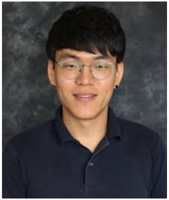
Towards Fast Mixing MCMC for Structure Learning

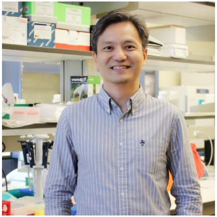
Presenter: Quan Zhou

Department of Statistics, Texas A&M University

Acknowledgment



Hyunwoong (Woody) Chang
PhD student
Department of Statistics



James Cai
Professor, Department of
Veterinary Integrative Biosciences

The research presented in this talk is supported by NSF DMS-2245591.

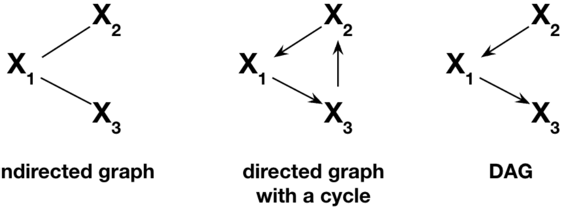
Outline of the talk

- Introduction
  - DAGs and Markov equivalence classes
  - Structure learning on three search spaces
- Rapid mixing of an equivalence class MCMC sampler
  - Construction of RW-GES
  - Rapid mixing of RW-GES
- MCMC sampling without score equivalence
  - Structure learning with equal error variance
  - Theoretical and practical advantages
  - Simulation studies and an example of single-cell data analysis

DAG models

DAG model

A p-node DAG model is a directed acyclic graph whose nodes are random variables $X_1, \ldots, X_p$. It encodes the conditional independence (CI) relations in the joint distribution of $(X_1, \ldots, X_p)$.



We only consider linear Gaussian DAG models in this talk.

Ordering of nodes

Ordering

Each DAG is consistent with at least one ordering: if i precedes j, then the edge between $\mathsf{X}_i, \mathsf{X}_j$ is directed as $\mathsf{X}_i \to \mathsf{X}_j$.

For the DAG $X_2 \to X_1 \leftarrow X_3$, the ordering can be $(2, 3, 1)$ or $(3, 2, 1)$.

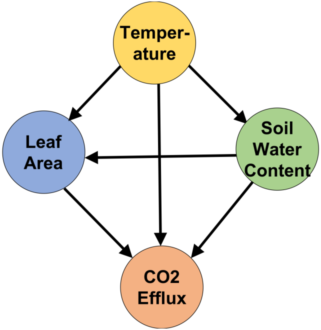
For linear Gaussian DAG models with ordering $(1, 2, \dots, p)$, we can write

$$X_j = \beta_{1j} X_1 + \cdots \beta_{(j-1)j} X_{j-1} + \epsilon_j, \qquad \text{for each } j,$$

where $\epsilon_1, \dots, \epsilon_p$ are ind. normal random variables.

Examples

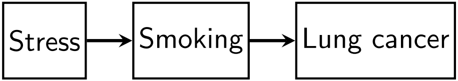
A hypothetical DAG model for soil respiration



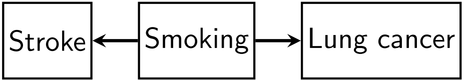
From my collaborator, Xuejun Dong, at Texas A&M University.

Ordering = (Temperature, Soil Water Content, Leaf Area, CO_2 Efflux).

