IX - Hierarchical models

We start with a definition.

Definition 9.1 *A hierarchical Bayes model is a Bayesian model where the prior distribution $\pi(\theta)$ is decomposed in conditional distributions*

$$\pi_{\theta|\psi_1}(\theta|\psi_1), \ldots, \pi_{\psi_{L-1}|\psi_L}(\psi_{L-1}|\psi_L)$$

and a marginal distribution $\pi_{\psi_L}(\psi_L)$ such that

$$\pi(\theta) = \int \pi_{\theta|\psi_1}(\theta|\psi_1)\Big(\prod_{l=1}^{L-1} \pi_{\psi_l|\psi_{l+1}}(\psi_l|\psi_{l+1})\Big)\pi_{\psi_L}(\psi_L)\mathrm{d}(\psi_1, \ldots, \psi_L).$$

The parameter ψ_l is called hyperparameter of level l.

As we will see in this short chapter, hierarchical models are particularly relevant to model observations that are "similar" but not identical, because they allow to easily control the interrelationship between the observations according to known group.

Having similar but not identical observations typically arises when we are interested in making inference on parameters corresponding to different sub-populations (or "units") of a global population.

Remark: As mentioned in Chapter 3, hierarchical models can also be used to reduce the impact of the prior distribution on the inference.

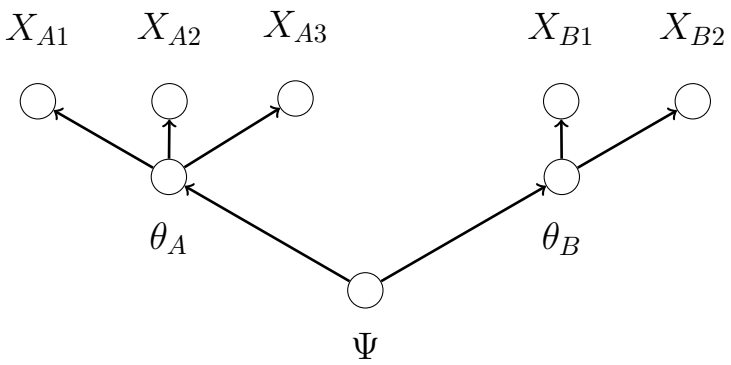
Modelling similar observations: An example

Consider a school with classes A and B. Class $j \in \{A, B\}$ contains n_j students, with $n_A = 3$ and $n_B = 2$. Let X_{ji} be the exam score of student $i \in \{1, \ldots, n_j\}$ in class $j \in \{A, B\}$. All the students are assumed to be "similar" (i.e. same age, similar socio-economic characteristics, etc.).

Assume that we have observed (X_{A1}, X_{A2}, X_{B1}) and that we want to predict X_{B2}. Then, it seems sensible to take into account that

- The distribution of the exam scores in the two classes should be "similar" since the two classes are in the same school. Hence, (X_{A1}, X_{A2}) is relevant to predict X_{B2}.
- However, it is unlikely that the distribution of the exam scores in the two classes is identical since e.g. the teacher is not the same in the two classes. Hence, the prediction of X_{B2} should be more influenced by X_{B1} than (X_{A1}, X_{A2}) .

A possible DAG in this context is therefore:



Remark: In this DAG, the distribution of X_{A1} and of X_{B1} are indeed not identical, since the former depends on θ_A while the latter depends on $\theta_B \neq \theta_A$, but are similar in the sense that they share the same hyperparameter Ψ.

Modelling similar observations: An example (continued)

Let $\theta = (\theta_A, \theta_B)$ and $X = (X_A, X_B)$ where, for $j = A, B$, $X_j = (X_{ji}, \ldots, X_{jn_j})$.

Then, from Chapter 8, we know that the above DAG implies that $\models \theta|\Psi$ while $\models X_j|\theta_j$ for $j = A, B$.

