Antal Jakovác

Causal methods in high-dimensional data analysis

AIME 2026
Budapest, Hungary, November 12-13, 2025

The setting

Many signals, one target

- collect facts:
  - time series: repeatedly measured quantities
  - events: occurrences recorded at specific times
- **goal**: predict a target Y from measured facts
  - simultaneous data: measured at the same time
  - historic data: near-past embeddings; distant history may require attention

Difficulty is not collecting variables: it is selecting the useful ones!

High dimensionality is everywhere

High dimensional data occur frequently

- finance and economics:
  - stock prices
  - quarterly reports, decisions
  - news, sentiment and political statements: irregular events
- medicine: multivariate physiological recordings
- physics and other complex systems
- etc.

Modelling

The conventional modelling pipeline

- gather potentially explanatory variables
- fit a predictive model: autoregression, trees, forests, SVMs, neural networks
- Rank relevance: feature importance, permutation tests, masking
- Reduce the feature set and retrain

But the first model may already be too difficult to estimate!

Feature selection

Why feature selection must come first

- Too many candidate inputs relative to observations
- Class imbalance and limited samples
- Even a simple model may be poorly constrained
- Goal: reveal dependence before committing to a predictive model

Pre-modelling relevance analysis is needed

Causal methods

In general: reveal causal relationship of time series

- Can be used to explore explanatory connection between variables
- Different methods exploit different signatures of causality:
  - Predictive improvement → Granger causality
  - Conditional independence → PC, FCI
  - Functional asymmetry → SCM, LiNGAM
  - Information flow → Transfer Entropy
  - Dynamical reconstruction → CCM
  - Response to manipulation → Interventions

Causal methods

Here we consider three simple approaches

- Mutual information: how much dependence exists?
- Conditional statistics: when and at what lag does dependence appear?
- Degrees of freedom: what dynamical structure is shared?

Mutual information

Measure independence of random variables

- independence means: $P(X,Y)=P_X(X)P_Y(Y)$
- distance of the two distributions → Kullback-Leibler divergence

$$I(X,Y)=\sum_{X,Y} P(X,Y)\log\frac{P(X,Y)}{P_X(X)P_Y(Y)}$$

- other interpretation: $I(X,Y)=H(X)+H(Y)-H(X,Y)$
- mutual information is zero for independent variables
- challenges:
  - more observable?
  - estimate joint distribution in high dimensions

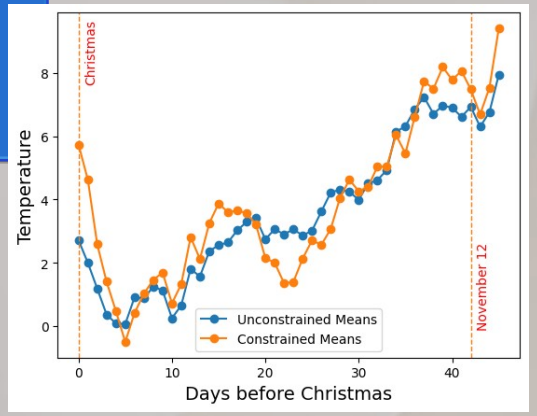
Conditional statistics

Robust estimate of dependence

- independence means: $P(Y \mid X) = P(Y) \;\Rightarrow\; E(f(Y) \mid g(X)) = E(f(Y))$
- Examples:
  - threshold events
  - means, variances
  - distribution bins
- A detected difference is evidence against independence

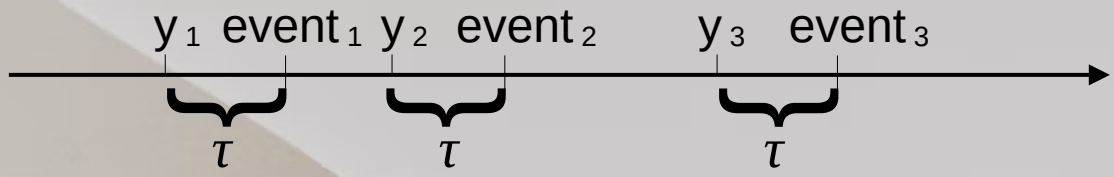
Choose observables that remain estimable with limited data.

Event-triggered temporal analysis



What can cause our observed events?

- events are defined by $g(X)=g_0 \Rightarrow t_1,t_2,\ldots,t_K$
- compute the event-conditioned average: $F(\tau)=\frac{1}{K}\sum_{k=1}^{K} f(Y(t_k-\tau))$
- compare with the unconditional baseline: $E(f(Y))$



Scanning tau reveals the temporal profile of dependence.

