# Model selection over partially ordered sets

Armeen Taeb*, Peter Bühlmann°, Venkat Chandrasekaran†

Department of Statistics, University of Washington *; Seminar for Statistics, ETH Zürich °; Departments of Computing and Mathematical Sciences and of Electrical Engineering, Caltech †

## Motivation

Model selection with Boolean-logical structure:

- formulate and test hypothesises, e.g. is this variable present?
- easy to define model complexity and false positives

What about for problems that lack Boolean-logical structure?

- ranking: global structure of transitivity
- clustering: global structure of set-partitions
- causal inference: global structure of acyclicity
- continuous problems, e.g. blind-source separation

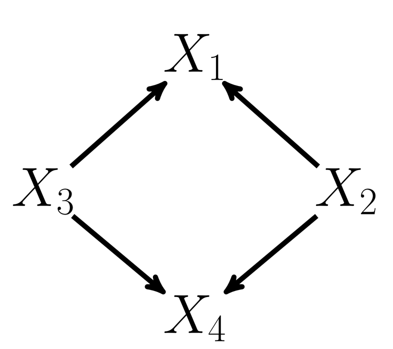
## Shortcomings of the standard perspective

Example I: clustering

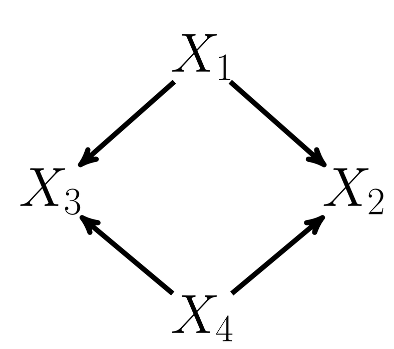
true clusters $= \{a, b\}, \{c\}$ estimated clusters $= \{a, b, c\}$

Boolean-logical perspective: FD = 2

Example II: causal structure learning




(a) true CPDAG




(b) estimated CPDAG

Boolean-logical perspective: FD = 4

## Model organization via posets

Models organized according to a poset $\mathcal{L}$ with relations $\preceq$:

| Attribute | Meaning |
|---|---|
| $\preceq$ | containment between simpler & more complex models |
| least element | the "null" model representing no discoveries |
| rank$(\cdot)$ | measures complexity of a model |

## False discovery framework

**Similarity valuation**: A symmetric function $\rho : \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ with:

- $0 \leq \rho(x, y) \leq \min\{\text{rank}(x), \text{rank}(y)\}$ for all $x, y \in \mathcal{L}$,
- $\rho(x, y) \leq \rho(z, y)$ for all $x \preceq z$,
- $\rho(x, y) = \text{rank}(x)$ if and only if $x \preceq y$.

### Definitions

Letting $x^\star \in \mathcal{L}$ be a true model and $\hat{x} \in \mathcal{L}$ be an estimate.

$$\text{TD}(\hat{x}, x^\star) \triangleq \rho(\hat{x}, x^\star),$$
$$\text{FD}(\hat{x}, x^\star) \triangleq \text{rank}(\hat{x}) - \rho(\hat{x}, x^\star) = \text{rank}(\hat{x}) - \text{TD}(\hat{x}, x^\star),$$
$$\text{FDP}(\hat{x}, x^\star) \triangleq \frac{\text{rank}(\hat{x}) - \rho(\hat{x}, x^\star)}{\text{rank}(\hat{x})} = \frac{\text{FD}(\hat{x}, x^\star)}{\text{rank}(\hat{x})}.$$

**Goal**: *maximize rank subject to false discovery control*

Suitable similarity valuation: $\rho_{\text{meet}}(\hat{x}, x^\star) \triangleq \max_{z \preceq \hat{x}, z \preceq x^\star} \text{rank}(z).$

- FD in clustering:

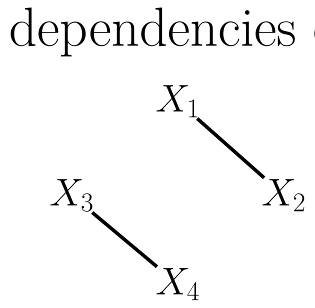
  \# groups in the coarsest common refinement minus \# groups in $\hat{x}$

