Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2021

Handout 6: Conjugate and semi-conjugate priors

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim: Explain, theorize, and construct conjugate and conditional conjugate prior distribution.

References:

- Raiffa, H., & Schlaifer, R. (1961; Sections 3.1-3.3). Applied statistical decision theory.
- Berger, J. O. (2013; Sections 4.2.2). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Sections 3.1 & 3.3). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

Web applets:

- `https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/`

1 Conjugate priors

Note 1. We aim at specifying a prior distribution family, which (i.) leads to a tractable (to some extend) posterior distribution, (ii.) is rich enough to allow us to quantify prior believe, and (iii.) has a reasonable interpretation.

Definition 2. Let $\mathcal{F} = \{F(y|\theta);\, \forall\theta \in \Theta\}$ be a family of sampling distributions. A family of prior distributions $\mathcal{P}$ on Θ is said to be (natural) conjugate for $\mathcal{F}$ if the posterior $\Pi(\theta|y)$ belongs to $\mathcal{P}$ for all prior $\Pi(\theta) \in \mathcal{P}$ and all $F(y|\theta) \in \mathcal{F}$; i.e.

$$\Pi(\theta|y) \in \mathcal{P},\ \ \forall F(y|\theta) \in \mathcal{F} \text{ and } \Pi(\theta) \in \mathcal{P}.$$

Note 3. By specifying a tractable conjugate prior distribution $\Pi(\theta)$, we can achieve tractability for the posterior $\Pi(\theta|y)$ since it belongs to the same distribution family as the prior.

1.1 General derivation

We restrict here the derivation of conjugate priors in statistical models satisfying the following assumptions:

Assumption 4. *Let $y = (y_1, ..., y_n)$ be observables. We assume:*

1. *the likelihood is such that $L(\theta|y_{1:n}) = L(\theta|y_{1:k})L(\theta|y_{k+1:n})$ for any $k \in \{1, ..., n-1\}$, and*
2. *there exists a parameteric sufficient statistic $t : \mathcal{Y} \to \mathbb{R}^k$ with $t(y_1, ..., y_n) = t \in \mathbb{R}^k$ where its dimension k is independent on the number of observables n.*

Lemma 5. *Let $t^{(1)} = t(y_1, ..., y_q)$ and $t^{(2)} = t(y_{q+1}, ..., y_n)$ be sufficient statistics of two data-sets. Then, under the conditions of Assumption 4, there exists a binary operator $*$ such that*

$$y^{(1)} * y^{(2)} = y^* := (y^*_1, ..., y^*_k) \tag{1}$$

such that

$$f(y_1, ..., y_n|\theta) \propto k(y^*|\theta) \textit{ and } \quad k(y^*|\theta) \propto k(y^{(1)}|\theta)k(y^{(2)}|\theta)$$

Note 6. Let statistic $t : \mathcal{Y} \to \mathbb{R}^k$ with $t(y_1, ..., y_n) = t \in \mathbb{R}^k$ be parametric sufficient, and let its dimension k be independent from data size n. Then from Neyman factorization theorem, the likelihood can be factorized as

$$f(y|\theta) = k(t(y)|\theta)\rho(y) \ \propto k(t(y)|\theta), \tag{2}$$

where $\rho(y)$ is the residual term of a likelihood kernel $k(t(y)|\theta)$ of θ. By Bayes theorem, the posterior distribution is such that

$$\mathrm{d}\Pi(\theta|y) = \frac{f(\theta|y)\mathrm{d}\Pi(\theta)}{\int f(\theta|y)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} \tag{3}$$

I can construct a family of pdf/pmf $\tilde{\pi}(\theta|t) = \frac{1}{N(t)}k(t|\theta)$ such that the normalising constant $N(t)$ is finite for $t \in \mathcal{T}$, i.e.

$$N(t) = \begin{cases} \int k(t|\theta)\mathrm{d}\theta < \infty & \text{cont.} \\ \sum_{\forall\theta\in\Theta} k(t|\theta) < \infty & \text{discr.} \end{cases}, \text{ for } t \in \mathcal{T}$$

