Stochastics meeting, Lunteren, November 1999

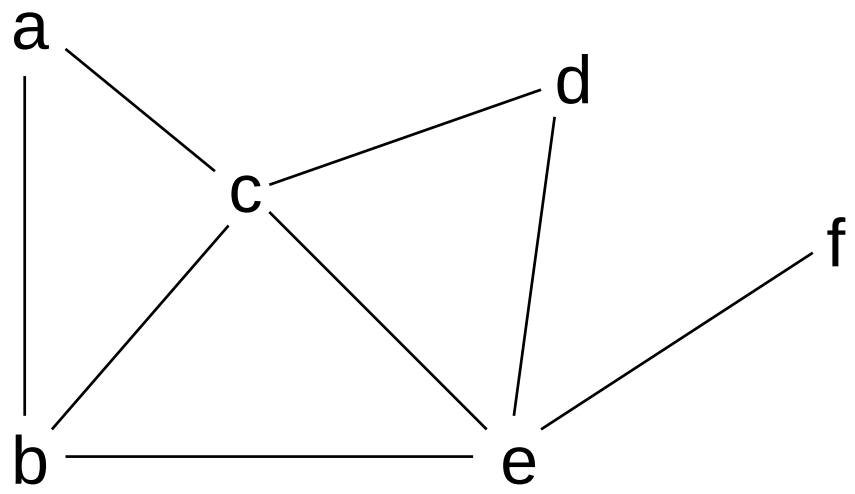
Model determination in decomposable graphical models

by Peter Green (University of Bristol)
and Paolo Giudici (Universitá di Pavia).

- introduction to conditional independence graphs
- Markov properties
- decomposable graphical models
- Bayesian model determination
- gaussian and multinomial cases
- MCMC computation

Conditional independence graph

given a graph whose vertices index the components of a random vector X, draw an (undirected) edge between vertices a and b if X_a and X_b are **not** conditionally independent given all other variables.



$$c \perp f | (a, b, d, e)$$

– being an abbreviation for

$$X_c \perp X_f | X_{\{a,b,d,e\}}$$

Conditional independence properties

The Markov property is familiar from temporal stochastic processes, where we learn that it may be expressed in several equivalent ways. For variables located on an arbitrary graph, the situation is more subtle: can distinguish 4 related properties, each capturing an aspect of Markovness.

Pairwise Non-adjacent pairs of variables are conditionally independent given the rest (see definition of graph).

Local Conditional only on adjacent variables (neighbours), each variable is independent of all others (this simplifies full conditionals).

Global Any two subsets of variables separated by a third are conditionally independent given the values of the third subset.

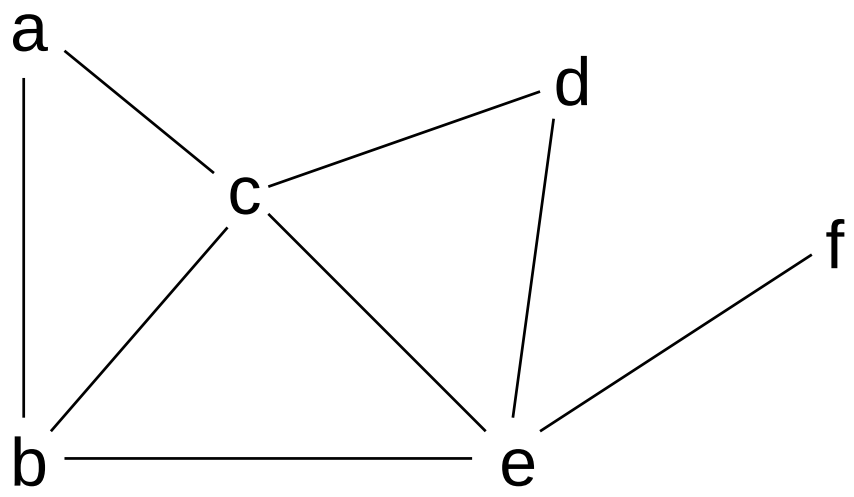
Factorisation The joint distribution factorises as a product of functions on cliques (=maximal complete subgraphs).

Mathematically, the interesting thing is that these are different, although $F \Rightarrow G \Rightarrow L \Rightarrow P$ always.

But for most statistical purposes, the important thing is that they are often the same; a sufficient but not necessary condition is that the joint distribution has the positivity property ("any values realisable individually are realisable jointly").

