More on Hyper Markov Laws

Lecture 7
Saint Flour Summerschool, July 13, 2006

Steffen L. Lauritzen, University of Oxford

Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and related algorithms
5. Log-linear and Gaussian graphical models
6. Hyper Markov laws
7. *More on Hyper Markov Laws*
8. Structure estimation and Bayes factors
9. More on structure estimation.

Laws and distributions

A statistical model involves a family $\mathcal{P}$ of distributions, often parametrized as

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$

We typically identify Θ with $\mathcal{P}$ when the parametrization

$$\theta \to P_\theta$$

is one-to-one and onto.

In a Gaussian graphical model, $\theta = K \in \mathcal{S}^+(\mathcal{G})$ is uniquely identifying any regular Gaussian distribution $\mathcal{N}_V(0, \Sigma)$, where $K = \Sigma^{-1}$, satisfying the Markov properties of $\mathcal{G}$.

The case when $\mathcal{P} = \mathcal{P}_{\mathcal{A}}$ is more complex, and a specific parametrization needs to be chosen to make a simple and one-to-one correspondence with a suitable parameter Θ.

A probability measure on $\mathcal{P}$ (or on Θ) represents a random element of $\mathcal{P}$.

We refer to a probability measure on $\mathcal{P}$ or Θ as a *law*, whereas a *distribution* is a probability measure on $\mathcal{X}$.

Thus we shall e.g. speak of the *Wishart law* as we think of W specifying a (random) distribution of X as $\mathcal{N}_V(0 \,|\, W)$.

Hyper Markov Laws

We identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. θ_A for $A \subseteq V$ denotes the marginal distribution of X_A under P_θ and $\theta_{A \mid B}$ the family of conditional distributions of X_A given X_B, etc.

For a law $\mathcal{L}$ on Θ we write

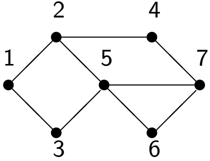
$$A \perp\!\!\!\perp_{\mathcal{L}} B \mid S \iff \theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} \mid \theta_S.$$

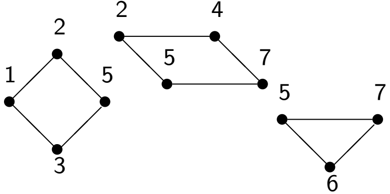
A law $\mathcal{L}$ on Θ is *hyper Markov* w.r.t. $\mathcal{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathcal{G}$;

(ii) $A \perp\!\!\!\perp_{\mathcal{L}} B \mid S$ whenever S is *complete* and $A \perp_{\mathcal{G}} B \mid S$.

Hyper Markov property





If θ follows a hyper Markov law for this graph, it holds for example that

$$\theta_{1235} \perp\!\!\!\perp \theta_{24567} \,|\, \theta_{25}.$$

We shall later see that *this is indeed true for $\hat{\theta} = \hat{p}$ or $\hat{\Sigma}$ in the graphical model with this graph,* i.e.

$$\hat{\Sigma}_{1235} \perp\!\!\!\perp \hat{\Sigma}_{24567} \,|\, \hat{\Sigma}_{25}.$$

Consequences of the hyper Markov property

We have

$$A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S \implies \theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B \,|\, \theta_S,$$

but *the converse is false!*

It generally holds that

$$A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S \iff \theta_{A\,|\,S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B\,|\,S} \,|\, \theta_S.$$

If $\mathcal{G}$ is chordal and $\mathcal{L}$ is hyper Markov on $\mathcal{G}$, it holds that

$$A \perp_{\mathcal{G}} B \,|\, S \implies A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S.$$

In general, if we form $\overline{\mathcal{G}}$ by completing all prime components of $\mathcal{G}$, then *if $\mathcal{L}$ is hyper Markov on $\mathcal{G}$*

$$A \perp_{\overline{\mathcal{G}}} B \,|\, S \implies A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S.$$

Directed hyper Markov property

$\mathcal{L} = \mathcal{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathcal{D}$ if θ is directed Markov on $\mathcal{D}$ for all $\theta \in \Theta$ and

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)},$$

or equivalently

$$\theta_{v \,|\, \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)},$$

or equivalently for a well-ordering of $\mathcal{D}$

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{pr}(v)} \,|\, \theta_{\mathrm{pa}(v)}.$$

If $\mathcal{D}$ is perfect, $\mathcal{L}$ is directed hyper Markov w.r.t. $\mathcal{D}$ if and only if $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$.

