Chapter 8

Some complements to Bayesian estimation

8.1 Choice of a prior law in the Bayesian estimation

One of the main difficulties in the application of Bayesian theory in practice is the choice or the attribution of the direct probabilities $f(x|\theta)$ and $\pi(\theta)$. In general, $f(x|\theta)$ is obtained via an appropriate model relating the observable quantity X to the parameters θ and is well accepted. The choice or the attribution of the prior $\pi(\theta)$ has been, and still is, the main subject of discussion and controversy between the Bayesian and orthodox statisticians.

Here, I will try to give a brief summary of different approaches and different tools that can be used to attribute a prior probability distribution. There are mainly four tools:

- use of some invariance principles
- use of maximum entropy (ME) principle
- use of conjugate and reference priors
- use of other information criteria

8.1.1 Invariance principles

Définition 1 [Group invariance] A probability model $f(x|\theta)$ is said to be invariant (or closed) under the action of a group of transformations $\mathcal{G}$ if, for every $g \in \mathcal{G}$, there exists a unique $\theta^* = \bar{g}(\theta) \in \mathcal{T}$ such that $y = g(x)$ is distributed according to $f(y|\theta^*)$.

Exemple 1 Any probability density function in the form $f(x|\theta) = f(x-\theta)$ is invariant under the *translation* group

$$\mathcal{G} : \{g_c(x) : g_c(x) = x + c, \quad c \in \mathbf{R}\} \tag{8.1}$$

This can be verified as follows

$$x \sim f(x-\theta) \longrightarrow y = x + c \sim f(y - \theta^*) \quad \text{with} \quad \theta^* = \theta + c$$

Exemple 2 Any probability density function in the form $f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta})$ is invariant under the *multiplicative* or *scale* transformation group

$$\mathcal{G} : \{g_s(x) : g_s(x) = s\,x, \quad s > 0\} \tag{8.2}$$

This can be verified as follows

$$x \sim \frac{1}{\theta} f(\frac{x}{\theta}) \longrightarrow y = s\,x \sim \frac{1}{\theta^*} f(\frac{y}{\theta^*}) \quad \text{with} \quad \theta^* = s\,\theta$$

Exemple 3 Any probability density function in the form $f(x|\theta_1, \theta_2) = \frac{1}{\theta_2} f(\frac{x-\theta_1}{\theta_2})$ is invariant under the *affine* transformation group

$$\mathcal{G} : \{g_{a,b}(x) : g_{a,b}(x) = a\,x + b, \quad a > 0, b \in \mathbb{R}\} \tag{8.3}$$

This can be verified as follows

$$x \sim \frac{1}{\theta_2} f(\frac{x-\theta_1}{\theta_2}) \longrightarrow y = a\,x + b \sim \frac{1}{\theta_2^*} f(\frac{y-\theta_1^*}{\theta_2^*}) \quad \text{with} \quad \theta_2^* = a\,\theta_2, \quad \theta_1^* = a\,\theta_1 + b.$$

Exemple 4 Any multi variable probability density function in the form $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}-\boldsymbol{\theta})$ is invariant under the translation group

$$\mathcal{G} : \{g_{\boldsymbol{c}}(\boldsymbol{x}) : g_{\boldsymbol{c}}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{c}, \quad \boldsymbol{c} \in \mathbf{R}^n\} \tag{8.4}$$

Exemple 5 Any multi variable probability density function in the form $f(\boldsymbol{x}) = f(\|\boldsymbol{x}\|)$ is invariant under the orthogonal transformation group

$$\mathcal{G} : \left\{g_{\boldsymbol{A}}(\boldsymbol{x}) : g_{\boldsymbol{A}}(\boldsymbol{x}) = \boldsymbol{A}\,\boldsymbol{x}, \quad \boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}\right\} \tag{8.5}$$

Exemple 6 Any multi variable probability density function in the form $f(\boldsymbol{x}|\theta) = \frac{1}{\theta} f(\frac{\|\boldsymbol{x}\|}{\theta})$ is invariant under the following transformation group

$$\mathcal{G} : \left\{g_{\boldsymbol{A},s}(\boldsymbol{x}) : g_{\boldsymbol{A},s}(\boldsymbol{x}) = s\,\boldsymbol{A}\,\boldsymbol{x}, \quad \boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}, \quad s > 0.\right\} \tag{8.6}$$

