Generalized Bayesian Inference with Sets of Conjugate Priors for Dealing with Prior-Data Conflict

Gero Walter

Lund University, 15.12.2015

This document is a step-by-step guide on how sets of priors can be used to better reflect prior-data conflict in the posterior. First we explain what conjugate priors are along an example. Then we show how conjugate priors can be constructed using a general result, and why they usually do not reflect prior-data conflict. In the last part, we see how to use sets of conjugate priors to deal with this problem.

1 Bayesian basics

Bayesian inference allows to combine information from data and information extraneous to data (e.g., expert information) into a 'complete picture'. Data $\boldsymbol{x}$ is assumed to be generated from a certain parametric distribution family, and information about unknown parameters is then expressed by a so-called prior distribution, a distribution over the parameter(s) of the data generating function.

As running example, let us consider an experiment with two possible outcomes, success and failure. The number of successes s in a series of n independent trials has then a Binomial distribution with parameters p and n, where n is known but $p \in [0, 1]$ is unknown. In short, $S \mid p \sim \text{Binomial}(n, p)$, which means

$$f(s \mid p) = P(S = s \mid p) = \binom{n}{s} p^s (1-p)^{n-s}, \quad s \in \{0, 1, \ldots, n\}\,. \tag{1}$$

Information about unknown parameters (here, p) is then expressed by a so-called prior distribution, some distribution with some pdf, here $f(p)$.

The 'complete picture' is then the so-called posterior distribution, here with pdf $f(p \mid s)$, expressing the state of knowledge after having seen the data. It encompasses information from the prior $f(p)$ and data and is obtained via Bayes' Rule,

$$f(p \mid s) = \frac{f(s \mid p)f(p)}{\int f(s \mid p)f(p)\,\mathrm{d}p} = \frac{f(s \mid p)f(p)}{f(s)} \propto f(s \mid p)f(p)\,, \tag{2}$$

where $f(s)$ is the so-called marginal distribution of the data S.

In general, the posterior distribution is hard to obtain, especially due to the integral in the denominator. The posterior can be approximated with numerical methods, like the Laplace approximation or simulation methods like MCMC (Markov chain Monte Carlo). There is a large literature dealing with computations of posteriors, and software like BUGS or JAGS has been developed which simplifies the creation of a sampler to approximate a posterior.

2 A conjugate prior

However, Bayesian inference not necessarily entails complex calculations and simulation methods. With a clever choice of parametric family for the prior distribution, the posterior distribution belongs to the same parametric family as the prior, just with updated parameters. Such prior distributions are called *conjugate* priors. Basically, with conjugate priors one trades flexibility for tractability: The parametric family restricts the form of the prior pdf, but with the advantage of much easier computations.[1]

The conjugate prior for the Binomial distribution is the Beta distribution, which is usually parametrised with parameters α and β.

$$f(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, p^{\alpha-1}\,(1-p)^{\beta-1}\,, \tag{3}$$

where $B(\cdot, \cdot)$ is the Beta function.[2] In short, we write $p \sim \mathrm{Beta}(\alpha, \beta)$.

From now on, we will denote prior parameter values by an upper index $^{(0)}$, and updated, posterior parameter values by an upper index $^{(n)}$. With this notational convention, let $S \mid p \sim \mathrm{Binomial}(n, p)$ and $p \sim \mathrm{Beta}(\alpha^{(0)}, \beta^{(0)})$.

[1] In fact, practical Bayesian inference was mostly restricted to conjugate priors before the advent of MCMC.

[2] The Beta function is defined as $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ and gives the inverse normalisation constant for the Beta distribution. It is related to the Gamma function through $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. We will not need to work with Beta functions here.

Then it holds that $p \mid s \sim \text{Beta}(\alpha^{(n)}, \beta^{(n)})$, where $\alpha^{(n)}$ and $\beta^{(n)}$ are updated, posterior parameters, obtained as

$$\alpha^{(n)} = \alpha^{(0)} + s\,, \qquad \beta^{(n)} = \beta^{(0)} + n - s\,. \tag{4}$$

From this we can see that $\alpha^{(0)}$ and $\beta^{(0)}$ can be interpreted as pseudocounts, forming a hypothetical sample with $\alpha^{(0)}$ sucesses and $\beta^{(0)}$ failures.

