ST 337 / ST 405
Bayesian Forecasting and Intervention

Jeremie Houssineau

Chapter 1

Bayesian inference

1.1 Background

We first consider a single-variate random variable $\boldsymbol{\theta}$ on a parameter space Θ representing the uncertainty about a parameter of interest. The probability distribution of $\boldsymbol{\theta}$ is denoted $p_{\boldsymbol{\theta}}(\cdot)$. In general, Θ can be a subset of the set of integers $\mathbb{Z}$ or a subset of the real line $\mathbb{R}$. These subsets will most commonly be the positive natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$, the non-negative natural numbers $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ or the interval $[0, 1]$.

A different font will be used to denote random variables and their realisations, e.g. $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}$ and x, y, θ respectively. The difference between these notations is small enough to be ignored if necessary and yet visible enough to highlight the difference between the two concepts in more difficult situations. For instance, when defining the expectation of $\boldsymbol{\theta}$ as

$$\mathbb{E}(\boldsymbol{\theta}) = \int_{\Theta} \theta \, p_{\boldsymbol{\theta}}(\theta) \, \mathrm{d}\theta,$$

it appears that the argument of $\mathbb{E}(\cdot)$ is different from the variable θ in the integral, as it should be for the equation to be well defined.

The objective is to learn about some unknown/random quantity of interest, modelled by the random variable $\boldsymbol{\theta}$ on Θ, by accumulating information related to $\boldsymbol{\theta}$. Before receiving any additional information, we often have some initial knowledge about the credibility of the different possible values of this parameter, that we encode into a probability distribution $p_{\boldsymbol{\theta}}(\cdot)$ on Θ.

Some textbooks indicate the argument when denoting probability distributions, for instance $p_{\boldsymbol{\theta}}(\theta)$. We consider a different approach and write $p_{\boldsymbol{\theta}}(\cdot)$, which underlines the fact that the considered probability distribution is indeed a function while highlighting the difference between the function and its value at a given parameter $\theta \in \Theta$.

We use the term *probability distribution* to refer to any *probability mass function* (p.m.f.) on a discrete space or to any *probability density function* (p.d.f.) otherwise. Note that if $\boldsymbol{x}$ is a random variable on $\mathbb{R}$ with probability density function $p_{\boldsymbol{x}}(\cdot)$, then there might be some $x \in \mathbb{R}$ such that $p_{\boldsymbol{x}}(x) > 1$. For instance, if any given $x \in \mathbb{R}$ denotes a distance from the origin measured in meters, then $p_{\boldsymbol{x}}(x)$ is the *density* of probability per meter. However, we indeed have

$$\int_B p_{\boldsymbol{x}}(x) \, \mathrm{d}x \in [0, 1]$$

for any subset B of $\mathbb{R}$. This is because $\mathrm{d}x$ can be seen as being expressed in meters, which makes the integral dimensionless.

Once a prior distribution representing our knowledge about $\boldsymbol{\theta}$ has been defined, we need to find a way to update this knowledge in the light of newly received information. We assume that this new piece of information takes the form of a point observation y in an observation space $\mathcal{Y}$. In general, an observation originates from a random experiment and is only indirectly related to $\boldsymbol{\theta}$ (otherwise there would be no inference problem). Therefore, we need to model how the different parameter values in Θ affect the distribution of the corresponding observation random variable $\boldsymbol{y}$. This is achieved via a

conditional probability distribution $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \cdot)$, which verifies

$$\int_{\mathcal{Y}} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) \, \mathrm{d}y = 1$$

for any $\theta \in \Theta$. The notation $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \theta)$ is a shorthand for $p_{\boldsymbol{y}}(\cdot \,|\, \boldsymbol{\theta} = \theta)$. In this context, $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \cdot)$ is usually referred to as the data distribution. The joint probability distribution $p_{\boldsymbol{\theta},\boldsymbol{y}}(\cdot)$ on $\Theta \times \mathcal{Y}$ can be expressed via conditional and prior distributions in two different ways:

$$\begin{aligned} p_{\boldsymbol{\theta},\boldsymbol{y}}(\theta, y) &= p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) p_{\boldsymbol{\theta}}(\theta) \\ &= p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) p_{\boldsymbol{y}}(y). \end{aligned}$$

This simple relation leads to the most fundamental result in Bayesian inference.

Theorem 1.1 (Bayes' rule)**.** *Let the prior distribution of $\boldsymbol{\theta}$ on Θ be $p_{\boldsymbol{\theta}}(\cdot)$ and the let the random variable $\boldsymbol{y}$ on $\mathcal{Y}$ have conditional distribution $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \cdot)$. If y is a given realisation of $\boldsymbol{y}$ then the posterior distribution of $\boldsymbol{\theta}$ given the observation y is*

$$p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) = \frac{p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) p_{\boldsymbol{\theta}}(\theta)}{p_{\boldsymbol{y}}(y)} \tag{1.2}$$

where $p_{\boldsymbol{y}}(\cdot)$ is the marginal distribution of $\boldsymbol{y}$ defined as

$$p_{\boldsymbol{y}}(y) = \int_{\Theta} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) p_{\boldsymbol{\theta}}(\theta) \, \mathrm{d}\theta.$$

A simple way to verify that there is a gain of information when going from the prior $p_{\boldsymbol{\theta}}(\cdot)$ to $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, \cdot)$ is to rewrite the variance of the prior distribution as

$$\underbrace{\mathrm{var}(\boldsymbol{\theta})}_{\text{prior variance}} = \underbrace{\mathbb{E}(\mathrm{var}(\boldsymbol{\theta} \,|\, \boldsymbol{y}))}_{\text{expected posterior variance}} + \mathrm{var}(\mathbb{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y})),$$

which shows an improvement in terms of variance since $\mathrm{var}(\mathbb{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y})) \geq 0$.

