A Tempering Algorithm For Large-sample Network Inference

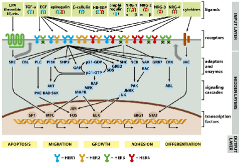
D. Barker[1], S. Hill[1] and S. Mukherjee[2]

[1]Complexity Science DTC, [2] Dept. of Statistics
University of Warwick, England
CV4 7AL

MIR@W - 7 March 2011

Background

- Networks (e.g. genes, proteins, metabolites) important notion in current biology.
- **Probabilistic Graphical Models** (PGM) are a key approach.
- RTK is an example of a signalling network.

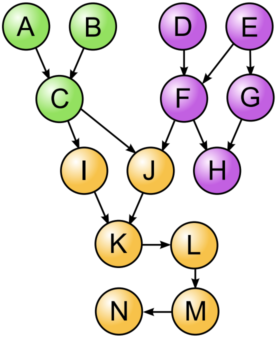


(Weinberg 2007, Yarden & Sliwkowski 2001)

Bayesian Networks

- **Stochastic Models** where a graph is used to describe probabilistic relationships between components.
- Graph specifies **conditional independence statements**.
- In some frameworks the graph must be directed and acyclic (DAG).

  Special cases include HMMs, Bayesian Networks (BN), Dynamic BNs.



Structural Inference

Interested in inference of graph G given some observed data $\mathbf{X}$.

Posterior probability over graphs G give by Bayes' Theorem

$$P(G|\mathbf{X}) \propto P(\mathbf{X}|G)P(G)$$

Has closed form *up to proportionality constant* for certain choices of underlying models.

Maximising $P(G|\mathbf{X})$ can have robustness problems; If posterior has several highly scoring graphs how do we choose between them?

- For this reason we use **model averaging**.

Model Averaging

Probability $E(e)$ of seeing an edge e averaged over all graphs G is more robust.

- Edges which repeatedly appear in likely graphs have high $E(e)$.

Knowledge of proportionality constant requires *enumeration* of whole p-node DAG space $\mathcal{G}$.

- $\mathcal{G}$ grows super-exponentially with p.

Thus we must use MCMC to estimate the posterior probabilities $P(G|\mathbf{X})$.

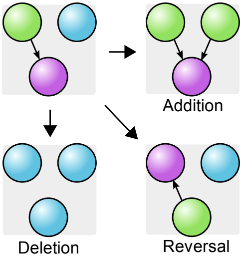
Monte Carlo

- Move around $\mathcal{G}$ by performing elementary moves on current graph G.
- Accept or reject new graphs G' based on MH acceptance probability;

$$\alpha = \frac{P(\mathbf{X}|G')|\eta(G)|}{P(\mathbf{X}|G)|\eta(G')|}$$

(for uniform priors)

Called MC^3 (Madigan & York 1995)

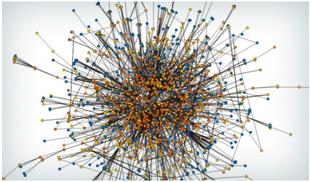


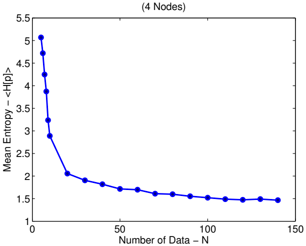
Neighbourhood $\eta(G)$ is all graphs reachable from G.

Estimate of posterior probability given by

$$\hat{P}(G|\mathbf{X}) = \frac{1}{t_{\text{max}}} \sum_{t=1}^{t_{\text{max}}} I(g^{(t)} = G)$$

Sample Size





Having more data is clearly a good thing.

- High throughput experiments, FACS, social science, etc...

Caution!

In certain situations large sample size N can cause problems.

Convergence to correct stationary distribution can be slow.

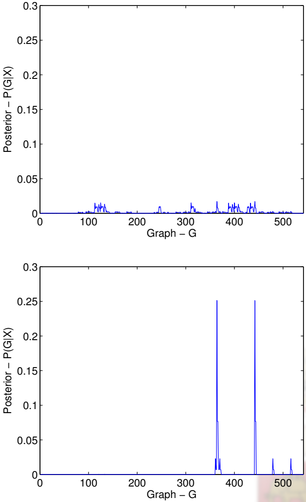
Motivation

- Posterior for p=4 node system with two different sample sizes $N = 5$ and $N = 10$

- Posterior mass concentrates on a few highly likely graphs.

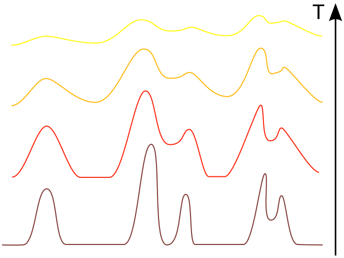
- If these are hard to get between Markov chain mixing is slow.

Note: As $N \to \infty$ we pick out all graphs from the correct data generating class.



Tempering

- Aim is to allow the Markov chain (MC) to move between high scoring graphs.
- Natural idea is to use tempering.
- Couple higher temperature MCs to the one with the desired posterior.



Temperature analogy achieved by raising posterior score to $\beta = \frac{1}{T}$:

$$P(G|\mathbf{X})^{\beta} \propto (\, P(\mathbf{X}|G)P(G) \,)^{\beta}$$

Tempering Scheme

Set up m MCs at temperatures $T_1, \dots T_m$.

MCs at higher temperature can explore the space more freely.

