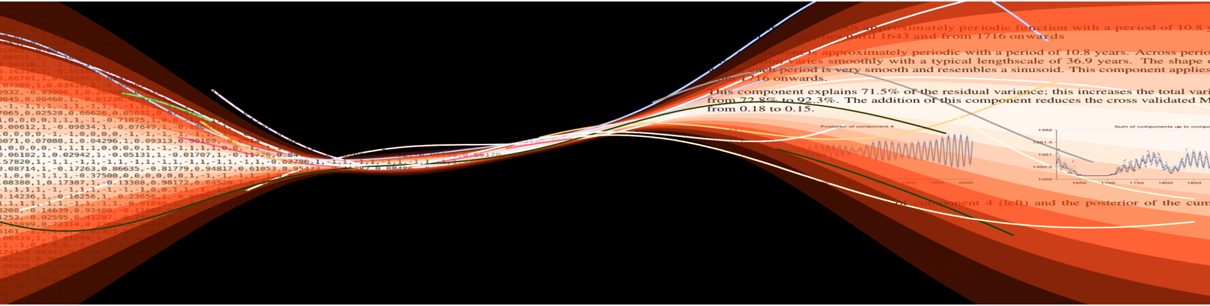
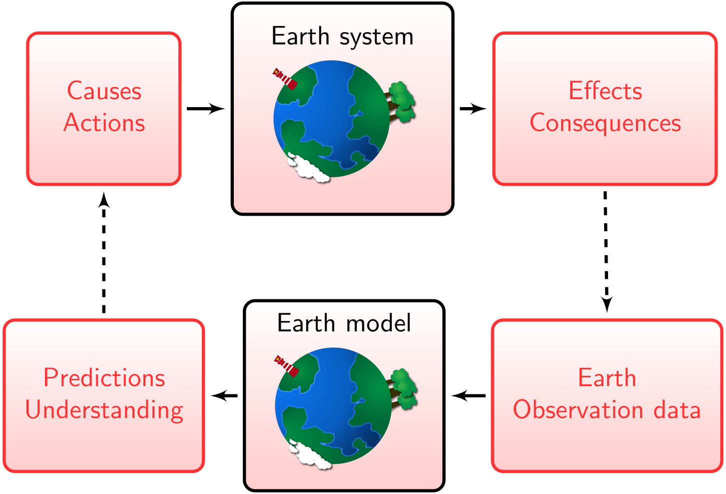
Causal inference in geosciences with multidimensional kernel deviance measures



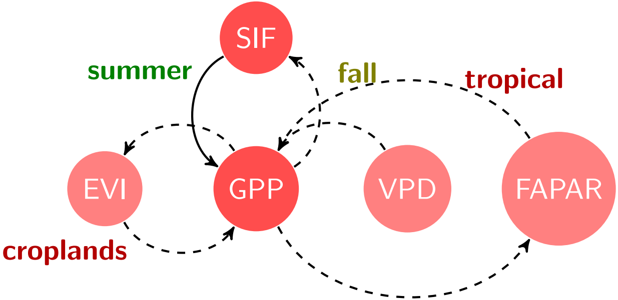
E. Díaz ,A. Pérez-Suay,
V. Laparra and G. Camps-Valls
Image Processing Laboratory (IPL)
Universitat de València, Spain
✉ emiliano.diaz@uv.es | http://isp.uv.es

Introduction



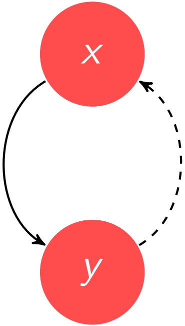
Outlook

- Infer causal relations between r.v. is challenging
- Even more from pure observational data: no models involved, no ground truth!
- In GRS, causality is key to understand the Earth's system



Outlook

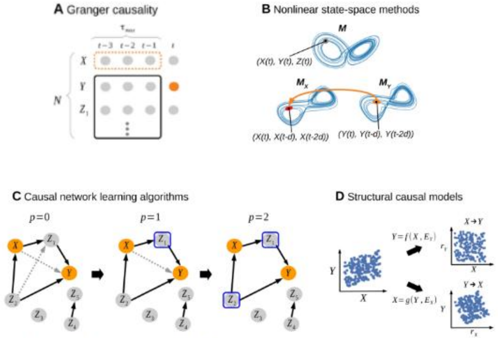
- Infer causal relations between r.v. is challenging
- Even more from pure observational data: no models involved, no ground truth!
- In GRS, causality is key to understand the Earth's system



