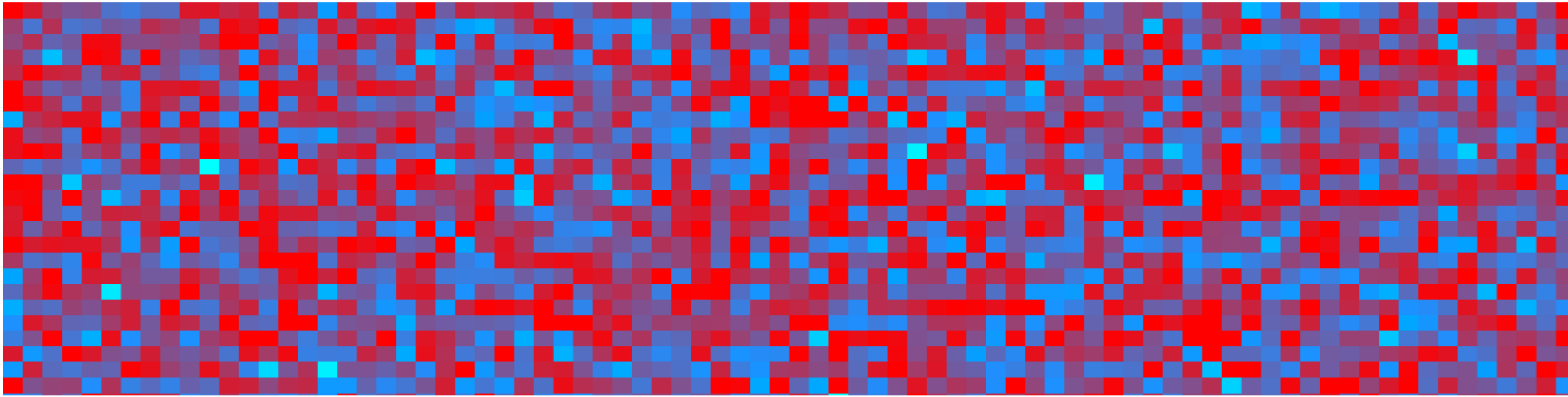
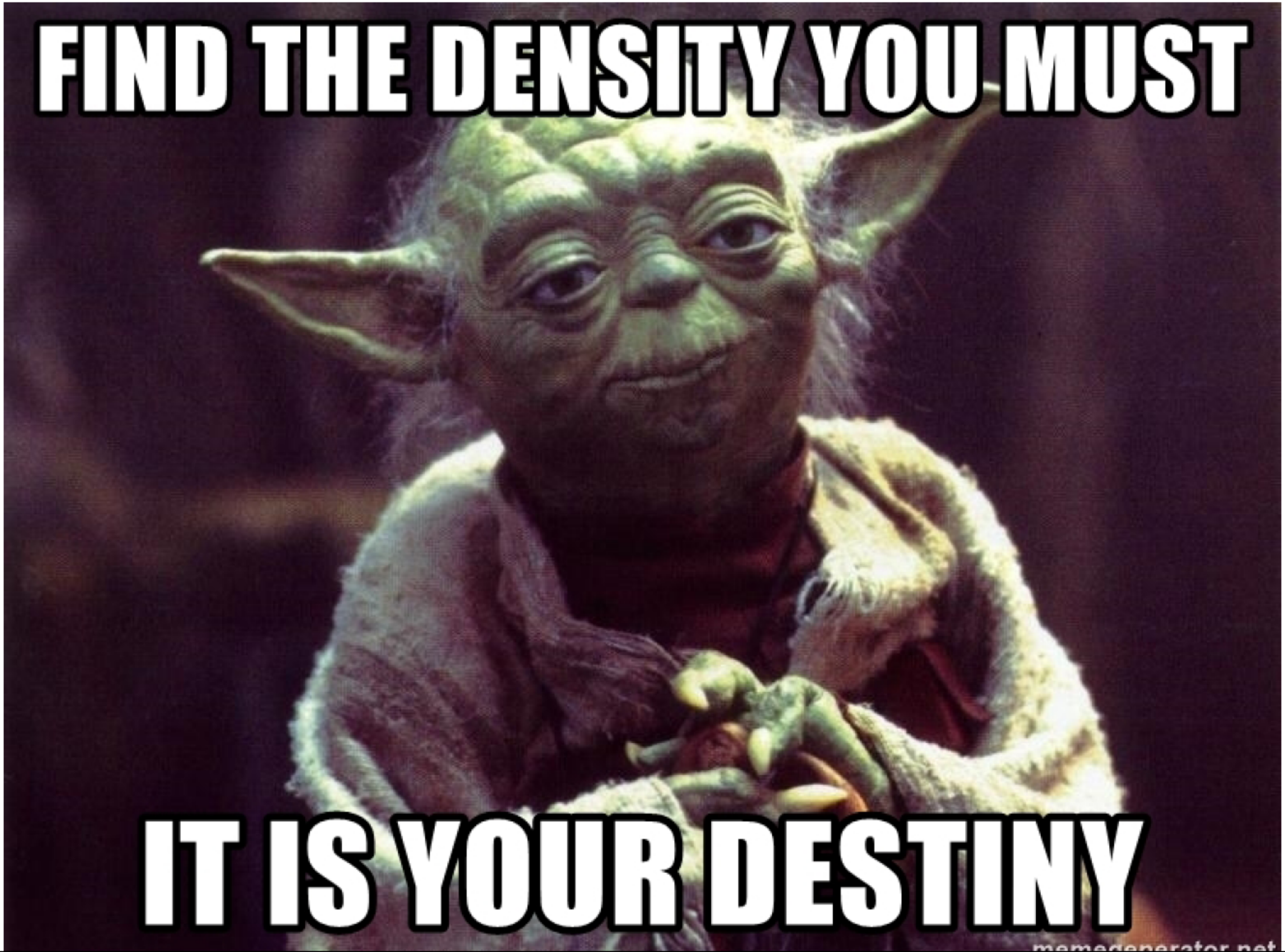
Gaussianizing the Earth



Gustau Camps-Valls
Image Processing Laboratory (IPL)
Universitat de València

gustau.camps@uv.es
http://isp.uv.es
@isp_uv_es


FIND THE DENSITY YOU MUST
IT IS YOUR DESTINY

PDF estimation is the core of statistics, machine learning and info theory

$$\sum_x p(x) \log p(x) \; \sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x) \; \sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2$$

$$\sum_x p(x) \log p(x) \; \sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x) \; \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \mathrm{d}x$$

$$\sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \; \sum_x p(x) \log p(x) \; \sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x)$$

$$\sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \; \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \mathrm{d}x \; \sum_x p(x) \log p(x)$$

$$\sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x) \; \sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \; \sum_x p(x) \log p(x)$$

$$\sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x) \; \sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \; \sum_x p(x) \log p(x)$$

$$\sum_{xy} p(x,y) \log p(y|x) \; \sum_x p(x) \log q(x) \; \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \mathrm{d}x$$

$$\sum_x p(x) \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2$$

Gaussianization for PDF estimation

Why? Statistical independence of features is useful to ...

- ... process dimensions independently, no dim curse
- ... tackle the PDF estimation problem directly
- ... and estimate multivariate IT measures

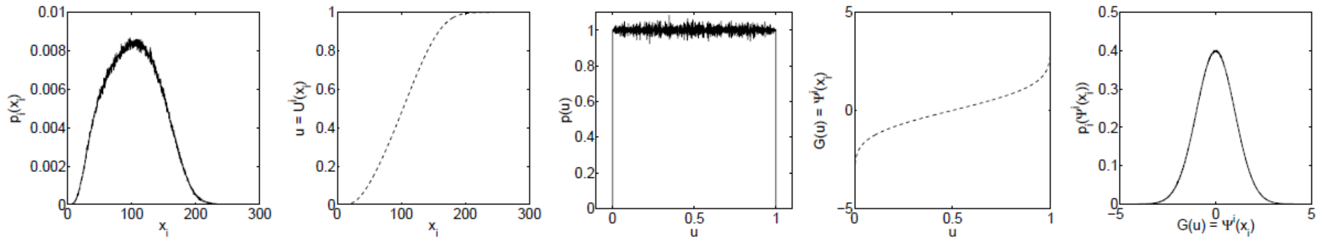
Marginal (univariate) Gaussianization is easy!

Gaussianization in each dimension, $\mathbf{\Psi}^i_{(k)}$, can be decomposed into two consecutive equalization transforms:

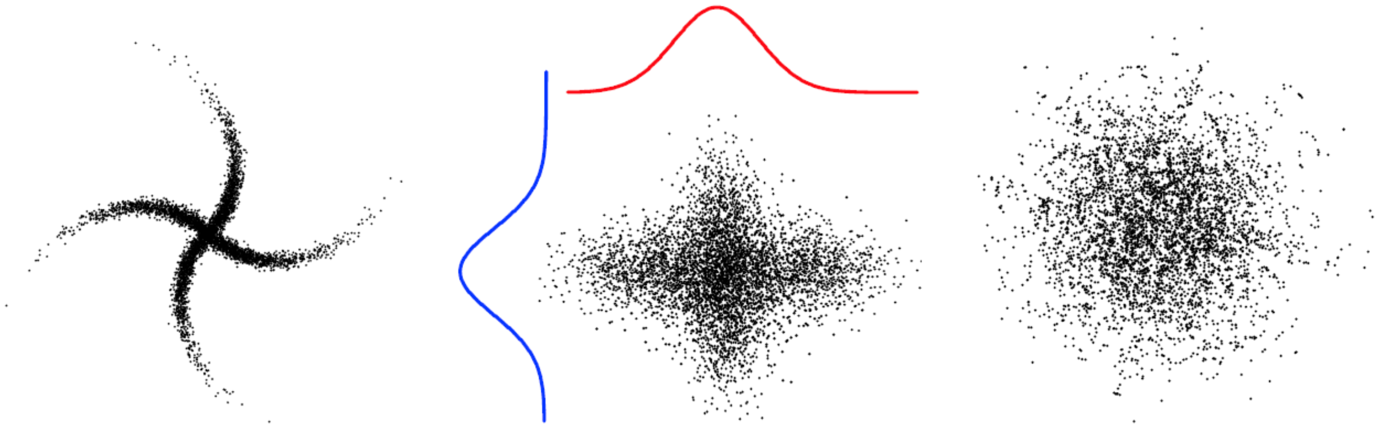
1. Marginal uniformization, $U^i_{(k)}$, based on the cdf of the marginal PDF
2. Gaussianization of a uniform variable, $G(u)$, based on the inverse of the cdf of a univariate Gaussian: $\mathbf{\Psi}^i_{(k)} = G \odot U^i_{(k)}$

$$u = U^i_{(k)}(x_i^{(k)}) \;=\; \int_{-\infty}^{x_i^{(k)}} p_i(x_i'^{(k)})\, dx_i'^{(k)}$$

$$G^{-1}(x_i) \;=\; \int_{-\infty}^{x_i} g(x_i')\, dx_i'$$



Original | **Marginal Gaussianization** | **Multivariate Gaussianization**



RBIG = Rotate and marginally Gaussianize

- √ An orthogonal transform does not affect Gaussianity
- √ Univariate Gaussianization is unique

