

# Hyper Markov Laws

## Lecture 6

**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.

## Log-linear models

$\mathcal{A}$  denotes a set of (pairwise incomparable) subsets of  $V$ .

A density  $f$  factorizes w.r.t.  $\mathcal{A}$

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x).$$

The set of distributions  $\mathcal{P}_{\mathcal{A}}$  which factorize w.r.t.  $\mathcal{A}$  is the *hierarchical log-linear model* generated by the  $\mathcal{A}$ .

$\mathcal{A}$  is the *generating class* of the log-linear model.

## Dependence graph

The *dependence graph*  $\mathcal{G}(\mathcal{P})$  for a family of distributions  $\mathcal{P}$  is the smallest graph  $\mathcal{G}$  so that

$$\alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathcal{P}.$$

The *dependence graph of a log-linear model*  $\mathcal{P}_{\mathcal{A}}$  is then determined by

$$\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$$

Sets in  $\mathcal{A}$  are complete in  $\mathcal{G}(\mathcal{A})$  and therefore *distributions in  $\mathcal{P}_{\mathcal{A}}$  factorize* according to  $\mathcal{G}(\mathcal{A})$ .

They are also global, local, and pairwise Markov w.r.t.  $\mathcal{G}(\mathcal{A})$ .

## Conformal log-linear model

The set  $\mathcal{C}(\mathcal{G})$  of cliques of  $\mathcal{G}$  is a generating class for the log-linear model of distributions which factorize w.r.t.  $\mathcal{G}$ .

If the dependence graph completely summarizes the restrictions imposed by  $\mathcal{A}$ , i.e. if  $\mathcal{A} = \mathcal{C}(\mathcal{G}(\mathcal{A}))$ ,  $\mathcal{A}$  is *conformal*.

*Conformal log-linear models can be completely described in terms of conditional independence.*

For more general log-linear models *factor graphs* are needed to yield a faithful representation of the factorization. MIM (software by David Edwards [www.hypergraph.dk](http://www.hypergraph.dk)), uses the term *interaction graph*.

## Likelihood equations

For any generating class  $\mathcal{A}$  it holds that *the maximum likelihood estimate  $\hat{p}$  of  $p$  is the unique element of  $\overline{\mathcal{P}_{\mathcal{A}}}$  which satisfies the system of equations*

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a. \quad (1)$$

The system of equations (1) expresses the *fitting of the marginals* in  $\mathcal{A}$ .

In general, the equations cannot be solved explicitly, but iterative methods are needed.

## Iterative Proportional Scaling

For  $a \in \mathcal{A}$  define the *scaling* operation on  $p$ :

$$(T_a p)(x) \leftarrow p(x) \frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}. \quad (2)$$

The operation  $T_a$  fits the  $a$ -marginal. Now, make any ordering of the generators  $\mathcal{A} = \{a_1, \dots, a_k\}$ . Define  $S$  by

$$Sp = T_{a_k} \cdots T_{a_2} T_{a_1} p.$$

Let  $p_0(x) \leftarrow 1/|\mathcal{X}|$ ,  $p_n = Sp_{n-1}$ ,  $n = 1, \dots$

It then holds that  $\lim_{n \rightarrow \infty} p_n = \hat{p}$  where  $\hat{p}$  is the unique maximum likelihood estimate of  $p \in \overline{\mathcal{P}_A}$ .

It is easy to show that  $\hat{p}(x) > 0$  for all  $x \in \mathcal{X}$  if and only if  $\hat{p} \in \mathcal{P}_A$ .

## IPS by probability propagation

A *chordal cover* of  $\mathcal{A}$  is a chordal graph  $\mathcal{G}$  so that for all  $a \in \mathcal{A}$ ,  $a$  are complete subsets of  $\mathcal{G}$ .

1. Find chordal cover  $\mathcal{G}$  of  $\mathcal{A}$  and arrange cliques  $\mathcal{C}$  of  $\mathcal{G}$  in a junction tree;
2. Represent  $p$  *implicitly* as  $p(x) = \frac{\prod_{C \in \mathcal{C}} \psi_C(x)}{\prod_{S \in \mathcal{S}} \psi_S(x)}$ ;
3. Replace (2) with

$$\psi_C(x_C) \leftarrow \psi_C(x_C) \frac{n(x_a)}{np(x_a)}, \quad x_C \in \mathcal{X}_C,$$

where  $a \subseteq C$  and  $p(x_a)$  is calculated by *probability propagation*.



## Closed form maximum likelihood

$\mathcal{A}$  is *decomposable* if  $\mathcal{A} = \mathcal{C}$  where  $\mathcal{C}$  are the cliques of a chordal graph.

