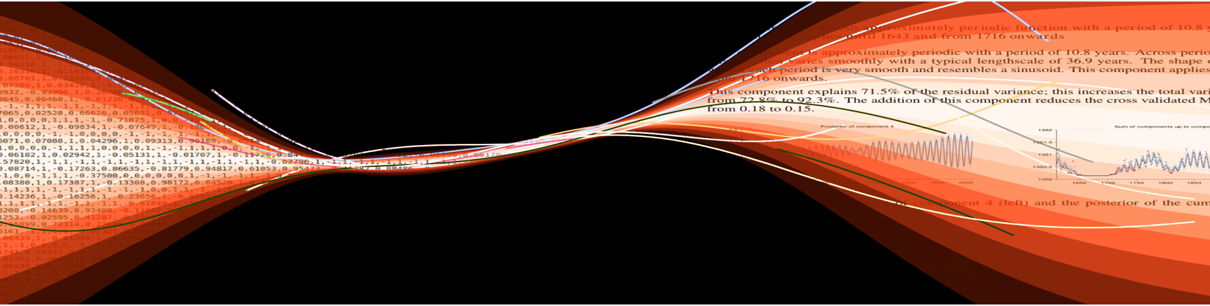
Causal inference in geosciences with multidimensional kernel deviance measures



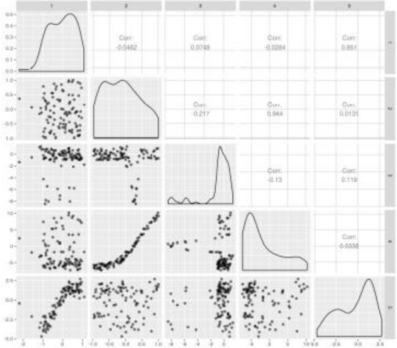
E. Díaz ,D. Sejdinovic, J. Ton
A. Pérez-Suay, V. Laparra and G. Camps-Valls
Image Processing Laboratory (IPL)
Universitat de València, Spain

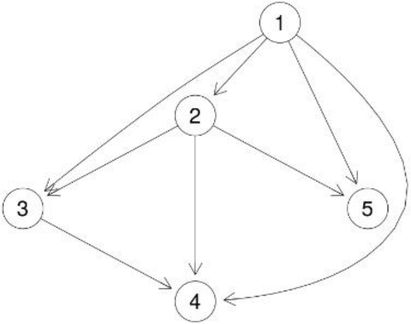
VNIVERSITAT DE VALÈNCIA

erc

Goal of Causal Inference for instantaneous observations

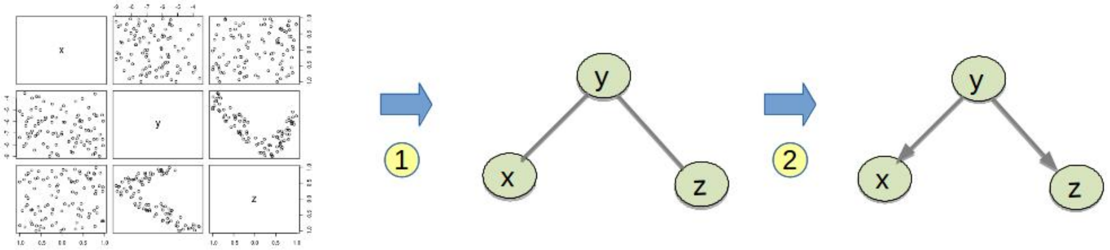
Given a system of p variables, with n observations available for each, learn underlying causal Directed Acyclic Graph (DAG)





