Measuring information in data: tools and models

Jaume Riba
Group. Signal Processing for Communications (SPCOM)
Dept. Teoria del Senyal i Communicacions (TSC)
Univ. Politècnica de Catalunya (UPC)
jaume.riba@upc.edu

Novembre 27, 2024 - Facultat de Matemàtiques i Estadística (FME-UPC)
Models Matemàtics de la Tecnologia

Funding

This work was funded by project MAYTE (PID2022-136512OB-C21) by MICIU/AEI/10.13039/501100011033 and ERDF/EU, grant 2021 SGR 01033 and grant 2022 FI SDUR 00164 by Departament de Recerca i Universitats de la Generalitat de Catalunya.

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN

Collaborators

Ferran de Cabrera Estanyol

Thesis (June 2023):
"Data-driven information-theoretic tools under a second-order statistic perspective"

Available at https://upcommons.upc.edu

PhD. Students

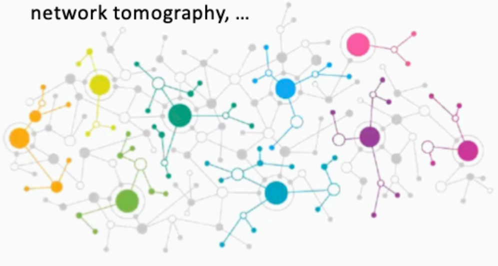
Carlos Alejandro López Molina
Thesis (by 2025)
"Majorization-Minimization framework in non-convex and Grassmann g-convex optimization: Exploiting diversity using sparsity and entropic criteria"

Marc Vilà Insa
Thesis (by 2026)
"Uniquely Factorable Constellations for Noncoherent Communications"

Aniol Martí Espelt
Thesis (by 2026)
"Communication and sensing with large arrays and statistical channel state information"

Explosive growth of generated data:

internet
communications (IoT)
more computational capacity of electronics
biomedical advances
network tomography, …

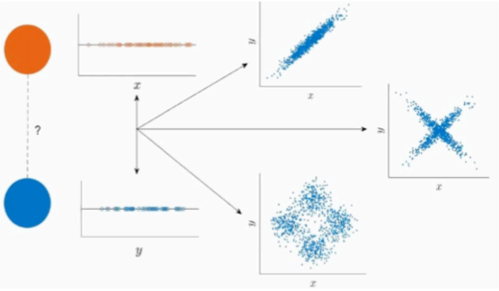


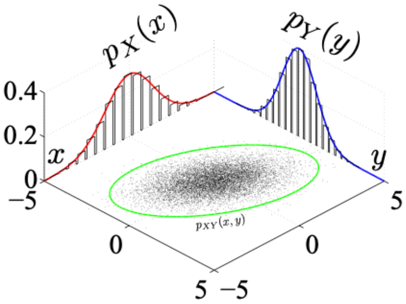
$x(l)$ $y(l)$

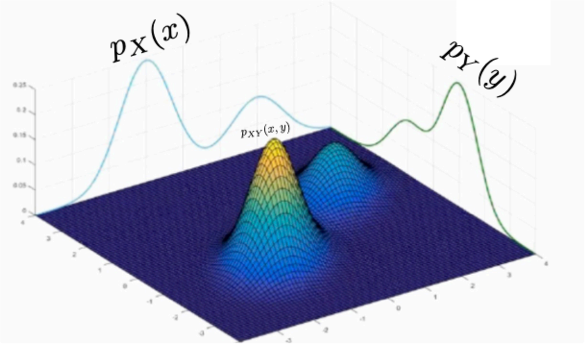
Are they related?

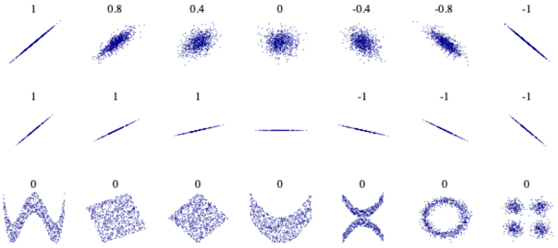
How much?

(causality?)









uncorrelation

$$\int\int xyp_{xy}(x,y)dxdy \;\overset{=}{\neq}\; \left(\int xp_x(x)dx\right)\left(\int yp_y(y)dy\right)$$

correlation

← value → ☺ ← value →

independence

$$p_{xy}(x,y) \;\overset{=}{\neq}\; p_x(x)p_y(y) \qquad \forall x,\, \forall y$$

dependence

← function → ☹ ← function →

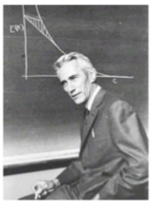
Shannon mutual information:
Kullback-Leibler divergence between densities

independence

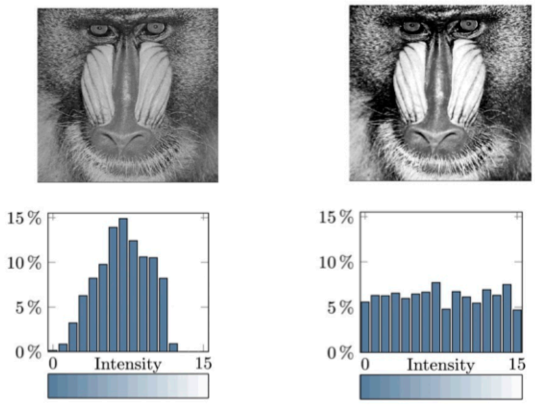
$$I_{xy} = \int\int p_{xy}(x,y)\ln\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}dxdy \;\overset{=}{>}\; 0$$