Since the observation y is given, it is also natural to see the data distribution as a function of the parameter with y fixed: the function $\theta \mapsto p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta)$ can then be referred to as the *likelihood function*. Also, since y is fixed and since the term $p_{\boldsymbol{y}}(y)$ in the denominator of (1.2) does not depend on θ, it follows that this term can be considered as a normalising constant and Bayes' rule can be simply expressed as

$$p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) \propto p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) p_{\boldsymbol{\theta}}(\theta).$$

Summarising the information a posteriori. The posterior $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, y)$ contains all the information one might be interested in, which is sometimes more than required. The simplest way of summarising a posterior distribution is via point estimates, for instance the mean $\mathbb{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y} = y)$, the median or the mode, defined as

$$\theta_{\mathrm{MAP}} = \operatorname*{argmax}_{\theta \in \Theta} p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y),$$

assuming it exists, where MAP stands for Maximum A Posteriori. However, summarising a posterior distribution by a single point can be misleading as it does not allow for gauging the underlying amount of uncertainty. This concern can be addressed by also considering posterior quantiles or intervals. Of particular interest is the $100(1-\alpha)\%$ *highest posterior density region* which is defined as the subset $B \subseteq \Theta$ that contains $100(1-\alpha)\%$ of the posterior probability, i.e.

$$\int_B p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) \, \mathrm{d}\theta = 1 - \alpha,$$

and such that $p_{\boldsymbol{\theta}}(\theta) \geq p_{\boldsymbol{\theta}}(\theta')$ for any $\theta \in B$ and any $\theta' \in \Theta \setminus B$. The advantage of the highest posterior density region is best seen on multimodal probability distributions, as illustrated on Figure 1.1.

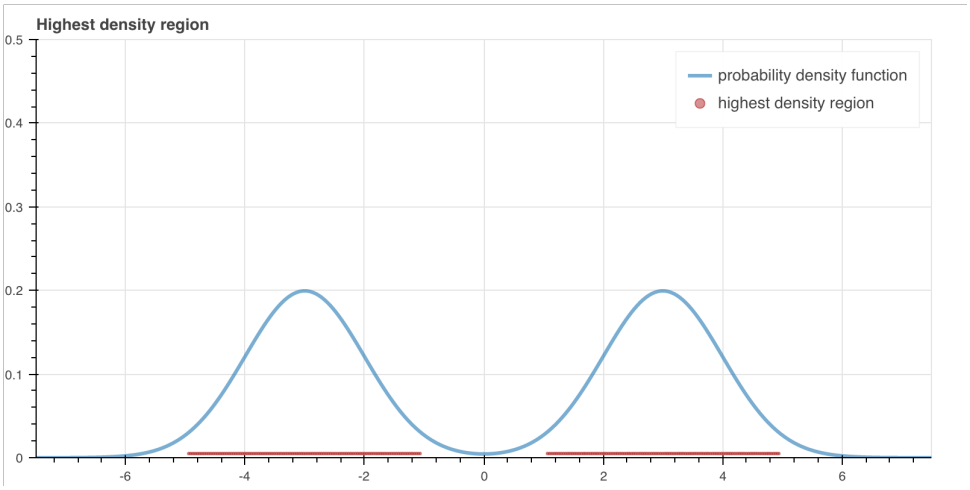

Figure 1.1: 95% highest density region for a Gaussian mixture.

Conditional independence of observations. In some cases, the objective is to predict the distribution of another observation originated from the same experimental setting. Let $\boldsymbol{y}'$ be the random variable describing this experiment. It is common (and convenient) to assume that $\boldsymbol{y}'$ does not depend on $\boldsymbol{y}$ for any given value θ of the parameter of interest that is, formally

$$p_{\boldsymbol{y},\boldsymbol{y}'|\boldsymbol{\theta}}(y, y' \,|\, \theta) = p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) p_{\boldsymbol{y}'|\boldsymbol{\theta}}(y' \,|\, \theta).$$

In particular, this assumption allows for applying Bayes' rule once more directly on the posterior distribution $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, y)$, which yields

$$p_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{y}'}(\theta \,|\, y, y') \propto p_{\boldsymbol{y}'|\boldsymbol{\theta}}(y'|\theta) p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y).$$

This assumption of conditional independence between observations will generally apply to the model we will consider and, if we receive a sequence $(y_1, \dots, y_n)$ of n observations resulting from the sequence of random variables $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, then it will be assumed that $\boldsymbol{y}_k$ and $\boldsymbol{y}_{k'}$ will be conditionally independent given $\boldsymbol{\theta}$ for any $k, k' \in \{1, \dots, n\}$. Formally, this assumption implies that

$$p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1, \dots, y_n \,|\, \theta) = \prod_{k=1}^{n} p_{\boldsymbol{y}_k|\boldsymbol{\theta}}(y_k \,|\, \theta)$$

so that the posterior distribution of $\boldsymbol{\theta}$ given $(y_1, \dots, y_n)$ is characterised by

$$p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\theta \,|\, y_1, \dots, y_n) \propto \prod_{k=1}^{n} p_{\boldsymbol{y}_k|\boldsymbol{\theta}}(y_k \,|\, \theta) p_{\boldsymbol{\theta}}(\theta).$$

This expression can be slightly simplified by assuming that the conditional distribution of the observation random variables is the same for any index in $\{1, \dots, n\}$, which will be denoted $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \cdot)$. In this case, it follows that

$$p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\theta \,|\, y_1, \dots, y_n) = \frac{\prod_{k=1}^{n} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y_k \,|\, \theta) p_{\boldsymbol{\theta}}(\theta)}{p_{\boldsymbol{y}_{1:n}}(y_1, \dots, y_n)} \tag{1.3}$$

with

$$p_{\boldsymbol{y}_{1:n}}(y_1, \dots, y_n) = \int \prod_{k=1}^{n} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y_k \,|\, \theta) p_{\boldsymbol{\theta}}(\theta) \, \mathrm{d}\theta.$$

Note that in this case, the sequence of random variables $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ is *exchangeable*: any reordering of the sequence would not change the associated joint probability distribution, that is, formally

$$p_{\boldsymbol{y}_1, \dots, \boldsymbol{y}_n}(\cdot) = p_{\boldsymbol{y}_{\sigma(1)}, \dots, \boldsymbol{y}_{\sigma(n)}}(\cdot)$$

for any permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$.