This can be verified as follows

$$\boldsymbol{x} \sim \frac{1}{\theta} f(\frac{\|\boldsymbol{x}\|}{\theta}) \longrightarrow \boldsymbol{y} = s\,\boldsymbol{A}\,\boldsymbol{x} \sim \frac{1}{\theta^*} f(\frac{\|\boldsymbol{y}\|}{\theta^*}) \quad \text{with} \quad \theta^* = s\,\theta.$$

From these examples we see also that any invariance transformation group $\mathcal{G}$ on $x \in \mathcal{X}$ induces a corresponding transformation group $\bar{\mathcal{G}}$ on $\theta \in \mathcal{T}$. For example for the translation invariance $\mathcal{G}$ on $x \in \mathcal{X}$ induces the following translation group on $\theta \in \mathcal{T}$

$$\bar{\mathcal{G}} : \{\bar{g}_c(\theta) : \bar{g}_c(\theta) = \theta + c, \quad c \in \mathbf{R}\} \tag{8.7}$$

and the scale invariance $\mathcal{G}$ on $x \in \mathcal{X}$ induces the following translation groupe on $\theta \in \mathcal{T}$

$$\bar{\mathcal{G}} : \{\bar{g}_s(\theta) : \bar{g}_s(\theta) = s\,\theta, \quad s > 0\} \tag{8.8}$$

We just see that for an invariant family of $f(x|\theta)$ we have a corresponding invariant family of prior laws $\pi(\theta)$. To be complete, we have also to consider the cost function to be able to define the Bayesian estimate.

Définition 2 [Invariant cost functions] Assume a probability model $f(x|\theta)$ is invariant under the action of the group of transformations $\mathcal{G}$. Then the cost function $c[\theta, \widehat{\theta}]$ is said to be invariant under the group of transformations $\tilde{\mathcal{G}}$ if, for every $g \in \mathcal{G}$ and $\widehat{\theta} \in \mathcal{T}$, there exists a unique $\widehat{\theta}^* = \tilde{g}(\widehat{\theta}) \in \mathcal{T}$ with $\tilde{g} \in \tilde{\mathcal{G}}$ such that

$$c[\theta, \widehat{\theta}] = c[\bar{g}(\theta), \widehat{\theta}^*] \qquad \text{for every } \theta \in \mathcal{T}.$$

Définition 3 [Invariant estimate] For an invariant probability model $f(x|\theta)$ under the group of transformation Gc and an invariant cost function $c[\theta, \widehat{\theta}]$ under the corresponding group of transformation $\bar{\mathcal{G}}$, an estimate $\widehat{\theta}$ is said to be invariant or *equivariant* if

$$\widehat{\theta}(g(x)) = \tilde{g}\left(\widehat{\theta}(x)\right)$$

Exemple 7 Estimation of θ from the data coming from any model of the kind $f(x|\theta) = f(x-\theta)$ with a quadratic cost function $c[\theta, \widehat{\theta}] = (\theta - \widehat{\theta})^2$ is equivariant and we have

$$\mathcal{G} = \bar{\mathcal{G}} = \tilde{\mathcal{G}} = \{g_c(x) : g_c(x) = x - c, \quad c \in \mathbb{R}\}$$

Exemple 8 Estimation of θ from the data coming from any model of the kind $f(x|\theta) = \frac{1}{\theta} f(\frac{1}{\theta})$ with the entropy cost function

$$c[\theta, \widehat{\theta}] = \frac{\theta}{\widehat{\theta}} - \ln(\frac{\theta}{\widehat{\theta}}) - 1$$

is equivariant and we have

$$\begin{aligned} \mathcal{G} &= \{g_s(x) : g_s(x) = s\,x, \quad s > 0\} \\ \bar{\mathcal{G}} = \tilde{\mathcal{G}} &= \{g_s(\theta) : g_s(\theta) = s\,\theta, \quad s > 0\} \end{aligned}$$