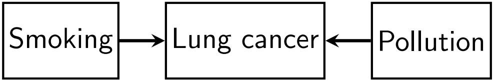
Examples for $p = 3$



Stress $\perp\!\!\!\perp$ Lung cancer | Smoking

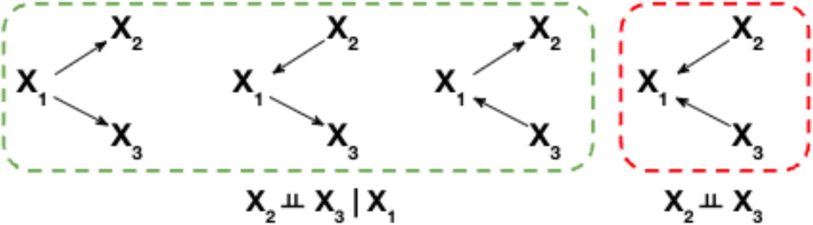


Stroke $\perp\!\!\!\perp$ Lung cancer | Smoking



Smoking $\perp\!\!\!\perp$ Pollution

Examples for $p = 3$


X_1
X_2
X_3
$X_2 \perp\!\!\!\perp X_3 \mid X_1$
$X_2 \perp\!\!\!\perp X_3$

So $X_2 \to X_1 \leftarrow X_3$ encodes one CI relation: $X_2 \perp\!\!\!\perp X_3$. This is called a "v-structure".

The other three DAGs all encode the CI relation $X_2 \perp\!\!\!\perp X_3 \mid X_1$; we say they are Markov equivalent.

Markov equivalence class

Markov equivalence class (MEC)

Two DAGs are Markov equivalent and belong to the same MEC if they encode the same set of CI relations.

Lemma

Two DAGs are Markov equivalent if and only if they have the same skeleton and v-structures.

For example, $X_1 \to X_2$ and $X_1 \leftarrow X_2$ are also Markov equivalent.

Given only observational data and no prior knowledge, Markov equivalent linear Gaussian DAG models are not distinguishable.

Score-based structure learning

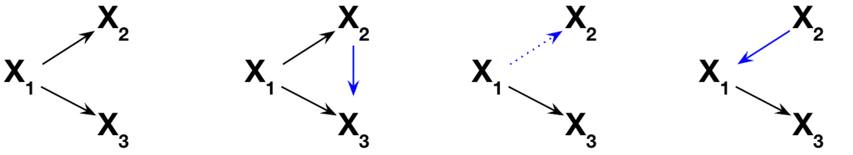
Structure learning

Learn the underlying DAG of a p-variate probability distribution from n i.i.d. observations.

Suppose we have a function ψ (called "score") such that a larger value of $\psi(G)$ indicates that the DAG G is more likely. We can run a greedy local search to find what DAG has the largest score.

Examples of local moves

Typical local operators for modifying a DAG



current DAG

edge addition

edge deletion

edge reversal

Consistency of the score and search algorithm

Local consistency of ψ

We say ψ is locally consistent if for any distinct DAGs G, G' that satisfy

$$G' = G \cup \{X_i \to X_j\},$$

we have
(i) $\psi(G) > \psi(G')$ if $X_i \perp\!\!\!\perp X_j \mid \mathrm{Pa}_j(G)$, and
(ii) $\psi(G') > \psi(G)$ if $X_i \not\perp\!\!\!\perp X_j \mid \mathrm{Pa}_j(G)$,
where $\mathrm{Pa}_j(G)$ denotes the parent set of node X_j in G.

If p is fixed and $n \to \infty$, we expect ψ will become locally consistent. Then will a local search algorithm always return the true DAG (regardless of the initial state)?

Three search spaces

Let $\mathcal{G}_p$ be the space of all p-node DAGs. In addition to $\mathcal{G}_p$, one can also perform local search on

- $\mathcal{E}_p$: the space of all p-node MECs;
- $\mathbb{S}_p$: symmetric group on $\{1, 2, \ldots, p\}$, i.e., the space of all orderings.

Directly searching $\mathcal{E}_p$ bypasses the need of traversing MECs, but the implementation of local moves on $\mathcal{E}_p$ can be complicated.

$\mathbb{S}_p$ is sometimes desirable since given the ordering, we can identify the parent set for each node separately by variable selection.

Bayesian structure learning

A standard Bayesian method is to use the prior of Geiger and Heckerman [6], calculate a posterior on $\mathcal{G}_p$ and define the score ψ to be the log-posterior. This approach satisfies the following.

- *Score equivalence:* $\psi(G_1) = \psi(G_2)$ if G_1, G_2 are Markov equivalent.
- *Modularity/decomposable score:* We can write

$$\psi(G) = \sum_{j=1}^{p} \psi_j(\mathsf{X}_j, \mathrm{Pa}_j(G))$$

for some functions $\psi_1, \ldots, \psi_p$ (dependency on the data is omitted).

