula transform

REGO Algorithm
Rank-based **E**uclidean **G**raph **O**ptimization



1., **Empirical copula transformation**

The input $\mathbf{X}^1, \ldots \mathbf{X}^n$ is mapped into the unit hypercube $\hat{\mathbf{Z}}^1, \ldots, \hat{\mathbf{Z}}^n$ so that the marginals become approximately uniform.

$$\mathbf{X}^t = \left[X_1^t, \ldots, X_d^t\right] \to [\underbrace{F_1^n(X_1^t)}_{\hat{Z}_1^t}, \ldots, \underbrace{F_d^n(X_d^t)}_{\hat{Z}_d^t}] = \left[\hat{Z}_1^t, \ldots, \hat{Z}_d^t\right] = \hat{\mathbf{Z}}^{\mathbf{t}}$$

2., **Rényi entropy calculation**

The transformed sample $(\hat{\mathbf{Z}}^1, \ldots, \hat{\mathbf{Z}}^n)$ is sent to an algorithm that estimates the α-entropy of it.

Theoretical Results
Consistency

Main theorem:

Let $d \geq 3$, $1/2 < \alpha < 1$.

Let $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^n$ be i.i.d. random variables, supported on $[0,1]^d$ with density $f = f_{\mathbf{X}}$.

Assume that $\mathfrak{G} \in \{$ TSP, MST, MM, k-NN $\}$ and consider the corresponding estimator $H_n = H_n(\hat{\mathbf{Z}}^{1:n}; \mathfrak{G})$ obtained by running the REGO algorithm on $\mathbf{X}^{1:n} = (\mathbf{X}^1, \ldots, \mathbf{X}^n)$.

$\Rightarrow -H_n \to I_\alpha(\mathbf{X})$ almost surely as $n \to \infty$.

Note:

- Consistent MI estimator using ranks only
- We don't have theoretical results for other d and α values…
- Ranks only $\Rightarrow$ Robust

Theoretical Results
Infinitesimal robustness

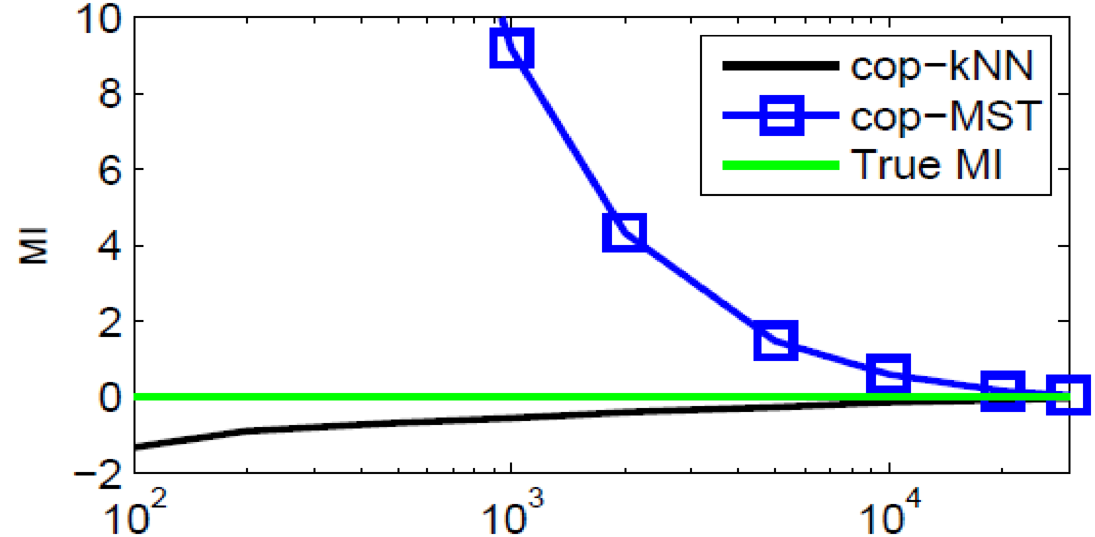
The finite sample influence functions (with some modification)

