732A96 Advanced Machine Learning

Graphical Models and Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 4: Structure Learning

Contents

Literature

- Main sources
  - Koski, T. J. T. and Noble, J. M. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda* 40, 51-103, 2012.
- Other sources
  - Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
  - Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- R resources
  - Nagarajan, R., Scutari, M. and Lébre, S. *Bayesian Networks in R*. Springer, 2013.
  - Scutari, S. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35, 2010.

Structure Learning for BNs: Independence Test Based Approach

- We can get a DAG H such that $p(X) = \prod_i p(X_i | pa_H(X_i))$ and, thus, that we can use for probabilistic reasoning as follows:

Let $Y_{1:n}$ be any ordering of the random variables $X_{1:n}$
For each Y_i do
    Set $pa_H(Y_i)$ to be any minimal subset of $Y_{1:i-1}$ such that
    $Y_i \perp_p Y_{1:i-1} \setminus pa_H(Y_i) | pa_H(Y_i)$

- **Exercise**. Prove the previous statement.
- Note that H has the minimum number of edges among the DAGs that are consistent with the ordering considered.
- However, H may not have the minimum number of edges among all the DAGs, i.e. the ordering considered may not be optimal.

$A \perp_p B$	H with ordering A, B, C	H with ordering C, A, B
$\begin{tikzcd} A \arrow[dr] & & B \arrow[dl] \\ & C & \end{tikzcd}$	$\begin{tikzcd} A \arrow[dr] & & B \arrow[dl] \\ & C & \end{tikzcd}$	$\begin{tikzcd} A \arrow[rr] & & B \\ & C \arrow[ul] \arrow[ur] & \end{tikzcd}$

- We can get one such optimal DAG H without searching over the $n!$ orderings **assuming** faithfulness, i.e. $I(p) = I(G)$ for some DAG G.

Structure Learning for BNs: Independence Test Based Approach

Parents and children (PC) algorithm

Let H be the complete undirected graph
$l := 0$
Repeat while $l \leq n-2$
    For each ordered pair of nodes X_i and X_j in H such that $X_i \in ad_H(X_j)$ and $|ad_H(X_i) \setminus X_j| \geq l$
        If there is some $S \subseteq ad_H(X_i) \setminus X_j$ such that $|S| = l$ and $X_i \perp_P X_j | S$, then
            $S_{ij} := S_{ji} := S$
            Remove the edge $X_i - X_j$ from H
    $l := l + 1$
Apply the rule R1 to H while possible
Apply the rules R2-R3 to H while possible

R1: $X_i \longrightarrow X_j \longrightarrow X_k$ with X_i — X_j — X_k undirected:

$$X_i - X_j - X_k \quad \Rightarrow \quad X_i \longrightarrow X_j \longleftarrow X_k$$
$$\wedge \ B \notin S_{ik}$$

R2: $$X_i \longrightarrow X_j - X_k \quad \Rightarrow \quad X_i \longrightarrow X_j \longrightarrow X_k$$

R3: $$\begin{tikzcd} X_i \arrow[r] \arrow[rr, bend left, dash] & X_j \arrow[r] & X_k \end{tikzcd} \quad \Rightarrow \quad \begin{tikzcd} X_i \arrow[r] \arrow[rr, bend left] & X_j \arrow[r] & X_k \end{tikzcd}$$

R4: $$\begin{tikzcd} & X_k \arrow[dr] & \\ X_i \arrow[ur, dash] \arrow[rr, dash] \arrow[dr, dash] & & X_j \\ & X_l \arrow[ur] & \end{tikzcd} \quad \Rightarrow \quad \begin{tikzcd} & X_k \arrow[dr] & \\ X_i \arrow[ur, dash] \arrow[rr] \arrow[dr, dash] & & X_j \\ & X_l \arrow[ur] & \end{tikzcd}$$

Structure Learning for BNs: Independence Test Based Approach

- In practice, we do not have access to p but to a finite sample from it. Then, replace $X_i \perp_p X_j | S$ in the PC algorithm with an independece test, preferably with one that is consistent so that the algorithm is **asymptotically** correct.
- Let $d_{1:N}$ be a complete sample. Then, $X_i \perp_p X_j | S$ implies that

$$N_{x_i, x_j, s} \approx N\, p(x_i|s) p(x_j|s)$$

  where $N_{x_i, x_j, s}$ is the number of instances in $d_{1:N}$ where x_i, x_j and s.
- We can measure the deviance from the expected situation above by

$$deviance = \sum_{a,b,s} \frac{[N_{x_i, x_j, s} - N\, p(x_i|s) p(x_j|s)]^2}{N\, p(x_i|s) p(x_j|s)}$$

- If the deviance is too large, then reject the hypothesis that $X_i \perp_p X_j | S$.
- Asymptotically, the deviance follows a χ^2 distribution with the appropriate number of degrees of freedom. Then, we can control the probability of falsely rejecting the hypothesis, a.k.a. p-value.