Metropolis-Hastings algorithm

It is often straightforward to transform a greedy local search algorithm to a local Metropolis-Hastings (MH) algorithm.

In each iteration, given the current DAG G,

1. propose a local move from G to some G',
2. accept the proposal with probability

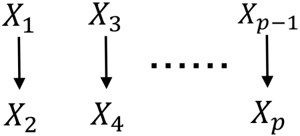
$$\alpha(G, G') = \min\left\{1, \frac{e^{\phi(G')} q(G \mid G')}{e^{\phi(G)} q(G' \mid G)}\right\},$$

where $q(G' \mid G)$ denotes the probability of proposing G' at G.

An example is the structure MCMC [13], which uses single-edge addition, deletion and reversal as the proposal; more sophisticated versions have also been developed [8, 9, 19].

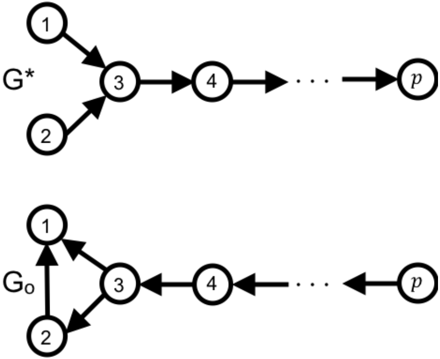
Challenges of MCMC sampling

- $|\mathcal{G}_p|$ is enormous and grows *super-exponentially* in p [18], e.g. $|\mathcal{G}_{10}| \approx 4 \times 10^{18}$.
- Traversing large MECs can be very difficult.



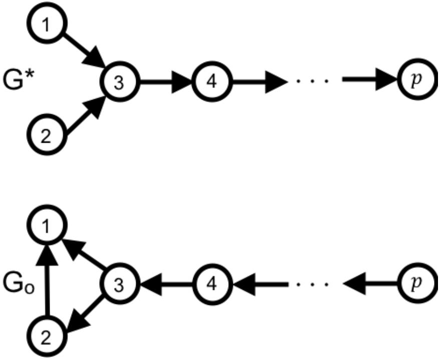
The MEC of this DAG (which is sparse) has $2^{p/2}$ member DAGs.

Traversing MECs can be difficult



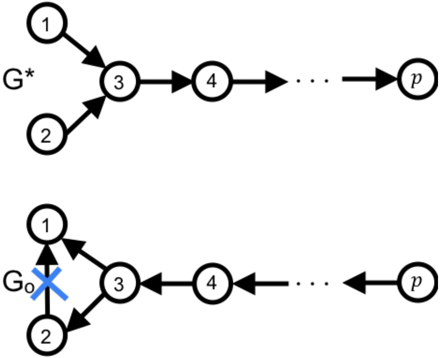
Suppose G^* is the true DAG, and n is sufficiently large so that all CI relations can be correctly inferred. Can the structure MCMC sampler quickly move from G_0 to G^*?

Traversing MECs can be difficult



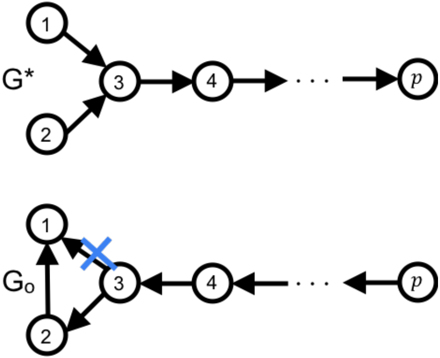
We only need to remove the edge $2 \to 1$ and reverse all the other edges.

Traversing MECs can be difficult



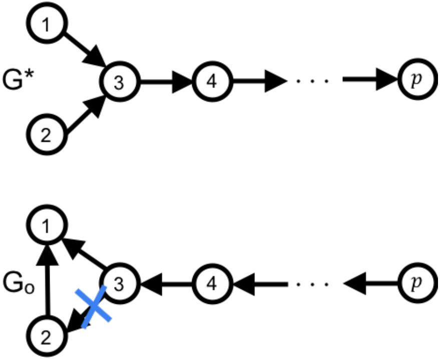
Cannot remove $2 \to 1$ since $2 \not\perp\!\!\!\perp 1 \mid 3$.