Rotation-based Iterative Gaussianization (RBIG)

Definition

Given a D-dimensional random variable $\mathbf{x}^{(0)} = [x_1, \ldots, x_D]^\top$ do

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{R}_{(k)} \boldsymbol{\Psi}_{(k)}(\mathbf{x}^{(k)}), \quad k = 1, \ldots, K$$

Rotation-based Iterative Gaussianization (RBIG)

Definition

Given a D-dimensional random variable $\mathbf{x}^{(0)} = [x_1, \dots, x_D]^\top$ do

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{R}_{(k)} \mathbf{\Psi}_{(k)}(\mathbf{x}^{(k)}), \quad k = 1, \dots, K$$

Properties

- √ is invertible and differentiable
- √ is valid under any rotation transform (PCA, ICA, random!)
- √ converges! (negentropy and MI reduce in each iteration)
- √ is fast (only marginal Gaussianization and rotations needed)
- √ is a deep neural net! (normalizing flow)
- × is relatively robust to high-dim spaces
- × is a meaningless transform

The change-of-variables formula

Let $\mathbf{x} \in \mathbb{R}^D$ be a r.v. with PDF $p_{\mathbf{x}}(\mathbf{x})$. Given some bijective, differentiable transform of $\mathbf{x}$ into $\mathbf{y}$ using $\mathcal{G} : \mathbb{R}^D \rightarrow \mathbb{R}^D$, $\mathbf{y} = \mathcal{G}(\mathbf{x})$, the PDFs are related:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x})) \left| \frac{d\mathcal{G}(\mathbf{x})}{d\mathbf{x}} \right| = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x})) \left| \nabla_{\mathbf{x}} \mathcal{G}(\mathbf{x}) \right|$$

where $|\nabla_{\mathbf{x}} \mathcal{G}|$ is the determinant of the transform's Jacobian matrix

RBIG for density estimation, $p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x}))\ |\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x})|$

- PDF of a multivariate Gaussian:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x})) \propto \exp(-\frac{1}{2}(\mathcal{G}(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{y}})^\top \boldsymbol{\Sigma}^{-1}(\mathcal{G}(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{y}}))$$

- Jacobian is the product of Jacobians:

$$\nabla_{\mathbf{x}}\mathcal{G} = \prod_{k=1}^{K} \mathbf{R}_{(k)} \nabla_{\mathbf{x}^{(k)}} \boldsymbol{\Psi}_{(k)}$$

$$\nabla_{\mathbf{x}^{(k)}} \boldsymbol{\Psi}_{(k)} = \begin{pmatrix} \frac{\partial \boldsymbol{\Psi}^1_{(k)}}{\partial x_1^{(k)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial \boldsymbol{\Psi}^d_{(k)}}{\partial x_d^{(k)}} \end{pmatrix}, \quad \frac{\partial \boldsymbol{\Psi}^i_{(k)}}{\partial x_i^{(k)}} = \frac{\partial \mathcal{G}}{\partial u} \frac{\partial u}{\partial x_i^{(k)}}$$

- Invertible:

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{R}_{(k)} \boldsymbol{\Psi}_{(k)}(\mathbf{x}^{(k)}) \rightarrow \mathcal{G}^{-1} : \mathbf{x}^{(k)} = \boldsymbol{\Psi}^{-1}_{(k)}(\mathbf{R}^\top_{(k)} \mathbf{x}^{(k+1)})$$

Theorem 1: negentropy reduces independently of the rotation

$$\Delta J = J(\mathbf{x}) - J(\mathbf{R}\boldsymbol{\Psi}(\mathbf{x})) \geq 0, \; \forall\, \mathbf{R}$$

Proof.

Divergence to a factorized PDF written in terms of its marginal PDFs [Cardoso03]:

$$\begin{aligned} \mathrm{D_{KL}}(p(\mathbf{x}) \mid \textstyle\prod_i q_i(x_i)) &= \mathrm{D_{KL}}(p(\mathbf{x}) \mid \textstyle\prod_i p_i(x_i)) + \mathrm{D_{KL}}(\textstyle\prod_i p_i(x_i) \mid \textstyle\prod_i q_i(x_i)) \\ &= I(\mathbf{x}) + \mathrm{D_{KL}}(\textstyle\prod_i p_i(x_i) \mid \textstyle\prod_i q_i(x_i)) \end{aligned}$$

If $q_i(x_i)$ are univariate Gaussian PDFs, $\prod_i q_i(x_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then:

$$J(\mathbf{x}) = I(\mathbf{x}) + J_m(\mathbf{x})$$

The negentropy reduction in our transform is:

$$\begin{aligned} \Delta J &= J(\mathbf{x}) - J(\mathbf{R}\boldsymbol{\Psi}(\mathbf{x})) = J(\mathbf{x}) - J(\boldsymbol{\Psi}(\mathbf{x})) \\ &= I(\mathbf{x}) + J_m(\mathbf{x}) - I(\boldsymbol{\Psi}(\mathbf{x})) - J_m(\boldsymbol{\Psi}(\mathbf{x})) = J_m(\mathbf{x}) \geq 0, \; \forall\, \mathbf{R} \end{aligned}$$

since: (1) $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is rotation invariant; (2) the I is invariant under dim-wise transforms [Studeny98]; and (3) the J_m of a marginally Gaussianized r.v. is 0. □

Theorem 2: redundancy reduces independently of the rotation

Given a marginally Gaussianized variable, $\mathbf{\Psi}(\mathbf{x})$, any rotation reduces the redundancy:

$$\Delta I = I(\mathbf{\Psi}(\mathbf{x})) - I(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \geq 0, \; \forall \, \mathbf{R}$$

Proof.

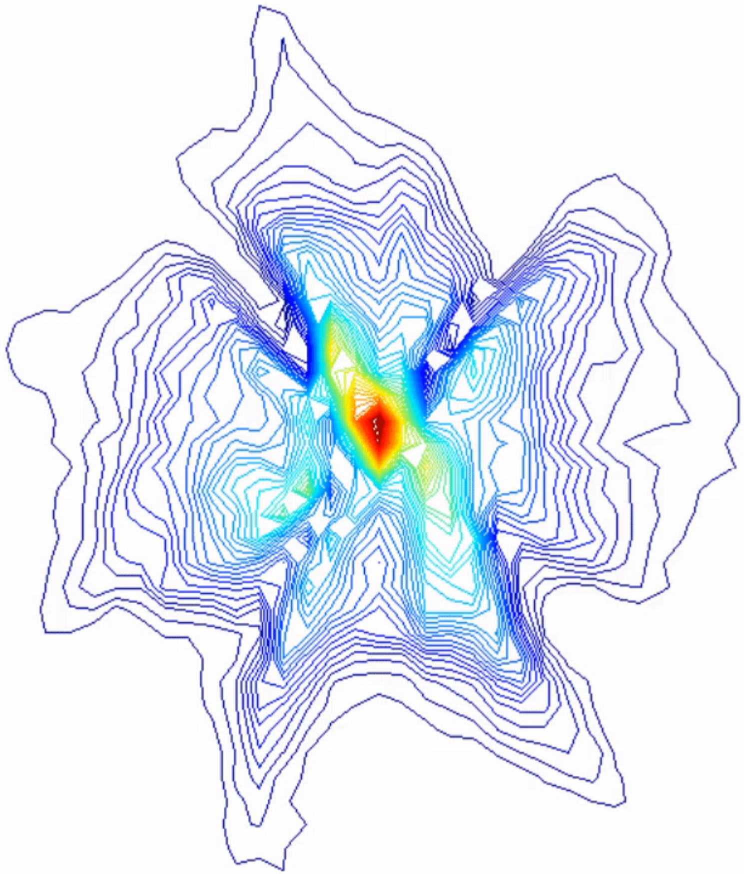
Remember

$$J(\mathbf{x}) = I(\mathbf{x}) + J_m(\mathbf{x}) \rightarrow I(\mathbf{x}) = J(\mathbf{x}) - J_m(\mathbf{x})$$

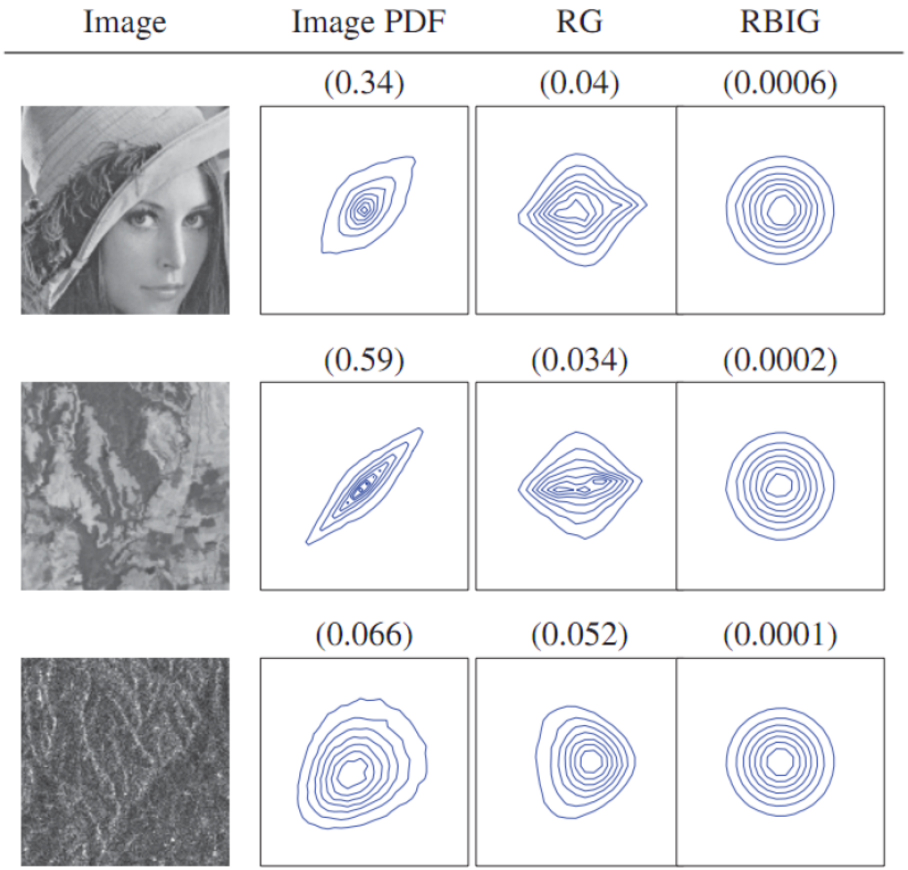
Apply it on $I(\mathbf{\Psi}(\mathbf{x}))$ and $I(\mathbf{R}\mathbf{\Psi}(\mathbf{x}))$:

$$\begin{aligned} \Delta I &= J(\mathbf{\Psi}(\mathbf{x})) - J_m(\mathbf{\Psi}(\mathbf{x})) - J(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) + J_m(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \\ &= J_m(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \geq 0, \; \forall \, \mathbf{R} \end{aligned}$$