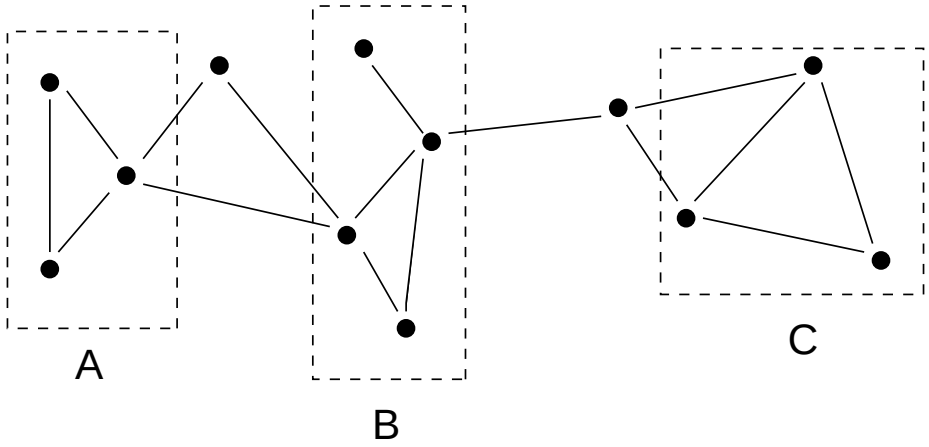
This result includes the Hammersley-Clifford theorem (Markov random field = Gibbs distribution, $L = F$).

For directed acyclic graphs, the situation is simpler: the directed local Markov property is *always* equivalent to the directed graph factorisation criterion: $DL = DF$.



$$P : c \perp f | (a, b, d, e) \qquad L : d \perp (a, b, f) | (c, e)$$

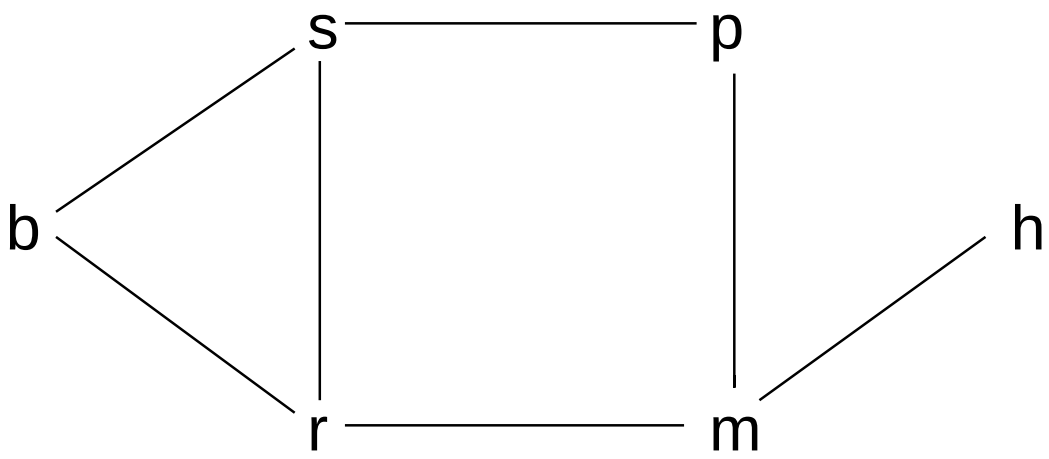
$$F \quad : \quad p(a, b, c, d, e, f) =$$
$$\psi_1(a, b, c)\psi_2(b, c, e)\psi_3(c, d, e)\psi_4(e, f)$$



$$G : A \perp C | B$$

Example: using CI graphs to summarise dependencies

Edwards and Havranek (*Biometrics*, 1985) summarise their conclusions from a 2^6 contingency table like this:



b	blood pressure > 140?
s	smoking?
r	ratio of α and β lipoproteins > 3?
p	strenuous physical work?
m	strenuous mental work?
h	family history of coronary heart disease?

Structural and quantitative learning

Given independent observations on a random vector X, drawn from a parametric statistical model, we wish to make inference

- about the parameters of its distribution (*quantitative learning*), and in particular
- about its conditional independence graph (*structural learning*).

In the Bayesian world, these will be done simultaneously, based on the joint posterior for parameters θ and graph g, derived from an appropriate prior and likelihood.

$$p(\theta, g|X) \propto p(g)p(\theta|g)p(X|\theta, g)$$

Graphical preliminaries: decomposable graphs

$g = (V, E)$ is an undirected graph with vertex set V, $\#V = p$, and edge set E.

A subgraph of g is *complete* if every pair of vertices is joined by an edge. A *maximal* complete subgraph is called a *clique*.

