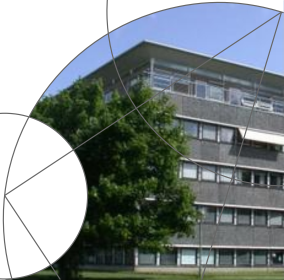
Likelihood Analysis of Gaussian Graphical Models

Steffen Lauritzen
Department of Mathematical Sciences



Overview of lectures

Lecture 1 Markov Properties and the Multivariate Gaussian Distribution

Lecture 2 *Likelihood Analysis of Gaussian Graphical Models*

Lecture 3 Gaussian Graphical Models with Additional Restrictions; structure identification.

For reference, if nothing else is mentioned, see Lauritzen (1996), Chapters 3 and 4.

Fundamental properties

For random variables X, Y, Z, and W it holds

(C1) If $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$;

(C2) If $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \,|\, Z$;

(C3) If $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \,|\, (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \,|\, Z$;

If density w.r.t. product measure $f(x, y, z, w) > 0$ also

(C5) If $X \perp\!\!\!\perp Y \,|\, (Z, W)$ and $X \perp\!\!\!\perp Z \,|\, (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \,|\, W$.

Semi-graphoid

An *independence model* (Studený, 2005) $\perp_\sigma$ is a ternary relation over subsets of a finite set V. It is a *graphoid* if for all disjoint subsets A, B, C, D:

(S1) if $A \perp_\sigma B \mid C$ then $B \perp_\sigma A \mid C$ (symmetry);

(S2) if $A \perp_\sigma (B \cup D) \mid C$ then $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid C$ (decomposition);

(S3) if $A \perp_\sigma (B \cup D) \mid C$ then $A \perp_\sigma B \mid (C \cup D)$ (weak union);

(S4) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid (B \cup C)$, then $A \perp_\sigma (B \cup D) \mid C$ (contraction);

(S5) if $A \perp_\sigma B \mid (C \cup D)$ and $A \perp_\sigma C \mid (B \cup D)$ then $A \perp_\sigma (B \cup C) \mid D$ (intersection).

Semigraphoid if only (S1)–(S4). It is *compositional* if also

(S6) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid C$ then $A \perp_\sigma (B \cup D) \mid C$ (composition).

Separation in undirected graphs

Let $\mathcal{G} = (V, E)$ be finite and simple undirected graph (no self-loops, no multiple edges).

For subsets A, B, S of V, let $A \perp_{\mathcal{G}} B \mid S$ denote that S *separates A from B in $\mathcal{G}$*, i.e. that all paths from A to B intersect S.

Fact: *The relation $\perp_{\mathcal{G}}$ on subsets of V is a compositional graphoid.*

This fact is the reason for choosing the name 'graphoid' for such independence model.

Probabilistic Independence Model

For a system V of *labeled random variables* $X_v, v \in V$, we use

$$A \perp\!\!\!\perp B \,|\, C \iff X_A \perp\!\!\!\perp X_B \,|\, X_C,$$

where $X_A = (X_v, v \in A)$ denotes the variables with labels in A.

The properties (C1)–(C4) imply that $\perp\!\!\!\perp$ *satisfies the semi-graphoid axioms* and the *graphoid axioms if the joint density of the variables is strictly positive.*

A regular multivariate Gaussian distribution defines a compositional graphoid independence model, as we shall see later.

Markov properties for undirected graphs

$\mathcal{G} = (V, E)$ simple undirected graph; An independence model $\perp_\sigma$ satisfies

(P) *the pairwise Markov property* if

$$\alpha \not\sim \beta \implies \alpha \perp_\sigma \beta \mid V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property* if

$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

(G) *the global Markov property* if

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

Structural relations among Markov properties

For any semigraphoid it holds that

$$(\mathrm{G}) \implies (\mathrm{L}) \implies (\mathrm{P})$$

If $\perp_\sigma$ satisfies graphoid axioms it further holds that

$$(\mathrm{P}) \implies (\mathrm{G})$$

so that *in the graphoid case*

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

The latter holds in particular for $\perp\!\!\!\perp$, when $f(x) > 0$.

The multivariate Gaussian

A d-dimensional random vector $X = (X_1, \dots, X_d)$ has a *multivariate Gaussian distribution* or *normal* distribution on $\mathcal{R}^d$ if there is a vector $\xi \in \mathcal{R}^d$ and a $d \times d$ matrix Σ such that

$$\lambda^\top X \sim \mathcal{N}(\lambda^\top \xi, \lambda^\top \Sigma \lambda) \quad \text{for all } \lambda \in R^d. \tag{1}$$

We then write $X \sim \mathcal{N}_d(\xi, \Sigma)$. Then

$$X_i \sim \mathcal{N}(\xi_i, \sigma_{ii}), \quad \mathrm{Cov}(X_i, X_j) = \sigma_{ij}.$$

Hence ξ is the *mean vector* and Σ the *covariance matrix* of the distribution.

