

MCMC for Directed Acyclic Graphs via Birth-and-Death Processes

Radu Craiu

Department of Statistical Sciences
University of Toronto



Morris Greenberg



Kieran Campbell

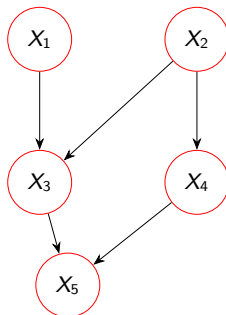
ICSA - Montreal - Aug 7, 2026

Outline

- ▶ Motivation: cancer cell interdependence
- ▶ Aim: infer a DAG from data
- ▶ Develop MCMC in the space of restricted orders
- ▶ Allow the restricted space to evolve
- ▶ Validity follows from the work of [Geyer and Møller \(1994\)](#)

Inferring a DAG from data

- ▶ Multivariate distributions can be represented through a directed acyclic graph (DAG) plus parameters; each vertex is a unique variable.
- ▶ Under a multivariate normal, the DAG dictates the sparsity pattern of the precision matrix. Here we work **outside** the multivariate normal setup.



- ▶ A DAG factorizes the joint dist'n over each variable's parents:

$$P(X_1, \dots, X_5) = \prod_{j=1}^5 P(X_j \mid \mathbf{Pa}(j)).$$

- ▶ For this graph:

$$P(X_1) \cdot P(X_2) \cdot P(X_3 \mid X_1, X_2) \\ P(X_4 \mid X_2) \cdot P(X_5 \mid X_3, X_4).$$

DAG notation

- ▶ A graph is $G = (V, E_G)$ where V is set of vertices (or nodes) and set of edges is E_G
- ▶ In a directed G , edges are ordered: $i \rightarrow j$ indicates i is a parent of j and $i \leftarrow j$ implies the reverse.
- ▶ We denote the parents of node i as $pa_G(i)$ (or $pa(i)$)
- ▶ No node is both an ancestor and a descendant of another).
- ▶ Every DAG G is compatible with at least one **topological order**, $\prec \in \mathbb{S}^p$, i.e. a permutation of $\{1, \dots, p\}$ that exhibits the ancestry of the vertices.
- ▶ If $\prec_{[i]} < \prec_{[j]}$ for all $i \rightarrow j$ in G , then $G \in \mathcal{G}_\prec$. Additionally, i is a *predecessor* of j in $\prec$, or $i \in pr_\prec(j)$ for short.

Graph inference methods

- ▶ Three main methods: *constraint-based (CB)*, *score-based (SB)*, and *hybrid (H)* methods
- ▶ CB infer graph structure by estimating conditional dependencies
 - ▶ Yield consistent estimators of the true DAG under mild assumptions ([Kalisch and Bühlmann 2007](#))
 - ▶ Require a large number of conditional independence tests ([Spirites et al. 2000](#))
- ▶ SB infer structure via an objective function (*score*) that measures the relative merits of each graph
 - ▶ Score ties the graph directly to model quality
 - ▶ Scalability as sample space is exponential in p (no. of nodes)
- ▶ H combine strategically CB and SB
 - ▶ Use CB to construct a reduced space of graphs
 - ▶ Use SB to explore the space

Ingredients: Scores

- ▶ A score $Q(G, D)$ measures how well graph G explains the data D .
- ▶ Common choices include:
 - ▶ **Information criteria:** BIC or AIC-type scores, usually based on penalized likelihood.
 - ▶ **Bayesian posterior scores:** unnormalized posterior scores such as

$$Q(G, D) \propto P(D \mid G)\pi(G).$$

- ▶ **Bayesian network scores:** BDe for multinomial data and BGe for Gaussian data.
- ▶ Decomposable scores are especially useful:

$$Q(G, D) = \prod_{i=1}^p S(X_i, \mathbf{Pa}_G(i) \mid D),$$

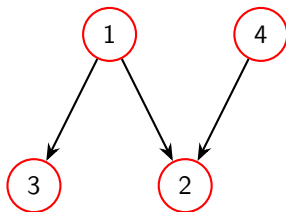
so local graph updates only require rescoreing affected nodes.