M • MENÚ DEL DÍA

1. Causal inference for instantaneous observations (CIIO)
2. Kernel Conditional Deviance for Causal Inference (KCDC)
3. Results for:
   - Simulated data: bi-variate and multivariate
   - Data from RTM model PROSAIL
   - Data from RTM emulator
   - 30 GRS causal inference problems

Overview of causal inference methods



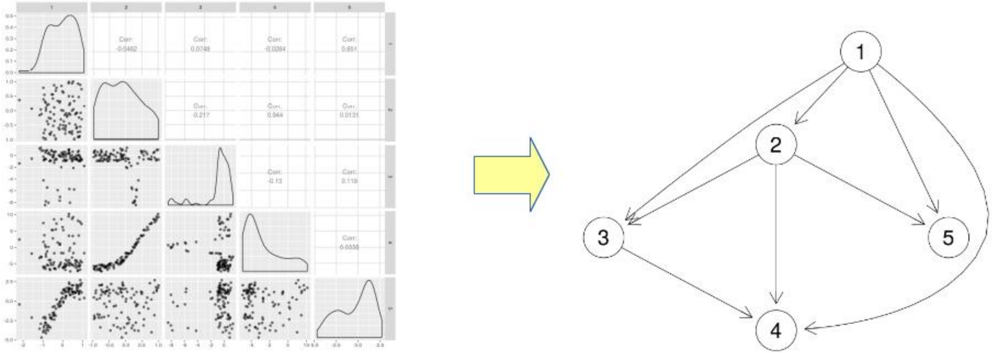
Acks: Runge et al 2019

A. Multivariate granger causality tests
B. Nonlinear state-space method CCM
C. Causal network learning algorithms (conditional independendence testing)
D. Structural Causal models

KCDC, the method studied here and ANMs, the method to which we will compare its performance belong to group D.

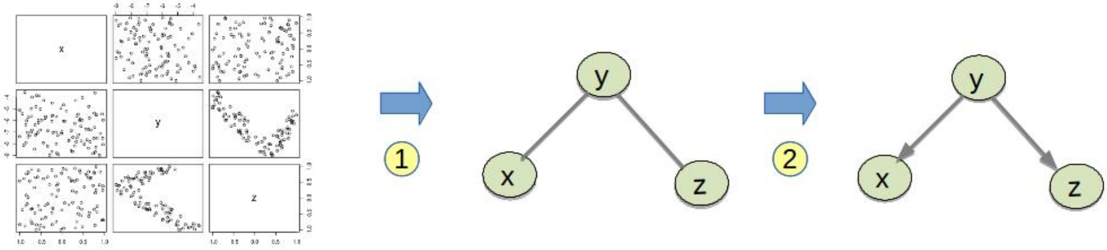
Goal of Causal Inference for instantaneous observations

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



Two step learning process

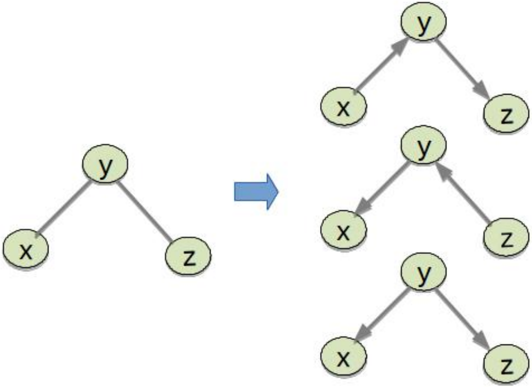
Learning a DAG can be separated into two steps:



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.

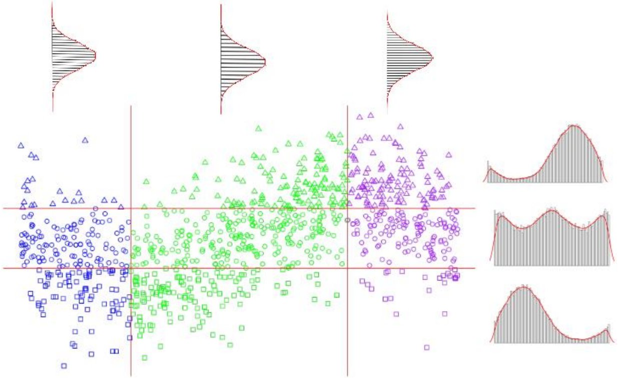
Our task for two and three variable examples



