VIII - Bayesian networks

Bayesian networks belong to the family of probabilistic graphical models.

In a probabilistic graphical model, each node corresponds to a random variable and an edge between two nodes represents the probabilistic dependence between the two corresponding random variables.

Bayesian networks correspond to a type of probabilistic graphical models known as directed acyclic graphs (DAGs), that can be used to provide a compact representation of the conditional independence assumptions for (X, θ).

DAGs are particularly useful

(i) To model complex systems;

(ii) To reduce the complexity of $f(x, \theta) := f(x|\theta)\pi(\theta)$ (and hence to facilitate the approximation of $\pi(\theta|x) \propto f(x, \theta)$) without modifying the conditional independence assumptions for (X, θ).

The main difference between these two points is that in (ii) the statistical model $\{f(\cdot|\theta);\ \theta \in \Theta\}$ is taken as given while in (i) this latter is derived from the DAG.

Informally speaking, the rational behind (ii) is that adding a random variable Ψ may simplify the network (so that $f(x, \theta, \psi)$ is "simpler" than $f(x, \theta)$) without modifying the conditional independence assumptions for (X, θ). This point is explored in more details at the end of the chapter while (i) is discussed in Chapter 9.

Notation and convention

For a set $A \subset \{1, \ldots, m\}$ let $y_A = (y_i,\, i \in A)$ and, for integers $1 \leq i \leq j \leq m$, let $i : j = \{i, \ldots, j\}$.

To simplify the presentation all the random variables that appear below are assumed to be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$ with strictly positive probability density functions.

For a probability density function $p : \mathbb{R}^m \to (0, +\infty)$ and non-empty disjoint sets $A, B \subset 1 : m$, let

$$p_{Y_A}(y_A) = \int_{\mathbb{R}^{m-|A|}} p(y_{1:d})\mathrm{d}y_{1:d\setminus A}, \quad p_{Y_A|Y_B}(y_A|y_B) = \frac{p_{Y_{A\cup B}}(y_{A\cup B})}{p_{Y_B}(y_B)}.$$

Remark: $p_{Y_A}(y_A|y_B)$ is well-defined as B is non-empty and $p(y) > 0$ for all $y \in \mathbb{R}^m$.

To simplify the notation we will sometimes write $p(y_A)$ instead of $p_{Y_A}(y_A)$, $p(y_A|y_B)$ instead of $p_{Y_A|Y_B}(y_A|y_B)$, etc.

The following simple result will play a key role in what follows.

Theorem 8.1 (Telescopic theorem) *Let $p : \mathbb{R}^m \to (0, +\infty)$ be a p.d.f. on $\mathbb{R}^m$. Then,*

$$p(y) = p(y_1) \prod_{i=2}^{m} p(y_i|y_{1:(i-1)}), \quad \forall (y_1, \ldots, y_m) \in \mathbb{R}^m.$$

Proof: Left as an exercise.

Directed acyclic graphs (DAGs)

For $i \in 1:m$ let pa_i be the smallest subset of $0:(i-1)$ for which

$$p(y) = \prod_{i=1}^{m} p(y_i|y_{\text{pa}_i}) \tag{1}$$

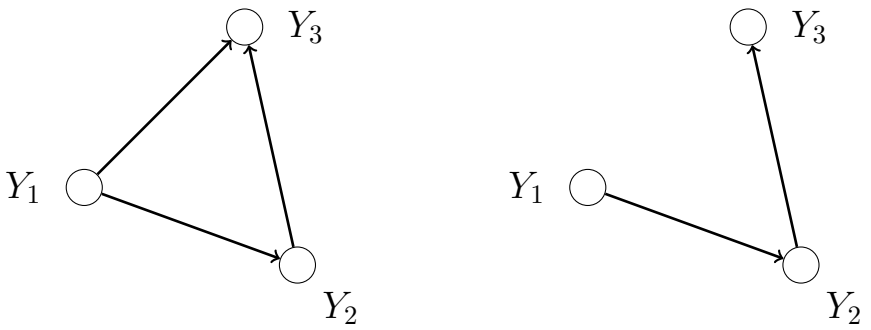
with the convention $p(y_i|y_{\text{pa}_i}) = p(y_i)$ when $\text{pa}_i = \{0\}$.

