Log-Linear and Gaussian Graphical Models

Lecture 5
Saint Flour Summerschool, July 10, 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. Conjugate prior families for graphical models
7. Hyper Markov laws
8. Structure learning and Bayes factors
9. More on structure learning.

Log–linear models

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

A density f (or function) *factorizes* w.r.t. $\mathcal{A}$ if there exist functions $\psi_a(x)$ which depend on x_a only and

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

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

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

No specific need to demand sets in $\mathcal{A}$ to be incomparable.
Only to avoid redundancy.

Traditional notation

Traditionally used for contingency tables, where e.g. m_{ijk} denotes the mean of the counts N_{ijk} in the cell (i, j, k) which has then been expanded as e.g.

$$\log m_{ijk} = \alpha_i + \beta_j + \gamma_k \tag{1}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} \tag{2}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \tag{3}$$

or (with redundancy)

$$\log m_{ijk} = \gamma + \delta_i + \phi_j + \eta_k + \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \tag{4}$$

etc.

Connecting to tradition

This largely a matter of different notation.

Assume data $X^1 = x^1, \ldots, X^n = x^n$ and $V = \{I, J, K\}$.

Write $i = 1, \ldots, |I|$ for the possible values of X_I etc. and

$$N_{ijk} = |\{\nu : x^\nu = (i, j, k)\}|,$$

etc. Then $m_{ijk} = nf(x)$ and *if* $f(x) > 0$ and factorizes w.r.t. $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$

$$\log f(x) = \log \psi_{IJ}(x_I, x_J) + \log \psi_{JK}(x_J, x_K).$$

Thus if we let

$$\alpha_{ij} = \log n + \log \psi_{IJ}(x_I, x_J), \quad \beta_{jk} = \log \psi_{JK}(x_J, x_K)$$

we have

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk}.$$

The only difference is the assumption of positivity which is not necessary when using the multiplicative definition.

It is typically an advantage to relax the restriction of positivity although it also creates technical difficulties.

The logarithm of the factors $\phi_a = \log \psi_a$ are known as *interaction terms of order* $|a| - 1$ or $|a|$*-factor interactions*.

Interaction terms of 0th order are called *main effects*.

We also refer to the factors themselves using the same terms.

Dependence graph

Any joint probability distribution P of $X = (X_v, v \in V)$ has a *dependence graph* $G = G(P) = (V, E(P))$.

This is defined by letting $\alpha \not\sim \beta$ in $G(P)$ exactly when

$$\alpha \perp\!\!\!\perp_P \beta \,|\, V \setminus \{\alpha, \beta\}.$$

X will then satisfy the pairwise Markov w.r.t. $G(P)$ and $G(P)$ is smallest with this property, i.e. *P is pairwise Markov w.r.t. $\mathcal{G}$ iff*

$$G(P) \subseteq \mathcal{G}.$$

The *dependence graph $G(\mathcal{P})$ for a family $\mathcal{P}$* is the smallest graph $\mathcal{G}$ so that all $P \in \mathcal{P}$ are pairwise Markov w.r.t. $\mathcal{G}$:

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

Dependence graph of log–linear model

For any generating class $\mathcal{A}$ we construct the dependence graph $G(\mathcal{A}) = G(\mathcal{P}_{\mathcal{A}})$ of the log–linear model $\mathcal{P}_{\mathcal{A}}$.

This is determined by the relation

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

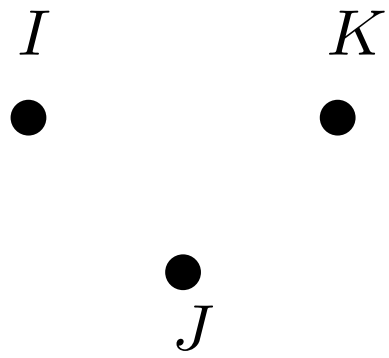
Sets in $\mathcal{A}$ are clearly complete in $G(\mathcal{A})$ and therefore *distributions in $\mathcal{P}_{\mathcal{A}}$ factorize according to $G(\mathcal{A})$*.

