A SINful approach to Gaussian graphical model selection

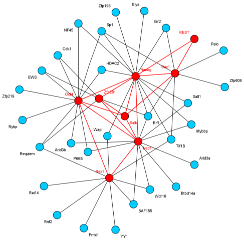
Mathias Drton and Michael Perlman

Adam Gustafson

Stat 572: Methodology Project
University of Washington

April 16, 2013

- In many data sets, characterizing the conditional independencies is of great interest.
- Gene Regulatory Networks: collection of DNA segments which interact with each other indirectly through RNA and protein expression.
- Of great scientific interest to learn the structure of such networks.



- Authors methods allow one to infer such networks for Gaussian data.

Graphical Models

- A *graphical model* is a probabilistic model in which conditional independencies are encoded via graphical properties.
- Formally, we have
  - A Graph $G = (V, E)$ where V is set of nodes and $E \subset V \times V$ is a set of edges.
  - If $i, j \in V$, then $(i, j) \in E$ means that nodes i and j are connected (may or may not be directed edge).
  - One-to-one correspondence between nodes and a set of random variables
  - A graphical model describes a family of distributions $p(x_V)$ over X_V.
  - A rule $r \in R$ is a predicate on a graph: $r(p, G) \in \{\text{true, false}\}$.
  - The set of distributions which a graphical model describes is

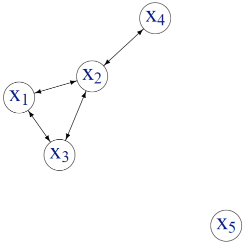
$$\mathcal{F}(G, R) = \{p : p \text{ is a distribution over } X_V \text{ and } r(p, G) = \text{true}, \forall r \in R\}$$

Bidirected Graphs

- Bidirected graphs encode marginal independence properties. In particular, we have the pairwise bidirected Markov property:

$$\mathcal{F}(G, R^{bg}) = \{p : X_u \perp\!\!\!\perp X_v \quad \text{for all non-adjacent pairs } u, v \in V(G)\}$$

- Example below has $X_5 \perp\!\!\!\perp X_{1:4}$, $X_3 \perp\!\!\!\perp X_{4,5}$, etc.

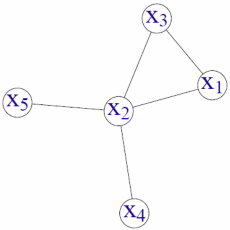


Undirected Graphical Models

- There are many rules for undirected graphs, all equivalent for positive distributions.
- Authors use *pairwise Markov property* rule:

$$\mathcal{F}(G, R^p) = \{p : X_u \perp\!\!\!\perp X_v | X_{V \setminus \{u,v\}} \text{ for all non-adjacent pairs } u, v \in V(G)\}$$

- Example below has $X_3 \perp\!\!\!\perp X_4 | \{X_1, X_2, X_5\}$, $X_3 \perp\!\!\!\perp X_5 | \{X_1, X_2, X_4\}$, etc.

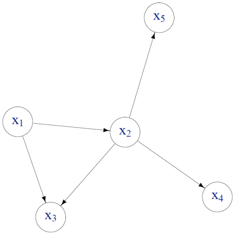


Directed Acyclic Graphical Models

- Conditional independence properties for directed acyclic graphs (DAGs) can be stated in terms of conditioning on parents, or directed factorization:

$$\mathcal{F}(G, R^{df}) = \{p : p(x) = \prod_{v \in V} p(x_v | x_{pa(v)})\}$$

- Example below has $X_3 \perp\!\!\!\perp X_4 | \{X_1, X_2\}$, $X_3 \perp\!\!\!\perp X_5 | \{X_1, X_2\}$, etc.