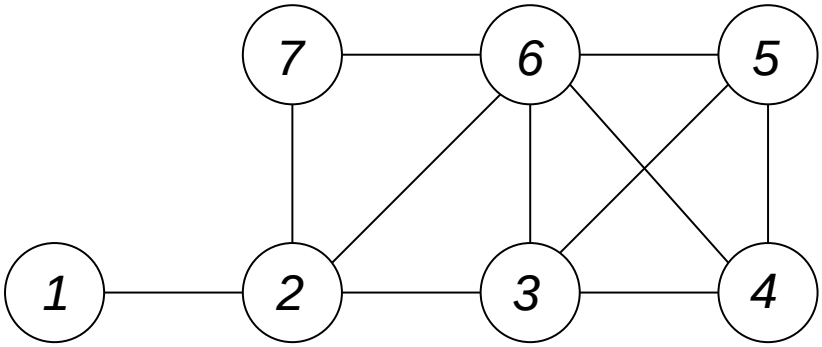
An ordering of the cliques $(C_1, C_2, \ldots, C_n)$ is *perfect* if, for each i, all of the vertices of C_i that are also contained in any previous clique are all contained in just one clique, that is,

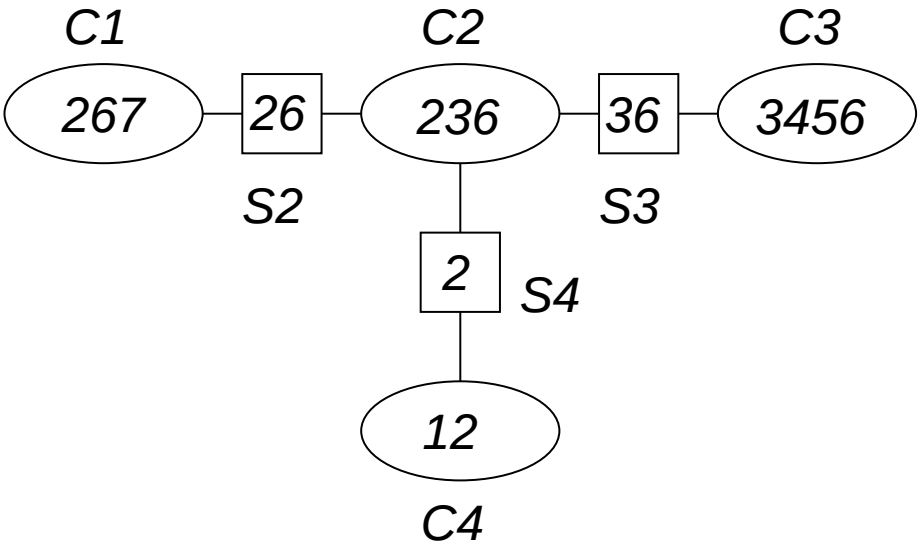
$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \subseteq C_h$$

for some $h = h(i) < i$. The sets S_i are called *separators*. Write $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$, $\mathcal{S} = \{S_2, S_3, \ldots, S_n\}$.

A graph admitting a perfect ordering is said to be *decomposable*; g is decomposable if and only if it has no chordless cycles of length greater than 3 – i.e. g is *triangulated*.

A triangulated graph:





A perfect ordering is given by:

$C_1 = \{2, 6, 7\}$

$C_2 = \{2, 3, 6\}$ $S_2 = \{2, 6\}$

$C_3 = \{3, 4, 5, 6\}$ $S_3 = \{3, 6\}$

$C_4 = \{1, 2\}$ $S_4 = \{2\}$

Decompositions, and the Markov property

A pair (A, B) of subsets of V is a decomposition of g if $V = A \cup B$, $A \cap B$ is complete, and $A \cap B$ *separates* A from B (i.e. every path $v_1 \leftrightarrow v_2 \cdots \leftrightarrow v_k$ from a vertex $v_1 \in A$ to a $v_k \in B$ has a $v_j \in A \cap B$.)

Let $X = X_V$ be a random vector with components indexed by V. For any $A \subseteq V$, X_A denotes the correponding subvector.

The distribution of X is (globally) Markov with respect to g, if for every decomposition (A, B) of g, $X_A \perp X_B | X_{A \cap B}$ ($\perp$ = 'is independent of'). For a decomposable graph, this simpler test implies the usual definition.

The Gaussian case

Suppose that $X_i \sim N_p(\mu, \Sigma)$, independently, for $i = 1, 2, \ldots, n$. We do not have anything special to say about μ; suppose $\mu = 0$.

In this Gaussian case, the global, local and pairwise Markov properties are all equivalent. In fact, writing $K = \Sigma^{-1}$, we have

$$X_i \perp X_j | X_{V\{i,j\}} \Leftrightarrow k_{ij} = 0,$$

that is the graph g is determined by Σ, by joining i and j by an edge if and only if the (i, j) element of the *inverse* of Σ is non-zero.