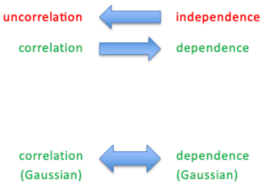
dependence

← value → ☺ ←→



Claude Shannon (1916-2001)





uncorrelation ⇐ independence

correlation ⇒ dependence

correlation (Gaussian) ⇔ dependence (Gaussian)

INDEX AND RATIONALE FOLLOWED

- Correlation between scalars

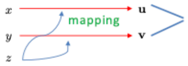
- Correlation between vectors via correlation between scalars

Proposal of 3 "natural" measures of correlation

- Dependence between scalars via correlation between vectors

Apply the 3 "natural" measures of correlation

- Conditional dependence



Extension of the ideas

- Proposal of toy problems for team work

- Correlation between scalars

Correlation

$$\sigma_{xy} = E\left[xy\right] - E[x]E[y]$$

$$\sigma^2_{xy} \begin{array}{l} = 0 \\ > 0 \end{array}$$

$= 0$ ← uncorrelation

> 0 ← correlation

- Correlation between scalars

Squared Pearson coefficient (coherence)

$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$-1 \leq r_{xy} \leq 1$$

$$r^2_{xy} \begin{array}{l} = 0 \\ > 0 \\ = 1 \end{array}$$

$= 0$ ← uncorrelation

> 0 ← correlation

$= 1$ ← deterministic & affine dependence

- Correlation between scalars

Estimation

$$\hat{\sigma}_{xy} = \frac{1}{\underbrace{L-1}_{\text{data size}}} \sum_{l=1}^{L} \left(x(l) - \bar{x}\right)\left(y(l) - \bar{y}\right)$$

data centering

$$\bar{x} = \frac{1}{L}\sum_{l=1}^{L} x(l)$$

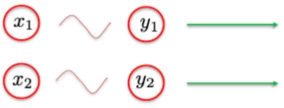
$$\bar{y} = \frac{1}{L}\sum_{l=1}^{L} y(l)$$

sample averages

- Correlation between scalars

Two pairs of scalars

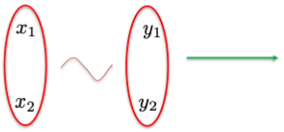
Individual measures



$r^2_{x_1 y_1}$

$r^2_{x_2 y_2}$

Global measure?



$r^2_{x_1 y_1} + r^2_{x_2 y_2}$

Is this intuition meaningful?

- Correlation between scalars

Gaussian case

$$I_{x_n y_n} = \ln \frac{1}{1 - r^2_{x_n y_n}}$$

Gaussian case (small correlation)

$$I_{x_n y_n} \approx r^2_{x_n y_n}$$

Local measure of information

- Correlation between scalars

Independent Gaussian pairs

$$I_{\mathbf{xy}} = \sum_n I_{x_n y_n} = \sum_n \ln \frac{1}{1 - r^2_{x_n y_n}} = \ln \frac{1}{\prod_n \left(1 - r^2_{x_n y_n}\right)}$$

$$= \ln \frac{1}{1 - \left(1 - \prod_n \left(1 - r^2_{x_n y_n}\right)\right)}$$

$$\boxed{1 - \prod_n \left(1 - r^2_{x_n y_n}\right)}$$

Global coherence in the range 0 to 1

- Correlation between scalars

Independent Gaussian pairs (small correlation)

$$I_{\mathbf{xy}} = \sum_n I_{x_n y_n} \approx \sum_n r^2_{x_n y_n}$$

… so

adding coherences is locally meaningful

- Correlation between vectors via correlation between scalars

$$\mathbf{u} = \begin{pmatrix} u(1) \\ u(2) \\ \vdots \\ u(N_u) \end{pmatrix} \qquad \mathbf{v} = \begin{pmatrix} v(1) \\ v(2) \\ \vdots \\ v(N_v) \end{pmatrix}$$

$$\mathbf{C}_{uv} = E\left[\left(\mathbf{u} - E[\mathbf{u}]\right)\left(\mathbf{v} - E[\mathbf{v}]\right)^H\right]$$

$$\mathbf{C}_{uv} \overset{=}{\underset{\neq}{}} \underset{\text{matrix}}{\mathbf{0}} \qquad \underset{\text{Frobenius norm}}{\|\mathbf{C}_{uv}\|_F^2} \overset{=}{\underset{>}{}} \underset{\text{scalar}}{0}$$

correlation

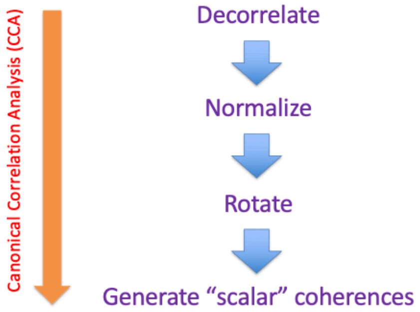
uncorrelation

correlation

uncorrelation

- Correlation between vectors via correlation between scalars

Coherence matrix?