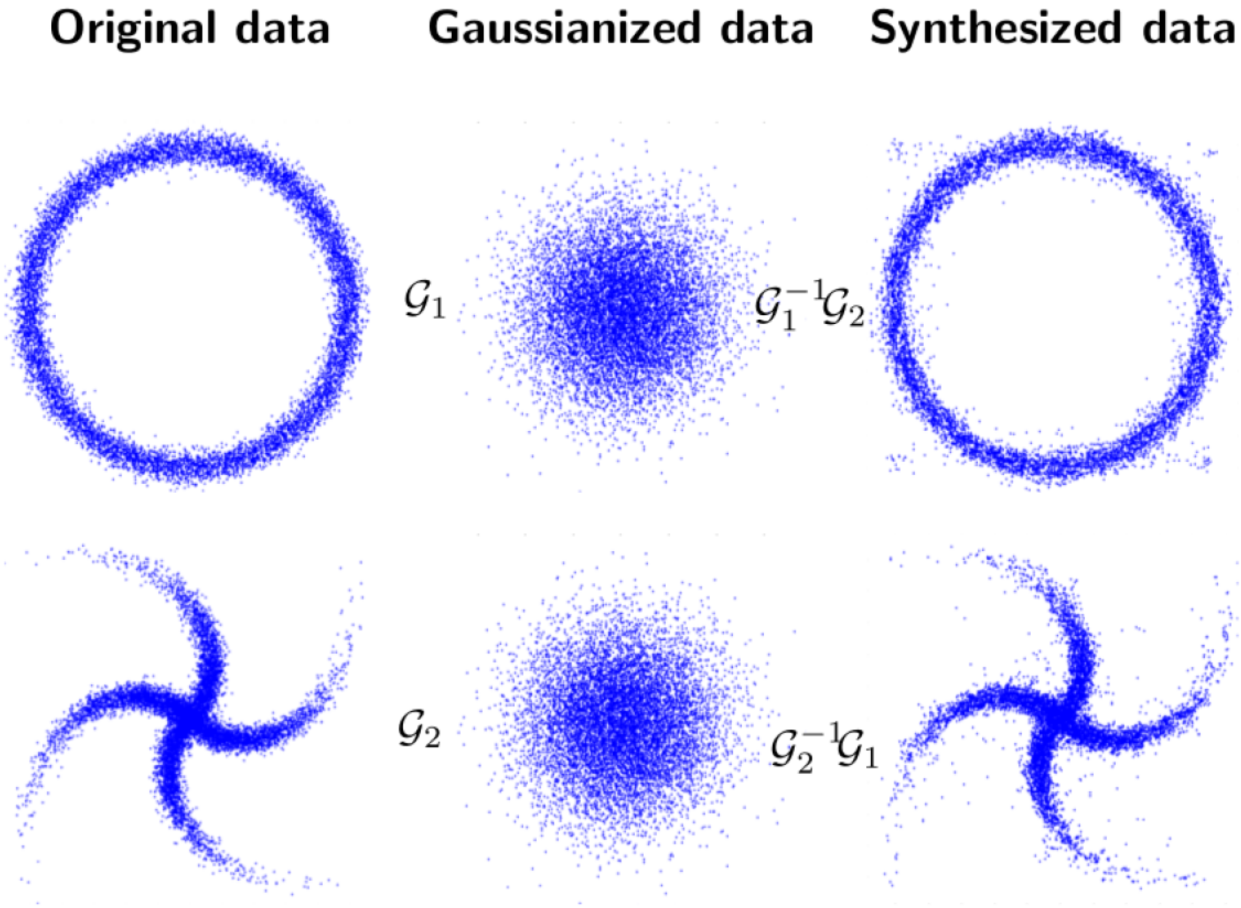
since (1) negentropy is rotation invariant, and (2) the marginal negentropy of a marginally Gaussianized r.v. is 0. □



