Graphical methods for efficient likelihood inference in Gaussian covariance models

Mathias Drton

Department of Statistics
University of Chicago

(joint work with Thomas Richardson)

July 1, 2008

Outline

1. Covariance graphs

- Covariance graph $G = (V, E)$ is simple undirected graph
  (we draw edges as $\longleftrightarrow$ and also speak of a bi-directed graph)
- Associated set of covariance matrices

$$\mathcal{C}(G) = \{\Sigma = (\sigma_{vw}) \in PD(V) \ : \ \sigma_{vw} = 0 \text{ if } (v, w) \notin E\}$$

Example

Graph G:

$$X_1 \longleftrightarrow X_2 \longleftrightarrow X_3 \longleftrightarrow X_4$$

Associated covariance matrices in $\mathcal{C}(G)$ are tridiagonal:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & 0 \\ 0 & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_{44} \end{pmatrix}$$

ML estimation in Gaussian covariance model

- Gaussian covariance model $\{\mathcal{N}(0, \Sigma) \ : \ \Sigma \in \mathcal{C}(G)\}$
- Observe n-sample giving rise to a data matrix

$$X = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ \vdots & & \vdots \\ X_{V1} & \dots & X_{Vn} \end{pmatrix}$$

- Sample covariance matrix

$$S = \frac{1}{n} X X^t$$

- Compute MLE $\hat{\Sigma}$ by maximizing

$$\ell(\Sigma) = -\log\det\Sigma - \mathrm{tr}(\Sigma^{-1}S) \quad \text{subject to } \Sigma \in \mathcal{C}(G).$$

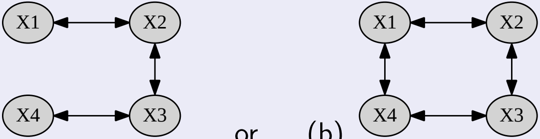
- Assume sample size $n \geq V$ s.t. $P(S \text{ positive definite}) = 1$.

Computational algebra

Rational likelihood equations can be solved algebraically for small graphs

Proposition

If $V \leq 4$, then the likelihood equations almost surely have one feasible solution (ML degree 1) except when G is

(a)  or (b)

Case (a): ML degree 5 (all 5 solutions may be feasible)
Case (b): ML degree 17.

In both cases (a) and (b) there exist positive definite sample covariance matrices such that MLE $\hat{\Sigma}$ cannot be expressed in radicals.

Iterative conditional fitting

Goal

Compute MLE of joint distribution of $X = (X_v \mid v \in V)$ in the model associated with a covariance graph $G = (V, E)$.

Outline of algorithm

Initialization: Choose feasible joint distribution of X and a vertex $w \in V$.

Iterations: Repeat the following steps until convergence

> *Step 1*: Fix marginal distribution of $X_{V \setminus \{w\}}$.
>
> *Step 2*: Estimate conditional distribution of X_w given $X_{V \setminus \{w\}}$ under the constraints implied by the graphical model.
>
> *Step 3*: Compute estimate of joint distribution of X by multiplying estimated conditional and fixed marginal distribution.
>
> *Step 4*: Set $w = w + 1 \mod V$.

Iterative conditional fitting for Gaussian models

Algorithm

Input: Graph G, sample covariance matrix S

Output: MLE $\hat{\Sigma}$ in Gaussian covariance model $\{\mathcal{N}(0,\Sigma) \ : \ \Sigma \in \mathcal{C}(G)\}$

Initialization: Choose starting value $\hat{\Sigma} \in \mathcal{C}(G)$ and a vertex $w \in V$.

Iterations: Repeat the following steps until convergence

Step 1: Fix submatrix $\hat{\Sigma}_{R\times R}$ where $R := V \setminus \{w\}$.

Step 2: Estimate, by maximum likelihood, w-th row and column of Σ subject to $\Sigma \in \mathcal{C}(G)$ and $\Sigma_{R\times R} = \hat{\Sigma}_{R\times R}$.

Step 3: Set $w = w + 1 \mod V$.

