Structural inference of DAGs (with MCMC)

Robert Goudie

SGX Meeting

March 11, 2015

Aim

Aim to estimate the structure of dependence between various components

Want to estimate the structure of a directed acyclic graph (DAG)

$$
\begin{tikzcd}
y_1 \arrow[r] & y_2 \arrow[rr] \arrow[dr] & & y_4 & \\
 & & y_3 \arrow[ur] \arrow[rr] & & y_5
\end{tikzcd}
$$

Essentially exploratory analysis

There are a large number of DAGs — for $p = 13$ nodes, 18676600744432035186664816926721 DAGs

Acyclicity restriction is awkward

Constraint-based (frequentist):

PC-algorithm (and similar) test for (conditional) independence of each pair (y_i, y_j) of variables.

Bayesian approaches:

Treat the DAG as just another parameter

Bayesian networks

Graph: nodes $v \in V$, edges $e \in V \times V$

Random vector Y, with components y_v for $v \in V$,

identify each y_v with node v.

Acyclic: no cycles/loops. Need a Directed Acyclic Graph (DAG)

$$
\begin{tikzcd}
y_1 \arrow[rr] & & y_2 \arrow[rr] \arrow[dr] & & y_4 & \\
 & & & y_3 \arrow[ur] \arrow[rr] & & y_5
\end{tikzcd}
$$

The graph specifies joint distribution can be factorised as

$$p(y_v) = \prod_{v \in V} p(y_v \mid y_{parents(v)})$$

Notation

Set of all DAGs $\mathcal{G} = \{G_g : g = 1, \ldots, |\mathcal{G}|\}$

Prior on DAGs $G_1, \ldots, G_{|\mathcal{G}|}$

$$\Pr(G_g) = p_g, \qquad g = 1, \ldots, |\mathcal{G}|$$

$$\text{where } p_g > 0 \text{ and } \sum_{g=1}^{|\mathcal{G}|} p_g = 1$$

Observations y

Each model G_g has parameters $\theta_g \in \Theta_g$, with prior $p(\theta_g)$

Likelihood under model G_g is $p(y \mid G_g, \theta_g)$

Model selection/averaging

Posterior distribution for DAGs

$$\Pr(G_g \mid y) \propto p(y \mid G_g)\Pr(G_g)$$

where $p(y \mid G_g)$ is the marginal likelihood of G_g

$$p(y \mid G_g) = \int_{\Theta_g} p(y \mid G_g, \theta_g) p(\theta_g) d\theta_g.$$

Difficulties:

- Evaluating the marginal likelihood – but use conjugate prior
- Normalising constant for $\Pr(G_g \mid y)$ is $\sum_{G_k \in \mathcal{G}} p(y \mid G_k)\Pr(G_k)$

Posterior computation

Hill-climbing algorithms etc for the MAP

Exact inference for the full posterior

- Direct enumeration
- Dynamic programming: Tian and He (2009); Koivisto and Sood (2004)

MCMC algorithms

A Metropolis-Hasting algorithm (MC^3; Madigan & York, 1995)

Construct Markov Chain $M(t), t = 1, 2, \ldots$

State space $\mathcal{G}$, the space of Bayesian Networks. ie DAGs

Target distribution $\Pr(M \mid x)$

$$\begin{tikzcd}
 & y_1 \arrow[dl] & \\
y_2 \arrow[rr] & & y_3 \arrow[dl] \\
 & y_4 &
\end{tikzcd}$$

current

$$\begin{tikzcd}
 & y_1 \arrow[dl] \arrow[dr] & \\
y_2 \arrow[rr] & & y_3 \arrow[dl] \\
 & y_4 &
\end{tikzcd}$$

add $y_1 \rightarrow y_3$

✓

$$\begin{tikzcd}
 & y_1 \arrow[dr] & \\
y_2 \arrow[rr] & & y_3 \arrow[dl] \\
 & y_4 &
\end{tikzcd}$$

remove $y_1 \rightarrow y_2$

✓

$$\begin{tikzcd}
 & y_1 \arrow[dr] & \\
y_2 \arrow[rr] & & y_3 \arrow[dl] \\
 & y_4 \arrow[ul] &
\end{tikzcd}$$

add $y_4 \rightarrow y_2$

✗

A Metropolis-Hasting algorithm (MC^3)

Neighbourhood $\eta(G)$ is set of DAGs with an edge added or removed.

