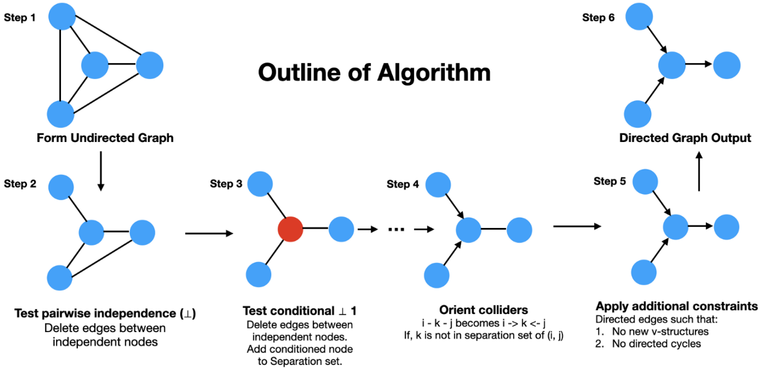
Identifiability of Additive Noise Models Using Conditional Variances

Hamed Vaheb

University of Luxembourg

November 3, 2022

Motivation



- **Assumption:** Faithfulness
- Recover at most Markov Equivalence Class

Motivation

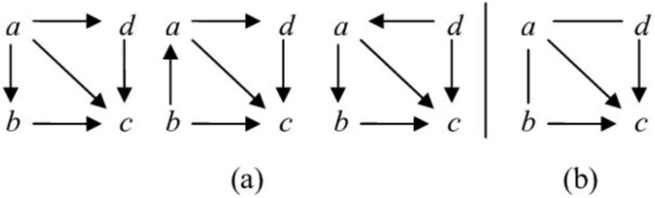
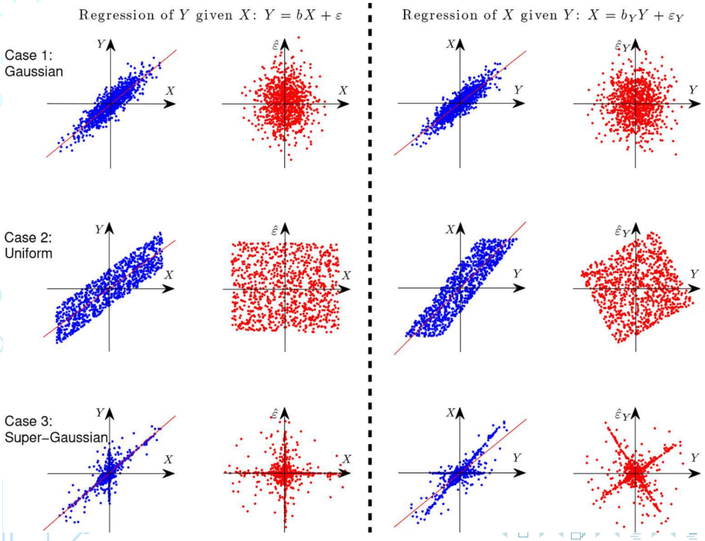
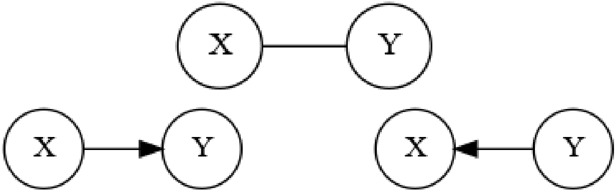


Figure: Markov Equivalence Class

Motivation


Regression of Y given X: $Y = bX + \varepsilon$
Regression of X given Y: $X = b_Y Y + \varepsilon_Y$
Case 1:
Gaussian
Case 2:
Uniform
Case 3:
Super–Gaussian
Y
X
$\hat{\varepsilon}$
$\hat{\varepsilon}_Y$

Motivation



(a) $Y := f(X) + \varepsilon$
$X \perp\!\!\!\perp \varepsilon$

(b) $X := f(Y) + \varepsilon$
$Y \perp\!\!\!\perp \varepsilon$

- **Identification:** asymmetry direction.