Traversing MECs can be difficult



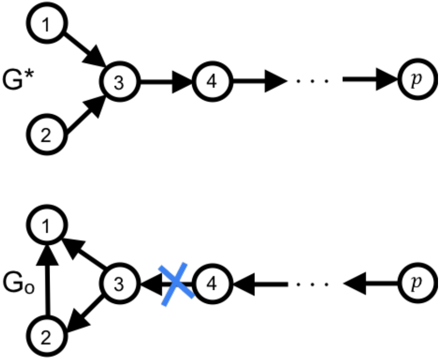
Cannot reverse $3 \to 1$ since that would result in a cycle.

Traversing MECs can be difficult



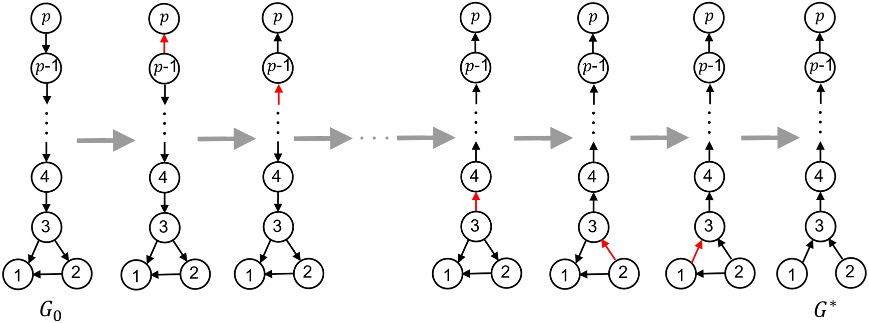
Cannot reverse $3 \to 2$ since $2 \perp\!\!\!\perp 4 \mid 3$.

Traversing MECs can be difficult



Cannot reverse $4 \to 3$ since $3 \perp\!\!\!\perp 5 \mid 4$.

Traversing MECs can be difficult



Have to first reverse $p \to p-1$, then $p-1 \to p-2$, and so on. (All these edge reversals result in Markov equivalent DAGs.)

Traversing MECs can be difficult

Can we introduce a new type of proposal that allows us to jump from one DAG to another random DAG in the same MEC?

Answer: Very difficult in practice, since counting or enumerating an MEC is highly time-consuming. The counting algorithm of Ghassami et al. [7] has complexity $O(p^{d+2})$, where d is the graph degree.

Possible solution 1: We can directly construct a local MH algorithm on $\mathcal{E}_p$, the space of MECs.

Possible solution 2: Choose some score that distinguishes between Markov equivalent DAGs.

Questions to be addressed

- In high-dimensional settings, do we have any theoretical guarantee for the complexity of MCMC algorithms (or greedy local search algorithms) for structure learning?
- If traversing MECs causes slow mixing, can we sacrifice score equivalence for faster mixing?
- How important is the prior knowledge to the mixing of MCMC algorithms?

Constructing a rapidly mixing MEC sampler

Our goal is to construct an MH sampler on $\mathcal{E}_p$ with rapid mixing guarantee under some high-dimensional assumptions (both $n, p \to \infty$).

Rapid mixing

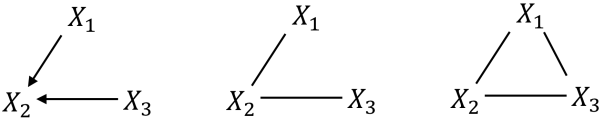
An MCMC algorithm is rapidly mixing if its mixing time grows polynomially with n and p.

Existing MEC samplers

Existing samplers on $\mathcal{E}_p$ use CPDAG operators to propose local moves [14, 16, 10, 2]. They can be slowly mixing when $n \to \infty$ and p is fixed.

CPDAG

Each MEC can be uniquely represented by a CPDAG (completed partially directed acyclic graph), also called essential graph.