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

Idea behind KCDC

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

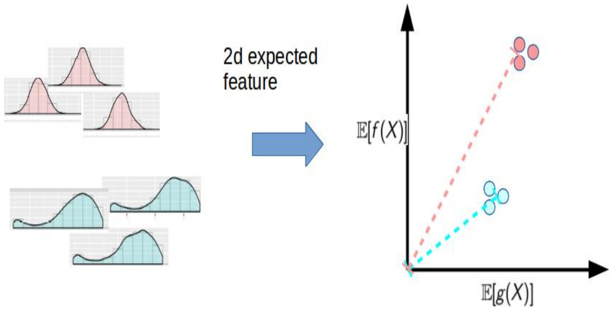


In causal direction $(x \rightarrow y)$ complexity of $p(y|x)$ does not depend on x whereas in anticausal direction $(y \rightarrow x)$ complexity of $p(x|y)$ varies more.

How do we measure complexity?

Use the norm of vector of expected features as a proxy for complexity.

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

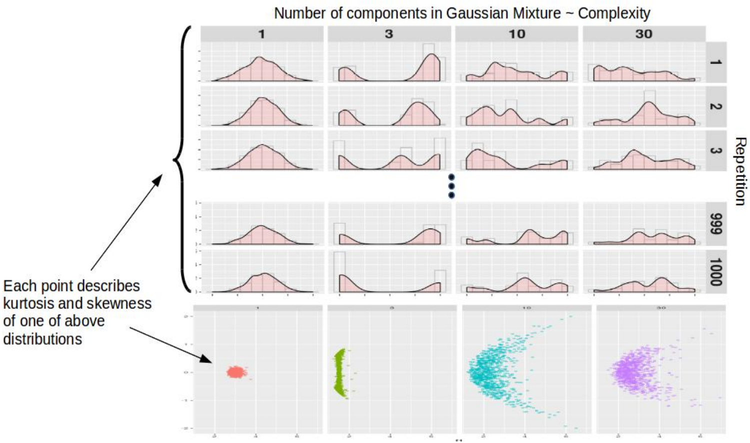


Intuition:

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.

Kernel Conditional Deviance for Causal Inference (KCDC)

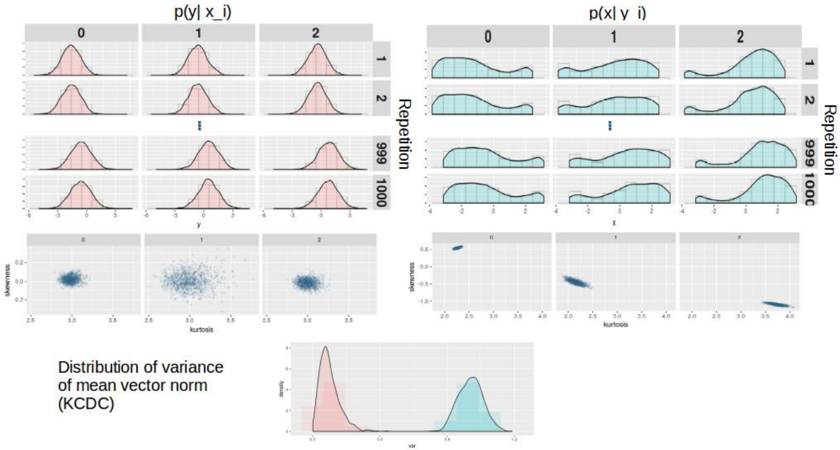
Based on this idea [Mitrovic et al, 2018] introduced KCDC to infer direction of causality for pairs of variables.

KCDC

$$S_{x \to y} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(||\mu_{y|x \in b_i}||_2 - \frac{1}{|B|} \sum_{j=1}^{|B|} ||\mu_{y|x \in b_j}||_2 \right)^2$$

- $B = b_1, ..., b_m$ are the bins that x is split into,
- KCDC is the variance of mean feature norms, corresponding to different bins
- Measure in both directions, direction of minimum variance is causal direction.

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



KCDC distinguishes causal direction for all 1000 repetitions.

How do we choose bins?

$$S_{x \to y} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(||\mu_{y|x \in b_i}||_2 - \frac{1}{|B|} \sum_{j=1}^{|B|} ||\mu_{y|x \in b_j}||_2 \right)^2$$

$$\mu_{y|x \in b_j} = \frac{1}{|b_j|} \sum_{i=1}^{|b_j|} \phi(y_i)$$

Instead of using bins, computing weighted feature norms allows us to calculate a mean feature norm for each data point and spares us choosing bins (important for extending to multivariate case).

