Bayesian constrained-based structure learning

Background

1. Multivariate Gaussian data

$$\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n \mid \Omega \overset{iid}{\sim} N_q(\underset{\sim}{0}, \Omega^{-1}) \qquad \Omega \in \mathcal{P} \quad \text{s.p.d. matrices}$$

$$\Omega \sim W_q(a, U)$$

$$p(\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n \mid \Omega) = (2\pi)^{-\frac{np}{2}} |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}(S\Omega)\right\}$$

$$S = \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^T$$

$$p(\Omega) = c(a, U) |\Omega|^{\frac{a-p-1}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}(U\Omega)\right\}$$

$$c(a, U) = \frac{|U|^{\frac{a}{2}}}{2^{\frac{aq}{2}} \Gamma_q\left(\frac{a}{2}\right)}$$

prior normalizing constant

↳ multivariate Gamma

The marginal data distribution (i.e. the marginal likelihood) is :

$$p(\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n) = \int p(\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n \mid \Omega)\, p(\Omega)\, d\Omega$$

$$= \frac{c(a, U)}{c(a+n, U+S)} \cdot (2\pi)^{-\frac{nq}{2}} := p(X)$$

$c(a+n, U+S)$ posterior normalising constant

Now consider $A \subseteq \{1, \dots, q\}$. We have:

$$\underset{\sim}{x}_1^A, \dots, \underset{\sim}{x}_n^A \mid \Omega_{A|\bar{A}} \overset{iid}{\sim} \mathcal{N}_{|A|}\left(\underset{\sim}{0}, (\Omega_{A|\bar{A}})^{-1}\right)$$

$$\Omega_{A|\bar{A}} \sim W_{|A|}\left(a - |\bar{A}|, U_{AA}\right)$$

with $x_i^A = (x_{ij})_{j \in A}$ $\quad U_{AA} = [U_{jj}]_{j \in A}$

$$\Omega_{A|\bar{A}} = \Omega_{AA} - \Omega_{A\bar{A}} (\Omega_{\bar{A}\bar{A}})^{-1} \Omega_{\bar{A}A}$$

$$= (\Sigma_{AA})^{-1}$$

i.e. $(\Omega_{A|\bar{A}})^{-1} = \Sigma_{AA}$
the marginal covariance matrix

Therefore we obtain :

$$p(\underset{\sim}{x}_1^A, \ldots, \underset{\sim}{x}_n^A) = \int p(\underset{\sim}{x}_1^A, \ldots, \underset{\sim}{x}_n^A \mid \Omega_{A|\bar{A}})\, p(\Omega_{A|\bar{A}})\, d\Omega_{A|\bar{A}}$$

$$= \frac{c(a-|\bar{A}|, U_{AA})}{c(a-|\bar{A}|+n, U_{AA}+S_{AA})} \cdot (2\pi)^{-\frac{n|A|}{2}}$$

$$:= p(X_A) \qquad \ast$$

Refs : Press (1982) "Applied Multivariate Analysis"

Consonni & La Rocca (2012, SJS)

2. Bayesian DAG-model selection

$D = (V, E)$ a DAG with $V = \{1, \dots, p\}$
$E \subseteq V \times V$

$pa_D(j) = \{u : (u,j) \in E\}$ parents of j in D

$fa_D(j) = j \cup pa_D(j)$ family of j in D

Under D we have:

$$p(x_1, \dots, x_p \mid D) = \prod_{j=1}^{p} p(x_j \mid \underset{\sim}{x}_{pa_D(j)})$$

In a Gaussian DAG model:

$$\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n \mid \Omega_D \sim \mathcal{N}_p(\underset{\sim}{0}, \Omega_D^{-1})$$

$$\Omega_D \sim p(\Omega_D)$$

$\Omega_D \in \mathcal{P}_D$ precision matrices Markov w.r.t. D

Which prior for Ω_D?

If the goal is to compute the DAG marginal likelihood

$$p(X \mid D) = \int p(\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n \mid \Omega_D)\, p(\Omega_D)\, d\Omega_D$$

then there is no need to specify "directly" $p(\Omega_D)$ but rather a prior on Ω just s.p.d. and some assumptions under which we recover $p(X \mid D)$ as:

$$p(X \mid D) = \prod_{j=1}^{q} \left\{ \frac{p(X_{fa_D(j)})}{p(X_{pa_D(j)})} \right\}$$

and $p(X_{fa_D(j)})$, $p(X_{pa_D(j)})$ are as in

with $A = fa_D(j)$, $A = pa_D(j)$ (*)

Refs: Geiger & Heckerman (2002, AOS)

Consonni & La Rocca (2012, SJS)

Created with Notewise

3. The PC Algorithm for DAG estimation

Refs: Kalisch & Buhlmann (2007, JMLR)

Slides on Graphical Models

Idea in PC Algorithm is to recover a CPDAG (Completed Partially DAG) through a sequence of Conditional Independence (CI) tests

Specifically, in Step 1 (skeleton estimation) the PC algorithm remove an edge $u - v$ if

$$X_u \perp\!\!\!\perp X_v \mid X_S \qquad \text{for some } S \subseteq V \setminus \{u,v\}$$

i.e. for at least one set S not including u and v

Idea in practice is to start from a set S of size 0 (i.e. $\emptyset$), then increase it by one and so on; we stop when we find a set S for which $X_u \perp\!\!\!\perp X_v \mid X_S$.

Which CI test?

For Gaussian data, a test based on partial correlation coefficients $\rho_{uv|S}$.

What is $\rho_{uv|S}$?