*The IPS-algorithm converges after at a finite number of cycles (at most two) if and only if  $\mathcal{A}$  is decomposable.*

*The MLE for  $p$  under the log-linear model  $\mathcal{A} = \mathcal{C}(\mathcal{G})$  is*

$$\hat{p}(x) = \frac{\prod_{C \in \mathcal{C}} n(x_C)}{n \prod_{S \in \mathcal{S}} n(x_S)^{\nu(S)},}$$

where  $\nu(S)$  is the usual multiplicity of a separator.

*In fact, with a suitably chosen ordering (e.g. MCS) of the cliques, the IPS-algorithm converges in a single cycle.*

## Gaussian likelihood function

The likelihood function based on a sample of size  $n$  is

$$L(K) \propto (\det K)^{n/2} e^{-\text{tr}(KW)/2},$$

where  $W$  is the Wishart matrix of sums of squares and products,  $W \sim \mathcal{W}_{|V|}(n, \Sigma)$  with  $\Sigma^{-1} = K \in \mathcal{S}^+(\mathcal{G})$ , where  $\mathcal{S}^+(\mathcal{G})$  are the positive definite matrices with  $\alpha \not\sim \beta \implies k_{\alpha\beta} = 0$ .

*The MLE of  $\hat{K}$  is the unique element of  $\mathcal{S}^+(\mathcal{G})$  satisfying*

$$n\hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in \mathcal{C}(\mathcal{G}).$$

## Iterative Proportional Scaling

For  $K \in \mathcal{S}^+(\mathcal{G})$  and  $c \in \mathcal{C}$ , define the operation of 'adjusting the  $c$ -marginal' as follows. Let  $a = V \setminus c$  and

$$T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \quad (3)$$

Next we choose any ordering  $(c_1, \dots, c_k)$  of the cliques in  $\mathcal{G}$ . Choose further  $K_0 = I$  and define for  $r = 0, 1, \dots$

$$K_{r+1} = (T_{c_1} \cdots T_{c_k})K_r.$$

*It then holds that  $\hat{K} = \lim_{r \rightarrow \infty} K_r$ , provided the maximum likelihood estimate  $\hat{K}$  of  $K$  exists.*

## Chordal graphs

If the graph  $\mathcal{G}$  is chordal, we say that the graphical model is *decomposable*.

In this case, *the IPS-algorithm converges in at most two cycles*, as in the discrete case.

*The maximum likelihood estimates exists if and only if  $n \geq |C|$  for all  $C \in \mathcal{C}$ . Then*

$$\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} [(w_C)^{-1}]^V - \sum_{S \in \mathcal{S}} \nu(S) [(w_S)^{-1}]^V \right\}.$$

the symbol  $[A]^V$  denotes for  $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$  the matrix obtained from  $A$  by filling up with zero entries to obtain full dimension.

## Existence of the MLE

The general problem of existence of the MLE is non-trivial:

*If  $n < \sup_{a \in \mathcal{A}} |a|$  the MLE does not exist.*

*If  $n \geq \sup_{C \in \mathcal{C}} |C|$ , where  $\mathcal{C}$  are the cliques of a chordal cover of  $\mathcal{A}$  the MLE exists with probability one.*

For  $n$  between these values the general situation is unclear.

For the  $k$ -cycle it holds (Buhl 1993) that for  $n = 2$ ,

$$P\{\text{MLE exists} \mid \Sigma = I\} = 1 - \frac{2}{(k-1)!},$$

whereas for  $n = 1$  the MLE does not exist and for  $n \geq 3$  the MLE exists with probability one, as a  $k$ -cycle has a chordal cover with maximal clique size 3.

## Special Wishart distributions

The formula

$$\hat{\Sigma} = n \left\{ \sum_{C \in \mathcal{C}} [(W_C)^{-1}]^V - \sum_{S \in \mathcal{S}} \nu(S) [(W_S)^{-1}]^V \right\}^{-1}$$

specifies  $\hat{\Sigma}$  as a random matrix.

The distribution of this random Wishart-type matrix is partly reflecting Markov properties of the graph  $\mathcal{G}$ .

This is also true for the distribution of  $\hat{\Sigma}$  for a non-chordal graph  $\mathcal{G}$  but not to the same degree.

Before we delve further into this, we shall need some more terminology.

## Laws and distributions

Families of distributions may not always be simply parameterized, or we may want to describe the families without specific reference to a parametrization.

Generally we think of

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

and sometimes identify  $\Theta$  with  $\mathcal{P}$  which is justified when the parametrization

$$\theta \rightarrow P_\theta$$

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 satisfying the Markov properties w.r.t.  $\mathcal{G}$ .

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.

In any case, any probability measure on  $\mathcal{P}$  (or on  $\Theta$ ) represents a random element of  $\mathcal{P}$ , i.e. a random distribution. The *sampling distribution of the MLE*  $\hat{p}$  is an example of such a measure.