Two step learning process

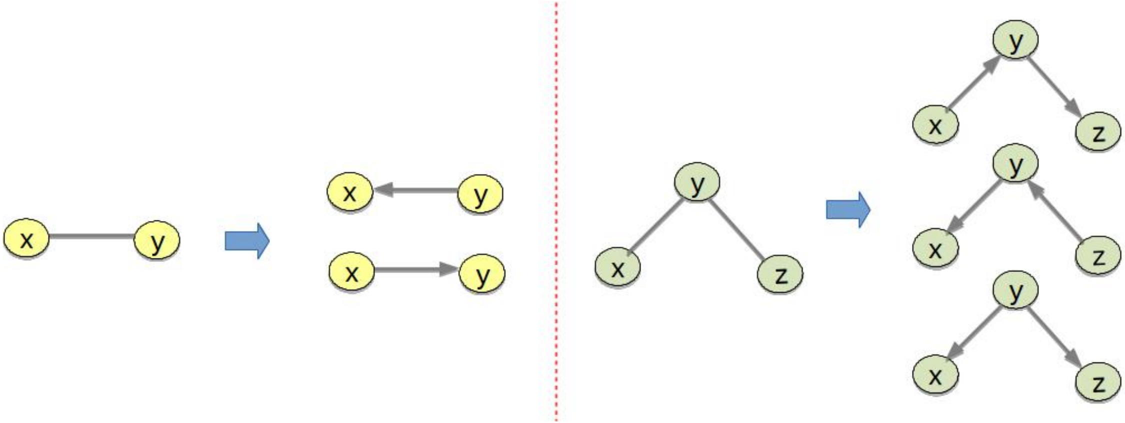
Learning a DAG can be separated into two steps:



1. Learn conditional independencies (learn dag skeleton and colliders)
2. Learn directions (learn undetermined causal relations)

Work presented here focuses on second part of learning process.

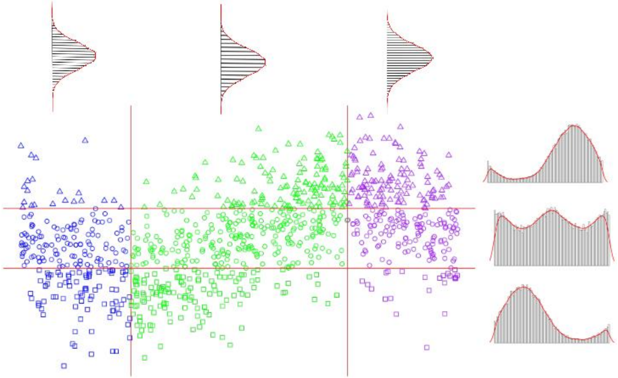
Our task for two and three variable examples



- Given that we know x and y dependent ($x \not\perp\!\!\!\perp y$): choose between $x \rightarrow y$ or $y \rightarrow x$
- Given that we know x and z conditionally independent given y ($x \perp\!\!\!\perp z | y$): choose between $x \rightarrow y \rightarrow z$, $x \leftarrow y \leftarrow z$ or $x \leftarrow y \rightarrow z$.

Idea behind KCDC

Following figure shows observations from model $y = sin(x) + n$ where $n \sim N(0, 1)$

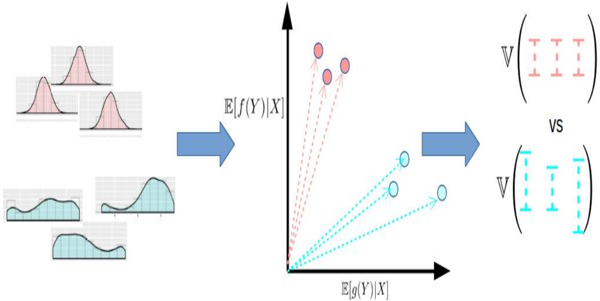


In causal direction $(x \to y)$ complexity of $p(y|x)$ does not depend on x whereas in anticausal direction $(y \to x)$ complexity of $p(x|y)$ varies more.

How do we measure complexity?

Use the *spread* of the norm of vector of expected features as a proxy for the complexity of a set of distributions.

$$\mu = \mathbb{E}\,\phi(Y) = \mathbb{E}\begin{bmatrix} f(Y) \\ Y^2 \\ \sin(Y) \\ \vdots \\ g(Y) \end{bmatrix}$$