Let $\Sigma_A = \mathrm{cov}(\mathrm{X_A})$ for any $A \subseteq V$. When g is decomposable, it can be shown that the multivariate Gaussian density can be factorised:

$$p(x|\Sigma, g) = \frac{\prod_{C \in \mathcal{C}} p(x_C|\Sigma_C)}{\prod_{S \in \mathcal{S}} p(x_S|\Sigma_S)}$$

as a ratio of products of the marginal densities on the cliques and separators.

Representing Σ

Need to store Σ_C for all $C \in \mathcal{C}$ and Σ_S for all $S \in \mathcal{S}$, where $\mathcal{C}$ and $\mathcal{S}$ change dynamically as g does. Fortunately, these can all be held within a single overall matrix Σ ($\Sigma_{i,j} = \text{cov}(X_i, X_j)$ makes sense irrespective of clique membership of i and j).

$\Rightarrow$ deal with *partially-specified* covariance matrices, Γ, say, where $\Gamma_{i,j} = \Sigma_{i,j}$ if $(i,j) \in E$, the edge set of g, and is otherwise unspecified. The interpretation of Γ is clarified through the notion of matrix completion (Dempster, 1972; Grone *et al.*, 1984):

1. if a positive definite completion Σ of Γ exists, it is the unique completion that satisfies $(\Sigma^{-1})_{i,j} = 0$ if $(i,j) \notin E$.

2. if g is decomposable, and Γ_C is p.d. for all $C \in \mathcal{C}$, then a p.d. completion exists.

Prior modelling

Our prior belief is that there **is** some structure – the conditional independence graph is not trivial. So we do not simply want to model Σ by a continuous probability model – Σ determines g, and *a posteriori* g would be complete a.s.

Thus we build a more structured model through $p(g)p(\Sigma|g)$, where of course the second factor respects the conditional independence properties specified by g.

The prior on g is trivial – uniform on all decomposable graphs. Of course, we do not know *how many* decomposable graphs there are (except for very small p), but this does not matter – the normalising constant is not needed. Importance sampling allows us in principle to reweight MCMC results to correspond to any other prior on g.

The prior on Σ given g is taken to be hyper-Markov (Dawid and Lauritzen, 1993): that is, for any decomposition (A, B) of g,

$$\Sigma_A \perp \Sigma_B | \Sigma_{A \cap B}$$

This is constructed using the notion of hyperconsistency – two laws on Σ_A and Σ_B are hyperconsistent in they induce the same marginal on $\Sigma_{A \cap B}$. Given the cliques and separators, and a pairwise hyperconsistent collection of laws on Σ_C, there is a unique hyper-Markov law on Σ, that puts all its probability on Σ respecting g.

We take Σ_C to be inverse Wishart $IW(\alpha, \Phi^C)$. Denote the corresponding density by $l_C(\Sigma_C)$; then the distribution for Σ is given by

$$l(\Sigma) = \frac{\prod_{C \in \mathcal{C}} l_C(\Sigma_C)}{\prod_{S \in \mathcal{S}} l_S(\Sigma_S)}$$

– perfectly matching the likelihood factorisation.

How to choose the hyperparameters?

We need to choose the hyperparameters $\{\Phi^A\}$. Both to simplify this choice and to ensure compatibility – we probably want to assign the same $p(\Sigma_C|g)$ for every g for which C is a clique – it is convenient to simply take a single n.n.d. matrix Φ, and set Φ^C to be the submatrix corresponding to variables in C, for each C.

We take Φ to be one of

- fixed, $= \tau(1-\rho)I + \rho J$
- random, as above with random (τ, ρ), or
- random, $\sim W(d, T)$, T diagonal.

The hierarchical versions alleviate some of the difficulties of prior specification, and provide greater robustness to the prior.

Why restrict to decomposable graphs?

If g is decomposable, we have seen that both likelihood and prior distribution factorise according to the clique–separator structure. This structure allows localisation of computations on the posterior, and in particular supports *local updating* in a MCMC sampler.

The decomposable graph is represented, not directly, but by means of its *junction tree* (or *forest* if not connected). The junction tree has

- vertices correponding to the cliques of g
- edges corresponding to the inclusions defining the perfect ordering

Recall

$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \subseteq C_h$$

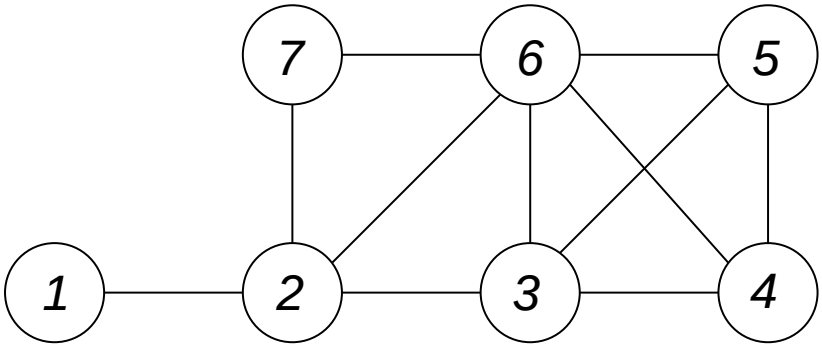
for some h. In the junction tree for this ordering, cliques C_i and C_h are connected by an edge, which can be labelled by the separator S_i. The ordering is not necessarily unique, so neither is the junction tree.