A multivariate Gaussian distribution is determined by its mean vector and covariance matrix.

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 on $\mathcal{R}^d$

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

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. Since a positive semidefinite matrix is positive definite if and only if it is invertible, we then also say that Σ is *regular*.

Adding two independent Gaussians yields a Gaussian:

If $X \sim \mathcal{N}_d(\xi_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_d(\xi_2, \Sigma_2)$ and $X_1 \perp\!\!\!\perp X_2$

$$X_1 + X_2 \sim \mathcal{N}_d(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).$$

Affine transformations preserve multivariate normality:

If L is an $r \times d$ matrix, $b \in \mathcal{R}^r$ and $X \sim \mathcal{N}_d(\xi, \Sigma)$, then

$$Y = LX + b \sim \mathcal{N}_r(L\xi + b, L\Sigma L^\top).$$

Marginal and conditional distributions

Partition X into into X_A and X_B, where $X_A \in \mathcal{R}^A$ and $X_B \in \mathcal{R}^B$ with $A \cup B = V$. Partition mean vector, concentration and covariance matrix accordingly as

$$\xi = \begin{pmatrix} \xi_A \\ \xi_B \end{pmatrix}, \quad K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}.$$

Then, if $X \sim \mathcal{N}(\xi, \Sigma)$ it holds that

$$X_B \sim \mathcal{N}_s(\xi_B, \Sigma_{BB}).$$

Also

$$X_A \,|\, X_B = x_B \sim \mathcal{N}_A(\xi_{A|B}, \Sigma_{A|B}).$$

Here

$$\xi_{A|B} = \xi_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \xi_B) \quad \text{and} \quad \Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.$$

Using the matrix identities

$$K_{AA}^{-1} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \tag{3}$$

and

$$K_{AA}^{-1}K_{AB} = -\Sigma_{AB}\Sigma_{BB}^{-1}, \tag{4}$$

it follows that

$$\xi_{A|B} = \xi_A - K_{AA}^{-1}K_{AB}(x_B - \xi_B) \quad \text{and} \quad K_{A|B} = K_{AA}.$$

Note that the *marginal covariance is simply expressed in terms of Σ* whereas the *conditional concentration is simply expressed in terms of K.*

Further, since

$$\xi_{A|B} = \xi_A - K_{AA}^{-1} K_{AB}(x_B - \xi_B) \quad \text{and} \quad K_{A|B} = K_{AA},$$

X_A and X_B are independent if and only if $K_{AB} = 0$, giving *$K_{AB} = 0$ if and only if $\Sigma_{AB} = 0$*.

More generally, if we partition X into X_A, X_B, X_C, the conditional concentration of $X_{A\cup B}$ given $X_C = x_C$ is

$$K_{A\cup B|C} = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix},$$

so

$$X_A \perp\!\!\!\perp X_B \,|\, X_C \iff K_{AB} = 0.$$

It follows that *a Gaussian independence model is a compositional graphoid.*

Gaussian graphical model

$\mathcal{S}(\mathcal{G})$ 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.*.

Likelihood with restrictions

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 and $\Sigma^{-1} = K \in \mathcal{S}^+(\mathcal{G})$.

Define the matrices $T^u, u \in V \cup E$ as those with elements

$$T^u_{ij} = \begin{cases} 1 & \text{if } u \in V \text{ and } i = j = u \\ 1 & \text{if } u \in E \text{ and } u = \{i, j\} \\ 0 & \text{otherwise.} \end{cases};$$

then *$T^u, u \in V \cup E$ forms a basis* for the linear space $\mathcal{S}(\mathcal{G})$ of symmetric matrices over V which have zero entries ij whenever i and j are non-adjacent in $\mathcal{G}$.