The update step in iterative conditional fitting

- Recall conditional distribution:

$$(X_w \mid X_R) \sim \mathcal{N}(\Sigma_{\{w\}\times R}\, \Sigma_{R\times R}^{-1} X_R \,,\, \sigma_{ww.R})$$

- Define pseudo-variables

$$Z_R = \Sigma_{R\times R}^{-1} X_R$$

- Spouses of w are the neighbors in the covariance graphs

$$\mathrm{sp}(w) = \{v \in V \ :\ (v, w) \in E\}$$

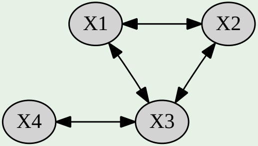
- Problem of estimating constrained conditional distribution in Step 2 has closed form solution (rational in S):

  Least squares regression of X_w on the pseudo-variables Z_u, $u \in \mathrm{sp}(w)$.

Fitting a 4-variable graph

Example

- Covariance graph G:



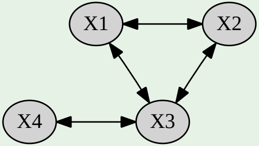
- Sample covariance matrix

$$S = \begin{pmatrix} 1 & 0.13 & 0.31 & -0.67 \\ & 1 & -0.43 & 0.23 \\ & & 1 & 0.17 \\ & & & 1 \end{pmatrix}$$

- Iterative conditional fitting takes 140 iterations (using defaults for `fitCovGraph` in R package ggm)

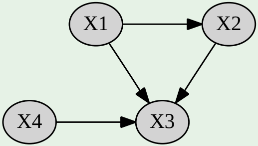
Example (cont.)

- Covariance graph G:



$$(X_1, X_2) \perp\!\!\!\perp X_4$$

- MLE $\hat{\Sigma}$ is rational in S because G is *Markov equivalent* to the DAG



$$(X_1, X_2) \perp\!\!\!\perp X_4$$

- Four least squares regressions suffice to compute $\hat{\Sigma}$!

What's coming next...

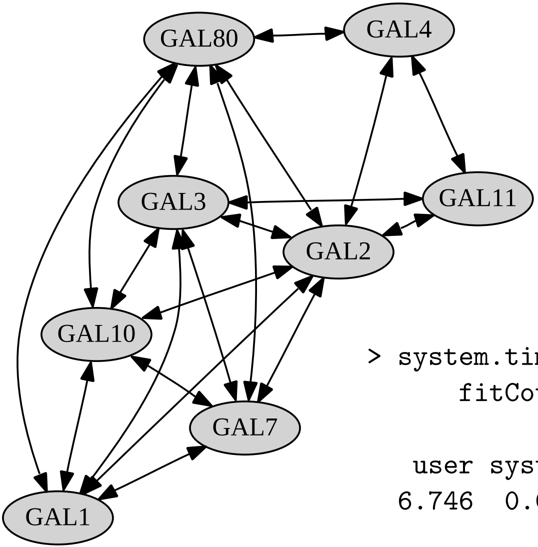
Goal

By removing arrowheads, transform covariance graph into another Markov equivalent graph such that associated model is easier to fit!

Tools

- Ancestral graphs:
  - removing arrowheads gives mixed graph
  - maximal ancestral graphs define (conditional) independence models
  - d-separation
- Residual iterative conditional fitting:
  - can be applied to ancestral graphs
  - reduces to least squares regression for DAGs

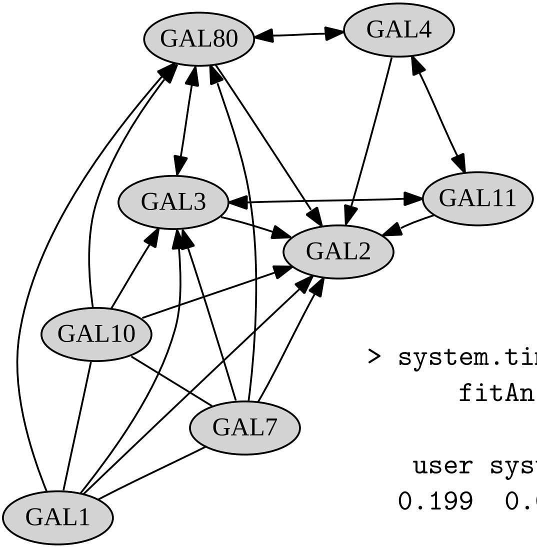
Gene expression data




```
> system.time( for(i in 1:10)
      fitCovGraph(G,S,n) )

  user system elapsed
 6.746  0.014   6.762
```

Gene expression data




```
> system.time( for(i in 1:10)
     fitAncestralGraph(G,S,n) )

 user system elapsed
0.199  0.002   0.201
```

2. Ancestral graphs

Consider simple mixed graphs with edges of 3 types,

$$\text{undirected } (\!-\!\!-\!), \text{ directed } (\longrightarrow) \text{ and bi-directed } (\longleftrightarrow)$$

but no loops or multiple edges between two vertices.