Chain Graphical Models

- For each $v \in V$, we now have a vector $U(v)$ of variables.
- First factorization is directed factorization over U; second factorization is clique factorization representation of undirected graphical model:

$$\mathcal{F}(G, R^{cg}) = \{p : p(x_U) = \prod_{v \in V} p(x_{U(v)} | x_{pa(v)})\}$$

where for each $v \in V$ we may write:

$$p(x_{U(v)} | x_{U(pa(v))}) \propto \prod_{c \in C(v)} \phi_c(x_c).$$

- Generalizes undirected and directed graphical models – more specificity.
- Authors give two different versions of chain graphs which I'll need to learn.

Gaussian Graphical Models

- Gaussian data can be viewed as any type of the preceeding graphical models.
- In particular, 0's in the precision matrix Σ^{-1} correspond to missing links in the undirected graphical model.
- Taking the Cholesky decomposition of Σ^{-1} yields $U'DU$, where U is upper triangular with ones on the diagonal, and D is diagonal.
- Let $U = (I - B)$. $B_{ij} = 0$ for $j > i$ implies no edge from node j to node i.
- Previous methods: backwards selection starting with full graph, controlling error at each stage.
- *Problem*: Overall error rate for false edge inclusion not controlled.

Authors' Method

- Undirect Case: Authors propose testing $H_{ij} : X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}$ vs. the general alternative.
- Maximum likelihood estimate of covariance matrix in zero mean case is $S = \frac{1}{n} X' X$ where $X \in \mathbb{R}^{n \times p}$ is a matrix of observations.
- ML estimate asymptotically normal.
- Delta Method implies inverse is asymptotically normal. Let r_{ij} be the sample partial correlation corresponding to link ij, and ρ_{ij} the corresponding population quantity.
- r_{ij} asymptotically normal, and Fisher's z-transform improves asymptotics:

$$z_{ij} = \frac{1}{2} \log \left(\frac{1 + r_{ij}}{1 - r_{ij}} \right)$$

Authors' Method 2

- Authors derive p-values which control overall error of false edge inclusion at level α.

$$\pi_{ij} = 1 - \left(2\Phi(\sqrt{n_p}|z_{ij}| - 1\right)^{p(p-1)/2}$$

- Estimation procedure for estimating presence of edge e_{ij}:

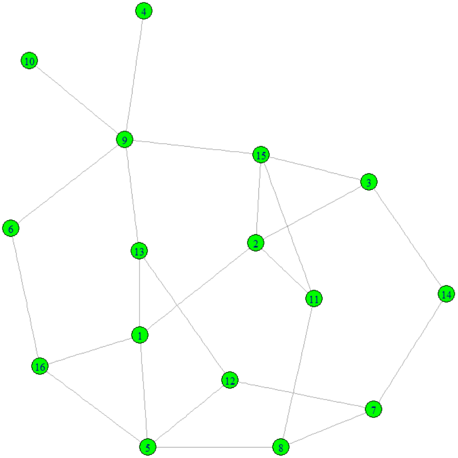
$$\widehat{e}_{ij}(\alpha) = \begin{cases} 0, & \pi_{ij} \geq \alpha \\ 1, & \pi_{ij} < \alpha. \end{cases}$$

- Authors state estimation procedure is conservative: can be improved with Holm's step-down procedure to form adjusted p-values.
- $1 - \alpha$ 'consistency' result:

$$\Pr(\widehat{G}(\alpha) \subseteq G) \geq 1 - \alpha$$

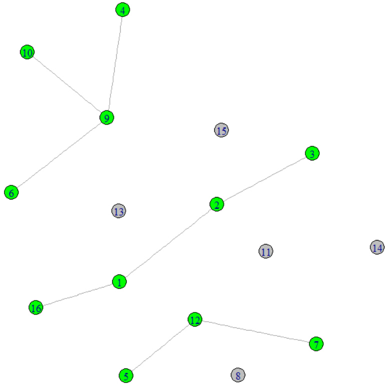
True (Population) Graph

- $p = 16$ in this case.



Estimated Graph: $n = 128$

- Very conservative estimate for $\alpha = 0.05$.



Error Rates

- ‘1’ is Gaussian data. ‘2’ is Nonparanormal data.