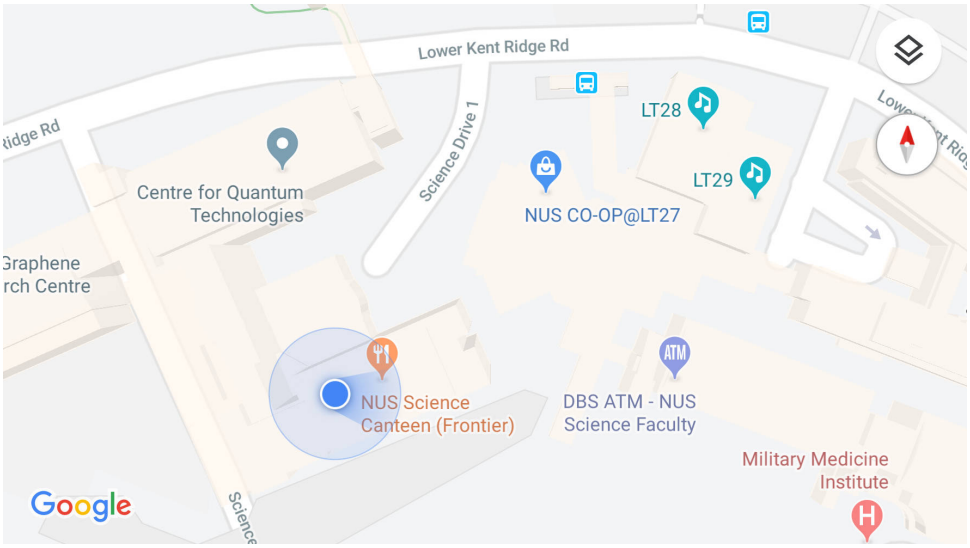

Figure 1.2: Estimating an unknown position from GPS data in Google Maps®.

Example 1.1. To determine your position, your smartphone (or any GPS-enabled device), uses a sequences of observations from satellites. Figure 1.2 illustrates the output of some estimation algorithm used by Google to determine an unknown/uncertain position from GPS data. The larger blue circle gives an idea about the uncertainty in the current estimate, which we expect to shrink in time (when additional data is received).

1.2 Standard distributions and conjugate priors

Equations of the form (1.3) can be difficult to work with in practice. The first step to address that difficulty is to assume that the data distribution $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot\,|\,\cdot)$ and the prior distribution $p_{\boldsymbol{\theta}}(\cdot)$ take a particular parametric form such as

Bernoulli p.m.f. on $\{0,1\}$ with parameter $p \in (0,1)$

$$\mathrm{Ber}(k;p) = p^k(1-p)^{1-k} = \begin{cases} 1-p & \text{if } k=0 \\ p & \text{if } k=1 \end{cases}$$

The parameter p is the probability of success.

Binomial p.m.f. on $\{0,\dots,n\}$ with parameter $n \in \mathbb{N}_0$ and $p \in (0,1)$

$$\mathrm{Bi}(k;n,p) = \binom{n}{k} p^k(1-p)^{n-k}$$

The parameter p is the probability of success and n is the number of trials.

Poisson p.m.f. on $\mathbb{N}_0$ with parameter $\lambda > 0$

$$\mathrm{Po}(k;\lambda) = \frac{\lambda^k}{k!}\exp(-\lambda)$$

Gamma p.d.f. on $(0,\infty)$ with parameters $\alpha > 0$ and $\beta > 0$

$$\mathrm{Ga}(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

where $\Gamma(\cdot)$ is the gamma function, satisfying $\Gamma(k+1) = k\Gamma(k)$ for any $k > 0$. The positive scalars α and β are respectively called the *shape parameter* and the *rate parameter*. If $\boldsymbol{x} \sim \mathrm{Ga}(\cdot\,;\alpha,\beta)$ then $\mathbb{E}(\boldsymbol{x}) = \alpha/\beta$ and $\mathrm{var}(\boldsymbol{x}) = \alpha/\beta^2$.

Beta p.d.f. on $[0,1]$ with shape parameters $\alpha > 0$ and $\beta > 0$

$$\mathrm{Be}(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\mathrm{B}(\cdot)$ is the beta function, satisfying

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Normal p.d.f. on $\mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$

$$\mathrm{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big).$$

Uniform p.d.f. on $\mathbb{R}$ with parameters $a, b \in \mathbb{R}$ such that $a < b$

$$\mathrm{U}(x;a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Student's t p.d.f. on $\mathbb{R}$ with $\nu > 0$ degrees of freedom

$$\mathrm{St}(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \Big(1 + \frac{x^2}{\nu}\Big)^{-\frac{\nu+1}{2}}$$

Generalised Student's t p.d.f. on $\mathbb{R}$ with $\nu > 0$ degrees of freedom and with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$

$$\mathrm{St}(x;\nu,\mu,\sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma} \Big(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\Big)^{-\frac{\nu+1}{2}}$$

In the standard distributions described above, a semicolon ";" was used instead of the conditioning bar "|" to indicate that one or several of the arguments might not be part of the inference problem and might be known instead.