Proposition 1 [Invariant Bayesian estimate] Suppose that a probability model $f(x|\theta)$ is invariant under the group of transformations $\mathcal{G}$ and that there exists a probability distribution $\pi^*(\theta)$ on $\mathcal{T}$ which is invariant under the group of transformations $\bar{\mathcal{G}}$, *i.e.*,

$$\pi^*(\bar{g}(A)) = \pi^*(A)$$

for any measurable set $A \in \mathcal{T}$. Then the Bayes estimator associated with π^*, noted $\widehat{\theta}^*$ minimizes

$$\int R\left(\theta, \widehat{\theta}\right) \pi^*(\theta)\,\mathrm{d}\theta = \int R\left(\theta, \bar{g}(\widehat{\theta})\right) \pi^*(\theta)\,\mathrm{d}\theta = \int \mathrm{E}\left[c\left[\theta, \bar{g}\left(\widehat{\theta}(X)\right)\right]\right] \pi^*(\theta)\,\mathrm{d}\theta \quad \text{over } \widehat{\theta}.$$

If this Bayes estimator is unique, it satisfies

$$\widehat{\theta}^*(x) = \tilde{g}^{-1}\left(\widehat{\theta}^*(g(x))\right)$$

Therefore, a Bayes estimator associated with an invariant prior and a strictly convex invariant cost function is almost equivariant.

Actually, invariant probability distributions are rare. The following are some examples:

Exemple 9 If $\pi(\theta)$ is invariant under the translation group $\mathcal{G}_c$, it satisfies $\pi(\theta) = \pi(\theta + c)$ for every θ and for every c, which implies that $\pi(\theta) = \pi(0)$ uniformly on $\mathbf{R}$ and this leads to the Lebesgue measure as an invariant measure.

Exemple 10 If $\theta > 0$ and $\pi(\theta)$ is invariant under the scale group $\mathcal{G}_s$, it satisfies $\pi(\theta) = s\,\pi(s\theta)$ for every $\theta > 0$ and for every $s > 0$, which implies that $\pi(\theta) = 1/\theta$.

Note that in both cases the invariant laws are improper.

8.2 Conjugate priors

The conjugate prior concept is tightly related to the sufficient statistic and exponential families.

Définition 4 [Sufficient statistics] When $X \sim P_\theta(x)$, a function $h(X)$ is said to be a sufficient statistic for $\{P_\theta(x), \theta \in \mathcal{T}\}$ if the distribution of X conditioned on $h(X)$ does not depend on θ for $\theta \in \mathcal{T}$.

Définition 5 [Minimal sufficiency] A function $h(X)$ is said to be minimal sufficient for $\{P_\theta(x), \theta \in \mathcal{T}\}$ if it is a function of every other sufficient statistic for $P_\theta(x)$.

A minimal sufficient statistic contains the whole information brought by the observation $X = x$ about θ.

Proposition 2 [Factorization theorem] Suppose that $\{P_\theta(x), \theta \in \mathcal{T}\}$ has a corresponding family of densities $\{p_\theta(x), \theta \in \mathcal{T}\}$. A statistic T is sufficient for θ if and only if there exist functions g_θ and h such that

$$p_\theta(x) = g_\theta(T(x))\, h(x) \tag{8.9}$$

for all $x \in \Gamma$ and $\theta \in \mathcal{T}$.

Exemple 11 If $X \sim \mathcal{N}(\theta, 1)$ then $T(x) = x$ can be chosen as a sufficient statistic.

Exemple 12 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathcal{N}(\theta, 1)$ then

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right] \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right](2\pi)^{-n/2}\exp\left[-\frac{n}{2}\theta^2\right]\exp\left[\theta\sum_{i=1}^{n} x_i\right] \end{aligned}$$

and we have $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$.

Note that, in this case, we need to know n and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Note also that we can write

$$f(\boldsymbol{x}|\theta) = a(\boldsymbol{x})\, g(\theta) \exp\left[\theta T(\boldsymbol{x})\right]$$

where

$$g(\theta) = (2\pi)^{-n/2}\exp\left[-\frac{n}{2}\theta^2\right] \quad \text{and} \quad a(\boldsymbol{x}) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right]$$

Exemple 13 If $X \sim \mathcal{N}(0, \theta)$ then $T(x) = x^2$ can be chosen as a sufficient statistic.

Exemple 14 If $X \sim \mathcal{N}(\theta_1, \theta_2)$ then $T_1(x) = x^2$ and $T_2(x) = x$ can be chosen as a set of sufficient statistics.