Therefore, the above DAG implies the following decomposition for the joint distribution of (X, θ, Ψ):

$$f(x, \theta, \psi) = \pi_\Psi(\psi) \prod_{j \in \{A,B\}} \pi_{\theta_j|\Psi}(\theta_j|\psi) \prod_{i=1}^{n_j} f_{X_{ji}|\theta_j}(x_{ji}|\theta_j).$$

Typically, we choose $\pi_{\theta_B|\Psi} = \pi_{\theta_A|\Psi}$ and $f_{X_{ji}|\theta_j} = f_{X_{A1}|\theta_A}$ for all i, j so that, to complete the Bayesian network (i.e. the Bayesian model) we just need to specify

$$\pi_\Psi, \quad \pi_{\theta_A|\Psi}, \quad f_{X_{A1}|\theta_A}.$$

Remark: With these three distributions we can handle an arbitrary large number of groups (i.e. classes) and an arbitrary large number of cases per group (i.e. students per class).

Remark: Assuming $\pi_{\theta_B|\Psi} = \pi_{\theta_A|\Psi}$ amounts to assuming that (θ_A, θ_B) is exchangeable (see Theorem 8.4). Therefore, hierarchical models are particularly useful when the units' specific parameters are exchangeable, since in this case the model is built from only a small number of probability density functions.

Remark: In the notation of Definition 9.1, $L = 1$ and $\psi_1 = \psi$.

Modelling similar observations: An example (end)

Remark that if the prior distribution for θ is $\pi(\theta) = \pi_A(\theta_A)\pi_B(\theta_B)$ then the resulting posterior distribution is given by

$$\begin{aligned}\pi(\theta|x) &\propto \left(\pi_A(\theta_A)\prod_{i=1}^{n_A} f_{X_{Ai}|\theta_A}(x_{Ai}|\theta_A)\right)\left(\pi(\theta_B)\prod_{i=1}^{n_B} f_{X_{Bi}|\theta_B}(x_{Bi}|\theta_B)\right)\\ &\propto \pi_A(\theta_A|x_A)\pi_B(\theta_B|x_B)\end{aligned}$$

where

$$\pi_j(\theta_j|x_j) \propto \pi_j(\theta_j)\prod_{i=1}^{n_j} f_{X_{ji}|\theta_j}(x_{ji}|\theta_j), \quad j \in \{A, B\}.$$

Therefore, when choosing $\pi(\theta) = \pi_A(\theta_A)\pi_B(\theta_B)$ the parameters θ_A and θ_B are estimated separately, the former using only x_A and the latter only x_B. Consequently, the resulting posterior distribution for θ_j will be more dispersed (i.e. the estimate of θ_j will be less "precise") than the one obtained with the hierarchical approach used above.

Remark: Using a hierarchical approach however makes sense only if we are ready to assume that X_A and X_B are similar.

Some final comments about hierarchical models

In practice it is quite frequent to encounter situations where it makes sense to assume that $f(x|\theta) = \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)$ so that we have as many parameters as observations.

In this scenario,

1. Considering a prior distribution of the form $\pi(\theta) = \prod_{i=1}^{d} \pi_i(\theta_i)$ would result in a posterior distribution for θ_i that depends on x_i only (that is, our inference on θ_i would be based on a single observation).
2. Considering a model of the form

$$f(x, \theta, \psi) = \pi_\Psi(\psi) \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)\pi_{\theta_i|\Psi}(\theta_i|\psi)$$

would result in a posterior distribution for θ_i that depends on all the observations $(x_1, \ldots, x_d)$ and therefore that contains more information about θ_i than in case 1.

An alternative modelling strategy in this context is to specify a model $\{f'(\cdot|, \lambda),\ \lambda \in \Lambda\}$ where λ is a low dimensional parameter (i.e. the dimension of λ is smaller than that of x).

- In a frequentist setting, specifying such a model is far from trivial as this requires to come up with a p.d.f. $f'(x|\lambda)$ that captures the dependence among the different components of x.
- In a Bayesian setting, the hierarchical approach allows to easily specify $f'(x|\lambda)$ as

$$f'(x|\lambda) = \int \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)\pi_{\theta_i|\Psi}(\theta_i|\lambda)\mathrm{d}(\theta_1, \ldots, \theta_d).$$

Remark: This approach is not allowed in a pure frequentist setting as it requires to treat the parameters as random variables.