Statistical Modelling - Module I

Graphical models

Lecture 2

Federico Castelletti

Department of Statistical Sciences
Università Cattolica del Sacro Cuore
Milan

Frequentist methods for structure learning

Frequentist methods for graph estimation

Traditionally, two types of methods for DAG model selection:

Constraint-based methods
estimate a DAG (or EG) using a sequence of conditional independence tests
example: PC algorithm for EG estimation

Score-based methods
define a score criterion for each DAG (EG) (e.g. based on penalized likelihood)
and search over the space of DAGs (EGs) for the graph with highest score
example: Greedy Equivalence Search (Chickering, 2002)

For UG model selection

methods based on lasso-type regularization (Graphical Lasso)

Focus on the Gaussian case

Frequentist methods for graph estimation

DAGs: Gaussian data

$\mathcal{D} = (V, E)$ DAG

$X_1, \ldots, X_q \mid \boldsymbol{\Sigma} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}) \quad \boldsymbol{\Sigma} \in M_{\mathcal{D}}$

$M_{\mathcal{D}}$ space of all s.p.d. covariance matrices Markov w.r.t. DAG $\mathcal{D}$

Consider the alternative Structural Equation Model (SEM) representation

$$\boldsymbol{x} = \boldsymbol{B}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B}^\top)^{-1}\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{B}^\top)^{-\top}$$

$\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$

$\boldsymbol{B}$ (q, q) matrix of (regression) coefficients with $\boldsymbol{B}_{u,v} \neq 0$ iff $u \to v$

This follows from the factorization property of a DAG (see next)

Frequentist methods for graph estimation

DAGs: Gaussian data

Remember the DAG factorization which under the Gaussian assumption becomes

$$\begin{aligned} f(x_1, \dots, x_q \,|\, \boldsymbol{D}, \boldsymbol{B}, \mathcal{D}) &= \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}) \\ &= \prod_{j=1}^{q} d\,\mathcal{N}\left(x_j \,|\, \boldsymbol{B}_j^\top \boldsymbol{x}, \sigma_j^2\right) \end{aligned}$$

$\boldsymbol{B}_j$ j-th column of matrix $\boldsymbol{B}$

Since each X_j depends only on variables in the set $\boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}$
only the elements of $\boldsymbol{B}_j$ indexed by $\mathrm{pa}_{\mathcal{D}}(j)$ are $\neq 0$
Accordingly, $\boldsymbol{B}_{u,v} \neq 0$ iff $u \to v$ in $\mathcal{D}$

So, missing edges in $\mathcal{D}$ corresponds to 0 elements in $\boldsymbol{B}$

Frequentist methods for graph estimation

UGs: Gaussian data

$\mathcal{G} = (V, E)$ undirected graph

$X_1, \dots, X_q \,|\, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}) \quad \boldsymbol{\Sigma} \in M_{\mathcal{G}}$

$M_{\mathcal{G}}$ space of all s.p.d. covariance matrices Markov w.r.t. UG $\mathcal{G}$

Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix

It is possible to show that

$X_u \perp\!\!\!\perp X_v \,|\, \boldsymbol{x}_{V \setminus \{u,v\}}$ iff $\boldsymbol{\Omega}_{u,v} = 0$

The PC algorithm

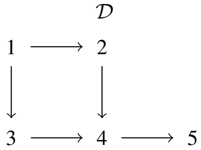
DAG model selection with the PC algorithm

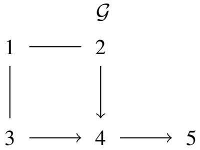
PC algorithm (Spirtes et al, 2000) is a constraint-based method which outputs an EG estimate
It is based on the following steps

Starting from the complete (undirected) graph

1. Identify the skeleton
2. Find the *v*-structures
3. Orient other edges if possible (e.g. edges whose orientation avoids *v*-structures)

Each step is performed using conditional independence tests of "pairwise type"
that is of the form $u \perp\!\!\!\perp v \mid S$, for each pair of nodes $\{u, v\}$ and conditioning set S
Consider the following example with true DAG $\mathcal{D}$ and true EG $\mathcal{G}$





DAG model selection with the PC algorithm

Step 1.

Consider the complete UG

For each pair $\{u, v\}$

remove $u - v$ if $u \perp\!\!\!\perp v \mid S$ for some $S \subset V$

In practice:

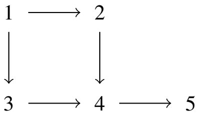
a) Start from conditioning sets of size 0 and test each $u \perp\!\!\!\perp v \mid \emptyset$

b) Move to conditioning sets of size 1 and test each $u \perp\!\!\!\perp v \mid z$ for each $\{u, v\}$ still connected

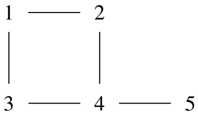
c) [...]