Structure Learning for BNs: Independence Test Based Approach

- **Exercise**. Run the PC algorithm assuming that p is faithful to the following DAGs.

$$\begin{tikzcd}
A \arrow[d] & B \arrow[dl] \arrow[dr] & C \arrow[d] \\
D \arrow[d] & & E \arrow[d] \\
F \arrow[d] \arrow[dr] & & G \arrow[dl] \arrow[d] \\
H & I & J
\end{tikzcd}
\qquad
\begin{tikzcd}
A & B \arrow[dl] \arrow[dr] & C \arrow[d] \\
D \arrow[u] \arrow[d] & & E \arrow[d] \\
F \arrow[d] \arrow[dr] & & G \arrow[dl] \arrow[d] \\
H & I & J
\end{tikzcd}$$

Structure Learning for BNs: Independence Test Based Approach

- The PC algorithm relies on the faithfulness assumption. Specifically
  - if $X_i \in ad_G(X_j)$ then $X_i \not\perp_p X_j | S$ for all S and, then, $X_i \in ad_H(X_j)$ **always**,
  - otherwise $X_i \perp_p X_j | pa_G(X_i)$ or $X_i \perp_p X_j | pa_G(X_j)$ and, then, $X_i \notin ad_H(X_j)$ at some point because $pa_G(X_i) \subseteq ad_H(X_i)$ and $pa_G(X_j) \subseteq ad_H(X_j)$ **always**.
- Two DAGs represent the same independencies (i.e. they are **equivalent**) if and only if they have the same adjacencies and **unshielded colliders**, i.e. subgraphs $X_i \rightarrow X_k \leftarrow X_j$ where X_i and X_j are not adjacent.
- The output of the PC algorithm is not a DAG in general, but an **essential graph** (EG):
  - H has an edge $X_i \rightarrow X_j$ if and only if $X_i \rightarrow X_j$ is in **every** DAG that is equivalent to G.
  - In other words, H has an edge $X_i - X_j$ if and only if $X_i \rightarrow X_j$ is in some DAG that is equivalent to G and $X_i \leftarrow X_j$ is in some other DAG that is equivalent to G.
- A naive way to convert H into a DAG that is equivalent to G is as follows:

Repeat while possible

  Replace any edge $X_i - X_j$ in H with $X_i \rightarrow X_j$ if this does not create a directed cycle or a new unshielded collider

If H is not a DAG, then backtrack

- **Exercise**. Prove that the previous algorithm is correct.

Structure Learning for BNs: Score Based Approach

- Alternatively, we can choose the DAG G with maximum posterior probability (a.k.a **Bayesian score**):

$$p(G|d_{1:N}) = p(d_{1:N}|G)p(G)/P(d_{1:N}) \propto p(d_{1:N}|G)p(G)$$

  where $p(d_{1:N}|G)$ is the marginal likelihood of $d_{1:N}$ given G, $p(G)$ is a prior probability distribution, and $p(d_{1:N})$ is a normalization constant.
- The Bayesian score is **consistent**. That is, given two DAGs G and H:
  - If $I(G) \subseteq I(p)$ but $I(H) \nsubseteq I(p)$, then $p(G|d_{1:N}) > p(H|d_{1:N})$ asymptotically.
  - If $I(G) \subseteq I(p)$ and $I(H) \subseteq I(p)$ but G implies fewer parameters than H, then $p(G|d_{1:N}) > p(H|d_{1:N})$ asymptotically.
- This justifies choosing the DAG with maximum Bayesian score to be used for probabilistic reasoning: **Asymptotically**, this results in the simplest DAG G such that $p(X) = \prod_i p(X_i|pa_G(X_i))$.