Use weights instead

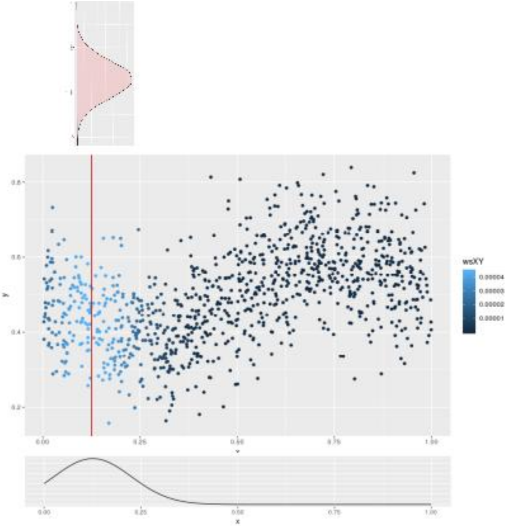
$$S_{x \to y} = \frac{1}{n} \sum_{i=1}^{n} \left(||\mu_{y|x_i}||_2 - \frac{1}{n} \sum_{j=1}^{n} ||\mu_{y|x_j}||_2 \right)^2$$

$$\mu_{y|x_i} = \sum_{i=1}^{n} w_i \phi(y_i)$$

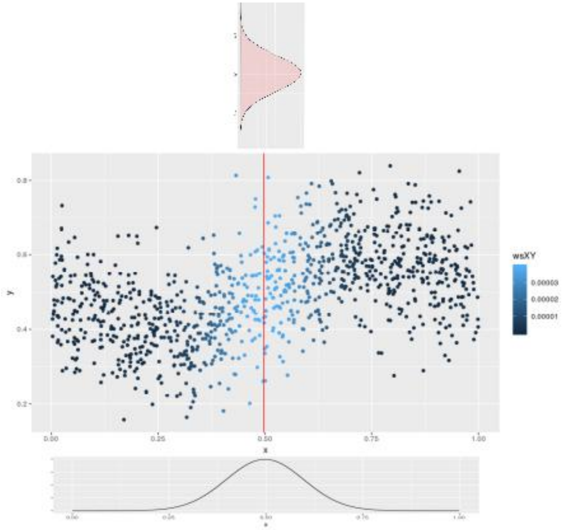
$$w_i = f(||x_i - x_j||_2)$$

Instead of using bins, computing weighted feature norms allows us to calculate a mean feature norm for each data point and spares us choosing bins (important for extending to multivariate case).

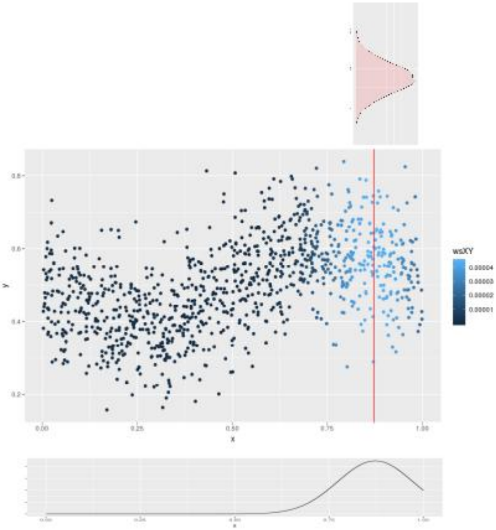
Back to $sin(x) + n$ example...



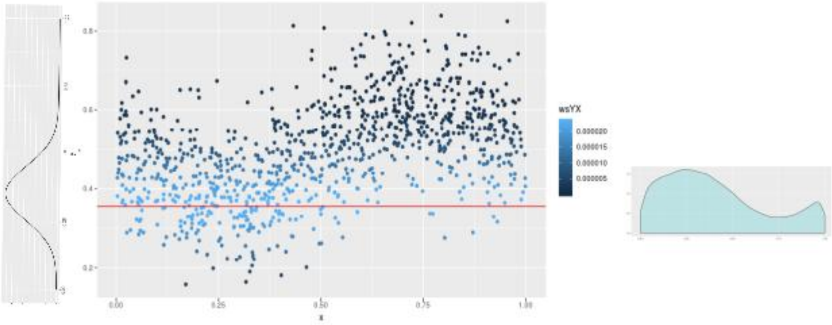
Back to $sin(x) + n$ example...



