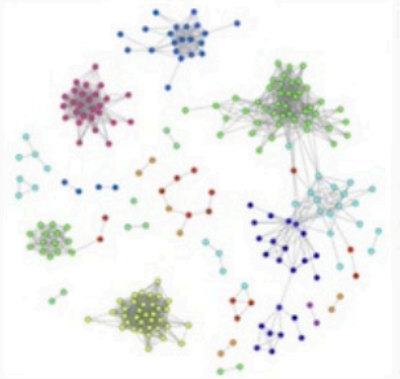
Learning Sparse Nonparametric DAGs

Dongbang and Reza

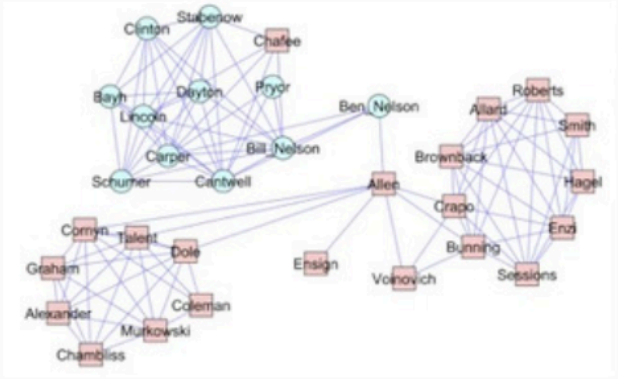
Graphical Model Class Project

Dr. Ni

Discovering Relationship Between Variables

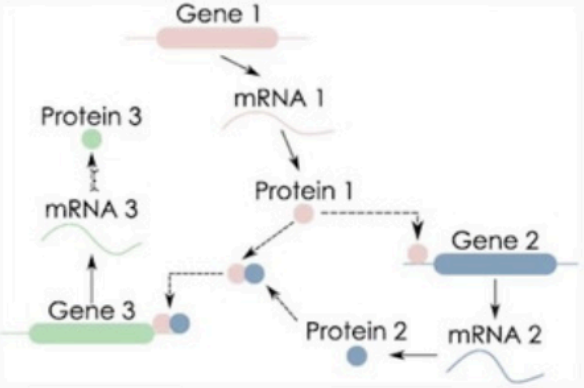


S&P 500 stock networks



Political voting patterns

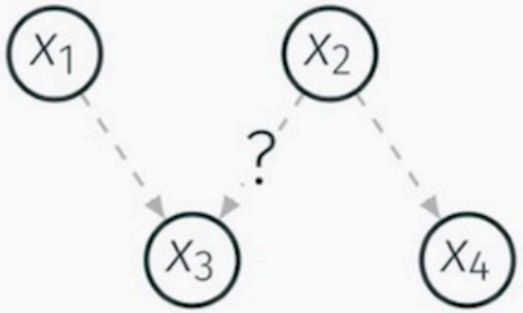
Brain connectivity



Gene regulatory networks

Bayesian Network Structure Learning

Which DAG fits the data best?



estimate ←

$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ 4.00 & -1.14 & 0.20 & -2.37 \\ -1.05 & 0.35 & -0.66 & -0.39 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Bayesian Networks and Structural Equation Models

Bayesian network (BN) G with d nodes:

$$P(\mathbf{x}; G) = \prod_{j=1}^{d} P(x_j | \mathbf{x}_{pa(j)})$$

Equivalently expressed as structural equation model (SEM):

$$x_j \sim P(x_j | \mathbf{x}_{pa(j)}) \iff \begin{aligned} & x_j = f_j(\mathbf{x}, \varepsilon_j) \\ & f_j \text{ only depends on } \mathbf{x}_{pa(j)} \\ & \varepsilon_j \sim \text{noise distribution} \\ & \varepsilon_1, \ldots, \varepsilon_d \text{ independent} \end{aligned}$$

Score-based Learning: General BN

M-estimation for general BN

Given $n \times d$ observations X, solve

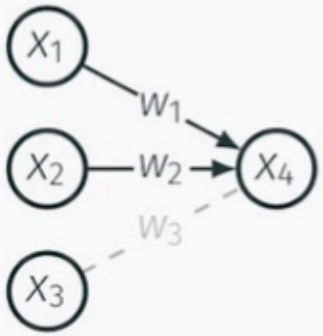
$$\min_{f=(f_1,\ldots,f_d)} \ell(f; X)$$
$$s.t. \quad G(f) \in DAG$$

Dependency induced by f forms a DAG

Simple Case: Linear Additive Noise Model

Linear parameterization of P:

$$X_j = f_j(\mathbf{X}, \varepsilon_j) := \underbrace{\mathbf{X}^T \mathbf{w}_j}_{\text{linear}} + \underbrace{\varepsilon_j}_{\text{zero mean}}$$