Structure Learning for BNs: Score Based Approach

- Moreover
$$p(d_{1:N}|G) = \int p(d_{1:N}|\theta_G, G)p(\theta_G|G)d\theta_G$$
where $p(d_{1:N}|\theta_G, G)$ is the likelihood function of $d_{1:N}$ given G and θ_G, and $p(\theta_G|G)$ is a prior probability distribution.
- **Assuming** that $p(\theta_G|G) = \prod_i \prod_j p(\theta_{X_i|pa_G(X_i)=j}|G)$ and $p(\theta_{X_i|pa_G(X_i)=j}|G) \sim Dirichlet(\alpha_{ij1}, \ldots, \alpha_{ijk_i})$, we have that
$$p(d_{1:N}|G) = \prod_i \prod_j \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_k \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$
where $\alpha_{ij} = \sum_k \alpha_{ijk}$, N_{ijk} is the number of instances in $d_{1:N}$ where $X_i = k$ and $pa_G(X_i) = j$, and $N_{ij} = \sum_k N_{ijk}$.
- The Bayesian score is **score equivalent** (i.e. it gives the same score to equivalent DAGs) if and only if
$$\alpha_{ijk} = \alpha p'(ijk)$$
where α is the user-defined imaginary sample size (the higher the more regularization) and $p'(ijk)$ is a prior probability distribution. For instance, $p'(ijk) = 1/[k_i \prod_{X_l \in pa_G(X_i)} k_l]$ results in the so-called BDeu score.

Structure Learning for BNs: Score Based Approach

- Under the Dirichlet parameter prior assumption and when $N \to \infty$, we have that

$$\log p(d_{1:N}|G) \approx \log p(d_{1:N}|\theta_G^{ML}, G) - \frac{\log M}{2} dim(G)$$

  where $dim(G)$ is the dimension or number of free parameters of G, i.e. $\sum_i (k_i - 1) \prod_{X_l \in pa_G(X_i)} k_l$.
- This approximation is called **Bayesian information criterion** (BIC), and it shows that the Bayesian score favours models that trade off fit of data and model complexity.

Structure Learning for BNs: Score Based Approach

- Number of DAGs with 1-12 nodes: 1, 3, 25, 543, 29281, 3781503, 1138779265, 783702329343, 1213442454842881, 4175098976430598143, 31603459396418917607425, 521939651343829405020504063
- Then, an exhaustive search is prohibitive. Moreover, there is no efficient alternative as the task is NP-hard. Then, a heuristic search must be performed instead.

Hill-climbing (HC)

Let G be the empty DAG
Repeat until no change occurs
  Add, remove or reverse any edge in G that improves the Bayesian score the most

---

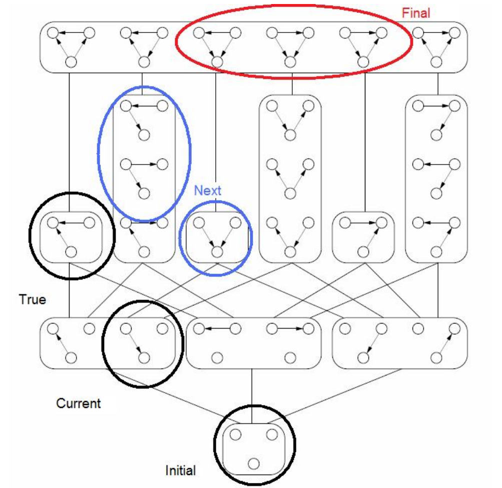
- The log Bayesian score is **decomposable** if $\log p(G)$ is so. That is

$$\log p(G|d_{1:N}) = \sum_i f(X_i, pa_G(X_i), d_{1:N})$$

- Then, adding, removing or reversing a edge in G implies recomputing only one or two factors.

Structure Learning for BNs: Score Based Approach

- Unfortunately, HC is not asymptotically correct.



Structure Learning for BNs: Score Based Approach

k-Greedy equivalence search (KES)

Let G be the empty DAG
Repeat until all the DAGs that are equivalent to G have been considered
    Repeat until no change occurs
        Add or remove any edge in G that improves the Bayesian score the most
    Replace G with any other DAG that is equivalent to G

- An arc $X_i \to X_j$ in G is **covered** when $pa_G(X_i) = pa_G(X_j) \setminus X_i$.
- There is a sequence of covered arc reversal between two equivalent DAGs. This helps in the last line of the KES algorithm.
- Alternatively, a heuristic search in the space of EGs can be performed.