It works in arbitrary natural and remote sensing images

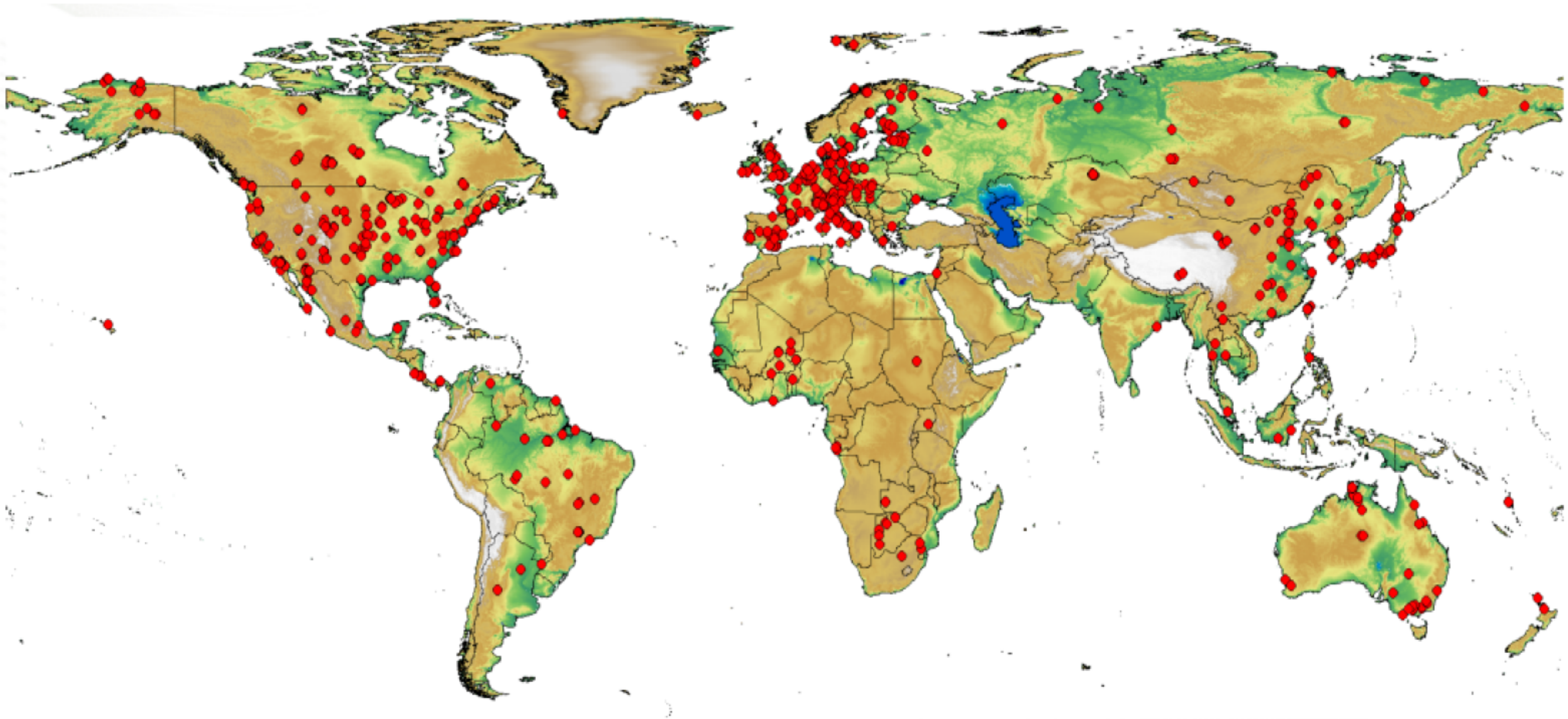


Synthesis in low dimensions



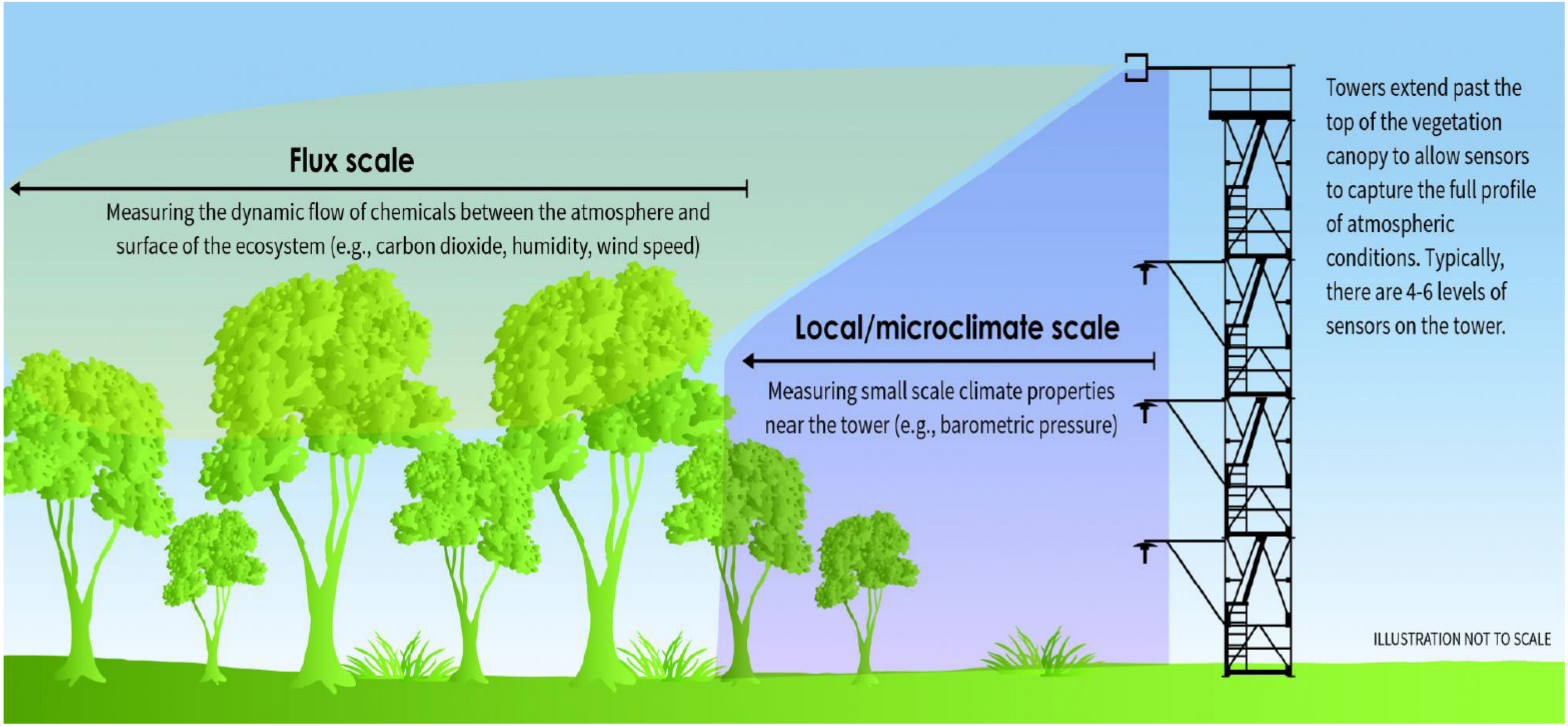
Synthesis

Synthesis in moderate dimensions



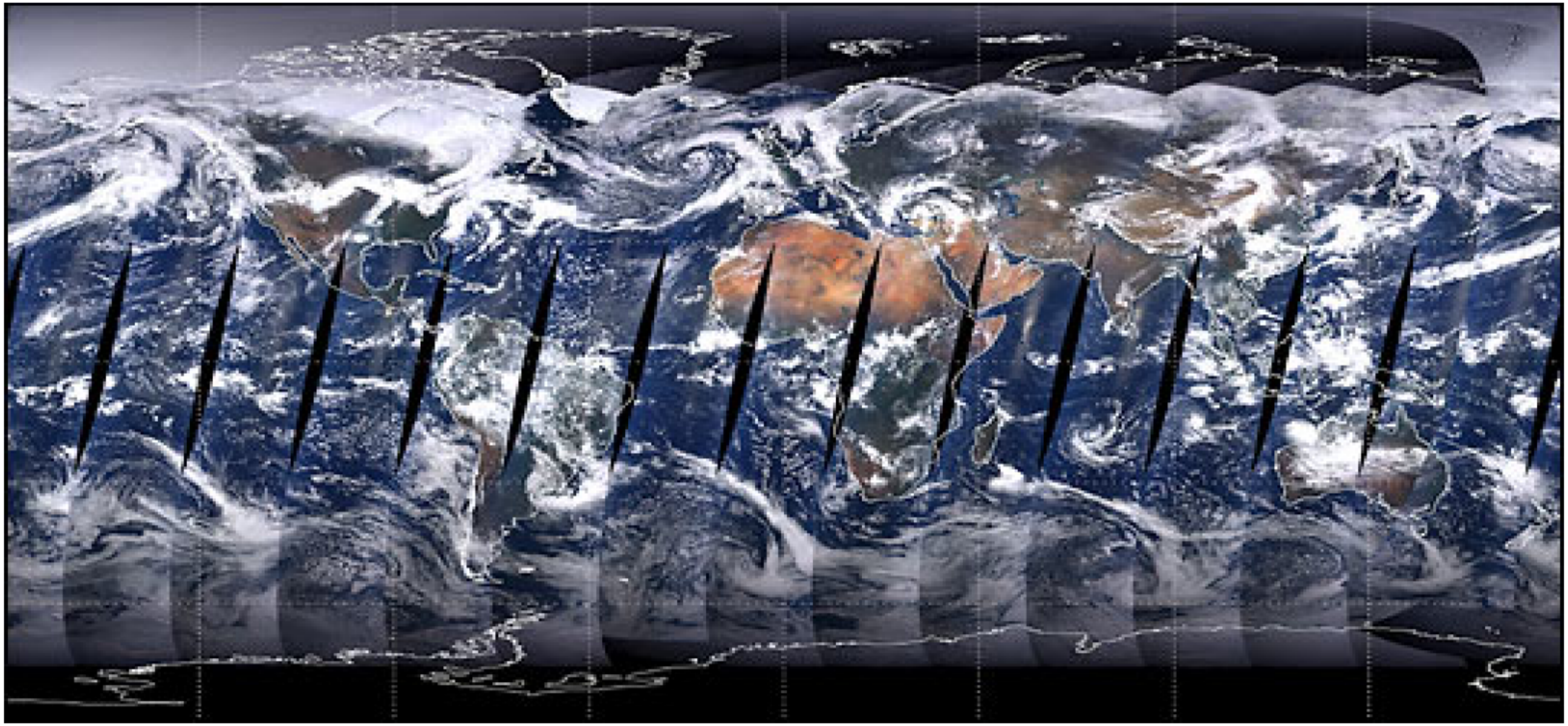
FLUXNET: a network for micro-meteorological tower sites

Synthesis in moderate dimensions



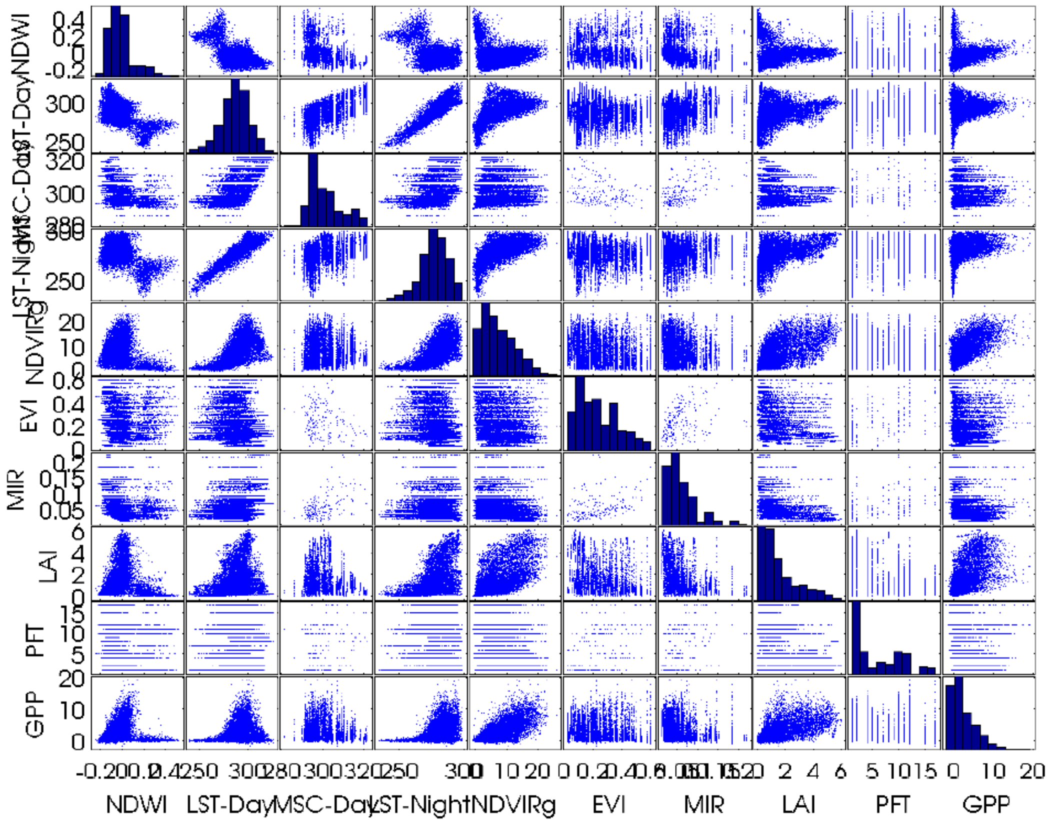
Sensors allow estimating turbulent exchange of carbon dioxide (CO_2), latent and sensible heat, CO_2 storage, net ecosystem exchange, energy balance, ...