Definition 8.1 pa_i *is the set of parents of* Y_i.

A DAG is a graph where each vertex (or node) represents a coordinate of Y and there is an edge from Y_i to Y_j if and only if $i \in \text{pa}_j$.

Remark: A DAG is therefore fully characterized by the pair $(Y, \{\text{pa}_i\}_{i=1}^m)$.

Example 8.1 *Here are two examples of DAGs (with* $m = 3$*):*



- *In the left-hand case* $\text{pa}_i = 1:(i-1)$ *for* $i = 2, 3$ *so that* Y *has p.d.f.*

$$p(y) = p(y_1)p(y_2|y_1)p(y_3|y_1, y_2).$$

- *In the right-hand case,* $\text{pa}_2 = \{1\}$ *and* $\text{pa}_3 = \{2\}$ *so that* Y *has p.d.f.*

$$p(y) = p(y_1)p(y_2|y_1)p(y_3|y_2).$$

DAGs and conditional independence

Definition 8.2 *Let $A, B, C \subset \{1, \ldots, m\}$ be non-empty disjoint sets. Then, Y_A is conditionally independent of Y_B given Y_C if*

$$p(y_A, y_B|y_C) = p(y_A|y_C)p(y_B|y_C), \quad \forall y \in \mathbb{R}^m$$

and we write $Y_A \perp\!\!\!\perp Y_B \,|\, Y_C$.

The practical implications of conditional independence are captured in the following result.

Theorem 8.2 *The following statements are equivalent:*

1. $Y_A \perp\!\!\!\perp Y_B \,|\, Y_C$
2. $p(y_A|y_B, y_C) = p(y_A|y_C)$ *for all* $y \in \mathbb{R}^m$.

Proof: Done in class.

For $i \geq 1$ Let $\overline{\text{pa}}_i = \{1, \ldots, i-1\} \setminus \text{pa}_i$. Then, Theorem 8.2 implies that (assuming that $\overline{\text{pa}}_i$ is not empty)

$$Y_i \perp\!\!\!\perp Y_{\overline{\text{pa}}_i} \,|\, Y_{\text{pa}_i} \tag{2}$$

and therefore, when building a DAG $(Y, \{\text{pa}_i\}_{i=1}^m)$ for Y, the set $\{\text{pa}_i\}_{i=1}^m$ represents the conditional independence assumptions for Y.

Formal description of a Bayesian network[a]

A Bayesian network B is an annotated directed acyclic graph that represents a joint probability distribution over a set of random variables Y.

The network is defined by a pair $B = (G, P)$, where $G = (Y, \{\mathrm{pa}_i\}_{i=1}^m)$ is the DAG whose nodes $Y_1, \ldots, Y_m$ represent random variables and whose edges (fully characterized by the set $\{\mathrm{pa}_i\}_{i=1}^m$) represent the direct dependencies between these variables.

The graph G encodes independence assumptions, by which each variable Y_i is independent of its nondescendents given its parents in G; that is, $Y_i \perp\!\!\!\perp Y_{\overline{\mathrm{pa}}_i} \,|\, Y_{\mathrm{pa}_i}$.

The second component P denotes the set of parameters of the network. This set contains the "parameter" $p_{Y_i|Y_{\mathrm{pa}_i}}$ for all $i = 1, \ldots m$; that is, $P = \{p_{Y_i|Y_{\mathrm{pa}_i}}\}_{i=1}^m$.

Accordingly, by (1), B defines a unique joint probability distribution for Y, namely:

$$p(y) = \prod_{i=1}^{m} p_{Y_i|Y_{\mathrm{pa}_i}}(y_i|y_{\mathrm{pa}_i}).$$

[a]This definition of Bayesian networks is taken from Ben-Gal I., "Bayesian Networks", in Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007).