The amount of change caused by adding one outlier, **x**

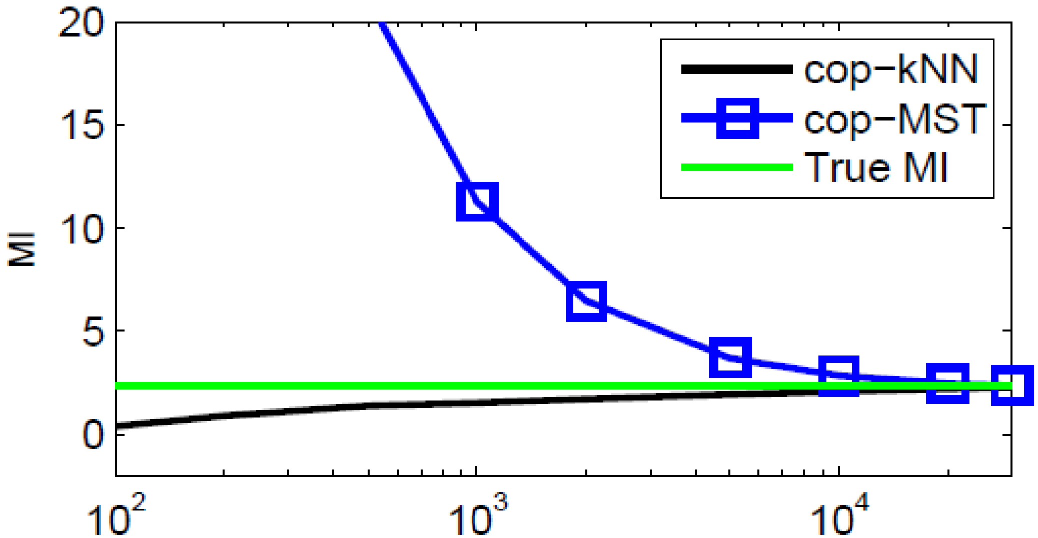
$$\Delta_n(\mathbf{x}) = |H_{n+1}(\mathbf{X}_{1:n}, \mathbf{x}) - H_n(\mathbf{X}_{1:n})| = O(n^{-\alpha}) \qquad 1/2 < \alpha < 1.$$

It cannot be arbitrarily big!

Empirical results, consistency in 10D

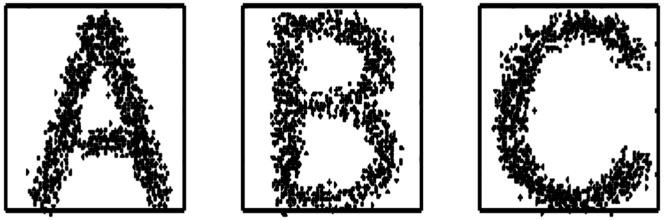


(a) 10D uniform



(b) 10D Gauss copula

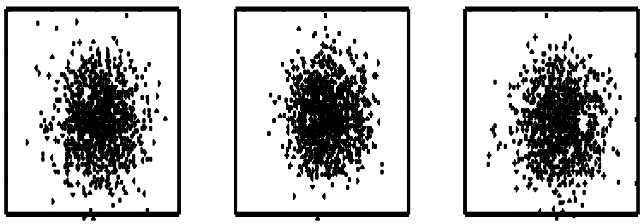
Independent Subspace Analysis



$\boldsymbol{S}^1 \in \mathbb{R}^2$ $\boldsymbol{S}^2 \in \mathbb{R}^2$ $\boldsymbol{S}^3 \in \mathbb{R}^2$

Hidden, independent sources
(subspaces)

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}^1 \\ \boldsymbol{S}^2 \\ \boldsymbol{S}^3 \end{pmatrix} \in \mathbb{R}^6$$



$\boldsymbol{X}^1 \in \mathbb{R}^2$ $\boldsymbol{X}^2 \in \mathbb{R}^2$ $\boldsymbol{X}^3 \in \mathbb{R}^2$

Observation

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}^1 \\ \boldsymbol{X}^2 \\ \boldsymbol{X}^3 \end{pmatrix} = \mathbf{AS} \in \mathbb{R}^6$$

$\mathbf{A} \in \mathbb{R}^{6\times 6}$ unknown mixing matrix

$$\underbrace{\mathbf{X}}_{2000} \;(2,2,2) = \underbrace{\mathbf{A}}_{6} \; \underbrace{\mathbf{S}}_{2000} \;(2,2,2)$$