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

Independence

The log–linear model specified by (1) is known as the *main effects model.*

It has generating class consisting of singletons only $\mathcal{A} = \{\{I\}, \{J\}, \{K\}\}$. It has dependence graph

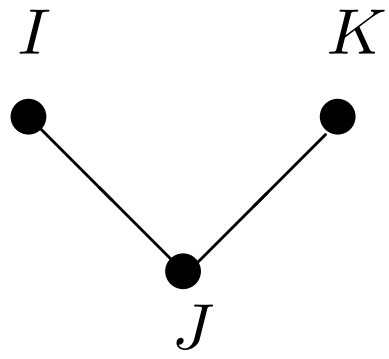


Thus it corresponds to *complete independence*.

Conditional independence

The log–linear model specified by (2) has no interaction between I and K.

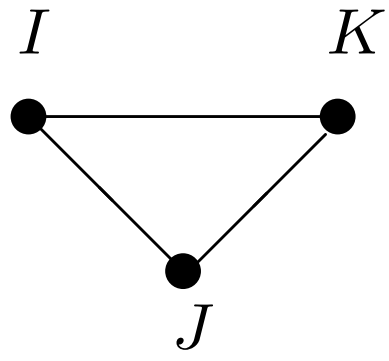
It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$ and dependence graph



Thus it corresponds to the *conditional independence* $I \perp\!\!\!\perp K \,|\, J$.

No interaction of second order

The log–linear model specified by (3) has no second-order interaction. It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}, \{I, K\}\}$ and its dependence graph



is the complete graph. Thus it has no conditional independence interpretation.

Conformal log-linear models

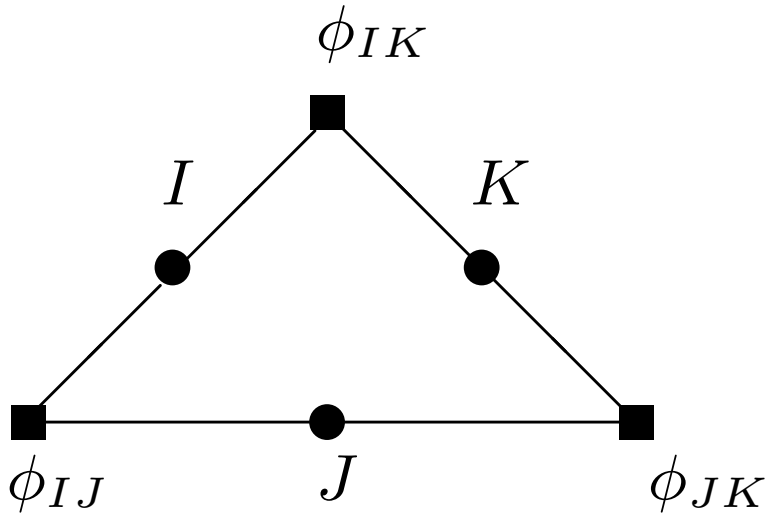
As a generating class defines a dependence graph $G(\mathcal{A})$, the reverse is also true.

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 $\mathcal{A} = \mathcal{C}(G(\mathcal{A}))$, $\mathcal{A}$ is *conformal*.

The generating classes for the models given by (1) and (2) are conformal, whereas this is not the case for (3).

Factor graphs



The *factor graph* of $\mathcal{A}$ is the bipartite graph with vertices $V \cup \mathcal{A}$ and edges define by

$$\alpha \sim a \iff \alpha \in a.$$

Using this graph even non-conformal log–linear models admit a simple visual representation.

Separation in factor graphs

If $\mathcal{F} = F(\mathcal{A})$ is the factor graph for $\mathcal{A}$ and $\mathcal{G} = G(\mathcal{A})$ the corresponding dependence graph, it is not difficult to see that for A, B, S being subsets of V

$$A \perp_{\mathcal{G}} B \,|\, S \iff A \perp_{\mathcal{F}} B \,|\, S$$

and hence conditional independence properties can be read directly off the factor graph also.

In that sense, the factor graph is more informative than the dependence graph.