Finally, an undirected graph $\mathcal{U}$ is recovered

$\mathcal{D}$



$\mathcal{U}$



DAG model selection with the PC algorithm

Step 2.
Consider all structures of type $u - z - v$ (undirected v-structures)
where u and v are not connected

If the missing edge between u and v was identified from a cond. indep. test $u \perp\!\!\!\perp v \mid S$ with $z \notin S$
(i.e. z was not in the conditioning set which made u and v independent)
then we orient $u \to z \leftarrow v$

Finally, a partially directed graph $\mathcal{U}^*$ with v-structures is recovered

$\mathcal{D}$

$$\begin{tikzcd}
1 \arrow[r] \arrow[d] & 2 \arrow[d] & \\
3 \arrow[r] & 4 \arrow[r] & 5
\end{tikzcd}$$

$\mathcal{U}^*$

$$\begin{tikzcd}
1 \arrow[r, dash] \arrow[d, dash] & 2 \arrow[d] & \\
3 \arrow[r] & 4 \arrow[r, dash] & 5
\end{tikzcd}$$

DAG model selection with the PC algorithm

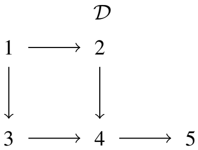
Step 3.
There are other directed edges (not directly involved in v-structures) that can be recovered
In particular those edges whose orientation can produce further v-structures

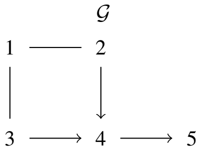
Hence, all edges $a - z$, with

- z involved in $u \rightarrow z \leftarrow v$
- a not connected with u and v

can be oriented as $a \leftarrow z$

Finally, an EG estimate $\mathcal{G}$ is recovered





DAG model selection with the PC algorithm

If the conditional independence decisions are "correct" in the large sample limit, the PC algorithm is guaranteed to converge to the true Markov Equivalence Class in the large sample limit, assuming

faithfulness, i.e $p(\cdot)$ perfect Markovian

i.i.d. samples

no latent variables

Which conditional independence tests?

1) Gaussian data

Tests on partial correlation coefficients

2) Categorical data

G-tests of (conditional) independence

DAG model selection with the PC algorithm

Gaussian data

At step 1. PC estimates an UG
by testing conditional independencies between two variables given a set of variables

In the Gaussian case, tests are based on partial correlation coefficients
In general, partial correlation between u and v, given the "rest" of variables is

$$\rho_{uv \,|\, \text{rest}} = -\frac{\omega_{uv}}{\sqrt{\omega_{uu}\omega_{vv}}}$$

where ω_{uv} is (u, v)-element of $\mathbf{\Omega}$
(0 iff u and v are conditional indep. given the rest)

More specifically, PC requires partial correlations between u and v given a set S, that is $\rho_{uv \,|\, S}$
for "some" set $S \subseteq V \setminus \{u, v\}$

Hence, $\rho_{uv \,|\, S}$ can be computed from $(\mathbf{\Sigma}_{UU})^{-1}$ with $U = u \cup v \cup S$

DAG model selection with the PC algorithm

Gaussian data

Hypothesis test for conditional independence $X_u \perp\!\!\!\perp X_v \,|\, X_S$ is then

$$H_0 \;:\; \rho_{u,v\,|\,S} = 0 \quad vs \quad H_1 \;=\; \rho_{u,v\,|\,S} \neq 0$$

Fisher's z statistics is generally used

$$z(\widehat{\rho}_{u,v\,|\,S}) = \tfrac{1}{2}\ln\left(\tfrac{1+\widehat{\rho}_{u,v\,|\,S}}{1-\widehat{\rho}_{u,v\,|\,S}}\right)$$

where $\widehat{\rho}_{u,v\,|\,S}$ is the sample partial correlation

Given α significance level, H_0 (conditional independence) accepted if

$$\sqrt{n-|S|-3}\,|z(\widehat{\rho}_{u,v\,|\,S})| < \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right)$$

where

Φ : c.d.f. of $\mathcal{N}(0,1)$

n : sample size

$\alpha \in (0,1)$ significance level

DAG model selection with the PC algorithm

Categorical data

Remember that for categorical variables conditional independencies correspond to factorizations of conditional probabilities in conditional contingency tables

In particular, $X_u \perp\!\!\!\perp X_v \mid X_z$ implies that
$p_{x_{uv}|x_z} = p_{x_{u.}|x_z} \, p_{x_{.v}|x_z}$ for each $x_{uv} \in \mathcal{X}_{u,v}, x_z \in \mathcal{X}_z$
where $p_{x_{uv}|x_z} = P(X_u = x_u, X_v = x_v \mid X_z = x_z)$

To test it we can adopt Chi-squared tests or G-tests of (conditional) independence between X_u, X_v where null hypothesis H_0 is conditional independence

