Statistical Modelling - Module I
Graphical models
Lecture 4

Federico Castelletti

Department of Statistical Sciences
Università Cattolica del Sacro Cuore
Milan

Introduction

DAGs (aka Bayesian networks)
correspond to a broader class of Conditional Independence (CI) models
w.r.t. decomposable graphs

In particular, the space of decomposable graphs
is a subset of the DAG space (in terms of CI models)

DAGs are also widely employed in biology, social science and psychology
as an alternative to Structural Equation Models (SEMs)
and for causal inference purposes

Literature on statistical methodologies for DAG structure learning has grown starting from the 90's

Introduction

Heckerman & Geiger (1995) and Geiger & Heckerman (2002)
propose a constructive method to assign priors to categorical and Gaussian DAG models
starting from a small number of assumptions

They derive two scores named

- BDe (Bayesian Dirichlet equivalent, for categorical DAGs)
- BGe (Bayesian Gaussian equivalent, for Gaussian DAGs)

which importantly are equivalent for any two Markov equivalent DAGs

Implementation of these scores within MCMC schemes (such as the one that we discussed in Slides 3) allows for structure learning of DAGs

Alternatively, one can consider other scores (e.g. based on penalized likelihoods, such as the BIC) and implement them into frequentist score-based algorithms

We describe both approaches in the next slides

Structure learning of DAGs

$\mathcal{D} = (V, E)$ DAG with nodes $V = \{1, \ldots, q\}$

$\boldsymbol{x} = (X_1, \ldots, X_q)$ random variables

Under $\mathcal{D}$ their joint distribution $f(\cdot)$ factorizes as

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{\theta}, \mathcal{D}) = \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{\theta}_j)$$

$\mathrm{pa}_{\mathcal{D}}(j)$ are the parents of node j in DAG $\mathcal{D}$

$\boldsymbol{\theta}$ is a *global* parameter indexing the DAG (and so it is DAG-dependent)
e.g. a covariance matrix or a contingency table of probabilities

$\boldsymbol{\theta}_j$ is a *local* (node-specific) parameter indexing the conditional distribution of variable X_j

DAG structure learning

Likelihood

Suppose we have n i.i.d. samples from a DAG model

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ with $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,q})^\top$

and let $\boldsymbol{X}$ be the (n, q) data matrix whose (i, j)-element is $x_{i,j}$

The likelihood function is then

$$
\begin{aligned}
f(\boldsymbol{X} \mid \boldsymbol{\theta}, \mathcal{D}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{q} f(x_{i,j} \mid \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{\theta}_j) \right\} \\
&= \prod_{j=1}^{q} f(\boldsymbol{X}_j \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{\theta}_j)
\end{aligned}
$$

with $\boldsymbol{X}_A$ sub-matrix of $\boldsymbol{X}$ with columns indexed by $A \subseteq \{1, \dots, q\}$

We would like to assign a score to any $\mathcal{D} \in \mathcal{S}_q$, say $\mathrm{Sc}(\mathcal{D})$,

where $\mathcal{S}_q$ is the set of all DAGs on q nodes

This score will depend on the specific form of the joint distribution $f(\cdot)$

for which we will (again) distinguish between

- Gaussian DAGs
- Categorical DAGs

Also, we can derive "scores" following both

- a frequentist approach
- a Bayesian approach

In the former case, such score can be implemented in an optimization (score-based) algorithm for DAG estimation

In the second case, such score (marginal likelihood) can be used within an MCMC scheme for posterior inference on DAGs

Scoring DAGs: the Bayesian way

To start with, we need to assign a prior to $\boldsymbol{\theta}$

Since $\boldsymbol{\theta}$ is DAG-dependent we write it as $p(\boldsymbol{\theta} \mid \mathcal{D})$

Geiger & Heckerman (1995, 2002) show that under some general assumptions (details later on)

it is possible to induce $p(\boldsymbol{\theta} \mid \mathcal{D})$ from a single (unique) prior
on the parameter of an unconstrained (complete) DAG $p(\boldsymbol{\theta})$