Exemple 15 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathcal{N}(\theta_1, \theta_2)$ then

$$\begin{aligned} f(\boldsymbol{x}|\theta_1,\theta_2) &= (2\pi)^{-n/2}\,\theta_2^{-1/2}\,\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right] \\ &= (2\pi)^{-n/2}\,\theta_2^{-1/2}\,\exp\left[-\frac{n\theta_1^2}{2\theta_2}\right]\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}x_i^2+\frac{\theta_1}{\theta_2}\sum_{i=1}^{n}x_i\right] \end{aligned}$$

and we have $T_1(\boldsymbol{x}) = \sum_{i=1}^n x_i$ and $T_2(\boldsymbol{x}) = \sum_{i=1}^n x_i^2$.

Note also that we can write

$$f(\boldsymbol{x}|\theta) = a(\boldsymbol{x})\,g(\theta_1,\theta_2)\exp\left[\frac{\theta_1}{\theta_2}T_1(\boldsymbol{x}) - \frac{1}{2\theta_2}T_2(\boldsymbol{x})\right]$$

where

$$g(\theta_1,\theta_2) = (2\pi)^{-n/2}\theta_2^{-1/2}\exp\left[-\frac{n\theta_1^2}{2\theta_2}\right] \quad \text{and} \quad a(\boldsymbol{x}) = 1.$$

In this case, $\frac{\theta_1}{\theta_2}$ and $\frac{-1}{2\theta_2}$ are called canonical parametrization. It is also usual to use n, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$ as the sufficient statistics.

Exemple 16 If $X \sim \mathbf{Gam}(\alpha, \theta)$ then $T(x) = x$ can be chosen as a sufficient statistic.

Exemple 17 If $X \sim \mathbf{Gam}(\theta, \beta)$ then $T(x) = \ln x$ can be chosen as a sufficient statistic.

Exemple 18 If $X \sim \mathbf{Gam}(\theta_1, \theta_2)$ then $T_1(x) = \ln x$ and $T_2(x) = x$ can be chosen as a set of sufficient statistics.

Exemple 19 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathbf{Gam}(\theta_1, \theta_2)$ then it is easy to show that $T_1(\boldsymbol{x}) = \sum_{i=1}^n \ln x_i$ and $T_2(\boldsymbol{x}) = \sum_{i=1}^n x_i$.

Définition 6 [Exponential family] A class of distributions $\{P_\theta(\boldsymbol{x}), \theta \in \mathcal{T}\}$ is said to be an exponential family if there exist: $a(\boldsymbol{x})$ a function of Γ on $\mathbb{R}$, $g(\boldsymbol{\theta})$ a function of $\mathcal{T}$ on $\mathbb{R}^+$, $\phi_k(\boldsymbol{\theta})$ functions of $\mathcal{T}$ on $\mathbf{R}$, and $h_k(\boldsymbol{x})$ functions of Γ on $\mathbf{R}$ such that

$$\begin{aligned} p_\theta(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}) &= a(\boldsymbol{x})\,g(\boldsymbol{\theta})\exp\left[\sum_{k=1}^{K}\phi_k(\boldsymbol{\theta})\,h_k(\boldsymbol{x})\right] \\ &= a(\boldsymbol{x})\,g(\boldsymbol{\theta})\exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{x})\right] \end{aligned}$$

for all $\theta \in \mathcal{T}$ and $x \in \Gamma$. This family is entirely determined by $a(\boldsymbol{x})$, $g(\boldsymbol{\theta})$, and $\{\phi_k(\boldsymbol{\theta}), h_k(\boldsymbol{x}),\ k = 1, \cdots, K\}$ and is noted $\mathbf{Exfn}(\boldsymbol{x}|a, g, \boldsymbol{\phi}, \boldsymbol{h})$

Particular cases:

- When $a(\boldsymbol{x}) = 1$ and $g(\boldsymbol{\theta}) = \exp[-b(\boldsymbol{\theta})]$ we have

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{x}) - b(\boldsymbol{\theta})\right]$$

  and is noted $\mathbf{CExf}(\boldsymbol{x}|b, \boldsymbol{\phi}, \boldsymbol{h})$.

- Natural exponential family:
  When $a(\boldsymbol{x}) = 1$, $g(\boldsymbol{\theta}) = \exp[-b(\boldsymbol{\theta})]$, $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ we have

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{\theta}^t\boldsymbol{x} - b(\boldsymbol{\theta})\right] \mathbf{Exf}(\boldsymbol{x}|b).$$

  and is noted $\mathbf{NExf}(\boldsymbol{x}|b)$.