Sample proposal G' uniformly from $\eta(G)$

Accept proposal with probability $\min(1, \alpha)$, where

$$\alpha = \frac{p(y \mid G')\,\Pr(G')}{p(y \mid G)\,\Pr(G)} \frac{|\eta(G')|^{-1}}{|\eta(G)|^{-1}}$$

Checking for cycles

Simple: Try all edge additions. Cycles check using DFS. $= \mathcal{O}(p^3)$

Instead use the transitive closure.

$$\begin{tikzcd}
 & y_1 \arrow[dl] & \\
y_2 \arrow[rr] & & y_3 \arrow[dl] \\
 & y_4 &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & y_1 \arrow[dl] \arrow[dr] \arrow[dd] & \\
y_2 \arrow[rr] \arrow[dr] & & y_3 \arrow[dl] \\
 & y_4 &
\end{tikzcd}$$

Adding an edge $i \to j$ introduces a cycle iff $j \to i$ is in the transitive closure of the initial DAG.

Query in $\mathcal{O}(1)$

Can update a matrix of path counts C_{ij} incrementally: adding an edge $i \to j$ increases the number of path from $k \to l$ by $C_{ki} C_{jl}$ (King and Sagert, 2002) – updates in $\mathcal{O}(p^2)$

Problem with MC^3

- Can be very slow to converge.
- Problem is combination of a large space and multimodality
- MC^3 moves are too 'small' and 'local'.
- The posterior is 'peakier' as sample size n increases.

Troublesome situations

- Reversing an edge
- 'Any 2 of 4' etc
- Near cyclic loops

Graph (a)

$$\begin{tikzcd} y_1 \arrow[r] \arrow[d] & y_2 \\ y_3 \arrow[ur] & \end{tikzcd}$$

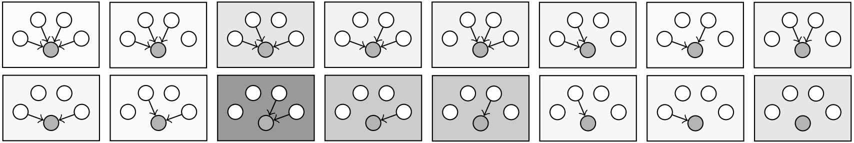
Graph (b)

$$\begin{tikzcd} y_1 \arrow[d] & y_2 \arrow[l] \arrow[dl] \\ y_3 & \end{tikzcd}$$

Regression - variable selection

Idea: use the connection with variable selection.

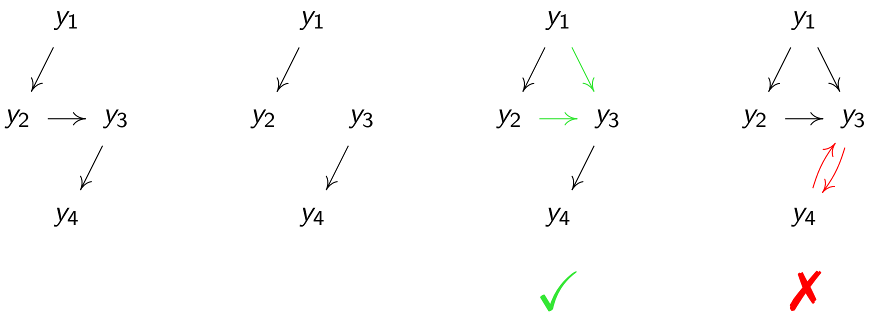
In regression there are 2^p models. For $p = 4$,



A Gibbs sampler

At each step sample a new set of parents of a particular node.

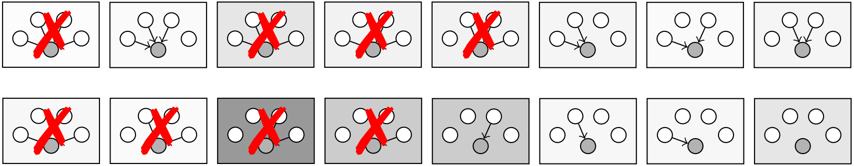
0. Current | 1. Remove $y_\bullet \to y_3$ | 2. Sample new parents for y_3



Each step is then a *constrained* variable selection problem.