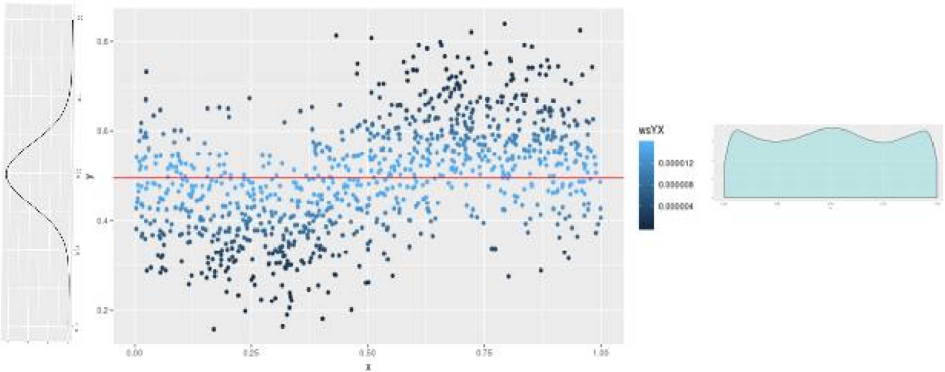
Back to $sin(x) + n$ example...



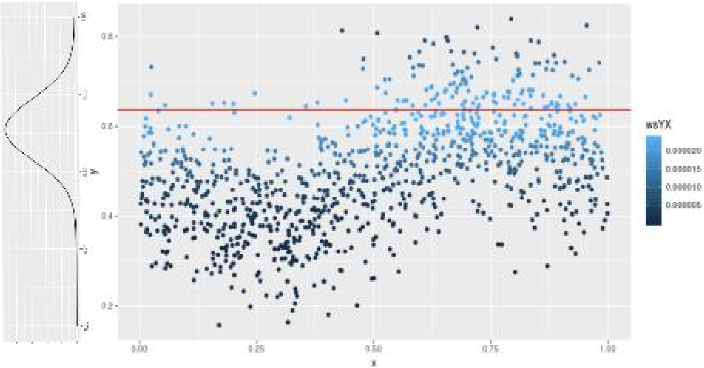
Back to $sin(x) + n$ example...

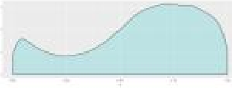


Back to $sin(x) + n$ example...



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





Kernel trick

$$S_{x\to y} = \frac{1}{n}\sum_{i=1}^{n}\left(||\mu_{y|x_i}||_2 - \frac{1}{n}\sum_{j=1}^{n}||\mu_{y|x_j}||_2\right)^2$$

$$\mu_{y|x_i} = \sum_{i=1}^{n} w_i\phi(y_i) \in \mathbb{R}^n$$

$$w_i = f(||x_i - x_j||_2)$$

- Kernel trick replaces explicit mean feature vector with implicit calculation of longer (possibly infinite) mean feature vector.
- This allows a more detailed description of $p(x|y_i)$ and $p(y|x_i)$, (sufficient and adequate number of features to properly represent distribution)

Kernel trick

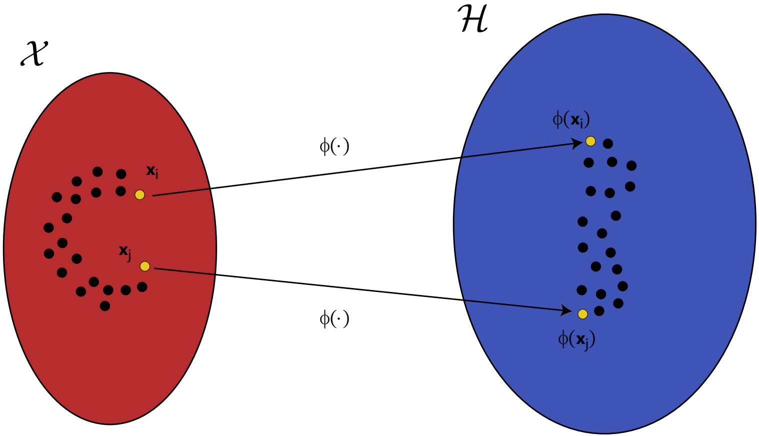
$$S_{x\to y} = \frac{1}{n}\sum_{i=1}^{n}\left(||\mu_{y|x_i}||_{\mathcal{H}_y} - \frac{1}{n}\sum_{j=1}^{n}||\mu_{y|x_j}||_{\mathcal{H}_y}\right)^2$$