Ingredients: Orders

- ▶ Two DAGs are in an *equivalence class* if encode the same set of joint distributions.
- ▶ When the true DAG belongs to a nontrivial equivalence class, structure learning from observational data alone becomes unidentifiable.
- ▶ We move away from MCMC on the space of graphs to the space of orders.
- ▶ Using the order space:
 - ▶ Remedies unidentifiability within equivalence classes
 - ▶ Allows for "larger" moves than using the graph space.
 - ▶ Reduces the computation complexity to $O(p^K)$

Multiple orders for the same DAG

(a)



(b)

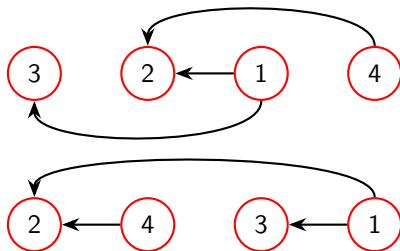
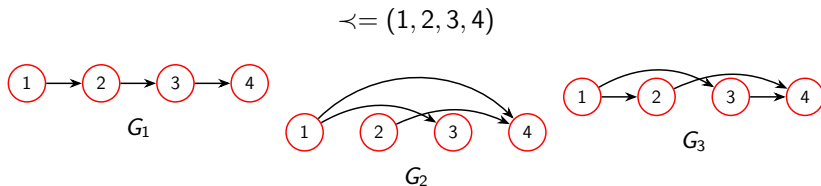


Figure: (a) Example DAG with 4 nodes (b) Two examples of topological order representation of the DAG in (a). In both topological orders, parents are always to the right of children.

Multiple DAGs for the same order



- ▶ Each graph has arrows only from earlier to later nodes in $\prec$.
- ▶ Hence $G_1, G_2, G_3 \in \mathcal{G}_{\prec}$ even though their edge sets differ.

Ingredients: Restricted spaces

- ▶ Kuipers et al. 2021 use a *restricted search space* estimated via constraint-based methods in a hybrid setup.
- ▶ Let $\mathcal{H} = (V, E_{\mathcal{H}}) \in \mathbb{D}_p$ denote the restricted space, with corresponding adjacency matrix H (where $H_{ji} = 1$ if $j \rightarrow i \in \mathcal{H}$) and $\mathbf{h}^i = \{X_j : H_{ji} = 1\}$.
- ▶ Index (number of parents) is restricted to be less than K
- ▶ The admissible DAGs are
$$\mathcal{G}_{\mathcal{H}} := \{G : G = (V, E_G) \in \mathbb{G}_p, E_G \subseteq E_{\mathcal{H}}\}.$$

Sampling via MCMC

- ▶ MCMC algorithms for graphs can vary in 3 key ways
- ▶ The score function used
 - ▶ We use decomposable scores
- ▶ The state space used
 - ▶ The state space is restricted ($\mathcal{H}$)
 - ▶ We work in the order space and the graph space alternating
 - ▶ Sample $\prec \sim \pi_{Ord}(\prec | D, \mathcal{H})$
 - ▶ **Sample $G \sim \pi_{DAG}(G | \prec, D, \mathcal{H})$**
- ▶ The move proposal setup is to alternate between moves in:
 - ▶ the order space given a restricted space
 - ▶ the set of restricted spaces

Hybrid sampler in the order space

- ▶ The posterior for order $\prec$ is proportional to the score $R(\prec | D)$

$$\begin{aligned}\pi_{Ord}(\prec | D) &\propto R(\prec | D) \\ &\propto \sum_{G \in \mathcal{G}_{\prec}} \pi_{DAG}(G | D) \propto \prod_{i=1}^p \sum_{\mathbf{Pa} \in \mathbf{Pa}_{\prec}(i)} S(X_i, \mathbf{Pa} | D)\end{aligned}$$

- ▶ When using a restricted space $\mathcal{H}$ the posterior becomes

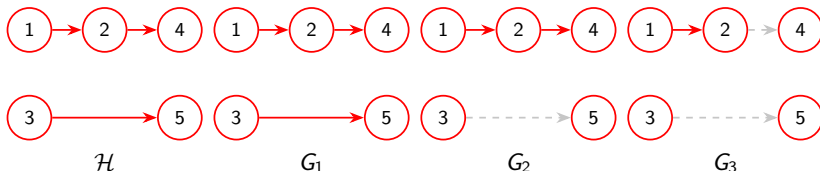
$$R_{\mathcal{H}}(\prec | D) \propto \prod_{i=1}^p \sum_{\substack{\mathbf{Pa} \subseteq \mathbf{h}^i \\ \mathbf{Pa} \in \mathbf{Pa}_{\prec}(i)}} S(X_i, \mathbf{Pa} | D)$$

- ▶ Constraining scores to be on the restricted space with graph index at most K **reduces the cost of MCMC steps from $O(p^K)$ to $O(Kp)$** .
- ▶ Caveat: $R_{\mathcal{H}}$ will not be the same target as π_{Ord} .