Assumptions above refer to both the likelihood and the prior

They show that they are satisfied by

- Gaussian DAG models with (Normal)-Wishart priors
- Categorical DAG models with Dirichlet priors

Scoring DAGs: the Bayesian way

Assumptions (just a sketch) are the following

On the sampling distribution:

1. *complete model equivalence*
2. *regularity*
3. *likelihood modularity*

On the prior:

4. *prior modularity*
5. *global parameter independence*

Main result is summarized in Theorem 2 of Geiger & Heckerman (2002) showing that
it is sufficient to specify a prior for the parameter of a complete DAG model:
the parameter prior for any other (not complete) DAG model can be derived automatically
(as in the next slide)

Scoring DAGs: the Bayesian way

If Assumptions 1:5 hold, the marginal likelihood can be recovered as:

$$\begin{aligned} m(\boldsymbol{X} \mid \mathcal{D}) &= \int f(\boldsymbol{X} \mid \boldsymbol{\theta}, \mathcal{D})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \prod_{j=1}^{q} \int f(\boldsymbol{X}_j \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{\theta}_j)\, p(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j \\ &= \prod_{j=1}^{q} m(\boldsymbol{X}_j \mid \boldsymbol{X}_{\mathrm{pa}(j)}, \mathcal{D}) \;=\; \prod_{j=1}^{q} \frac{m\big(\boldsymbol{X}_{\mathrm{fa}_{\mathcal{D}}(j)}\big)}{m\big(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}\big)} \end{aligned}$$

with $\mathrm{fa}_{\mathcal{D}}(j) = j \cup \mathrm{pa}_{\mathcal{D}}(j)$ family of node j in $\mathcal{D}$

and where (this is the important point)

each term $m(\boldsymbol{X}_A)$ corresponds to the marginal likelihood for variables in A
computed under an *unconstrained* (complete) DAG

Obs: the same kind of object that we used to compute the marginal likelihood of decomposable graphs

Instead of cliques and separators we now have for each node its *family* and *parents*

Scoring DAGs: the Bayesian way

Gaussian DAGs

Model assumption is

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \mid \boldsymbol{\Omega} \overset{\text{iid}}{\sim} \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1})$

Prior on $\boldsymbol{\Omega}$ under a complete DAG is a Wishart:

$\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U})$

Remember (Lecture 3) that under such model and prior
we have a closed-form expression for the marginal likelihood relative to any $\boldsymbol{X}_A$, $A \subseteq V$:

$$m(\boldsymbol{X}_A) = (\pi)^{-\frac{n|A|}{2}} \frac{|\boldsymbol{U}_A|^{\frac{a}{2}}}{|\boldsymbol{U}_A + \boldsymbol{S}_A|^{\frac{a-|\bar{A}|+n}{2}}} \frac{\Gamma_{|A|}\left(\frac{a-|\bar{A}|+n}{2}\right)}{\Gamma_{|A|}\left(\frac{a}{2}\right)}$$

$\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$
$\boldsymbol{S}_A$ and $\boldsymbol{U}_A$ submatrices with cols/rows indexed by A

By plugging in this formula in $m(\boldsymbol{X} \mid \mathcal{D})$ we specialize the marginal likelihood to the Gaussian case

The resulting score is called BGe (Bayesian Gaussian equivalent score)

Scoring DAGs: the Bayesian way

Categorical DAGs

Model assumption is

$$p(\boldsymbol{x}^{(i)}) = \prod_{x \in \mathcal{X}} \left\{ \Pr\left(X_1^{(i)} = x_1^{(i)}, \ldots, X_q^{(i)} = x_q^{(i)} \,|\, \boldsymbol{\Pi}\right) \right\}^{\mathbb{1}(\boldsymbol{x}^{(i)} = x)} \quad i = 1, \ldots, n$$

(a generalized Bernoulli over the product space $\mathcal{X}$)

for any level $x \in \mathcal{X}$ of $(X_1, \ldots, X_q)$ and independently across i