  Example I: common refinement $= \{a, b\}, \{c\} \Rightarrow \text{FD}(\hat{x}, x^\star) = 1$

- FD in causal:

  \# edges in $\hat{x}$ minus \#edges in a densest CPDAG that contains conditional dependencies encoded in both $\hat{x}, x^\star$

Example II: densest =




$\Rightarrow \text{FD}(\hat{x}, x^\star) = 2$

Other appropriate similarity valuations in e.g. total ranking, subspace selection and blind source separation

## Greedy approaches to model selection

Starting from least model, greedily grow model complexity

Key ingredients:

- data-driven function $\Psi$: measures statistical significance for moving between neighboring models
- minimal set of neighboring models $\mathcal{S}$: accounting for invariances

**Theorem:** $\Psi_{\text{stable}}$: based on subsampling and stability of a base procedure, and $\Psi_{\text{test}}$: based on testing; used-specified $\alpha \in (0, 1)$

$$\Psi_{\text{stable}} : \quad \mathbb{E}[\text{FD}(\hat{x}, x^\star)] \leq \sum_k \frac{q_k^2}{|\mathcal{S}_k|(1 - 2\alpha)},$$
$$\Psi_{\text{test}} : \quad \mathbb{P}\left(\text{FD}(\hat{x}, x^\star) > 0\right) \leq \alpha|\mathcal{S}|.$$

- $\mathcal{S}_k =$ restriction of $\mathcal{S}$ to a specific rank
- $q_k =$ avg. discoveries by base procedure w.r.t. specific rank

## Experiments

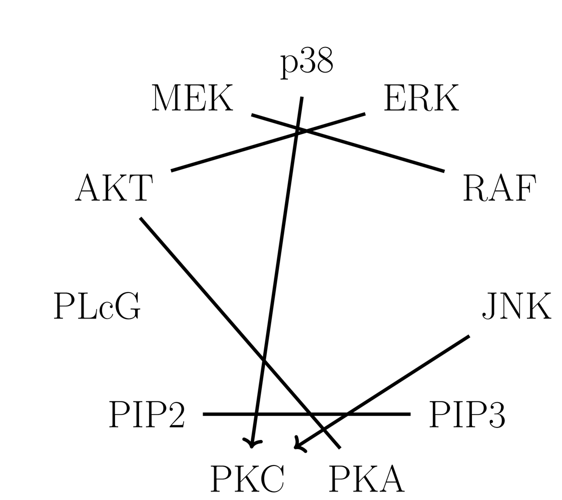
**Ranking educational systems:** improving ranking of countries based on new PISA test scores: base ranking from 2015 scores

- new ranking from 2018 test scores using our algorithm with $\Psi_{\text{test}}$ with family-wise-error control at level 0.05

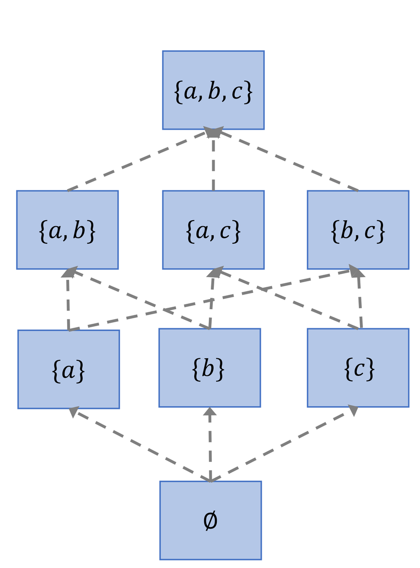
**Causal discovery from biological data:** identifying causal relationships among proteins from Sachs dataset