Data in list form

Consider a sample $X^1 = x^1, \ldots, X^n = x^n$ from a distribution with probability mass function p. We refer to such data as being in *list form*, e.g. as

case	Admitted	Sex
1	Yes	Male
2	Yes	Female
3	No	Male
4	Yes	Male
⋮	⋮	⋮

Contingency Table

Data often presented in the form of a *contingency table* or *cross-classification*, obtained from the list by sorting according to category:

	Sex	
Admitted	Male	Female
Yes	1198	557
No	1493	1278

The numerical entries are *cell counts*

$$n(x) = |\{\nu : x^\nu = x\}|$$

and the total number of observations is $n = \sum_{x \in \mathcal{X}} n_(x)$.

Likelihood function

Assume now $p \in \mathcal{P}_{\mathcal{A}}$ but otherwise unknown. The likelihood function can be expressed as

$$L(p) = \prod_{\nu=1}^{n} p(x^{\nu}) = \prod_{x \in \mathcal{X}} p(x)^{n(x)}.$$

In contingency table form the data follow a multinomial distribution

$$P\{N(x) = n(x), x \in \mathcal{X}\} = \frac{n!}{\prod_{x \in X} n(x)!} \prod_{x \in \mathcal{X}} p(x)^{n(x)}$$

but this only affects the likelihood function by a constant factor.

Properties of the likelihood function

The likelihood function

$$L(p) = \prod_{x \in \mathcal{X}} p(x)^{n(x)},$$

is continuous as a function of the ($|\mathcal{X}|$-dimensional vector) unknown probability distribution p.

Since the *closure* $\overline{\mathcal{P}_{\mathcal{A}}}$ is compact (bounded and closed), L *attains its maximum on* $\overline{\mathcal{P}_{\mathcal{A}}}$ (not necessarily on $\mathcal{P}_{\mathcal{A}}$ itself).

Indeed, it is also true that L *has a unique maximum over* $\overline{\mathcal{P}_{\mathcal{A}}}$, essentially because the likelihood function is log-concave.

Uniqueness of the MLE

For simplicity, we only establish uniqueness within $\mathcal{P}_{\mathcal{A}}$. The proof is indirect.

Assume $p_1, p_2 \in \mathcal{P}_{\mathcal{A}}$ with $p_1 \neq p_2$ and

$$L(p_1) = L(p_2) = \sup_{p \in \mathcal{P}_{\mathcal{A}}} L(p). \tag{5}$$

Define

$$p_{12}(x) = c\sqrt{p_1(x)p_2(x)},$$

where $c^{-1} = \{\sum_x \sqrt{p_1(x)p_2(x)}\}$ is a normalizing constant.

Then $p_{12} \in \mathcal{P}_{\mathcal{A}}$ because

$$\begin{aligned} p_{12}(x) &= c\sqrt{p_1(x)p_2(x)} \\ &= c\prod_{a\in\mathcal{A}} \sqrt{\psi_a^1(x)\psi_a^2(x)} = \prod_{a\in\mathcal{A}} \psi_a^{12}(x), \end{aligned}$$

where e.g. $\psi_a^{12} = c^{1/|\mathcal{A}|}\sqrt{\psi_a^1(x)\psi_a^2(x)}$.

The Cauchy–Schwarz inequality yields

$$c^{-1} = \sum_x \sqrt{p_1(x)p_2(x)} < \sqrt{\sum_x p_1(x)}\sqrt{\sum_x p_2(x)} = 1.$$

Hence

$$\begin{aligned} L(p_{12}) &= \prod_x p_{12}(x)^{n(x)} \\ &= \prod_x \left\{ c\{\sqrt{p_1(x)p_2(x)} \right\}^{n(x)} \\ &= c^n \prod_x \sqrt{p_1(x)}^{n(x)} \prod_x \sqrt{p_2(x)}^{n(x)} \\ &= c^n \sqrt{L(p_1)L(p_2)} \\ &> \sqrt{L(p_1)L(p_2)} = L(p_1) = L(p_2), \end{aligned}$$

which contradicts (5). Hence we conclude $p_1 = p_2$.

The extension to $\overline{\mathcal{P}_{\mathcal{A}}}$ is almost identical. It just needs a limit argument to establish $p_1, p_2 \in \overline{\mathcal{P}_{\mathcal{A}}} \implies p_{12} \in \overline{\mathcal{P}_{\mathcal{A}}}$.