Exercise 1. *Confirm Eq.* (4)*, i.e., show that, when* $S \mid p \sim \text{Binomial}(n, p)$ *and* $p \sim \text{Beta}(\alpha^{(0)}, \beta^{(0)})$*, the density of the posterior distribution for* p *is of the form Eq.* (3) *but with updated parameters. (Hint: use the last expression in Eq.* (2) *and consider for the posterior the terms related to* p *only.)*

You have seen in the talk that we considered a different parametrisation of the Beta distribution in terms of $n^{(0)}$ and $y^{(0)}$, defined as

$$n^{(0)} = \alpha^{(0)} + \beta^{(0)}\,, \qquad y^{(0)} = \frac{\alpha^{(0)}}{\alpha^{(0)} + \beta^{(0)}}\,, \tag{5}$$

such that writing $p \sim \text{Beta}(n^{(0)}, y^{(0)})$ corresponds to

$$f(p \mid n^{(0)}, y^{(0)}) = \frac{p^{n^{(0)}y^{(0)}-1}\,(1-p)^{n^{(0)}(1-y^{(0)})-1}}{B(n^{(0)}y^{(0)}, n^{(0)}(1-y^{(0)}))}\,. \tag{6}$$

In this parametrisation, the updated, posterior parameters are given by

$$n^{(n)} = n^{(0)} + n\,, \qquad y^{(n)} = \frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{s}{n}\,, \tag{7}$$

and we write $p \mid s \sim \text{Beta}(n^{(n)}, y^{(n)})$.

Exercise 2. *Confirm the equations for updating* $n^{(0)}$ *to* $n^{(n)}$ *and* $y^{(0)}$ *to* $y^{(n)}$*. (Hint: Find expressions for* $\alpha^{(0)}$ *and* $\beta^{(0)}$ *in terms of* $n^{(0)}$ *and* $y^{(0)}$*, then use Eq.* (4) *and solve for* $n^{(n)}$ *and* $y^{(n)}$*.)*

From the properties of the Beta distribution, it follows that $y^{(0)} = \frac{\alpha^{(0)}}{\alpha^{(0)}+\beta^{(0)}} = \text{E}[p]$ is the prior expectation for the success probability p, and that the higher $n^{(0)}$, the more probability weight will be concentrated around $y^{(0)}$, as $\text{Var}(p) = \frac{y^{(0)}(1-y^{(0)})}{n^{(0)}+1}$. From the interpretation of α and β and Eq. (5), we see that $n^{(0)}$ can also be interpreted as a (total) pseudocount or prior strength.

Exercise 3. *Write a function* `dbetany(x,n,y, ...)` *that returns the value of the Beta density function at* `x` *for parameters* $n^{(0)}$ *and* $y^{(0)}$ *instead of* `shape1` *(=* α*) and* `shape2` *(=* β*) as in* `dbeta(x, shape1, shape2, ...)`*. Use your function to plot the Beta pdf for different values of* $n^{(0)}$ *and* $y^{(0)}$ *to see how the pdf changes according to the parameter values.*

The formula for $y^{(n)}$ in Eq. (7) is not written in the most compact form in order to emphasize that $y^{(n)}$, the posterior expectation of p, is a weighted average of the prior expectation $y^{(0)}$ and s/n (the fraction of successes in the data), with the weights $n^{(0)}$ and n, respectively. We see that $n^{(0)}$ plays the same role for the prior mean $y^{(0)}$ as the sample size n for the observed mean s/n, reinforcing the interpretation as pseudocount. Indeed, the higher $n^{(0)}$, the higher the weight for $y^{(0)}$ in the weighted average calculation of $y^{(n)}$, so $n^{(0)}$ gives the strength of the prior as compared to the sample size n.

Exercise 4. *Give a ceteris paribus analysis for* $\mathrm{E}[p \mid s] = y^{(n)}$ *and* $\mathrm{Var}(p \mid s) = \frac{y^{(n)}(1-y^{(n)})}{n^{(n)}+1}$ *(i.e, discuss how* $\mathrm{E}[p \mid x]$ *and* $\mathrm{Var}(p \mid s)$ *behave) when*

(i) $n^{(0)} \to 0$*,*

(ii) $n^{(0)} \to \infty$*, and*

(iii) $n \to \infty$ *when* $s/n = const.$

and consider also the form of $f(p \mid s)$ *based on* $\mathrm{E}[p \mid s]$ *and* $\mathrm{Var}(p \mid s)$*.*

3 Conjugate priors for canonical exponential families

Fortunately it is not necessary to search or guess to find a conjugate prior to a certain data distribution, as there is a general result on how to construct conjugate priors when the sample distribution belongs to a so-called *canonical exponential family* (e.g., Bernardo and Smith 2000, pp. 202 and 272f). This result covers many sample distributions, like Normal and Multinomial models, Poisson models, or Exponential and Gamma models, and gives a common structure to all conjugate priors constructed in this way.