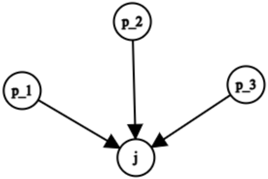
Generalize

- **Nonlinear, Non-Gaussian:**
  - Additive Noise Model (ANM)

$$Y := f_{AN}(X) + \varepsilon$$

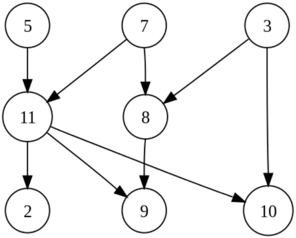
- **Multivariate:**

$$X_j = f_j(X_{Pa(j)}) + \varepsilon_j$$



- We need assumptions on the model!
- **Goal:**
  - Identification with less restrictive assumptions
  - Construct causal graph

Directed Graph Ordering



The graph shown to the left has many valid topological sorts, including:

- 5, 7, 3, 11, 8, 2, 9, 10 (visual top-to-bottom, left-to-right)
- 3, 5, 7, 8, 11, 2, 9, 10 (smallest-numbered available vertex first)
- 5, 7, 3, 8, 11, 10, 9, 2 (fewest edges first)
- 7, 5, 11, 3, 10, 8, 9, 2 (largest-numbered available vertex first)
- 5, 7, 11, 2, 3, 8, 9, 10 (attempting top-to-bottom, left-to-right)
- 3, 7, 8, 5, 11, 10, 2, 9 (arbitrary)

- **Ordering:** Directions of edges s.t.$(j,k) \in E, j$ comes before k
- True Ordering: $(\pi_1, \pi_2, ..., \pi_m)$
- Estimated Ordering: $(\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_m)$
- **Learning graph**: learning the ordering and the skeleton

Outline

Outline

Prior Identifiability Conditions

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$
- $var(\varepsilon_j) = \sigma_j$

Lemma

Set of dentifiable ANMs:

- *non-linear ANMs:*
  $(f_j)_{j \in V}$ *are not linear,*
- *non-Gaussian linear ANMs:*
  $(f_j)_{j \in V}$ *are linear*
  $(X_j)_{j \in V}$ *or* $(\varepsilon_j)_{j \in V} \not\sim \mathcal{N}$
- *(Gaussian) linear ANMs:*
  $(f_j)_{j \in V}$ *are linear*
  $(\sigma_j)_{j \in V}$ *are the similar or known.*

Conditional Variance Identifiability

- Introduced by Park [3]
- Without constraints on the form of $(f_j)_{j\in V}$ and $(\varepsilon_j)_{j\in V}$
- $(\varepsilon_j)_{j\in V}$ independent, but different distributions with heterogenous variance $(\sigma_j^2)_{j\in V}$
- Only scale of error variances, influence of parents.
- **Assumption**: Causal Minimality

Bivariate Example

$$X_1 \longrightarrow X_2$$
G_1

(a)
$X_1 = \varepsilon_1$
$X_2 = \beta_1 X_1 + \varepsilon_2$

$$X_1 \longleftarrow X_2$$
G_2

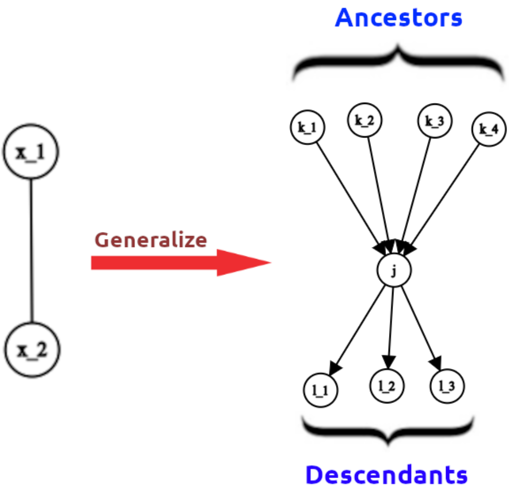
(b)
$X_1 = \beta_2 X_2 + \varepsilon_1$
$X_2 = \varepsilon_2$

