Bayesian Hierarchical Models

In Bayesian inference, model parameters, $\boldsymbol{\theta}$, are considered as random variables, and their probability distribution is sought. In hierarchical models, several "layers" of model parameters, $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_h,\ldots,\boldsymbol{\theta}_H$, are present. Importantly, there exists dependence between these parameter groups in the sense that the probability distribution for $\boldsymbol{\theta}_h$ depends on the outcomes of $\boldsymbol{\theta}_{h+1}$. As a result of this dependence structure, the joint distribution for all intervening parameters, including the measurable random variables, X, is written:

$$f(\mathbf{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_H)=f(\mathbf{x}|\boldsymbol{\theta}_1)\cdot f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)\cdot f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3)\cdots f(\boldsymbol{\theta}_{H-1}|\boldsymbol{\theta}_H)\cdot f(\boldsymbol{\theta}_H) \tag{1}$$

where the parameters $\boldsymbol{\theta}_h$, for h>1, are called hyperparameters. It is understood that conditional probabilities in terms of hyperparameters, specifically $f(\boldsymbol{\theta}_h|\boldsymbol{\theta}_{h+1})$, form an important ingredient in this modelling approach. In fact, Bayesian hierarchical models is the continuous version of Bayesian networks, which are usually formulated in term of discrete random variables. In this document, the ultimate objective is the same as in ordinary Bayesian inference: to determine the probability distribution of the model parameters given observations of the measurable random variable(s), $\mathbf{X}$. That is, the probability distributions $f(\boldsymbol{\theta}_h|\mathbf{x})$ are ultimately sought. To understand how such results are obtained, consider the derivation of ordinary Bayesian updating, which starts with the conditional rule of probability applied to random variables:

$$f(\theta|\mathbf{x})=\frac{f(\mathbf{x},\theta)}{f(\mathbf{x})} \tag{2}$$

By substitution of the multiplication rule of probability Eq. (2) turns into:

$$f(\theta|\mathbf{x})=\frac{f(\mathbf{x}|\theta)}{f(\mathbf{x})}\cdot f(\theta) \tag{3}$$

Furthermore, Eq. (3) is usually reformulated in terms of the likelihood function, to avoid having to interpret the numerator in Eq. (3) as a probability, but rather as proportional to the probability of observing the observations, which leads to the traditional Bayesian form of Eq. (3):

$$f(\theta|\mathbf{x})=\frac{L(\mathbf{x}|\theta)}{c}\cdot f(\theta) \tag{4}$$

It is recalled that the denominator in Eqs. (2), (3), (4) requires integration over the model parameters, which will be reiterated shortly. To understand the hierarchical Bayes approach it is now useful to return to Eq. (2), which in the presence of the model parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, … $\boldsymbol{\theta}_h$, … $\boldsymbol{\theta}_H$ turns into

$$f(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_H|\mathbf{x})=\frac{f(\mathbf{x},\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_H)}{f(\mathbf{x})} \tag{5}$$

Again it is noted that the denominator is obtained by a multi-fold integral, which in this case reads

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_H) d\boldsymbol{\theta}_1, \cdots, d\boldsymbol{\theta}_H \tag{6}$$

While Eq. (5) is analogous to Eq. (2), it contains more information than the sought results, namely the distribution for only $\boldsymbol{\theta}_h$, i.e., $f(\boldsymbol{\theta}_h|\mathbf{x})$. This result is obtained by "integrating out" the other $\boldsymbol{\theta}$-variables from Eq. (5), which together with Eq. (6) then turns into:

$$f(\boldsymbol{\theta}_h|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta}_h)}{f(\mathbf{x})} = \frac{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_H) d\boldsymbol{\theta}_1, \cdots, d\boldsymbol{\theta}_{h-1}, d\boldsymbol{\theta}_{h+1}, \cdots, d\boldsymbol{\theta}_H}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_H) d\boldsymbol{\theta}_1, \cdots, d\boldsymbol{\theta}_H} \tag{7}$$