Likelihood equations

The maximum likelihood estimate $\hat{p}$ of p is the unique element of $\overline{\mathcal{P}_{\mathcal{A}}}$ which satisfies the system of equations

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a. \tag{6}$$

Here $g(x_a) = \sum_{y:y_a=x_a} g(y)$ is the *a-marginal* of the function g.

The system of equations (6) expresses the *fitting of the marginals* in $\mathcal{A}$.

This is also an instance of the familiar result that in an exponential family (log-linear $\sim$ exponential), the MLE is found by equating the sufficient statistics (marginal counts) to their expectation.

Proportional scaling

To show that the equations (6) indeed have a solution, we simply describe a convergent algorithm which solves it. This cycles (repeatedly) through all the a-marginals in $\mathcal{A}$ and fit them one by one.

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

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

where $0/0 = 0$ and $b/0$ is undefined if $b \neq 0$.

Fitting the marginals

The operation T_a *fits the a-marginal* if $p(x_a) > 0$ when $n(x_a) > 0$:

$$\begin{aligned} n(T_a p)(x_a) &= n \sum_{y : y_a = x_a} p(y) \frac{n(y_a)}{np(y_a)} \\ &= n \frac{n(x_a)}{np(x_a)} \sum_{y : y_a = x_a} p(y) \\ &= n \frac{n(x_a)}{np(x_a)} p(x_a) = n(x_a). \end{aligned}$$

Iterative Proportional Scaling

Make an ordering of the generators $\mathcal{A} = \{a_1, \ldots, a_k\}$.
Define S by a full cycle of scalings

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

Define the iteration

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

It then holds that

$$\lim_{n \to \infty} p_n = \hat{p}$$

where $\hat{p}$ is the unique maximum likelihood estimate of $p \in \overline{\mathcal{P}_{\mathcal{A}}}$, i.e. the solution of the equation system (6).

Iterative Proportional Fitting

Known as the *IPS*-algorithm or *IPF*-algorithm, or as a variety of other names. Implemented e.g. (inefficiently) in *R* in `loglin` with front end `loglm` in MASS.

Key elements in proof:

1. If $p \in \overline{\mathcal{P}_{\mathcal{A}}}$, so is $T_a p$;
2. T_a is continuous at any point p of $\overline{\mathcal{P}_{\mathcal{A}}}$ with $p(x_a) \neq 0$ whenever $n(x_a) = 0$;
3. $L(T_a p) \geq L(p)$ so likelihood always increases;
4. $\hat{p}$ is the unique fixpoint for T (and S);
5. $\overline{\mathcal{P}_{\mathcal{A}}}$ is compact.

A simple example

	Admitted		
Sex	Yes	No	S-marginal
Male	1198	1493	2691
Female	557	1278	1835
A-marginal	1755	2771	4526

Admissions data from Berkeley. Consider $A \perp\!\!\!\perp S$, corresponding to $\mathcal{A} = \{\{A\}, \{S\}\}$.

We should fit A-marginal and S-marginal iteratively.

Initial values

	Admitted		
Sex	Yes	No	S-marginal
Male	1131.5	1131.5	2691
Female	1131.5	1131.5	1835
A-marginal	1755	2771	4526

Entries all equal to $4526/4$. Gives initial values of np_0.

Fitting S-marginal

	Admitted		
Sex	Yes	No	S-marginal
Male	1345.5	1345.5	2691
Female	917.5	917.5	1835
A-marginal	1755	2771	4526

For example

$$1345.5 = 1131.5 \frac{2691}{1131.5 + 1131.5}$$

and so on.

Fitting A-marginal

Sex	Admitted: Yes	Admitted: No	S-marginal
Male	1043.46	1647.54	2691
Female	711.54	1123.46	1835
A-marginal	1755	2771	4526

For example

$$711.54 = 917.5 \frac{1755}{917.5 + 1345.5}$$

and so on.

Algorithm has converged, as both marginals now fit!

Normalised to probabilities

	Admitted		
Sex	Yes	No	S-marginal
Male	0.231	0.364	0.595
Female	0.157	0.248	0.405
A-marginal	0.388	0.612	1

Dividing everything by 4526 yields $\hat{p}$.

It is overkill to use the IPS algorithm as there is an explicit formula in this case.