Prior on $\boldsymbol{\Pi}$ under a complete DAG is

$\boldsymbol{\Pi} \sim \text{Dirichlet}(\boldsymbol{a})$

Still (Lecture 3), under such model and prior
we have a closed-form expression for the marginal likelihood relative to any $\boldsymbol{X}_A$, $A \subseteq V$:

$$m(\boldsymbol{X}_A) \;=\; \frac{\Gamma\big(\sum_{x_A \in \mathcal{X}_A} a(x_A)\big)}{\Gamma\big(\sum_{x_A \in \mathcal{X}_A} a(x_A) + n(x_A)\big)} \frac{\prod_{x_A \in \mathcal{X}_A} \Gamma\big(a(x_A) + n(x_A)\big)}{\prod_{x_A \in \mathcal{X}_A} \Gamma\big(a(x_A)\big)}$$

By plugging in this formula in $m(\boldsymbol{X} \,|\, \mathcal{D})$ we specialize the marginal likelihood to the categorical case

The resulting score is called BDe (Bayesian Dirichlet equivalent score)

For both (Gaussian and categorical) settings we can structure an MCMC to target

$$p(\mathcal{D} \mid \boldsymbol{X}) \propto m(\boldsymbol{X} \mid \mathcal{D})\, p(\mathcal{D}) \quad \mathcal{D} \in \mathcal{S}_q$$

the posterior of $\mathcal{D} \in \mathcal{S}_q$

We can consider, as in Lecture 3, a Metropolis Hastings algorithm
with acceptance probability of $\mathcal{D}'$ given $\mathcal{D}$

$$\alpha(\mathcal{D}' \mid \mathcal{D}) = \min\left\{1; \frac{m(\boldsymbol{X} \mid \mathcal{D}')}{m(\boldsymbol{X} \mid \mathcal{D})} \cdot \frac{p(\mathcal{D}')}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \mid \mathcal{D}')}{q(\mathcal{D}' \mid \mathcal{D})}\right\}$$

$q(\mathcal{D}' \mid \mathcal{D})$ suitable proposal distribution when the chain is at DAG $\mathcal{D}$

$p(\mathcal{D})$ prior on $\mathcal{D}$ (can be assigned "similarly" as in the UG case)

See Algorithm 1 in Lecture 3

For Gaussian DAGs, the previous MCMC scheme (together with many other functions)

is implemented in the `R` package `BCDAG`

`Bayesian Structure and Causal Learning of Gaussian Directed Graphs`

by the `learnDAG` function under the setting `collapse = TRUE`

Obs: with `collapse = FALSE` you can also sample from the posterior of the DAG-parameters

Other inputs are:

- `data` the (n, q) data matrix $\boldsymbol{X}$
- `S` the number of MCMC iterations
- `burn` the burn-in period
- `a,U` hyperparameters of the DAG-Wishart prior
- `w` prior probability of edge inclusion of $p(\mathcal{D})$

See `R` code

Scoring DAGs: the frequentist way

In general, as a frequentist score, we can use a penalized likelihood leading to

- an Akaike Information Criterion (AIC)
- a Bayesian Information Criterion (BIC)

Both consider the likelihood function evaluated at the ML estimator of the DAG-parameter $\boldsymbol{\theta}$ to which a (different) penalty term is applied

$$\begin{aligned}\mathrm{AIC}(\mathcal{D}) &= \log f(\boldsymbol{X} \mid \widehat{\boldsymbol{\theta}}, \mathcal{D}) - d \\ \mathrm{BIC}(\mathcal{D}) &= \log f(\boldsymbol{X} \mid \widehat{\boldsymbol{\theta}}, \mathcal{D}) - \frac{d}{2}\log(n)\end{aligned}$$

d number of parameters

n number of observations

$\widehat{\boldsymbol{\theta}}$ MLE of $\boldsymbol{\theta}$

We can specialize each score to categorical and Gaussian DAGs

Scoring DAGs: the frequentist way

Categorical DAGs

Remember our notation for categorical data (Lecture 3) and also the following

$\pi^{j \mid S}_{x_j \mid x_S} = \Pr(X_j = x_j \mid X_S = x_S)$ conditional probability for

variable X_j evaluated at x_j

given configuration x_S of variables in S, $j \notin S$

$n_x = \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}^{(i)} = x)$ number of observations that are equal to x