- CPDAG estimated using our algorithm with desired FD level = 2

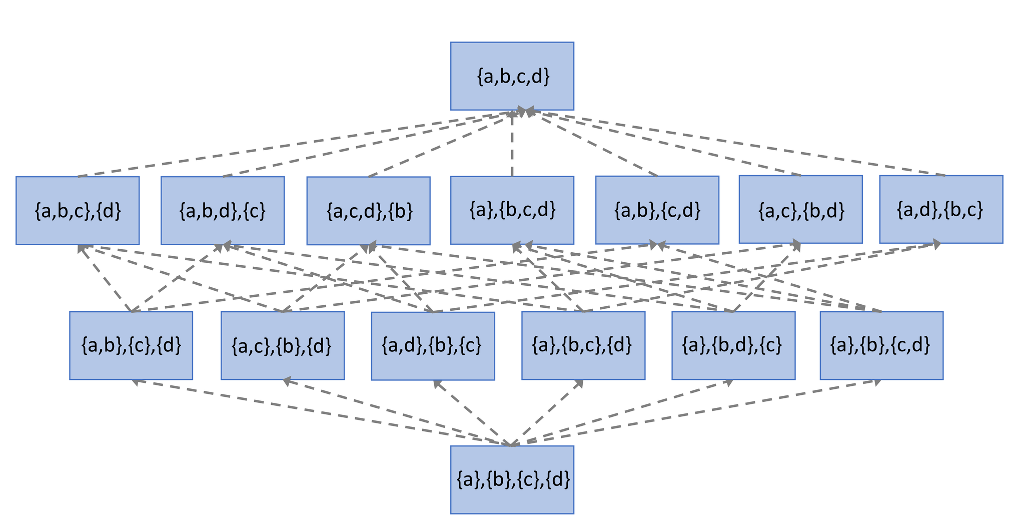
| | 2015 base ranking | testing approach |
|---|---|---|
| 1 | CAN | FIN |
| 2 | FIN | IRL |
| 3 | IRL | EST |
| 4 | EST | CAN |
| 5 | KOR | KOR |
| 6 | JPN | JPN |
| 7 | NOR | NOR |
| 8 | NZL | NZL |
| 9 | DEU | POL |
| 10 | POL | DEU |
| 11 | SvN | AUS |
| 12 | NLD | SWE |
| 13 | AUS | SvN |
| 14 | SWE | DNK |
| 15 | DNK | NLD |



