Perturbations and causality in Gaussian latent variable models

Armeen Taeb

ETH Zürich

Collaborators

Peter Bühlmann

Juan Gamella

Christina Heinze-Deml

CA water reservoir network

Reservoirs: central source of water

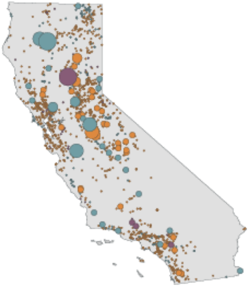
- buffer against severe drought
- hydroelectric power, agriculture, etc.

Water managers must assess:

- likelihood of system-wide failure
- effectiveness of potential policies

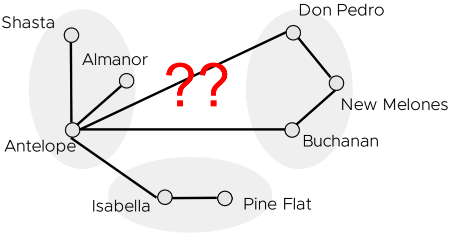
Shasta lake

CA reservoir system



Highly interconnected network

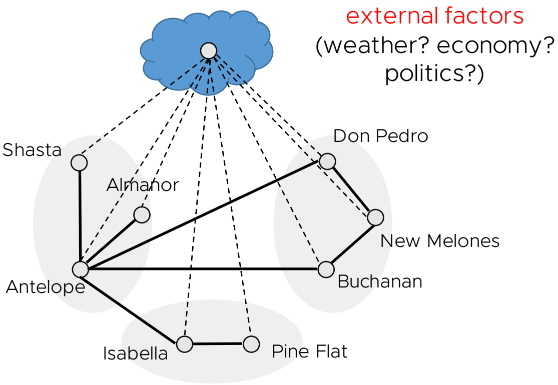
Understand reservoir interdependencies



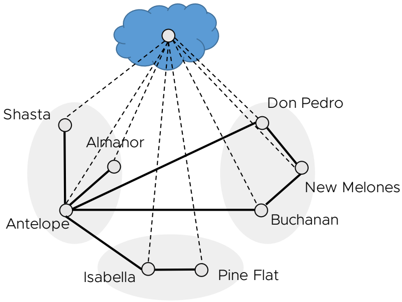
$\Rightarrow$ Graphical models

Challenge

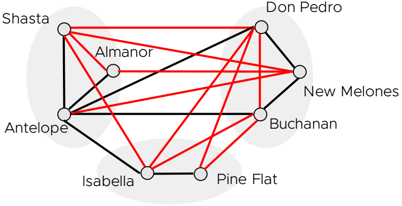
Many driving factors are latent



Challenge



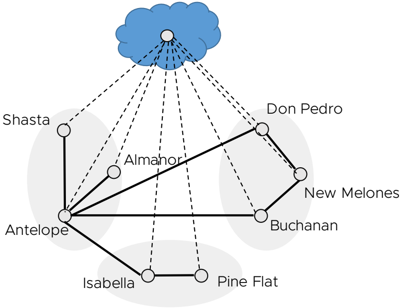
accounting for
latent variables



without accounting
for latent variables

Policy decisions based on models without latent variables
$\rightarrow$ sub-optimal management

Accounting for latent variables



Undirected Gaussian graphical model with latent variables:

- **Taeb** et al., *Water Resources Research*, 2017

Causal inference with latent variables?

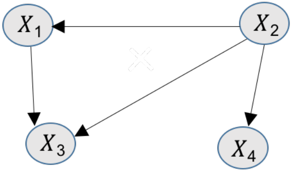
Causal questions

Causal questions are about the mechanism generating the data or predictions after external interventions

E.g. how does change in water level of one reservoir affects water levels of other reservoirs?

Directed acyclic graphs

- Directed acyclic graphs (DAGs) represent causal relations



- Pearl, Spirtes-Glymour-Scheines, Bollen, Dawid, Robins, Richardson, Didelez, Maathuis, ...

DAGs are useful

- Encode conditional dependency relationships
- Causal interpretation: extrapolate to unseen environments

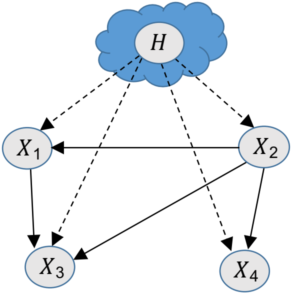
DAGs are useful

- Encode conditional dependency relationships
- Causal interpretation: extrapolate to unseen environments