Example 1.2. We consider a large population for which we want to know the proportion of individuals with a given trait. We denote by $\boldsymbol{\theta}$ the random variable on $\Theta = [0,1]$ describing the uncertainty about this proportion of interest and by $p_{\boldsymbol{\theta}}$ the associated probability distribution. In order to learn about $\boldsymbol{\theta}$, we observe the presence of the trait in n individuals in the population and denote by $y_1, \dots, y_n$ the corresponding observations taking value in $\{0,1\}$, 1 for *success*, that is the trait is present, and 0 for *failure*, the trait is absent. For a given parameter value $\theta \in \Theta$, the conditional distribution of the corresponding random variables $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ is deduced to be

$$p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1, \dots, y_n \,|\, \theta) = \prod_{k=1}^{n} \mathrm{Ber}(y_k;\theta).$$

However, since the random variables $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are exchangeable, the data provided by the observations $y_1, \dots, y_n$ can be more simply expressed as a single piece of data $y = y_1 + \cdots + y_k$ with data distribution $p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) = \mathrm{Bi}(y;n,\theta)$. The beta distribution is defined on $[0,1]$, can be used to express a wide range of prior knowledge on $\boldsymbol{\theta}$ and, as such, is a suitable prior distribution. Formally, we consider $p_{\boldsymbol{\theta}}(\theta) = \mathrm{Be}(\theta;\alpha,\beta)$ for some $\alpha > 0$ and $\beta > 0$ and it follows that

$$\begin{aligned} p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) &\propto \theta^y (1-\theta)^{n-y} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \end{aligned}$$

so that $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) = \mathrm{Be}(\theta;\alpha+y,\beta+n-y)$.

There are two important aspects in Example 1.2: a) we did not have to compute the normalising constant $p_{\boldsymbol{y}}(y)$ to find the posterior distribution, and b) the posterior distribution is of the same form as the prior distribution, that is, they are both beta distributions (although with different parameters). This second aspect is very important in Bayesian inference and bears a specific name as follows.

Definition 1.1. A class of prior distributions $\mathcal{P}$ is said to be *conjugate* for a class of likelihood functions $\mathcal{F}$ if $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot\,|\,\cdot) \in \mathcal{P}$ for all $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot\,|\,\cdot) \in \mathcal{F}$ and all $p_{\boldsymbol{\theta}}(\cdot) \in \mathcal{P}$.

Here are few useful examples of conjugacy:

Binomial-Beta The class of priors $\mathcal{P} = \{\mathrm{Be}(\cdot\,;\alpha,\beta) : \alpha > 0, \beta > 0\}$ is conjugate for the class of likelihoods $\mathcal{F} = \{\theta \mapsto \mathrm{Bi}(\cdot\,;n,\theta) : n \in \mathbb{N}_0\}$

Normal-Normal The class of priors $\mathcal{P} = \{\mathrm{N}(\cdot\,;\mu_0,\sigma_0^2) : \mu_0 \in \mathbb{R}, \sigma_0^2 > 0\}$ is conjugate for the class of likelihoods $\mathcal{F} = \{\theta \mapsto \mathrm{N}(\cdot\,;\theta,\sigma^2) : \sigma^2 > 0\}$

Gamma-Gamma The class of priors $\mathcal{P} = \{\mathrm{Ga}(\cdot\,;\alpha_0,\beta_0) : \alpha_0 > 0, \beta_0 > 0\}$ is conjugate for the class of likelihoods $\mathcal{F} = \{\theta \mapsto \mathrm{Ga}(\cdot\,;\alpha,\theta) : \alpha > 0\}$

Another illustration of this concept is given in the following example based on the Poisson p.m.f. and the gamma p.d.f.

Example 1.3. We now assume that the random experiments $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ yielding the observations $y_1, \ldots, y_n$ are independent integer-valued random variables distributed according to a Poisson p.m.f. with unknown parameter $\boldsymbol{\theta}$, that is

$$p_{\boldsymbol{y}_i|\boldsymbol{\theta}}(y_i\,|\,\theta) = \mathrm{Po}(y_i;\theta) = \frac{\theta^{y_i}}{y_i!}\exp(-\theta)$$

for any $i \in \{1,\ldots,n\}$. The likelihood for the n observations can be then be expressed as

$$p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1,\ldots,y_n\,|\,\theta) = \frac{\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}\exp(-n\theta).$$

The parameter of a Poisson p.m.f. is a positive scalar so that the random variable $\boldsymbol{\theta}$ can be assumed to have a gamma p.d.f. a priori

$$p_{\boldsymbol{\theta}}(\theta) = \mathrm{Ga}(\theta;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$$

for some parameters $\alpha, \beta > 0$. It follows that the posterior distribution of $\boldsymbol{\theta}$ given the observations $y_1, \ldots, y_n$ take the form

$$\begin{aligned} p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\theta\,|\,y_1,\ldots,y_n) &\propto p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1,\ldots,y_n\,|\,\theta)p_{\boldsymbol{\theta}}(\theta) \\ &\propto \theta^{\alpha+n\bar{y}_n-1}\exp(-(\beta+n)\theta), \end{aligned}$$

with $\bar{y}_n = n^{-1}\sum_{i=1}^n y_i$. We can recognise in this last expression that the posterior distribution takes the same form as a gamma distribution with parameters $\alpha' = \alpha + n\bar{y}_n$ and $\beta' = \beta + n$. It can be easily verified that these posterior parameters are still positive scalars. Since $p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\cdot\,|\,\cdot)$ integrates to 1 by definition, it follows that

$$p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\theta\,|\,y_1,\ldots,y_n) = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\theta^{\alpha'-1}\exp(-\beta'\theta),$$

so that the class of priors $\mathcal{P} = \{\mathrm{Ga}(\cdot\,;\alpha,\beta) : \alpha > 0, \beta > 0\}$ is conjugate for the class of likelihoods $\mathcal{F} = \{\theta \mapsto \mathrm{Po}(\cdot\,;\theta)\}$.