It is the $\mathrm{Corr}(X_u, X_v \mid X_S)$ that is the correlation coefficient between X_u and X_v in the joint, conditional distribution

$$(X_u, X_v) \mid X_S$$

Start from $X_A = (X_j)_{j \in A}$ with $A = \{u, v, S\}$

for which we know that

$$X_A \mid \Omega_{A|\bar{A}} \sim N_{|A|}\left(\underset{\sim}{0}, (\Omega_{A|\bar{A}})^{-1}\right)$$

and $\Omega_{A|\bar{A}}^{-1} = \Sigma_{AA}$

Then, we partition A into $\{u, v\}$ and S and Σ_{AA} accordingly

We consider the conditional distribution of $X_{\{u,v\}}$ given X_S :

$$X_{\{u,v\}} \mid X_S \sim \mathcal{N}_2\left(\mu_{\{u,v\}|S} \,;\, \Sigma_{\{u,v\}|S}\right)$$

with $\mu_{\{u,v\}|S} = \Sigma_{S,\{u,v\}}\left(\Sigma_{\{u,v\}\{u,v\}}\right)^{-1}\Sigma_{\{u,v\},S}$

$$\Sigma_{\{u,v\}|S} = \Sigma_{\{u,v\}\{u,v\}} - \Sigma_{\{u,v\},S}\left(\Sigma_{S,S}\right)^{-1}\Sigma_{S,\{u,v\}}$$

Remark : the out-diagonal element of $\Sigma_{\{u,v\}|S}$ is the covariance between X_u and X_v in the conditional distribution of (X_u, X_v) given X_S : $\underline{\mathrm{Cov}(X_u, X_v \mid X_S)}$

The partial correlation coefficient is then :

$$\mathrm{Corr}(X_u, X_v \mid X_S) \overset{\text{def.}}{=} \frac{\left[\Sigma_{\{u,v\}|S}\right]_{u,v}^{\;1,2}}{\sqrt{\underset{1,1}{\left[\Sigma_{\{u,v\}|S}\right]_{u,u}}\;\underset{2,2}{\left[\Sigma_{\{u,v\}|S}\right]_{v,v}}}}$$

$$\Vert$$

$$\rho_{\{u,v\}|S}$$

$\rho_{\{u,v\}|S} = 0 \quad \text{iff} \quad X_u \perp\!\!\!\perp X_v \mid X_S$

and $\rho_{\{u,v\}|S} = 0 \quad \text{iff} \quad [\Sigma_{\{u,v\}|S}]_{u,v} = 0$

Now, the joint distribution of $X_{\{u,v\}} \mid X_S$ can be equivalently parameterize as:

$$\Sigma_{\{u,v\}|S} \mapsto \{ L_{u|v,S} \,;\, D_{u|v,S} , D_{v|S} \}$$

where $L_{u|v,S}$ is the regression coefficient in the regression of X_u on $X_v \cup X_S$, while $D_{u|v,S}$ and $D_{v|S}$ are the conditional variances.

$$L_{u|v,S} = [\Sigma_{\{u,v\}|S}]_{u,\{v,S\}} \left[[\Sigma_{\{u,v\}|S}]_{\{v,S\}\{v,S\}} \right]^{-1}$$

This should be equivalent as doing the following.

Consider

$$\mathcal{D}_0 \qquad \begin{tikzcd} & S \arrow[dl] \arrow[dr] & \\ u & & v \end{tikzcd} \qquad\qquad \mathcal{D}_1 \qquad \begin{tikzcd} & S \arrow[dl] \arrow[dr] & \\ u \arrow[rr] & & v \end{tikzcd}$$

$$BF_{01} = \frac{m(X \mid \mathcal{D}_0)}{m(X \mid \mathcal{D}_1)} = \frac{\cancel{p(X_u \mid X_s)}\, p(X_v \mid X_s)\, \cancel{p(X_s)}}{\cancel{p(X_u \mid X_s)}\, p(X_v \mid X_u, X_s)\, \cancel{p(X_s)}} =$$

$$= \frac{p(X_v, X_s) / p(X_s)}{p(X_v, X_u, X_s) / p(X_u, X_s)} =$$

$$= \frac{p(X_{fa_{\mathcal{D}_0}(v)}) / p(X_{pa_{\mathcal{D}_0}(v)})}{p(X_{fa_{\mathcal{D}_1}(v)}) / p(X_{pa_{\mathcal{D}_1}(v)})}$$

Similarly to what is given in Geiger & Heckermann (2002) it should coincide with the ratio of two marginal likelihoods of two complete DAGs.

Uncertainty quantification

PCalg does not allow for uncertainty quantification underlying the estimated graphical structure.

Here instead we have $p(u \to v \mid X)$

Let $p(u \to v)$ and $p(u \not\to v)$ be prior probabilities of having or not an edge between u and v.

$$p(u \to v \mid X) = \frac{\cancel{m(X \mid u \to v)\, p(u \to v)}}{m(X \mid u \to v)\, p(u \to v) + m(X \mid u \not\to v)\, p(u \not\to v)} \times \frac{m(X \mid u \to v)\, p(u \to v)}{\cancel{m(X \mid u \to v)\, p(u \to v)}} =$$

$$= \frac{1}{\dfrac{m(X \mid u \to v)\, p(u \to v) + m(X \mid u \not\to v)\, p(u \not\to v)}{m(X \mid u \to v)\, p(u \to v)}} =$$

$$= \frac{1}{1 + BF_{01} \dfrac{p(u \not\to v)}{p(u \to v)}}$$

ie given prior probabilities of edges and the BF we can compute $p(u \to v \mid X)$

If :

- $BF_{01} = 0 \Rightarrow m(X \mid \mathcal{D}_0) = 0 \Rightarrow p(u \to v \mid X) = 1$

- $BF_{01} \to +\infty \Rightarrow m(X \mid \mathcal{D}_1) = 0 \Rightarrow p(u \to v \mid X) = 0$