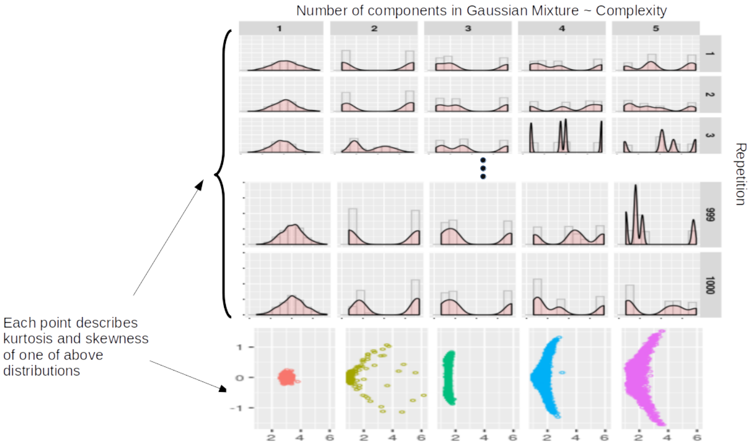


Intuition:

1. Expected feature vector represents distribution if adequate features chosen.
2. Expected feature vector of similar distributions constrained to subspace of feature space and so have similar norms.

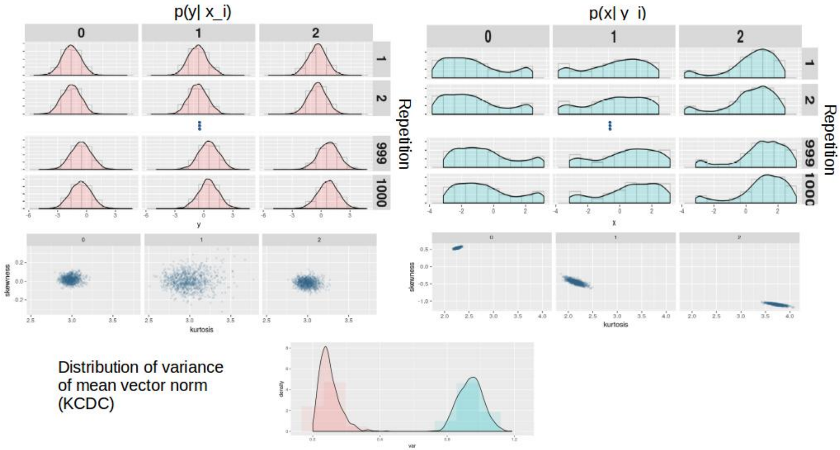
Illustration: gaussian mixtures

The higher the number of components in a gaussian mixture the more complex it is.



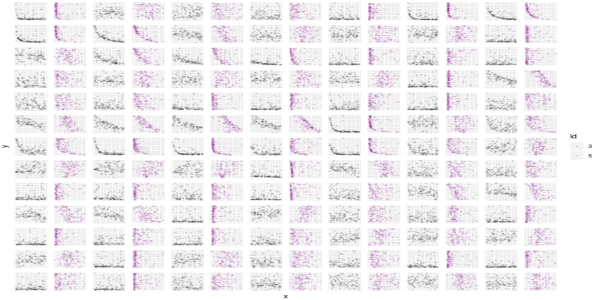
Norm of mean vector can help us distinguish between distributions.

Back to $sin(x) + n$ example...



KCDC distinguishes causal direction for all 1000 repetitions.

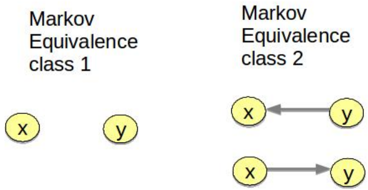
Experiment 3: RTM Prosail Simulated Pairs



- 182 data sets with 1000 pairs of points each
- max 100 points used
- causes consist of **7** biological parameters
- effects consist of reflectances for **13** different bands

measure	ccr	auc
ANM	62.6 %	60.2 %
KCDC	97.8 %	99.3 %
SHSIC	-	65.0 %

Extending KCDC to systems with more than two variables



To extend KCDC to DAGs with more than two nodes (higher dimensional systems) we note that:

- KCDC only serves to distinguish between DAGs in the same Markov Equivalence class (those graphs with same set of conditional independencies).
- The distribution of nodes with no parents is not taken into account since the causal mechanism is encoded in the conditional distributions of nodes with parents.