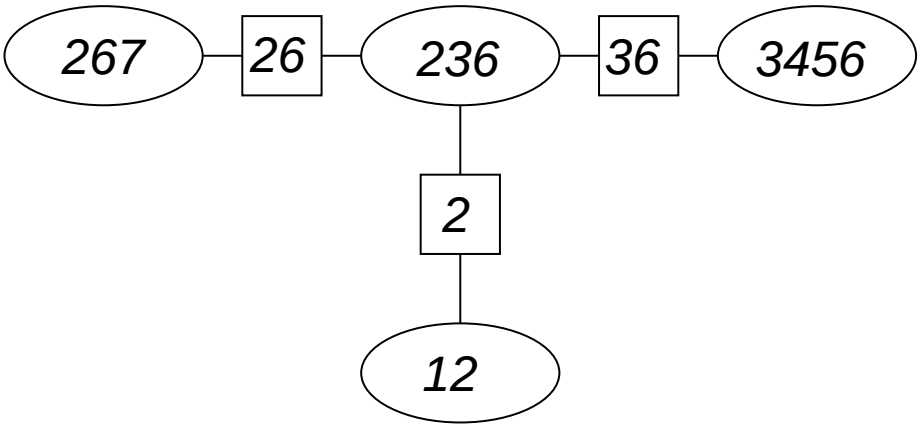
Is it a big restriction?

In many settings, there are results showing arbitrary graphical models are well-approximated by decomposable ones.

Example of a junction tree

A graph on 7 vertices, with 4 cliques, and a junction tree representation (one of two equivalent ones):





The sets of cliques and separators *are* unique.
Note the running intersection property: all cliques containing any specified set of vertices are connected.

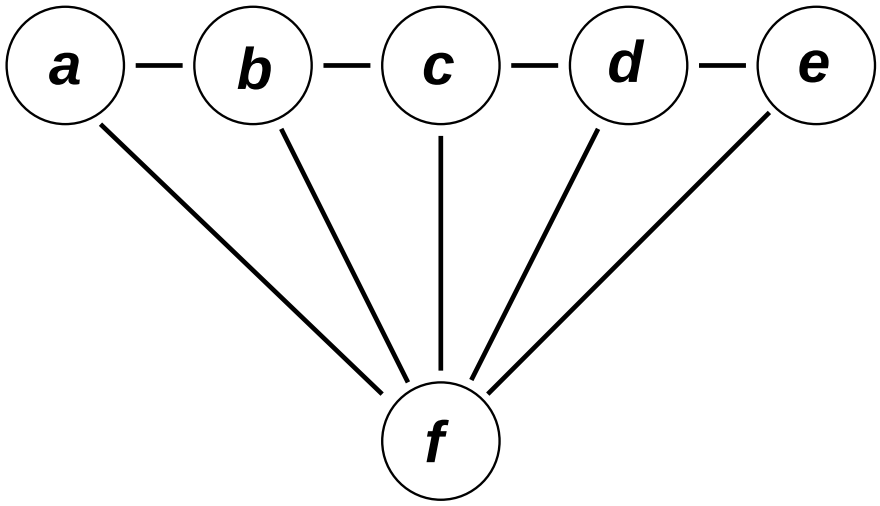
Modifying graphs to preserve decomposability

If we add or delete an edge to a decomposable graph, the result may not be decomposable. Can we check that decomposability is not lost, without doing a global computation?

Legal deletions. An edge can be deleted from g if and only if it is contained in exactly one clique of g. (Frydenberg and Lauritzen, 1989)

Legal additions. An edge (a, b) can be added to g if and only if *either* a and b are in different connected components of g, *or* there exist $R, T \subset V$ such that $a \cup R$ and $b \cup T$ are cliques, and $S = R \cap T$ is a separator on the path between $a \cup R$ and $b \cup T$ in a junction forest corresponding to g.