IPS by probability propagation

The IPS-algorithm performs the scaling operations T_a:

$$p(x) \leftarrow p(x)\frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}. \tag{7}$$

This moves through all possible values of $x \in \mathcal{X}$, which in general can be *huge*, hence impossible.

Jiroušek and Přeučil (1995) realized that the algorithm could be implemented using probability propagation:

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

1. Find chordal cover $\mathcal{G}$ of $\mathcal{A}$;

2. Arrange cliques $\mathcal{C}$ of $\mathcal{G}$ in a junction tree;

3. Represent p *implicitly* as

$$p(x) = \frac{\prod_{C \in \mathcal{C}} \psi_C(x)}{\prod_{S \in \mathcal{S}} \psi_S(x)};$$

4. Replace the step (7) 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*.

Since the scaling only involves $\mathcal{X}_C$, this is possible just if $\max_{C \in \mathcal{C}} |\mathcal{X}_C|$ is of a reasonable size.

Closed form maximum likelihood

In some cases the IPS algorithm converges after a finite number of cycles.

An explicit formula is then available for the MLE of $p \in \mathcal{P}_{\mathcal{A}}$.

A generating class $\mathcal{A}$ is called *decomposable* if $\mathcal{A} = \mathcal{C}$ (i.e. $\mathcal{A}$ is conformal) and $\mathcal{C}$ are the cliques of a chordal graph $\mathcal{G}$.

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

$\mathcal{A} = \{\{1,2\},\{2,3\},\{1,3\}\}$ is the smallest non-conformal generating class, demanding proper iteration.

Explicit formula for MLE

Let $\mathcal{S}$ be the set of *minimal separators* of the chordal graph $\mathcal{G}$. The MLE for p under the log-linear model with generating class $\mathcal{A} = \mathcal{C}(\mathcal{G})$ 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 number of times S appears as an intersection $a \cap b$ of neighbours in a junction tree $\mathcal{T}$ with $\mathcal{A}$ as vertex set.

Contrast this with the factorization of the probability function itself:

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

Density of multivariate Gaussian

If Σ is *positive definite*, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density w.r.t. Lebesgue measure on $\mathcal{R}^d$

$$f(x \,|\, \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2}, \quad (8)$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. We then also say that Σ is *regular*.

Marginal and conditional distributions

Partition X into X_1 and X_2, where $X_1 \in \mathcal{R}^r$ and $X_2 \in \mathcal{R}^s$ with $r + s = d$.

Partition mean vector, concentration and covariance matrix accordingly as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

so that Σ_{11} is $r \times r$ and so on. *Then, if* $X \sim \mathcal{N}_d(\xi, \Sigma)$

$$X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22})$$

and

$$X_1 \,|\, X_2 = x_2 \sim \mathcal{N}_r(\xi_{1|2}, \Sigma_{1|2}),$$

where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-}\Sigma_{21}.$$

Σ_{22}^{-} is an arbitrary generalized inverse to Σ_{22}.

In the regular case it also holds that

$$K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{9}$$

and

$$K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}, \tag{10}$$

so then,

$$\xi_{1|2} = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = K_{11}^{-1}.$$

In particular, *if $\Sigma_{12} = 0$, X_1 and X_2 are independent.*

Gaussian likelihoods

Consider $\xi = 0$ and a sample $X^1 = x^1, \ldots, X^n = x^n$ $\mathcal{N}_d(0, \Sigma)$ with Σ regular.

Using (8), we get the likelihood function

$$\begin{aligned} L(K) &= (2\pi)^{-nd/2} (\det K)^{n/2} e^{-\sum_{\nu=1}^n (x^\nu)^\top K x^\nu/2} \\ &\propto (\det K)^{n/2} e^{-\operatorname{tr}\{K \sum_{\nu=1}^n x^\nu (x^\nu)^\top\}/2} \\ &= (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2}. \end{aligned} \tag{11}$$

where

$$W = \sum_{\nu=1}^n x^\nu (x^\nu)^\top$$

is the matrix of *sums of squares and products*.