Canonical Correlation Analysis (CCA)

Decorrelate

Normalize

Rotate

Generate "scalar" coherences

- Correlation between vectors via correlation between scalars

Decorrelate & normalize

$$\mathbf{u}' = \mathbf{C}_u^{-1/2}\mathbf{u}$$

$$\mathbf{v}' = \mathbf{C}_v^{-1/2}\mathbf{v}$$

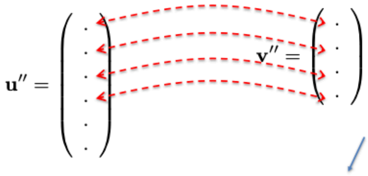
Rotate

$$\mathbf{u}'' = \mathbf{O}_u^H\mathbf{u}'$$

$$\mathbf{v}'' = \mathbf{O}_v^H\mathbf{v}'$$

- Correlation between vectors via correlation between scalars

Impose correlation "by pairs"



diagonal

$$\mathbf{C}_{u''v''} = \mathbf{O}_u^H\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\mathbf{O}_v = \mathbf{D}$$

$$\mathbf{\Gamma} = \boxed{\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}} \overset{\text{SVD}}{=} \mathbf{O}_u\mathbf{D}\mathbf{O}_v^H$$

Coherence matrix

The singular values of the coherence matrix are Pearson coefficients of pairs

Measure from the coherence matrix:

$$\sum_{n=1}^{\text{mín}(N_u,N_v)} \ln \frac{1}{1-d_n^2} = -\ln\det\left(\mathbf{I}-\boldsymbol{\Gamma}^H\boldsymbol{\Gamma}\right) \approx \operatorname{tr}\left(\boldsymbol{\Gamma}^H\boldsymbol{\Gamma}\right) = \|\boldsymbol{\Gamma}\|^2$$

$$\boxed{\sum_{n=1}^{\text{mín}(N_u,N_v)} \lambda_n = \sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2}$$

$k_n =$ singular values of $\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$

$\lambda_n =$ eigen values of $\left(\mathbf{C}_{uv}^H\mathbf{C}_u^{-1}\mathbf{C}_{uv}\mathbf{C}_v^{-1}\right)$

(Linear) maximal correlation problem:

$u'' = \mathbf{f}^H\mathbf{u}$

$\max_{\mathbf{f},\mathbf{g}} r^2_{u''v''}$

$\|\mathbf{f}\|^2 = 1$

$\|\mathbf{g}\|^2 = 1$

$v'' = \mathbf{g}^H\mathbf{v}$

$=$ maximum eigenvector of the coherence matrix $= \lambda_1$

$\mathbf{f} =$ first column of $\mathbf{O}_u$

$\mathbf{g} =$ first column of $\mathbf{O}_v$

Estimation

$$\boxed{\hat{\mathbf{C}}_{uv} = \frac{1}{L-1}\sum_{l=1}^{L}\left(\mathbf{u}(l)-\bar{\mathbf{u}}\right)\left(\mathbf{v}(l)-\bar{\mathbf{v}}\right)^H} \quad \in \mathbb{C}^{N_u\times N_v}$$

(data centering)

more compactly...

$$\boxed{\hat{\mathbf{C}}_{uv} = \frac{1}{L-1}\mathbf{U}\mathbf{P}^{\perp}\mathbf{V}^H} \quad \in \mathbb{C}^{N_u\times N_v}$$

$\mathbf{U} = [\mathbf{u}(1),\mathbf{u}(2),\ldots,\mathbf{u}(L)]$
Data matrix U

$\mathbf{P}^{\perp} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^H}{L}$
Orthogonal projector
(data centering)

$\mathbf{V} = [\mathbf{v}(1),\mathbf{v}(2),\ldots,\mathbf{v}(L)]$
Data matrix V

In summary: three measures

σ_{xy}^2

① $\|\mathbf{C}_{uv}\|^2$

Intuitive and simple.

$\left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right|^2$

② $k^2_{\max}\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$ Left and right singular vectors associated to $k_{\max}$ discover the best linear relations

Normalized in the range 0 to 1. Requires inverses.

③ $\displaystyle\sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right) = \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$

Aims at measuring information. Requires inverses.

Axioms for a measure of information

Rényi

1 $\delta_{xy} = \delta_{yx}$

2 $\delta_{xy} = 0$ iff x and y are independent

3 $0 \le \delta_{xy} \le 1$

4 $\delta_{xy} = 1$ if x and y are strictly dependent

5 $\delta_{xy} = \psi\left(r_{xy}\right)$ if x and y are jointly Gaussian

6 $\delta_{uv} = \delta_{xy}$

Granger

$u = f(x)$
$v = g(y)$
Strictly monotonic functions

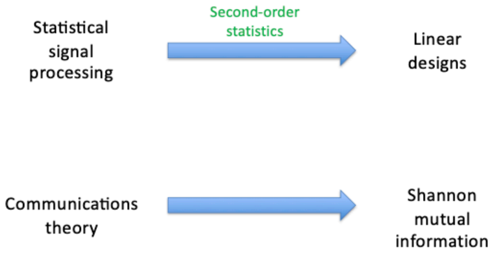
Information-theoretic squared coherence

$I_{xy} = E\ln\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}$ Shannon mutual information

$\rho_{xy}^2 = 1 - e^{-I_{xy}}$

IT coherence

$$\rho_{xy}^2 \begin{array}{l} = 0 \leftarrow \text{independence} \\ > 0 \leftarrow \text{dependence} \\ = 1 \leftarrow \text{deterministic \& functional dependence} \end{array}$$



Statistical signal processing

Second-order statistics

Linear designs

Communications theory

Shannon mutual information

Second-order statistics:

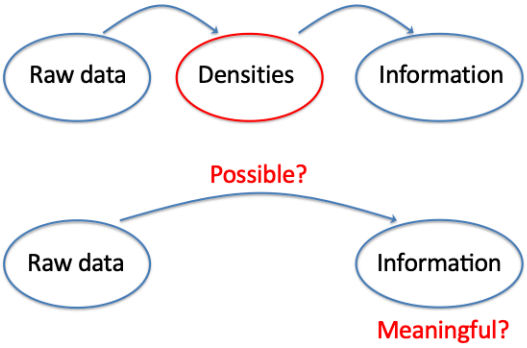
it is directly estimable from data

it captures only linear (affine) relations

Should we estimate first densities and then (Shannon) information?