All the 3 graphs are CPDAGs. How to move from the 3rd to the 1st?

How to define the neighborhood?

Challenges:

- For MCMC samplers based on CPDAG operators, the “neighborhood” of each MEC is too small, giving rise to local modes. (Neighborhood: the set of MECs that can be reached by one proposal.)
- But for rapid mixing to be possible, the neighborhood size needs to be polynomial in p.

Constructing the search space and neighborhood

We say a DAG G is sparse if its in-degree is bounded by d_{in} and out-degree is bounded by d_{out}.

Search space of our algorithm

The set of all MECs that contain at least one sparse member DAG.

Neighborhood of our algorithm

An MEC $\mathcal{E}'$ is a neighbor of MEC $\mathcal{E}$ if there exist sparse $G' \in \mathcal{E}'$ and sparse $G \in \mathcal{E}$ such that G' can be obtained from G by adding, deleting or “swapping” an edge.

”Swap” means to delete an edge $j \to i$ and add $k \to i$.

Constructing the search space and neighborhood

- The choice of the neighborhood is very similar to that of GES (greedy equivalence search), a classical structure learning algorithm with consistency guarantee in low-dimensional settings; see Chickering [5]. (GES doesn't use swap moves.)
- This neighborhood is much larger than those used in existing MEC samplers.
- If $d_{\text{in}} + d_{\text{out}} = O(\log p)$, the neighborhood size is *polynomial* in p; see Lemma 1 of our paper [21].
- Efficient implementation of the proposal can be done by using the operators introduced in Chickering [5].

Rapid mixing of RW-GES sampler

We define ψ (log-posterior) using an empirical Bayes model (extending a DAG selection model of [12]) which assigns same score to Markov equivalent DAGs.

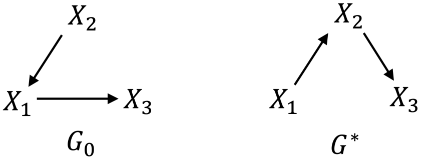
Theorem 6 of Zhou and Chang [21]

Under some high-dimensional assumptions, our MCMC sampler RW-GES (random walk GES sampler) is *rapidly mixing* with high probability.

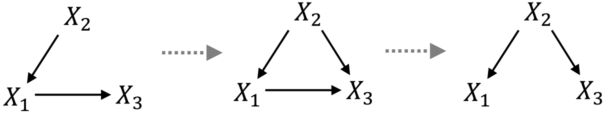
This result is obtained by first proving the consistency of the greedy local search. Challenge: The low-dimensional consistency result of GES cannot be extended to the high-dimensional case due to node degree constraints.

Example

Assume all CI relations can be inferred correctly. How to move from the MEC of G_0 to the MEC of the true DAG G^*?



Since $X_1 \not\perp\!\!\!\perp X_2$ and $X_1 \not\perp\!\!\!\perp X_3$, in GES we have to add an edge first.



If one imposes $d_{\text{in}} = 1$ or $d_{\text{out}} = 1$, this path is not allowed.

Consistency of greedy local search

Solution: introduce *swap* proposals, require the “true maximum degree” $d^* = O(\sqrt{\log p})$ and use $d_{\text{in}} = O(\sqrt{\log p}), d_{\text{out}} = O(\log p)$.

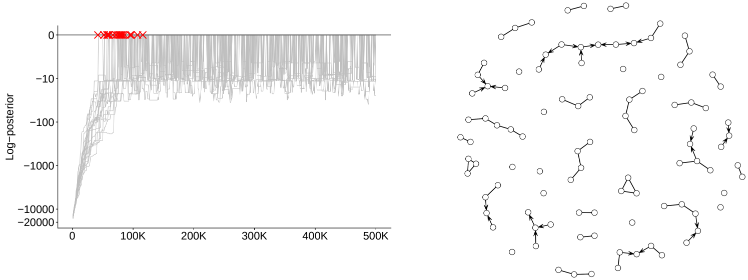
We define d^* as the maximum degree of minimal I-maps of the true DAG.