Meta independence

In the following we shall for $A, B \subseteq V$ identify

$$\theta_{A \cup B} = (\theta_{B \,|\, A}, \theta_A) = (\theta_{A \,|\, B}, \theta_B),$$

i.e. any joint distribution of $X_{A \cup B}$ is identified with a pair of further marginal and conditional distributions.

Define for $S \subseteq V$ the S-section $\Theta^{\theta^*_S}$ of Θ as

$$\Theta^{\theta^*_S} = \{\theta \in \Theta : \theta_S = \theta^*_S, \theta \in \Theta\}.$$

The *meta independence relation* $\ddagger_{\mathcal{P}}$ is defined as

$$A \ddagger_{\mathcal{P}} B \,|\, S \iff \forall \theta^*_S \in \Theta_S : \Theta^{\theta^*_S} = \Theta^{\theta^*_S}_{A \,|\, S} \times \Theta^{\theta^*_S}_{B \,|\, S},$$

In words, A and B are *meta independent* w.r.t. $\mathcal{P}$ given S, if the pair of conditional distributions $(\theta_{A \mid S}, \theta_{B \mid S})$ vary in a product space when θ_S is fixed.

Equivalently, fixing the values of $\theta_{B \mid S}$ and θ_S places the same restriction on $\theta_{A \mid S}$ as just fixing θ_S.

The relation $\ddagger_{\mathcal{P}}$ satisfies the semigraphoid axioms as it is a special instance of variation independence.

Note also that *for any triple (A, B, S) and any law $\mathcal{L}$ on Θ it holds that*

$$A \perp\!\!\!\perp_{\mathcal{L}} B \mid S \implies A \ddagger_{\mathcal{P}} B \mid S$$

for if $\theta_{A \mid S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \mid S} \mid \theta_S$ it must in particular be true that $(\theta_{A \mid S}, \theta_{B \mid S})$ vary in a product space for every fixed value of θ_S.

Meta Markov models

The family $\mathcal{P}$, or Θ, is said to be *meta Markov* w.r.t. $\mathcal{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathcal{G}$;

(ii) $A \perp_{\mathcal{G}} B \mid S \implies A \ddagger_{\mathcal{P}} B \mid S$ whenever S is complete.

A Markov model is meta Markov if and only if

$$A \perp_{\overline{\mathcal{G}}} B \mid S \implies A \ddagger_{\mathcal{P}} B \mid S,$$

where $\overline{\mathcal{G}}$ is obtained from $\mathcal{G}$ by completing all prime components,

If $\mathcal{G}$ is chordal, $\overline{\mathcal{G}} = \mathcal{G}$ and hence for any meta Markov model $\mathcal{P}$

$$A \perp_{\mathcal{G}} B \mid S \implies A \ddagger_{\mathcal{P}} B \mid S.$$

Hyper Markov laws and meta Markov models

Since it for any law $\mathcal{L}$ on Θ holds that

$$A \perp\!\!\!\perp_{\mathcal{L}} B \mid S \implies A \ddagger_{\mathcal{P}} B \mid S,$$

hyper Markov laws live on meta Markov models: *If a law $\mathcal{L}$ on Θ is hyper Markov w.r.t. $\mathcal{G}$, Θ is meta Markov w.r.t. $\mathcal{G}$.*

In particular, *if a Markov model is not meta Markov, it cannot carry a hyper Markov law without further restricting to $\Theta_0 \subset \Theta$.*

A Gaussian graphical model with graph $\mathcal{G}$ is meta Markov on $\mathcal{G}$.

