Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families

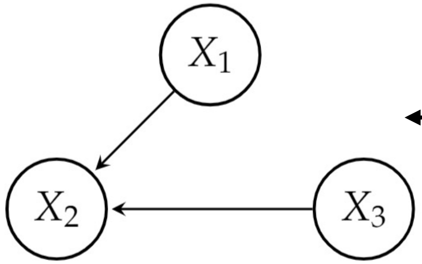
Goutham Rajendran, Bohdan Kivva, Ming Gao, Bryon Aragam

THE UNIVERSITY OF CHICAGO

NeurIPS 2021

Background

- Graphical models: Graphs that compactly represent probability distributions



$$\longleftrightarrow \quad \begin{array}{ccc} X_1 & X_2 & X_3 \end{array}$$

$$\begin{bmatrix} 0.89 & -3.17 & 1.65 \\ -1.34 & -4.59 & 0.64 \\ 2.11 & 5.05 & -0.28 \\ -0.23 & 0.56 & -0.34 \\ 0.29 & -4.40 & 1.66 \\ & \ldots & \end{bmatrix}$$

- The structure learning problem: Given data, find the best fit graph
- Applications: Machine learning, genetics, medicine, physics, etc.

Approaches

We will focus on learning directed graphs, also known as Bayesian networks.

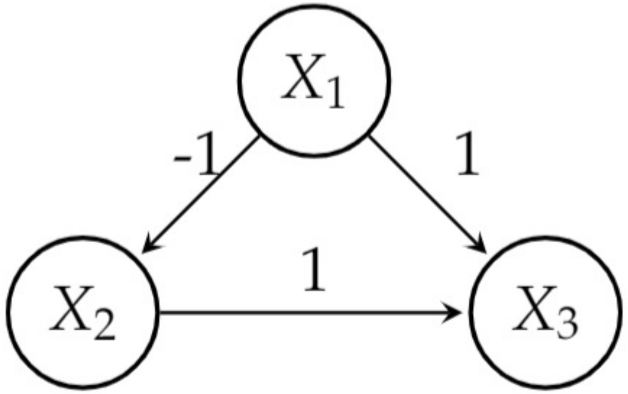
- Constraint based: Based on independence tests.
  - Example - PC algorithm
- Score based: Define a score and optimize it over all graphs.
  - Example - GES algorithm
- We will focus on score-based methods in this talk.

Example

- Let N_1, N_2, N_3 be i.i.d. standard Gaussians N(0, 1)

$$X_1 = N_1$$
$$X_2 = -X_1 + N_2 = N_2 - N_1$$
$$X_3 = X_1 + X_2 + N_3 = N_2 + N_3$$

$\longleftrightarrow$



- This special case is also known as a structural equation model.

Score based learning

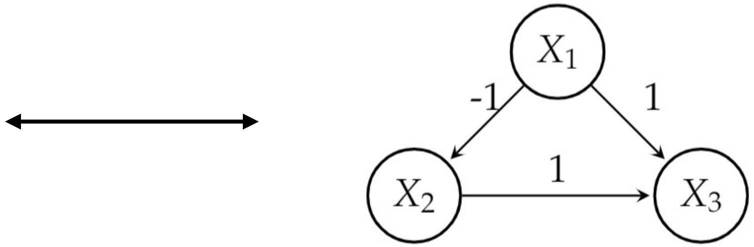
- Problem: Given a random vector $X = (X_1,.., X_d)$, want to learn a directed acyclic graph (DAG) W for X.
- Score function S: A function that maps DAG W to a number.
  - Think of the score as a measure of fit.
- Score based approach is to solve

$$\min_{W \in \mathsf{DAG}} S(W).$$

Example: Least-squares score

- Recall

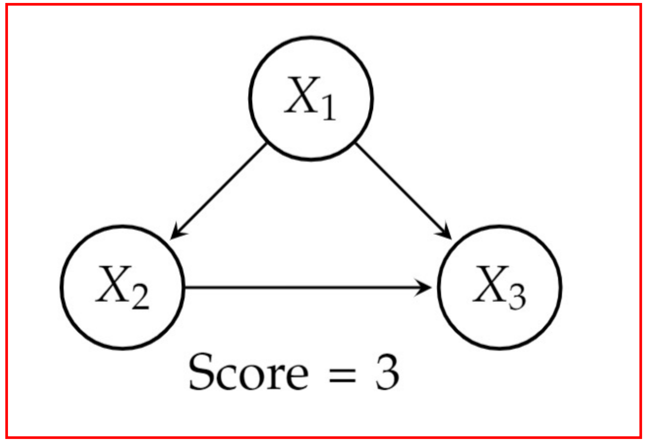
$$X_1 = N_1$$
$$X_2 = -X_1 + N_2 = N_2 - N_1$$
$$X_3 = X_1 + X_2 + N_3 = N_2 + N_3$$