$$X_1 \qquad X_2$$
G_3

(c)
$X_1 = \varepsilon_1$
$X_2 = \varepsilon_2$

- **G1**
  - $Var(X_2) > Var(X_1)$
  - $\mathbb{E}[Var(X_1|X_2)] < \mathbb{E}[Var(X_2|X_1)]$
- **Asymmetry**: amount of uncertainty projected from the model

Generalize



Conditional Variance Identifiability Condition - Forward

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$

Conditional Variance Identifiability Condition - Forward

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$
- $Var(X_2) = \underbrace{(E[Var(X_2|X_1)])}_{\sigma_2^2} + \underbrace{(Var(E[X_2|X_1]))}_{\beta_1^2\sigma_1^2} > \sigma_1^2 = Var(X_1)$

Conditional Variance Identifiability Condition - Forward

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$
- $Var(X_2) = \underbrace{(E[Var(X_2|X_1)])}_{\sigma_2^2} + \underbrace{(Var(E[X_2|X_1]))}_{\beta_1^2\sigma_1^2} > \sigma_1^2 = Var(X_1)$

Theorem

Let $P(X)$ be generated from an ANM with DAG G and true ordering π. Suppose that causal minimality holds. Then, DAG G is uniquely identifiable if for any node $j = \pi_m \in V, k \in De(j)$,

- *Forward stepwise selection:*
  $Var(X_j|X_{\pi_1}, ..., X_{\pi_{m-1}}) < Var(X_k|X_{\pi_1}, ..., X_{\pi_{m-1}})$
  $\sigma_j^2 < \sigma_k^2 + \mathbb{E}[Var(\mathbb{E}[X_k|X_{Pa(k)}]|X_{\pi_1}, ..., X_{\pi_{m-1}})]$

Conditional Variance Identifiability Condition - Forward

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$
- $Var(X_2) = \underbrace{(E[Var(X_2|X_1)])}_{\sigma_2^2} + \underbrace{(Var(E[X_2|X_1]))}_{\beta_1^2\sigma_1^2} > \sigma_1^2 = Var(X_1)$
- **Step 1**: $k \in De(\pi_1) = V \setminus \{\pi_1\}$
  $Var(X_{\pi_1}) = \sigma_{\pi_1}^2 < \sigma_k^2 + var(E[X_k|X_{Pa(K)}]) = var(X_k)$

Theorem

Let $P(X)$ be generated from an ANM with DAG G and true ordering π. Suppose that causal minimality holds. Then, DAG G is uniquely identifiable if for any node $j = \pi_m \in V, k \in De(j)$,

- *Forward stepwise selection:*
  $Var(X_j|X_{\pi_1}, ..., X_{\pi_{m-1}}) < Var(X_k|X_{\pi_1}, ..., X_{\pi_{m-1}})$
  $\sigma_j^2 < \sigma_k^2 + \mathbb{E}[Var(\mathbb{E}[X_k|X_{Pa(k)}]|X_{\pi_1}, ..., X_{\pi_{m-1}})]$