DAGs and full conditional distributions

It should be clear that a DAG is constructed with respect to a specific ordering of the components of Y. Hence, for arbitrary $i \in \{1, \ldots, m\}$ and non-empty disjoint sets $B, C \subset \{1, \ldots, m\} \setminus \{i\}$, a DAG cannot be directly used to figure out if Y_i is conditionally independent of Y_B given Y_C.

To address this kind of questions we need to know $p(y_i|y_{-i})$, the full conditional distribution of Y_i (see Chapter 7).

The next result provides an expression of $p(y_i|y_{-i})$ that depends only on the quantities that need to be specified in a Bayesian network.

Proposition 8.1 *We have*

$$p(y_i|y_{-i}) \propto p_{Y_i|Y_{\mathrm{pa}_i}}(y_i|y_{\mathrm{pa}_i}) \prod_{j \in \{k:\, i \in \mathrm{pa}_k\}} p_{Y_j|Y_{\mathrm{pa}_j}}(y_j|y_{\mathrm{pa}_j}), \quad i = 1, \ldots, m.$$

Proof: This is a direct consequence of (1).

Remark: This result shows that knowing $\{\mathrm{pa}_i\}_{i=1}^m$ allows to find the smallest set $C_i \subset \{1, \ldots, m\}$ such that

$$p(y_i|y_{-i}) \propto \prod_{j \in C_i} p_{Y_j|Y_{\mathrm{pa}_j}}(y_j|y_{\mathrm{pa}_j}), \quad \forall y \in \mathbb{R}^m. \tag{3}$$

Some comments about JAGS

In JAGS the joint distribution we want to sample from is actually specified through a Bayesian network $B = (G, P)$. Proposition 8.1 allows JAGS to efficiently compute $p(y_i|y_{-i})$ as in (3) and, as the elements of P are taken from a given list of probability distributions, to recognize whenever it is possible to sample from $p(y_i|y_{-i})$.

Bayesian networks and prior specification

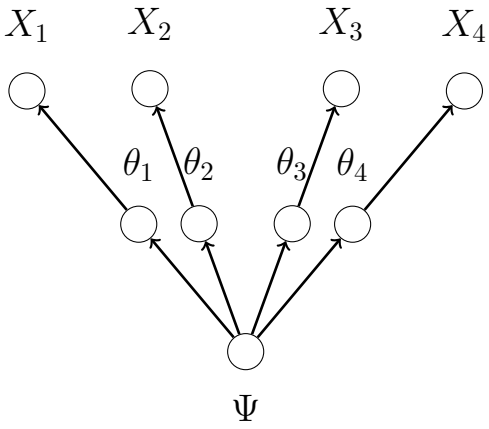
Bayesian networks may be useful to simplify the expression of $\pi(\theta)$ and hence of the joint distribution $f(x, \theta)$.

The idea is to start with the DAG of $Y = (X, \theta)$ and then to add a random variable Ψ to the DAG in order to simplify its structure.

Example 8.2 *Let $k = d = 4$, $m = 2d$ and suppose that*

$$f(x|\theta) = \prod_{i=1}^{d} f(x_i|\theta_i), \quad \pi(\theta) = \prod_{i=1}^{d} \pi_{\theta_i|\theta_{1:(i-1)}}(\theta_i|\theta_{1:(i-1)}) \tag{4}$$

so that there is no conditional independence assumptions on θ. Let Ψ be an additional random variable and consider the following DAG for (X, θ, Ψ)



Then, in the above DAG, the joint distribution of (θ, Ψ) is

$$\tilde{\pi}(\theta, \psi) = \tilde{\pi}_{\Psi}(\psi) \prod_{i=1}^{d} \tilde{\pi}_{\theta_i|\Psi}(\theta_i|\psi). \tag{5}$$

The expression of $\tilde{\pi}(\theta, \psi)$ is simpler than the expression of $\pi(\theta)$ since in (4) $\pi_{\theta_i|\theta_{1:(i-1)}}(\theta_i|\theta_{1:(i-1)})$ is a function of i variables while in (5) $\tilde{\pi}_{\theta_i|\Psi}(\theta_i|\psi)$ depends only on two variables. To simplify further the modelling process it is common to take $\tilde{\pi}_{\theta_i|\Psi} = \tilde{\pi}_{\theta_1|\Psi}$ for all i.