To keep heads straight we refer to a probability measure on  $\mathcal{P}$  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 it specifying a distribution of  $f(\cdot | \Sigma)$ .



## Hyper Markov Laws

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

For a law  $\mathcal{L}$  on  $\Theta$  we write

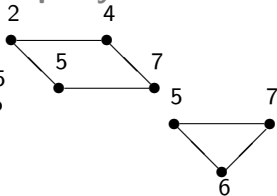
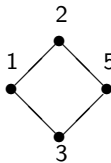
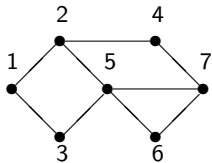
$$A \perp\!\!\!\perp_{\mathcal{L}} B | S \iff \theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} | \theta_S.$$

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

- (i) All  $\theta \in \Theta$  are globally Markov w.r.t.  $\mathcal{G}$ ;
- (ii)  $A \perp\!\!\!\perp_{\mathcal{L}} B | S$  whenever  $S$  is complete and  $A \perp_{\mathcal{G}} B | S$ .

Note the conditional independence is only required to hold for *graph decompositions*.

## Hyper Markov property



If  $\theta$  follows a hyper Markov law for this graph, it holds for example that

$$\theta_{1235} \perp\!\!\!\perp \theta_{24567} \mid \theta_{25}.$$

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

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

## Consequences of the hyper Markov property

Clearly, if  $A \perp\!\!\!\perp_{\mathcal{L}} B \mid S$ , we have for example also (using property (C2) of conditional independence)

$$\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B \mid \theta_S$$

since  $\theta_A$  and  $\theta_B$  are functions of  $\theta_{A \cup S}$  and  $\theta_{B \cup S}$  respectively.

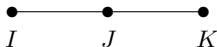
But *the converse is false!*  $\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B \mid \theta_S$  does *not* imply  $\theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} \mid \theta_S$ , since  $\theta_{A \cup S}$  is *not* a function of  $(\theta_A, \theta_S)$ . In contrast,  $X_{A \cup B}$  is indeed a (one-to-one) function of  $(X_A, X_B)$ .

However *it generally holds* that

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

## Simple example

Consider the conditional independence model with graph



Here the MLE based on data  $X^{(n)} = (X^1, \dots, X^n)$  is

$$\hat{p}_{ijk} = \frac{N_{ij+}N_{+jk}}{nN_{+j+}}$$

and

$$\hat{p}_{ij+} = \frac{N_{ij+}}{n}, \quad \hat{p}_{+jk} = \frac{N_{+jk}}{n}, \quad \hat{p}_{+j+} = \frac{N_{+j+}}{n}.$$

Clearly, it holds that  $\hat{p}$  is Markov on  $\mathcal{G}$  and

$$\{N_{ij+}\} \perp\!\!\!\perp \{N_{+jk}\} \mid \{X_j^{(n)}\}.$$

But since e.g.

$$P(\{N_{ij+} = n_{ij}\} \mid \{X_j^{(n)}\}) = \prod_j \left( \frac{n_{+j+}!}{\prod_i n_{ij+}!} \prod_i p_{ij+}^{n_{ij+}} \right),$$

we have

$$\{N_{ij+}\} \perp\!\!\!\perp \{X_j^{(n)}\} \mid \{N_{+j+}\}$$

and hence

$$\{N_{ij+}\} \perp\!\!\!\perp \{N_{+jk}\} \mid \{N_{+j+}\},$$

which yields the hyper Markov property.

## Chordal graphs

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

$$A \perp_{\mathcal{G}} B | S \implies A \perp_{\mathcal{L}} B | S$$

i.e. it is not necessary to specify that  $S$  is a complete separator to obtain the relevant conditional independence.

This follows essentially because for a chordal graph it holds that

$$A \perp_{\mathcal{G}} B | S \implies \exists S^* \subseteq S : A \perp_{\mathcal{G}} B | S^* \text{ with } S^* \text{ complete.}$$

If  $\mathcal{G}$  is not chordal, we can form  $\overline{\mathcal{G}}$  by completing all prime components of  $\mathcal{G}$ .

Then if  $\theta$  is hyper Markov on  $\mathcal{G}$ , it is also hyper Markov on  $\overline{\mathcal{G}}$ , and thus

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

But the similar result would be *false* for an arbitrary chordal cover of  $\mathcal{G}$ .

## Directed hyper Markov property

We have similar notions and results in the directed case.

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

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

or equivalently  $\theta_{v \mid \text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{nd}(v)} \mid \theta_{\text{pa}(v)}$ , or equivalently for a well-ordering

$$\theta_{v \cup \text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{pr}(v)} \mid \theta_{\text{pa}(v)}.$$

In general there is no similar statement corresponding to the global property and  $d$ -separation.

However, 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.  $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$ .



## References

- Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, **20**, 263–70.