A Gibbs sampler

Sample parents from conditional distribution



1. Choose node
2. Identify ‘parent sets’ that are non-cycle-forming
3. Renormalise the non-cycle-forming parent sets
4. Sample new parents for the selected node according to this distribution

Notes

- Can be thought of as ‘blocking’ – a standard trick for Gibbs sampling.
- Correctness doesn’t follow from usual proof of Gibbs sampling (Hammersley-Clifford’s postivity condition does not apply)
- Need to constrain in-degree for feasibility in large graphs
- Larger blocks. Product distribution of constrained variable selection problems. Blocks of 3 nodes seemed to work well.

Comparison of methods

Convergence

- trace plots of marginal likelihoods
- comparing posterior edge probabilities between runs

Accuracy

- Absolute errors in posterior edge probabilities
- ROC curves

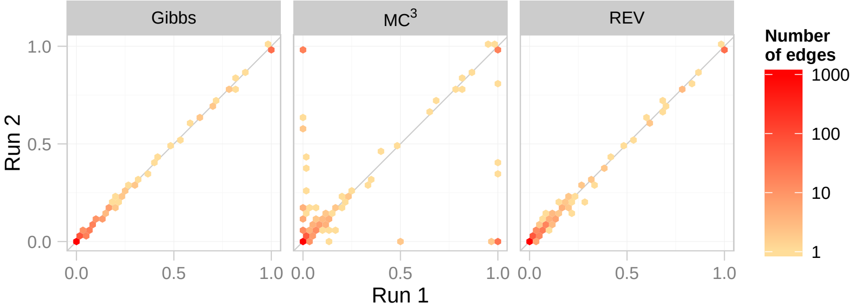
REV sampler

Grzegorczyk and Husmeier (2008) – Metropolis-Hastings sampler

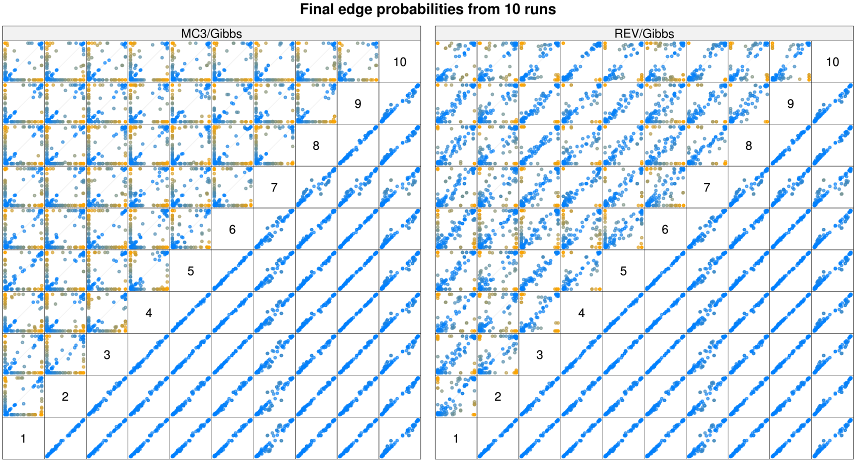
- Select an existing edge $i \to j$
- Generate a proposal graph in which edge is reversed, and new parents for i and j
- Accept proposal as per Metropolis-Hastings