- Scalar random variable with a vector parameter:

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \mathbf{Exf}(x|a, g, \boldsymbol{\phi}, \boldsymbol{h}) \\ &= a(x)g(\boldsymbol{\theta}) \exp\left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta})h_k(x)\right] \\ &= a(x)g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(x)\right] \end{aligned}$$

  and is noted $\mathbf{Exfk}(\boldsymbol{x}|a, g, \boldsymbol{\phi}, \boldsymbol{h})$.

- Scalar random variable with a scalar parameter:

$$p(x|\theta) = \mathbf{Exf}(x|a, g, \phi, h) = a(x)g(\theta) \exp[\phi(\theta)h(x)]$$

  and is noted $\mathbf{Exf}(\boldsymbol{x}|a, g, \phi, h)$.

- Simple scalar exponential family:

$$p(x|\theta) = \theta \exp[-\theta x] = \exp[-\theta x + \ln\theta], \quad x \geq 0, \quad \theta \geq 0.$$

Définition 7 [Conjugate distributions] A family $\mathcal{F}$ of probability distributions $\pi(\boldsymbol{\theta})$ on $\mathcal{T}$ is said to be conjugate (or closed under sampling) if, for every $\pi(\boldsymbol{\theta}) \in \mathcal{F}$, the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ also belongs to $\mathcal{F}$.

The main argument for the development of the conjugate priors is the following: When the observation of a variable X with a probability law $f(x|\theta)$ modifies the prior $\pi(\theta)$ to a posterior $\pi(\theta|x)$, the information conveyed by x about θ is obviously limited, therefore it should not lead to a modification of the whole structure of $\pi(\theta)$, but only of its parameters.

Définition 8 [Conjugate priors] Assume that $f(\boldsymbol{x}|\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x}))$ where $\boldsymbol{t} = \{n, \boldsymbol{s}\} = \{n, s_1, \ldots, s_k\}$ is a vector of dimension $k+1$ and is sufficient statistic for $f(\boldsymbol{x}|\boldsymbol{\theta})$. Then, if there exists a vector $\{\tau_0, \boldsymbol{\tau}\} = \{\tau_0, \tau_1, \ldots, \tau_k\}$ such that

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{f(\boldsymbol{s} = (\tau_1, \cdots, \tau_k)|\boldsymbol{\theta}, n = \tau_0)}{\displaystyle\int f(\boldsymbol{s} = (\tau_1, \cdots, \tau_k)|\boldsymbol{\theta}', n = \tau_0)\, \mathrm{d}\boldsymbol{\theta}'}$$

exists and defines a family $\mathcal{F}$ of distributions for $\boldsymbol{\theta} \in \mathcal{T}$, then the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau})$ will remain in the same family $\mathcal{F}$. The prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\tau})$ is then a conjugate prior for the sampling distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Proposition 3 [Sufficient statistics for the exponential family] For a set of n i.i.d. samples $\{x_1, \cdots, x_n\}$ of a random variable $X \sim \mathbf{Exf}(x|a, g, \boldsymbol{\theta}, \boldsymbol{h})$ we have

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j|\boldsymbol{\theta}) &= [g(\boldsymbol{\theta})]^n \left(\prod_{j=1}^{n} a(x_j) \right) \exp \left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta}) \sum_{j=1}^{n} h_k(x_j) \right] \\ &= g^n(\boldsymbol{\theta})\, a(\boldsymbol{x}) \exp \left[\boldsymbol{\phi}^t(\boldsymbol{\theta}) \sum_{j=1}^{n} \boldsymbol{h}(x_j) \right], \end{aligned}$$

where $a(\boldsymbol{x}) = \prod_{j=1}^{n} a(x_j)$. Then, using the factorization theorem it is easy to see that

$$\boldsymbol{t} = \left\{ n, \sum_{j=1}^{n} h_1(x_j), \cdots, \sum_{j=1}^{n} h_K(x_j) \right\}$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Proposition 4 [Conjugate priors of the Exponential family] A conjugate prior family for the exponential family