Wishart distribution

The Wishart distribution is the sampling distribution of the matrix of sums of squares and products. More precisely:

A random $d \times d$ matrix S has a *d-dimensional Wishart distribution* with parameter Σ and n *degrees of freedom* if

$$W \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} X^{\nu} (X^{\nu})^{\top}$$

where $X^{\nu} \sim \mathcal{N}_d(0, \Sigma)$. We then write

$$W \sim \mathcal{W}_d(n, \Sigma).$$

The Wishart is the multivariate analogue to the χ^2:

$$\mathcal{W}_1(n, \sigma^2) = \sigma^2 \chi^2(n).$$

If $W \sim \mathcal{W}_d(n, \Sigma)$ its mean is $\mathbf{E}(W) = n\Sigma$.

If W_1 and W_2 are independent with $W_i \sim \mathcal{W}_d(n_i, \Sigma)$, then

$$W_1 + W_2 \sim \mathcal{W}_d(n_1 + n_2, \Sigma).$$

If A is an $r \times d$ matrix and $W \sim \mathcal{W}_d(n, \Sigma)$, then

$$AWA^\top \sim \mathcal{W}_r(n, A\Sigma A^\top).$$

For $r = 1$ we get that when $W \sim \mathcal{W}_d(n, \Sigma)$ and $\lambda \in R^d$,

$$\lambda^\top W \lambda \sim \sigma_\lambda^2 \chi^2(n),$$

where $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$.

Wishart density

If $W \sim \mathcal{W}_d(n, \Sigma)$, where Σ is regular, then

W is regular with probability one if and only if $n \geq d$.

When $n \geq d$ the Wishart distribution has density

$$\begin{aligned} & f_d(w \,|\, n, \Sigma) \\ & \quad = \quad c(d,n)^{-1} (\det \Sigma)^{-n/2} (\det w)^{(n-d-1)/2} e^{-\operatorname{tr}(\Sigma^{-1} w)/2} \end{aligned}$$

w.r.t. Lebesgue measure on the set of positive definite matrices.

The Wishart constant $c(d,n)$ is

$$c(d,n) = 2^{nd/2} (2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(n+1-i)/2\}.$$

Conditional independence

Consider $X = (X_1, \ldots, X_V) \sim \mathcal{N}_{|V|}(0, \Sigma)$ with Σ regular and $K = \Sigma^{-1}$.

The concentration matrix of the conditional distribution of (X_α, X_β) given $X_{V \setminus \{\alpha,\beta\}}$ is

$$K_{\{\alpha,\beta\}} = \begin{pmatrix} k_{\alpha\alpha} & k_{\alpha\beta} \\ k_{\beta\alpha} & k_{\beta\beta} \end{pmatrix}.$$

Hence

$$\alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha, \beta\} \iff k_{\alpha\beta} = 0.$$

Thus *the dependence graph $\mathcal{G}(K)$ of a regular Gaussian distribution is given by*

$$\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$$

Graphical models

$\mathcal{S}(\mathcal{G})$ denotes the symmetric matrices A with $a_{\alpha\beta} = 0$ unless $\alpha \sim \beta$ and $\mathcal{S}^+(\mathcal{G})$ their positive definite elements.

A *Gaussian graphical model* for X specifies X as multivariate normal with $K \in \mathcal{S}^+(\mathcal{G})$ and otherwise unknown.

Note that the density then factorizes as

$$\log f(x) = \text{constant} - \frac{1}{2} \sum_{\alpha \in V} k_{\alpha\alpha} x_\alpha^2 - \sum_{\{\alpha,\beta\} \in E} k_{\alpha\beta} x_\alpha x_\beta,$$

hence *no interaction terms involve more than pairs.*.

This is different from the discrete case and generally makes things easier.

Likelihood function

The likelihood function based on a sample of size n is

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

where W is the Wishart matrix of sums of squares and products, $W \sim \mathcal{W}_{|V|}(n, \Sigma)$ with $\Sigma^{-1} = K \in \mathcal{S}^+(\mathcal{G})$.