We showed that a greedy local search returns the true MEC within $(3d^* + 2d_{\text{in}})p$ steps (see Theorem 3 in our paper).

Remarks

- Please see my other slides [link] for MCMC theory and methodology for general high-dimensional model selection problems.
- Discussion on the ARGES algorithm of Nandy et al. [15].
- Open problems: rapid mixing on the DAG or order space. (Caveat!)
- One assumption (permutation β-min condition) required to obtain the selection consistency or rapid mixing is restrictive [20]. In reality, the posterior distribution is often highly multimodal.
- The theory does yield useful insights (e.g. choice of hyperparameters, orders of growth of n, p and model sparsity).

A numerical example



Left: trajectories of 20 RW-GES runs on a simulated data set with $n = 800, p = 100$; red crosses mark the first time the true MEC is sampled.

Right: CPDAG of the true model used to simulate the data.

Equal error variance assumption

With only observational data, the true DAG model may be identifiable under additional assumptions, e.g. equal error variance [17].

Example: for $p = 3$ and ordering $(1, 2, 3)$, equal error variance means that we can express the joint distribution of $(\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ by

$$\begin{aligned} \mathsf{X}_1 &= \epsilon_1, \\ \mathsf{X}_2 &= \beta_{12}\mathsf{X}_1 + \epsilon 2, \\ \mathsf{X}_3 &= \beta_{13}\mathsf{X}_1 + \beta_{23}\mathsf{X}_2 + \epsilon 3, \end{aligned}$$

where $\epsilon_1, \epsilon_2, \epsilon_3 \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for some $\sigma^2 > 0$.

This essentially means that the error variances are known up to a constant multiplicative factor.

Why the equal variance assumption helps

Example (Gaussian DAG with $p = 2$)

Consider $X = (X_1, X_2)$ generated by the structure equation model

$$\begin{aligned} X_1 &= \epsilon_1, & \epsilon_1 &\sim N(0, \sigma^2), \\ X_2 &= \beta X_1 + \epsilon_2, & \epsilon_2 &\sim N(0, \sigma^2), \end{aligned}$$

where ϵ_1, ϵ_2 are independent; this corresponds to the DAG $X_1 \to X_2$.

If $\beta \neq 0$,

$$(\beta^2 + 1)\sigma^2 = \mathrm{Var}(X_2) > \mathrm{Var}(X_1) = \sigma^2.$$

If sample size is large, we should be able to tell whether $\mathrm{X}_1 \to \mathrm{X}_2$ or $\mathrm{X}_1 \leftarrow \mathrm{X}_2$ is the true model.

Non-decomposable posterior score

We build an empirical Bayes model and derive the score of a DAG G under the equal error variance assumption:

$$\psi_{\text{eev}}(G) = -|G|(c_1 + c_0 \log p) - \frac{\alpha p n + \kappa}{2} \log \left(\sum_{j=1}^{p} \hat{\omega}_j(G) \right).$$

- $|G|$ denotes the number of edges in G.
- c_0, c_1, α, κ are hyperparameters.
- $\hat{\omega}_j(G)$ is the maximum likelihood estimate of the error variance of node j given parent nodes in G.

ψ_{eev} is *non-decomposable* and this procedure is *not score-equivalent*.

Why we want to use this in practice

- We proved the high-dimensional selection consistency under a condition on the true model that is slightly weaker than the equal error variance assumption [3].
- MCMC algorithms targeting a score-equivalent posterior usually converge very slowly in practice due to the existence of large MECs.
- The posterior distribution derived from equal error variance, $e^{\psi_{\text{eev}}}$, is more concentrated and thus easier to sample from. A theoretical argument is given in our paper [3].

Hence, even if we have no knowledge about the error variances, using ψ_{eev} can be beneficial.