Compatible graphs for $\mathcal{H}$ and $\prec$

$$\prec = (1, 2, 3, 4, 5), \quad E_{\mathcal{H}} = \{1 \rightarrow 2, 2 \rightarrow 4, 3 \rightarrow 5\}.$$

- Think of $\mathcal{H}$ as a small graph of **allowed** edges.
- A compatible DAG $G \in \mathcal{G}_{\mathcal{H}, \prec}$ keeps a subset of the edges in $E_{\mathcal{H}}$: it may **drop** some allowed edges but never adds new ones.

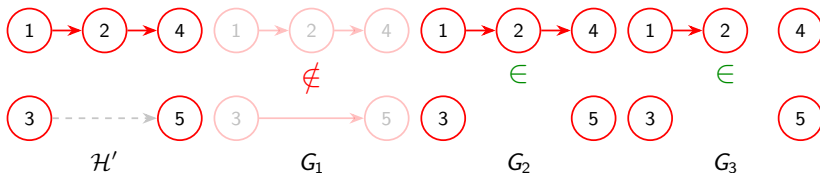


Solid red edges are kept; gray dashed edges are dropped. Each G keeps a subset of the edges of $\mathcal{H}$.

Changing the space: Removing one edge restricts the compatible set

$$E_{\mathcal{H}'} = E_{\mathcal{H}} \setminus \{3 \rightarrow 5\}.$$

- ▶ The edge $3 \rightarrow 5$ is no longer admissible.
- ▶ Every graph that **used** $3 \rightarrow 5$ now drops out of the compatible set; the rest survive.



The move $\mathcal{H} \mapsto \mathcal{H}'$ changes the sum from $\sum_{G \in \mathcal{G}_{\mathcal{H}}}$ to a smaller $\sum_{G \in \mathcal{G}_{\mathcal{H}'}}$.

Cost of restricting: Order space

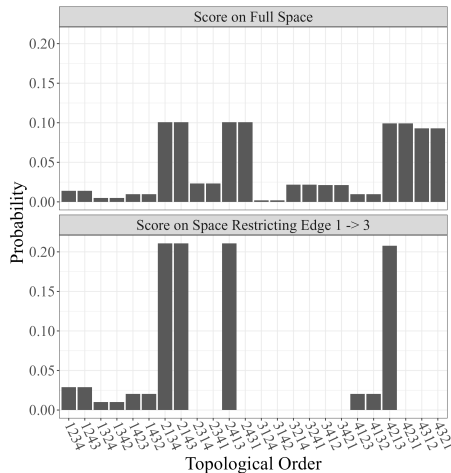


Figure: 4-node example: order posterior for the full search space (top) vs. the restricted space (bottom), where the edge $1 \rightarrow 3$ is excluded.

MCMC on restricted spaces

- **Theorem 1** Let $\mathcal{H} \in \mathbb{D}_p$ be a restricted space with corresponding edge set $E_{\mathcal{H}}$ and admissible DAG set $\mathcal{G}_{\mathcal{H}}$. Let π_{DAG}, π_{Ord} denote the posterior distributions over DAGs and orders, respectively, using the full DAG space $\mathbb{G}_p$, and let $\pi_{Ord}^{(\mathcal{H})}$ be the posterior over orders restricted to $\mathcal{G}_{\mathcal{H}}$. Define $\varepsilon_{\mathcal{H}} := 1 - \pi_{DAG}(\mathcal{H}|D)$ as the posterior mass assigned to DAGs outside of $\mathcal{H}$. Then, for some constant $c \in [0, 1]$,

$$1 - \sqrt{1 - c\varepsilon_{\mathcal{H}}} \leq D_{TV}(\pi_{Ord}^{(\mathcal{H})}, \pi_{Ord}) \leq \min \left(\sqrt{2 - 2\sqrt{1 - c\varepsilon_{\mathcal{H}}}}, 1 \right)$$

where D_{TV} denotes the total variation distance between distributions.