Example



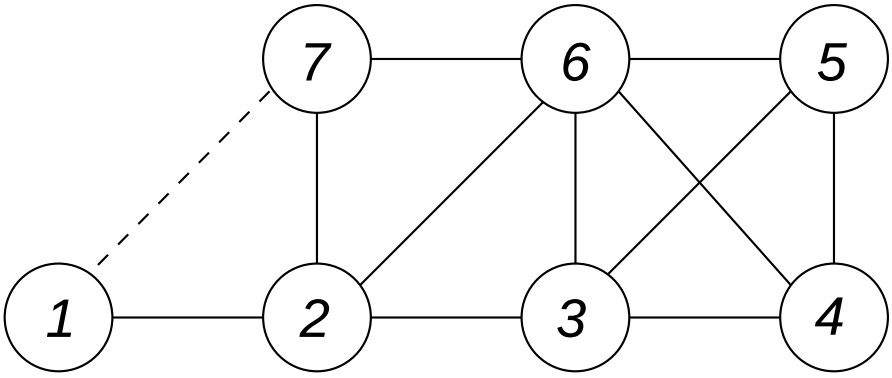
The cliques are $(a, b, f), (b, c, f), (c, d, f), (d, e, f)$.

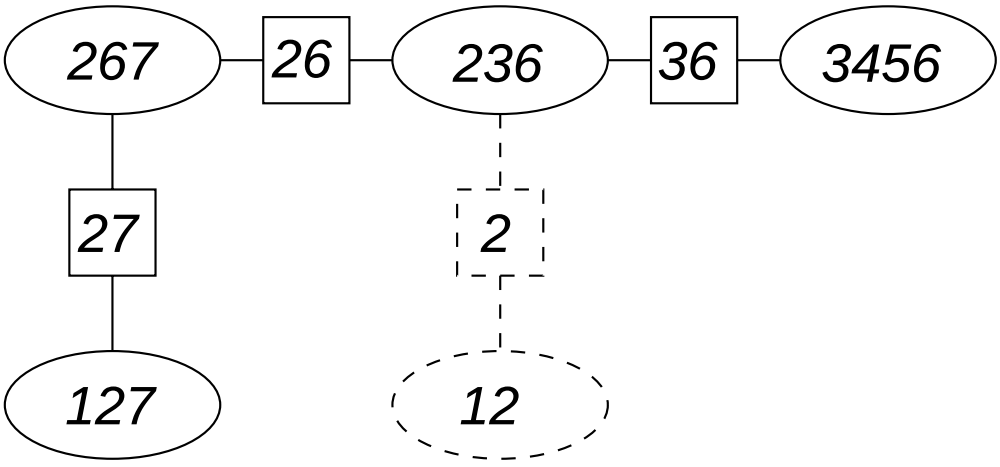
By the first result, edges $(b, f), (c, f), (d, f)$ cannot be deleted.

By the second result, edges $(a, e), (a, d), (b, e)$ cannot be added (for each, $R \cap T = \{f\}$, and $\{f\}$ is not a separator).

Local changes to a junction tree

If we add an edge (1,7), it makes a simple change to the junction tree (or forest):





but we may need to switch to a different equivalent junction tree first.

MCMC computation

Full posterior is computed by MCMC, involving a collection of reversible moves:

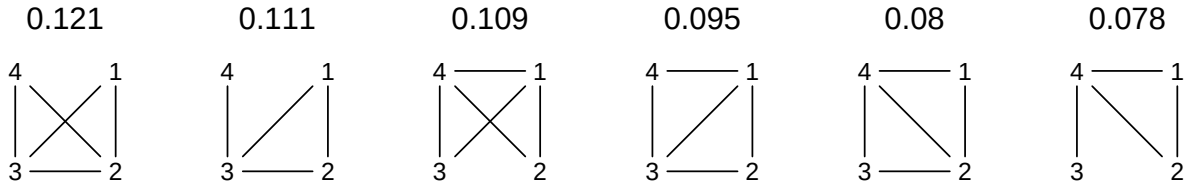
- updating g by adding or deleting an edge (and making minimal consequent changes to Σ) – a 'dimension-changing' move
- updating Σ for given g
- updating hyperparameter α
- updating hyperparameter Φ

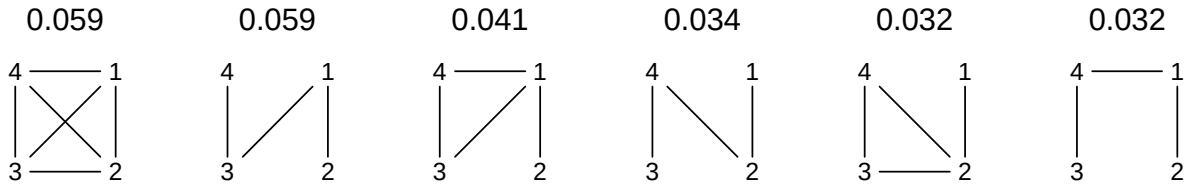
All are Metropolis-Hastings updates, and clique–separator factorisation facilitates computation of acceptance probabilities.