Synthesis in moderate dimensions

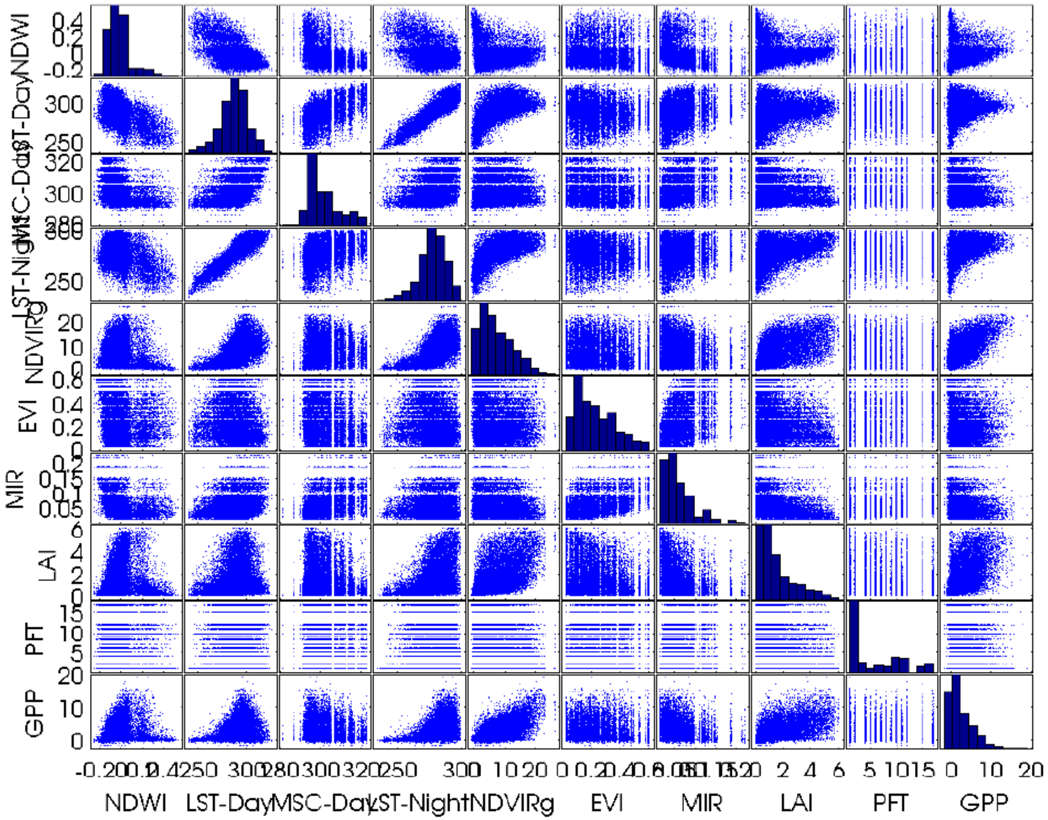


MODIS sensor: 36 channels, 8-daily, 500 m

Synthesis in moderate dimensions



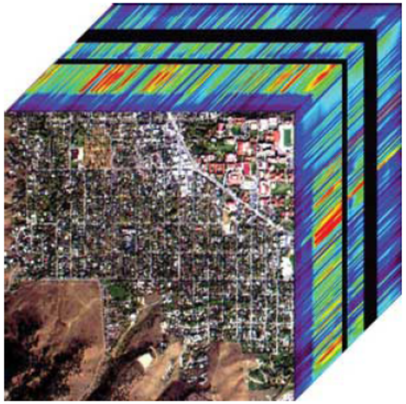
Synthesis in moderate dimensions

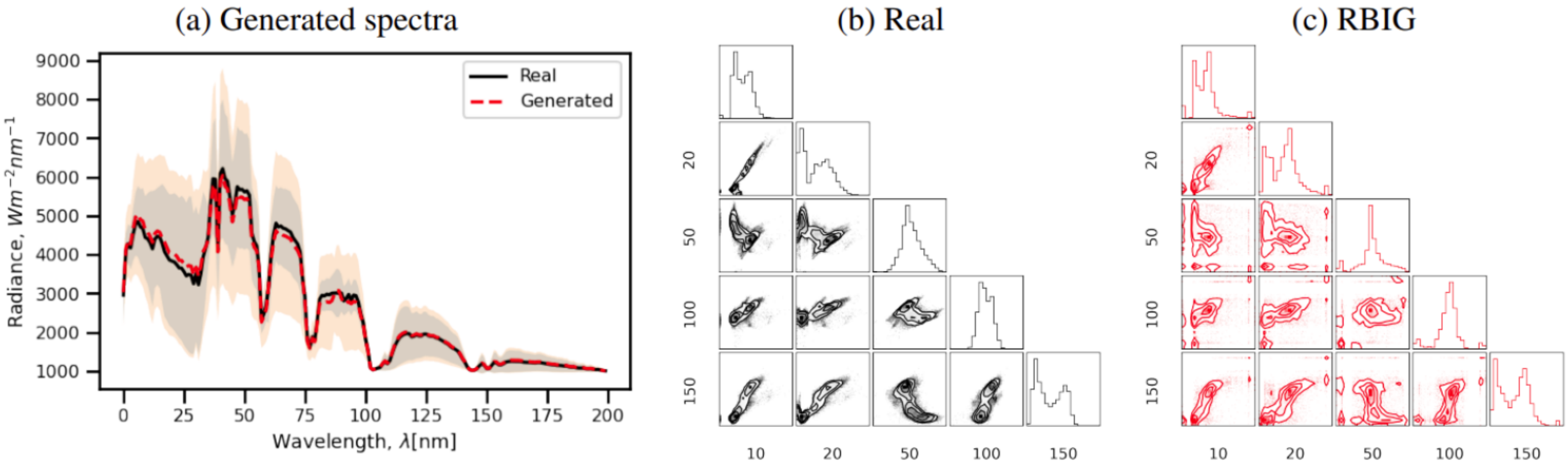


Synthesis in moderate dimensions

real, $n = 10^4$	ME	RMSE	MAE	R
LR	-0.01	1.82	1.28	0.78
GPR	+0.03	1.72	1.14	**0.81**
real+syn, $n = 10^6$	ME	RMSE	MAE	R
LR	-0.01	1.80	1.27	0.79
GPR	**-0.00**	**1.63**	**1.03**	**0.83**

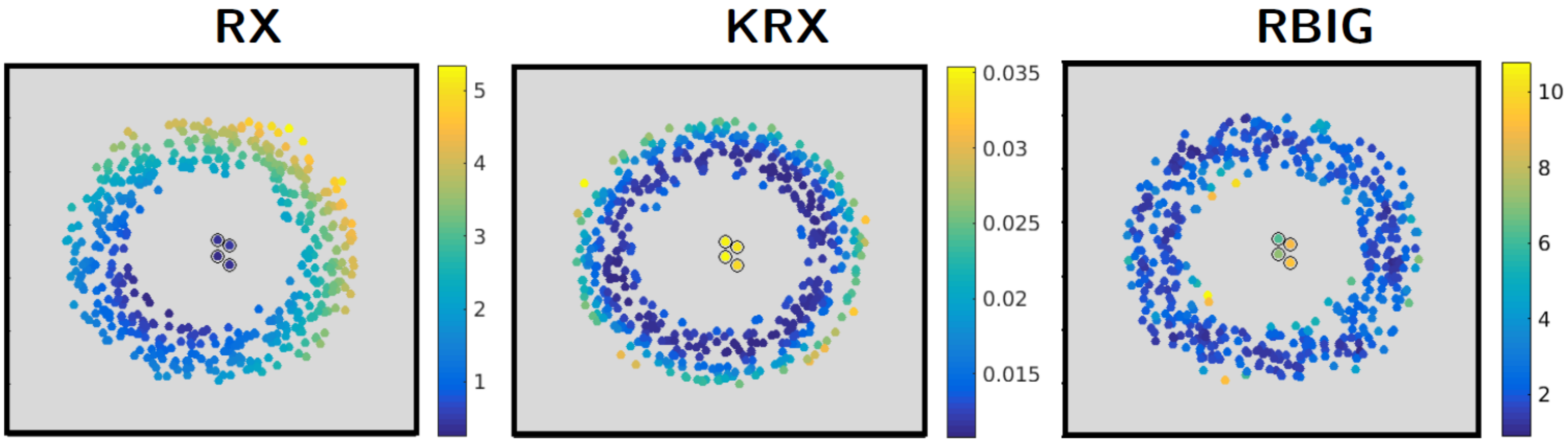
Synthesis in very high dimensions



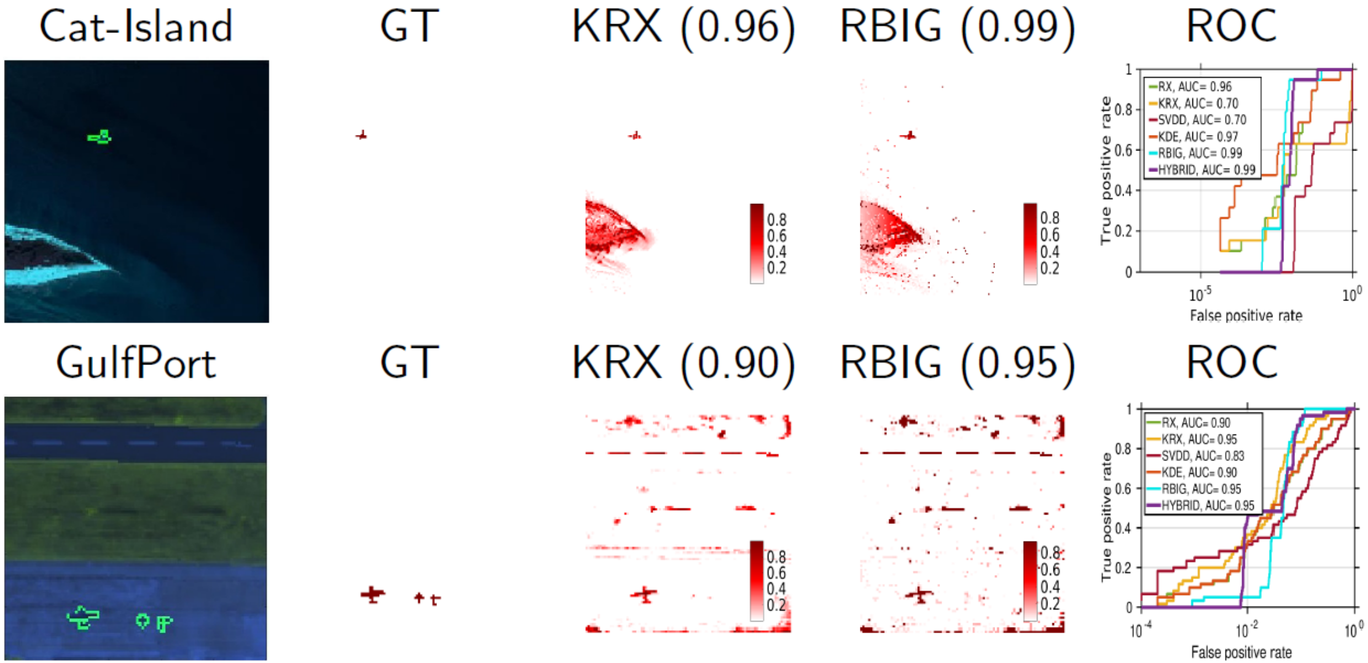


Anomaly and extreme detection

$$A_{\mathrm{RX}}(\mathbf{x}) = (\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)$$
$$A_{\mathrm{KRX}}(\mathbf{x}) = \tilde{k}(\mathbf{x},\cdot)^\top (\mathbf{K}\mathbf{K})^{-1}\tilde{k}(\mathbf{x},\cdot)$$
$$A_{\mathrm{RBIG}}(\mathbf{x}) \propto \frac{1}{p_{\mathrm{RBIG}}(\mathbf{x})}$$



Anomaly and extreme detection



RBIG framework allows to compute all IT measures

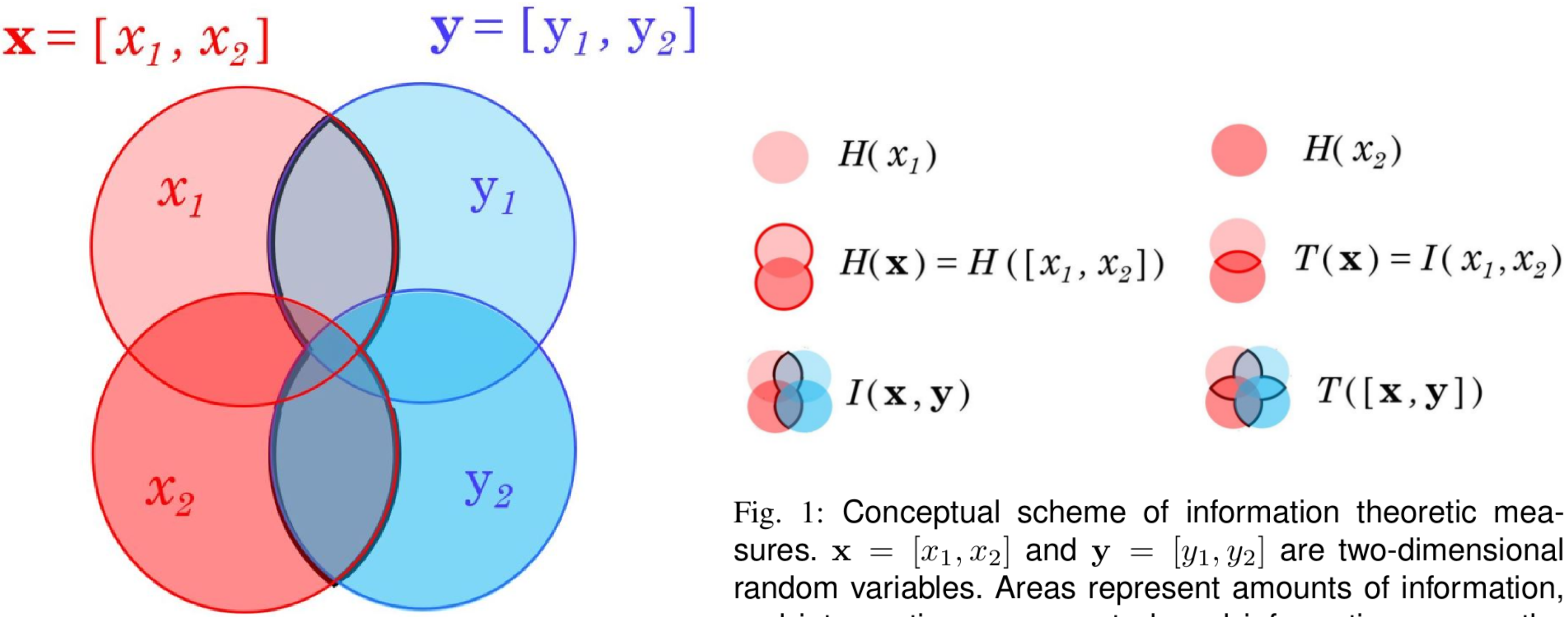


Fig. 1: Conceptual scheme of information theoretic measures. $\mathbf{x} = [x_1, x_2]$ and $\mathbf{y} = [y_1, y_2]$ are two-dimensional random variables. Areas represent amounts of information, and intersections represent shared information among the corresponding variables and dimensions. Examples of entropy, total correlation and mutual information are given.