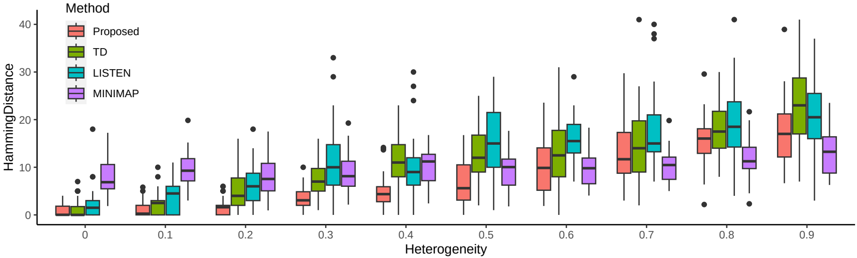
Order MCMC

We build an order MCMC sampler targeting the posterior $e^{\psi_{\mathrm{eev}}}$.

1. Similarly to minimal I-MAP MCMC [1], we approximate the posterior probability of each ordering using a single best DAG.
2. We develop an iterative generalization of the top-down algorithm of Chen et al. [4], which can be used to generate a warm start for the order MCMC sampler.
3. We use adjacent transpositions to make proposals, which appears to work well in our numerical experiments.

Simulation results

$n = 500, p = 40$, error variances drawn from $\mathrm{Unif}(1 - b, 1 + b)$.



TD and LISTEN are two frequentists' structure learning algorithms assuming equal error variance. MINIMAP denotes minimal I-MAP MCMC with a score-equivalent posterior (not assuming equal error variance).

Simulation results

Results for $p = 7, n = 100$. We exactly calculate the posterior distribution $e^{\psi_{\text{eev}}}$ (which is non-score-equivalent and assumes equal error variance) and e^{ψ} (which is score-equivalent and does not assume equal error variance). We draw error variances from $\text{Unif}(1-b, 1+b)$ or Inv-gamma$(3, 2)$.

Method		$b = 0$	$b = 0.3$	$b = 0.5$	$b = 0.7$	$b = 0.9$	$\text{IG}(3, 2)$
Non-score-equivalent	HD	0.1±0.0	0.5±0.2	1.6±0.4	2.1±0.5	2.6±0.5	3.3±0.8
	Flip%	1.1±0.7	4.0±1.5	10.0±2.4	13.4±3.0	18.5±3.9	21.1±4.1
Score-equivalent	HD	3.0±0.3	2.5±0.2	2.6±0.3	2.6±0.2	2.7±0.2	2.6±0.2
	Flip%	23.0±2.9	22.3±3.1	23.4±3.2	23.7±3.2	24.7±3.1	23.7±3.0

Even when $b = 0.9$, imposing equal variance assumption is helpful. The score-equivalent method makes more mistakes about edge directions.

Another interpretation

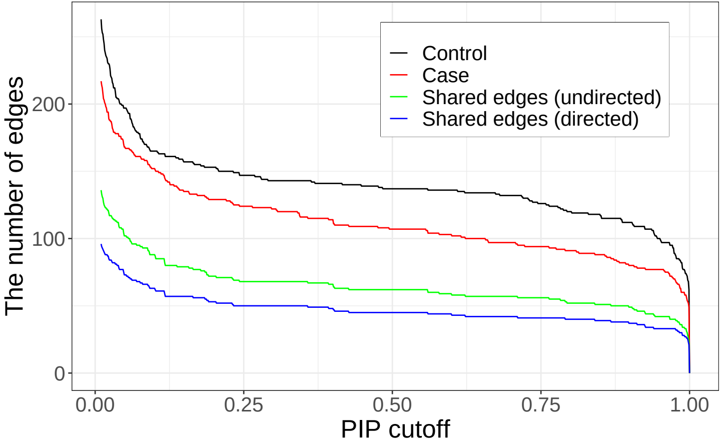
As long as we have a minimal amount of information about the error variances, we can probably obtain more accurate results by scaling the data and imposing the equal error variance assumption.

Single-cell data analysis

A single-cell RNA data set on Alzheimer's diseases [11].

- Control $n_0 = 2,300$, case $n_0 = 1,666$.
- Genes from BDNF (brain-derived neurotrophic factor) pathway: $p = 73$.
- Normalized log-transformed expression levels.
- We analyze case and control samples separately. For each we run order MCMC for 2×10^5 iterations (first half as burn-in).