It is here emphasized that the integration in the denominator in Eq. (7) is over all model parameters, while the integration in the numerator is over all model parameters except $\boldsymbol{\theta}_h$. Next, it is of interest to rewrite Eq. (7) in a way that explicitly identifies the prior distribution, similar to the transition from Eq. (2) to Eq. (3). Mirroring the use of the multiplication rule in Eq. (3), Eq. (7) is re-written as:

$$f(\boldsymbol{\theta}_h|\mathbf{x}) = \frac{L(\mathbf{x}|\boldsymbol{\theta}_h) \cdot f(\boldsymbol{\theta}_h)}{c} \tag{8}$$

where c as usual serves to normalize the distribution:

$$c = \int_{-\infty}^{\infty} L(\mathbf{x}|\boldsymbol{\theta}_h) \cdot f(\boldsymbol{\theta}_h) d\boldsymbol{\theta}_h \tag{9}$$

The likelihood function in Eq. (8) is:

$$L(\mathbf{x}|\boldsymbol{\theta}_h) = \begin{cases} f(\mathbf{x}|\boldsymbol{\theta}_1) & \text{for} \quad h = 1 \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{h-1}|\boldsymbol{\theta}_h) d\boldsymbol{\theta}_1, \cdots, d\boldsymbol{\theta}_{h-1} & \text{for} \quad h = 2, \ \ldots, \ H \end{cases} \tag{10}$$

where the conditional distribution is obtained directly from the joint distribution in Eq. (1):

$$f(\mathbf{x}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{h-1}|\boldsymbol{\theta}_h) = f(\mathbf{x}|\boldsymbol{\theta}_1) \cdot f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \cdot f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3) \cdots f(\boldsymbol{\theta}_{h-1}|\boldsymbol{\theta}_h) \tag{11}$$

The prior distribution in Eq. (8) is:

$$f(\boldsymbol{\theta}_h) = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{\theta}_h, \cdots, \boldsymbol{\theta}_H) d\boldsymbol{\theta}_{h+1}, \cdots, d\boldsymbol{\theta}_H & \text{for} \quad h = 1, \ \ldots, \ H-1 \\ f(\boldsymbol{\theta}_H) & \text{for} \quad h = H \end{cases} \tag{12}$$

where, from Eq. (1):

$$f(\boldsymbol{\theta}_h, \cdots, \boldsymbol{\theta}_H) = f(\boldsymbol{\theta}_h|\boldsymbol{\theta}_{h+1}) \cdot f(\boldsymbol{\theta}_{h+1}|\boldsymbol{\theta}_{h+2}) \cdots f(\boldsymbol{\theta}_H) \tag{13}$$

It is observed that both Eq. (7) and the rewritten version in Eq. (8) make use of the complete joint PDF in Eq. (1). However, the formulation in Eq. (8) is computationally advantageous because it splits large multi-fold integrals into smaller problems. An approach to further reduce the computational effort is to conduct the Bayesian updating in Eq. (8) conditional upon having done Bayesian updating at the previous level:

$$f(\boldsymbol{\theta}_h|\mathbf{x},\boldsymbol{\theta}_{h+1}) = \frac{L(\mathbf{x}|\boldsymbol{\theta}_h,\boldsymbol{\theta}_{h+1})\cdot f(\boldsymbol{\theta}_h|\boldsymbol{\theta}_{h+1})}{c} \qquad (14)$$

where c as usual is the integral of the numerator. Once the analysis in Eq. (14) is carried out, the sought distribution is obtained by "integrating out" $\boldsymbol{\theta}_{h+1}$:

$$f(\boldsymbol{\theta}_h|\mathbf{x}) = \int_{-\infty}^{\infty} f(\boldsymbol{\theta}_h|\mathbf{x},\boldsymbol{\theta}_{h+1})\cdot f(\boldsymbol{\theta}_{h+1}|\mathbf{x})\,d\boldsymbol{\theta}_{h+1} \qquad (15)$$

The Bayesian updating in Eq. (14) is aided by the fact that $\boldsymbol{\theta}_{h+1}$ is fixed so that the likelihood function is independent of $\boldsymbol{\theta}_{h+1}$ and thus the same as in Eq. (10), while the prior is one link in the joint PDF in Eq. (1).