For the construction, we will consider distributions of i.i.d. samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ of size n directly.[3] With the Binomial distribution, we did so indirectly only: The Binomial(n, p) distribution for S results from n independent trials with success probability p each. Encoding success as $x_i = 1$ and failure as $x_i = 0$ and collecting the n results in a vector $\boldsymbol{x}$, we get $s = \sum_{i=1}^{n} x_i$. It turns out that the sample distribution depends on $\boldsymbol{x}$ only

[3]It would be possible, and indeed is often done in the literature, to consider a single observation x in Eq. (9) only, as the conjugacy property does not depend on the sample size. However, we find our version with n-dimensional i.i.d. sample $\boldsymbol{x}$ more appropriate.

through s:

$$f(\boldsymbol{x} \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)} = p^s(1-p)^{n-s}\,, \tag{8}$$

so s summarizes the sample $\boldsymbol{x}$ without changing the pmf. Such a summary function of a sample $\boldsymbol{x}$ is called *a sufficient statistic of* $\boldsymbol{x}$.[4]

The construction has two steps. We first rewrite the sample distribution in a specific form to identify certain ingredients, and then construct the conjugate prior based on these ingredients. We will cover sampling distributions with only one parameter here; you can find the formulation for exponential family distributions with more than one parameter in Appendix B.

For step 1, a sample distribution is said to belong to the *canonical exponential family* if its density or mass function satisfies the decomposition

$$f(\boldsymbol{x} \mid \theta) = a(\boldsymbol{x}) \exp\left\{\psi \cdot \tau(\boldsymbol{x}) - n\mathbf{b}(\psi)\right\}. \tag{9}$$

The ingredients of this decomposition are:

- $\psi \in \Psi \subset \mathbb{R}$, a transformation of the distribution parameter $\theta \in \Theta$, called the *natural parameter* of the canonical exponential family;
- $\tau(\boldsymbol{x})$, a sufficient statistic of the sample $\boldsymbol{x}$. It holds that $\tau(\boldsymbol{x}) = \sum_{i=1}^{n} \tau^*(x_i)$, where $\tau^*(x_i) \in \mathcal{T} \subset \mathbb{R}$;
- $\mathbf{b}(\psi)$, some function of ψ (or, in turn, of θ);
- $a(\boldsymbol{x})$, some function of $\boldsymbol{x}$.

Let us do the decomposition for the Binomial distribution before we go to the second step. The Binomial pmf from Eq. (1) can be rewritten as follows:

$$f(s \mid p) = \binom{n}{s} p^s(1-p)^{n-s} \tag{10}$$
$$= \binom{n}{s} \exp\left\{ \log\Big(\frac{p}{1-p}\Big) s - n\big(-\log(1-p)\big)\right\}. \tag{11}$$

We have thus $\psi = \log(p/(1-p))$, $\tau(\boldsymbol{x}) = s$, $\mathbf{b}(\psi) = -\log(1-p)$, and $a(\boldsymbol{x}) = \binom{n}{s}$. The function $\log(p/(1-p))$ is known as the *logit*, denoted by $\text{logit}(p)$.

[4]There are $\binom{n}{s}$ 0/1 vectors $\boldsymbol{x}$ with s 1's, leading to the Binomial pmf Eq. (1). For a Bayesian analysis, such factors that do not depend on the parameter of interest do not matter. This is one of the central differences between Bayesian and Frequentist methods.

In step 2, a conjugate prior on ψ can be constructed from the ingredients identified in step 1 by

$$p(\psi \mid n^{(0)}, y^{(0)})\,\mathrm{d}\psi \propto \exp\left\{n^{(0)}\left[y^{(0)} \cdot \psi - \mathbf{b}(\psi)\right]\right\}\mathrm{d}\psi\,, \tag{12}$$

where $n^{(0)} > 0$ and $y^{(0)} \in \mathbb{R}$ are the parameters by which a certain prior can be specified.[5] We will refer to priors of the form Eq. (12) as *canonically constructed priors*. Note that Eq. (12) provides a distribution over the natural parameter ψ and not over the usual parameter θ. When $\psi \neq \theta$ it can be useful to transform the density over ψ to a density over θ.

Continuing our example, it turn out that the Beta$(n^{(0)}, y^{(0)})$ is the canonically constructed prior to the Binomial distribution. Constructing the prior from the ingredients $\psi = \log(p/(1-p))$, $\tau(\boldsymbol{x}) = s$, and $\mathbf{b}(\psi) = -\log(1-p)$ leads to

$$f\big(\psi \mid n^{(0)}, y^{(0)}\big)\,\mathrm{d}\psi \propto \exp\left\{n^{(0)}\left[y^{(0)}\log\left(\frac{p}{1-p}\right) + \log(1-p)\right]\right\}\mathrm{d}\psi\,. \tag{13}$$