Definition (Richardson & Spirtes, 2002)

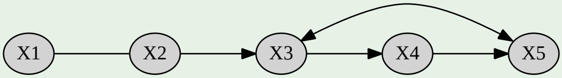
A simple mixed graph is ancestral if none of the following occurs:

(i) Undirected edge meets arrowhead: $- v \longleftarrow$, $- v \longleftrightarrow$

(ii) Directed cycle: $v \longrightarrow \ldots \longrightarrow v$

(iii) Spouse is an ancestor: $v \longleftrightarrow w \longrightarrow \ldots \longrightarrow v$

Example (Non-ancestral graph)



d-Separation in ancestral graphs

Definition

- A vertex v on a path is a collider if the incident edges are of the form:

$$\longrightarrow v \longleftarrow, \quad \longleftrightarrow v \longleftrightarrow, \quad \longleftrightarrow v \longleftarrow$$

- A path π d-connects two vertices $v, w \in V$ given $C \subseteq V \setminus \{v, w\}$ if:
  (i) if u is a non-collider on π, then $u \notin C$,
  (ii) if u is a collider on π, then

$$u \in \mathrm{An}(C) := C \cup \{s \in V \ : \ \exists t \in C \text{ s.t. } s \longrightarrow \ldots \longrightarrow t\}.$$

- Two disjoint and non-empty sets $A, B \subseteq V$ are d-connected given $C \subseteq V \setminus (A \cup B)$ if there is a path that d-connects a vertex $v \in A$ and a vertex $w \in B$ given C.

  If there is no such d-connecting path, then C d-separates A and B.

Global Markov property for ancestral graphs

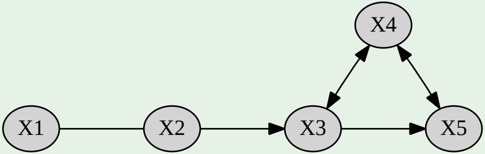
Definition

The joint distribution of a random vector $(X_v \mid v \in V)$ obeys the global Markov property for an ancestral graph $G = (V, E)$ if

$$C \text{ d-separates } A \text{ and } B \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C.$$

Example

Global Markov property for



yields e.g. $X_1 \perp\!\!\!\perp (X_3, X_4, X_5) \mid X_2, \quad (X_1, X_2) \perp\!\!\!\perp X_4, \quad (X_1, X_2) \perp\!\!\!\perp X_5 \mid X_3.$

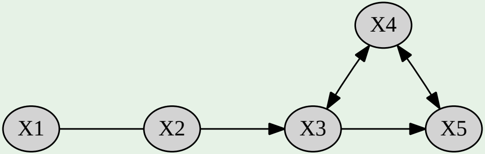
Maximal ancestral graphs

Definition

An ancestral graph is maximal if for every non-edge (v, w) there exists a set $C \subseteq V \setminus \{v, w\}$ such that C d-separates v and w.

Example

Global Markov property for



yields e.g. $X_1 \perp\!\!\!\perp (X_3, X_4, X_5) \mid X_2$, $\quad (X_1, X_2) \perp\!\!\!\perp X_4$, $\quad (X_1, X_2) \perp\!\!\!\perp X_5 \mid X_3$.
This graph is a maximal ancestral graph.

Gaussian covariance models

If G is a covariance graph (ancestral graph with only bi-directed edges), then:

- A path d-connecting v and w given C has all non-endpoint vertices in the conditioning set C.
- Global Markov property specializes to
$$V \setminus C \text{ separates } A \text{ and } B \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C .$$
- All distributions in the Gaussian model $\{\mathcal{N}(0,\Sigma) \,:\, \Sigma \in \mathcal{C}(G)\}$ obey the global Markov property for G (Kauermann, 1996).
- Any ancestral graph that is Markov equivalent to G and has the same adjacencies is also maximal.

