Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 6: Conjugate priors [a]

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

Aim: Explain the prior distribution. 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/`

[a] Author: Georgios P. Karagiannis.

1 Proper & improper priors

Note 1. Priors do not necessarily need to be probability distributions but they need to lead to posterior probability distributions.

Definition 2. The prior $\Pi(\theta)$ with pdf/pmf $\pi(\theta) > 0$ for $\theta \in \Theta$, is called proper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta < \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) < \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a probability distribution ; and improper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta = \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) = \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a not a probability distribution.

Note 3. An improper prior $\Pi(\theta)$ can only be used for inference if it leads to a well defined posterior probability distribution (aka proper posterior); namely if the 'Properness condition'

$$\int_{\Theta} f(y|\theta)\pi(\theta)\mathrm{d}\theta < \infty \tag{1}$$

is satisfied for the observable sequence y at hand. If (1) is not satisfied, posterior quantities like mean, median, variance have no meaning.

Proposition 4. *If the sampling distribution $F(\cdot|\theta)$ is discrete and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is always proper.*

Proof. Provided as an Exercise 34 in the Exercise sheet. □

Proposition 5. *If the sampling distribution $F(\cdot|\theta)$ is continuous and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is almost always proper.*

Proof. Provided as an Exercise 35 in the Exercise sheet. □

2 Conjugate priors

Note 6. 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 7. 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 8. 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.

2.1 General derivation

Note 9. Let $y = (y_1, ..., y_n)$ be observables. We restrict the derivation of conjugate priors in cases where y_i are drawn from $F(\cdot|\theta)$ conditionally independent of θ, and 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.

What is in the box describes the rational of the approach and can be skipped.

Fact 10. *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 Note 9, there exists a binary operator $*$ such that*

$$y^{(1)} * y^{(2)} = y^* := (y_1^*, ..., y_k^*) \tag{2}$$

such that

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

Note 11. 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

$$f(y|\theta) = k(t(y)|\theta)\rho(y) \ \propto k(t(y)|\theta), \tag{3}$$

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

$$\mathrm{d}\Pi(\theta|y) = \frac{f(\theta|y)\mathrm{d}\Pi(\theta)}{\int f(\theta|y)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} \tag{4}$$

Assuming fictitious observables y' quantifying Your prior believe about θ (prior to getting the experimental info from observable data y), such that the sufficient statistic is $t' = t(y') \in \mathbb{R}^k$, I could possibly choose prior $\Pi(\theta)$ such that

$$\mathrm{d}\Pi(\theta) = \underbrace{\frac{1}{N(\tau)}k(t' = \tau|\theta)}_{=\pi(\theta)}\mathrm{d}\theta \text{ with } \pi(\theta) \propto k(\tau|\theta) \tag{5}$$

by assigning some researcher specified fixed hyper-parameters $\tau = (\tau_0, ..., \tau_{k-1})$ on $t' = t(y')$ such that the normalising constant is finite

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

Then from Fact 10, the posterior (4) could get a form such that

$$\mathrm{d}\Pi(\theta|y) = \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta}{\int k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta} = \frac{k(t(y)*\tau|\theta)\mathrm{d}\theta}{\int k(t(y)*\tau|\theta)\mathrm{d}\theta} = \frac{1}{N(t(y)*\tau)}k(t(y)*\tau|\theta)\mathrm{d}\theta \tag{6}$$

where $*$ is the binary operator (2) that combines the two kernels $k(t(y)|\theta)k(\tau|\theta) = k(t(y)*\tau|\theta)$.
Essentially the prior $\Pi(\theta)$ (5) and the posterior $\Pi(\theta|y)$ (6) belong to the same distribution family $\mathcal{P}$. The only difference is in the hyper-parameter values. 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 12. *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*

$$\pi(\theta) := \tilde{\pi}(\theta|\tau) = \frac{1}{N(\tau)}k(\tau|\theta) \propto k(\tau|\theta) \tag{7}$$

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 13. Essentially, in Theorem 12, and Note 11, 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 14. 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 15. 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

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

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

$$\pi(\theta) := \pi(\theta|\tau) \propto \left(\frac{1}{\theta}\right)^{\tau_0} 1\left(\tau_1 \in [0, \theta]\right) \propto \mathrm{Pa}(\theta|a = \tau_0 - 1, b = \tau_1)$$