This follows for example from results of collapsibility of Gaussian graphical models (Frydenberg 1990).

Log-linear meta Markov models

Using results on collapsibility of log-linear models (Asmussen and Edwards 1983) that

A log-linear model $\mathcal{P}_{\mathcal{A}}$ is meta Markov on its dependence graph $\mathcal{G}(\mathcal{A})$ if and only if $S \in \mathcal{A}$ for any minimal complete separator S of $\mathcal{G}(\mathcal{A})$.

In particular, *if $\mathcal{A}$ is conformal, $\mathcal{P}_{\mathcal{A}}$ is meta Markov.*

For example, the log-linear model with generating class

$$\mathcal{A} = \{ab, ac, ad, bc, bd, be, cd, ce, de\}$$

has dependence graph with cliques $\mathcal{C} = \{abcd, bcde\}$. Since the complete separator bcd is not in $\mathcal{A}$, this model is *not* meta Markov.

The model with generating class

$$\mathcal{A}' = \{ab, ac, ad, bcd, be, ce, de\}$$

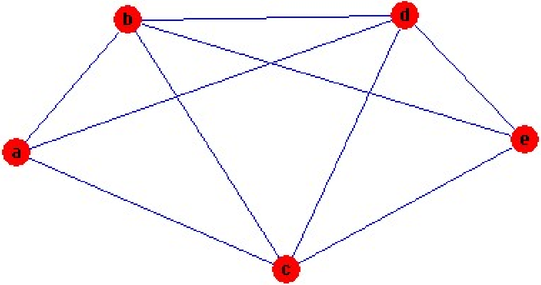
has the same dependence graph $\mathcal{G}(\mathcal{A}') = \mathcal{G}(\mathcal{A})$ but even though $\mathcal{A}'$ is not conformal, $\mathcal{P}_{\mathcal{A}'}$ is meta Markov on $\mathcal{G}(\mathcal{A}')$.

But also the model with generating class

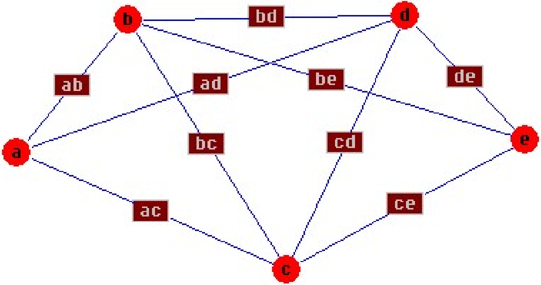
$$\mathcal{A}'' = \{ab, ac, bc, bd, cd, ce, de\}$$

has a different dependence graph $\mathcal{G}(\mathcal{A}'')$. The separator bcd is not in $\mathcal{A}''$, but $\mathcal{P}_{\mathcal{A}''}$ is meta Markov on $\mathcal{G}(\mathcal{A}'')$, as both *minimal* separators bc and cd are in $\mathcal{A}''$.

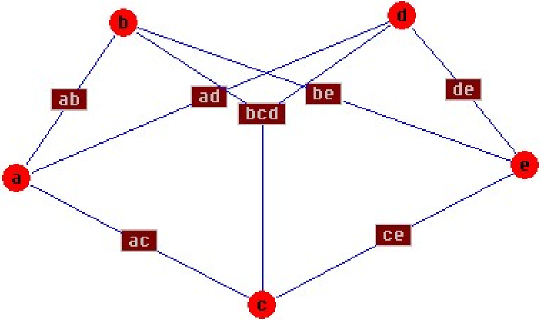
Dependence graph of $\mathcal{A}$ and $\mathcal{A}'$