Assessing convergence of the MCMC algorithm can be tricky

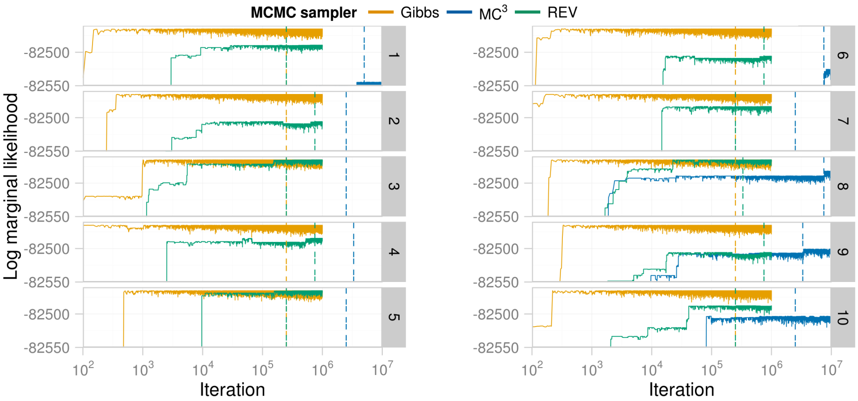
Stability of inclusion probabilities



Convergence assessment

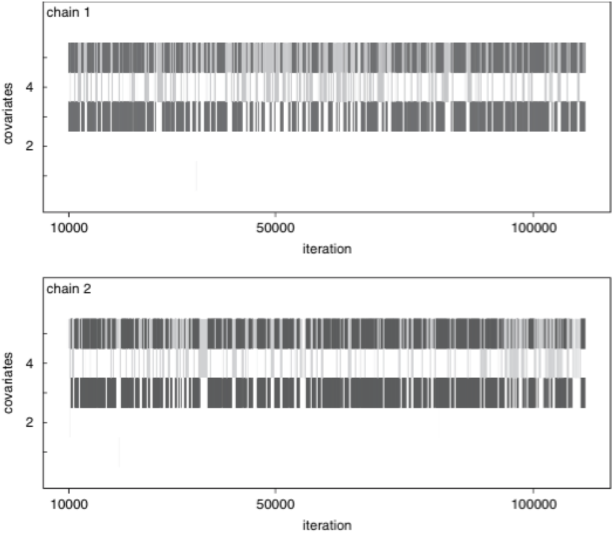


Convergence assessment



Convergence assessment

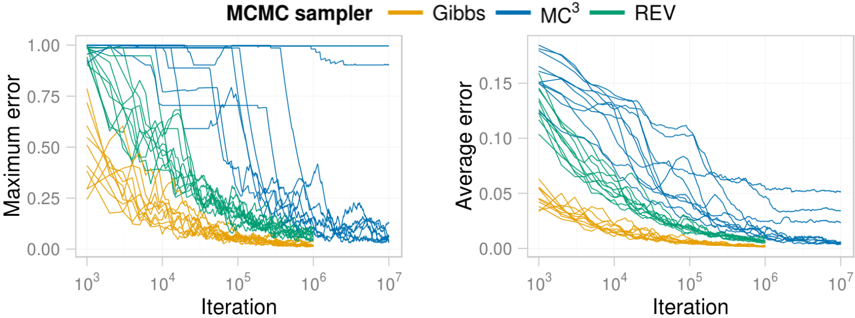
'Jump' extension to BUGS (Lunn, 2008)