Goal: Estimate **A** and **S** observing samples from **X** only

Independent Subspace Analysis

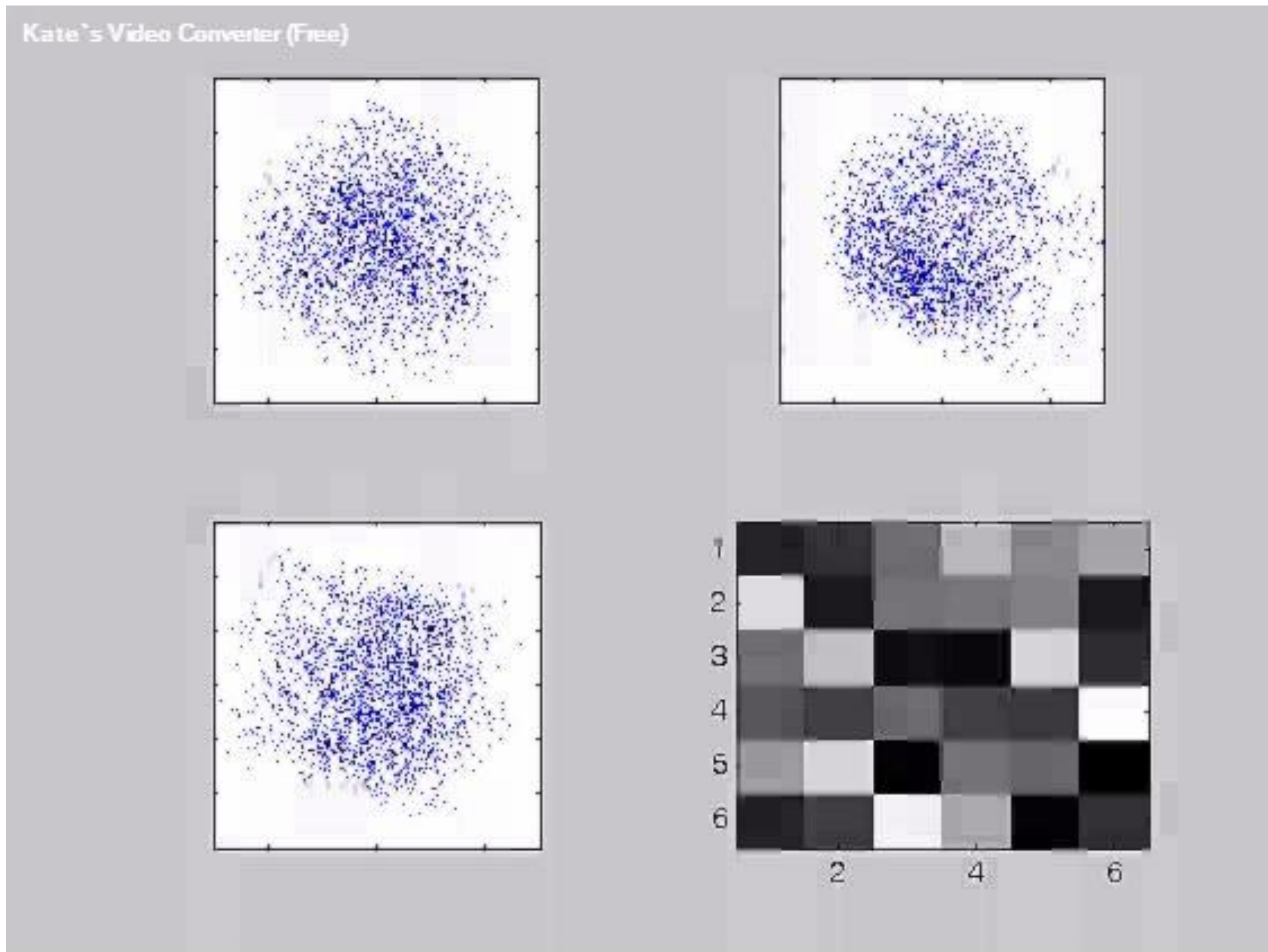
$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}^1 \\ \boldsymbol{X}^2 \\ \boldsymbol{X}^3 \end{pmatrix} = \mathbf{AS} \in \mathbb{R}^6$$

Objective: $\min_{\mathbf{W} \in \mathbb{R}^{6\times 6}} I(\boldsymbol{Y}^1, \boldsymbol{Y}^2, \boldsymbol{Y}^3)$