RBIG framework allows to compute all IT measures

1. Total Correlation (aka multi-information)

$$TC = \sum_{k=1}^{K} \left(D\ h(\mathcal{N}(0,1)) + \sum_{d=1}^{D} h(\mathbf{x}_d^{(k)}) \right)$$

RBIG framework allows to compute all IT measures

1. Total Correlation (aka multi-information)

$$TC = \sum_{k=1}^{K} \left(D\ h(\mathcal{N}(0,1)) + \sum_{d=1}^{D} h(\mathbf{x}_d^{(k)}) \right)$$

2. Multidimensional entropy (and negentropy):

$$H(\mathbf{x}) = \sum_{d=1}^{D} h(\mathbf{x}_d) - TC(\mathbf{x})$$

RBIG framework allows to compute all IT measures

1. Total Correlation (aka multi-information)

$$TC = \sum_{k=1}^{K} \left(D \ h(\mathcal{N}(0,1)) + \sum_{d=1}^{D} h(\mathbf{x}_d^{(k)}) \right)$$

2. Multidimensional entropy (and negentropy):

$$H(\mathbf{x}) = \sum_{d=1}^{D} h(\mathbf{x}_d) - TC(\mathbf{x})$$

3. Kullback-Leibler divergence: $D_{KL}(\mathbf{x}\|\mathbf{y}) = TC(\mathcal{G}_x(\mathbf{y}))$

RBIG framework allows to compute all IT measures

1. Total Correlation (aka multi-information)

$$TC = \sum_{k=1}^{K} \left(D\ h(\mathcal{N}(0,1)) + \sum_{d=1}^{D} h(\mathbf{x}_d^{(k)}) \right)$$

2. Multidimensional entropy (and negentropy):

$$H(\mathbf{x}) = \sum_{d=1}^{D} h(\mathbf{x}_d) - TC(\mathbf{x})$$

3. Kullback-Leibler divergence: $D_{KL}(\mathbf{x}\|\mathbf{y}) = TC(\mathcal{G}_x(\mathbf{y}))$
4. Conditional independence

$$\begin{aligned} I(\mathbf{x},\mathbf{y}|\mathbf{z}) &= H(\mathbf{x},\mathbf{z}) + H(\mathbf{y},\mathbf{z}) - H(\mathbf{x},\mathbf{y},\mathbf{z}) - H(\mathbf{z}) \\ &= TC(\mathbf{x},\mathbf{y},\mathbf{z}) - TC(\mathbf{x},\mathbf{z}) - TC(\mathbf{y},\mathbf{z}) \end{aligned}$$

with the null hypothesis distribution $p(I(\mathbf{x}, r(\mathbf{y})|\mathbf{z}))$

But ... how to estimate total correlation?

1: Given data $\mathbf{x}^{(0)} = [x_1, \dots, x_D]^\top \in \mathbb{R}^D$
2: Learn the sequence of Gaussianization transforms $\mathbf{y} = \mathcal{G}(\mathbf{x})$
3: Compute the cumulative reduction in mutual information

$$TC = \sum_{k=1}^{K} \left(D\ h(\mathcal{N}(0,1)) + \sum_{d=1}^{D} h(\mathbf{x}_d^{(k)}) \right)$$

Total Correlation

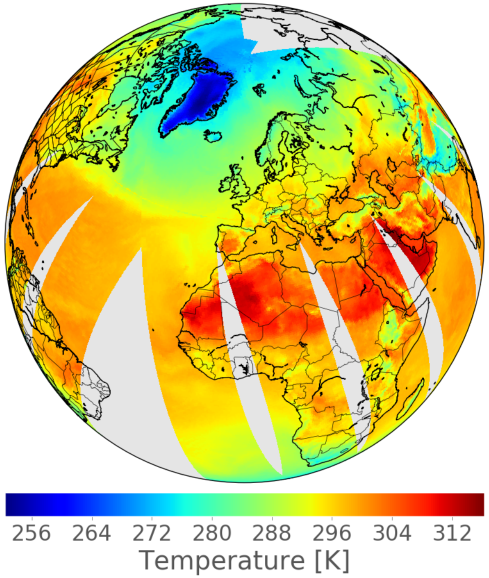
TABLE 1: Relative mean absolute errors in percentage for total correlation estimation on known PDFs. Best value in dark gray, second best value in bright gray.

		D_x	RBIG	kNN	KDP	expF	vME	Ens
Gaussian	-	3	1.5	2.5	159.2	1.2	8.5	9.8
		10	3.1	31.2	>100	0.2	33.9	44.9
		50	1.3	32.7	>100	0.1	>100	38.7
		100	0.8	31.0	89.9	0.1	94.2	34.9
Rotated	-	3	1.70	1.80	82.90	16.80	1.90	9.40
		10	8.30	27.20	>100	11.00	24.20	38.70
		50	7.70	51.10	>100	15.10	>100	59.40
		100	7.50	57.80	>100	15.50	>100	64.50
Student	$\nu = 3$	3	7.01	13.55	>100	94.03	>100	66.59
		10	32.93	16.73	>100	67.32	>100	15.27
		50	18.18	12.02	>100	29.44	>100	24.65
		100	12.71	17.41	>100	21.12	>100	28.63
	$\nu = 5$	3	26.61	52.76	>100	89.74	81.85	133.12
		10	23.94	19.74	>100	49.60	>100	12.31
		50	10.10	16.87	>100	20.29	>100	32.14
		100	7.10	22.53	>100	15.39	>100	34.96
	$\nu = 20$	3	88.27	>100	>100	48.56	>100	>100
		10	3.05	11.86	>100	10.51	>100	19.93
		50	3.07	33.17	>100	4.54	>100	52.62
		100	1.31	35.56	>100	3.43	>100	49.46

TABLE 2: Relative mean absolute errors in percentage for entropy estimation on known PDFs. Best value in dark gray, second best value in bright gray.

		D_x	RBIG	kNN	KDP	expF	vME	Ens
Gaussian		3	1.5	2.5	159.2	1.2	8.5	9.8
		10	3.1	31.2	>100	0.2	33.9	44.9
		50	1.3	32.7	>100	0.1	>100	38.7
		100	0.8	31.0	89.9	0.1	94.2	34.9
Rotated		3	2.8	4.7	127.2	37.2	3.6	22.7
		10	17.4	45.2	263.8	23.9	4.5	62.0
		50	7.6	46.0	140.2	14.2	87.6	53.1
		100	5.2	43.50	113.9	12.1	94.3	48.3
Student	$\nu = 3$	3	0.56	0.62	35.7	11.5	3.25	2.11
		10	2.81	1.45	138.2	15.9	52.9	1.80
		50	6.12	3.37	198.7	22.43	175.4	6.96
		100	6.88	8.45	237.3	25.34	164.9	13.59
	$\nu = 5$	3	0.27	0.66	24.9	3.50	1.24	2.00
		10	1.16	1.26	96.2	5.63	59.23	1.23
		50	2.80	4.77	147.5	9.61	202.3	8.81
		100	3.17	10.6	187.7	11.4	194.9	16.2
	$\nu = 20$	3	0.27	0.49	19.2	0.70	1.41	1.76
		10	0.54	0.82	70.6	1.6	46.6	0.30
		50	0.93	6.62	107.3	3.37	219.7	11.06
		100	0.69	13.4	139.6	4.23	214.2	19.24

Total Correlation



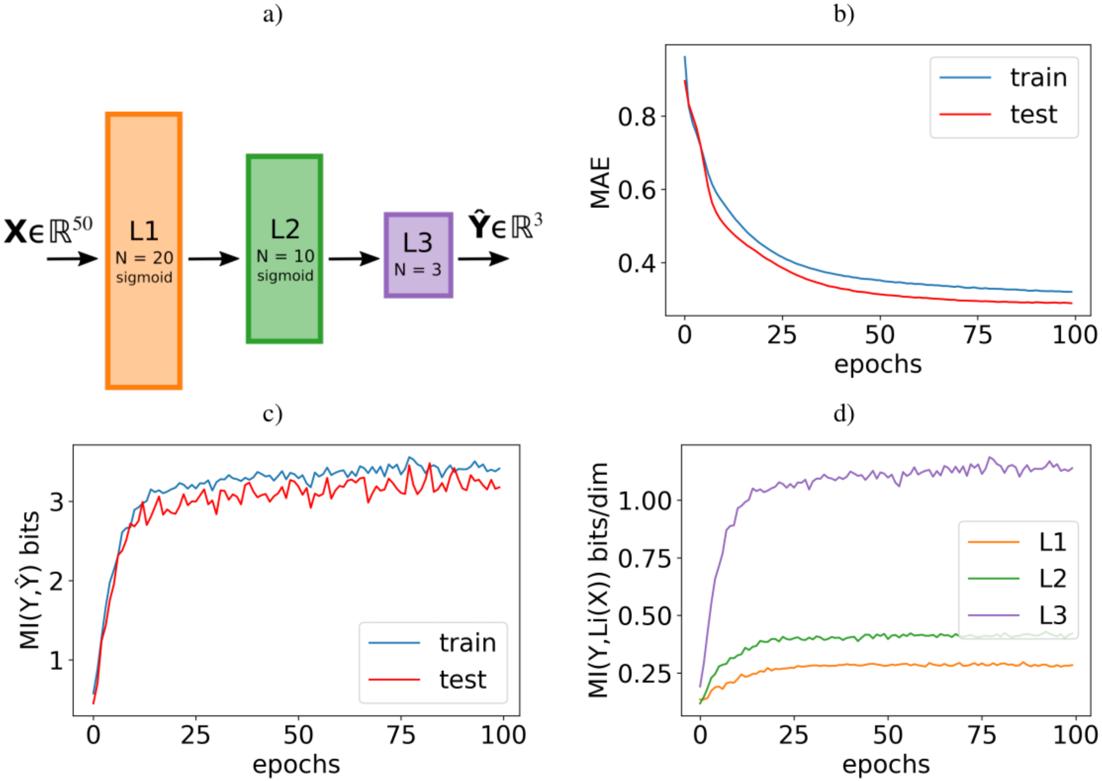
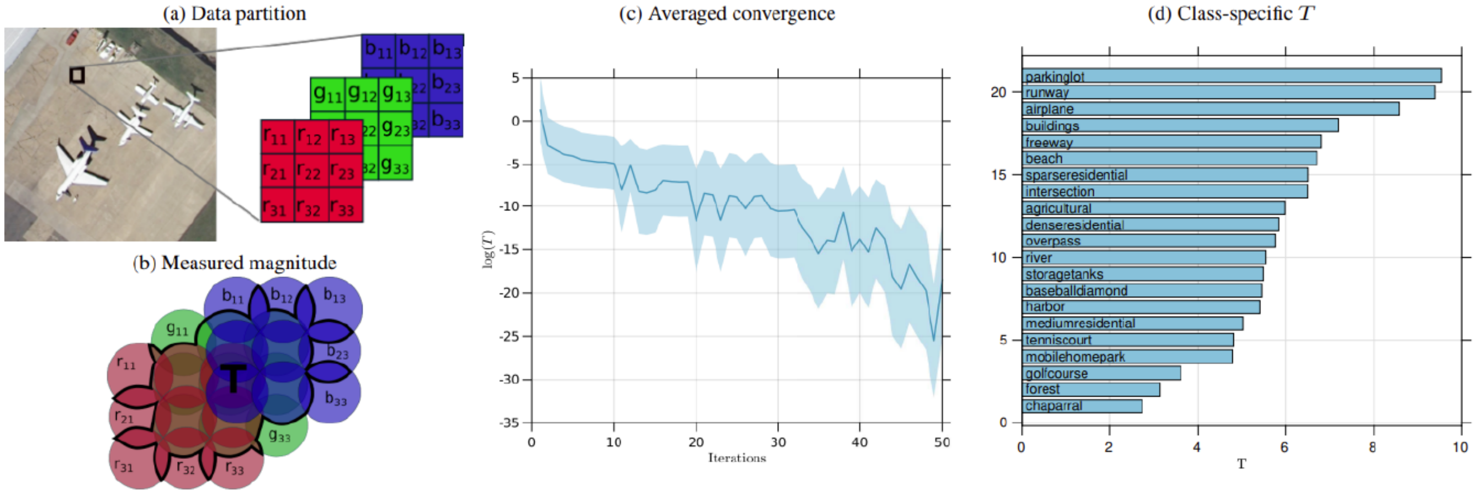
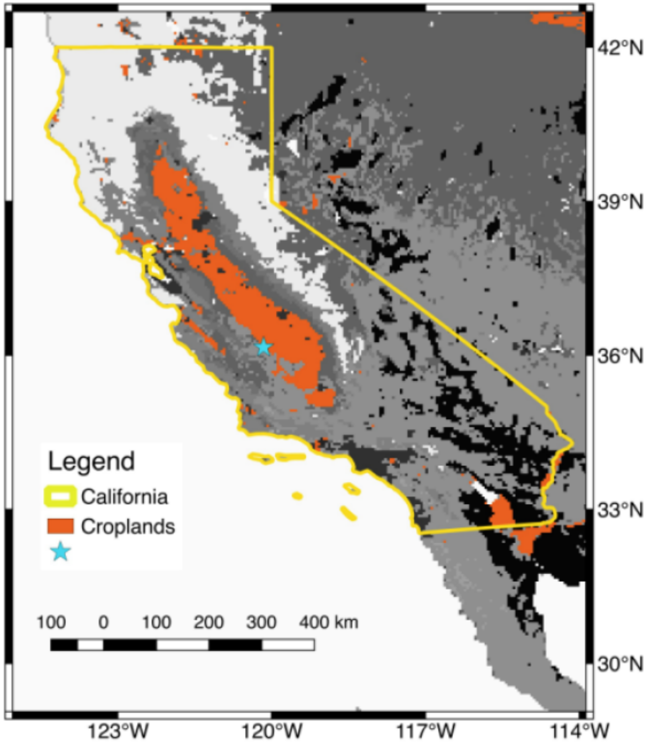