Absolute errors in posterior edge probabilities

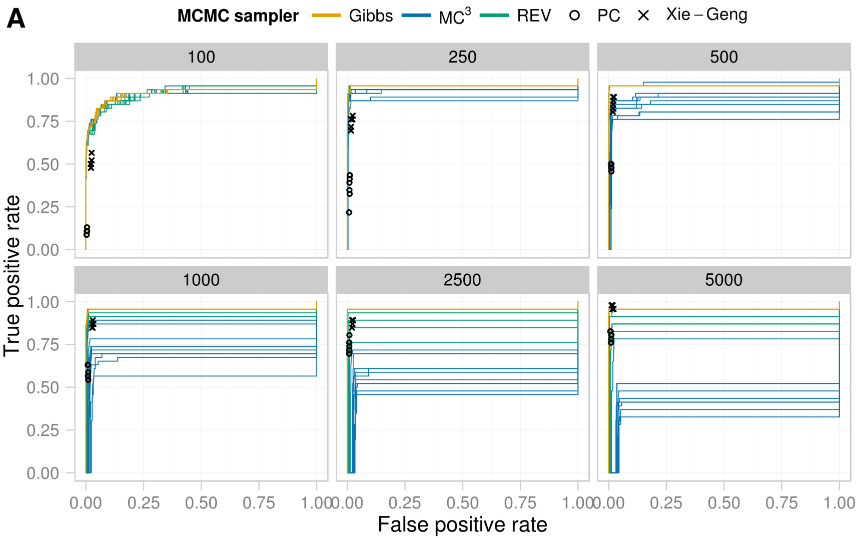
$p = 18$ example, $n = 101$

Compare to exact posterior edge probabilities via Tian and He (2009) method

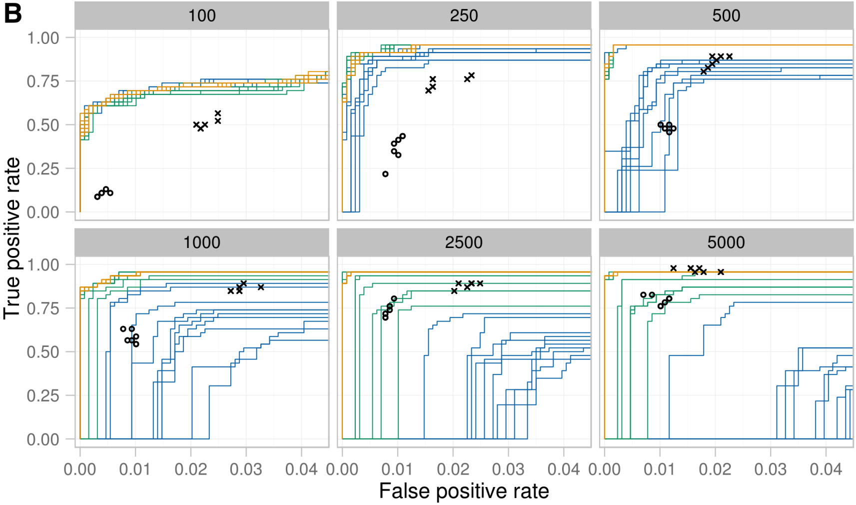


ROC curves

$p = 37,\ n = 100, 250, 500, 1000, 2500, 5000$

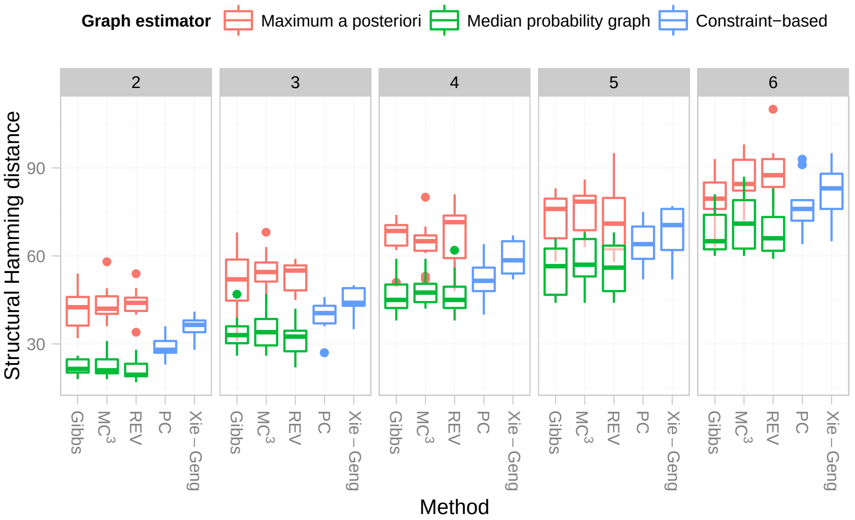


ROC curves


B
100
250
500
1000
2500
5000
1.00
0.75
0.50
0.25
0.00
True positive rate
0.00
0.01
0.02
0.03
0.04
False positive rate

Sparseness

'Random networks', $p = 25$, $n = 1000$



Conclusions

- MC^3 is fine for small p, small n problems
- For p in the hundreds or more, constraint-based methods are difficult to beat.
- For p in the tens (or low hundreds), full Bayesian solutions are possible. REV or Gibbs work well here.
- For $p < 20$ish exact Bayesian solutions available (remarkably)

- Worth looking for 'larger' Gibbs moves to get better mixing
- Implementation in R: *github.com/rjbgoudie/structmcmc*

References

- Hoeting, J., Madigan, D., Raftery, A. E. and Volinsky, C. (1999) *Bayesian model averaging: A tutorial*. Statistical Science, **14**, 382401.
- Madigan, D., & York, J. C. (1995). *Bayesian Graphical Models for Discrete Data.* International Statistical Review / Revue Internationale de Statistique, **63**, 215–232.
- Smith, M., & Kohn, R. (1996). *Nonparametric Regression Using Bayesian Variable Selection.* Journal of Econometrics, **75**, 317–343.
- Grzegorczyk, M., & Husmeier, D. (2008). *Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move.* Machine Learning, **71**, 265-305.
- Hobert, J. P., Robert, C. P. and Goutis, C. (1997) *Connectedness conditions for the convergence of the Gibbs sampler*. Statistics & Probability Letters, **33**, 235240
- King, V., & Sagert, G. (2002). *A Fully Dynamic Algorithm for Maintaining the Transitive Closure.* Journal of Computer and System Sciences, 65(1), 150–167.