Or perhaps estimate information directly by averaging transformed data samples?

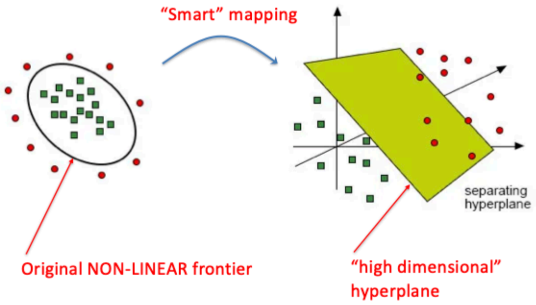
Are we obligated to Shannon information or perhaps we can accept other information measures?



Cover's theorem
(informal statement)

Non linearly separable sets can be linearly separable by "intentionally" increasing the dimensionality of the problem.

Thomas Cover (1938-2012)



Fact:
Second-order statistics are not able to capture non-linear relations.

Conjecture (inspiration from Cover's thm.)
Can we map data to vectors and then use second-order statistics to capture non-linear relations and statistical dependence manifested in the original data?

$$p_{xy}(x,y) = p_x(x)p_y(y)$$

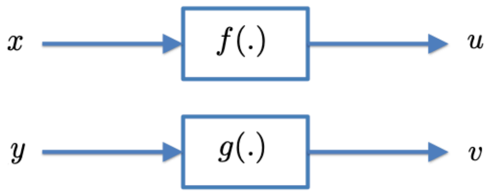
Mathematical

Equivalent statements of independence

$$E\left[f(x)g(y)\right] = E\left[f(x)\right]E\left[g(y)\right]$$

Operational

$$\forall f, \forall g$$



☹ Searching for all functions!!! $\forall f, \forall g$

🙂 Uncorrelation idea is reborn!!! $E\left[uv\right] = E\left[u\right]E\left[v\right]$

If $f_0 \& g_0$ $\exists$ such that $r^2_{u_0 v_0} > 0$

with $u_0 = f_0(x)$ then
$v_0 = g_0(y)$

x and y are statistically dependent

Hirschfeld-Gebelein-Rényi (HGR) maximal correlation coefficient

$$\mathrm{HGR}_{xy} = \text{máx}_{\underbrace{f,g}} r^2_{f(x),g(y)}$$

☹ Searching for all functions!!! $\forall f, \forall g$

$$\mathrm{HGR}_{xy} \begin{array}{l} = 0 \longleftarrow \text{independence} \\ \neq 0 \longleftarrow \text{dependence} \\ = 1 \longleftarrow \text{deterministic \& functional dependence} \end{array}$$

Characteristic function

$$\varphi_x(\omega_x) = E\left[e^{j\omega_x\, x}\right] = \underbrace{\int p_x(x)e^{j\omega_x\, x}dx}$$

Marginal CF

Inv. Fourier transform of the PDFs

$$\varphi_y(\omega_y) = E\left[e^{j\omega_y\, y}\right] = \overbrace{\int p_y(y)e^{j\omega_y\, y}dy}$$

Joint characteristic function

$$\varphi_{xy}(\omega_x, \omega_y) = E\left[e^{j(\omega_x\ x\ +\ \omega_y\ \mathrm{y})}\right]$$

$$= \int\int p_{xy}(x,y)e^{j(\omega_x\ x\ +\ \omega_y\ \mathrm{y})}dxdy$$

Note that:

$$\varphi_{xy}(\omega_x, -\omega_y) = E\left[e^{j(\omega_x\ x - \omega_y\ \mathrm{y})}\right] = E\left[\underbrace{e^{j\omega_x\ x}}\underbrace{\left(e^{j\omega_y\ y}\right)}^{*}\right]$$

Correlation between two complex random variables

$$p_{xy}(x,y) = p_x(x)p_y(y)$$

Mathematical

$$\Updownarrow$$

$$E\left[f(x)g(y)\right] = E\left[f(x)\right]E\left[g(y)\right]$$

Operational

$\forall f, \forall g$ ☹ Searching functions!!!

$$p_{xy}(x,y) = p_x(x)p_y(y)$$

Mathematical

$$\Updownarrow$$

$$\xcancel{E\left[f(x)g(y)\right] = E\left[f(x)\right]E\left[g(y)\right]}$$

~~Operational~~

~~$\forall f, \forall g$ ☹ Searching functions!!!~~

$$p_{xy}(x,y) = p_x(x)p_y(y)$$

Mathematical

$$\Updownarrow$$

Universal mapper

$s_\omega(.) = e^{j\omega\ \cdot}$

$$E\left[s_{\omega_x}(x)s^*_{\omega_y}(y)\right] = E\left[s_{\omega_x}(x)\right]E^*\left[s_{\omega_y}(y)\right]$$

Operational and practical

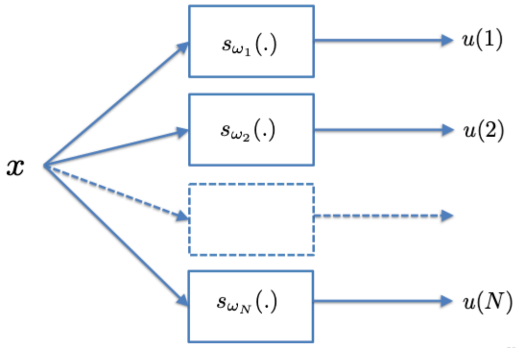
$\forall \omega_x, \omega_y \in \mathbb{R}$ 🙂 Searching reals!!!