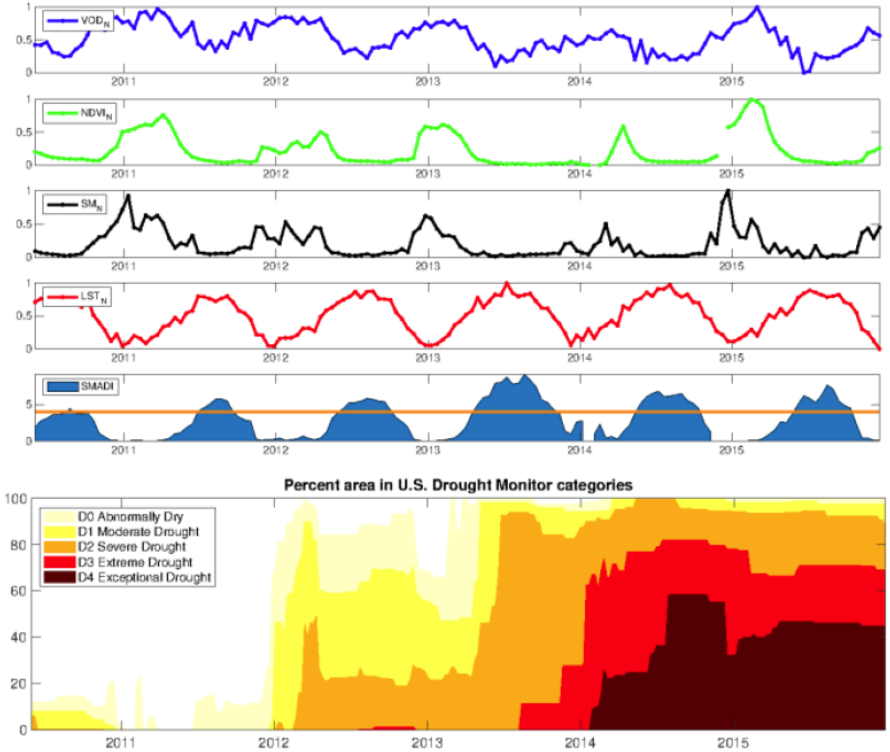
Fig. 8: Learning in artificial neural networks from RBIG estimations of mutual information: evolution of I during the training of an ANN. a) Configuration of the considered neural network. b) Error evolution. c) Evolution of I between the predicted output and the actual data. d) Evolution of I per dimension between the output of each layer and the actual data.

Information in high spatial resolution images

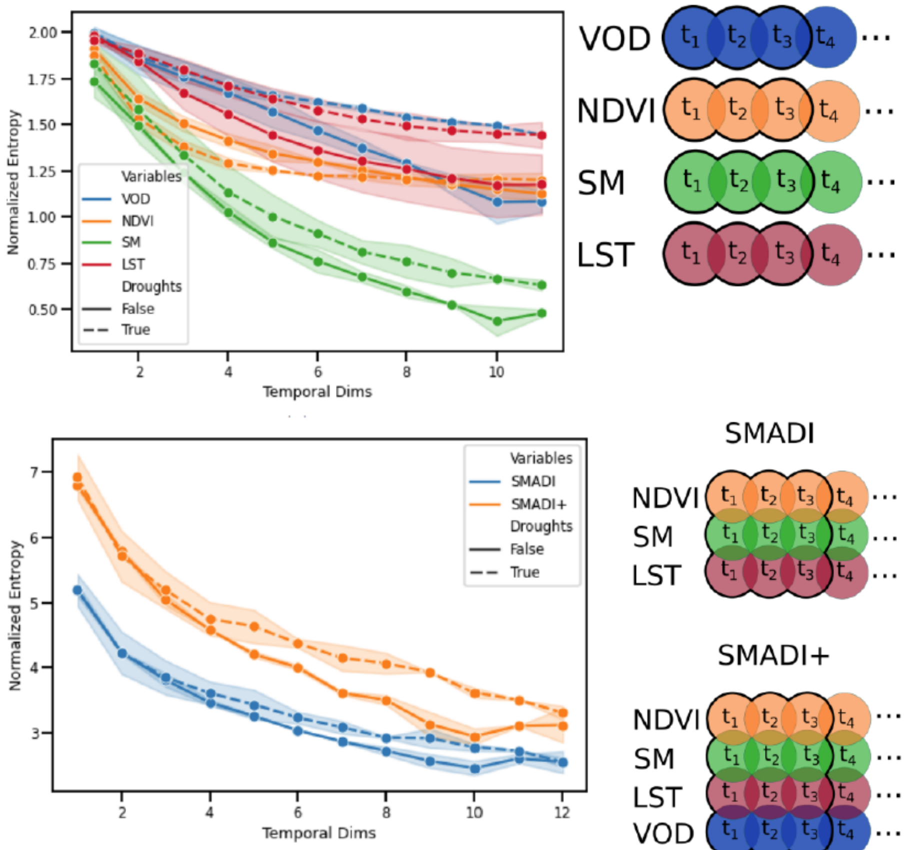


Information in terrestrial biosphere over time

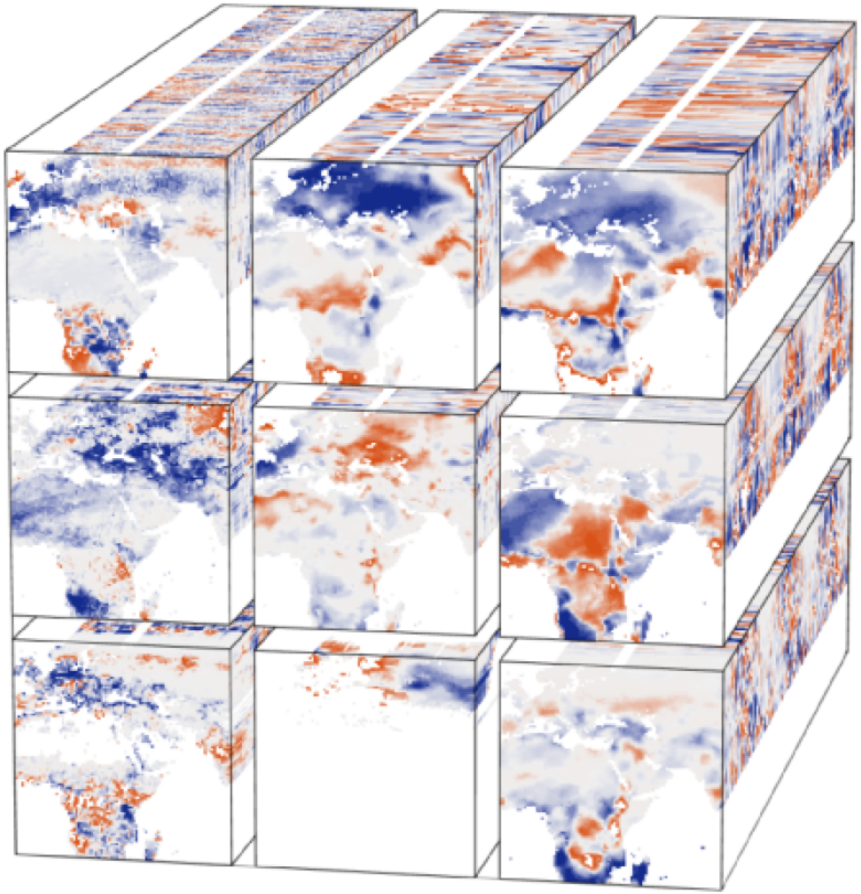




Information in terrestrial biosphere over time



Spatio-temporal information analysis



Spatio-temporal information analysis

Example Variables
Land Surface Temperature
Precipitation
Air Temperature
Soil Moisture
Evaporation
Water Vapour
⋮

Table 1: A few variables that can be found within the data cube (source: esdc.net).