Challenges

- ▶ We consider moves that grow AND shrink $\mathcal{H}_t$
- ▶ By varying $\mathcal{H}_t$ we aim to reduce the error established in Theorem 1
- ▶ The restricted spaces $\mathcal{H}_t$ are of varying dimension
- ▶ Of interest is the stationary distribution on $(\mathbf{H}, \mathbf{O}) \in \mathbb{D}_p \times \mathbb{S}^p$
- ▶ This requires theoretical validation.

A sampler with varying state spaces

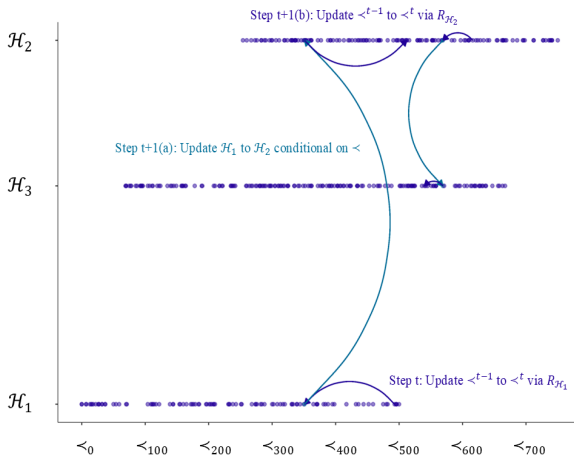


Figure: Illustration of the **B**irth-death processes **R**estricted **O**ver **O**rders **D**istributions (BROOD) Algorithm

Sampling via Restricted MCMC

- We need to sample:

Step H Challenging: sample the restricted space ($\mathcal{H}$) conditional on the order ($\prec$).

Step O Sample order given the restricted space

Step G Finally, given the restricted space and the order, sample from the set of graphs compatible with both $\mathcal{H}$ and $\prec$.

- We rely on the framework developed by [Geyer and Møller \(1994\)](#)
- Consider the MH algorithm sampling on $(\mathcal{H}, \prec) \in \mathbb{D}_p \times \mathbb{S}^p$ that is a mixture of 2 transition kernels:

$$Q_\ell(\mathcal{H}, \prec | \mathcal{H}_t, \prec_t) = (1 - \ell)Q_0(\prec | \mathcal{H}_t, \prec_t) + \ell Q_1(\mathcal{H} | \mathcal{H}_t, \prec_t) \quad 0 \leq \ell \leq 1$$

- Q_0 is a standard MH procedure concentrated on $\mathcal{H}_t$ for Step 0.
- Q_1 is concentrated on $H_{m_t-1} \cup H_{m_t} \cup H_{m_t+1}$ where $\mathcal{H}_t \in H_{m_t}$ and the index m_t denotes the number of edges.

The development steps

- ▶ Given the current state $(\mathcal{H}, \prec)$
- ▶ Construct a birth and death process with rates $\beta(\mathcal{H}, \prec)$ and $\delta(\mathcal{H}, \prec)$ that has a stationary distribution.
- ▶ Define a MH algorithm with acceptance probabilities that depend on β and δ
- ▶ Ergodicity follows from the work of [Geyer and Møller \(1994\)](#) ([Preston, 1977](#), see also).

Sampling via Restricted MCMC

Theorem 2: Let $\mathcal{H} \in \mathbb{D}_p$, and $\prec \in \mathbb{S}^p$. Define $B_e(\mathcal{H}, \prec)$ and $D_e(\mathcal{H}, \prec)$ as follows:

$$B_e(\mathcal{H}, \prec) \propto \frac{\frac{1}{|\mathcal{G}_{\mathcal{H}^+e}|} \pi_{DAG}(\mathcal{H}^{+e} | \prec, D)}{\frac{1}{|\mathcal{G}_{\mathcal{H}}|} \pi_{DAG}(\mathcal{H} | \prec, D)} = \frac{1}{2} \frac{\pi_{DAG}(\mathcal{H}^{+e} | \prec, D)}{\pi_{DAG}(\mathcal{H} | \prec, D)}$$

$$D_e(\mathcal{H}, \prec) \propto \frac{\frac{c^*}{|\mathcal{G}_{\mathcal{H}^-e}|} \pi_{DAG}(\mathcal{H}^{-e} | \prec, D)}{\frac{1}{|\mathcal{G}_{\mathcal{H}}|} \pi_{DAG}(\mathcal{H} | \prec, D)} = 2c^* \frac{\pi_{DAG}(\mathcal{H}^{-e} | \prec, D)}{\pi_{DAG}(\mathcal{H} | \prec, D)}$$

where $c^* \in (0, 1]$. Let $\beta(\mathcal{H}, \prec) = \sum_{e \notin E_{\mathcal{H}}} B_e(\mathcal{H}, \prec)$, and $\delta(\mathcal{H}, \prec) = \sum_{e \in E_{\mathcal{H}}} D_e(\mathcal{H}, \prec)$. Then, a birth-death process defined by the rates β, δ yields a unique solution with a limit that is a tractable probability distribution.

