CAM: Causal Additive Models, High-dimensional Order Search and Penalized Regression

Peter Bühlmann, Jonas Peters, Jan Ernest; AOS, 2014

September 19, 2023

Table of Contents

Table of Contents

Preliminaries: Directed Acyclic Graph (DAG)

- p variables: $\boldsymbol{X} = (X_1, \ldots, X_p)$.
- One node for each variable. Set of nodes: $V = \{1, \ldots, p\}$.
- Set of edges: $E = \{(i, j) \in V^2 : i \to j\}$.
- $D = (V, E)$ is a DAG if all the edges are directed and there are no cycles.

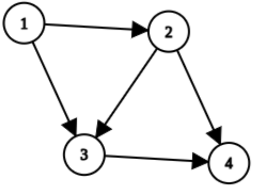

Figure: Example of a DAG

- Here, $V = \{1, \ldots, 4\}$.
- $E = \{(1,2),(1,3),(2,3),(2,4),(3,4)\}$.
- If we include another edge, $4 \to 1$, then it is no longer a DAG.

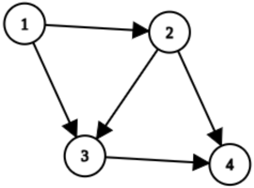

Figure: Parents of Nodes in a DAG

- Define, $\text{pa}_D(i) = \{k \in V : (k, i) \in E\}$.
- $\text{pa}_D(i)$ is the set of parents of node i in the DAG D.
- $\text{pa}_D(i)$ consists of all nodes that has direct edge to node i.
- For example, $\text{pa}_D(1) = \phi$ and $\text{pa}_D(3) = \{1, 2\}$, $\text{pa}_D(4) = \{2, 3\}$ etc.

Table of Contents

Basic Idea: Causal Discovery

- Suppose p variables, $\boldsymbol{X}$ has a joint distribution, $\mathbb{P}$.
- Assume, there exists a DAG D, which describe the **true** data generating process.
- Then, it is possible to infer **true** causal relationship between p variables.

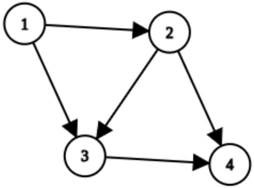

Figure: Inferring Causal Relationships using a DAG

- For instance, assume the DAG D represents $\mathbb{P}$ perfectly.
- Then, X_1 is a common cause for both X_2 and X_3.
- X_1 do not affect X_4 directly, but may have indirect causal effect which mediates through X_3, or X_2 or both.

Preliminaries: Causal Discovery

Given the observed data, how can we recover the DAG?

- Conditional Independent Tests:
    - PC (Spirtes et al., 2000) and its variants.
- Optimizing a Score:
    - GES (Chickering, 2002)
- Structural Equation Model (SEM):
    - Linear SEM (Peters and Bühlmann, 2013)
    - Non-linear SEM (Bühlmann et al., 2014)
    - Partially Linear SEM (Rothenhäusler et al., 2018)
- and other methods...

Structural Equation Model (SEM)

- General SEM:

$$X_j = f_j(\boldsymbol{X}_{\mathrm{pa}_D(j)}, \epsilon_j) \quad \epsilon_1, \ldots, \epsilon_p \text{ (mutually) independent}$$

  $\{f_j : 1 \leq j \leq p\}$ are unknown functions.
- Too general; lacks identifiability.

Additive SEM

- Functions $f_j(.)$ are additive in its arguments:

$$X_j = \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \epsilon_j \tag{1}$$

  $\epsilon_1, \ldots, \epsilon_p$ independent with $\epsilon_j \sim \mathrm{Normal}(0, \sigma_j^2)$.
- Too many structural assumptions.
- But, identifiability is achieved.

Additive SEM

Q: What do we mean by identifiability here?