Main lemma

In a simple mixed graph $G = (V, E)$, define the boundary of $A \subseteq V$ as

$$\mathrm{Bd}(A) = A \cup \{v \in V \;:\; (v, w) \in E \text{ for some } w \in A\}.$$

Definition

A simple mixed graph G has the boundary containment property if

$$v \;\text{——}\; w \text{ in } G \quad \Longrightarrow \quad \mathrm{Bd}(v) = \mathrm{Bd}(w)$$

$$v \longrightarrow w \text{ in } G \quad \Longrightarrow \quad \mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$$

(In other words: G has no unshielded non-colliders.)

Lemma

Suppose a bi-directed graph G and an ancestral graph H have the same adjacencies. Then G and H are Markov equivalent $\iff$ H has boundary containment property.

(Ancestral graphs are Markov equivalent if d-separation relations are the same.)

Simplicial graphs

Definition

A vertex $v \in V$ is simplicial, if $\mathrm{Bd}(v)$ is complete, i.e., every pair of vertices in $\mathrm{Bd}(v)$ are adjacent. A subset $A \subseteq V$ is simplicial, if $\mathrm{Bd}(A)$ is complete.

Drop the arrowhead at v:

Replace $v \longleftarrow w$ by $v \;—\; w$ or $v \longleftrightarrow w$ by $v \longrightarrow w$

Definition

Let G be a bi-directed graph. The simplicial graph G^s is the simple mixed graph obtained by dropping all the arrowheads at simplicial vertices of G.

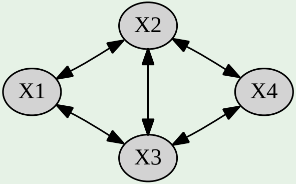
Theorem

The simplicial graph G^s of a bi-directed graph G is a maximal ancestral graph that is Markov equivalent to G.

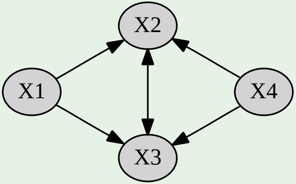
Example & Markov equivalence

Example

G :



G^s :



Proposition (Pearl & Wermuth, 1994)

The bi-directed graph G is Markov equivalent to an undirected graph
$\iff$ *Simplicial graph G^s is an undirected graph*
$\iff$ *G is disjoint union of complete (bi-directed) graphs.*

Minimally oriented graphs

Definition

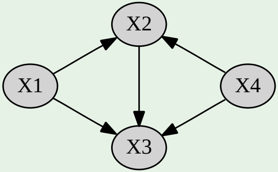
Let G be a bi-directed graph. A minimally oriented graph of G is a maximal ancestral graph $G^{\min}$ such that:

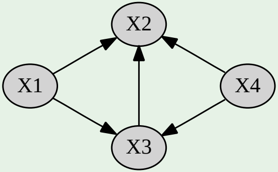
(i) G and $G^{\min}$ are Markov equivalent;

(ii) $G^{\min}$ has the minimum number of arrowheads of all maximal ancestral graphs that are Markov equivalent to G.

(A graph with d directed and b bi-directed edges has $d + 2b$ arrowheads.)

Example (Two minimally oriented graphs)





Construction of minimally oriented graphs

Algorithm

Let G be a bi-directed graph, and $\leq$ a total order on V that extends the partial order $\preccurlyeq_B$ obtained from strict boundary containment. Create a new graph $G^{\min}_{<}$ as follows:

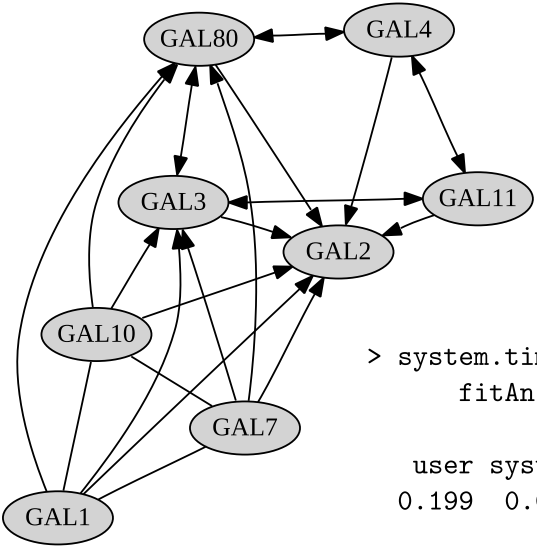
1. find the simplicial graph G^s of G;
2. set $G^{\min}_{<} = G^s$;
3. replace every bi-directed edge $v \longleftrightarrow w \in G^{\min}_{<}$ with $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ and $v < w$ by the directed edge $v \longrightarrow w$.