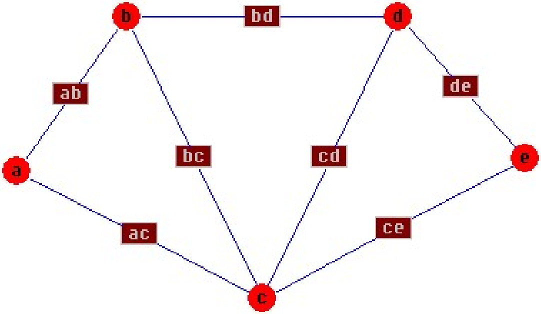
Factor graph of $\mathcal{A}$



Factor graph of $\mathcal{A}'$



Factor graph of $\mathcal{A}''$



Meta Markov properties on supergraphs

Clearly, if θ is globally Markov w.r.t. the graph $\mathcal{G}$, it is also Markov w.r.t. any super graph $\mathcal{G}' = (V, E')$ with $E \subseteq E'$.

The similar fact is *not* true for meta Markov models. For example, the Gaussian graphical model for the 4-cycle $\mathcal{G}$ with adjacencies $1 \sim 2 \sim 3 \sim 4 \sim 1$, is meta Markov on $\mathcal{G}$, because it has no complete separators.

But the same model is *not* meta Markov w.r.t. the larger graph $\mathcal{G}'$ with cliques $\{124, 234\}$, since for any $K \in \mathcal{S}^+(\mathcal{G})$,

$$\sigma_{24} = \frac{\sigma_{12}\sigma_{14}}{\sigma_{11}} + \frac{\sigma_{13}\sigma_{34}}{\sigma_{33}}.$$

So fixing the value of σ_{24} restricts the remaining parameters in a complex way.

Maximum likelihood in meta Markov models

Under certain conditions, *the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under P_θ.* These are

(i) Θ is meta Markov w.r.t. $\mathcal{G}$;

(ii) For any prime component Q of $\mathcal{G}$, the MLE $\hat{\theta}_Q$ for θ_Q based on $X_Q^{(n)}$ is *sufficient* for Θ_Q and *boundedly complete*.

A sufficient condition for (ii) is that Θ_Q *is a full and regular exponential family in the sense of Barndorff-Nielsen (1978).*

In particular, *these conditions are satisfied for any Gaussian graphical model and any meta Markov log-linear model.*

Canonical construction of hyper Markov laws

The distributions of maximum likelihood estimators are important examples of hyper Markov laws. But for *chordal graphs* there is a canonical construction of such laws.

Let $\mathcal{C}$ be the cliques of a chordal graph $\mathcal{G}$ and let $\mathcal{L}_C, C \in \mathcal{C}$ be a family of laws over $\Theta_C \subseteq \mathbb{P}(\mathcal{X}_C)$.

The family of laws are *hyperconsistent* if for any C and D with $C \cap D = S \neq \emptyset$, $\mathcal{L}_C$ and $\mathcal{L}_D$ induce the same law for θ_S.

If $\mathcal{L}_C, C \in \mathcal{C}$ are hyperconsistent, there is a unique hyper Markov law $\mathcal{L}$ over $\mathcal{G}$ with $\mathcal{L}(\theta_C) = \mathcal{L}_C, C \in \mathcal{C}$.

Strong hyper and meta Markov properties

In some cases it is of interest to consider a stronger version of the hyper and meta Markov properties.

A meta Markov model is *strongly meta Markov* if $\theta_{A\,|\,S}\ddagger_{\mathcal{P}}\theta_S$ for all complete separators S.

Similarly, a hyper Markov model is *strongly hyper Markov* if $\theta_{A\,|\,S} \perp\!\!\!\perp_{\mathcal{L}} \theta_S$ for all complete separators S.

A directed hyper Markov model is *strongly directed hyper Markov* if $\theta_{v\,|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{pa}(v)}$ for all $v \in V$.

Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.

Bayesian inference

Parameter $\theta \in \Theta$, data $X = x$, likelihood

$$L(\theta \,|\, x) \propto p(x \,|\, \theta) = \frac{dP_\theta(x)}{d\mu(x)}.$$

Express knowledge about θ through a *prior law* π on θ. Use also π to denote density of the prior law w.r.t. some measure ν on Θ.