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x})\, g(\boldsymbol{\theta}) \exp \left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta})\, h_k(\boldsymbol{x}) \right]$$

is given by

$$\pi(\boldsymbol{\theta}|\tau_0, \boldsymbol{\tau}) = z(\boldsymbol{\tau})[g(\boldsymbol{\theta})]^{\tau_0} \exp \left[\sum_{k=1}^{K} \tau_k \phi_k(\boldsymbol{\theta}) \right]$$

The associated posterior law is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \tau_0, \boldsymbol{\tau}) \propto [g(\boldsymbol{\theta})]^{n+\tau_0} a(\boldsymbol{x}) z(\boldsymbol{\tau}) \exp \left[\sum_{k=1}^{K} \left(\tau_k + \sum_{j=1}^{n} h_k(x_j) \right) \phi_k(\boldsymbol{\theta}) \right].$$

We can rewrite this in a more compact way:
If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \mathbf{Exfn}(\boldsymbol{x}|a(\boldsymbol{x}), g(\boldsymbol{\theta}), \boldsymbol{\phi}, \boldsymbol{h}),$$

then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) = \mathbf{Exfn}(\boldsymbol{\theta}|g^{\tau_0}, z(\boldsymbol{\tau}), \boldsymbol{\tau}, \boldsymbol{\phi}),$$

and the associated posterior law is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}) = \mathbf{Exfn}(\boldsymbol{\theta}|g^{n+\tau_0}, a(\boldsymbol{x})\, z(\boldsymbol{\tau}), \boldsymbol{\tau}', \boldsymbol{\phi})$$

where

$$\tau_k' = \tau_k + \sum_{j=1}^{n} h_k(x_j)$$

or

$$\boldsymbol{\tau}' = \boldsymbol{\tau} + \bar{\boldsymbol{h}}, \quad \text{with} \quad \bar{h}_k = \sum_{j=1}^{n} h_k(x_j).$$

Définition 9 [Conjugate priors of natural exponential family] If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta})\right]$$

Then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_0) = g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\tau}_0^t \boldsymbol{\theta} - d(\boldsymbol{\tau}_0)\right]$$

and the corresponding posterior is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}_0) = g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\tau}_n^t \boldsymbol{\theta} - d(\boldsymbol{\tau}_n)\right] \quad \text{with} \quad \boldsymbol{\tau}_n = \boldsymbol{\tau}_0 + \bar{\boldsymbol{x}}$$

where

$$\bar{\boldsymbol{x}}_n = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j$$

A slightly more general notation which gives some more explicit properties of the conjugate priors of the natural exponential family is the following:
If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta})\right]$$

Then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\alpha_0, \boldsymbol{\tau}_0) = g(\alpha_0, \boldsymbol{\tau}_0) \exp\left[\alpha_0 \boldsymbol{\tau}_0^t \boldsymbol{\theta} - \alpha_0 b(\boldsymbol{\tau}_0)\right]$$

The posterior is

$$\pi(\boldsymbol{\theta}|\alpha_0, \boldsymbol{\tau}_0, \boldsymbol{x}) = g(\alpha, \boldsymbol{\tau}) \exp\left[\alpha \boldsymbol{\tau}^t \boldsymbol{\theta} - \alpha b(\boldsymbol{\tau})\right]$$

with

$$\alpha = \alpha_0 + n \quad \text{and} \quad \boldsymbol{\tau} = \frac{\alpha_0 \boldsymbol{\tau}_0 + n\bar{\boldsymbol{x}}}{(\alpha_0 + n)})$$

and we have the following properties:

$$\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}\right] = \mathrm{E}\left[\bar{\boldsymbol{X}}|\boldsymbol{\theta}\right] = \nabla b(\boldsymbol{\theta})$$

$$\mathrm{E}\left[\nabla b(\boldsymbol{\Theta})|\alpha_0, \boldsymbol{\tau}_0\right] = \boldsymbol{\tau}_0$$

$$\mathrm{E}\left[\nabla b(\boldsymbol{\theta})|\alpha_0, \boldsymbol{\tau}_0, \boldsymbol{x}\right] = \frac{n\bar{\boldsymbol{x}} + \alpha_0 \boldsymbol{\tau}_0}{\alpha_0 + n} = \pi \bar{\boldsymbol{x}}_n + (1 - \pi)\boldsymbol{\tau}_0, \quad \text{with} \quad \pi = \frac{n}{\alpha_0 + n}$$