Further, as $K \in \mathcal{S}(\mathcal{G})$, we have

$$K = \sum_{v \in V} k_v T^v + \sum_{e \in E} k_e T^e \tag{5}$$

and hence

$$\operatorname{tr}(Kw) = \sum_{v \in V} k_v \operatorname{tr}(T^v w) + \sum_{e \in E} k_e \operatorname{tr}(T^e w);$$

leading to the log-likelihood function

$$\begin{aligned} l(K) &= \log L(K) \sim \frac{n}{2} \log(\det K) - \operatorname{tr}(Kw)/2 \\ &= \frac{n}{2} \log(\det K) \\ &\quad - \sum_{v \in V} k_v \operatorname{tr}(T^v w)/2 + \sum_{e \in E} k_e \operatorname{tr}(T^e w)/2. \end{aligned}$$

Hence we can identify the family as a (regular and canonical) exponential family with $-\mathrm{tr}(T^u W)/2, u \in V \cup E$ as canonical sufficient statistics.

The likelihood equations can be obtained from this fact or by differentiation, combining the fact that

$$\frac{\partial}{\partial k_u} \log \det(K) = \mathrm{tr}(T^u \Sigma)$$

with (5).

This eventually yields the *likelihood equations*

$$\mathrm{tr}(T^u w) = n \, \mathrm{tr}(T^u \Sigma), \quad u \in V \cup E.$$

The likelihood equations

$$\operatorname{tr}(T^u w) = n \operatorname{tr}(T^u \Sigma), \quad u \in V \cup E.$$

can also be expressed as

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

Remember the *model restriction* $K = \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$.

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

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

General theory of exponential families ensure the solution to be unique, provided it exists.

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

$$M_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \qquad (6)$$

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

$$(K_{AA})^{-1} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.$$

Switching the role of K and Σ yields

$$\Sigma_{AA} = (K^{-1})_{AA} = \left(K_{AA} - K_{AB}K_{BB}^{-1}K_{BA}\right)^{-1}.$$

Hence

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

Thus the C-marginal covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted concentration matrix becomes

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

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

From (6) it is seen that the pattern of zeros in K is preserved under the operation M_c, and it stays positive definite.

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

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

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

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

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.

This algorithm is also known as *Iterative Proportional Scaling* or *Iterative Marginal Fitting*.

Factorization

Assume density f w.r.t. product measure on $\mathcal{X}$.
For $a \subseteq V$, $\psi_a(x)$ denotes a function which depends on x_a only, i.e.

$$x_a = y_a \implies \psi_a(x) = \psi_a(y).$$

We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity.

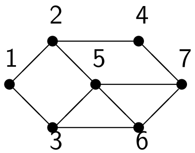
Definition

The distribution of X *factorizes w.r.t.* $\mathcal{G}$ or satisfies (F) if

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

where $\mathcal{A}$ are *complete* subsets of $\mathcal{G}$.

Complete subsets of a graph are sets with all elements pairwise neighbours.



The *cliques* of this graph are the maximal complete subsets $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, $\{2,5\}$, $\{3,5,6\}$, $\{4,7\}$, and $\{5,6,7\}$.
A complete set is any subset of these sets.
The graph above corresponds to a factorization as

$$\begin{aligned} f(x) &= \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{25}(x_2,x_5) \\ &\times \psi_{356}(x_3,x_5,x_6)\psi_{47}(x_4,x_7)\psi_{567}(x_5,x_6,x_7). \end{aligned}$$

Theorem

Let (F) denote the property that f factorizes w.r.t. $\mathcal{G}$ and let (G), (L) and (P) denote Markov properties for $\perp\!\!\!\perp$. It then holds that

$$(F) \implies (G).$$

If f is continuous and $f(x) > 0$ for all x, $(P) \implies (F)$.

The former of these is a simple direct consequence of the factorization whereas the second implication is more subtle.

Thus in the case of positive density (but typically only then), *all the properties coincide:*

$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

Graph decomposition

Consider an *undirected* graph $\mathcal{G} = (V, E)$. A partitioning of V into a triple (A, B, S) of subsets of V forms a *decomposition* of $\mathcal{G}$ if

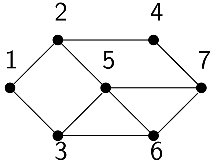
$$A \perp_{\mathcal{G}} B \mid S \text{ and } S \text{ is complete.}$$

The decomposition is *proper* if $A \neq \emptyset$ and $B \neq \emptyset$.

The *components* of $\mathcal{G}$ are the induced subgraphs $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$.

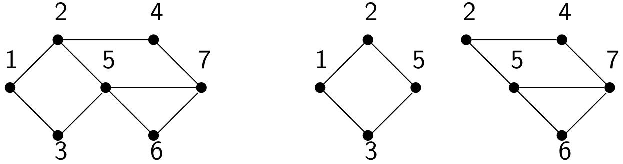
A graph is *prime* if no proper decomposition exists.