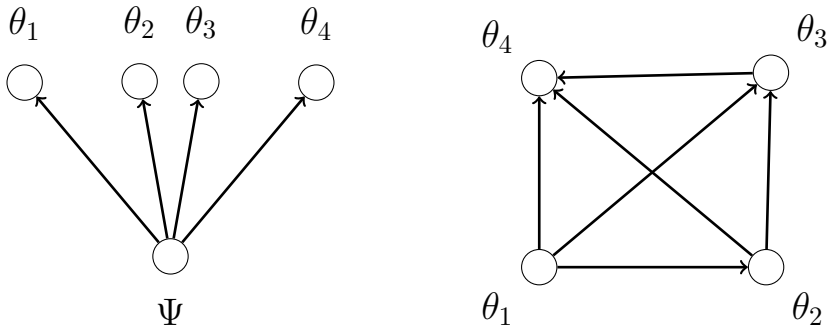
Mutually conditionally independent random variables

Example 8.2 (end) *For this to be useful it remains to show that the joint distribution* (5) *does not modify the conditional independence modelling assumption on* θ. *This is indeed the case since*

$$\tilde{\pi}(\theta_i|\theta_{1:(i-1)}) = \frac{\int \prod_{j=1}^{i} \tilde{\pi}_{\theta_1|\Psi}(\theta_j|\psi)\tilde{\pi}_{\Psi}(\psi)\mathrm{d}\psi}{\int \prod_{j=1}^{i-1} \tilde{\pi}_{\theta_1|\psi}(\theta_j|\psi)\tilde{\pi}_{\Psi}(\psi)\mathrm{d}\psi\mathrm{d}\theta_i} \tag{6}$$

depends on $\theta_{1:i}$.

Here is what happens to the DAG of (θ, Ψ) *when we integrate* Ψ *out:*



The joint density (5) corresponds to a special form of conditional independence defined in the next definition.

Definition 8.3 *The random variable* Y *is mutually conditionally independent given* Ψ *if* $Y_A \perp\!\!\!\perp Y_B \,|\, \Psi$ *for any disjoint sets* $A, B \subset \{1, \ldots, m\}$. *In this case we write* $\models Y \,|\, \Psi$.

This form of conditional independence is useful in modelling since this is the one that brings the greatest simplification in the expression of $f(x, \theta)$.

DAG of (Y, Ψ) versus DAG of Y

Example 8.2 shows that, for $m = 4$, adding a random variable Ψ such that $\models \theta \,|\, \Psi$ allows to remove two edges in the DAG.

The following proposition generalizes this observation.

Proposition 8.2 *Let* $\mathrm{qa}_i \subset 0 : (i-1)$ *be the set of parents of* Y_i *in the DAG of* (Y, Ψ) *(with the convention* $0 \in \mathrm{qa}_i$ *means that* Ψ *is a parent of* Y_i*) and* $\mathrm{pa}_i \subset 0 : (i-1)$ *be the set of parents of* Y_i *in the DAG of* Y*. Assume that for all* $i = 1, \ldots m$*, we have* $\mathrm{qa}_i = \{0\}$ *(so that* $\models Y \,|\, \Psi$*). Then,* $\mathrm{pa}_i = \{1, \ldots i-1\}$ *and therefore the DAG of* Y *has* $m(m-1)/2$ *edges while the DAG of* (Y, Ψ) *has* m *edges.*

Proposition 8.2 is a direct consequence of the following theorem that gives the general result for what happens to the DAG of Y when we integrate out Ψ in the DAG of (Y, Ψ).