Conjugate priors

Observation law $p(x\|\theta)$	Prior law $p(\theta\|\boldsymbol{\tau})$	Posterior law $p(\theta\|x,\boldsymbol{\tau}) \propto p(\theta\|\boldsymbol{\tau})p(x\|\theta)$
Discrete variables		
Binomial $\mathbf{Bin}(x\|n,\theta)$	Beta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Beta $\mathbf{Bet}(\theta\|\alpha+x,\beta+n-x)$
Negative Binomial $\mathbf{NegBin}(x\|n,\theta)$	Beta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Beta $\mathbf{Bet}(\theta\|\alpha+n,\beta+x)$
Multinomial $\mathbf{M}_k(x\|\theta_1,\cdots,\theta_k)$	Dirichlet $\mathbf{Di}_k(\theta\|\alpha_1,\cdots,\alpha_k)$	Dirichlet $\mathbf{Di}_k(\theta\|\alpha_1+x_1,\cdots,\alpha_k+x_k)$
Poisson $\mathbf{Pn}(x\|\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\|\alpha+x,\beta+1)$
Gamma $\mathbf{Gam}(x\|\nu,\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\|\alpha+\nu,\beta+x)$
Beta $\mathbf{Bet}(x\|\alpha,\theta)$	Exponential $\mathbf{Ex}(\theta\|\lambda)$	Exponential $\mathbf{Ex}(\theta\|\lambda-\log(1-x))$
Normal $\mathbf{N}(x\|\theta,\sigma^2)$	Normal $\mathbf{N}(\theta\|\mu,\tau^2)$	Normal $\mathbf{N}\left(\mu\|\frac{\mu\sigma^2+\tau^2x}{\sigma^2+\tau^2},\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
Continuous variables		
Normal $\mathbf{N}(x\|\mu,1/\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}\left(\theta\|\alpha+\frac{1}{2},\beta+\frac{1}{2}(\mu-x)^2\right)$
Normal $\mathbf{N}(x\|\theta,\theta^2)$	Generalized inverse Normal $\mathbf{INg}(\theta\|\alpha,\mu,\sigma) \propto$ $\|\theta\|^{-\alpha}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{1}{\theta}-\mu\right)^2\right]$	Generalized inverse Normal $\mathbf{INg}(\theta\|\alpha_n,\mu_n,\sigma_n)$

Table 8.1: Relation between the sampling distributions, their associated conjugate priors and their corresponding posteriors

8.3 Non informative priors based on Fisher information

Another notion of information related to the maximum likelihood estimation is the Fisher information. In this section, first we give some definitions and results related to this notion and we see how this is used to define non informative priors.

Proposition 5 [Information Inequality] Let $\widehat{\theta}$ be an estimate of the parameter θ in a family $\{P_\theta; \theta \in \mathcal{T}\}$ and assume that the following conditions hold:

1. The family $\{P_\theta; \theta \in \mathcal{T}\}$ has a corresponding family of densities $\{p_\theta(x); \theta \in \mathcal{T}\}$, all with the same support.
2. $p_\theta(x)$ is differentiable for all $\theta \in \mathcal{T}$ and all x in its support.
3. The integral
$$g(\theta) = \int_\Gamma h(x)\, p_\theta(x)\, \mu(\,\mathrm{d}x)$$
exists and is differentiable for $\theta \in \mathcal{T}$, for $h(x) = \widehat{\theta}(x)$ and for $h(x) = 1$ and
$$\frac{\partial g(\theta)}{\partial \theta} = \int_\Gamma h(x)\, \frac{\partial p_\theta(x)}{\partial \theta}\, \mu(\,\mathrm{d}x)$$

Then

$$\mathrm{Var}_\theta[\widehat{\theta}(X)] \geq \frac{\left[\frac{\partial}{\partial \theta}\mathrm{E}_\theta\left\{\widehat{\theta}(X)\right\}\right]^2}{I_\theta} \tag{8.10}$$

where

$$I_\theta \stackrel{\mathrm{def}}{=} \mathrm{E}_\theta\left\{\left[\frac{\partial}{\partial \theta}\ln p_\theta(X)\right]^2\right\} \tag{8.11}$$

Furthermore, if $\frac{\partial^2}{\partial \theta^2} p_\theta(x)$ exists for all $\theta \in \mathcal{T}$ and all x in the support of $p_\theta(x)$, and if

$$\int \frac{\partial^2}{\partial \theta^2} p_\theta(x)\, \mu(\,\mathrm{d}x) = \frac{\partial^2}{\partial \theta^2} \int p_\theta(x)\, \mu(\,\mathrm{d}x)$$

then I_θ can be computed via

$$I_\theta = -\mathrm{E}_\theta\left\{\frac{\partial^2}{\partial \theta^2}\ln p_\theta(X)\right\} \tag{8.12}$$

The quantity defined in (8.11) is known as *Fisher's information* for estimating θ from X, and (8.10) is called the *information inequality*.