$\boxed{\text{DAGs are typically unknown and must be learned from data}}$

Objective in a nutshell

Using data, find the causal relations (DAG) among observed X with latent confounding H



i.e. identify solid edges

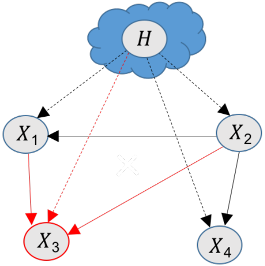
Linear structural causal models (SCM)

- Generating process for X via SCM [Pearl, 2000]:

$$\boxed{X_j \leftarrow b_j^\star X_{\mathrm{pa}(j)} + \gamma_j^\star H + \epsilon_j \quad \text{for all } j}$$

i.e. each X_j is expressed as a linear combination of its parents

- Example:



X_3 is a linear combination of X_1, X_2, H

Linear structural causal models (SCM)

Compact description:

$$\boxed{X \leftarrow {\color{red}B^\star} X + \Gamma^\star H + \epsilon} \quad ; \quad B^\star_{i,i} = 0 \;\forall\; i$$

- ${\color{red}B^\star}$: lower triangular (assuming X ordered) with

$$X_k \text{ parent of } X_j \iff B^\star_{j,k} \neq 0$$

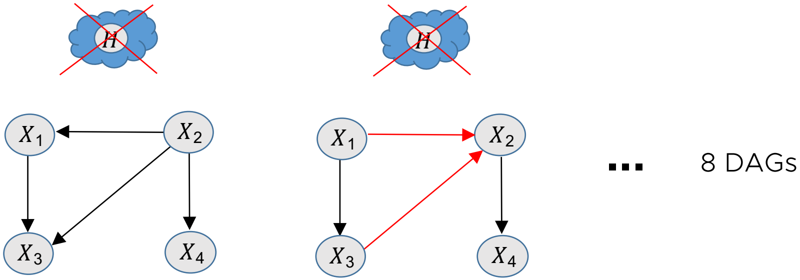
- $\Gamma^\star$: effect of latent variables
- ϵ: noise term with independent components; $\epsilon \perp\!\!\!\perp H$

Goal: identify ${\color{red}B^\star}$ from data of X

Challenging with observational data alone

$$X \leftarrow B^{\star}X + \Gamma^{\star}H + \epsilon$$

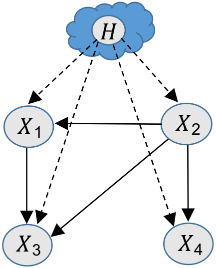
- $\Gamma^{\star} = 0$ (i.e. no latent effects): identifiability up to a "Markov equivalence class" [Pearl, 2000]

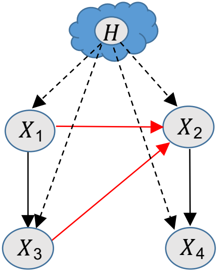


Challenging with observational data alone

$$X \leftarrow B^\star X + \Gamma^\star H + \epsilon$$

- $\Gamma^\star \neq 0$: even harder to identify $B^\star$



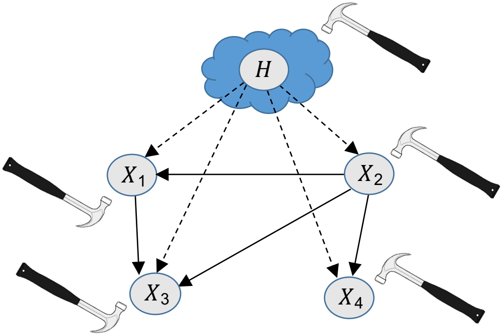


... $\gg$ 8 DAGs

- If dense latent effects and sparse DAG, identifiability up to "Markov equivalence class" [Frot et al., 2019]

Interventions

Direct external *perturbations* on the variables



Exploiting perturbations can be very useful

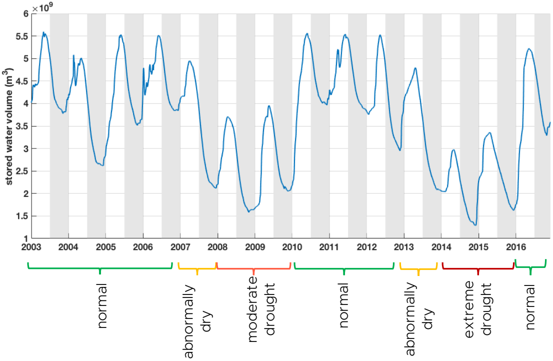
Improve identifiability and robustness to distributional changes
[Bühlmann, Meinshausen, Eberhardt, Schölkopf, Uhler, . . .]