1.3 A closer look at the normal distribution

The normal distribution plays a central role in Statistics and Probability since it is a natural choice for a distribution of errors. In particular, if we perform a random experiment $\boldsymbol{y}$ to measure a quantity of interest θ and if the measurement error can be expressed as the sum of a large quantity of small deviations, then the distribution of the error can often be assumed normally distributed.[1]

[1] as a consequence of the central limit theorem

1.3.1 Unknown mean and known variance

In this situation, it holds that

$$\begin{aligned} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) &= \mathrm{N}(y; \theta, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(y-\theta)^2 \Big) \end{aligned}$$

where the parameter σ, referred to as the *standard deviation* is assumed to be known. Another parametrisation of the normal distribution is based on the precision $\tau = 1/\sigma^2$, that is

$$\mathrm{N}(y; \theta, \tau^{-1}) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\Big(-\frac{\tau}{2}(y-\theta)^2 \Big)$$

Similarly, if we have some prior knowledge about $\boldsymbol{\theta}$ which can be informally expressed as "the parameter should be more or less $\mu_0 \in \mathbb{R}$ with no preference for lower or higher values", then we can choose the prior distribution

$$p_{\boldsymbol{\theta}}(\theta) = \mathrm{N}(\theta; \mu_0, \sigma_0^2)$$

where the standard deviation σ_0 quantifies the uncertainty.

Posterior distribution Introducing $\tau_0 = 1/\sigma_0^2$ as the prior precision, we want to gather the terms that depend on θ in the expression of $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, y)$ as follows

$$\begin{aligned} p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) &\propto \exp\Big(-\frac{\tau}{2}(y-\theta)^2 - \frac{\tau_0}{2}(\theta-\mu_0)^2 \Big) \\ &= \exp\Big(-\frac{1}{2}\Big[(\tau+\tau_0)\theta^2 + -2(\tau y + \tau_0\mu_0)\theta + \tau y^2 + \tau_0\mu_0^2 \Big] \Big) \\ &= \exp\Big(-\frac{1}{2}\Big[(\tau+\tau_0)\Big(\theta^2 + -2\frac{\tau y + \tau_0\mu_0}{\tau+\tau_0}\theta + \frac{(\tau y + \tau_0\mu_0)^2}{(\tau+\tau_0)^2} \Big) + R(y, \mu_0, \tau_0) \Big] \Big) \\ &= \exp\Big(-\frac{(\tau+\tau_0)}{2}\Big(\theta - \frac{\tau y + \tau_0\mu_0}{\tau+\tau_0} \Big)^2 \Big) \exp\Big(-\frac{1}{2} R(y, \mu_0, \tau_0) \Big) \end{aligned}$$

where $R(y, \mu_0, \tau_0)$ is a remainder which does not depend on θ. We find that the posterior distribution $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, y)$ is equal to $\mathrm{N}(\cdot\,; \mu_1, \tau_1^{-1})$ with

$$\mu_1 = \frac{\tau y + \tau_0\mu_0}{\tau_0 + \tau} \qquad \text{and} \qquad \tau_1 = \tau_0 + \tau.$$

Coming back to standard deviations, we have

$$\mu_1 = \frac{\sigma_0^2 y + \sigma^2 \mu_0}{\sigma_0^2 + \sigma^2} \qquad \text{and} \qquad \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma^2}.$$

It appears that the larger σ_0 is compared to σ the more importance will the observation have in the posterior mean μ_1, or simply put, that μ_1 is a weighted average between y and μ_0. The posterior mean can also be expressed as

$$\mu_1 = \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(y - \mu_0)$$

where the "update" of the prior mean by the observation is made more apparent. This expression will be important in more general cases as will become clear later on.

Predictive distribution If we conduct another random experiment $\boldsymbol{y}'$ independently and with the same characteristics, that is such that $\boldsymbol{y}'$ is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ and such that $p_{\boldsymbol{y}'|\boldsymbol{\theta}}(\cdot \,|\, \theta) = p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot \,|\, \theta)$, and if we try to predict its outcome y', we obtain the *predictive distribution*

$$\begin{aligned} p_{\boldsymbol{y}'|\boldsymbol{y}}(y' \,|\, y) &= \int p_{\boldsymbol{y}'|\boldsymbol{\theta},\boldsymbol{y}}(y' \,|\, \theta, y) p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) \,\mathrm{d}\theta \\ &= \int p_{\boldsymbol{y}'|\boldsymbol{\theta}}(y' \,|\, \theta) p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta \,|\, y) \,\mathrm{d}\theta \\ &\propto \int \exp\Big(-\frac{1}{2\sigma^2}(y'-\theta)^2 - \frac{1}{2\sigma_1^2}(\theta-\mu_1)^2 \Big) \,\mathrm{d}\theta. \end{aligned}$$

Instead of making an explicit computation of the integral, we can use the facts that $\mathbb{E}(\boldsymbol{y}' \,|\, \boldsymbol{\theta} = \theta) = \theta$ and $\mathrm{var}(\boldsymbol{y}' \,|\, \boldsymbol{\theta} = \theta) = \sigma^2$ to deduce the mean and variance of the predictive distribution $p_{\boldsymbol{y}'|\boldsymbol{y}}(\cdot \,|\, y)$ as

$$\mathbb{E}(\boldsymbol{y}' \,|\, \boldsymbol{y} = y) = \mathbb{E}(\mathbb{E}(\boldsymbol{y}' \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y} = y) = \mathbb{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y} = y) = \mu_1$$

and

$$\begin{aligned}\mathrm{var}(\boldsymbol{y}' \,|\, \boldsymbol{y} = y) &= \mathbb{E}(\mathrm{var}(\boldsymbol{y}' \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y} = y) + \mathrm{var}(\mathbb{E}(\boldsymbol{y}' \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y} = y) \\ &= \mathbb{E}(\sigma^2 \,|\, \boldsymbol{y} = y) + \mathrm{var}(\boldsymbol{\theta} \,|\, \boldsymbol{y} = y) \\ &= \sigma^2 + \sigma_1^2\end{aligned}$$