To transform this density over ψ to a density over p, we have to multiply it with

$$\left|\frac{\mathrm{d}\psi}{\mathrm{d}p}\right| = \left|\frac{\mathrm{d}}{\mathrm{d}p}\log\left(\frac{p}{1-p}\right)\right| = \left|\frac{1-p}{p}\left(\frac{(1-p)+p}{(1-p)^2}\right)\right| = \frac{1}{p(1-p)}\,, \tag{14}$$

and so we get

$$\begin{aligned}
&f(p \mid n^{(0)}, y^{(0)})\,\mathrm{d}p && (15)\\
&\quad = f(\psi \mid n^{(0)}, y^{(0)})\left|\frac{\mathrm{d}\psi}{\mathrm{d}p}\right|\mathrm{d}p && (16)\\
&\quad \propto \exp\left\{n^{(0)}y^{(0)}\log(p) + \big(n^{(0)} - n^{(0)}y^{(0)}\log(1-p)\big)\right\}\frac{1}{p(1-p)}\,\mathrm{d}p && (17)\\
&\quad = p^{n^{(0)}y^{(0)}-1}(1-p)^{n^{(0)}(1-y^{(0)})-1}\,\mathrm{d}p\,. && (18)
\end{aligned}$$

This is indeed the Beta distribution from Eq. (6).

For all canonically constructed priors, the prior parameters $n^{(0)}$ and $y^{(0)}$ are updated to their posterior values $n^{(n)}$ and $y^{(n)}$ by

$$n^{(n)} = n^{(0)} + n\,, \qquad y^{(n)} = \frac{n^{(0)}}{n^{(0)}+n}\cdot y^{(0)} + \frac{n}{n^{(0)}+n}\cdot\frac{\tau(\boldsymbol{x})}{n}\,, \tag{19}$$

[5] Actually, the domain of $y^{(0)}$ is $\mathcal{Y}$, defined as the interior of the convex hull of $\mathcal{T}$; these intricacies do not matter in this exercise however.

and the posterior can be written as

$$p(\psi \mid \boldsymbol{x}, n^{(0)}, y^{(0)}) = p(\psi \mid n^{(n)}, y^{(n)}) \propto \exp\left\{ n^{(n)} \left[\langle y^{(n)}, \psi \rangle - \mathbf{b}(\psi) \right] \right\} \mathrm{d}\psi \,. \tag{20}$$

Usually, $y^{(0)}$ and $y^{(n)}$ can be seen as the parameter describing the main characteristics of the prior and the posterior, and thus we will call them *main prior* and *main posterior parameter*, respectively.[6] $y^{(0)}$ can also be understood as a prior guess for the mean sufficient statistic $\tilde{\tau}(\boldsymbol{x}) := \tau(\boldsymbol{x})/n$. For all constructed priors, $y^{(n)}$ is a weighted average of this prior guess $y^{(0)}$ and the sample 'mean' $\tilde{\tau}(\boldsymbol{x})$, with weights $n^{(0)}$ and n, respectively; therefore, $n^{(0)}$ can be seen as "prior strength" or "pseudocount", reflecting the weight one gives to the prior as compared to the sample size n.

Exercise 5. *Confirm Eq.* (19)*, the equations for updating $n^{(0)}$ to $n^{(n)}$ and $y^{(0)}$ to $y^{(n)}$, for one-parametric exponential family distributions. (Hint: Use the last expression in Eq.* (2) *and consider only the terms related to ψ for the posterior.)*

Exercise 6. *Construct the canonical conjugate prior to a sample distribution of your choice. This works only for distributions forming a canonical exponential family! (As a counterexample, the Weibull distribution with unknown shape parameter does not form a canonical exponential family.) You can try, e.g., the Normal (Gaussian) distribution with fixed variance σ_0^2 or, if you want to avoid a density transformation, the Normal distribution with fixed variance* 1.[7]

4 Prior-data conflict

As discussed in the talk, the weighted average structure for $y^{(n)}$ in Eq. (19) is intuitive, but with it comes the problematic behaviour in case of prior-data conflict. In most parametric models, the spread of the posterior does not systematically depend on how far the prior guess $y^{(0)}$ diverges from the mean sufficient statistic $\tilde{\tau}(\boldsymbol{x})$. Then, conflict between prior assumptions and information from data is just averaged out, and the posterior gives a false impression of certainty, being more pointed around $y^{(n)}$ than the prior in spite of the conflict.

[6] Remember, for the Beta distribution, $y^{(0)}$ is the expected success probability p.

[7] You can find the solution for $\boldsymbol{X} \overset{\text{i.i.d}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$ in Appendix A. Furthermore, Table 1 in Quaeghebeur and Cooman (2005) gives ψ, $\tau^*(x_i)$, $a(x)$ and $\mathbf{b}(\psi)$ for the most common sample distributions that form a canonical exponential family.