Assuming imaginary observables y' (prior to getting the experimental info from observable data y) or equivalently a sufficient statistic t' such as $t' = t(y') \in \mathbb{R}^k$, then I could possibly specify a prior $\Pi(\theta)$ such that

$$\mathrm{d}\Pi(\theta) = \frac{1}{N(\tau)}k(t' = \tau|\theta)\mathrm{d}\theta \quad \text{with} \quad \pi(\theta) := \tilde{\pi}(\theta|\tau) \propto k(\tau|\theta) \tag{4}$$

by assigning researcher specific fixed hyper-parameters $\tau = (\tau_0, ..., \tau_{k-1})$. Then from Lemma 5, the posterior (3) could get a form such that

$$\begin{aligned} \mathrm{d}\Pi(\theta|y) &= \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta}{\int k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta} = \frac{k(t(y) * \tau|\theta)\mathrm{d}\theta}{\int k(t(y) * \tau|\theta)\mathrm{d}\theta} = \frac{1}{N(t(y) * \tau)}k(t(y) * \tau|\theta)\mathrm{d}\theta \\ &= \tilde{\pi}(\theta|t(y) * \tau)\mathrm{d}\theta, \quad \text{with} \quad \pi(\theta|y) := \tilde{\pi}(\theta|t(y) * \tau) \propto k(t(y) * \tau|\theta) \end{aligned} \tag{5}$$

where $*$ is the binary operator (1) that combines the two kernels $k(t(y)|\theta)k(\tau|\theta) = k(t(y) * \tau|\theta)$.

Note 7. Essentially the prior $\Pi(\theta)$ in (4) and the posterior in $\Pi(\theta|y)$ (5) belong to the same distribution family $\mathcal{P}$ due to (5). The only difference is in the hyper-parameter values which change from prior to posterior as $\tau \mapsto t(y) * \tau$. In the posterior distribution the hyper-parameters combine both the prior info quantified in τ and the experimental info quantified in $t(y)$ according to the binary operator $*$.

Theorem 8. *Let $y = (y_1, ..., y_n)$ be observable quantities drawn from $F(y|\theta)$ independently conditional on θ, and let $f(y|\theta)$ be the likelihood with sufficient statistic $t := t(y)$ of a fixed dimension k independent from n. The conjugate prior $\Pi(\theta)$ with hyper-parameter τ of the likelihood $f(y|\theta)$ can be specified by setting its pdf/pmf as*

$$\pi(\theta) := \tilde{\pi}(\theta|\tau) = \frac{1}{N(\tau)}k(\tau|\theta) \ \propto \ k(\tau|\theta) \tag{6}$$

where $k(\cdot|\theta)$ is a kernel of the likelihood from the Neyman factorization

$$f(y|\theta) = k(t(y)|\theta)\rho(y) \ \propto k(t(y)|\theta),$$

and τ are hyper-parameters such that $N(\tau) = \int k(\tau|\theta)d\theta < \infty$.

Note 9. Essentially, in Theorem 8, and Note 6, we expect that since the likelihood kernel $k(\tau|\theta)$ is tractable, it may lead to a tractable conjugate prior, and hence to a tractable posterior distribution.

Note 10. Once the conjugate family of prior distributions $\Pi(\theta)$ has been specified, You can assign values on the hyper-parameters τ based on your a priori information. In fact, the values assigned on τ do not necessarily need to lie in the support of the sufficient statistics $\mathcal{T}$; the only restriction is that τ has to lead to a proper posterior $N(t(y) * \tau) < \infty$.

Example 11. Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i \overset{\text{iid}}{\sim} \mathrm{U}(0, \theta)$ for all $i = 1, ..., n$. Specify the conjugate prior for θ.

Pareto distribution: If $x \sim \mathrm{Pa}(a, b)$, then it has a pdf $f(x) = ab^a(\frac{1}{\theta})^{a+1}1(b < \theta)$

Solution. The likelihood $f(y|\theta)$ can be factorized as

$$f(y|\theta) = \prod_{i=1}^{n} \mathrm{U}(y_i|0, \theta) = \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} 1\left(y_i \in [0, \theta]\right) = \underbrace{\left(\frac{1}{\theta}\right)^n 1\left(\max_{\forall i=1:n}(y_i) < \theta\right)}_{=k(t(y)|\theta)} \underbrace{\cancelto{=1}{1\left(\min_{\forall i=1:n}(y_i) > 0\right)}}_{=\rho(y)}$$