$$\mu_{y|x_i}(y) = \sum_{i=1}^{n} w_i k(y_i, y) \in \mathcal{H}_y$$

$$w_i = g(l(x_i, x_j))$$

- $k(y, y')$ kernel for output variable y and $l(y, y')$ kernel for output variable.
- Kernel trick replaces explicit mean feature vector with implicit calculation of longer (possibly infinite) mean feature vector.
- This allows a more detailed description of $p(x|y_i)$ and $p(y|x_i)$, (sufficient and adequate number of features to properly represent distribution)

Kernel trick [Schölkopf, 1998]



Similarity measure $\mathbf{x}_i^\top \mathbf{x}_j$

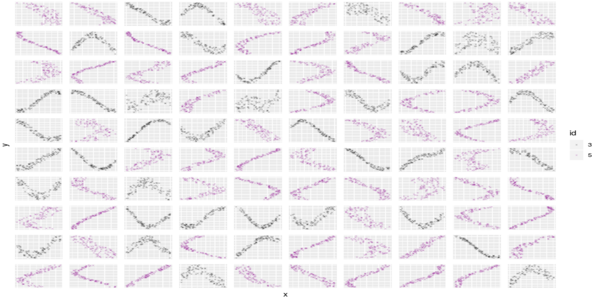
$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Our contribution

Up until now we have explored KCDC proposed by [Mitrovic et al, 2018] to infer direction of causality for pairs of variables. Our contribution consists of:

1. Test KCDC on GRS pairs to validate its effectiveness in geosciences,
2. Extend KCDC to multivariate systems of variables, and
3. Test multivariate KCDC on multivariate simulated datasets.

Experiment 1: Artificial Cause-Effect Pairs



- 100 data sets with 100 pairs of points each
- Additive noise models $y = f(x) + n$ with:
  - non-linear random function f
  - $x, n \sim U(-3, 3)$

measure	ccr	auc
ANM	60.0 %	57.7 %
KCDC	91.0 %	95.7 %

Cause-Effect Pairs database

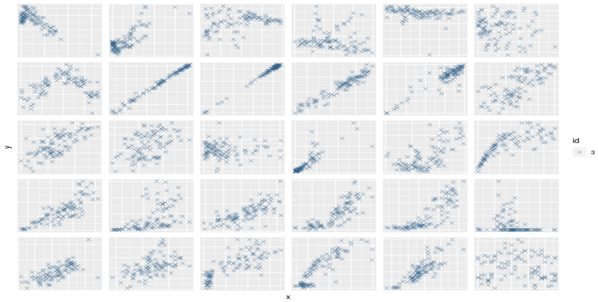
- Cause Effect Pairs (CEP) contains annotated 102 pairs[1]
- Unidimensional and GRS variables only (30 out of 100)

id	*x*	*y*	Cause
pair0001	Altitude	Temperature	→
pair0002	Altitude	Precipitation	→
pair0003	Longitude	Temperature	→
pair0004	Altitude	Sunshine hours	→
pair0020	Latitude	Temperature	→
pair0021	Longitude	Precipitation	→
pair0042	Day of the year	Temperature	→
pair0043	Temperature at t	Temperature at t+1	→
pair0044	Pressure at t	Pressure at t+1	→
pair0045	Sea level pressure at t	Sea level pressure at t+1	→
pair0046	Relative humidity at t	Relative humidity at t+1	→
pair0049	Ozone concentration	Temperature	←
pair0050	Ozone concentration	Temperature	←
pair0051	Ozone concentration	Temperature	←
pair0072	Sunspots	Global mean temperature	→

id	*x*	*y*	Cause
pair0073	CO2 emissions	Energy use	←
pair0077	Temperature	Solar radiation	←
pair0078	PPFD	Net Ecosystem Productivity	→
pair0079	Net Ecosystem Productivity	Diffuse PPFDdif	←
pair0080	Net Ecosystem Productivity	Diffuse PPFDdif	←
pair0081	Temperature	Local CO2 flux, BE-Bra	→
pair0082	Temperature	Local CO2 flux, DE-Har	→
pair0083	Temperature	Local CO2 flux, US-PFa	→
pair0087	Temperature	Total snow	→
pair0089	root decomposition Oct (grassl)	root decomposition Oct (grassl)	←
pair0090	root decomposition Oct (forest)	root decomposition Oct (forest)	←
pair0091	clay cont. in soil (forest)	soil moisture	→
pair0092	organic carbon in soil (forest)	clay cont. in soil (forest)	←
pair0093	precipitation	runoff	→
pair0094	hour of day	temperature	→