Theorem

(i) *The graph $G^{\min}_{<}$ constructed in the above algorithm is a minimally oriented graph for the bi-directed graph G.*

(ii) *If $G^{\min}$ is a minimally oriented graph for a bi-directed graph G, then there exists a total order $\leq$ on the vertex set such that $G^{\min} = G^{\min}_{<}$.*

Gene expression data




```
> system.time( for(i in 1:10)
      fitAncestralGraph(G,S,n) )

  user system elapsed
 0.199  0.002   0.201
```

Chordal cographs

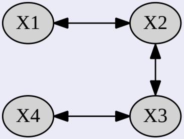
Let $G^{\min}$ be a minimally oriented graph for a bi-directed graph G.

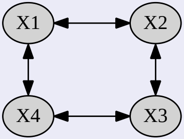
Theorem (Pearl & Wermuth, 1994)

G is Markov equivalent to a DAG $\iff$ $G^{\min}$ has no bi-directed edges.

Lemma

$G^{\min}$ has a bi-directed edge $\iff$ G has an induced subgraph equal to

(a) 

or *(b)* 

.

Chordal cograph: Graph containing neither the path (a) nor the 4-cycle (b)

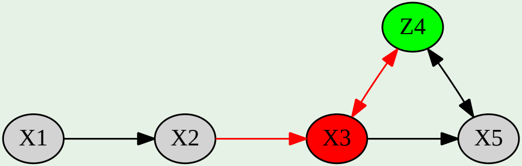
Residual iterative conditional fitting

Residual iterative conditional fitting is an iterative algorithm that can be used to calculate MLE in Gaussian ancestral graph models:

- Iterations cycle through vertices w and compute least squares regression of X_w on X_v, $v \in \mathrm{pa}(v)$, and Z_u, $u \in \mathrm{sp}(v)$.
- Pseudo-variables are Z_u now derived from residuals

Example

Update step for $w = 3$:



Least squares regression of X_3 on X_2 and Z_4

Likelihood inference in Gaussian covariance models

Theorem

(i) *If $A \subseteq V$ is simplicial, then MLE $\hat{\Sigma}$ in $\mathcal{C}(G)$ satisfies $\hat{\Sigma}_{A\times A} = S_{A\times A}$.*

(ii) *If v has no spouses in G^{min}, then*

$$\hat{\Sigma}_{v\times pa(v)}\hat{\Sigma}^{-1}_{pa(v)\times pa(v)} = S_{v\times pa(v)}S^{-1}_{pa(v)\times pa(v)}.$$

Computational algebra yields. . .

Theorem

The MLE $\hat{\Sigma}$ in $\mathcal{C}(G)$ is a rational function of the sample cov. matrix S $\iff$ G is a chordal cograph.

Conclusion

- Construction of minimally oriented graph is very similar to 'sink orientation' described by Pearl & Wermuth (1994):
  - Start with undirected graph
  - Add arrowheads at v if induced subgraph is $\longrightarrow v \longleftarrow$
  - Further directed edges may be needed to get ancestral graph (insertion of complete DAGs)
- Graphical constructions also useful for bi-directed graphical models when variables are categorical:
  - Binary bi-directed models and ICF (M.D. & Richardson, 2008)
  - Discrete models for chain graphs (M.D., 2008)

Some references

- Covariance graph models:
  - M.D. & Richardson (2008). JMLR
  - Pearl & Wermuth (1994). In Lecture Notes in Statistics
  - Kauermann (1996). Scand. J. Statist.
  - Butte et al. (2000). PNAS *('relevance networks')*
  - Wermuth, Cox & Marchetti (2006). Bernoulli
- Iterative conditional fitting:
  - M.D. & Richardson (2003,2004). UAI
  - Chaudhuri, M.D. & Richardson (2007). Biometrika
  - M.D., Eichler & Richardson (2008). Preprint on arXiv.
- Categorical variables:
  - M.D. & Richardson (2008). JRSS B
  - M.D. (2008). *Discrete chain graph models*, two preprints on my webpage (COMPSTAT)
  - Lupparelli, Marchetti & Bergsma (2008). Preprint on arXiv.