with sufficient statistic $t = (n, \max_{\forall i=1:n}(y_i))$. Hence, I set

$$\pi(\theta) := \pi(\theta|\tau) \propto \left(\frac{1}{\theta}\right)^{\tau_0} 1\left(\tau_1 < \theta\right) = \left(\frac{1}{\theta}\right)^{\overbrace{\tau_0 - 1}^{=a}+1} 1\left(\overbrace{\tau_1}^{=b} < \theta\right) \propto \mathrm{Pa}(\theta|a, b)$$

for some $a >$ and $b > 0$ for the prior distribution to be proper (aka integrate to 1). By Bayes theorem the posterior is

$$\begin{aligned}
\pi(\theta|y) \propto \prod_{i=1}^{n} \mathrm{Un}(y_i|0, \theta)\mathrm{Pa}(\theta|a, b) &\propto \overbrace{\left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} [1(y_i \in [0, \theta])]}^{=\prod_{i=1}^{n} \mathrm{Un}(y_i|0,\theta)} \times \overbrace{\left(\frac{1}{\theta}\right)^{a+1} 1(\theta > b)}^{\propto \mathrm{Pa}(\theta|a,b)} \\
&\propto \left(\frac{1}{\theta}\right)^{n+a+} \underbrace{\prod_{i=1}^{n} 1(\theta > y_i)1(\theta > b)}_{=1(\theta > \max(b, y_{(n)}))} \propto \mathrm{Pa}(\theta|a^* = n + a, b^* = \max(b, y_{(n)})).
\end{aligned}$$

where $\theta > \max(b, y_{(n)})$. Notice that we do not care about τ_0, τ_1 as out prior now parameterized with $a > 0$ and $b > 0$.

Example 12. Consider the model of Normal linear regression where the observables are pairs (ϕ_i, y_i) for $i = 1, ..., n$, assumed to be modeled according to the sampling distribution $y_i|\beta, \sigma^2 \sim \mathrm{N}(\phi_i^\top \beta, \sigma^2)$ for $i = 1, ..., n$ with unknown $(\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Find the conjugate prior for (β, σ^2).

Hint: $(y - \Phi\beta)^\top(y - \Phi\beta) = (\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right] (\beta - \hat{\beta}_n) + (n - d)\hat{\sigma}_n^2$;

$$\hat{\beta}_n = (\Phi^\top \Phi)^{-1}\Phi^\top y\,; \qquad \hat{\sigma}_n^2 = \frac{(y - \Phi\hat{\beta}_n)^\top(y - \Phi\hat{\beta}_n)}{n - d} = ... = \frac{y^\top y - y^\top \Phi(\Phi^\top \Phi)^{-1}\Phi^\top y}{n - d}$$

Solution. The likelihood is

$$\begin{aligned}
f(y|\beta, \sigma^2) &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - \Phi\beta)^\top(y - \Phi\beta)\right) = \\
&\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right] (\beta - \hat{\beta}_n) - \frac{1}{2\sigma^2}(n - d)\hat{\sigma}_n^2\right)}_{=k(t(y)|\beta, \sigma^2)}
\end{aligned}$$

where Φ is the design matrix and the sufficient statistic is $t = (n, \Phi^\top y, \Phi^\top \Phi, y^\top y)$. Then, given prior hyper-parameters $\tau = (\tau_0, \tau_1, \tau_2, \tau_3)$, I set as a conjugate prior

$$\begin{aligned}\pi(\beta, \sigma^2) &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\tau_0}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \tau_1)^\top \tau_2 (\beta - \tau_1) - \frac{1}{\sigma^2}\tau_3\right) \\ &\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \tau_1)^\top \tau_2 (\beta - \tau_1)\right)}_{\propto \mathrm{N}(\beta|\tau_1, \tau_2^{-1}\sigma^2)} \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{\tau_0 - d}{2} - 1 + 1} \left(-\frac{1}{\sigma^2}\tau_3\right)}_{= \mathrm{IG}(\sigma^2|\frac{\tau_0 - d}{2} - 1, \tau_3)}\end{aligned}$$

where $\tau = (\tau_0, \tau_1, \tau_2, \tau_3)$ are just arbitrary parameters set by the researcher. I can use a friendlier parametrization

$$\begin{aligned}\beta|\sigma^2 &\sim \mathrm{N}(\mu_0, V_0\sigma^2); && \text{cond. prior distr} \\ \sigma^2 &\sim \mathrm{IG}(a_0, \kappa_0); && \text{prior distr}\end{aligned} \tag{7}$$

