

Fitting Directed Acyclic Graphs in genomics: an empirical Bayes approach

Prediction and Marker Selection

Siem Heisterkamp^{1,2}

¹MSD Research Laboratories, department BARDS Oss, The Netherlands

²Groningen Bioinformatics Centre (GBIC), Groningen, The Netherlands

2011 Bayesian Biostatistics Conference, Houston

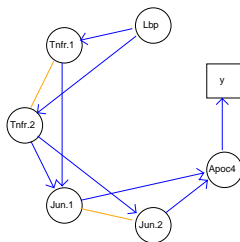


Liver Toxicity in rat

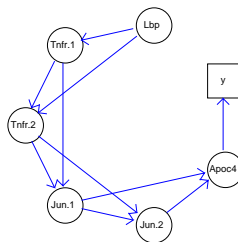
- ▶ Part of a large study (collaboration between MSD and University of Groningen)
 - ▶ Rats exposed to a range of toxic compounds, duration of time and doses
 - ▶ Gene-expression arrays applied on liver
- ▶ How to find biologically meaningful associations?

Directed Acyclic Graph

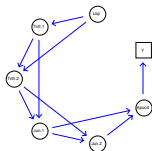
Full model with common edges



Full DAG model



Full DAG model



Adjacency matrix

	<i>Lbp</i>	<i>Tnfr.1</i>	<i>Tnfr.2</i>	<i>Jun.1</i>	<i>Jun.2</i>	<i>Apoc4</i>	<i>Y</i>
<i>Lbp</i>	0	1	1	0	0	0	0
<i>Tnfr.1</i>	0	0	1	1	0	0	0
<i>Tnfr.2</i>	0	0	0	1	1	0	0
<i>Jun.1</i>	0	0	0	0	1	1	0
<i>Jun.2</i>	0	0	0	0	0	1	0
<i>Apoc4</i>	0	0	0	0	0	0	1
D_{in}	0	1	2	2	2	2	1

DAG

Some properties

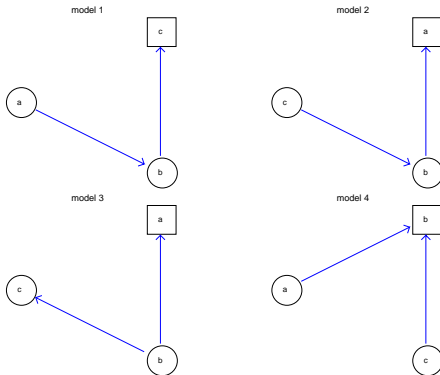
- ▶ **Directed**: directed *Edge* one *node* to another
- ▶ **Acyclic**: No path starts and stops in the same node
- ▶ **Graph**: representation of *nodes* and *edges*

How many DAG's are possible?

- ▶ With **4** nodes, **1** response and *no relabeling* **855**
- ▶ With **10** nodes, **1** response and *no relabeling* $1.3 \cdot 10^{23}$
- ▶ Compare without direction
 - ▶ 4 nodes, 1 response : 31 configurations
 - ▶ 10 nodes, 1 response : 2047 configurations

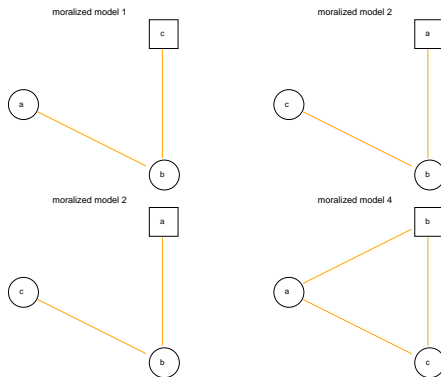
Markov-Equivalence of different DAG's

Example: DAG's with 3 nodes



Markov-Equivalence of different DAG's

Associated Moralized graphs



Seemingly different DAG's may be equivalent

- ▶ Two DAG's are **Markov Equivalent** *iff*
 1. Skeleton graph's are equal
 2. The 'amoralities' are the same
- ▶ In other words: the **moralized** graphs must be the same
- ▶ Causality can *only* be established in case of 'colliding' arrows

Directed Acyclic Graph

Finding the distributions

- ▶ $\log \text{lik}(\text{Dag}) = \sum_{\nu} \log P(\text{Child}|\nu)$
- ▶ Linear model; algorithm by Cox & Wermuth (1996)