- Let $\mathbb{P}$ is generated by model (1) with DAG D and functions $f_{j,k}$.
- And $\mathbb{Q}$ is generated by model (1) with a different DAG, $D'(\neq D)$ and different set of functions $f'_{j,k}$.
- Then, under some conditions on $f_{j,k}$ and $f'_{j,k}$,

$$\mathbb{Q} \neq \mathbb{P}$$

(See Lemma 1)

Rewrite model (1) as,

$$
\begin{aligned}
X_j &= \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \epsilon_j \\
&= \sum_{k \neq j} f_{j,k}(X_k) + \epsilon_j
\end{aligned}
$$

$$f_{j,k}(.) \neq 0 \text{ iff there is a directed edge } k \to j \text{ in } D$$

- Parameters to Estimate:

$$\theta = (f_{1,2}, \ldots, f_{1,p}, f_{2,1}, \ldots, f_{p-1,p}, \sigma_1, \ldots, \sigma_p)$$

- How?
  - ϵ_j's are normal; Can we use Likelihood?
  - Can we perform regressions?

- Regression is possible. But,
  - we have p regressions.
  - No defined set of response or covariate.
- We need some criterion to **order the variables**.

(Section 1.1) "If the order among the variables would be known, the problem boils down to variable selection in multivariate (potentially nonlinear) regression;"

Table of Contents

Causal Ordering

- Let, π is a permutation of $\{1, \ldots, p\}$.
- Define $\boldsymbol{X}^\pi = (X_1^\pi, \ldots, X_p^\pi)$, where,

$$X_j^\pi = X_{\pi(j)}$$

- Let, D^π be the fully connected DAG with edges $\pi(k) \rightarrow \pi(j)$ for all $k < j$.

- For instance, fix $p = 4$ and π such that,

$$\pi(1) = 2,\ \pi(2) = 3,\ \pi(3) = 4,\ \pi(4) = 1$$

- D^π consists of edges $\pi(k) \to \pi(j)$ for all $k < j$.
- D^π is a super-DAG of true DAG D.

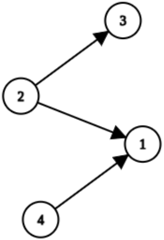

Figure: True DAG, D

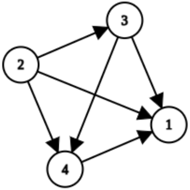

Figure: Fully connected DAG, D^π

- For any permutation π, we can construct a fully-connected DAG, D^{π}.
- But, D is not fully-connected. True order is not unique.
- For instance, the following two permutations respects the **causal ordering** in D.

$$\pi(1) = 2,\ \pi(2) = 3,\ \pi(3) = 4,\ \pi(4) = 1$$
$$\pi(1) = 2,\ \pi(2) = 4,\ \pi(3) = 3,\ \pi(4) = 1$$

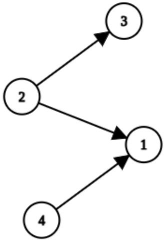

Figure: True DAG, D

- Let, D^0 is the true DAG.
- Define, the set of true ordering,

$$\Pi^0 = \{\pi^0 : \text{ the fully connected DAG } D^{\pi^0} \text{ is a super-DAG of } D^0\}$$

- If we can identify Π^0, then we need to remove edges to reach D^0 (why? - Next Slide).
- How to remove edges? - Regression + variable selection.

"Any true ordering of permutation π^0 allows for a lower-triangular representation" (See equation (5))

- Recall $\pi \in \Pi^0$ for the true DAG, $D^0 = D$, where

$$\pi(1) = 2, \; \pi(2) = 3, \; \pi(3) = 4, \; \pi(4) = 1$$

- Then, we can write,

$$\begin{aligned} X_2 &= \epsilon_2 \\ X_3 &= f_{3,2}(X_2) + \epsilon_3 \\ X_4 &= f_{4,2}(X_2) + f_{4,3}(X_3) + \epsilon_4 \\ X_1 &= f_{1,2}(X_2) + f_{1,3}(X_3) + f_{1,4}(X_4) + \epsilon_1 \end{aligned} \tag{2}$$

- Now, it is easier to recover D by (nonlinear) regression.