Good: 🙂

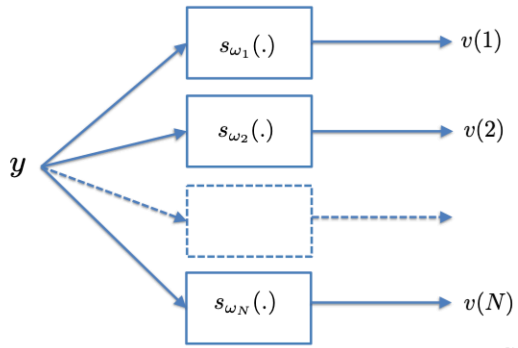
Dependence is discovered by discovering correlated "frequency" pairs.

But… ☹

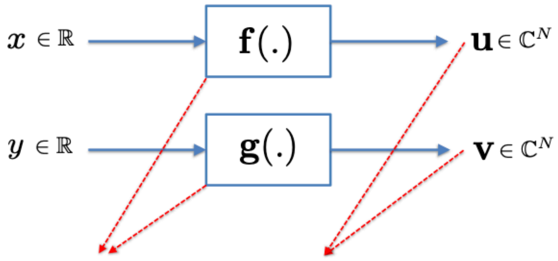
No single frequency pair will "capture" the total correlation discovered by the max. correlation coefficient (HGR).



- Dependence between scalars via correlation between vectors



- Dependence between scalars via correlation between vectors



Universal mappers

Are vectors correlated?
How much?

Apply the proposed three measures of correlation

Reminder of the three natural measures from random vectors

σ_{xy}^2

① $\|\mathbf{C}_{uv}\|^2$

Intuitive and simple.

$\left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right|^2$

② $k^2_{\max}\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$ Left and right singular vectors associated to $k_{\max}$ discover the best linear relations

Normalized in the range 0 to 1. Requires inverses.

③ $\sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right) = \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$

Aims at measuring information. Requires inverses.

Reminder of the three natural measures from random vectors

σ_{xy}^2

① $\boxed{\|\mathbf{C}_{uv}\|^2}$

Intuitive and simple.

$\left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right|^2$

② $k^2_{\max}\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$ Left and right singular vectors associated to $k_{\max}$ discover the best linear relations

Normalized in the range 0 to 1. Requires inverses.

③ $\sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right) = \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$

Aims at measuring information. Requires inverses.

- Dependence between scalars via correlation between vectors

① $\boxed{\|\mathbf{C}_{uv}\|^2}$

$$\text{tr}\left(\hat{\mathbf{C}}_{uv}^H\hat{\mathbf{C}}_{uv}\right) = \frac{1}{(L-1)^2}\,\text{tr}\left(\mathbf{V}\mathbf{P}^{\perp}\mathbf{U}^H\mathbf{U}\mathbf{P}^{\perp}\mathbf{V}^H\right)$$

$$= \frac{1}{(L-1)^2}\,\text{tr}\left(\mathbf{P}^{\perp}\underbrace{\mathbf{U}^H\mathbf{U}}\mathbf{P}^{\perp}\underbrace{\mathbf{V}^H\mathbf{V}}\right)$$

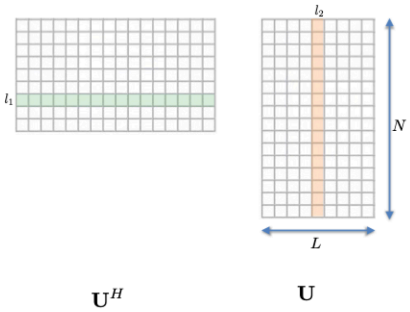
$$= \frac{1}{(L-1)^2}\,\text{tr}\left(\mathbf{P}^{\perp}\mathbf{K}_x\mathbf{P}^{\perp}\mathbf{K}_y\right)$$

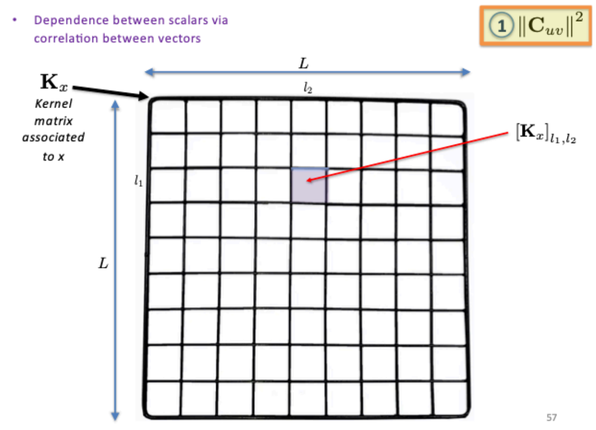
$\boxed{\hat{\mathbf{C}}_{uv} = \frac{1}{L-1}\mathbf{U}\mathbf{P}^{\perp}\mathbf{V}^H}$