Conditional Variance Identifiability Condition - Backward

- $X_j = f_j(X_{Pa(j)}) + \varepsilon_j$
- $\mathbb{E}[Var(X_1|X_2)] = \underbrace{Var(X_1)}_{\sigma_1^2} - \underbrace{Var(\mathbb{E}[X_1|X_2])}_{\frac{\beta_1^2\sigma_1^4}{\beta_1^2\sigma_1^2+\sigma_2^2}} < \underbrace{\mathbb{E}[Var(X_2|X_1)]}_{\sigma_2^2}$
- **Step 1**: $\ell \in An(\pi_p) = V \setminus \{\pi_p\}$

$$Var(X_{\pi_p}|X_{V\setminus\pi_p}) = \sigma_{\pi_p}^2 > \sigma_\ell^2 - E[Var(E[X_\ell|X_{V\setminus\pi_p}]|X_{Pa(\ell)})] = Var(X_\ell|X_{V\setminus\ell})$$

Theorem

Let $P(X)$ be generated from an ANM with DAG G and true ordering π. Suppose that causal minimality holds. Then, DAG G is uniquely identifiable if for any node $j = \pi_m \in V, \ell \in An(j)$,

- *Backward Stepwise Selection:*
  $Var(X_\ell|X_{\pi_m}, ..., X_{\pi_p} \setminus X_\ell) < Var(X_{\pi_{m-1}}|X_{\pi_m}, ..., X_{\pi_p} \setminus X_\ell)$
  $\sigma_j^2 > \sigma_\ell^2 - \mathbb{E}[Var(\mathbb{E}[X_\ell|X_{\pi_m}, ..., X_{\pi_p} \setminus X_\ell]|X_{Pa(\ell)})]$

Outline

Ordering Estimation (Forward)

Algorithm 1: Ordering estimation using the forward stepwise selection

Input : n i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated ordering, $\widehat{\pi} = (\widehat{\pi}_1, ..., \widehat{\pi}_p)$

Set $\widehat{\pi}_0 = \emptyset$
for $m = \{1, 2, \cdots, p\}$ **do**
  Set $S = \{\widehat{\pi}_0, ..., \widehat{\pi}_{m-1}\}$
  for $j \in \{1, 2, \cdots, p\} \setminus S$ **do**
    Estimate the conditional variance of X_j given X_S, $\widehat{\sigma}^2_{j|S}$
  end
  The m-th element of the ordering $\widehat{\pi}_m = \arg\min_j \widehat{\sigma}^2_{j|S}$
end

Ordering Estimation (Bakward)

Algorithm 2: Ordering estimation using the backward stepwise selection

Input : n i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated ordering, $\widehat{\pi} = (\widehat{\pi}_1, ..., \widehat{\pi}_p)$

Set $S = \{1, 2, \cdots, p\}$
for $m = \{p, p-1, \cdots, 1\}$ **do**
  for $j \in S$ **do**
    Estimate the conditional variance X_j given $X_{S \setminus j}$, $\widehat{\sigma}^2_{j|S\setminus j}$
  end
  The m-th element of the ordering $\widehat{\pi}_m = \arg\max_j \widehat{\sigma}^2_{j|S\setminus j}$
  Update $S = S \setminus \pi_m$
end

Steps

- **Goal:**
  - Identification $\equiv$ Ordering
  - Construct causal graph

1. Element-wise ordering estimation from the initial (π_1) using conditional variances
2. Parent estimation using the conditional independence relationships

Uncertainty Scoring Algorithm

Algorithm 3: Uncertainty Scoring (US) algorithm

Input : n i.i.d. samples from an ANM, $X^{1:n}$
Output: Estimated directed acyclic graph, $\widehat{G} = (V, \widehat{E})$

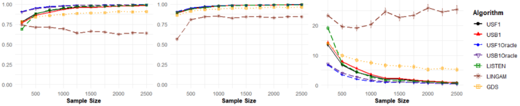
Step (1): Ordering Estimation
Estimate the ordering $\widehat{\pi}$ using Algorithm 1
Step (2): Parents Estimation
for $m = \{2, \cdots, p\}$ **do**
  for $j = \{1, \cdots, m-1\}$ **do**
    Perform a conditional independence test between $\widehat{\pi}_m$ and $\widehat{\pi}_j$ given $\{\widehat{\pi}_1, ..., \widehat{\pi}_{m-1}\} \setminus \widehat{\pi}_j$
    If dependent, include $\widehat{\pi}_j$ into $\widehat{\mathrm{Pa}}(\widehat{\pi}_m)$
  end
end
Estimate the edge set $\widehat{E} := \cup_{m \in \{2,3,\ldots,p\}} \cup_{k \in \widehat{\mathrm{Pa}}(\widehat{\pi}_m)} (k, \widehat{\pi}_m)$