However, this does not show that the predictive distribution is normal! Yet, this can help us for the calculation of the corresponding probability distribution. In particular, following the same approach as with the posterior distribution $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot \,|\, \cdot)$ above, we obtain a remainder $R(y', \mu_1, \tau_1)$ of the same form as before with $\tau_1 = 1/\sigma_1^2$, and it follows that

$$\begin{aligned}R(y', \mu_1, \tau_1) &= \tau y'^2 + \tau_1 \mu_1^2 - \frac{(\tau y' + \tau_1 \mu_1)^2}{(\tau + \tau_1)} \\ &= \frac{y'^2}{\sigma^2} + \frac{\mu_1^2}{\sigma_1^2} - \frac{(\sigma_1^2 y' + \sigma^2 \mu_1)^2}{\sigma^2 \sigma_1^2 (\sigma^2 + \sigma_1^2)} \\ &= \frac{1}{\sigma^2 + \sigma_1^2}\Big((\sigma^2 + \sigma_1^2) \frac{\sigma_1^2 y'^2 + \sigma^2 \mu_1^2}{\sigma^2 \sigma_1^2} - \frac{\sigma_1^4 y'^2 + 2\sigma^2 \sigma_1^2 y' \mu_1 + \sigma^4 \mu_1^2}{\sigma^2 \sigma_1^2} \Big) \\ &= \frac{1}{\sigma^2 + \sigma_1^2}\Big(y'^2 + \mu_1^2 - 2 y' \mu_1 \Big),\end{aligned}$$

and it holds that

$$p_{\boldsymbol{y}'|\boldsymbol{y}}(y' \,|\, y) \propto \exp\Big(-\frac{1}{2(\sigma^2 + \sigma_1^2)} (y' - \mu_1)^2 \Big),$$

from which we conclude that $p_{\boldsymbol{y}'|\boldsymbol{y}}(\cdot \,|\, y)$ is indeed normal for any observation y.

Multiple observations Consider the case where n observations $y_1, \ldots, y_n$ resulting from the sequence of conditionally independent random experiments $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ have been received. To determine the posterior distribution $p_{\boldsymbol{\theta}|\boldsymbol{y}_{1:n}}(\cdot \,|\, y_1, \ldots, y_n)$, it would be possible to iterate n times the calculations made for the posterior distribution given a single observation. However, we can also remark that

$$\begin{aligned}p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1, \ldots, y_n \,|\, \theta) &= \prod_{i=1}^n p_{\boldsymbol{y}_i|\boldsymbol{\theta}}(y_i \,|\, \theta) = \prod_{i=1}^n \mathrm{N}(y_i; \theta, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \Big) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^n \big((y_i - \bar{y}_n) - (\theta - \bar{y}_n) \big)^2 \Big)\end{aligned}$$

where $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$. This expression can then be simplified by expanding the square and noticing that $\sum_{i=1}^n (y_i - \bar{y}_n) = 0$, which yields

$$\begin{aligned}p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1, \ldots, y_n \,|\, \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \bar{y}_n)^2 - \frac{n}{2\sigma^2} (\theta - \bar{y}_n)^2 \Big) \\ &\propto \exp\Big(-\frac{n}{2\sigma^2} (\theta - \bar{y}_n)^2 \Big).\end{aligned}$$

Since multiplicative coefficients that are constant in θ have no effect on the posterior distribution when using Bayes' rule, we can replace $p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(\cdot \,|\, \theta)$ by the conditional distribution

$$p_{\bar{\boldsymbol{y}}_n|\boldsymbol{\theta}}(\bar{y}_n \,|\, \theta) = \mathrm{N}(\bar{y}_n; \theta, \sigma^2/n),$$

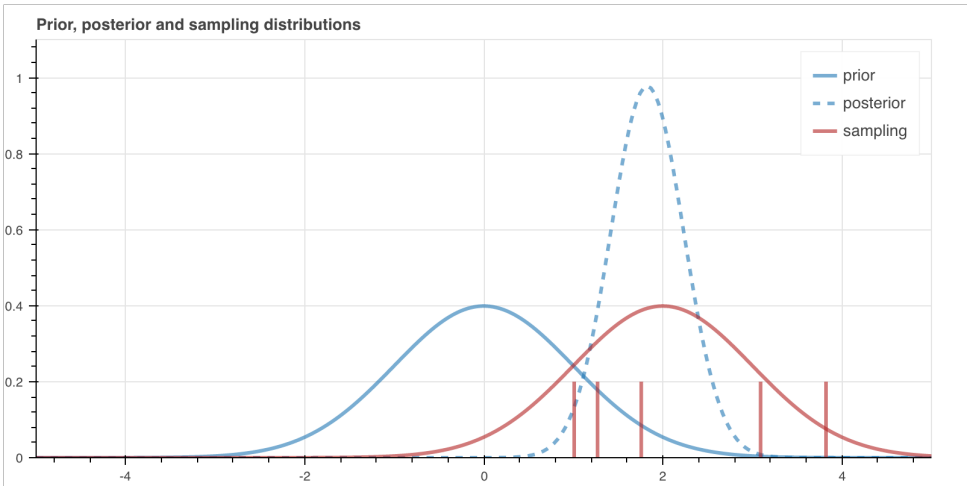

Figure 1.3: Normal sampling distribution with unknown mean.

where $\bar{\boldsymbol{y}}_n$ is the random experiment consisting of taking the average of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. The fact that the posterior distribution only depends on the observations via the average $\bar{y}_n$ shows that $\bar{y}_n$ is a *sufficient statistics*. The posterior mean and variance corresponding to this likelihood are respectively

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{y}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \qquad \text{and} \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

This result leads to some important conclusions: 1. the mean is more and more influenced by $\bar{y}_n$ when n increases and 2. the variance modelling our uncertainty about the true value of the parameter is inversely proportional to n, in particular, the variance goes to 0 when n tends to infinity. A illustration of the prior and posterior distributions for multiple observations is given in Figure 1.3.