Model and exploit perturbations
with unknown strengths and locations

Perturbation model

- Data (X^e, H^e) from observed environments, experimental conditions or sub-populations $e \in \mathcal{E}$

- Example: reservoir volumes in different "time blocks"



Perturbation model

For every $e \in \mathcal{E}$, X^e is generated according to the SCM:

$$\boxed{X^e \leftarrow B^\star X^e + \Gamma^\star H^e + \epsilon + \delta^e}$$

$(H^e, \delta^e, \epsilon)$ jointly independent

- δ^e: additive noise interventions independent across coordinates

Why should perturbations help?

$$\text{for all } e \in \mathcal{E}:$$
$$X^e \leftarrow B^\star X^e + \Gamma^\star H^e + \epsilon + \delta^e$$

$B^\star, \Gamma^\star$ are invariant across all $e \in \mathcal{E}$,

- same $B^\star, \Gamma^\star$ fit data equally well for all $e \in \mathcal{E}$

Data across all $e \in \mathcal{E}$ substantially improves identifiability

Previous work

Method	Perturbed response	Perturbed latents
Imbens, 1997 Peters et al., 2015 Rothenhäusler et al., 2019	×	×
Rothenhäusler et al., 2015	✓	×
Taeb and Bühlmann, 2021	✓	✓

Previously: output a single causal structure $\Rightarrow$

do not give an equivalence class when non-identifiable

Previously: do not directly model the latent effects $\Rightarrow$

reduced identifiability and worse small sample performance

Modeling simplifications

$$\text{for all } e \in \mathcal{E}:$$
$$X^e \leftarrow B^\star X^e + \Gamma^\star H^e + \epsilon + \delta^e$$

- $\epsilon + \delta^e \sim \mathcal{N}(0, \text{diag}(w^{e,\star}))$
- $H^e \sim \mathcal{N}(0, \psi^{e,\star}\mathcal{I})$: i.i.d. latent variables

Marginal of X^e specified by: $\underbrace{(B^\star, \Gamma^\star)}_{\text{invariant parameters}} \quad , \quad \underbrace{(w^{e,\star}, \psi^{e,\star})}_{\text{perturbation parameters}}$

Maximum-likelihood framework

Given: i.i.d. data of $\{X^e : e \in \mathcal{E}\}$ with $|\mathcal{E}| = m$

Given: candidate DAGs $\mathcal{D}_{\text{cand}}$

Solve the regularized MLE:

$$\begin{aligned} \arg\min \quad & \sum_{e=1}^{m} \ell(\cdot; X^e \text{ data}) + \lambda \text{ complexity}(\mathcal{D}) \\ \text{subject-to} \quad & \mathcal{D} \in \mathcal{D}_{\text{cand}} \end{aligned}$$

Maximum-likelihood framework

Given: i.i.d. data of $\{X^e : e \in \mathcal{E}\}$ with $|\mathcal{E}| = m$

Given: candidate DAGs $\mathcal{D}_{\mathsf{cand}}$

Solve the regularized MLE:

$$\begin{aligned} \arg\min \quad & \sum_{e=1}^{m} \ell(\cdot; X^e \text{ data}) + \lambda \text{ complexity}(\mathcal{D}) \\ \text{subject-to} \quad & \mathcal{D} \in \mathcal{D}_{\mathsf{cand}} \end{aligned}$$

- searching among candidate DAGs

Maximum-likelihood framework

Given: i.i.d. data of $\{X^e : e \in \mathcal{E}\}$ with $|\mathcal{E}| = m$

Given: candidate DAGs $\mathcal{D}_{\text{cand}}$

Solve the regularized MLE:

$$\begin{aligned} \arg\min \quad & \sum_{e=1}^{m} \ell(\cdot; X^e \text{ data}) + \lambda \, \text{complexity}(\mathcal{D}) \\ \text{subject-to} \quad & \mathcal{D} \in \mathcal{D}_{\text{cand}} \end{aligned}$$

- complexity($\mathcal{D}$): # edges in the moral graph of $\mathcal{D}$ (undirected)

Maximum-likelihood framework

Given: i.i.d. data of $\{X^e : e \in \mathcal{E}\}$ with $|\mathcal{E}| = m$

Given: candidate DAGs $\mathcal{D}_{\text{cand}}$

Solve the regularized MLE:

$$\begin{aligned} \arg\min \quad & \sum_{e=1}^{m} \ell(\cdot; X^e \text{ data}) + \lambda \text{ complexity}(\mathcal{D}) \\ \text{subject-to} \quad & \mathcal{D} \in \mathcal{D}_{\text{cand}} \end{aligned}$$

Tuning parameters: λ, # columns in Γ (# latent variables)

Maximum-likelihood framework

How do we obtain candidate DAGs?

- domain expertise
- DAG learning algorithm on the pooled data

How do we solve MLE?