Outline

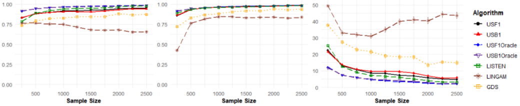
Comparison

- LISTEN: Linear model with heterogenous variance using backward selection
- LINGAM: Non-Gaussian linear model
- GS: Linear model with equal variance
- USF: Uncertainty scoring with forward selection
- USB: Uncertainty scoring with backward selection

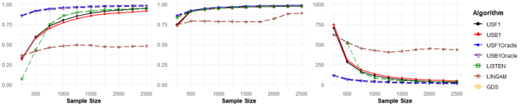
Linear Gaussian with Heterogeneous Variance



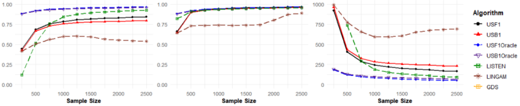
(a) Precision: p=20, d=2 (b) Recall: p=20, d=2 (c) Hamming: p=20, d=2



(d) Precision: p=20, d=4 (e) Recall: p=20, d=4 (f) Hamming: p=20, d=4



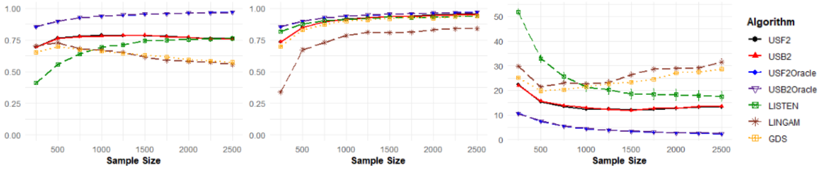
(g) Precision: p=200, d=2 (h) Recall: p=200, d=2 (i) Hamming: p=200, d=2



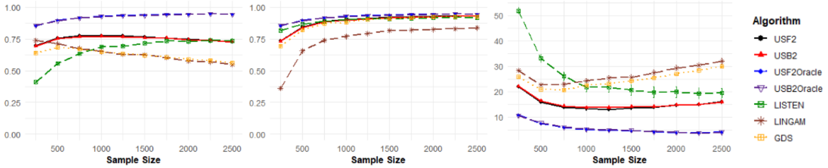
(j) Precision: p=200, d=4 (k) Recall: p=200, d=4 (l) Hamming: p=200, d=4

Gaussian Polynomial with Heterogeneous and Homogeneous Variance

- 20-node



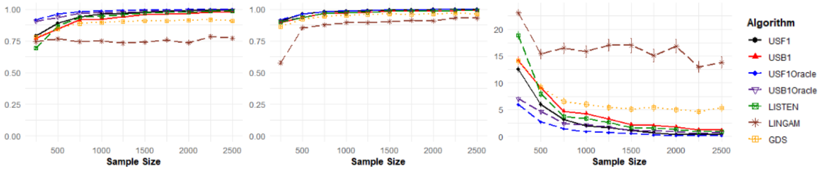
(a) Precision: Homogeneous (b) Recall: Homogeneous (c) Hamming: Homogeneous