$\boxed{\mathbf{K}_x = \mathbf{U}^H\mathbf{U}}$ $\boxed{\mathbf{K}_y = \mathbf{V}^H\mathbf{V}}$

- Dependence between scalars via correlation between vectors

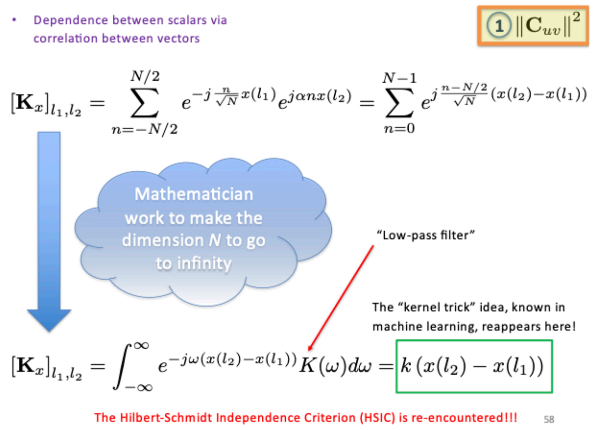
① $\boxed{\|\mathbf{C}_{uv}\|^2}$





- Dependence between scalars via correlation between vectors

① $\|\mathbf{C}_{uv}\|^2$



- Dependence between scalars via correlation between vectors

① $\|\mathbf{C}_{uv}\|^2$

$$[\mathbf{K}_x]_{l_1,l_2} = \sum_{n=-N/2}^{N/2} e^{-j\frac{n}{\sqrt{N}}x(l_1)} e^{j\alpha n x(l_2)} = \sum_{n=0}^{N-1} e^{j\frac{n-N/2}{\sqrt{N}}(x(l_2)-x(l_1))}$$

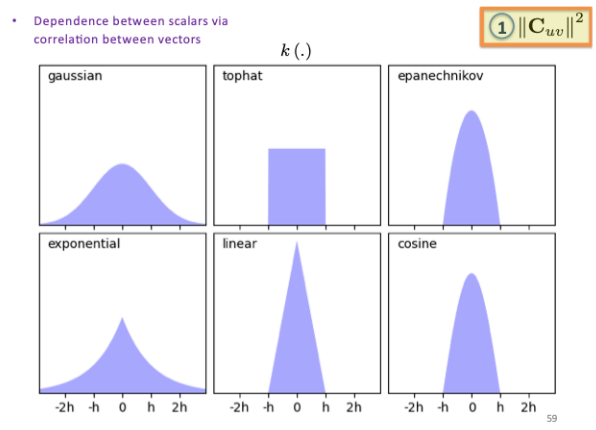
Mathematician work to make the dimension *N* to go to infinity

"Low-pass filter"

The "kernel trick" idea, known in machine learning, reappears here!

$$[\mathbf{K}_x]_{l_1,l_2} = \int_{-\infty}^{\infty} e^{-j\omega(x(l_2)-x(l_1))} K(\omega) d\omega = \boxed{k\left(x(l_2)-x(l_1)\right)}$$

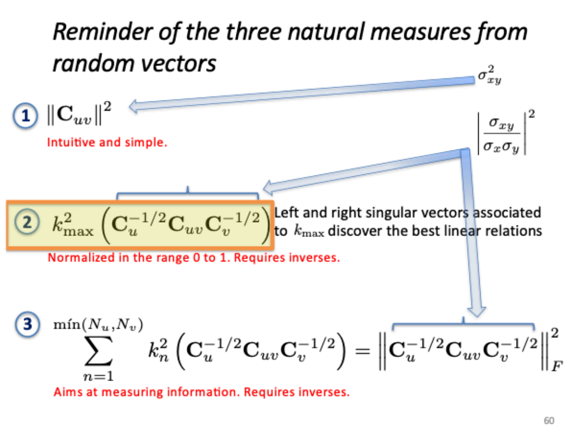
The Hilbert-Schmidt Independence Criterion (HSIC) is re-encountered!!!



- Dependence between scalars via correlation between vectors

① $\|\mathbf{C}_{uv}\|^2$

$k(.)$



Reminder of the three natural measures from random vectors

σ_{xy}^2

① $\|\mathbf{C}_{uv}\|^2$

Intuitive and simple.

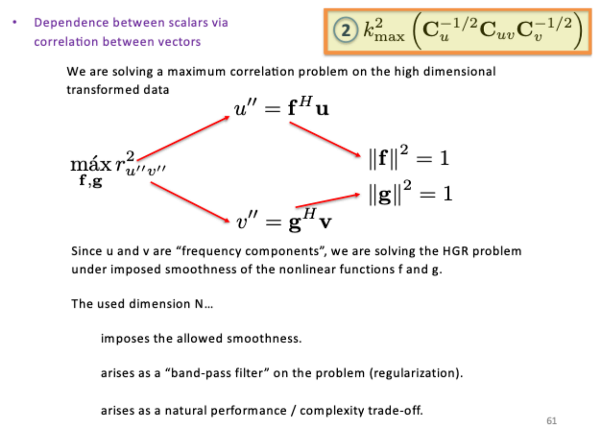
$\left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right|^2$

② $k_{\max}^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$ Left and right singular vectors associated to $k_{\max}$ discover the best linear relations

Normalized in the range 0 to 1. Requires inverses.