1. *Moralize* the graph

2. Trace-back from \mathbf{Y} $E[y|\nu_y] = \sum_j a_{jy} \beta_{y|\nu_j}$ with $\left(\sum_j a_{jy} - d_y\right) \neq 0$

3. Conditional logLik: $\propto \lambda (y - E[y|\nu_y])^2$
and precision $\lambda \left(d_y - \sum_j a_{jy}\right)$

4. Repeat 2 until last married grand-parents...

Simple Example: neighbour prior

Neighbour prior



$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -1 & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \Rightarrow \lambda \sum_{j>1} (\beta_j - \beta_{j-1})^2$$

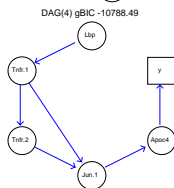
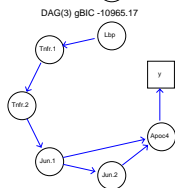
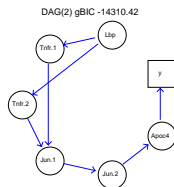
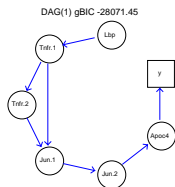
Linear Causal model

$$\log \text{lik} \left(\beta \mid Y, \mathbf{A}, \mathbf{X}, \lambda, \alpha, \underbrace{\dots}_{\text{usual}} \right) = \underbrace{\log \text{lik} \left(Y \mid \beta, \mathbf{A}, \mathbf{X}, \alpha, \underbrace{\dots}_{\text{usual}} \right)}_{\text{usual log lik}} +$$

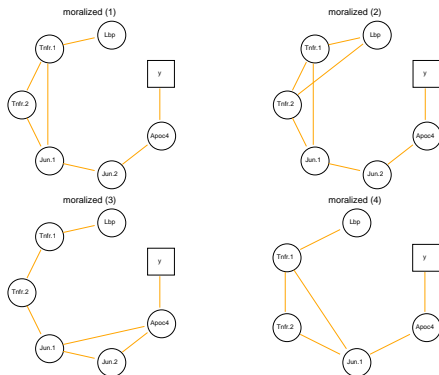
$$-0.5 n \underbrace{\left(\lambda \beta^t \mathbf{L}_A \beta - r_{\mathbf{L}_A} \log(\lambda) - \log(\det(\mathbf{L}_A)) \right)}_{\text{prior log lik}}$$

► r rank of \mathbf{L}_A

4 best DAG's using gBIC



4 best *moralized* DAG's



in-vivo liver toxicity in rat

validation by leave one out

Table: Prediction for 4 models

gBIC	<i>DAG(1)</i>		<i>DAG(2)</i>		<i>DAG(3)</i>		<i>DAG(4)</i>	
	-28071		-14310		-10965		-10788	
	T	C	T	C	T	C	T	C
pred. T	6	0	6	0	6	0	6	0
pred. C	4	10	4	10	4	10	4	10
Total	10	10	10	10	10	10	10	10

Note: *None* of the models are *Markov-equivalent*

How to search plausible models (2)?

- ▶ Double exponential prior on β (Lasso L_1)
- ▶ Poisson prior on the number of edges (L_0)
- ▶ Demanding that at *least* one connected path exists:
 \Rightarrow 1-prob(no pathway of any length exists)
- ▶ extra priors added (penalty functions)

$$\underbrace{-0.5 n \lambda_1 \left(\sum_{s \in E} |\beta_s| \right)}_{\text{prior} \quad \text{log lik } L_1 \quad \text{norm}} + n \log \underbrace{\left(1 - e^{-\sum_k \sum_{\substack{f \in F \\ g \in G}} \sum_s \pi(\lambda_1, \lambda_2, \beta_s) \cdot w_{s,f,g}^k} \right)}_{\text{prior} \quad \text{log lik } L_0 \quad \text{norm}}$$

For Further Reading I

Cox DR, Wermuth N, *Multivariate Dependencies: Models analysis and Interpretation*(1996). Chapman & Hall, London

Dawid, A.P., (2004) Probability, Causality and the Empirical World: A Bayes -de Finetti -Popper -Borel Synthesisism, *Statistical Science*, **19**, 44-57.

Dawid, A.P., (2007) Fundamentals of Statistical Causality *Research Report no 279*, Department of Statistical Sciences, University College London, September 2007

Pearl, J. *Causality: Models Reasoning and Inference*(2000). Cambridge, Cambridge University Press

Raftery, A. E., (1996) Hypothesis testing and model selection in Gilks, W. R., Richardson, S., Spiegelhalter, D.J. (eds.) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.