High Spatial Resolution Cube

0.083°, 5''

Low Spatial Resolution Cube

0.25°, 15''

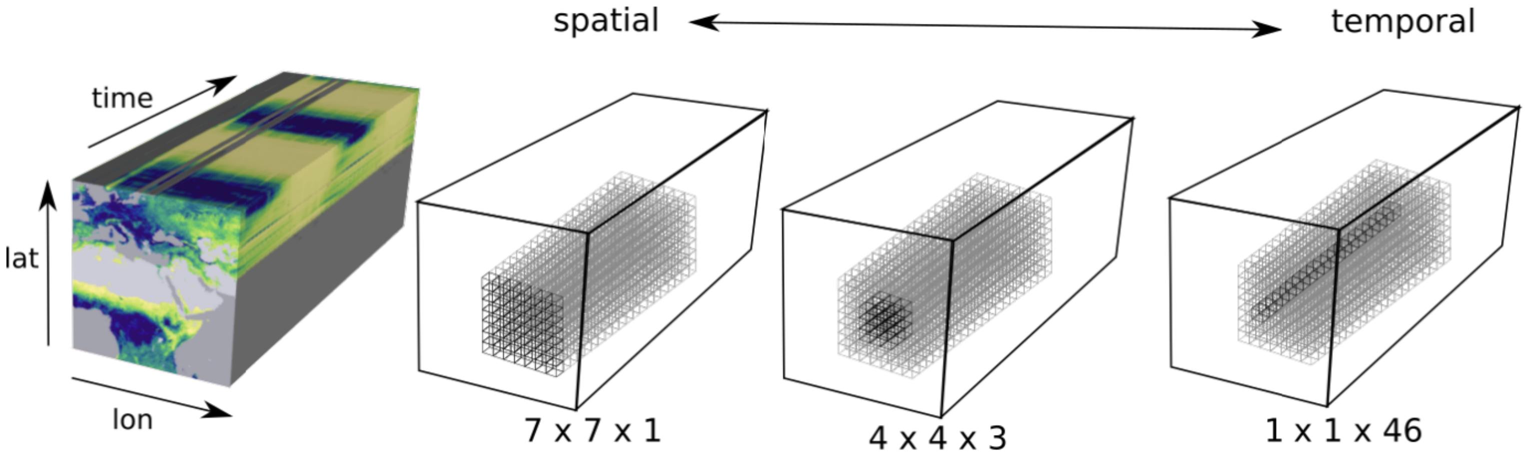
Temporal Resolution

8 days or 46 per year; 2001-2011

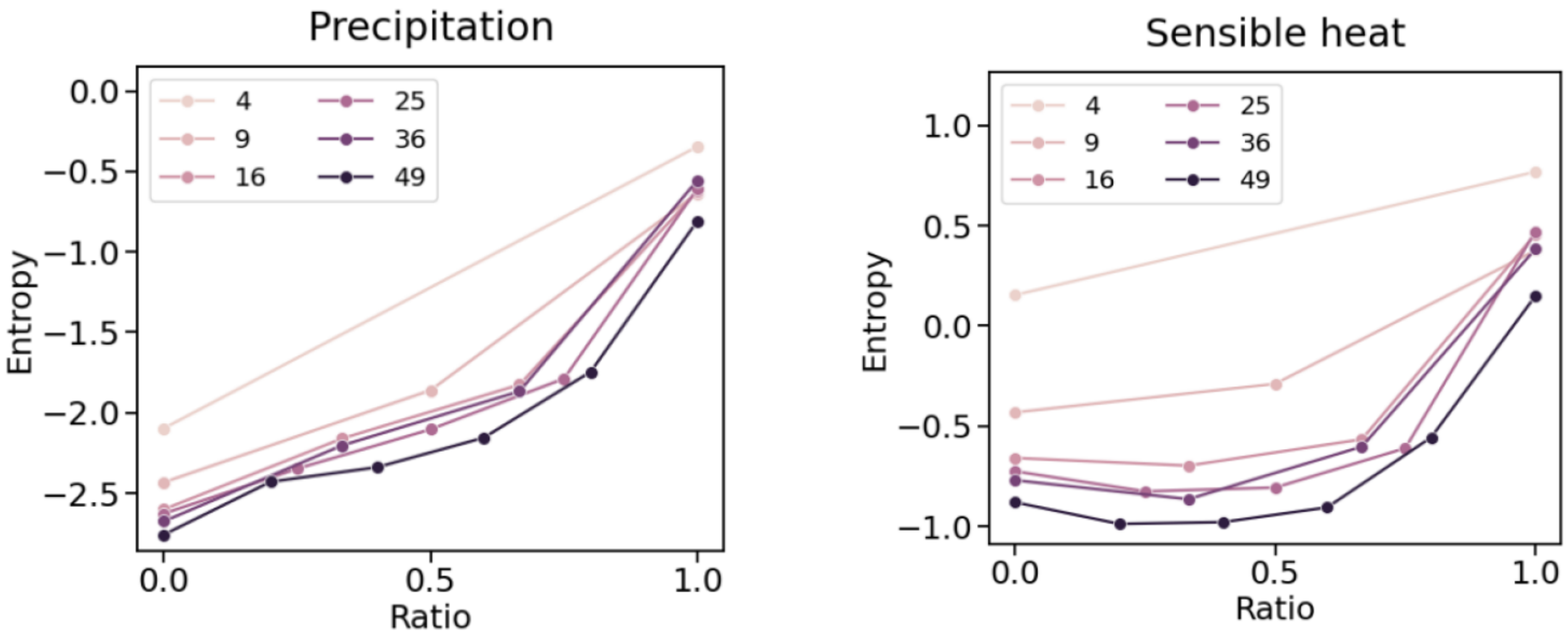
A Lot of Holes

e.g. between 10% to 50% for some datasets

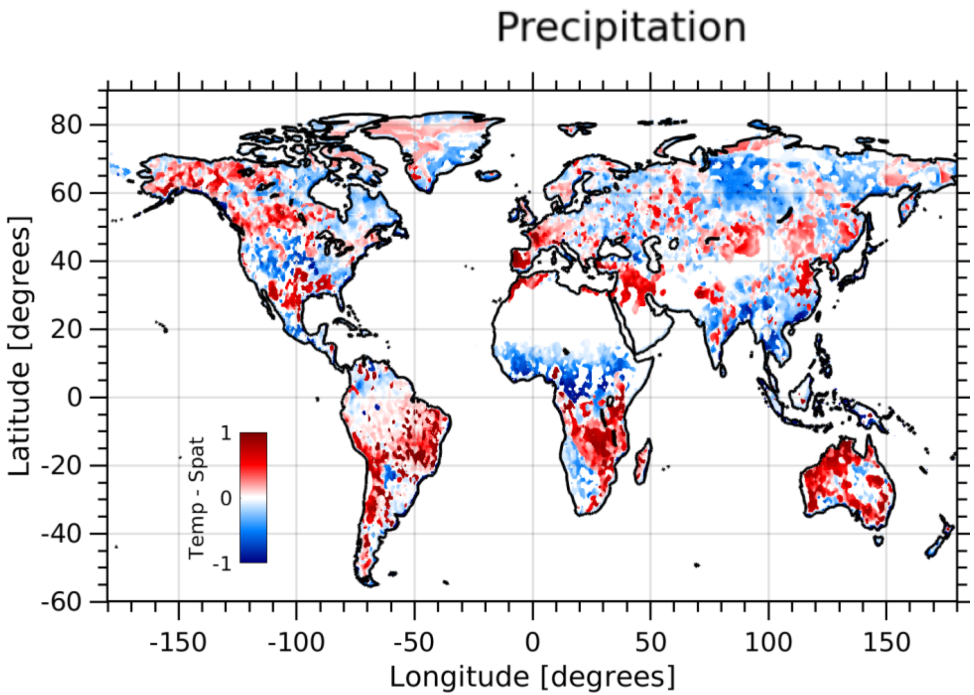
Spatio-temporal information analysis

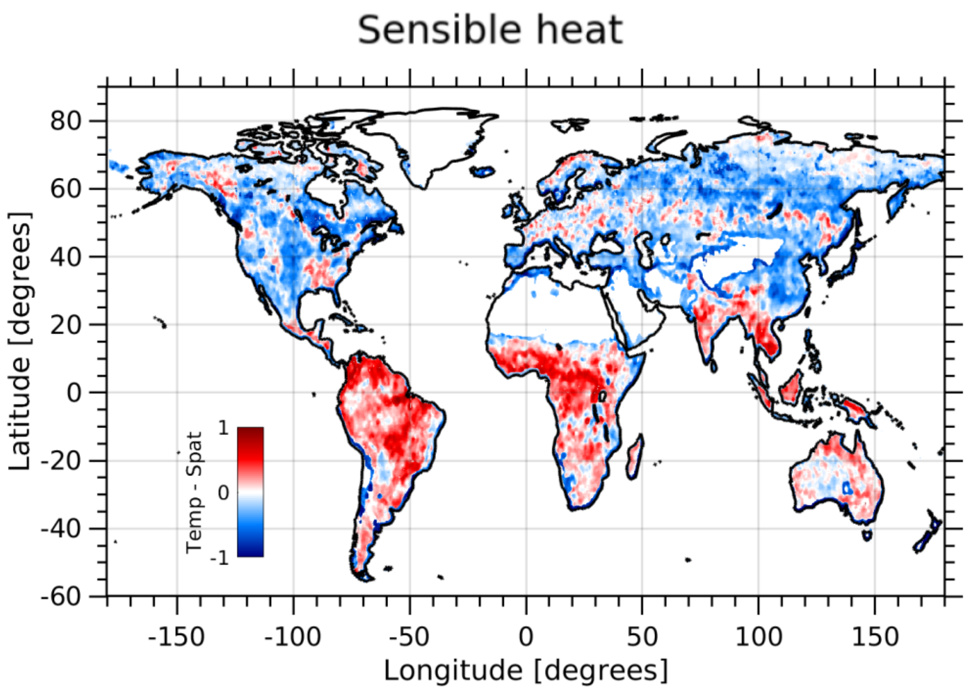


Spatio-temporal information analysis



Spatio-temporal information analysis





Conclusions

Take-home messages

- √ Simple, Fast, Versatile, Hyperparameter free
- √ Info bottleneck with multivariate measures
- √ Many applications possible, use it!
  - https://isp.uv.es/rbig.html
  - https://github.com/IPL-UV/rbig_jax

Future steps

- Train all layers at the same time
- Conditional Independence Test
- Conditional Density Estimation

References

- *"Iterative Gaussianization: From ICA to random rotations"* V. Laparra, G. Camps-Valls, J. Malo, IEEE Transactions on Neural Networks, 22(4):537549, Apr 2011
- *"Gaussianizing the Earth,"* J. Johnson, V. Laparra, M. Piles, and G. Camps-Valls, in IEEE Geoscience & Remote Sensing Magazine, 2020.
- *"Information Theory in Density Destructors,"* Johnson, J.E. Laparra, V. Santos-Rodriguez, R., Camps-Valls, G., Malo, J., International Conference on Machine Learning (ICML), 2019
- *"Information Theory Measures using Gaussianization,"* V. Laparra, E. Jonhson, G. Camps-Valls, R. Santos-Rodrguez, Jess Malo, IEEE Transactions on Information Theory, submitted, 2020