Exercise 7. *Write the functions* `nn(n0, n)` *and* `yn(n0, y0, s, n)` *implementing Eq.* (7)*, the update step for the Beta prior. Plot prior and posterior densities for different choices of* `n0, y0, s, n` *to see the effect (or the lack of it) of prior-data conflict on the posterior, using your code from Exercise 3. E.g, take* `n0 = 8`*,* `y0 = 0.75` *to fix a prior, and compare the posterior for* `s = 12, n = 16` *and for* `s = 0, n = 16`*.*

5 Sets of conjugate priors

Modelling prior information with sets of conjugate priors retains the tractability of conjugate analysis and ensures prior-data conflict sensitivity. It also allows to express prior knowledge more cautiously, or partial prior knowledge, and generally makes it possible to express the precision of probability statements encoded in the prior. The resulting imprecise/interval probability models can be seen as systematic sensitivity analysis, or as a kind of robust Bayesian method.

The central idea is to consider sets $\Pi^{(0)}$ of canonical parameters $(n^{(0)}, y^{(0)})$ which create corresponding sets of priors $\mathcal{M}^{(0)}$ via

$$\mathcal{M}^{(0)} = \left\{ f(\psi \mid n^{(0)}, y^{(0)}) : (n^{(0)}, y^{(0)}) \in \Pi^{(0)} \right\}. \tag{21}$$

Quaeghebeur and Cooman (2005) suggested sets $\Pi^{(0)} = n^{(0)} \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, but Walter and Augustin (2009) showed that the resulting sets of priors are still insensitive to prior-data conflict, and proposed instead to use sets $\Pi^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$.[8] Sets of priors generated by $\Pi^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$ were called 'generalized iLUCK-models', and were implemented in an **R** package `luck` (Walter and Krautenbacher 2013).

Exercise 8. *You can install the* `luck` *package by downloading* `http://download.r-forge.r-project.org/src/contrib/luck_0.9.tar.gz` *to your working directory and then executing*
`install.packages("luck_0.9.tar.gz", repos = NULL, type = "source")`
at the ***R*** *prompt. If there is a problem, try installing the package* `TeachingDemos` *first. After installation, load the package by executing* `library(luck)`*.*

(i) Use the functions `LuckModel()` *and* `LuckModelData()` *to create a* `LuckModel` *object corresponding to a parameter set* $\Pi^{(0)}$ *with an interval for both* $n^{(0)}$ *and* $y^{(0)}$*, and that contains data such that* $\tau(\boldsymbol{x})/n \in [\underline{y}^{(0)}, \overline{y}^{(0)}]$*. Create a second* `LuckModel` *object with the same* $\Pi^{(0)}$ *but for which* $\tau(\boldsymbol{x})/n \notin [\underline{y}^{(0)}, \overline{y}^{(0)}]$*.*

[8] A detailed discussion of different types of $\Pi^{(0)}$ is given in Walter (2013, §3.1).

(ii) Plot the prior and posterior parameter sets $\Pi^{(0)}$ and $\Pi^{(n)}$ for both objects. You can find the help for the plot function for `LuckModel` *objects via* `?luck::plot`*. To plot $\Pi^{(n)}$, you need to supply the option* `control=controlList(posterior=TRUE)`*; to plot a second parameter set in the same plot window, use* `add=TRUE`*. (You may need to set* `xlim` *and* `ylim` *to make the plotting region large enough!)*

(iii) Can you explain why the two $\Pi^{(n)}$'s have different shapes, and how their respective shapes come about? (Hint: Each point in the upper bound of $\Pi^{(n)}$ is a weighted average of $\overline{y}^{(0)}$ and $\tau(\boldsymbol{x})/n$.)

(iv) Vary the length of the $y^{(0)}$ interval and the $n^{(0)}$ interval, vary the sample statistic $\tau(\boldsymbol{x})$ and the sample size n. What is the effect on $\Pi^{(n)}$ for each change? E.g., what happens to the range of $y^{(n)}$ values when the $y^{(0)}$ interval gets larger? What happens to $\Pi^{(0)}$ when $n \to \infty$ with $\tilde{\tau}(\boldsymbol{x}) = \tau(\boldsymbol{x})/n$ constant?