③ $\sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right) = \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$

Aims at measuring information. Requires inverses.



- Dependence between scalars via correlation between vectors

② $k_{\max}^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$

We are solving a maximum correlation problem on the high dimensional transformed data

$u'' = \mathbf{f}^H\mathbf{u}$

$\max_{\mathbf{f},\mathbf{g}} r_{u''v''}^2$

$\|\mathbf{f}\|^2 = 1$

$\|\mathbf{g}\|^2 = 1$

$v'' = \mathbf{g}^H\mathbf{v}$

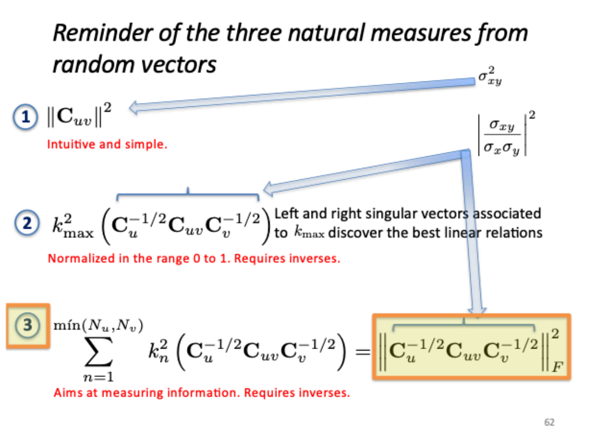
Since u and v are "frequency components", we are solving the HGR problem under imposed smoothness of the nonlinear functions f and g.

The used dimension N…

imposes the allowed smoothness.

arises as a "band-pass filter" on the problem (regularization).

arises as a natural performance / complexity trade-off.



Reminder of the three natural measures from random vectors

σ_{xy}^2

① $\|\mathbf{C}_{uv}\|^2$

Intuitive and simple.

$\left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right|^2$

② $k_{\max}^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right)$ Left and right singular vectors associated to $k_{\max}$ discover the best linear relations

Normalized in the range 0 to 1. Requires inverses.

③ $\sum_{n=1}^{\text{mín}(N_u,N_v)} k_n^2\left(\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right) = \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$

Aims at measuring information. Requires inverses.

- Dependence between scalars via correlation between vectors

$$③\ \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$$

Measure of information between two phenomena based on second order statistics only.

Is it Shannon mutual information?

No! (Thesis).
Instead, it measures the so-called **Squared-Loss Mutual Information**

- Dependence between scalars via correlation between vectors

$$③\ \left\|\mathbf{C}_u^{-1/2}\mathbf{C}_{uv}\mathbf{C}_v^{-1/2}\right\|_F^2$$

Shannon mutual information:

$$I_{xy} = \int\int p_{xy}(x,y)\ln\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}dxdy \begin{matrix}\text{independence}\\ \overset{=}{>}\ 0\\ \text{dependence}\end{matrix}$$

Squared-Loss Mutual Information

$$i_{xy} = \int\int \frac{\left(p_{xy}(x,y) - p_x(x)p_y(y)\right)^2}{p_x(x)p_y(y)}dxdy \begin{matrix}\text{independence}\\ \overset{=}{>}\ 0\\ \text{dependence}\end{matrix}$$

$i_{xy} \geq I_{xy}$	Upper-bounds Shannon MI
$\lim\limits_{I_{xy}\to 0}\frac{i_{xy}}{2I_{xy}} = 1$	Measures Shannon MI locally

---

- Conditional dependence

$$p_{xy|z}(x,y) \ \overset{=}{\neq}\ p_{x|z}(x)p_{y|z}(y)$$

$\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathbf{z}}$ Measure conditional dependence from the conditional covariance matrix.

- Conditional dependence

Virtual sources:

$$\mathring{\mathbf{u}} = \frac{\mathbf{u}_1 - \mathbf{u}_2}{\sqrt{2}} \qquad \mathring{\mathbf{v}} = \frac{\mathbf{v}_1 - \mathbf{v}_2}{\sqrt{2}} \qquad \mathring{z} = \frac{z_1 - z_2}{\sqrt{2}}$$

Main idea:

$$\begin{aligned}\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathbf{z}} &\triangleq \int_{\mathbb{R}}\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathbf{z}=z}\,\mathrm{d}F_{\mathbf{z}}(z) = \int_{\mathbb{R}}\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathbf{z}=z}\,\mathrm{d}F_{\mathbf{z}}(z) \quad ☹\\ &= \int_{\mathbb{R}^2}\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathbf{z}}=0}\,\mathrm{d}F_{\mathbf{z}}(z_1)\,\mathrm{d}F_{\mathbf{z}}(z_2)\\ &= \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathbf{z}}=0}\int_{\mathbb{R}^2}\mathrm{d}F_{\mathbf{z}}(z_1)\,\mathrm{d}F_{\mathbf{z}}(z_2)\\ &= \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathbf{z}}=0} \quad ☺\end{aligned}$$

- Conditional dependence

Idea: U-STATISTICS

$$\hat{\mathbf{C}}_{uv} = \frac{1}{\underbrace{L(L-1)}_{\text{data pairs!!!}}}\sum_{1\leq l_1<l_2\leq L}^{L}\left(\mathbf{u}(l_1) - \mathbf{u}(l_2)\right)\left(\mathbf{v}(l_1) - \mathbf{v}(l_2)\right)^H$$

data centering not needed!