Inference about θ from x is then represented through *posterior law* $\pi^*(\theta) = p(\theta \,|\, x)$. Then, from Bayes' formula

$$\pi^*(\theta) = p(x \,|\, \theta)\pi(\theta)/p(x) \propto L(\theta \,|\, x)\pi(\theta)$$

so the *likelihood function is equal to the density of the posterior w.r.t. the prior* modulo a constant.

Bernoulli experiments

Data $X_1 = x_1, \ldots, X_n = x_n$ independent and Bernoulli distributed with parameter θ, i.e.

$$P(X_i = 1 \,|\, \theta) = 1 - P(X_i = 0) = \theta.$$

Use a beta prior:

$$\pi(\theta \,|\, a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

If we let $x = \sum x_i$, we get the posterior:

$$\begin{aligned} \pi^*(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{n-x+b-1} \end{aligned}$$

So the posterior is also beta with parameters $(a + x, b + n - x)$.

Conjugate families

A family $\mathcal{P}$ of laws on Θ is said to be *conjugate* under sampling from x if

$$\pi \in \mathcal{P} \implies \pi^* \in \mathcal{P}.$$

The family of beta laws is conjugate under Bernoulli sampling.

If the family of priors is parametrised:

$$\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$$

we sometimes say that α is a *hyperparameter.* Then, Bayesian inference can be made by just updating hyperparameters. Terminology of hyperparameter breaks down in more complex models.

Conjugacy of hyper Markov properties

If $\mathcal{L}$ is a prior law over Θ and $X = x$ is an observation from θ, $\mathcal{L}^* = \mathcal{L}(\theta \mid X = x)$ denotes the *posterior law* over Θ.

If $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

If $\mathcal{L}$ is strongly hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

In the latter case, *the update of $\mathcal{L}$ is local to prime components*, i.e.

$$\mathcal{L}^*(\theta_Q) = \mathcal{L}^*_Q(\theta_Q) = \mathcal{L}_Q(\theta_Q \mid X_Q = x_Q)$$

and *the marginal distribution p of X is globally Markov w.r.t. $\overline{\mathcal{G}}$*, where

$$p(x) = \int_\Theta P(X = x \mid \theta)\mathcal{L}(d\theta).$$

Conjugate exponential families

For a k-dimensional exponential family

$$p(x \mid \theta) = b(x) e^{\theta^\top t(x) - \psi(\theta)}$$

the *standard conjugate family* is given as

$$\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa\psi(\theta)}$$

for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where $\mathcal{A}$ is determined so that the normalisation constant is finite.

Posterior updating from $(x_1, \ldots, x_n)$ with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.

The family of Beta laws is an example of a standard conjugate family.

Hyper inverse Wishart and Dirichlet laws

Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

$$\pi(K \mid \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}, K \in \mathcal{S}^{+}(\mathcal{G}).$$

These laws are termed *hyper inverse Wishart laws* as Σ follows an inverse Wishart law for complete graphs.

For chordal graphs, each marginal law $\mathcal{L}_C$ of Σ_C is inverse Wishart.

For any meta Markov model where Θ and Θ_Q are full and regular exponential families for all prime components Q, it follows directly from Barndorff-Nielsen (1978), page 149, that *the standard conjugate prior law is strongly hyper Markov w.r.t.* $\mathcal{G}$.

This is in particular true for the hyper inverse Wishart laws.

The analogous prior distribution for log-linear meta Markov models are likewise termed *hyper Dirichlet laws*.

They are also strongly hyper Markov and if $\mathcal{G}$ is *chordal, each induced marginal law* $\mathcal{L}_C$ *is a standard Dirichlet law*.

References

Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika*, **70**, 567–78.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. John Wiley and Sons, New York.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *Annals of Statistics*, **18**, 790–805.