- alternating minimization: convex w.r.t. B (fixing other parameters)

Remarks:

- candidate DAGs may be dense $\rightarrow$ greedy backward deletion
- under dense latents $+$ sparse DAG, our algorithm obtains $B^\star$ as one of the solution asymptotically

Impossibility results

Thm. [**Taeb** et al, 2021]: Without constraining the latent effects, the problem is ill-posed:

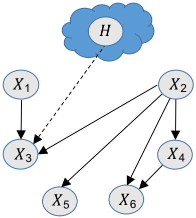
$$\mathcal{D}_{\mathsf{opt}} = \{\text{all DAGs}\}$$

Result holds regardless of amount or size of interventions on the observed variables

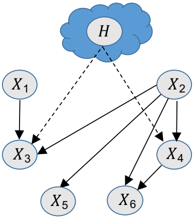
Sufficient conditions for identifiability

1. latent variables induce $\boxed{\text{some spurious dependencies}}$

Bad:

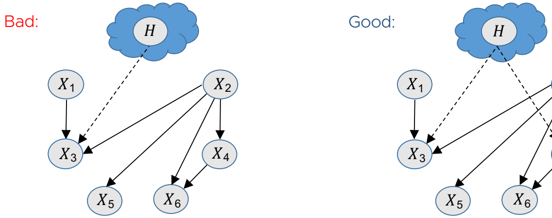


Good:



Sufficient conditions for identifiability

1. latent variables induce $\boxed{\text{some spurious dependencies}}$



Two environments where:

2. $\boxed{\text{strong perturbations}}$ on all observed variables where:
   perturbation stren. on observed $\gg$ perturbation stren. on latents
3. $\boxed{\text{heterogeneous perturbations}}$ across environments

Identifiability guarantees

B_{opt}: optimal solution(s) of MLE in population

Thm. [**Taeb** et al, 2021]: under Assumptions 1-4:

$$B^\star = \arg\min_B \quad \#\ \text{edges}[\text{moral}(B)] \ \ \text{subject-to} \quad B \in B_{\text{opt}}$$

Identifiability guarantees

B_{opt}: optimal solution(s) of MLE in population

Corr: regularized MLE is consistent

Corr: consistent with linear (non-Gaussian) SCM

Corr: $B^\star$ minimizes the worst-case risk over distributional shifts

Equivalence classes

If assumptions are not satisfied $\Rightarrow$ equivalence class of DAGs

DAGs in the equivalence have the same likelihood score
$\Rightarrow$ can be determined by finding best scoring DAGs

Equivalence classes

If assumptions are not satisfied $\Rightarrow$ equivalence class of DAGs

DAGs in the equivalence have the same likelihood score
$\Rightarrow$ can be determined by finding best scoring DAGs

Thm. [**Taeb** et al] Graphical characterization[a][b]:

$$\text{Equivalence Class} = \{\mathcal{D} \in \mathrm{MEC} : \mathtt{PA}_{\mathcal{D}}(i) = \mathtt{PA}_{\mathcal{D}^\star}(i) \text{ for all } i \in \mathrm{I}^\star\}$$

where:

- MEC: Markov equivalence class of population DAG $\mathcal{D}^\star$
- $\mathrm{I}^\star$: intervened variables

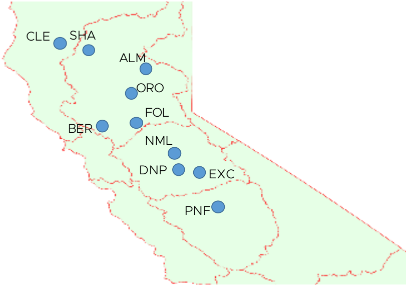
[a] Assumes dense latent effects and sparse underlying DAG
[b] Generally: equivalence class $\supseteq \{\mathcal{D} \in \mathrm{MEC} : \mathtt{PA}_{\mathcal{D}}(i) = \mathtt{PA}_{\mathcal{D}^\star}(i) \text{ for all } i \in \mathrm{I}^\star\}$

Reservoirs

10 largest reservoirs in California
with monthly volume data

4 environments corresponding to
drought severity:

condition	period
normal	2003-2006, 2010-2012, 2016
abnormally dry	2007, 2013
moderate drought	2008-2009
extreme drought	2014-2015



Reservoirs

Interventions on the latent variables in all environments

Interventions on: FOL, NML, ORO (strongest in the extreme drought environment)

In 2014-2015:

- NML released large amount of water to farmers
- FOL released large amount of cold water for salmon-spawning

Summary

This talk: perturbation model for Gaussian observed and latent variables

- regularized maximum-likelihood estimator
- identifiability guarantees and equivalence class characterizations

Ongoing work:

- more efficient algorithms (e.g. greedy methods)?
- high-dimensional consistency guarantees?
- designing active intervention strategies?

paper: *arXiv 2101.06950*