where $\mu_0 \in \mathbb{R}$,$V_0 > 0$, $a_0 > 0$, and $\kappa_0 > 0$. Recall from Exercise 26 in Exercise sheet that prior (7) updates to the posterior $\pi(\beta, \sigma^2|y)$ as

$$\begin{aligned}\beta|y, \sigma^2 &\sim \mathrm{N}(\mu_n, V_n\sigma^2); \\ \sigma^2|y &\sim \mathrm{IG}(a_n, \kappa_n)\end{aligned}$$

with some hyper-parameters $\mu_n, V_n, a_n, \kappa_n$ computed in Exercise 26.

1.2 Conjugate priors for Exponential families [1]

Note 13. Exponential family of distributions cover a large range of distributions satisfying the conditions in Note 4.

Fact 14. *(Pitman-Koopman-Lemma) If a distribution family $\{F(y|\theta), \forall\theta \in \Theta\}$ is such that there exists a sufficient statistic whose dimension is independent on the number of observations and the support $y \in \mathcal{Y}$ of $F(y|\theta)$ does not depend on θ, then it is an exponential family.*

Note 15. When the parametric model is member of the Exponential family, a conjugate prior distribution on its uncertain parameters can be specified as shown in Theorem 16.

Theorem 16. *Let $y = (y_1, ..., y_n)$ be observable quantities generated from an exponential family distribution as*

$$y_i|\theta \overset{iid}{\sim} \mathit{Ef}_k(u, g, h, c, \phi, \theta, c),\ i = 1, ..., n$$

with pdf/pmf

$$f(y_i|\theta) = \mathit{Ef}_k(y_i|u, g, h, c, \phi, \theta, c) = u(y_i)g(\theta)\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(y_i)\right).$$

Then the conjugate prior distribution $\Pi(\theta)$ for the likelihood has pdf/pmf of the form

$$\pi(\theta) := \tilde{\pi}(\theta|\tau) = \frac{1}{K(\tau)}g(\theta)^{\tau_0}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right) \propto g(\theta)^{\tau_0}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right)$$

for $\theta \in \Theta$, where $\tau = (\tau_0, \tau_1, ..., \tau_k)$ are hyper-parameters is such that

$$K(\tau) = \begin{cases}\int_\Theta g(\theta)^{\tau_0}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right)d\theta < \infty & \textit{cont.} \\ \\ \sum_{\forall\theta\in\Theta} g(\theta)^{\tau_0}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right) < \infty & \textit{discr.}\end{cases}$$

[1] Web applet: `https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/`

Proof. The likelihood is

$$f(y|\theta) = \prod_{i=1}^{n} \mathrm{Ef}(y_i|u, g, h, c, \phi, \theta, c) = \prod_{i=1}^{n} u(y_i) \underbrace{g(\theta)^n \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \left(\sum_{i=1}^{n} h_j(y_i)\right)\right)}_{=k(t(y)|\theta)}.$$

with sufficient statistic for θ

$$t(y) = \left(n, \sum_{i=1}^{n} h_1(y_i), ..., \sum_{i=1}^{n} h_k(y_i)\right) = (t_0, ..., t_k)$$

Let $\tau = (\tau_0, \tau_1, ..., \tau_n)$. The conjugate prior form has the form

$$\pi(\theta) := \tilde{\pi}(\theta|\tau) \propto k(t(y) = \tau|\theta) = g(\theta)^{\tau_0} \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j\right).$$

□

Note 17. Intuition about the prior in Theorem 16: τ_0 replaces the sample size n, and hence τ_0 can be thought of as being the weight of prior info or 'quality of prior info'; i.e. the larger the value the stronger the prior info. The rest $\tau_1, ...\tau_k$ can be thought of as summarizing the prior info.