$\boldsymbol{N} = \{n_x, x \in \mathcal{X}\}$ resulting q-way contingency table of counts

For any $S \subseteq V$ and level $x_S \in \mathcal{X}_S$

$n^S_{x_S} = \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}^{(i)}_S = x_S)$

$\boldsymbol{N}_S = \{n^S_{x_S}, x_S \in \mathcal{X}_S\}$ allied $|S|$-way marginal contingency table of counts

Scoring DAGs: the frequentist way

Categorical DAGs

Then, the likelihood can be written as

$$
\begin{aligned}
p(\boldsymbol{X} \mid \boldsymbol{\Pi}, \mathcal{D}) &= \prod_{i=1}^{n} \left\{ \prod_{x \in \mathcal{X}} \left\{ \Pr\left(X_1^{(i)} = x_1^{(i)}, \ldots, X_q^{(i)} = x_q^{(i)} \mid \boldsymbol{\Pi} \right) \right\}^{\mathbb{1}(\boldsymbol{x}^{(i)} = x)} \right\} \\
&= \prod_{j=1}^{q} \left\{ \prod_{k \in \mathcal{X}_{\mathrm{pa}(j)}} \left\{ \prod_{m \in \mathcal{X}_j} \left\{ \pi_{m \mid k}^{j \mid \mathrm{pa}(j)} \right\}^{n_{(m,k)}^{\mathrm{fa}(j)}} \right\} \right\}
\end{aligned}
$$

Each conditional probability

can be estimated using the corresponding sample proportion estimator

$$
\widehat{\pi}_{m \mid k}^{j \mid \mathrm{pa}(j)} = \frac{\#(X_j = m \wedge X_{\mathrm{pa}(j)} = k)}{\#(X_{\mathrm{pa}(j)} = k)}
$$

to obtain $p(\boldsymbol{X} \mid \widehat{\boldsymbol{\Pi}}, \mathcal{D})$, which is finally plugged in into the AIC or BIC

Scoring DAGs: the frequentist way

Gaussian DAGs

Given the SEM reparameterization of $\boldsymbol{\Sigma}$ with

- $\boldsymbol{B}$ (q, q) matrix of regression coefficients
- $\boldsymbol{D}$ (q, q) diagonal matrix with conditional variances $\sigma_1^2, \ldots, \sigma_q^2$

the likelihood can be written as

$$\begin{aligned} f(\boldsymbol{X} \mid \boldsymbol{D}, \boldsymbol{B}, \mathcal{D}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{q} d\,\mathcal{N}\big(x_{i,j} \mid \boldsymbol{B}_{\mathrm{pa}(j),j}^{\top} \boldsymbol{x}_{i,\mathrm{pa}(j)}, \sigma_j^2\big) \right\} \\ &= \prod_{j=1}^{q} d\,\mathcal{N}_n\big(\boldsymbol{X}_j \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{B}_{\mathrm{pa}(j),j}, \sigma_j^2 \boldsymbol{I}_n\big) \end{aligned}$$

$\boldsymbol{B}_{\mathrm{pa}(j),j}$ elements of $\boldsymbol{B}$ with rows indexed by $\mathrm{pa}(j)$ and column j

Each of the q terms is the likelihood of a normal linear regression model
for which we can compute the MLEs $\widehat{\boldsymbol{B}}_{\mathrm{pa}(j),j}$ and $\widehat{\sigma}_j^2$, $j = 1, \ldots, q$
and so obtain $f(\boldsymbol{X} \mid \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{D}}, \mathcal{D})$, which is plugged in into the AIC or BIC

Scoring DAGs: the frequentist way

Hill Climbing algorithm

For frequentist DAG estimation we can consider a score-based approach

based on the optimization of the AIC or BIC (or even the marginal likelihood)