Example



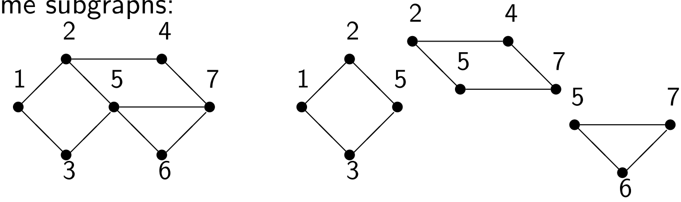
The graph to the left is prime

Decomposition with $A = \{1, 3\}$, $B = \{4, 6, 7\}$ and $S = \{2, 5\}$



Decomposability

Any graph can be recursively decomposed into its maximal prime subgraphs:



A graph is *decomposable* (or rather fully decomposable) if it is complete or admits a proper decomposition into *decomposable* subgraphs.

Definition is recursive. Alternatively this means that *all maximal prime subgraphs are cliques.*

Factorization of Markov distributions

Suppose P satisfies (F) w.r.t. $\mathcal{G}$ and (A, B, S) is a decomposition. Then

(i) *$P_{A\cup S}$ and $P_{B\cup S}$ satisfy (F) w.r.t. $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively;*

(ii) $f(x)f_S(x_S) = f_{A\cup S}(x_{A\cup S})f_{B\cup S}(x_{B\cup S})$.

The converse also holds in the sense that *if (i) and (ii) hold, and (A, B, S) is a decomposition of $\mathcal{G}$, then P factorizes w.r.t. $\mathcal{G}$.*

Recursive decomposition of a decomposable graph into cliques yields the formula:

$$f(x)\prod_{S\in\mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{C\in\mathcal{C}} f_C(x_C).$$

Here $\mathcal{S}$ is the set of *minimal complete separators* occurring in the decomposition process and $\nu(S)$ the number of times such a separator appears in this process.

Characterizing decomposable graphs

A graph is *chordal* if all cycles of length ≥ 4 have chords.

The following are equivalent for any undirected graph $\mathcal{G}$.

(i) *$\mathcal{G}$ is chordal;*

(ii) *$\mathcal{G}$ is decomposable;*

(iii) *All maximal prime subgraphs of $\mathcal{G}$ are cliques;*

There are also many other useful characterizations of chordal graphs and algorithms that identify them.

Trees are chordal graphs and thus decomposable.

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*.

We also have the *factorization of densities*

$$f(x \mid \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \Sigma_S)^{\nu(S)}} \tag{7}$$

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 (7) we can for example match the expressions for the trace and determinant of Σ

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

and further

$$\det\Sigma \quad = \quad \{\det(K)\}^{-1} = \frac{\prod_{C\in\mathcal{C}} \det\{\Sigma_C\}}{\prod_{S\in\mathcal{S}} \{\det(\Sigma_S)\}^{\nu(S)}}$$

These are some of many relations that can be derived using the decomposition property of chordal graphs.

The same factorization clearly holds for the maximum likelihood estimates:

$$f(x \,|\, \hat{\Sigma}) = \frac{\prod_{C \in \mathcal{C}} f(x_C \,|\, \hat{\Sigma}_C)}{\prod_{S \in \mathcal{S}} f(x_S \,|\, \hat{\Sigma}_S)^{\nu(S)}} \tag{8}$$

Moreover, it follows from the general likelihood equations that

$$\hat{\Sigma}_A = W_A / n \text{ whenever } A \text{ is complete.}$$

Exploiting this, we can obtain an explicit formula for the maximum likelihood estimate in the case of a chordal graph.

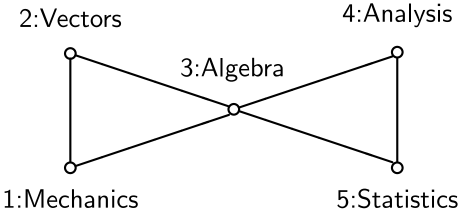
For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V| \times |V|$, i.e.

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

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:

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

Mathematics marks



This graph is chordal with cliques $\{1, 2, 3\}$, $\{3, 4, 5\}$ with separator $S = \{3\}$ having $\nu(\{3\}) = 1$.