Single-cell data analysis



PIP: posterior inclusion probability of each edge. Most edges have the same direction in both data sets.

Comparison with the score-equivalent approach

At PIP cutoff $= 0.5$, for our method, 41% of edges in G^{case} are also in G^{cont}. For minimal I-MAP MCMC (score-equivalent), this ratio is 26%.

Stability analysis: repeat the same analysis 30 times and calculate the Gelman-Rubin scale factor for each edge.

- For our method, 99.7% edges have GR ≤ 1.1. For minimal I-MAP MCMC, this ratio is 93.7%.
- For minimal I-MAP MCMC, 90 edges have GR $= \infty$.
- For our method, maximum GR $= 2.56$ in control samples and 1.26 in case samples.

Concluding remarks

- We obtain the first rapid mixing guarantee for high-dimensional structure learning via MCMC sampling. A random walk MH sampler on the MEC space that attains this bound is constructed.
- To obtain the consistency of GES in high-dimensional settings, we introduce swap moves and find sufficient sparsity conditions.
- We show that imposing the equal error variance assumption is likely to improve the mixing of MCMC algorithms and thus increase the estimation accuracy. An order MCMC sampler is developed.
- Mixing time of the MCMC sampler should probably be taken into account when we choose the statistical model.
- Instead of trying to improve the MCMC algorithm, sometimes it may help to "modify" the target posterior.
- Expert knowledge is important, even if it is inaccurate.

Thank you!

Slides available at `https://web.stat.tamu.edu/~quan/papers.html`

- Q. Zhou and H. Chang. "Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes." *Annals of Statistics*, `arXiv:2101.04084`.
- H. Chang, J. Cai and Q. Zhou "Order-based structure learning without score equivalence", *Biometrika*, `arXiv:2202.05150`.

References I

[1] Raj Agrawal, Caroline Uhler, and Tamara Broderick. Minimal I-MAP MCMC for scalable structure discovery in causal dag models. In *International Conference on Machine Learning*, pages 89–98, 2018.

[2] Federico Castelletti, Guido Consonni, Marco L Della Vedova, and Stefano Peluso. Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

[3] Hyunwoong Chang, James Cai, and Quan Zhou. Order-based structure learning without score equivalence. *arXiv preprint arXiv:2202.05150*, 2022.

[4] Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

[5] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[6] Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440, 2002.

[7] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Counting and sampling from markov equivalent dags using clique trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3664–3671, 2019.

References II

[8] Paolo Giudici and Robert Castelo. Improving Markov chain Monte Carlo model search for data mining. *Machine learning*, 50(1-2):127–158, 2003.

[9] Marco Grzegorczyk and Dirk Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71(2-3):265, 2008.

[10] Yangbo He, Jinzhu Jia, and Bin Yu. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *The Annals of Statistics*, 41(4): 1742–1779, 2013.

[11] Jing Jiang, Cankun Wang, Ren Qi, Hongjun Fu, and Qin Ma. scREAD: a single-cell RNA-Seq database for Alzheimer's disease. *Iscience*, 23(11):101769, 2020.

[12] Kyoungjae Lee, Jaeyong Lee, and Lizhen Lin. Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse cholesky factors. *The Annals of Statistics*, 47(6):3413–3437, 2019.

[13] David Madigan, Jeremy York, and Denis Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

[14] David Madigan, Steen A Andersson, Michael D Perlman, and Chris T Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics–Theory and Methods*, 25(11): 2493–2519, 1996.

References III

[15] Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

[16] Michael D Perlman. Graphical model search via essential graphs. *Contemporary Mathematics*, 287:255–266, 2001.

[17] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[18] Robert W Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V: Proceedings of the Fifth Australian Conference, Held at the Royal Melbourne Institute of Technology, August 24–26, 1976*, pages 28–43. Springer, 1977.

[19] Chengwei Su and Mark E Borsuk. Improving structure MCMC for Bayesian networks through Markov blanket resampling. *The Journal of Machine Learning Research*, 17(1):4042–4061, 2016.

[20] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

[21] Quan Zhou and Hyunwoong Chang. Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *Annals of Statistics*, to appear, 2023.