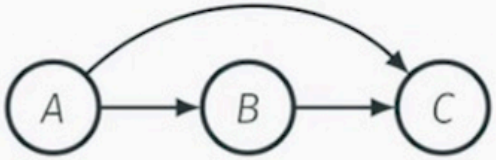
Smooth Optimization (Zheng et al., 2018)

$$\min_W \ \ell(W;X) \quad \text{s.t.} \quad G(W) \in DAG \qquad \Longleftrightarrow \qquad \min_W \ \ell(W;X) \quad \text{s.t.} \quad h(W) = 0$$

(combinatorial 😱) (smooth 😎)

$h(W) := tr\exp(W \circ W) - d.$

Linearity Can Hurt

Ground truth	Linear model estimate	Why?
		$corr(A, B) = 0$
		$corr(A, C\|B) \neq 0$

Nonparametric Additive Noise Model (ANM)

$$x_j = f_j(\mathbf{x}) + \varepsilon_j, \quad \textit{where} \quad \begin{cases} f_j \text{ only depends on } \mathbf{x}_{pa(j)} \\ \mathbb{E}[f_j(\mathbf{x})] = 0 \\ \varepsilon_j \sim \text{noise distribution} \\ \varepsilon_1, \ldots, \varepsilon_d \text{ independent} \end{cases}$$

M-estimation for Nonparametric ANM

Given $n \times d$ observations X, solve

$$\min_{f=(f_1,\ldots,f_d)} \ell(f; X)$$
$$\text{s.t.} \quad G(f) \in DAG$$

How to express acyclic dependency?

Nonparametric Sparsity $\rightarrow$ Nonparametric Acyclicity

Zero partial derivative means input-output is independent:

$$\left\| \frac{\partial f}{\partial x_j} \right\|_{L^2} = \sqrt{\int_{\mathcal{X}} \left(\frac{\partial f(\mathbf{x})}{\partial x_j} \right)^2 d\mathbf{x}} = 0 \iff x_j \perp\!\!\!\perp f(\mathbf{x})$$

Construct a proxy $W(\mathbf{f}) = W(f_1, \ldots, f_d) \in \mathbb{R}^{d \times d}$:

$$[W(\mathbf{f})]_{kj} := \left\| \frac{\partial f_j}{\partial x_k} \right\|_{L^2}$$

Plug in $h(W)$:

$$h(W(\mathbf{f})) = tr(e^{W(\mathbf{f}) \circ W(\mathbf{f})}) - d = 0 \iff G(\mathbf{f}) \in DAG$$

What approximation class $\mathcal{F}$ to choose?

MLPs as $\mathcal{F}$

Multilayer perceptrons (MLP) with h layers:

$$f_j(x) = MLP(x; A_j^{(1)}, \ldots, A_j^{(h)}) = \sigma(A_j^{(h)}\sigma(\cdots A_j^{(2)}\sigma(A_j^{(1)}x)))$$

Each $[W(f)]^2_{kj}$ can be obtained by chain rule:

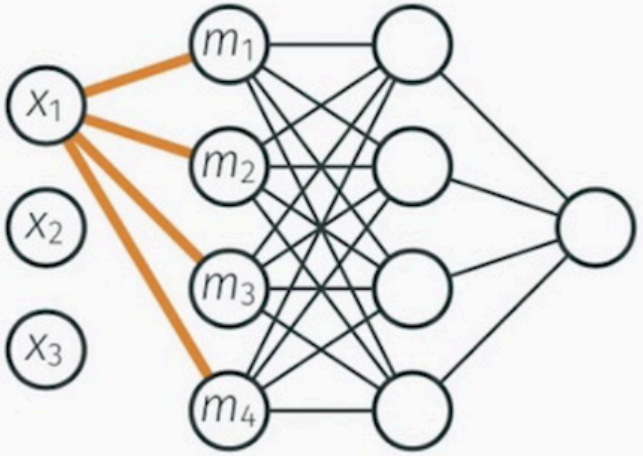
$$[W(f)]^2_{kj} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial f_j}{\partial x_k}(x^{(i)})\right)^2$$

In fact, can be done independent of depth!

MLPs as $\mathcal{F}$: Focus on First Layer

If the first layer is disconnected, $x_1 \perp\!\!\!\perp f(\mathbf{x})$.

$$A^{(1)} = \begin{pmatrix} 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix}$$



Proposition (Only need to constrain $A^{(1)}$)

Let

$\mathcal{F} = \{f \mid f(\mathbf{x}) = MLP(\mathbf{x}; A^{(1)}, \ldots, A^{(h)}), x_k \perp\!\!\!\perp f(\mathbf{x})\}$ *and*

$\mathcal{F}_0 = \{f \mid f(\mathbf{x}) = MLP(\mathbf{x}; A^{(1)}, \ldots, A^{(h)}), kth\text{-}column(A^{(1)}) = 0\}$.

Then $\mathcal{F} = \mathcal{F}_0$.