For the particular case in which $\widehat{\theta}$ is unbiased $\mathrm{E}_\theta\left\{\widehat{\theta}(X)\right\} = \theta$, the information inequality becomes

$$\mathrm{Var}_\theta[\widehat{\theta}(X)] \geq \frac{1}{I_\theta} \tag{8.13}$$

Expression $\frac{1}{I_\theta}$ is known as the *Cramer-Rao lower bound* (CRLB).

Exemple 20 [The information Inequality for exponential families] Assume that $\mathcal{T}$ is open and p_θ is given by

$$p_\theta(x) = a(x)\, g(\theta) \exp\left[g(\theta)\, h(x)\right]$$

Then it can be shown that

$$I_\theta \stackrel{\text{def}}{=} \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln p_\theta(X) \right]^2 \right\} = \left| g'(\theta) \right|^2 \mathrm{Var}_\theta(h(X)) \tag{8.14}$$

and

$$\frac{\partial}{\partial \theta} \mathrm{E}_\theta \left\{ h(X) \right\} = g'(\theta)\, \mathrm{Var}_\theta(h(X)) \tag{8.15}$$

and thus, if we choose $\widehat{\theta}(x) = h(x)$ we obtain the lower bound in the information inequality (8.10)

$$\mathrm{Var}_\theta[\widehat{\theta}(X)] = \frac{\left[\frac{\partial}{\partial \theta} \mathrm{E}_\theta \left\{ \widehat{\theta}(X) \right\} \right]^2}{I_\theta} \tag{8.16}$$

Définition 10 [Non informative priors] Assume $X \sim f(x|\theta) = p_\theta(x)$ and assume that

$$I_\theta \stackrel{\text{def}}{=} \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln p_\theta(X) \right]^2 \right\} = -\mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \ln p_\theta(X) \right\} \tag{8.17}$$

Then, a non informative prior $\pi(\theta)$ is defined as

$$\pi(\theta) \propto I_\theta^{1/2} \tag{8.18}$$

Définition 11 [Non informative priors, case of vector parameters]
Assume $X \sim f(x|\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(x)$ and assume that

$$I_{ij}(\theta) \stackrel{\text{def}}{=} -\mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_{\boldsymbol{\theta}}(X) \right\} \tag{8.19}$$

Then, a non informative prior $\pi(\boldsymbol{\theta})$ is defined as

$$\pi(\boldsymbol{\theta}) \propto \left| \boldsymbol{I}(\boldsymbol{\theta}) \right|^{1/2} \tag{8.20}$$

where $\boldsymbol{I}(\theta)$ is the Fisher information matrix with the elements $I_{ij}(\boldsymbol{\theta})$.

Exemple 21 If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta}) \right]$$

then

$$\boldsymbol{I}(\boldsymbol{\theta}) = \nabla \nabla^t b(\boldsymbol{\theta})$$

and

$$\pi(\boldsymbol{\theta}) \propto \left| \boldsymbol{I}(\boldsymbol{\theta}) \right|^{1/2} = \left| \prod_{i=1}^{n} \frac{\partial^2 \theta_i}{\partial b(\boldsymbol{\theta})^2} \right|^{1/2}$$

Exemple 22 If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}\left(\mu, \sigma^2\right), \quad \boldsymbol{\theta} = (\mu, \sigma^2)$$

then

$$I(\boldsymbol{\theta}) = \mathrm{E}_\theta \left\{ \begin{pmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^3} \\ \frac{2(X-\mu)}{\sigma^3} & \frac{3(X-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \end{pmatrix} \right\} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

and

$$\pi(\boldsymbol{\theta}) = \pi(\mu, \sigma^2) \propto \frac{1}{\sigma^4}$$