# Lecture 7: January 17

*Lecturer: Donlapark Pornnopparath*

here are two major approaches in finding a Bayesian network from the data:

- Constraint-based approach: we start with the complete graph, then a number of hypothesis tests are made to find independence conditions, which are used as constraints on the edges of the graph.
- Score-based approach: we first define a score that evaluates how well a Bayesian network fits the data, and then search over a space of all possible graphs to find the one with maximal score.

We will start with the constraint-based approach. This approach does not require us to find all possible conditional independence relations. Instead, the corresponding algorithm contains a set of queries that ask for only some of these relations. Before anything else, we start with some assumptions:

- The network $G$ is sparse in a sense that for all node $X_i$, $|\text{Pa}(X_i)| \leq d$ for some $d$.
- The underlying distribution $P$ is faithful to $G$ i.e. any independence in $P$ can be reflected by a d-separation implied from $G$ .

## 7.1 Hypothesis test

To approximate an independence relation $X \perp\!\!\!\perp Y|\boldsymbol{Z}$ from the data $D$, we need a tool to help us decide if the relation hold or not; this is where a hypothesis test comes into play. Suppose that $X$ and $Y$ are discrete, then we use the chi-squared test:

$$d_{\chi^2}(D) = \sum_{x,y,\boldsymbol{z}} \frac{(n(x,y,\boldsymbol{z}) - n \cdot \widehat{P}(\boldsymbol{z})\widehat{P}(x|\boldsymbol{z})\widehat{P}(y|\boldsymbol{z}))^2}{n \cdot \widehat{P}(\boldsymbol{z})\widehat{P}(x|\boldsymbol{z})\widehat{P}(y|\boldsymbol{z})},$$

where $\widehat{P}$ is the *empirical distribution* e.g. $\widehat{P}(\boldsymbol{z}) = n(\boldsymbol{z})/n$ and $\widehat{P}(X|\boldsymbol{z}) = n(X,\boldsymbol{z})/n(\boldsymbol{z})$. This is a test on datasets of size $n$ , so the $p$–value of a $\chi^2$ statistic is given by

$$p(t) = P(\{D : d_{\chi^2}(D) > t\}|H_0, n)$$

Since it is infeasible to compute this, we instead approximate the distribution of $d_{\chi_2}(D)$ by a chi-squared distribution with $(|\Omega_X| - 1)(|\Omega_Y| - 1)\prod_{Z\in\boldsymbol{Z}} |\Omega_Z|$, where $\Omega_X$ is the sample space of $X$ and so on.

## 7.2 The PC algorithm

Let $I(D)$ be the set of independence conditions from the hypothesis test on data $D$. Under some assumptions on the desired graph $G$ and the underlying probability distribution $P$, we can build $G$ using the PC (Peter & Clark) algorithm.

### 7.2.1 Finding the graph skeleton

This step is based on the following results about d-separation in Bayesian network:

**Theorem 7.1.** Let $G$ be a DAG. Two nodes $X$ and $Y$ in $G$ are not adjacent if and only if there exists a set of nodes that d-separates $X$ and $Y$.

**Proof:** $\leftarrow$ If $X$ and $Y$ were adjacent, then they could not be d-separated.
$\rightarrow$ Without loss of generality, assume that there is no directed path from $Y$ to $X$. We will show that $\text{Pa}(Y)$ d-separates $X$ and $Y$. If there is no chain from $X$ to $Y$ then we are done.

- If $X - \ldots - Z \rightarrow Y$ is a trail in $G$ then the corresponding path is blocked by $Z$ which is a parent of $Y$.
- If $X - \ldots - Z \leftarrow Y$ then there exists a v–structure closest to $Y$, $X - \ldots \rightarrow W \leftarrow -Z \leftarrow Y$ since there is no directed path from $Y$ to $X$. Here, $W$ cannot be a parent of $Y$, otherwise there would be a cycle. In particular, conditioning on $W$ would lead to the path being blocked by $W$.

■

The proof of Theorem 7.1 leads to the following corollary:

**Corollary 7.2.** Let $G$ be a DAG. two nodes $X$ and $Y$ are d-separated by some set of nodes if and only if they are d-separated by either the parents of $X$ or the parents of $Y$.

These results allow us to construct a skeleton of the graph as follows:

1. Start with the complete graph $G^*$.
2. Fix $d > 0$ and make hypothesis test for $X \perp\!\!\!\perp Y | \mathbf{W}$ for all $\mathbf{W}$ of size $1, 2, \ldots, d$. If we accept any of the null hypotheses then the edge between $X$ and $Y$ is removed.

Note that by Corollary 7.2, we only need to consider the neighborhoods of $X$ and $Y$ in the current $G^*$.

### 7.2.2 Finding v-structures

Now we find all possible edge orientations that conform with the independence conditions.

**Theorem 7.3.** Let $G$ be a DAG. If $X$ and $Y$ are adjacent to $Z$, but they are not adjacent to each other, then $G$ has the v-structure $X \rightarrow Z \leftarrow Y$ if and only if there is a subset of nodes $\mathbf{W}$ that d-separates $X$ and $Y$ and does not contain $Z$.

**Proof:** $\rightarrow$ Suppose that $X \rightarrow Z \leftarrow Y$, then $Z$ is not a parent of $X$ and $Y$. Since $X$ and $Y$ are not adjacent, Theorem 7.1 implies that $X$ and $Y$ are d-separated by some set of nodes $\mathbf{W}$ which does not contain $Z$.
$\leftarrow$ Suppose that $X - Z - Y$ is of any other orientation i.e $X \leftarrow Z \leftarrow Y$, $X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \rightarrow Y$, then $X$ can travel to $Y$ through $Z$, and so there is no such $W$ that d-separates $X$ and $Y$. ■

This allows us to find v-structures in $G$.

- For any skeleton $X - Z - Y$ where the path $X - Y$ was removed by accepting the hypothesis test $X \perp\!\!\!\perp Y | \mathbf{W}$ where $\mathbf{W}$ does not contain $Z$, we orient $X \rightarrow Z \leftarrow Y$.

### 7.2.3 Orienting the undirected edges

In the last step, we orient any undirected edge that is a part of the following subgraphs:

1.
2.
3.

There are some upsides and downsides of the constraint-based approach.

#### Upsides

- It is generally faster than the score-based approach.
- Allow us to inject prior expert knowledge into the model.

#### Downsides

- The independence test is not reliable on small datasets.
- One false positive or negative can ruin the structure of the graph.