Wassily Hoeffding (1914-1991)

- Conditional dependence

Use incomplete U-statistics to compute correlations

$$\hat{\mathbf{C}}_{uv} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{u}(l_1(k)) - \mathbf{u}(l_2(k))\right)\left(\mathbf{u}(l_1(k)) - \mathbf{u}(l_2(k))\right)^H$$

$$K < \frac{L(L-1)}{2}$$

Pairs of data selected according to small values of $\|\mathring{z}\|^2$, identified by sorting.

Sorting distance pairs is one of the fundamental problems in computer science.

Conditional dependence: modeling example

Co-information: $\mathrm{I}(x;y;z) = \mathrm{I}(x;y) - \mathrm{I}(x;y|z)$

$$\mathcal{M}^+ : \begin{cases} x = \sqrt{\gamma}ap + v \\ y = \sqrt{\gamma}aq + w \\ z = a \end{cases} \qquad \mathcal{M}^- : \begin{cases} x = \sqrt{\gamma}bp + v \\ y = \sqrt{\gamma}cq + w \\ z = b - c \end{cases}$$

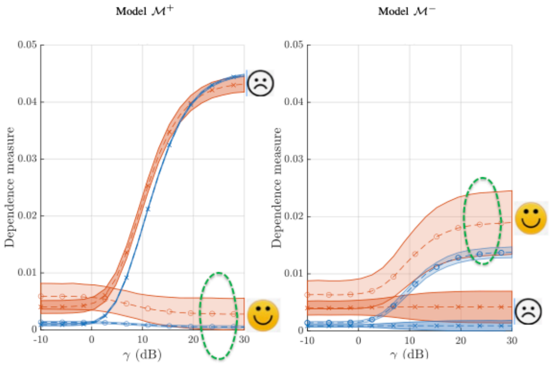
$a, b, c \sim \mathcal{U}(0, \sqrt{3})$

$v, w \sim \mathcal{N}(0, 1)$

$p, q \sim \mathrm{Bern}_{1/2}\{-1, 1\}$

Parameter γ controls the total amount of absolute co-information.

Conditional dependence: modeling example



Conditional dependence

- Submitted to IEEE Signal Processing Letters:

Conditional Dependence via U-Statistics Pruning

Ferran de Cabrera, Marc Vilà-Insa, *Graduate Student Member, IEEE*, and Jaume Riba, *Senior Member, IEEE*

Available at: https://arxiv.org

Proposal of toy problems for team work

- Proposed parameters:
  - Number of pairs: K = 100
  - Number of samples per pair: L = 100
  - Oracle quality: $\dfrac{I_{xy}}{\sigma_{\hat{I}_{xy,\mathrm{oracle}}}} = 10$
- Form T teams of two persons each.

Proposal of toy problems for team work

- Generation task for each team:
  - Invention of a model to generate dependent data.
  - Generate K pairs of data (L samples each), all with identical marginal statistics:
    - 50% are independent pairs
    - 50% are dependent pairs
  - Calibrate the difficulty of the model with an oracle estimator.

Proposal of toy problems for team work

- Analysis task for each team:
  - Chose another team to interchange the generated data pairs, without telling which pairs are dependent.
  - Develop some measure of statistical dependence from the explained ideas.
  - Apply the developed measure to the available data.
  - Classify the pairs in independent and dependent.
  - Provide your results to the team that generated the data.

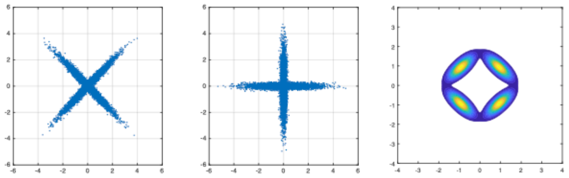
- Evaluation of the analysis task for each team:
  - Evaluate the task of the team that has provided results based on your data, by telling the percentage of error in the classification.
  - Make all the analysis scores public to all.
- Evaluation of the modelling task for each team:
  - The worst is the score that you give to the other team, the best is the score that you receiver on modelling.
  - But other team can now evaluate your data to check your model and make you to decrease your score on modelling.

$$I_{xy} = E \ln \frac{p_{xy}(x,y)}{p_x(x)p_y(y)}$$

An oracle (pdfs known) estimator (benchmark) is based on the sample average from the data obtained from dependent pairs:

$$\hat{I}_{xy,\text{oracle}} = \frac{1}{L}\sum_{l=1}^{L} \ln \frac{p_{xy}(x_l,y_l)}{p_x(x_l)p_y(y_l)}$$

Example: Gaussian Mixture Models (GMM)



Dependent data sequences are obtained from X and Y samples at same time.

Independent data sequences are obtained from X and Y samples at different time.

- ***Contributions***
  - Increase the data dimensionality and find **linear dependences** there *instead of* find **non-linear dependences** directly.
  - Linear-phase complex vectors "steerings" emerge as **universal maps** to increase dimensionality in a "regularized manner".
  - Classical **second-order** analysis schemes are **reborn** as natural tools for measuring information in data.
  - **Kernel methods appear** when the dimension tends to infinity.
  - Measures of **information** (different from Shannon) emerge as natural **surrogates** for handling data.
  - **Doctoral thesis**: F. De Cabrera: "Data-driven information-theoretic tools under a second-order statistic perspective".
  - Interest in **cross-disciplinary links** and **fresh perspectives**.
- **Mathematicians, please feel free to contact us (*Jaume.riba@upc.edu*) for exploring potential links!!!**

"all things are delicately interconnected", J. Holzer