- Each chain simulated using often used MH scheme.

Every iteration with probability p_{swap} swap graphs between randomly chosen neighbouring chains i and j

- Accept the swap with probability ρ.

Swapping probability

$$\rho = \frac{(P(\mathbf{X}|G_j)P(G_j))^{\beta_i}\,(P(\mathbf{X}|G_i)P(G_i))^{\beta_j}}{(P(\mathbf{X}|G_i)P(G_i))^{\beta_i}\,(P(\mathbf{X}|G_j)P(G_j))^{\beta_j}}$$

Simulation

First we examine performance on synthetic data generated from the known network shown earlier.

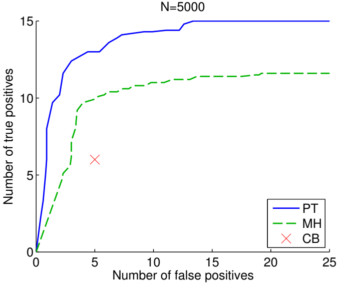
Data is generated using

- $A \sim N(0, \sigma)$ for parent nodes.
- $C \sim N(A + B + \gamma AB, \sigma)$ for child nodes.
  (with parents A and B)

Since we know the underlying graph from which the data were generated we can draw ROC curves...

ROC Curves

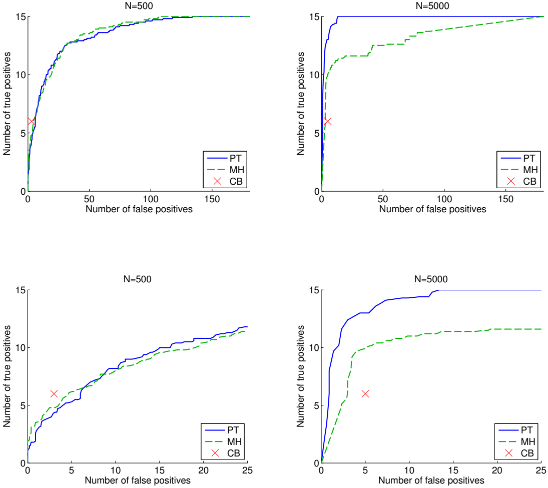
Curves parametrised by threshold t; keep in output graph all edges with $E(e) > t$.



Tempering has picked up fewer false positive edges compared to standard MC^3 for the same number of true edges.

(Xie & Geng, JMLR 2008)

ROC Curves


N=500
N=5000
N=500
N=5000
Number of true positives
Number of false positives
PT
MH
CB

Proteomic Data

We examine here the application to inferring the underlying DBN from a set of proteomic data.

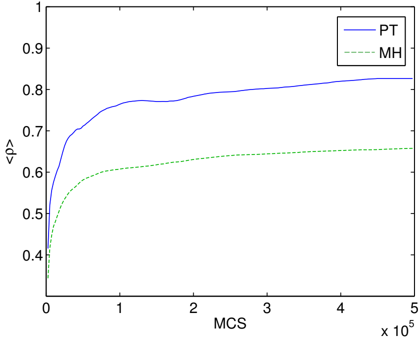
Due to certain factorisation for DBNs we can calculate **exact edge probabilities**.

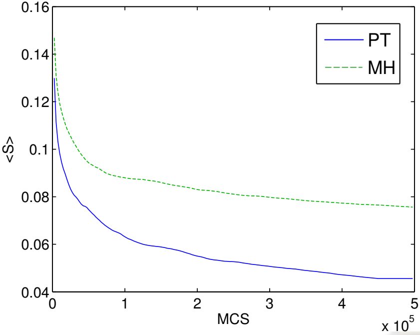
- Gives us gold standard comparison!

We examine;

- **Correlation** ρ between the exact and MC estimated edge probabilities.
- Normalised **sum difference** s between the exact and MC estimated edge probs.

Edge Probabilities

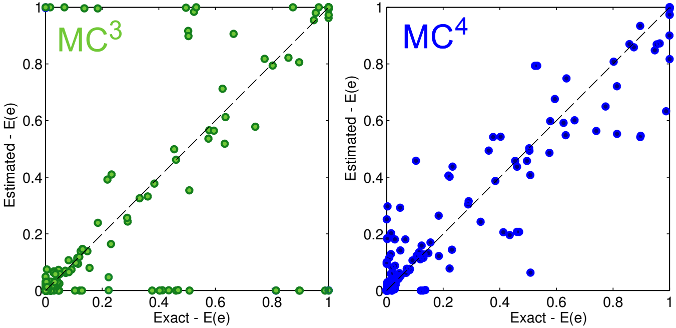




T =1.0,1.25,1.5,1.75,2.0 and $p_{\text{swap}} = 0.1$, averaged over 4 runs.

Edge Probabilities

If we look at the individual edge probabilities we see better performance (closer to $x = y$) for tempering:



Toughest edges to infer are significantly better estimated by using tempering.

Conclusions

- As sample size increases posterior mass can concentrate around several hard to move between graphs.
- Widely employed MCMC schemes can fail to estimate edges properly in these increasingly common situations.
- Counter this by using higher temperature chains coupled to desired posterior: PT.
- Important to draw robust conclusions from data in a wide range of fields.

Acknowledgements

I would like to thank my co-authors

- Sach Mukherjee,
- Steven Hill,

as well as

- R. Goudie
- M. Nicodemi
- N. Podd

for useful discussions, EPSRC for funding...

And finally, thank you for listening.