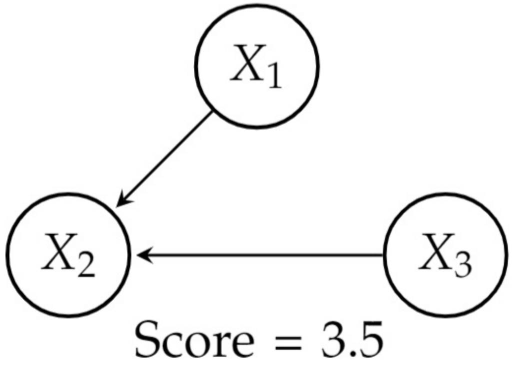


- Least-squares score:

$$S(W) = \sum_{i \le n} (X_i - \sum_{j \in pa(i)} W_{ji} X_j)^2$$

- Find best model fit





Prior works on structure learning

- Exactly solving the minimization problem is NP-hard.
- Approaches include greedy algorithms such as GES, or inefficient dynamic programming algorithms.
- Intriguing recent works: Under some conditions, simple greedy algorithms (not score-based) output the true model.
- This work: A general score-based algorithm that subsumes and generalizes many of these works.

Setting

- Given dataset which are samples of random vector $X = (X_1,\ldots, X_d)$
- Assume we have a decomposable score

$$S(W) = \sum_{i \leq d} S_i(W^{(i)})$$

- Additional notation: W^{-e} zeroes out edge e; $W[T \rightarrow i]$ sets parents of vertex i to be T.

Greedy forward-backward Search (GFBS)

Algorithm 1: Greedy Forward-Backward Search

Input: Dataset X, tolerance parameter $\gamma \geq 0$

Output: DAG W

$W = \emptyset$ `// n-vertex graph with no edges`

$T = []$ `// The ordering`

`// Forward phase`

for *iter* $= 1$ *to* d **do**

    $i = \arg\min_{i \notin T} S_i(e_T)$ `// Minimize jump in score`

    $W = W[T \to i]$

    $T.append(i)$

`// Backward phase`

for *edge* e *in* W **do**

    if $S(W^{-e}) - S(W) \leq \gamma$ **then**

        $W = W^{-e}$ `// Delete the edge e`

return W `// Guaranteed to be a DAG`

A summary of highlights

- Output is always a DAG
- Running time: Polynomial in d and time to evaluate score.
- Statistical guarantees: Under some assumptions, GFBS always outputs the true DAG (generalizes several prior works)
- Sample complexity guarantees
- Different from GES because GES is edge-greedy whereas GFBS is vertex-greedy.

The Bregman-score – A generalization of least-squares

- Let ϕ be strictly convex and differentiable.
- Define Bregman-divergence $d_\phi(x, y) = \phi(x) - \phi(y) - (x - y)\phi'(y)$
  - Generalizes Euclidean distance, Logistic loss, KL-divergence, etc.
- Define Bregman-information of a distribution $I_\phi(D) = \mathbb{E}_{x \sim \mathcal{D}}[d_\phi(x, \mu)]$
- Define the Bregman-score as

$$S_\phi(W) = \sum_{i \leq d} \mathbb{E}[I_\phi(X_i | \, \mathrm{pa}_W(i))]$$

  - Generalizes the least-squares score (the special case $\phi(x) = x^2$).
  - For exponential family models, this is the expected negative log-likelihood.

Assumptions for the statistical guarantee

- Assumption 1: For any vertex i, if Y is a set of non-descendants,

$$\mathbb{E}[I_\phi(X_i|Y)] > \mathbb{E}[I_\phi(X_i|\,\mathrm{pa}(i))]$$

  - Informally, expected Bregman-information drops as more parents are conditioned.
  - Similar to causal minimality
- Assumption 2: There is a constant $\tau > 0$ such that for all vertices i,

$$\mathbb{E}[I_\phi(X_i|\,\mathrm{pa}(i))] = \tau$$

  - Informally, if all parents have been conditioned upon, then expected Bregman-information is the same across all vertices.
  - Generalizes the equal variance assumption from prior works