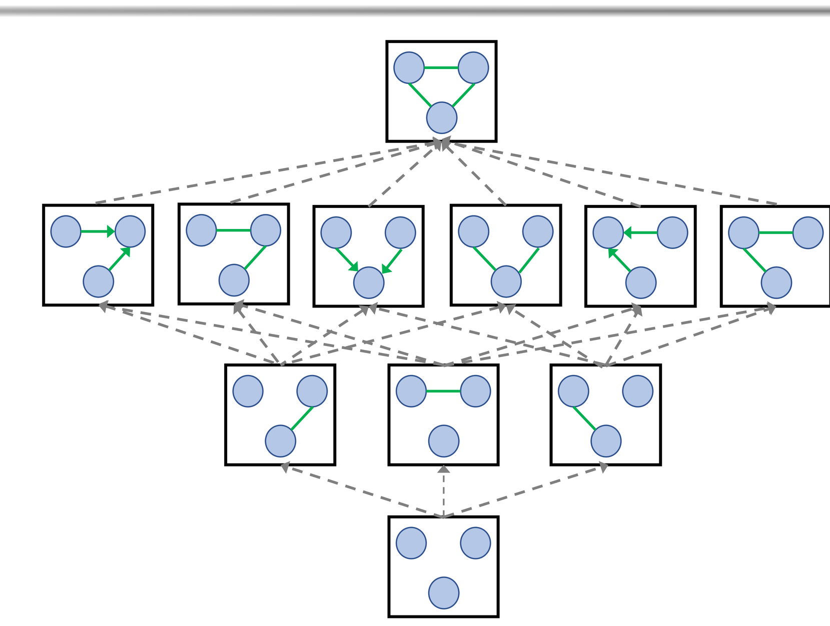

left: ranking
right: causal



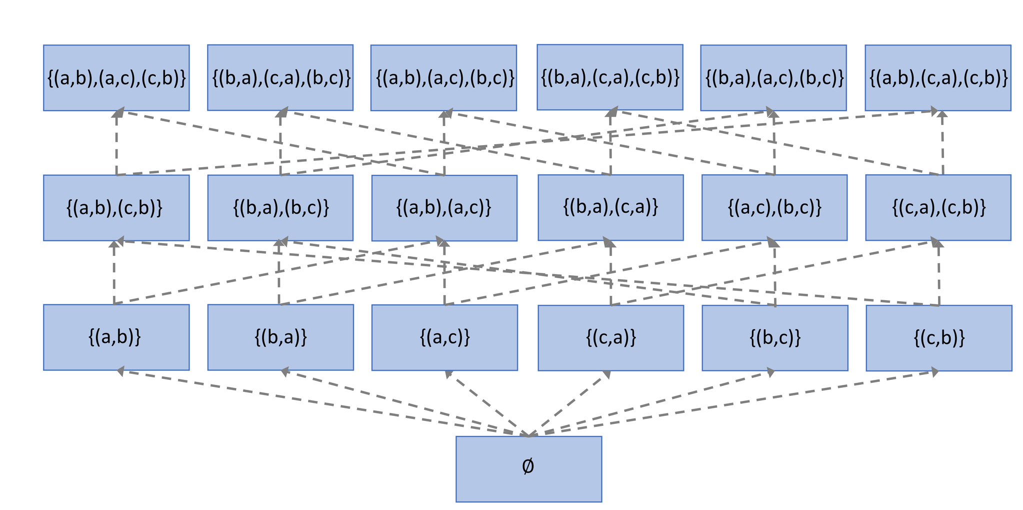

(a) Variable selection



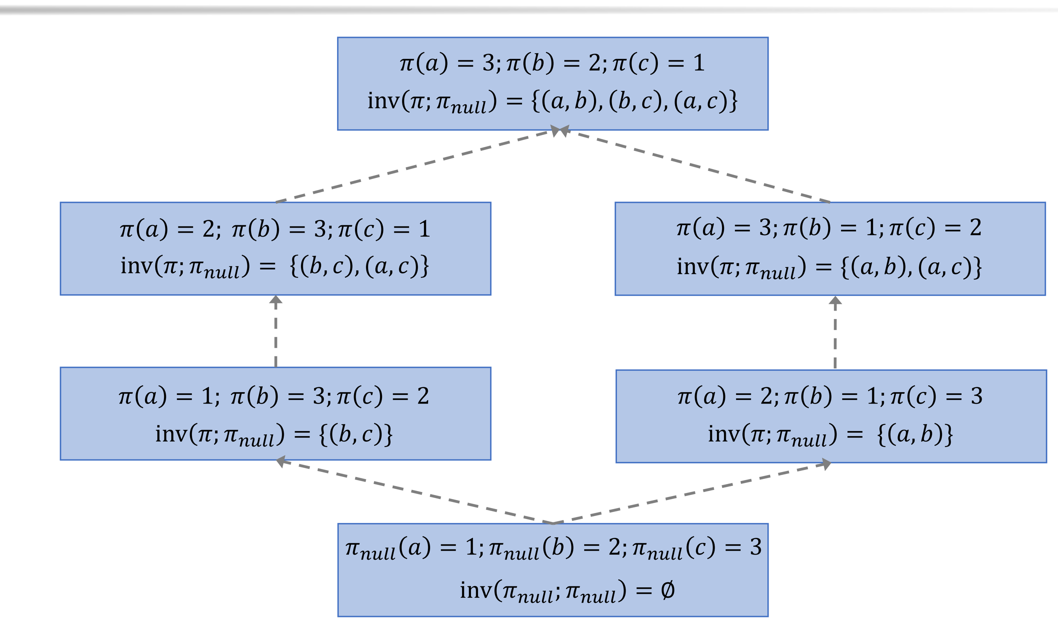
(b) Clustering



(c) Causal inference



(d) Partial ranking



(e) Total ranking