Following the same type of calculations as before, we also find that

$$p_{\boldsymbol{y}_{1:n}}(y_1, \ldots, y_n) \propto \exp\Big(-\frac{n}{2(\sigma^2 + n\sigma_0^2)}(\bar{y}_n - \mu_0)^2\Big).$$

1.3.2 Known mean and unknown variance

In this situation, we consider that the parameter is equal to the precision of the observations, that is

$$p_{\boldsymbol{y}|\boldsymbol{\theta}}(y \,|\, \theta) = \mathrm{N}(y; \mu, \theta^{-1}) \propto \sqrt{\theta}\exp\Big(-\frac{\theta}{2}(y-\mu)^2\Big)$$

for a given mean μ. As before, we would like to have a conjugate prior for this likelihood; however, assuming that $\boldsymbol{\theta}$ is normally distributed is not going to work since we clearly want to avoid non-positive values. To make an educated guess about the type of prior distribution we are looking for, it is useful to consider again the case of n observations $y_1, \ldots, y_n$. Indeed, we find that

$$\begin{aligned} p_{\boldsymbol{y}_{1:n}|\boldsymbol{\theta}}(y_1, \ldots, y_n \,|\, \theta) &= \frac{\theta^{n/2}}{(2\pi)^{n/2}}\exp\Big(-\frac{\theta}{2}\sum_{i=1}^n (y_i - \mu)^2\Big) \\ &\propto \theta^{n/2}\exp\Big(-\frac{n\theta}{2}v\Big) \end{aligned}$$

with $v = n^{-1}\sum_{i=1}^n (y_i - \mu)^2$ the sufficient statistics for this model. This suggests the use of a gamma distribution as a prior, that is

$$p_{\boldsymbol{\theta}}(\theta) = \mathrm{Ga}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}\exp(-\beta\theta)$$

for some parameters $\alpha > 0$ and $\beta > 0$. This choice of prior also addresses our concern about positivity of the parameter. It follows easily that the posterior distribution is gamma with parameters

$$\alpha_n = \alpha + \frac{n}{2} \qquad \text{and} \qquad \beta_n = \beta + \frac{nv}{2}.$$

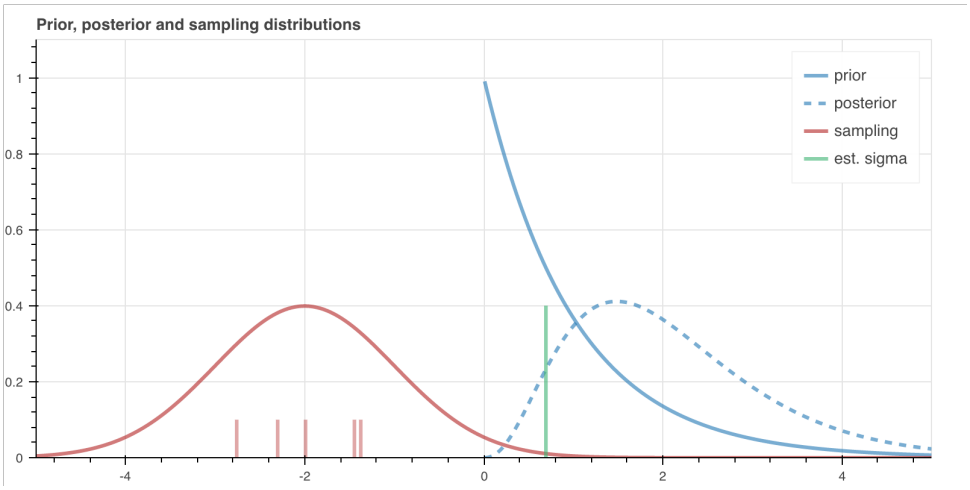

Figure 1.4: Normal sampling distribution with unknown variance.

Note that the prior mean $\mathbb{E}(\boldsymbol{\theta}) = \alpha/\beta$ and the prior variance $\text{var}(\boldsymbol{\theta}) = \alpha/\beta^2$ induced by the gamma distribution can be used to set the values of α and β. A illustration of the prior and posterior distributions for multiple observations is given in Figure 1.4.

1.3.3 Unknown mean and variance

In the case where both the mean and variance are unknown, the random variable $\boldsymbol{\theta}$ has two components: the random variable $\boldsymbol{\mu}$ describing the unknown mean and the random variable $\boldsymbol{\tau}$ describing the unknown precision, that is $\boldsymbol{\theta}$ is the random vector $(\boldsymbol{\mu}, \boldsymbol{\tau})^\intercal$ with $\cdot^\intercal$ denoting the transposition. Proceeding as before, we write the likelihood and study its form as a function of the unknown mean and variance

$$p_{\boldsymbol{y}_{1:n}|\boldsymbol{\mu},\boldsymbol{\tau}}(y_1, \dots, y_n \,|\, \mu, \tau) = \frac{\tau^{n/2}}{(2\pi)^{n/2}} \exp\Big(-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \Big) \tag{1.15a}$$

$$\propto \tau^{n/2} \exp\Big(-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 - \frac{n\tau}{2} (\mu - \bar{y}_n)^2 \Big) \tag{1.15b}$$

$$\propto \Big(\exp\Big(-\frac{n\tau}{2} (\mu - \bar{y}_n)^2 \Big) \Big) \Big(\tau^{n/2} \exp\Big(-\frac{n\hat{v}}{2} \tau \Big) \Big) \tag{1.15c}$$

with $\hat{v} = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$ the sample variance, which suggests a normal prior for $\boldsymbol{\mu} \,|\, \boldsymbol{\tau}$ and a gamma prior for $\boldsymbol{\tau}$. In particular, we set

$$p_{\boldsymbol{\mu}|\boldsymbol{\tau}}(\mu \,|\, \tau) = \text{N}\big(\mu; \mu_0, (k\tau)^{-1}\big)$$