[1] `https://webdav.tuebingen.mpg.de/cause-effect/`

Experiment 2: Cause-Effect Pairs database

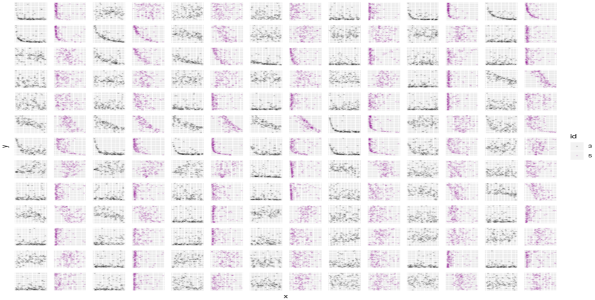


- 30 data sets with 126-10369 pairs of points each
- max 100 points used
- Non-linear, non-additive examples included

measure	ccr	auc
ANM	60.0 %	55.7 %
KCDC	66.7 %	70.2 %
SHSIC	-	70.0 %

- SHSIC result from [Pérez-Suay et al, 2019]

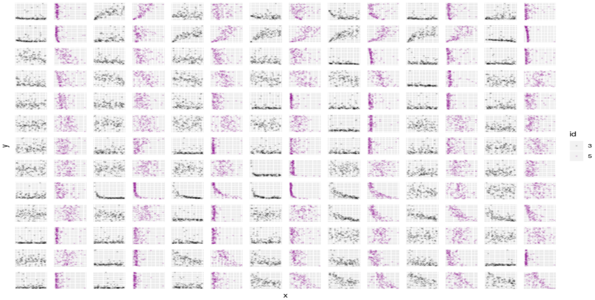
Experiment 3: RTM Prosail Simulated Pairs



- 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 %

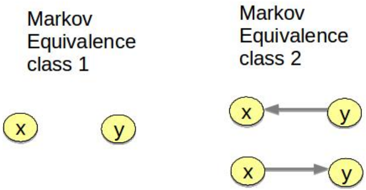
Experiment 4: RTM Prosail Emulator Pairs



- 182 data sets with 500,000 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	58.8 %	60.4 %
KCDC	97.3 %	99.4 %
SHSIC	-	80.0 %

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

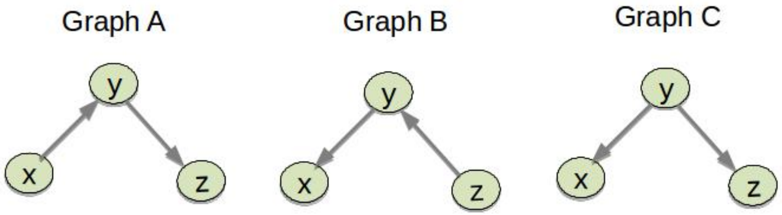
Taking this into account we write the KCDC of a general p-node DAG as:

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

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

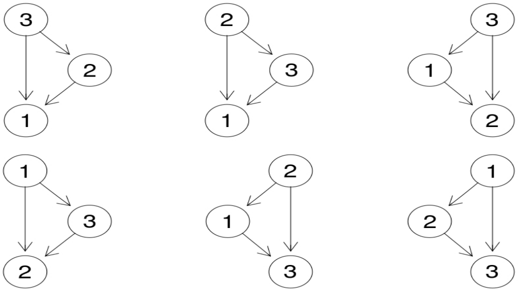


With previous definition:

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

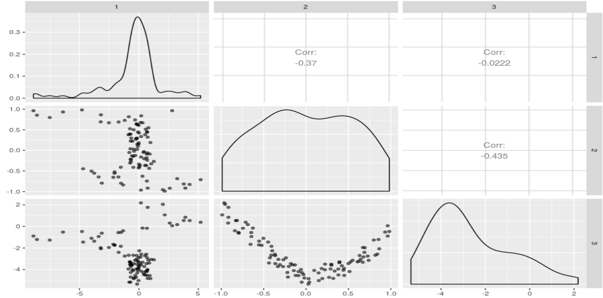
Lets see some experimental results for multi-variate KCDC.