The main statistical guarantee

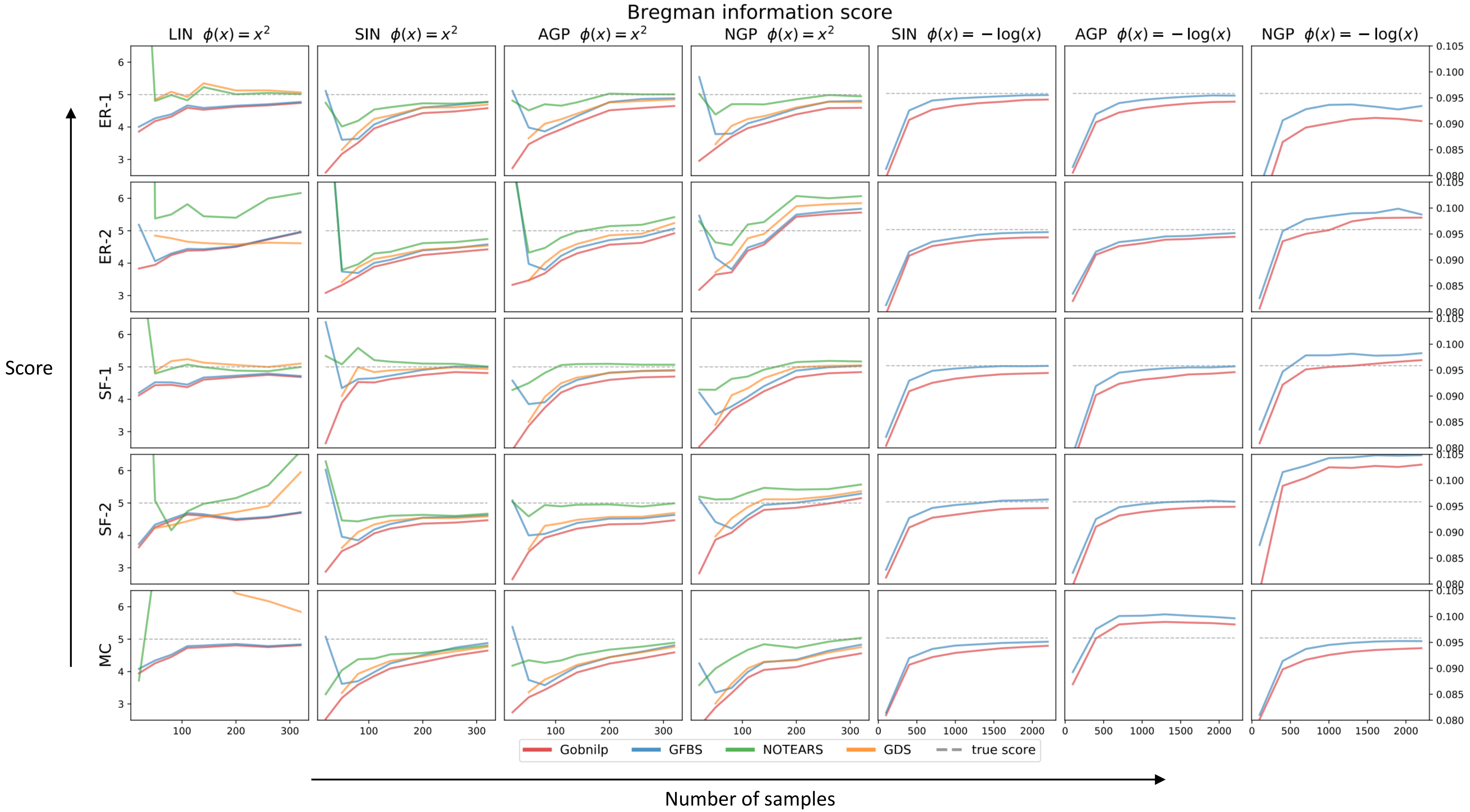
- Main Theorem: Under the above assumptions, GFBS returns the true model.
- Corollary (Identifiability): Under the above assumptions, the model is identifiable.
- This generalizes and subsumes prior works
- Also suggests the Itakuro-Saito score for multiplicative structural equation models

Experimental setup

- Bregman-score:
  - $\phi(x) = x^2$
  - $\phi(x) = -\log(x)$
- Graphs
  - Markov Chains
  - Erdős-Rényi graphs
  - Scale-Free graphs
- Model: $X_i = f\big(pa(i)\big) + Z_i$ where
  - f is linear (LIN), sine (SIN) or additive/non-additive Gaussian process (AGP/NGP)
  - Z_i is the t-distribution with unit variance or uniform [1, 2].
- Algorithms:
  - GFBS
  - GOBNILP (optimum score)
  - NOTEARS
  - GDS

Experiments on optimizing score

Grey line – True optimal score



Future directions

- Under what conditions will GFBS globally optimize the score?
- Can we compare GFBS and GES?
- Can we formally compare Assumption 1 to causal minimality?
- In the finite sample case, GFBS returns a non-optimal score, can we somehow regularize the backward phase?

Thank you