Example 18. Let $y = (y_1, ..., y_n)$ be observable quantities, generated from an exponential family of distributions as

$$y_i|\theta \overset{\text{iid}}{\sim} \mathrm{Ef}(u, g, h, c, \phi, \theta, c),\ i = 1, ..., n$$

with density

$$\mathrm{Ef}(y_i|u, g, h, c, \phi, \theta, c) = u(y_i) g(\theta) \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) h_j(y_i)\right)$$

and assume a conjugate prior $\Pi(\theta)$ with pdf/pmf

$$\pi(\theta) = \tilde{\pi}(\theta|\tau)\ \propto g(\theta)^{\tau_0} \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j\right)$$

Show that the posterior $\mathrm{d}\Pi(\theta|y)$ of θ has pdf/pmf $\pi(\theta|y) = \tilde{\pi}(\theta|\tau^*)$ with $\tau^* = (\tau_0^*, \tau_1^*, ..., \tau_k^*)$, $\tau_0^* = \tau_0 + n$, and $\tau_j^* = \sum_{i=1}^{n} h_j(x_i) + \tau_j$ for $j = 1, ..., k$, and pdf/pmf

$$\pi(\theta|y) = \pi(\theta|\tau^*)\ \propto g(\theta)^{\tau^*} \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j^*\right) \tag{8}$$

- Comment: Here, the operation $*$ is addition i.e. $\tau * t(y) \longmapsto \tau + t(y) = \tau^*$

Solution. According to the Bayes theorem, where

$$\begin{aligned}\pi(\theta|y) \propto f(y|\theta)\pi(\theta) &\propto g(\theta)^n \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \left(\sum_{i=1}^{n} h_j(y_i)\right)\right) g(\theta)^{\tau_0} \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j\right) \\ &\propto g(\theta)^{n+\tau_0} \exp\left(\sum_{j=1}^{k} c_j \phi_j(\theta) \left(\sum_{i=1}^{n} h_j(y_i) + \tau_j\right)\right) \propto \tilde{\pi}(\theta|\tau + t(y)).\end{aligned}$$

Example 19. Let $y = (y_1, ..., y_n)$ be observables iid from a Bernoulli sampling distribution $y_i \overset{\text{iid}}{\sim} \text{Br}(\theta)$ for all $i = 1, ..., n$ where $\theta \in [0, 1]$. Specify a conjugate prior distribution for θ.

Hint: Beta distribution: if $x \sim \text{Be}(a, b)$, then $f(x) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}1(x \in [0, 1])$

Solution. The sampling distribution $f(x|\theta)$ is the Bernoulli distribution which belongs to the exponential family as

$$f(y_i|\theta) = \text{Br}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i} = (1-\theta)\exp\left(\log\left(\frac{\theta}{1-\theta}\right)y_i\right)$$

with $u(y_i) = 1$, $g(\theta) = (1-\theta)$, $c_1 = 1$, $\phi_1(\theta) = \log(\frac{\theta}{1-\theta})$, $h_1(y_i) = y_i$. The corresponding conjugate prior has pdf such as

$$\pi(\theta) \propto g(\theta)^{\tau_0}\exp\left(c_1\phi_1(\theta)\tau_j\right) = (1-\theta)^{\tau_0}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_1\right) = \theta^{(\tau_1+1)-1}(1-\theta)^{(\tau_0-\tau_1+1)-1}$$

Since we recognize that the prior distribution is Beta, we perform a re-parametrization, as

$$\theta \sim \text{Be}(a, b)$$

where $a = \tau_1 + 1 > 0$, $b = \tau_0 - \tau_1 + 1 > 0$.

Example 20. Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$ for all $i = 1, ..., n$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)$ is unknown. Specify a conjugate prior distribution for $\theta = (\mu, \sigma^2)$.

Solution. The sampling distribution $f(x|\mu, \sigma^2)$ is Normal distribution which is member of the regular 2-parameter exponential family, since

$$f(y_i|\mu, \sigma^2) = \text{N}(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(y_i-\mu)^2) = (\frac{1}{2\pi})^{\frac{1}{2}}(\frac{1}{\sigma^2})^{\frac{1}{2}}\exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2})\exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i)$$

$$\text{with} \quad u(y_i) = (\frac{1}{2\pi})^{\frac{1}{2}}, \quad g(\theta) = (\frac{1}{\sigma^2})^{\frac{1}{2}}\exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2}), \quad h(y_i) = (y_i, y_i^2), \quad \phi(\theta) = (\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}), \quad c = (1, -\frac{1}{2})$$