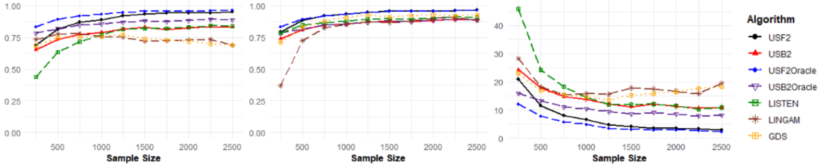
(d) Precision: Heterogeneous (e) Recall: Heterogeneous (f) Hamming: Heterogeneous

ANMs with Gaussian and Non-Gaussian Errors

- Error distributions were sequentially uniform, $U(-1, 1)$, Gaussian $\mathcal{N}(0, 1/3)$, and half of t-distribution with 10 degrees of freedom



(a) Precision: Linear (b) Recall: Linear (c) Hamming: Linear

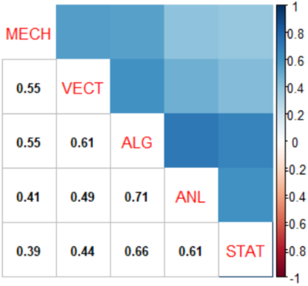


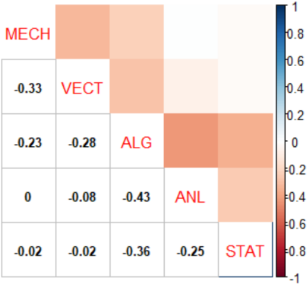
(d) Precision: Quadratic (e) Recall: Quadratic (f) Hamming: Quadratic

Description

- Examination marks for 88 students
- Subjects: mechanics, vectors, algebra, analysis, and statistics.
- Multivariate Gaussian
- Bnlearn R package

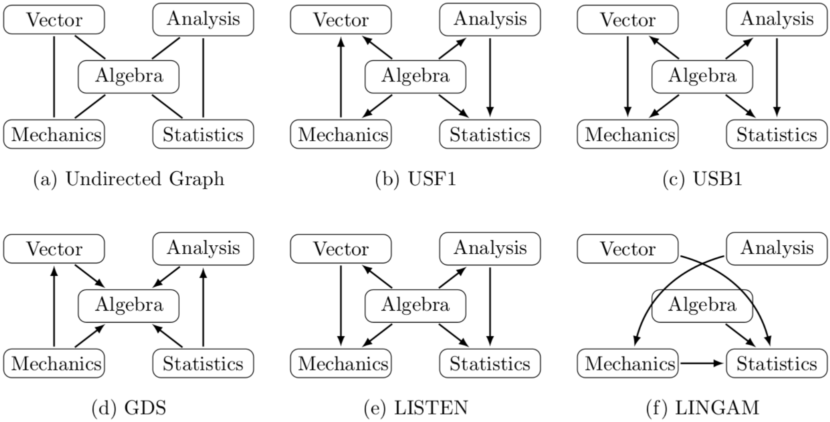
Correlation and Partial Correlation Plots





- The scores for analysis and statistics are conditionally independent of mechanics and vectors, given algebra

Results



(a) Undirected Graph (b) USF1 (c) USB1

(d) GDS (e) LISTEN (f) LINGAM

- Knowledge of analysis and vectors is prerequisite for statistics and mechanics, respectively

Outline

Conclusion

- Conditional variance is a valid identifiability condition
- It helps to estimate true ordering of a graph
- Uncertainty scoring algorithm: Estimating orders & Finding parents
- Works better on less dense data

References

[1] Clark Glymour, Kun Zhang, and Peter L. Spirtes. "Review of Causal Discovery Methods Based on Graphical Models". In: *Frontiers in Genetics* 10 (2019).

[2] Benjamin Kap. *The Effect of Noise Level on Causal Identification with Additive Noise Models*. Aug. 2021.

[3] Gunwoong Park. "Identifiability of Additive Noise Models Using Conditional Variances". In: *Journal of Machine Learning Research* 21.75 (2020), pp. 1–34. URL: http://jmlr.org/papers/v21/19-664.html.

[4] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN: 0262037319.

Thank You.