Greedy equivalence search (GES)

Let G be the empty EG
Repeat until no change occurs
    Add or remove any edge in G that improves the Bayesian score the most

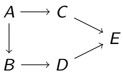
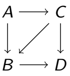
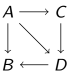
Structure Learning for BNs: Score Based Approach

- KES and GES are asymptotically correct not only under the faithfulness assumption but also under the milder **composition property** assumption:

$$U \perp_p V \mid Z \wedge U \perp_p W \mid Z \Rightarrow U \perp_p V \cup W \mid Z$$

- Not all the probability distributions satisfy the composition property, e.g. XOR. However,
  - all the probability distributions resulting from marginalization and conditioning in a faithful probability distribution satisfy the composition property, and
  - all the regular Gaussian probability distributions satisfy the composition property.
- Correctness under the composition property assumption means that the resulting DAG G is **inclusion optimal** with respect to p. That is:

$$I(G) \subseteq I(p) \text{ and there is no DAG } H \text{ such that } I(G) \subset I(H) \subseteq I(p)$$

Faithful to p	Faithful to $p(A,B,C,D \mid E)$	Inclusion optimal with respect to $p(A,B,C,D \mid E)$
	None	 and  and equivalent

Structure Learning for MNs: Independence Test Based Approach

- We can get an UG H such that $p(X) = \prod_i \varphi(C_i)/Z$ and, thus, that we can use for probabilistic reasoning as follows:

For each X_i do
  Set $ad_H(X_i)$ to be any minimal subset of $X \setminus X_i$ such that $X_i \perp_p X \setminus ad_H(X_i) | ad_H(X_i)$

- **Exercise**. Prove the previous statement.
- Assuming composition, we can get H without searching over the 2^{n-1} possible adjacent sets for each node.

Incremental associative Markov boundary algorithm (IAMB)

For each X_i do
  $ad_H(X_i) := \varnothing$
  Repeat until no change occurs
    if there exists $X_j \notin ad_H(X_i) \cup X_i$ such that $X_i \not\perp_p X_j | ad_H(X_i)$ then
      $ad_H(X_i) := ad_H(X_i) \cup X_j$
  Repeat until no change occurs
    if there exists $X_j \in ad_H(X_i)$ such that $X_i \perp_p X_j | ad_H(X_i) \setminus X_j$ then
      $ad_H(X_i) := ad_H(X_i) \setminus X_j$

Structure Learning for MNs: Independence Test Based Approach

- We prove below the inclusion optimality of the IAMB algorithm **assuming** that all the independence tests are correct.
- Firstly, we show that $X_i \perp_p X \setminus (ad_H(X_i) \cup X_i) | ad_H(X_i)$ after the first repeat.
  - Assume to the contrary that $X_i \not\perp_p X \setminus (ad_H(X_i) \cup X_i) | ad_H(X_i)$. Then, there exists $X_j \notin ad_H(X_i) \cup X_i$ such that $X_i \not\perp_p X_j | ad_H(X_i)$ by the composition property. This contradicts the fact that the first repeat ended.
- Secondly, we show that $X_i \perp_p X \setminus ((ad_H(X_i) \setminus X_j) \cup X_i) | ad_H(X_i) \setminus X_j$ after each node removal in the second repeat.
  - To see it, recall from above that $X_i \perp_p X \setminus (ad_H(X_i) \cup X_i) | ad_H(X_i)$ which together with $X_i \perp_p X_j | ad_H(X_i) \setminus X_j$ imply the desired result by the contraction property.
- Finally, we show that $ad_H(X_i)$ is minimal with respect to $X_i \perp_p X \setminus (ad_H(X_i) \cup X_i) | ad_H(X_i)$ after the second repeat.
  - Assume to the contrary that there exists $M \subset ad_H(X_i)$ such that $X_i \perp_p X \setminus (M \cup X_i) | M$. Then, $X_i \perp_p X \setminus ((ad_H(X_i) \setminus X_j) \cup X_i) | ad_H(X_i) \setminus X_j$ with $X_j \in ad_H(X_i) \setminus M$ by the weak union property, and $X_i \perp_p X_j | ad_H(X_i) \setminus X_j$ by the decomposition property. This contradicts the fact that the second repeat ended.

Structure Learning for MNs: Score Based Approach

- **Exercise**. Sketch how to perform structure learning for MNs. Consider issues such as score decomposability, existence of closed form expressions, and problems due to equivalent MNs.
- **Exercise**. Sketch how to perform structure learning for BNs and MNs from an incomplete sample. Consider issues such as score decomposability, existence of closed form expressions, and problems due to equivalent BNs and MNs.