The corresponding conjugate prior has pdf such as

$$\begin{aligned}\pi(\mu, \sigma^2) &\propto \left(\sqrt{\frac{1}{\sigma^2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2}\mu^2)\right)^{\tau_0}\exp\left(\mu\frac{1}{\sigma^2}\tau_1 - \frac{1}{2}\frac{1}{\sigma^2}\tau_2\right) \\ &\propto \underbrace{\left(\frac{1}{\sigma^2/\tau_0}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{1}{\sigma^2/\tau_0}(\mu-\frac{\tau_1}{\tau_0})^2\right)}_{\propto \text{N}(\mu|\frac{\tau_1}{\tau_0}, \frac{\sigma^2}{\tau_0})}\underbrace{(\frac{1}{\sigma^2})^{\frac{(\tau_0-3)}{2}+1}\exp\left(-\frac{1}{\sigma^2}\frac{1}{2}(\tau_2-\frac{\tau_1^2}{\tau_0})\right)}_{\propto \text{IG}(\sigma^2|\frac{\tau_0-3}{2}, \frac{1}{2}(\tau_2-\frac{\tau_1^2}{\tau_0}))}\end{aligned}$$

where $\tau = (\tau_0, \tau_1, \tau_2)$. I recognize that the prior distribution is of standard form $\pi(\theta|\mu_0, n_0, a_0, \kappa_0) = \text{N}(\mu|\mu_0, \frac{\sigma^2}{\lambda_0})\text{IG}(\sigma^2|a_0, \kappa_0)$, with $\mu_0 = \frac{\tau_1}{\tau_0}$, $\lambda_0 = \tau_0$, $a_0 = \frac{\tau_0-3}{2}$, and $b_0 = \frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})$.

2 Conditional conjugate priors

Note 21. In some problems involving more realistic/complicated statistical models, certain computational tools, (e.g., the Gibbs sampler in Term 2), require the availability of tractable posterior conditionals instead of that of the full joint posterior. Specifying, conditional conjugate priors is a way to achieve this.

Definition 22. Let $\mathcal{F} = \{F(y|\theta_1, \theta_2); \forall\theta_1 \in \Theta_1, \forall\theta_2 \in \Theta_2\}$ be a family of sampling distributions. A family of prior distributions $\mathcal{P}_{\theta_1}$ for θ_2 conditional on θ_1 is said to be conditional conjugate for $\mathcal{F}$ if the posterior $\Pi(\theta_2|y, \theta_1)$ belongs to $\mathcal{P}_{\theta_1}$ for all prior $\Pi(\theta_2|\theta_1) \in \mathcal{P}_{\theta_1}$ and all $F(y|\theta_1, \theta_2) \in \mathcal{F}$; i.e.

$$\Pi(\theta_2|y, \theta_1) \in \mathcal{P}_{\theta_1}, \ \ \forall F(y|\theta_1, \theta_2) \in \mathcal{F} \text{ and } \Pi(\theta_2|\theta_1) \in \mathcal{P}_{\theta_1}.$$

Note 23. The conditional conjugate prior for θ_2 conditional θ_1 is specified from Theorem 8 as the conjugate prior of θ_2 on $F(y|\theta_1, \theta_2)$ given that parameter θ_1 is known/fixed. Based on this, Neyman factorization is applied as

$$f(y|\theta_1, \theta_2) = k_1(t(y)|\theta_1)\rho(y|\theta_1) \ \propto k(t(y)|\theta_1),$$

and the prior is specified according to

$$\pi(\theta_1) \propto k_1(\tau_1|\theta_1) \tag{9}$$

for some researchers specified prior hyper-parameter vector τ_1. Likewise, I get the conditional conjugate prior for θ_1 conditional θ_2 as $\pi(\theta_2) \propto k_2(\tau_2|\theta_2)$. The join prior $\Pi(\theta)$ satisfying conditional conjugation for θ_1 and θ_2 is

$$\pi(\theta) = \pi(\theta_1)\pi(\theta_2) \propto k_1(\tau_1|\theta_1)k_2(\tau_2|\theta_2).$$

This derivation is extendable to any number of blocks $\theta = (\theta_1, ..., \theta_B)$.

Example 24. Consider the the model of Normal linear regression where the observables are pairs (ϕ_i, y_i) for $i = 1, ..., n$, assumed to be modeled according to the sampling distribution $y_i|\beta, \sigma^2 \sim \mathrm{N}(\phi_i^\top \beta, \sigma^2)$ for $i = 1, ..., n$ with unknown $(\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Find the conditional conjugate priors for (β, σ^2).