The resulting algorithm is called *Hill Climbing* (HC) and works as follows:

Algorithm 1 Hill Climbing

Input: $\mathcal{D}^{(0)}$ (an arbitrary initial DAG)
Output: $\widehat{\mathcal{D}}$ (the estimated DAG)

1. Set $\mathcal{D} = \mathcal{D}^{(0)}$
2. Compute the score of $\mathcal{D}$, $\text{Sc}(\mathcal{D})$
3. Repeat the following steps, as long as $\text{Sc}(\mathcal{D})$ increases

    3.1 For every possible arc addition, deletion or reversal not resulting in a cyclic network:

    3.1.1 Compute the score of the modified DAG $\mathcal{D}^*$, $\text{Sc}(\mathcal{D}^*)$

    3.1.2 If $\text{Sc}(\mathcal{D}^*) > \text{Sc}(\mathcal{D})$ set $\mathcal{D} = \mathcal{D}^*$ and $\text{Sc}(\mathcal{D}) = \text{Sc}(\mathcal{D}^*)$

4. Return $\mathcal{D}$

where $\text{Sc}(\mathcal{D})$ is a DAG score (AIC, BIC, . . .)

References

CAO, X., KHARE, K. & GHOSH, M. (2019).
Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models.
The Annals of Statistics, **47(1)**, 319–348.

CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M.L. & PELUSO, S. (2018).
Learning Markov Equivalence Classes of Directed Acyclic Graphs: An Objective Bayes Approach.
Bayesian Analysis, **13(4)**, 1235–1260.

CASTELLETTI, F. and CONSONNI, G. (2021).
Bayesian inference of causal effects from observational data in Gaussian graphical models.
Biometrics **77(1)**, 136–149.

CASTELLETTI, F. & CONSONNI, G. (2019).
Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways.
The Annals of Applied Statistics, **13(4)**, 2289–2311.

CASTELLETTI, F. & CONSONNI, G. (2022+).
Bayesian sample size determination for causal discovery.
arXiv preprint `https://arxiv.org/abs/2206.00755`

CASTELLETTI, F. & MASCARO, A. (2022+).
BCDAG: An R package for Bayesian structure and Causal learning of Gaussian DAGs.
arXiv preprint `https://arxiv.org/abs/2201.12003`

References

CASTELLETTI, F. & PELUSO, S. (2021).
Equivalence class selection of categorical graphical models.
Computational Statistics & Data Analysis, **164**, 107304.

CONSONNI, G. & PELUSO, S. (2020).
Compatible priors for model selection of high-dimensional Gaussian DAGs.
Electronic Journal of Statistics, **14(2)**, 4110–4132.

GEIGER, D. & HECKERMAN, D. (2002).
Parameter priors for directed acyclic graphical models and the characterization of several probability distributions.
The Annals of Statistics, **30**, 1412–1440.

HAUSER, A. & BÜHLMANN, P. (2015).
Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs.
Journal of the Royal Statistical Society Series B, **77(1)**, 291–318.

HECKERMAN, D., GEIGER, D. & CHICKERING, M.D. (1995).
Learning Bayesian networks: the combination of knowledge and statistical data.
Machine Learning, **20**, 197–243.

HEINZE-DEML, C., MAATHUIS, M. & MEINSHAUSEN, N. (2018).
Causal structure learning.
Annual Review of Statistics and Its Application **5**, 371–391.

References

MAATHUIS, M.H., KALISCH, M. & BÜHLMANN, P. (2009).
Estimating high-dimensional intervention effects from observational data.
The Annals of Statistics **37(6A)**, 3133–3164.

PEARL, J. (2000).
Causality: models, reasoning, and inference.
Cambridge University Press.

PEARL, J. (2003).
Statistics and causal inference: A review.
Test **12**, 281–345.

SCUTARI, M. (2003).
Learning Bayesian Networks with the bnlearn R Package.
Journal of Statistical Software **35(3)**, 1–22.