Since one degree of freedom is lost by subtracting the average, we get in this example

$$\hat{K} = 87 \begin{pmatrix} w^{11}_{[123]} & w^{12}_{[123]} & w^{13}_{[123]} & 0 & 0 \\ w^{21}_{[123]} & w^{22}_{[123]} & w^{23}_{[123]} & 0 & 0 \\ w^{31}_{[123]} & w^{32}_{[123]} & w^{33}_{[123]} + w^{33}_{[345]} - 1/w_{33} & w^{34}_{[345]} & w^{35}_{[345]} \\ 0 & 0 & w^{43}_{[345]} & w^{44}_{[345]} & w^{45}_{[345]} \\ 0 & 0 & w^{53}_{[345]} & w^{54}_{[345]} & w^{55}_{[345]} \end{pmatrix}$$

where $w^{ij}_{[123]}$ is the ijth element of the inverse of

$$W_{[123]} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix}$$

and so on.

Existence of the MLE

The IPS algorithm converges to the maximum likelihood estimator of $\hat{K}$ of K *provided that the likelihood function does attain its maximum.*

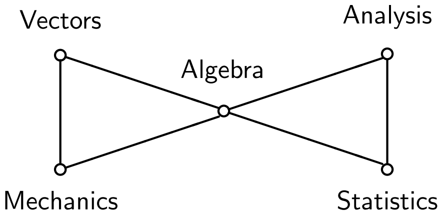
The question of existence is non-trivial.

A *chordal cover* of $\mathcal{G}$ is a chordal graph (no cycles without chords) $\mathcal{G}'$ of which $\mathcal{G}$ is a subgraph.

Let $n' = \max_{C \in \mathcal{C}'} |C|$, where $\mathcal{C}'$ is the set of cliques in $\mathcal{G}'$ and let n^+ *denote smallest possible value* of n'.

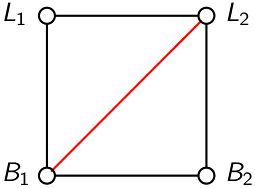
The quantity $\tau(\mathcal{G}) = n^+ - 1$ is known as the *treewidth* of $\mathcal{G}$ (Halin, 1976; Robertson and Seymour, 1984).

The condition $n > \tau(\mathcal{G})$ is sufficient for the existence of the MLE.



This graph has treewidth $\tau(\mathcal{G})=2$ since it is itself chordal and the largest clique has size 3.

Hence *$n = 3$ observations is sufficient for the existence of the MLE.*



This graph has also treewidth $\tau(\mathcal{G})=2$ since a chordal cover can be obtained by adding a diagonal edge.

Hence also here *$n = 3$ observations is sufficient for the existence of the MLE.*

Determining the treewidth $\tau(\mathcal{G})$ is a difficult combinatorial problem (Robertson and Seymour, 1986), but for any n it can be *decided with complexity $O(|V|)$ whether $\tau(\mathcal{G}) < n$* (Bodlaender, 1997).

If we let n^{-} denote the maximal clique size of $\mathcal{G}$, *a necessary condition is that $n \geq n^{-}$.*

For $n^{-} \leq n \leq \tau(\mathcal{G})$ it is unclear.

Buhl (1993) shows for a p-cycle, we have $n^{-} = 2$ and $\tau(\mathcal{G}) = 2$. If now $n = 2$, the probability that the MLE exists is strictly between 0 and 1. In fact,

$$P\{\text{MLE exists} \mid K = I\} = 1 - \frac{2}{(p-1)!}.$$

Similar results hold for the bipartite graphs $K_{2,m}$ (Uhler, 2012) and other special cases, but general case is unclear.

Recently there has been considerable progress (Gross and Sullivant, 2015), for example it can be shown that *$n = 4$ observations suffice for any planar graph*, an interesting parallel to the four-colour theorem.

The *r-core* of a graph $\mathcal{G}$ is obtained by repeatedly removing all vertices with less than r neighbours.

It then holds (Gross and Sullivant, 2015) that *if the r-core of $\mathcal{G}$ is empty, then $n = r$ observations are enough.*

Bodlaender, H. L. (1997). Treewidth: Algorithmic techniques and results. In Prívara, I. and Ručička, P., editors, *Mathematical Foundations of Computer Science 1997*, volume 1295 of *Lecture Notes in Computer Science*, pages 19–36. Springer Berlin Heidelberg.

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, 20:263–270.

Gross, E. and Sullivant, S. (2015). The maximum likelihood threshold of a graph. ArXiv:1404.6989.

Halin, R. (1976). S-functions for graphs. *Journal of Geometry*, 8:171–186.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

Robertson, N. and Seymour, P. D. (1984). Graph minors III. Planar tree-width. *Journal of Combinatorial Theory, Series B*, 36:49–64.

Robertson, N. and Seymour, P. D. (1986). Graph minors II. Algorithmic aspects of tree-width. *Journal of Algorithms*, 7:309–322.

Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer-Verlag, London.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics*, 40:238–261.