Solution. The likelihood kernel is

$$f(y|\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right](\beta - \hat{\beta}_n) - \frac{1}{2\sigma^2}(n-d)\hat{\sigma}_n^2\right) \tag{10}$$

To find the conditional conjugate $\pi(\beta)$: I consider σ^2 as fixed/known/nuisance, and hence the kernel in 9 is

$$f(y|\beta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right](\beta - \hat{\beta}_n)\right)$$

leading to a conjugate prior

$$\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta - \mu_0)^\top \overbrace{V_0^{-1}}^{\text{absorbs constant } \sigma^2} (\beta - \mu_0)\right) \propto \mathrm{N}(\beta|\mu_0, V_0)$$

To find the conditional conjugate $\pi(\sigma^2)$: I consider β as fixed/known/nuisance, and hence the likelihood kernel in 9 is

$$f(y|\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\overbrace{\frac{n}{2}}^{\text{data}}} \exp\left(-\frac{1}{\sigma^2}\overbrace{\frac{(\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right](\beta - \hat{\beta}_n) + (n-d)\hat{\sigma}_n^2}{2}}^{\text{data/constants}}\right)$$

leading to a conjugate conditional prior

$$\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a_0+1} \exp\left(-\frac{1}{\sigma^2}\kappa_0\right) \propto \mathrm{IG}(\sigma^2|a_0, \kappa_0)$$

Then the conditional conjugate $\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2)$ is

$$\begin{cases} \beta \sim \mathrm{N}(\mu_0, V_0); \\ \sigma^2 \sim \mathrm{IG}(a_0, \kappa_0) \end{cases} \tag{11}$$

The conditional posterior distribution $\Pi(\beta|y,\sigma^2)$ is $\beta|y,\sigma^2 \sim \mathrm{N}(\mu_n, V_n)$, computed by Bayesian theorem as

$$\pi(\beta|y,\sigma^2) \propto f(y|\beta,\sigma^2)\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\frac{\Phi^\top\Phi}{\sigma^2}\right](\beta-\hat{\beta}_n)-\frac{1}{2}(\beta-\mu_0)^\top V_0^{-1}(\beta-\mu_0)\right) \propto \mathrm{N}(\mu_n, V_n')$$

$$\text{with} \quad V_n' = \left[\frac{\Phi^\top\Phi}{\sigma^2}+V_0^{-1}\right]^{-1} \quad \text{and} \quad \mu_n = V_n'\left[\frac{\Phi^\top\Phi}{\sigma^2}\hat{\beta}_n + V_0^{-1}\mu_0\right]$$

and $\Pi(\sigma^2|y,\beta)$ is $\sigma^2|y,\beta \sim \mathrm{IG}(a_n,\kappa_n)$, computed by Bayesian theorem as

$$\pi(\sigma^2|y,\beta) \propto f(y|\beta,\sigma^2)\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a_0+1}\exp\left(-\frac{1}{\sigma^2}\left[\frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n)+(n-d)\hat{\sigma}_n^2+\kappa_0\right]\right)$$
$$\propto \mathrm{IG}(\sigma^2|a_n,\kappa_n)$$

with $\kappa_n = \frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n)+(n-d)\hat{\sigma}_n^2+\kappa_0$ and $a_n = \frac{n}{2}+a_0$. Hence, according to Definition 22, we verified conditional conjugation of (11) an computed the set of the associated full conditional posteriors

$$\begin{cases}\beta|y,\sigma^2 \sim & \mathrm{N}(\mu_n, V_n') \\ \sigma^2|y,\beta \sim & \mathrm{IG}(a_n,\kappa_n)\end{cases}$$

3 Practice

Question 25. *For practice try Exercises 46, 33, and 47 from the Exercise sheet.*

Question 26. *Consider a Normal regression problem,*

$$y_i = \phi_i^\top\beta = \beta_0 + \left(\frac{x_i-\bar{x}}{s_x}\right)\beta_1 + \epsilon_i\,,\ \ \epsilon_i \sim N(0,\sigma^2)\,,\ \textit{for } i = 1,...,n,$$

where let's say y_i denotes the length (in cm), with x_i denoting the temperature (in Celsius degrees) of water the i-th fish swims. Between the priors specified in Example 12 and Example 24, which one (and why) is more reasonable from the modeling point of view?