PC algorithm implements G-tests by considering the following G-statistic

$$G = 2 \sum_{x_z \in \mathcal{X}_z} \sum_{x_{uv} \in \mathcal{X}_{uv}} n_{x_{uv} \,|\, x_z} \ln \left(\frac{n_{x_{uv} \,|\, x_z}}{\widetilde{n}_{x_{uv} \,|\, x_z}} \right)$$

where
$n_{x_{uv}|x_z}$ is the number of observations with $X_u = x_u$, $X_v = x_v$, $X_z = x_z$ i.e count in cell $\{x_u, x_v, x_z\}$
$\widetilde{n}_{x_{uv}|x_z}$ is the corresponding *expected* frequency
i.e. $\widetilde{n}_{x_{uv}|x_z} = \frac{1}{n_{x_z}} n_{x_{u.}|x_z} \, n_{x_{.v}|x_z}$ (what I would observe if indep. holds)

DAG model selection with the PC algorithm

Categorical data

G-test is basically a likelihood ratio test for categorical (multinomial) data

$G \xrightarrow{D} \chi_d$ Chi-squared distribution with d.f. $d = (l_u - 1)(l_v - 1)l_z$
where $l_u = |\mathcal{X}_u|$ is the number of levels of X_u

Therefore, the null hypothesis of independence is accepted iff $G < \chi_{d,1-\alpha}$, $\alpha \in (0, 1)$

Similarly, the test can be extended for conditioning sets of size > 1

Graphical Lasso

Graphical Lasso for UG model selection

$X_1, \dots, X_q \mid \boldsymbol{\Omega} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1})$

$\boldsymbol{\Omega}$ Markov w.r.t. an unknown UG $\mathcal{G}$

Given n i.i.d. q-variate observations, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and $\boldsymbol{X}$ the (n, q) data matrix

likelihood is

$$p(\boldsymbol{X}; \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Omega}\boldsymbol{S}\big)\right\} \qquad \boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$$

Goal is to estimate a *sparse* precision matrix $\boldsymbol{\Omega}$

GLasso borrows ideas from the lasso (for regularized linear regression) to encourage zero entries in $\boldsymbol{\Omega}$

Therefore, it simultaneously performs (sparse) parameter estimation and model (graph) selection

Remember: zero entries in $\boldsymbol{\Omega} \implies$ missing edges in $\mathcal{G}$

Graphical Lasso for UG model selection

Problem is:

$\max_{\boldsymbol{\Omega}} \;\log p(\boldsymbol{X};\boldsymbol{\Omega}) - \lambda||\boldsymbol{\Omega}||_1$

- $||\boldsymbol{\Omega}||_1 = \sum_{u \neq v} |\omega_{u,v}|$ is the L1 norm
- $\lambda > 0$ tuning parameter controlling the amount of shrinkage

Solution to this problem was provided in an algorithm by Friedman et al. (2008)

$$\max_{\boldsymbol{\Omega}} \;\log p(\boldsymbol{X};\boldsymbol{\Omega}) - \lambda||\boldsymbol{\Omega}||_1 \quad \Longleftrightarrow \quad \min_{\boldsymbol{\Omega}} \; -n\log|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{S}) + \lambda||\boldsymbol{\Omega}||_1$$

The idea is to formulate the problem a "collection" node-wise lasso regression problems

Starting from $u = 1$, consider $\widetilde{\boldsymbol{\Sigma}}$ (an "initial" estimate of $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$) and partition $\widetilde{\boldsymbol{\Sigma}}$ and $\boldsymbol{S}$ as

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \widetilde{\boldsymbol{\Sigma}}_{11} & \widetilde{\boldsymbol{\Sigma}}_{12} \\ \widetilde{\boldsymbol{\Sigma}}_{21} & \widetilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix} \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}$$

Graphical Lasso for UG model selection

Banerjee et al. (2007) show that the GLasso problem is equivalent to

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}||\widetilde{\boldsymbol{\Sigma}}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{b}||^2 + \lambda||\boldsymbol{\beta}||_1$$

where $\boldsymbol{b} = \widetilde{\boldsymbol{\Sigma}}_{11}^{-1}\boldsymbol{S}_{12}$ and $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$

Same procedure is repeated over $u \in \{1, \dots, q\}$

From the above collection of estimated "regression" coefficients,

elements of a sparse inverse covariance matrix $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Sigma}}^{-1}$ can be recovered using $\widehat{\boldsymbol{\Sigma}}_{12} = \widehat{\boldsymbol{\Sigma}}_{11}\boldsymbol{\beta}$

References

Banerjee, O., Ghaoui, L.E. & d'Aspremont, A. (2007).
Model selection through sparse maximum likelihood.
Journal of Machine Learning Research **9**, 485-516.

Chickering, D.M. (2002).
Optimal Structure Identification With Greedy Search.
Journal of Machine Learning Research **3(3)**, 507-554.

Friedman, J., Hastie, T. & Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics **9(3)**, 432–441.

Kalisch, M. & Bühlmann, P. (2007).
Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
Journal of Machine Learning Research **8**, 613-636.

Spirtes, P., Glymour, C. & Scheines, R. (2000).
Causation, Prediction and Search (2nd edition).
Cambridge, MA: The MIT Press.