Theorem 8.3 *Let* $\{\mathrm{qa}_i\}_{i=1}^m$ *and* $\{\mathrm{pa}_i\}_{i=1}^m$ *be as in Proposition 8.2. Then, for* $i = 2, \ldots, m$*, we have*

$$\mathrm{pa}_i = \begin{cases} A_{i-1} \cup \left\{ \cup_{j \in A_i} \mathrm{qa}_j \right\} \setminus \{0\}, & 0 \in \mathrm{qa}_i \\ \mathrm{qa}_i, & 0 \notin \mathrm{qa}_i \end{cases}$$

where $A_i = \{j : 1 \leq j \leq i \text{ and } 0 \in \mathrm{qa}_j\}$.

Proof: Done in class.

In words: If Ψ is a parent of Y_i and Ψ is marginalized out then Y_i gets an edge from each Y_j $(i < j)$ for which $0 \in \mathrm{qa}_j$ and also an edge from all the Y_k's that are parents of these Y_j's in the DAG of (Y, Ψ).

Exchangeablility and de Finetti's theorem

In Example 8.2 we describe a way to simplify the expression of the prior distribution $\pi(\theta)$ without imposing any conditional independence assumptions on θ.

However, it is clear that probability density functions of the form

$$\pi(\theta) = \int \prod_{i=1}^{d} \pi(\theta_i|\psi)\pi_\Psi(\psi)\mathrm{d}\psi \tag{7}$$

form only a subset of all the probability density functions on Θ that makes no conditional independence assumptions. Therefore, prior distributions of this form impose conditions on $(\theta_1, \ldots, \theta_d)$ that need to be understood.

As shown by de Finetti's theorem (Theorem 8.4), this modelling strategy implicitly assumes that the sequence $(\theta_1, \ldots, \theta_d)$ is exchangeable.

Definition 8.4 *A sequence of random variables $Y_1, \ldots, Y_m$ is said to be exchangeable if for any permutation σ of $\{1, \ldots, m\}$ the joint distribution of $(Y_{\sigma(1)}, \ldots, Y_{\sigma(m)})$ is the same as the joint distribution of $(Y_1, \ldots, Y_m)$.*

Remark: If $Y_1, \ldots, Y_m$ are i.i.d. then $(Y_1, \ldots, Y_m)$ is exchangeable.

Theorem 8.4 (Hewitt and Savage, 1955) *A sequence of $\mathcal{Y}$-valued random variables $Y_1, \ldots, Y_m$ is exchangeable if and only if there exists a unique measure Π on $\mathcal{P}(\mathcal{Y})$ such that, for any measurable sets $B_i \subset \mathcal{Y}$, $i = 1, \ldots, m$, we have*

$$\mathbb{P}\big(Y_1 \in B_1, \ldots, Y_m \in B_m\big) = \int_{\mathcal{P}(\mathcal{Y})} \prod_{i=1}^{m} P(B_i)\Pi(\mathrm{d}P). \tag{8}$$

A simpler version of de Finetti's theorem

Theorem 8.4 is the general version of de Finetti's theorem. However, it is not the easiest one to understand as it involves an integral over a space of probability measures.

Although much more restrictive, the next result has the advantage to be easier to understand.

Theorem 8.5 (De Finetti, 1931) *A sequence of $\{0,1\}$-valued random variables $Y_1, \ldots, Y_m$ is exchangeable if and only if there exists a probability measure $\pi_\Psi \in \mathcal{P}([0,1])$ such that, for any $y \in \{0,1\}^m$, we have*

$$\mathbb{P}\big(Y_1 = y_1, \ldots, Y_m = y_m\big) = \int_0^1 \prod_{i=1}^m \big(\psi^{y_i}(1-\psi)^{1-y_i}\big)\pi_\Psi(\mathrm{d}\psi).$$

In words, if $Y = (Y_1, \ldots, Y_m)$ is as in Theorem 8.5, then there exists a random variable Ψ on $[0,1]$ such that $\models Y \,|\, \Psi$. Moreover, $Y_i|\Psi \sim \text{Bernoulli}(\Psi)$.

Consequently, prior distributions of the form (7) assume that $(\theta_1, \ldots, \theta_d)$ is exchangeable.