Extending KCDC to systems with more than two variables

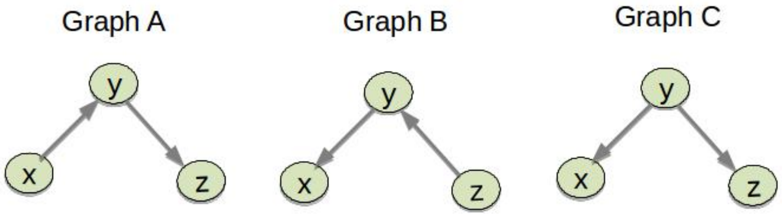
Taking this into account we write the KCDC of a general p-node DAG as:

$$KCDC(\mathcal{G}) = \sum_{i \in \mathcal{A}} KCDC\Big(p(x_i|pa(x_i))\Big) \tag{1}$$

where

- A is the set of nodes in the dag $\mathcal{G}$ that have at least one parent, and
- $pa(x_i)$ is the set of parents of node x_i.

An example

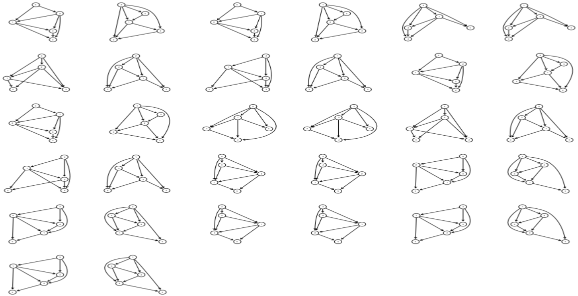


With previous definition:

- $KCDC(\mathcal{G}_A) = KCDC(p(y|x)) + KCDC(p(z|y))$
- $KCDC(\mathcal{G}_B) = KCDC(p(x|y)) + KCDC(p(y|z))$
- $KCDC(\mathcal{G}_C) = KCDC(p(x|y)) + KCDC(p(z|y))$

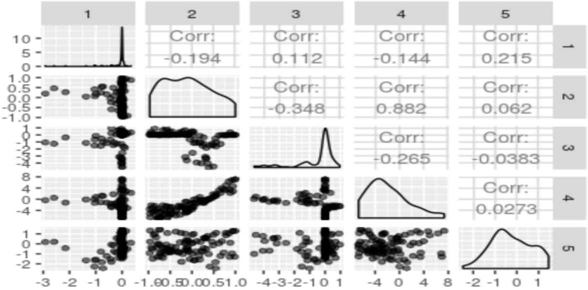
Lets see some experimental results for multi-variate KCDC.

Experiment 5: Artificial Cause-Effect 5-tuples



- 100 datasets with 100 5-tuples each
- Non-additive noise models $e = f(a, b, c, n,)$ with:
    - non-linear random function f
    - $a, b, c, d, n \sim U(-1, 1)$
- true causal structure one of 32 dags on the left.

Experiment 5: Artificial Cause-Effect 5-tuples



- Data for 1 of 100 datasets plotted on left.

measure	ccr	edgeCCR
ANM	8.0 %	67.3 %
KCDC	23.0 %	80 %
Rnd	6.0 %	63.3 %

References

- Mitrovic et al., ‘Causal Inference via Kernel Deviance Measures,’ NIPS 2018.
- Ton et al., ‘Noise contrastative estimation for meta learning using kernel mean embeddings’ AISTATS 2019 submitted.
- Pérez-Suay and Camps-Valls, ‘Causal Inference in Geoscience and Remote Sensing From Observationl Data,’ IIIE-GRSL, 2019.
- Pérez-Suay and Camps-Valls, ‘Sensitivity Maps of the Hilbert-Schmidt Independence Criterion,’ Applied Soft Computing, 2017.
- Mooij et al., ‘Distinguishing cause from effect using observational data,’ JMLR, 17(1), 2016.
- Gretton et al., ‘Measuring statistical dependence with Hilbert-Schmidt norms,’ ALT, 2005
- Hoyer et al., ‘Nonlinear causal discovery with additive noise models,’ NIPS 2008.
- Camps-Valls, Mooij and Schölkopf, ‘Remote sensing feature selection by kernel dependence measures,’ IEEE-GRSL 2010