with μ_0 the prior mean and $k \in \mathbb{N}$, and

$$p_{\boldsymbol{\tau}}(\tau) = \text{Ga}(\tau; \alpha, \beta),$$

for some $\alpha > 0$ and $\beta > 0$. The parameter k can be interpreted as the number of observations that the prior distribution is equivalent to, in terms of information. This prior yields the following posterior distribution

$$p_{\boldsymbol{\mu},\boldsymbol{\tau}|\boldsymbol{y}_{1:n}}(\mu, \tau \,|\, y_1, \dots, y_n) \propto p_{\boldsymbol{y}_{1:n}|\boldsymbol{\mu},\boldsymbol{\tau}}(y_1, \dots, y_n \,|\, \mu, \tau) \text{N}\big(\mu; \mu_0, (k\tau)^{-1}\big) \text{Ga}(\tau; \alpha, \beta).$$

Considering the first term in the likelihood in (1.15c) and the conditional distribution of $\boldsymbol{\mu}$, we find that

$$\exp\Big(-\frac{n\tau}{2} (\mu - \bar{y}_n)^2 \Big) \text{N}\big(\mu; \mu_0, (k\tau)^{-1}\big) \propto \text{N}\big(\mu; \mu_n, ((k+n)\tau)^{-1}\big) \exp\Big(-\frac{nk\tau}{2(n+k)} (\mu_0 - \bar{y}_n)^2 \Big)$$

so that the posterior distribution of $\boldsymbol{\mu}$ given $\boldsymbol{\tau} = \tau$ is also normal with mean

$$\mu_n = \frac{k\mu_0 + n\bar{y}_n}{k+n}$$

and variance $((k+n)\tau)^{-1}$. Now considering the second term in the likelihood in (1.15c), the remaining term in the above equation and the prior on $\boldsymbol{\tau}$, we find that

$$\tau^{n/2} \exp\Big(-\frac{n\hat{v}}{2}\tau\Big) \exp\Big(-\frac{nk\tau}{2(n+k)}(\mu_0 - \bar{y}_n)^2\Big) \mathrm{Ga}(\tau;\alpha,\beta)$$
$$\propto \tau^{\alpha+n/2-1} \exp\bigg(-\tau\Big(\beta + \frac{n\hat{v}}{2} + \frac{nk}{2(n+k)}(\mu_0 - \bar{y}_n)^2\Big)\bigg)$$

so that the posterior distribution of $\boldsymbol{\tau}$ is also gamma, with parameters

$$\alpha_n = \alpha + \frac{n}{2} \qquad \text{and} \qquad \beta_n = \beta + \frac{n\hat{v}}{2} + \frac{nk}{2(n+k)}(\mu_0 - \bar{y}_n)^2.$$

We can deduce from these results that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\mu} \,|\, \boldsymbol{y}_{1:n} = y_{1:n}) &= \mathbb{E}\big(\mathbb{E}(\boldsymbol{\mu} \,|\, \boldsymbol{\tau}, \boldsymbol{y}_{1:n} = y_{1:n}) \,|\, \boldsymbol{y}_{1:n} = y_{1:n}\big) \\ &= \mu_n\end{aligned}$$

and that

$$\begin{aligned}\mathrm{var}(\boldsymbol{\mu} \,|\, \boldsymbol{y}_{1:n} = y_{1:n}) &= \mathbb{E}\big(\,\mathrm{var}(\boldsymbol{\mu} \,|\, \boldsymbol{\tau}, \boldsymbol{y}_{1:n} = y_{1:n}) \,|\, \boldsymbol{y}_{1:n} = y_{1:n}\big) \\ &= \mathbb{E}\big(((k+n)\boldsymbol{\tau})^{-1} \,|\, \boldsymbol{y}_{1:n} = y_{1:n}\big) \\ &= \frac{\beta_n}{(k+n)(\alpha_n - 1)}.\end{aligned}$$

It follows that $\mathbb{E}(\boldsymbol{\mu} \,|\, \boldsymbol{y}_{1:n} = y_{1:n})$ tends to $\bar{y}_n$ as n tends to infinity and the posterior variance $\mathrm{var}(\boldsymbol{\mu} \,|\, \boldsymbol{y}_{1:n} = y_{1:n})$ tends to 0. Although we have computed the posterior mean and variance of the random variable $\boldsymbol{\mu}$, this does not imply that its distribution is normal. We can find the form of this distribution by computing the following integral directly:

$$p_{\boldsymbol{\mu}|\boldsymbol{y}_{1:n}}(\mu \,|\, y_1, \ldots, y_n) \propto \int_0^\infty \tau^{\alpha_n + \frac{1}{2} - 1} \exp\bigg(-\tau\Big(\beta_n + \frac{k+n}{2}(\mu - \mu_n)^2\Big)\bigg)\, \mathrm{d}\tau.$$

This expression can be recognised as yet another beta distribution with some new parameter $\alpha' > 0$ and $\beta' > 0$ so we know it integrates to $\Gamma(\alpha')\beta'^{-\alpha'}$. Since only β' depends on μ, it follows that

$$\begin{aligned}p_{\boldsymbol{\mu}|\boldsymbol{y}_{1:n}}(\mu \,|\, y_1, \ldots, y_n) &\propto \Big(\beta_n + \frac{k+n}{2}(\mu - \mu_n)^2\Big)^{-\alpha_n - \frac{1}{2}} \\ &\propto \Big(1 + \frac{1}{2\alpha_n}\frac{(k+n)\alpha_n}{\beta_n}(\mu - \mu_n)^2\Big)^{-\frac{2\alpha_n+1}{2}},\end{aligned}$$

which is a generalised Student's t distribution with location parameter μ_n, scale parameter $\beta_n/((k+n)\alpha_n)$ and $2\alpha_n$ degrees of freedom.