Results on Fret's data

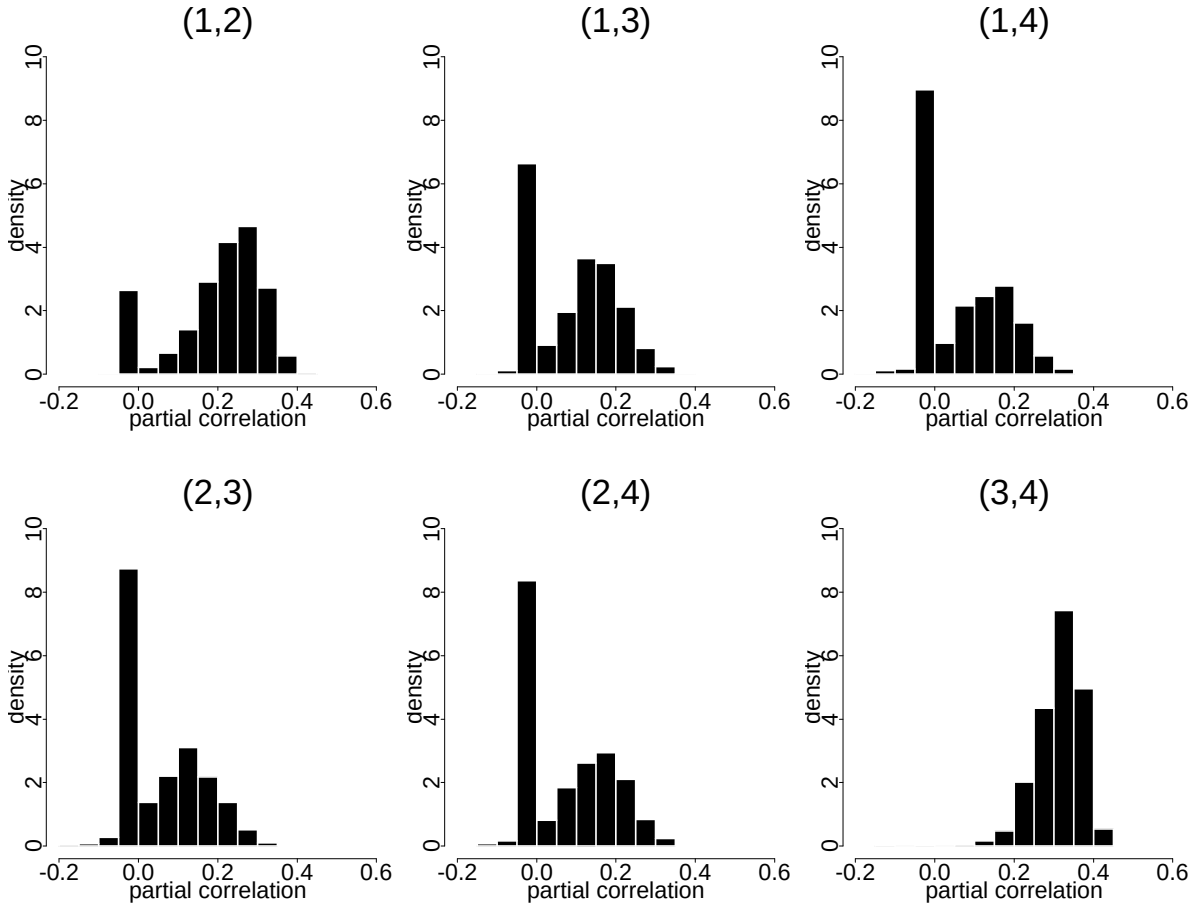
Only 4 variables – 64 graphs, of which 61 are decomposable. All pairs of variables highly correlated marginally, no evident pattern in inverse sample covariance matrix.

Most probable graphs *a posteriori*





Posterior distributions of partial correlation coefficients, for Fret's data



Convergence of ergodic averages, plotted every 200 sweeps:

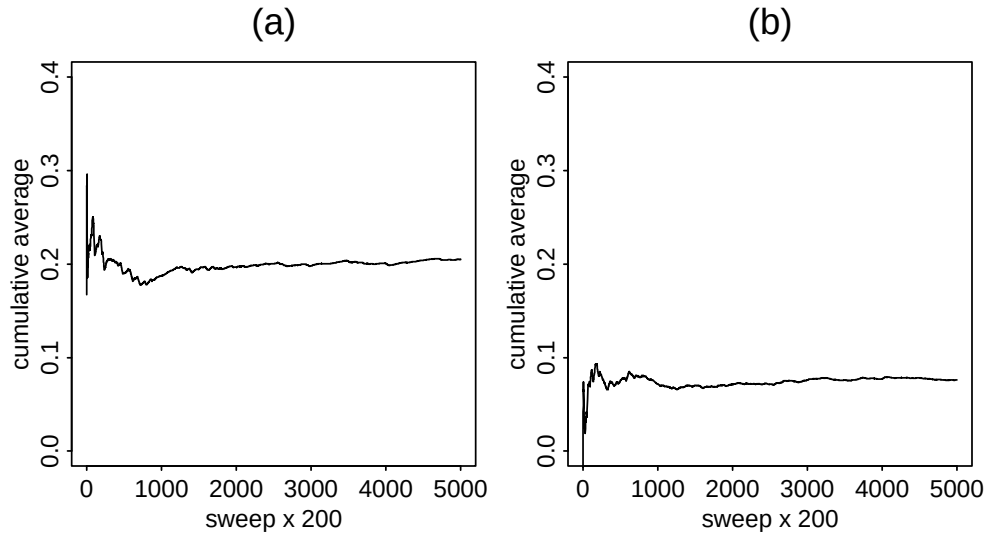


Table 1: Sensitivity of structural learning, with respect to the prior, using Fret's data, for a non-hierarchical model.

	$\alpha = p + 1$		
	$\rho = -0.3$	$\rho = 0$	$\rho = 0.9$
g_0	100011	111001	100011
$p(g_0\|x)$	0.1165	0.1267	0.2645
$E(\text{n.edges}\|x)$	3.63	4.16	3.13

	$\alpha = 2p$		
	$\rho = -0.3$	$\rho = 0$	$\rho = 0.9$
g_0	110001	111011	110001
$p(g_0\|x)$	0.1415	0.1171	0.2142
$E(\text{n.edges}\|x)$	3.52	4.21	3.04

Table 2: Sensitivity of model averaged inference on a partial correlation coefficient with respect to the prior, for a non-hierarchical model, using Fret's data.

	$\alpha = p + 1$		
	$\rho = -0.3$	$\rho = 0.9$	$\rho = 0.9$
$E(\rho_{12}\|x)$	0.204	0.207	0.241

	$\alpha = 2p$		
	$\rho = -0.3$	$\rho = 0$	$\rho = 0.9$
$E(\rho_{12}\|x)$	0.204	0.208	0.239

Table 3: Sensitivity of structural learning with respect to the prior, for a hierarchical model, using Fret's data.

	$f = p + 1$		
	$d = 2$	$d = p$	$d = 2p$
g_0	110111	111011	110111
$p(g_0 \mid x)$	0.1383	0.1304	0.1422
$E(\text{n.edges} \mid x)$	4.41	4.40	4.46

	$f = 2p$		
	$d = 2$	$d = p$	$d = 2p$
g_0	111011	111011	111011
$p(g_0 \mid x)$	0.1317	0.1412	0.1517
$E(\text{n.edges} \mid x)$	4.45	4.54	4.53

Table 4: Performance of the Markov chain Monte Carlo samplers: acceptance fractions and computation times, in minutes and seconds for 100 000 sweeps.

Move type	Fret's	Spatial lattice
g	0.022	0.002
Σ	0.573	0.379
Φ	0.566	0.642
α	0.518	0.577
Time	2:16	22:03

Decomposable contingency tables – the multinomial case

– work recently submitted, with Claudia Tarantola and Paolo Giudici.

The Dawid & Lauritzen theory also applies to the Dirichlet / Multinomial set-up.

Again, likelihood and prior can be represented as ratios of products of terms indexed by cliques and separators.

Differences

A new problem, *cf.* the gaussian case: find a parsimonious clique-based parameterisation.

Gaussian case: store covariance matrix for each clique: these can all be held as an ‘active subset’ of the overall covariance matrix.

Multinomial case: need to store cell probability matrix for each clique.

$\Rightarrow$ new data structures are required.

Otherwise, the methodology is largely unchanged, except that for within-graph MCMC updates, we (?) use a Metropolis-Hastings update, propagated out cumulatively along the junction tree.

References and further reading

Dawid, A. P. and Lauritzen, S. L. (1993) Hyper Markov laws in the statistical analysis of decomposable statistical models. *Annals of Statistics*, **21**, 1272–1317.

Frydenberg, M. and Lauritzen, S. L. (1989) Decomposition of maximum likelihood in mixed interaction models. *Biometrika*, **76**, 539–555.

Giudici, P. and Green, P. J. (1999) Decomposable graphical Gaussian model determination. *Biometrika,* **86**.

Lauritzen, S. L. (1996) *Graphical models*, Oxford University Press.