- But how to identify the set Π^0?
  - Section 2.4: Maximum Likelihood Estimation for Order (Low Dimension)
  - Section 3: Restricted MLE (High Dimension)

Table of Contents

Unrestricted MLE for Order Search

- For p variables, consider all possible permutations.
- For each permutation π, we have a lower triangular representation (similar to (2)).
- Estimate the functions $\hat{f}_j^\pi$ by some non-linear regression method (e.g., boosting).

$$\hat{f}_j^\pi = \underset{g_j}{\operatorname{argmin}} \left\| \boldsymbol{X}_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi) \right\|_2^2$$

- Estimate the variances

$$(\hat{\sigma}_j^\pi)^2 = \left\| \boldsymbol{X}_j^\pi - \sum_{k=1}^{j-1} \hat{f}_{j,k}^\pi(X_k^\pi) \right\|_2^2$$

- Maximize the unpenalized negative log-likelihood:

$$\hat{\pi} \in \underset{\pi}{\operatorname{argmin}} \sum_{j=1}^{p} \log(\hat{\sigma}_j^\pi)$$

- After estimating $\hat{\pi}$, we can construct, the fully connected DAG, $D^{\hat{\pi}}$.
- Variable selection on $D^{\hat{\pi}}$ to obtain the final estimated DAG, $\hat{D}^{\hat{\pi}}$.
- If $p = 10$, the number of possible permutations are $10! > 3 \times 10^6$.
- For each permutation, it involves $p - 1$ regressions.
- How to perform order search for high-dimensional data?

Restricted MLE on a Preliminary Neighborhood

- Regress X_j on $X_{-j} = \{X_k : k \neq j\}$ (Additive Regression; Group Lasso).
- $$\hat{\mathbb{E}}_{add}[X_j | X_{-j}] = \sum_{k \in \hat{A}_j} \hat{h}_{jk}(X_k)$$

  with, $\hat{A}_j = \{k : k \neq j, \hat{h}_{j,k} \neq 0\}$.
- $\hat{A}_j$: preliminary neighborhood of node j.
- Previously, we regress,

  $$X_{\pi(j)} \text{ on } \{X_k : k \in \{\pi(1), \ldots, \pi(j-1)\}\}$$
- Now we regress,

  $$X_{\pi(j)} \text{ on } \{X_k : k \in \{\pi(1), \ldots, \pi(j-1)\}\} \cap \hat{A}_{\pi(j)}$$

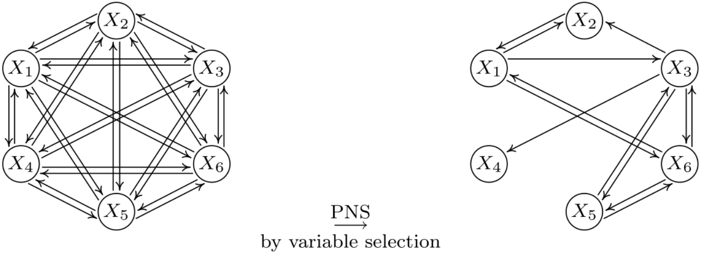

FIG. 1. *Step PNS. For each variable the set of possible parents is reduced (in this plot, a directed edge from X_k to X_j indicates that X_k is a selected variable in $\hat{A}_j$ and a possible parent of X_j). This reduction leads to a considerable computational gain in the remaining steps of the procedure.*

Here are some of the recent review papers on Causal Discovery:

- Heinze-Deml et al. (2018)
- Glymour et al. (2019)
- Vowels et al. (2022)

and reference therein...

References

Bühlmann, P., Peters, J., and Ernest, J. (2014). "Cam: Causal additive models, high-dimensional order search and penalized regression."

Chickering, D.M. (2002). "Optimal structure identification with greedy search." *Journal of machine learning research*, **3(Nov)**, 507–554.

Glymour, C., Zhang, K., and Spirtes, P. (2019). "Review of causal discovery methods based on graphical models." *Frontiers in genetics*, **10**, 524.

Heinze-Deml, C., Maathuis, M.H., and Meinshausen, N. (2018). "Causal structure learning." *Annual Review of Statistics and Its Application*, **5(1)**, 371–391. doi:10.1146/annurev-statistics-031017-100630.

Peters, J. and Bühlmann, P. (2013). "Identifiability of Gaussian structural equation models with equal error variances." *Biometrika*, **101(1)**, 219–228. ISSN 0006-3444. doi:10.1093/biomet/ast043.

Rothenhäusler, D., Ernest, J., and Bühlmann, P. (2018). "Causal inference in partially linear structural equation models." *The Annals of Statistics*, **46(6A)**, 2904–2938.

Spirtes, P., Glymour, C.N., and Scheines, R. (2000). *Causation, prediction,*

Thank You