Putting it all together

Theorem 3: Let μ be a trans-dimensional probability measure that is compatible with a birth-death process defined by rates β, δ . Then $r(x, \xi)$, the acceptance ratio in the mixture kernel sampler of [Geyer and Møller \(1994\)](#), is defined by the ratio of waiting times $w(x \cup \xi), w(x)$ of the jump process:

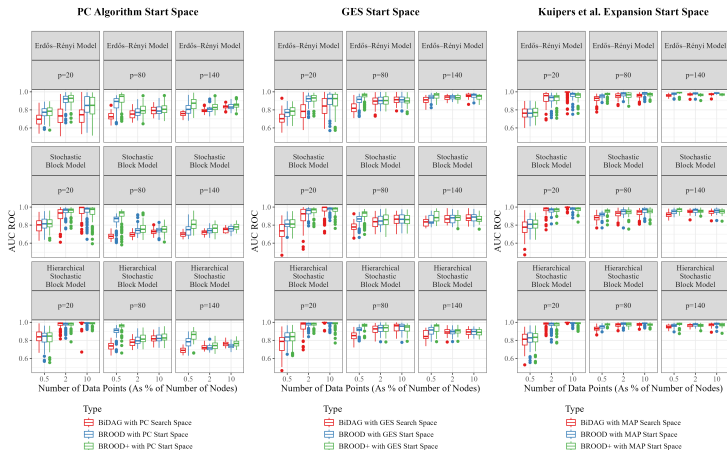
$$r(x, \xi) = \frac{\frac{1}{\beta(x \cup \xi) + \delta(x \cup \xi)}}{\frac{1}{\beta(x) + \delta(x)}} = \frac{w(x \cup \xi)}{w(x)}$$

Time reversibility is ensured with the following acceptance rates:

$$\begin{aligned} \text{death} \rightarrow A_1(x \mid x \cup \xi) &= \begin{cases} \min\{1, 1/r(x, \xi)\} & \text{if } x \cup \xi \in H, \\ 0 & \text{otherwise} \end{cases} \\ \text{birth} \rightarrow A_1(x \cup \xi \mid x) &= \begin{cases} \min\{1, r(x, \xi)\} & \text{if } x \cup \xi \in H, \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Simulation results - fixed sparsity

ROC AUC for a Variety of Models
Data generated from Structural Equation Models



- ▶ What did we learn?
 - ▶ Designing good MCMC samplers for DAGs is hard
 - ▶ We lose the intuition given in Euclidian spaces by distances
 - ▶ Space is discrete but enormous and missing "parts" of it can be consequential
 - ▶ Theory for such samplers is challenging, but badly needed.
- ▶ What do we want to know?
 - ▶ Can we design an adaptive version so tuning parameters, say c^* , are optimized?
 - ▶ Downstream data analysis...
- ▶ Paper available here: raducraiu.com/publications/

References

- GEYER, C. J. and MØLLER, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics* **21** 359–373.
URL <http://www.jstor.org/stable/4616323>
- KALISCH, M. and BÜHLMANN, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research* **8** 613–636.
URL <http://jmlr.org/papers/v8/kalisch07a.html>
- KUIPERS, J., SUTER, P. and MOFFA, G. (2021). Efficient Sampling and Structure Learning of Bayesian Networks. *Journal of Computational and Graphical Statistics* **0** 1–12.
URL <https://doi.org/10.1080/10618600.2021.2020127>
- PRESTON, C. (1977). Spatial birth-and-death processes. *Bulletin of the International Statistical Institute* **46** 371–391.
URL <https://mathscinet.ams.org/mathscinet/relay-station?mr=0474532>
- SPIERTES, P., GLYMOUR, C. and SCHEINES, R. (2000). Causation, Prediction, and Search.
URL <https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/>