MLPs as $\mathcal{F}$

Let $\mathbf{A} = \{A_j^{(l)} : j = 1, \ldots, d, l = 1, \ldots, h\}$.

$$[W(f)]_{kj}^2 := [W(\mathbf{A}^{(1)})]_{kj}^2 = \|k\text{th-column}(A_j^{(1)})\|_2$$

Continuous M-estimation for BN with MLP

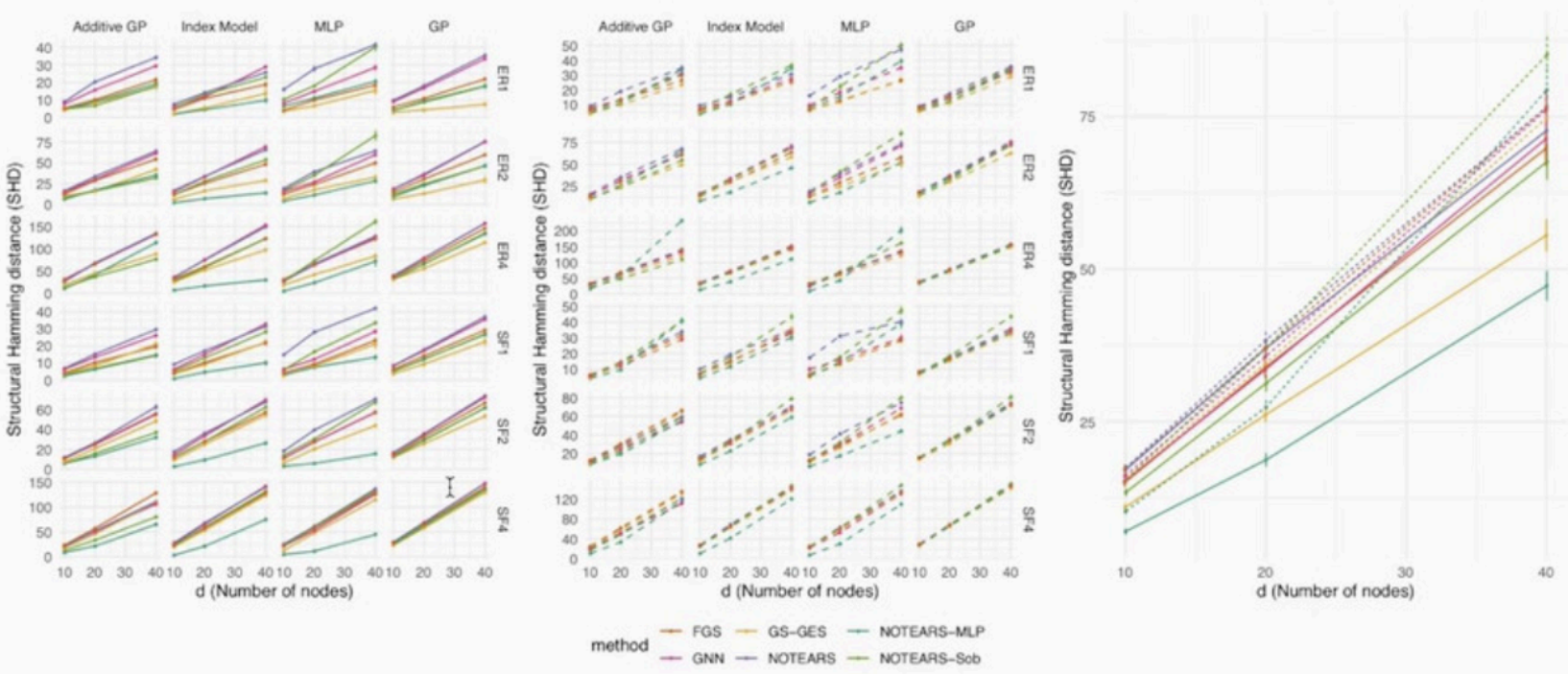
Given $n \times d$ observations X, solve

$$\min_{\mathbf{A}} \ell(\mathbf{A}; X)$$
$$s.t.\ h(W(\mathbf{A}^{(1)})) = 0$$

Experiments

- Ground truth: ER, SF
- Samples $n = 1000, 200$
- Variables $d = 10, 20, 40$
- SEM: Additive GP, Index model, MLP, GP
- Methods: NOTEARS-MLP, NOTEARS-Sobolev
- Baselines:
  - NOTEARS (Zheng et al., 2018)
  - FGS (Ramsey et al., 2017)
  - GS-GES (Huang et al., 2018)
  - DAG-GNN (Yu et al., 2019)

Structure Recovery: SHD (lower is better)



Left: $n = 1000$. Middle: $n = 200$. Right: Average all configurations.

Summary

- Score-based learning of sparse nonparametric DAGs
- Nonparametric sparsity $\rightarrow$ nonparametric acyclicity
- Continuous (nonconvex) optimization