Experiment 5: Artificial Cause-Effect 3-tuples



- 100 datasets with 100 3-tuples each
- Additive noise models $z = f(x, y) + n$ with:
  - non-linear random function f
  - $x, y, n \sim U(-1, 1)$
- true causal structure one of 6 dags on the left.

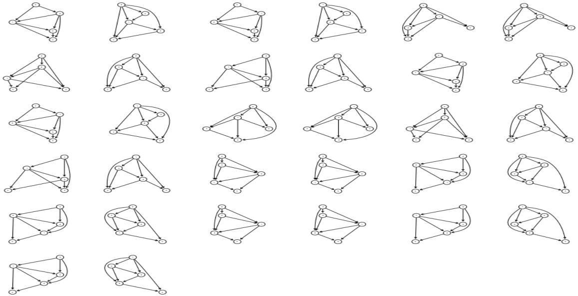
Experiment 5: Artificial Cause-Effect 3-tuples



- Data for 1 of 100 datasets plotted on left.

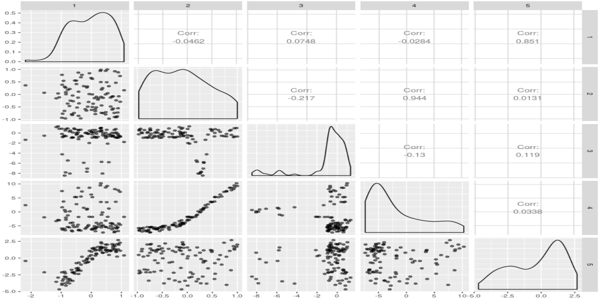
measure	ccr	edgeCCR
ANM	30.0 %	59.0 %
KCDC	73.0 %	88.3 %
Rnd	23.0 %	55.0 %

Experiment 5: Artificial Cause-Effect 5-tuples



- 100 datasets with 100 5-tuples each
- Additive noise models $e = f(a, b, c, d) + n$ 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

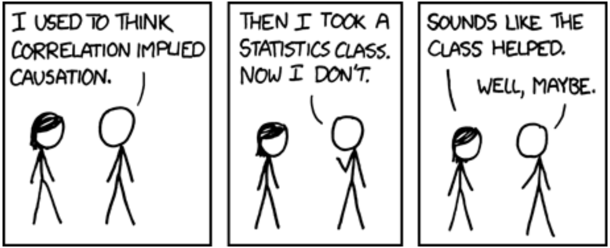


- Data for 1 of 100 datasets plotted on left.

measure	ccr	edgeCCR
ANM	8.0 %	67.3 %
KCDC	43.0 %	88.0 %
Rnd	6.0 %	63.3 %

Take-home messages

- State-of-the art method for observational causal inference
- Physical models assessment
- Many potential GRS apps to explore
- Multivariate problems and cond. indep.



References

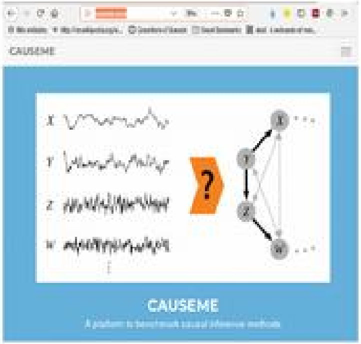
- Mitrovic et al., ‘Causal Inference via Kernel Deviance Measures,’ NIPS 2018.
- Pérez-Suay and Camps-Valls, ‘Causal Inference in Geoscience and Remote Sensing From Observationl Data,’ IIIE-GRSL, 2019.
- Pérez-Suay and Camps-Valls, ‘Sensitivity Maps of the Hilbert-Schmidt Independence Criterion,’ Applied 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
- Camps-Valls and Bruzzone, *Kernel methods for Remote Sensing Data Analysis*, Wiley and Sons, 2009.

Cause me: test your causal discovery algorithm online.

Extension to causal discovery

- **CauseMe: http://causeme.uv.es**
  - Download time series with ground truth
  - Run your causal discovery algorithm offline
  - Upload your causal graph
  - Get your results!

CauseMe: http://causeme.uv.es



"Inferring causation from time series with perspectives in Earth system sciences"
Runge, Bathiany, Bollt, Camps-Valls, et al. Nat Comm (submitted), 2018.
"Causal Inference in Geoscience and Remote Sensing from Observational Data,"
Pérez-Suay and Camps-Valls, IEEE Trans. Geosc. Rem. Sens, 2018