For any matrix A we let $A(\mathcal{G}) = \{a(\mathcal{G})_{\alpha\beta}\}$ where

$$a(\mathcal{G})_{\alpha\beta} = \begin{cases} a_{\alpha\beta} & \text{if } \alpha = \beta \text{ or } \alpha \sim \beta \\ 0 & \text{otherwise.} \end{cases}$$

Then, as $K \in \mathcal{S}(\mathcal{G})$

$$\operatorname{tr}(KW) = \operatorname{tr}\{KW(\mathcal{G})\}.$$

Hence we can identify the family as a (regular and canonical) exponential family with elements of $W(\mathcal{G})$ as canonical sufficient statistics and the likelihood equations

$$\mathbf{E}\{W(\mathcal{G})\} = n\Sigma(\mathcal{G}) = w(\mathcal{G})_{\text{obs}}.$$

Alternatively we can write the equations as

$$n\hat{\sigma}_{vv} = w_{vv}, \quad n\hat{\sigma}_{\alpha\beta} = w_{\alpha\beta}, \quad v \in V, \{\alpha, \beta\} \in E,$$

with the model restriction $\Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$.

This 'fits variances and covariances along nodes and edges in $\mathcal{G}$' so we can write the equations as

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

hence making the equations analogous to the discrete case.

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

$$T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \quad (12)$$

This operation is clearly well defined if w_{cc} is positive definite.

Exploiting that it holds in general that

$$(K^{-1})_{cc} = \Sigma_{cc} = \left\{K_{cc} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1},$$

we find the covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted

concentration matrix becomes

$$\begin{aligned}\tilde{\Sigma}_{cc} &= \{(T_c K)^{-1}\}_{cc} \\ &= \left\{n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1} \\ &= w_{cc}/n,\end{aligned}$$

hence $T_c K$ *does indeed adjust the marginals*.

From (12) it is seen that the pattern of zeros in K is preserved under the operation T_c, and it can also be seen to stay positive definite.

In fact, T_c *scales proportionally* in the sense that

$$f\{x \,|\, (T_c K)^{-1}\} = f(x \,|\, K^{-1}) \frac{f(x_c \,|\, w_{cc}/n)}{f(x_c \,|\, \Sigma_{cc})}.$$

This clearly demonstrates the analogy to the discrete case.

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

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

Then we have: *Consider a sample from a covariance selection model with graph $\mathcal{G}$. Then*

$$\hat{K} = \lim_{r \to \infty} K_r,$$

provided the maximum likelihood estimate $\hat{K}$ of K exists.

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 $n = 2$,

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

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.

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 a finite number of steps*, as in the discrete case.

We also have the familiar *factorization of densities*

$$f(x \,|\, \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \,|\, \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \,|\, \Sigma_S)^{\nu(S)}} \tag{13}$$

where $\nu(S)$ is the number of times S appear as intersection between neighbouring cliques of a junction tree for $\mathcal{C}$.

Relations for trace and determinant

Using the factorization (13) we can match the expressions for the trace and determinant to obtain

$$\operatorname{tr}(KW) = \sum_{C \in \mathcal{C}} \operatorname{tr}(K_C W_C) - \sum_{S \in \mathcal{S}} \nu(S) \operatorname{tr}(K_S W_S)$$

and further

$$\begin{aligned} \det \Sigma &= \{\det(K)\}^{-1} = \frac{\prod_{C \in \mathcal{C}} \det\{(K^{-1})_C\}}{\prod_{S \in \mathcal{S}} [\det\{(K^{-1})_S\}]^{\nu(S)}} \\ &= \frac{\prod_{C \in \mathcal{C}} \det\{\Sigma_C\}}{\prod_{S \in \mathcal{S}} \{\det(\Sigma_S)\}^{\nu(S)}} \end{aligned}$$

Maximum likelihood estimates

For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V| \times |V|$, i.e.

$$\left([A]^V\right)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} & \text{if } \gamma \in d, \mu \in e \\ 0 & \text{otherwise.} \end{cases}$$

The maximum likelihood estimates exists if and only if $n \geq C$ for all $C \in \mathcal{C}$. Then the following simple formula holds for the maximum likelihood estimate of K:

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

The determinant of the MLE is

$$\det(\hat{K}) = \frac{\prod_{S \in \mathcal{S}} \{\det(w_S)\}^{\nu(S)}}{\prod_{C \in \mathcal{C}} \det(w_C)} n^{|V|}.$$

References

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

Jiroušek, R. and Přeučil, R. (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics and Data Analysis*, **19**, 177–89.