A model with $\Pi^{(0)} = n^{(0)} \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$ corresponds to a vertical slice of the plotted sets $\Pi^{(0)}$. The posterior parameter set is a vertical slice as well and can be expressed as $\Pi^{(n)} = n^{(n)} \times [\underline{y}^{(n)}, \overline{y}^{(n)}]$, where $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ result from updating $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$, respectively:

$$\underline{y}^{(n)} = \frac{n^{(0)}\underline{y}^{(0)} + \tau(\boldsymbol{x})}{n^{(0)} + n}\,, \qquad \overline{y}^{(n)} = \frac{n^{(0)}\overline{y}^{(0)} + \tau(\boldsymbol{x})}{n^{(0)} + n}\,. \tag{22}$$

The posterior imprecision in the y dimension, denoted by $\Delta_y(\Pi^{(n)})$, is

$$\Delta_y(\Pi^{(n)}) = \overline{y}^{(n)} - \underline{y}^{(n)} = \frac{n^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{n^{(0)} + n}\,. \tag{23}$$

Exercise 9. *Do you see from Eq.* (23) *why models with $\Pi^{(0)} = n^{(0)} \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$ are insensitive to prior-data conflict?*

When $\Pi^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, things are different, as you have seen. Then, the lower and upper bound in the y dimension are given by

$$\underline{y}^{(n)} = \inf_{\Pi^{(n)}} y^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}}{\overline{n}^{(0)} + n}\,\underline{y}^{(0)} + \dfrac{n}{\overline{n}^{(0)} + n}\,\tilde{\tau}(\boldsymbol{x}) & \tilde{\tau}(\boldsymbol{x}) \geq \underline{y}^{(0)} \\ \dfrac{\underline{n}^{(0)}}{\underline{n}^{(0)} + n}\,\underline{y}^{(0)} + \dfrac{n}{\underline{n}^{(0)} + n}\,\tilde{\tau}(\boldsymbol{x}) & \tilde{\tau}(\boldsymbol{x}) < \underline{y}^{(0)} \end{cases}, \tag{24}$$

$$\overline{y}^{(n)} = \sup_{\Pi^{(n)}} y^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}}{\overline{n}^{(0)} + n}\,\overline{y}^{(0)} + \dfrac{n}{\overline{n}^{(0)} + n}\,\tilde{\tau}(\boldsymbol{x}) & \tilde{\tau}(\boldsymbol{x}) \leq \overline{y}^{(0)} \\ \dfrac{\underline{n}^{(0)}}{\underline{n}^{(0)} + n}\,\overline{y}^{(0)} + \dfrac{n}{\underline{n}^{(0)} + n}\,\tilde{\tau}(\boldsymbol{x}) & \tilde{\tau}(\boldsymbol{x}) > \overline{y}^{(0)} \end{cases}. \tag{25}$$

Exercise 10. *Which cases in Eq.* (24) *and Eq.* (25) *correspond to prior-data conflict? What do they have in common, and how does this link to what you saw in the* $\Pi^{(n)}$ *plots?*

The posterior imprecision in the y dimension can be expressed by

$$\Delta_y(\Pi^{(n)}) = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)} + n} + \inf_{y^{(0)} \in [\underline{y}^{(0)}, \overline{y}^{(0)}]} |\tilde{\tau}(\boldsymbol{x}) - y^{(0)}| \frac{n(\overline{n}^{(0)} - \underline{n}^{(0)})}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)} \,. \tag{26}$$

Note that the expression $\inf_{y^{(0)} \in [\underline{y}^{(0)}, \overline{y}^{(0)}]} |\tilde{\tau}(\boldsymbol{x}) - y^{(0)}| = 0$ when $\tilde{\tau}(\boldsymbol{x}) \in [\underline{y}^{(0)}, \overline{y}^{(0)}]$, otherwise it gives the distance of $\tilde{\tau}(\boldsymbol{x})$ to the $y^{(0)}$ interval.

Exercise 11. *How does the shape of* $\Pi^{(n)}$ *reflect Eq.* (26) *when* $\tilde{\tau}(\boldsymbol{x}) \in [\underline{y}^{(0)}, \overline{y}^{(0)}]$*?*

6 Sets of conjugate priors for scaled normal data

The prior for the *scaled normal distribution*, a normal distribution with variance 1, is implemented in the `luck` package. For $X \sim \mathrm{N}(\mu, 1)$, the canonically constructed prior is $\mu \sim \mathrm{N}(y^{(0)}, 1/n^{(0)})$; the sufficient statistic is $\tau(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. For a derivation, see Appendix A below and set $\sigma_0^2 = 1$.

