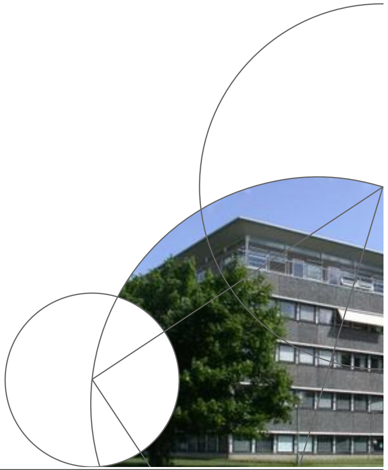
Faculty of Science

Likelihood Analysis of Gaussian Graphical Models

Steffen Lauritzen
Department of Mathematical Sciences



UNIVERSITY OF COPENHAGEN DEPARTMENT OF MATHEMATICAL SCIENCES

Consider the case where $\xi = 0$ and a sample $X^1 = x^1, \dots, X^n = x^n$ from a multivariate Gaussian distribution $\mathcal{N}_d(0, \Sigma)$ with Σ regular. Using the expression for the density, we get the likelihood function

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

where

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

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF MATHEMATICAL SCIENCES

Writing the trace out

$$\operatorname{tr}(KW) = \sum_i \sum_j k_{ij} W_{ji}$$

emphasizes that it is linear in both K and W and we can recognize this as a linear and canonical exponential family (Barndorff-Nielsen, 1978) with K as the canonical parameter and $-W/2$ as the canonical sufficient statistic. Thus, the likelihood equation becomes

$$\mathbf{E}(-W/2) = -n\Sigma/2 = -w/2$$

since $\mathbf{E}(W) = n\Sigma$. Solving, we get

$$\hat{K}^{-1} = \hat{\Sigma} = w/n$$

in analogy with the univariate case.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF MATHEMATICAL SCIENCES

Rewriting the likelihood function as

$$\log L(K) = \frac{n}{2}\log(\det K) - \operatorname{tr}(Kw)/2$$

we can of course also differentiate to find the maximum, leading to the equation

$$\frac{\partial}{\partial k_{ij}} \log(\det K) = w_{ij}/n,$$

which in combination with the previous result yields

$$\frac{\partial}{\partial K} \log(\det K) = K^{-1}.$$

The latter can also be derived directly by writing out the determinant, and it holds for any non-singular square matrix, i.e. one which is not necessarily positive definite.

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

$$K = \sum_{v \in V} k_v T^v + \sum_{e \in E} k_e T^e \qquad (2)$$

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 $-\operatorname{tr}(T^u W)/2, u \in V \cup E$ as canonical sufficient statistics.

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

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

with (2). This eventually yields the likelihood equations

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

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

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

We should 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.

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

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

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

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

and 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 (3) it is seen that the pattern of zeros in K is preserved under the operation M_c, and it can also be seen to stay positive definite.

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

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

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*.

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

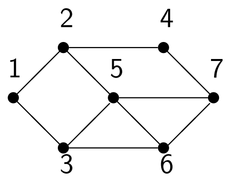
$$A \perp_{\mathcal{G}} B \mid S \text{ and } S \text{ is complete.}$$

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 $\mathcal{G}_{B\cup S}$.

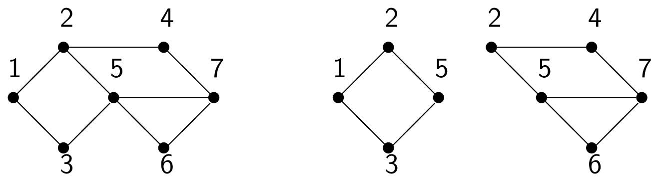
A graph is *prime* if no proper decomposition exists.

Example



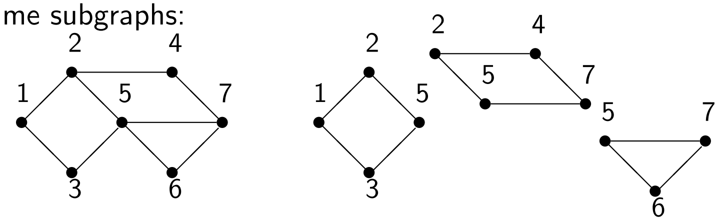
The graph to the left is prime

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*

$$f(x\,|\,\Sigma) = \frac{\prod_{C\in\mathcal{C}} f(x_C\,|\,\Sigma_C)}{\prod_{S\in\mathcal{S}} f(x_S\,|\,\Sigma_S)^{\nu(S)}} \qquad (4)$$

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 (4) 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:

$$f(x\,|\,\hat{\Sigma}) = \frac{\prod_{C\in\mathcal{C}} f(x_C\,|\,\hat{\Sigma}_C)}{\prod_{S\in\mathcal{S}} f(x_S\,|\,\hat{\Sigma}_S)^{\nu(S)}} \qquad (5)$$

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.

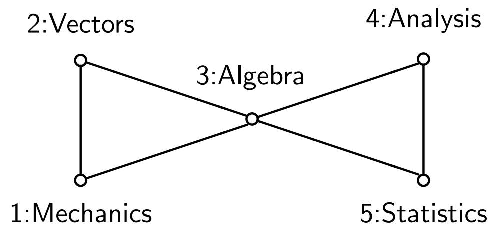
For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V| \times |V|$, i.e.

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

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.*

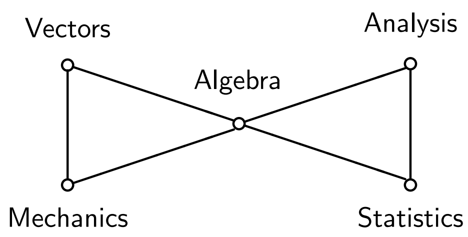
The question of existence is non-trivial.

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'.

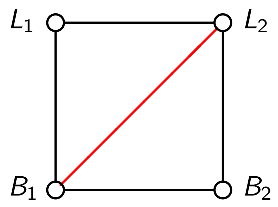
The quantity $\tau(\mathcal{G}) = n^+ - 1$ is known as the *treewidth* of $\mathcal{G}$ (Halin, 1976; Robertson and Seymour, 1984).

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, 2014), for example it can be shown that *$n = 4$ observations suffice for any planar graph*, an interesting parallel to the four-colour theorem.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley and Sons, New York.

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. (2014). The maximum likelihood threshold of a graph. ArXiv:1404.6989.

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

Robertson, N. and Seymour, P. D. (1984). Graph minors III.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF MATHEMATICAL SCIENCES

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.

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

Steffen Lauritzen — Likelihood Analysis of Gaussian Graphical Models — Swiss Winterschool 2015, Lecture 2
Slide 28/28