Degrees of freedom method

Having more information we can address questions about the relationship of two data series

- independence: $X \perp Y$
- mutual interaction: $X \Leftrightarrow Y$
- directed influence (drive): $X \rightarrow Y$ or $Y \rightarrow X$
- common cause: $\exists Z : Z \rightarrow X,\ Z \rightarrow Y$

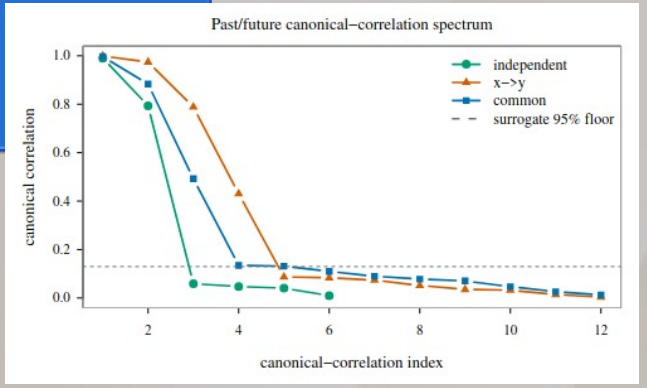
Degrees of freedom method

Idea to reveal the causal relationships

- if X → Y, then all information present in X shows up in Y, too!
- in deterministic case initial conditions determine the complete time series → **degrees of freedom (df)**
  - for example a 2nd order differential equation needs 2 initial conditions (x_0, v_0) → df=2
- in autonomous systems it is enough to fix df conditions anyhow
  - example: x(t=0) and x(t=1) is enough

Information content can be identified with the degrees of freedom.

Degrees of freedom method



How do we measure degrees of freedom?

- estimate $P(x_n \mid \text{ history of length } m)$
- if m ≥ df, then all values are fixed → distribution is a delta-function
- if m < df, then we get a broader distribution
- estimate information content from the (conditional) entropy
- works also in stochastic case, up to
  - there always remains an uncertainty due to the noise
  - change from m<df to m≥df regime is not sharp (elbow behavior)

Degrees of freedom method

How can we compute the causal categories?

- consider two time series X, Y
- estimate the dimension dX, dY, dXY
- causal relationships
  - $\mathrm{df}_X + \mathrm{df}_Y = \mathrm{df}_{XY} \Leftrightarrow X \perp Y$
  - $\mathrm{df}_X = \mathrm{df}_Y = \mathrm{df}_{XY} \Leftrightarrow X \Leftrightarrow Y$
  - $\mathrm{df}_X \leq \mathrm{df}_Y = \mathrm{df}_{XY} \Leftrightarrow X \Rightarrow Y$ and $\mathrm{df}_Y \leq \mathrm{df}_X = \mathrm{df}_{XY} \Leftrightarrow Y \Rightarrow X$
  - $max(\mathrm{df}_X, \mathrm{df}_Y) < \mathrm{df}_{XY} < df_X + df_Y \Leftrightarrow \exists Z : Y \Rightarrow X, Z \Rightarrow Y$

Degrees of freedom method

Implementation → Scientific Reports, 2026

- fixing more and more data → less and less data remains
- use simple observables → binarize data
- measure conditional entropy on binarized samples
- use AI (random forest) to tell the df
- train on robust models (Markov processes, Hénon process)
- test effectivity on simulated and real data → state-of-art results

Degrees of freedom method

Applications → **Scientific Reports, 2026**

- Controlled examples: Markov processes, Hénon maps, coupled Lorenz systems
- Temporal questions: the chicken-and-egg problem
- Observed systems: sunspots and global temperature
- Further examples: fish populations / sea surface temperature; epileptic EEG

Causal relationships

What should we trust?

- In reality, everything acts on everything else (up to physical constraints)
- df depends on the observational precision → noise level
- causal relations can depend on the precision → e.g. superluminar light propagation
- for two generic time series the causal relation is not necessarily homogeneous
  - time dependence
  - observational variable dependence

Limitation on modelling

A fundamental limitation may show up

- Each individual input can be independent of the target
  - examples: image recognition, sound recognition
- To increase relevance we need combinations of the inputs
- But in principle any kind of combination can occur! → number of possibilities grow rapidly
- External structure — such as locality in CNNs — restricts the search

With few observations, hidden-feature causality may be indistinguishable from randomness.

Conclusions

Take home messages

- High-dimensional prediction needs relevance screening
- causality methods help to understand possible relations in a model independent way
  - mutual information: measures statistical dependence
  - conditional statistics: yields information about dependence
  - degrees of freedom method: reveals causal relationships

Find structure first — build predictive models second.

The end