Exercise 12. *Use the functions* `ScaledNormalLuckModel()` *and* `ScaledNormalData()` *to create a* `LuckModel` *for scaled normal data. Plot the set of prior and posterior cdfs using* `cdfplot()`*, and observe how this changes depending on* $\tilde{\tau}(\boldsymbol{x}) = \bar{x}$ *being inside our outside the* $y^{(0)}$ *interval. How are the ranges for* $n^{(0)}$ *and* $y^{(0)}$ *reflected in the set of cdfs?*

The Bayesian equivalent to frequentist confidence intervals are credible intervals. A 95% posterior credible interval is an interval for θ covering a probability weight of $\gamma = 95\%$ according to the posterior over θ. It can be obtained, e.g., as the interval from the 2.5% quantile to the 97.5% quantile of the posterior. Highest density intervals are credible intervals that consist of the θ values with the highest cdf values. Often denoted as HPD (for highest posterior density) intervals, they are more difficult to obtain than quantile-based credible intervals, but give the shortest interval among all level γ intervals when the distribution is unimodal. For sets of priors, we can consider the union of all highest density intervals corresponding to all priors in the set, and likewise for the set of posteriors.

Exercise 13. *Calculate prior and posterior union of highest density intervals using* `unionHdi()` *for a* `ScaledNormalLuckModel`*. Compare its length when prior-data conflict is or is not present.*

Experienced **R** programmers can work to extend the `luck` package (I'm very happy to help):

Exercise 14. *Write your own subclasses to implement the conjugate prior to a sample distribution of your choice. (You may have derived a conjugate prior in Exercise 6 already.) Take the code files* `01-01_ScaledNormalData.r` *and* `01-02_ScaledNormal.r` *as a blueprint. You can find these files in the* `R` *folder of the package sources. The constructor functions can be much simpler than those in the files which are written to accommodate all kinds of inputs.*

A The Normal distribution with known variance as canonically constructed conjugate prior

Consider the normal or Gaussian distribution with known variance σ_0^2. The pdf for n independent samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ can be written as

$$\begin{aligned} f(\boldsymbol{x} \mid \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\Big\{ -\frac{1}{2\sigma_0^2}(x_i - \mu)^2 \Big\} \\ &= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\} \qquad (27) \\ &= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma_0^2} \Big[\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \Big] \Big\} \\ &= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma_0^2} \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma_0^2} \Big\}. \qquad (28) \end{aligned}$$

So we have here $\psi = \frac{\mu}{\sigma_0^2}$, $\mathbf{b}(\psi) = \frac{\mu^2}{2\sigma_0^2}$, and $\tau(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. From these ingredients, a conjugate prior can be constructed with (12), leading to

$$p\Big(\frac{\mu}{\sigma_0^2} \mid n^{(0)}, y^{(0)}\Big)\, \mathrm{d}\frac{\mu}{\sigma_0^2} \propto \exp\Big\{ n^{(0)} \Big(y^{(0)} \frac{\mu}{\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2} \Big) \Big\}\, \mathrm{d}\frac{\mu}{\sigma_0^2}\,. \qquad (29)$$

This prior, transformed to the parameter of interest μ and with the square completed,

$$p(\mu \mid n^{(0)}, y^{(0)})\, \mathrm{d}\mu \propto \frac{1}{\sigma_0^2} \exp\Big\{ -\frac{n^{(0)}}{2\sigma_0^2} \big(-2\mu y^{(0)} + \mu^2 \big) \Big\}\, \mathrm{d}\mu$$

$$\propto \exp\Big\{ -\frac{n^{(0)}}{2\sigma_0^2}(\mu - y^{(0)})^2 \Big\}\, \mathrm{d}\mu\,, \tag{30}$$

is a normal distribution with mean $y^{(0)}$ and variance $\frac{\sigma_0^2}{n^{(0)}}$, i.e. $\mu \sim \mathrm{N}(y^{(0)}, \frac{\sigma_0^2}{n^{(0)}})$. With (19), the parameters for the posterior distribution are

$$y^{(n)} = \mathrm{E}[\mu \mid n^{(n)}, y^{(n)}] = \frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \bar{x} \tag{31}$$

$$\frac{\sigma_0^2}{n^{(n)}} = \mathrm{Var}(\mu \mid n^{(n)}, y^{(n)}) = \frac{\sigma_0^2}{n^{(0)}+n}\,. \tag{32}$$

The posterior expectation of μ thus is a weighted average of the prior expectation $y^{(0)}$ and the sample mean $\bar{x}$, with weights $n^{(0)}$ and n, respectively. The effect of the update step on the variance is that it decreases by the factor $n^{(0)}/(n^{(0)}+n)$, for any sample of size n.