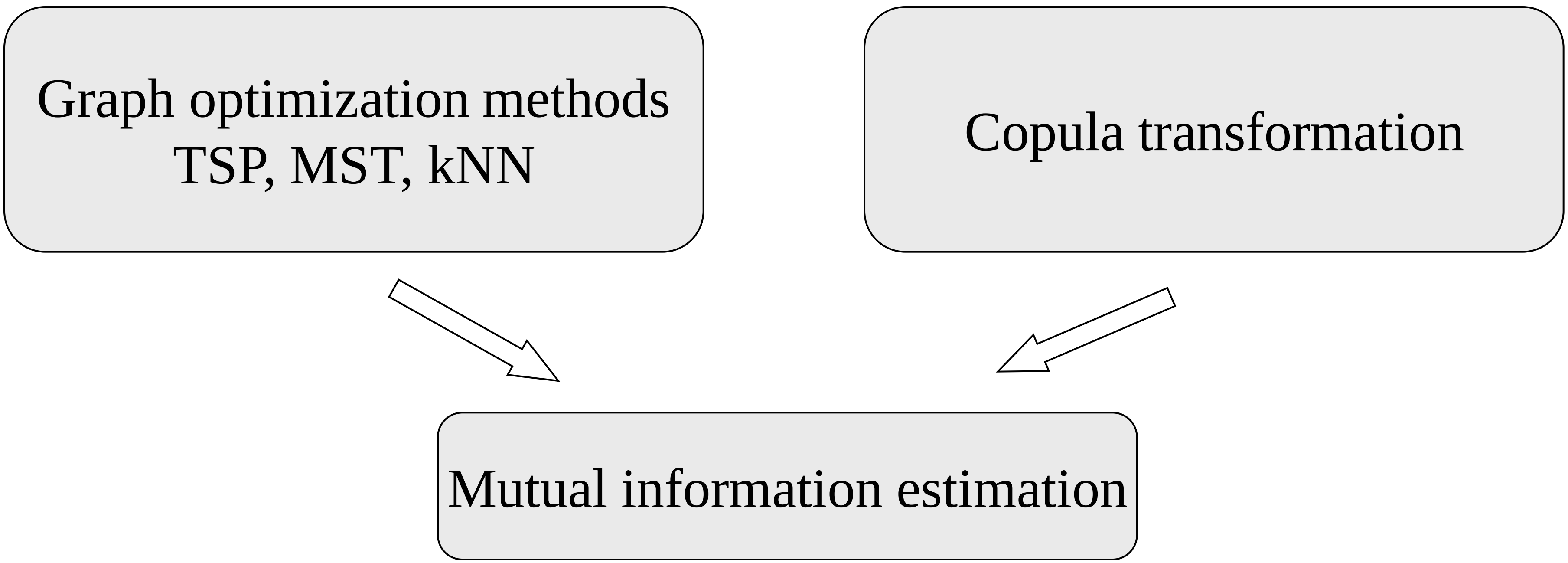
$$\boldsymbol{Y} = \mathbf{WX} = \mathbf{WAS} \in \mathbb{R}^6$$

$$\mathbf{W} \in \mathbb{R}^{6\times 6}$$

In case of perfect separation **WA** is a block permutation matrix

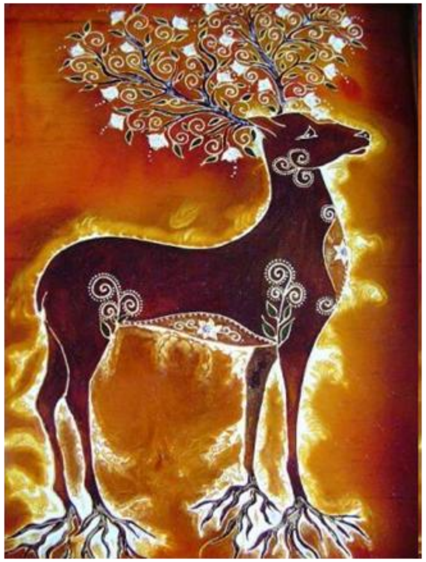


Take me home!



Marriage of seemingly unrelated mathematical areas produced a **consistent** and **robust** Rényi's mutual information estimator

Thanks for your attention! ☺



“Hey Rege, Róka Rege, Hey REGŐ Rejtem…
I am calling my grandfather, I sense his soul here.
I am listening to his words, they are breaking the silence.
Impart your knowledge to your grandson please,
1000 years have already passed…,
Knowledge equals life, hey!”