By Bayes theorem the posterior is

$$\pi(\theta|y) \propto \prod_{i=1}^{n} \text{Un}(y_i|0,\theta)\text{Pa}(\theta|\tau_0-1,\tau_1) \propto \overbrace{\left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} 1(y_i<\theta)}^{=\prod_{i=1}^n \text{Un}(y_i|0,\theta)} \times \overbrace{\left(\frac{1}{\theta}\right)^{\tau_0} 1(\theta>\tau_1)}^{\propto \text{Pa}(\theta|\tau_0-1,\tau_1)}$$
$$\propto \left(\frac{1}{\theta}\right)^{n+\tau_0} \underbrace{\prod_{i=1}^{n} 1(\theta>x_i)1(\theta>\tau_1)}_{=1(\theta>\max(\tau_1,x_{(n)}))} \propto \text{Pa}(\theta|a^*=n+\tau_0-1, b^*=\max(\tau_1,x_{(n)})).$$

where $\theta > \max(\tau_1, x_{(n)})$.

Example 16. 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 $y_i|\beta, \sigma^2 \sim \text{N}(\phi_i^\top \beta, \sigma^2)$ 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: $(y-\Phi\beta)^\top(y-\Phi\beta) = (\beta-\hat{\beta}_n)^\top \left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n) + (n+d-2)\hat{\sigma}_n^2$;

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

Solution. The likelihood is

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

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

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

but as $\tau = (\tau_0, \tau_1, \tau_2, \tau_3)$ are just arbitrary parameters set by the researcher, I can use a friendlier parametrization[1]

$$\beta|\sigma^2 \sim \text{N}(\mu_0, V_0\sigma^2)\text{;prior distr}$$
$$\sigma^2 \sim \text{IG}(a_0, \kappa_0)\text{prior distr}$$

[1] In Exercise 26 of theExercise sheet, the posterior $\pi(\beta, \sigma^2|y)$ is derived as

$$\beta|y,\sigma^2 \sim \text{N}(\mu_n, V_n\sigma^2);$$
$$\sigma^2|y \sim \text{IG}(a_n, \kappa_n)$$

with some hyper-parameters $\mu_n, V_n, a_n, \kappa_n$ computed there.

2.2 Conjugate priors for Exponential families [2]

Note 17. Exponential family of distributions cover a large range of distributions satisfying the conditions in Note 9.

Fact 18. *(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 19. When the parametric model is member of the Exponential family, a conjugate prior distribution on its uncertain parameters can be specified.

Theorem 20. *Let* $y = (y_1, ..., y_n)$ *be observable quantities generated from an exponential family distribution as*

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

with pdf/pmf

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

Then the conjugate prior distribution $\mathrm{d}\Pi(\theta)$ *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*

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

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 21. Intuition about the prior in Theorem 20: τ_0 replaces the sample size n, and hence τ 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.

[2] https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/

Example 22. Let $y = (y_1, ..., y_n)$ be observable quantities, generated from an exponential family of distributions as

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

with density

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

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 $\text{d}\Pi(\theta|y)$ 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: The operation $*$ here is addition $\tau * t(y) \longmapsto \tau + t(y) = \tau^*$

Solution. According to the Bayes theorem, where

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) \propto g(\theta)^n \exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\left(\sum_{i=1}^{n} h_j(y_i)\right)\right)\ g(\theta)^{\tau_0} \exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right)$$

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

Example 23. Let $y = (y_1, ..., y_n)$ be observables drawn 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

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

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 24. 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

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

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

The corresponding conjugate prior has pdf such as

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

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

3 Conditional conjugate

Note 25. 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 26. 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.

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

Note 27. The conditional conjugate prior for θ_2 conditional θ_1 is specified from Theorem 12 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 28. 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

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

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

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

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 full 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

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

with $\kappa_n = \frac{1}{2}(\beta-\hat{\beta}_n)^\top \left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n) + \kappa_0$ and $a_n = \frac{n}{2} + a_0$. Hence, according to Definition 26, we verified conditional conjugation of (11) with the associated full conditional posteriors

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

4 Practice

Question 29. *For practice try Exercises 36, 33, and 37 from the Exercise sheet.*

Question 30. *Consider a Normal regression problem,*

$$y_i = \phi_i^\top \beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2)$$

where let's say y_i denotes the length (in cm), and $\phi_i = (1, x_i)$ with x_i denoting the temperature (in Celsius degrees) of water the i-th fish swims. Between the priors specified in Example 16 and Example 28, which one (and why) is more reasonable from the modeling point of view?