B Canonical exponential families with more than one parameter

A sample distribution is said to belong to the *q-parametric canonical exponential family* if its density or mass function satisfies the decomposition

$$f(\boldsymbol{x} \mid \theta) = a(\boldsymbol{x}) \exp\big\{ \langle \psi, \tau(\boldsymbol{x}) \rangle - n\mathbf{b}(\psi) \big\}\,. \tag{33}$$

The ingredients of the decomposition are:

- $\psi \in \Psi \subset \mathbb{R}^q$, a transformation of the (vectorial) parameter $\theta \in \Theta$, called the *natural parameter* of the canonical exponential family;
- $\mathbf{b}(\psi)$, a scalar function of ψ (or, in turn, of θ);
- $a(\boldsymbol{x})$, a scalar function of $\boldsymbol{x}$;
- $\tau(\boldsymbol{x})$, a sufficient statistic of the sample $\boldsymbol{x}$ which has dimension q (the same as ψ). It holds that $\tau(\boldsymbol{x}) = \sum_{i=1}^n \tau^*(x_i)$, where $\tau^*(x_i) \in \mathcal{T} \subset \mathbb{R}^q$.
- $\langle \cdot, \cdot \rangle$ denotes the scalar product, i.e., for $u, v \in \mathbb{R}^q$ is $\langle u, v \rangle = \sum_{j=1}^q u_j \cdot v_j$.

From these ingredients, a conjugate prior on ψ can be constructed as

$$p(\psi \mid n^{(0)}, y^{(0)})\, \mathrm{d}\psi \propto \exp\Big\{ n^{(0)} \Big[\langle y^{(0)}, \psi \rangle - \mathbf{b}(\psi) \Big] \Big\}\, \mathrm{d}\psi\,, \tag{34}$$

where $n^{(0)} > 0$ and $y^{(0)} \in \mathcal{Y}$ are the parameters by which a certain prior can be specified. $\mathcal{Y}$, the domain of $y^{(0)}$, is defined as the interior of the convex hull of $\mathcal{T}$. We refer to priors of the form in Eq. (34) as *canonically constructed priors*. Note that Eq. (34) provides a distribution over the natural parameter ψ and not over the usual parameter θ. When $\psi \neq \theta$ it can be useful to transform the density over ψ to a density over θ.

The prior parameters $n^{(0)}$ and $y_j^{(0)}$, $j = 1, \ldots, q$, are updated to their posterior values $n^{(n)}$ and $y_j^{(n)}$ by

$$n^{(n)} = n^{(0)} + n\,, \qquad y_j^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y_j^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{\tau(\boldsymbol{x})_j}{n}\,, \tag{35}$$

and the posterior can be written as

$$\begin{aligned} p(\psi \mid \boldsymbol{x}, n^{(0)}, y^{(0)}) &= p(\psi \mid n^{(n)}, y^{(n)}) \\ &\propto \exp\left\{ n^{(n)} \left[\langle y^{(n)}, \psi \rangle - \mathbf{b}(\psi) \right] \right\} \mathrm{d}\psi\,. \end{aligned} \tag{36}$$

$y^{(0)}$ and $y^{(n)}$ can be seen as the parameter vectors describing the main characteristics of the prior and the posterior, and thus we will call them *main prior* and *main posterior parameter*, respectively. $y^{(0)}$ can also be understood as a prior guess for the mean sufficient statistic $\tilde{\tau}(\boldsymbol{x}) := \tau(\boldsymbol{x})/n$.[9] $y_j^{(n)}$ is a weighted average of this prior guess $y_j^{(0)}$ and the sample 'mean' $\tilde{\tau}(\boldsymbol{x})_j$, with weights $n^{(0)}$ and n, respectively. $n^{(0)}$ can be seen as "prior strength" or "pseudocount", reflecting the weight one gives to the prior as compared to the sample size n.

References

Bernardo, J. and A. Smith (2000). *Bayesian Theory*. Chichester: Wiley.

Quaeghebeur, E. and G. de Cooman (2005). "Imprecise probability models for inference in exponential families". In: *ISIPTA '05. Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*. Ed. by F. Cozman, R. Nau, and T. Seidenfeld. Manno: SIPTA, pp. 287–296. URL: `http://leo.ugr.es/sipta/isipta05/proceedings/papers/s019.pdf`.

Walter, G. (2013). "Generalized Bayesian Inference under Prior-Data Conflict". PhD thesis. Department of Statistics, LMU Munich. URL: `http://edoc.ub.uni-muenchen.de/17059/`.

[9] This is because $\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi] = \nabla \mathbf{b}(\psi)$, where in turn $\mathrm{E}[\nabla \mathbf{b}(\psi) \mid n^{(0)}, y^{(0)}] = y^{(0)}$, see Bernardo and Smith (2000, Prop. 5.7, p. 275).

Walter, G. and N. Krautenbacher (2013). `luck`*: **R** package for Generalized iLUCK-models*. URL: `http://luck.r-forge.r-project.org/`.

Walter, G. and T. Augustin (2009). "Imprecision and Prior-data Conflict in Generalized Bayesian Inference". In: *Journal of Statistical Theory and Practice* 3, pp. 255–271. DOI: `10.1080/15598608.2009.10411924`.