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

Handout 8: Non-informative priors

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

Aim: Explain the non-informative priors. Explain, theorize, and construct Laplace and Jeffreys' priors.

References:

- Robert, C. (2007; Sections 3.5.1 - 3.5.3.). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Berger, J. O. (2013; Sections 3.3, & 4.2.3). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Web applets: `https://georgios-stats-1.shinyapps.io/demo_conjugatejeffreyslaplacepriors/`

1 Non-informative priors

Note 1. Non-informative priors (or objective priors) are often specified (in practice) when no prior information is available.

Note 2. Non-informative prior distributions can be hardly justified in the subjective Bayesian stats because probabilities are considered to be subjective, and hence every researcher has some personal believe about the unknowns. Non-informative priors are used in the Subjective Bayes framework as a last resort when no prior information exist or when the specific application requires them.

Note 3. Objective Bayes and objective probability are variants of the subjective Bayesian stats and probability, where the probability is assumed to represent degree of believe about a proposition (similar to Subjective framework) but this is not a matter of an individual's personal degree of believe (in contrast to Subjective framework). There is no room for personal belief, hence everyone should assign the same prior probabilities, hence the posterior probability should be the same for different researchers. This Bayesian philosophical variation can be hardly justified, no?

Note 4. Objective Bayes requires the specification of non-informative priors as a mean to eliminate subjective/individual believes to the priors, and assign priors generally accepted by everybody. So these non-informative priors are also called objective priors.

Note 5. Objective Bayes is an arguable variation of the Bayesian framework. Different people may have different prior degrees of believe and hence the use of a prior accepted by everybody is arguable. Also, prior degree of believe of a group or several groups of researchers can still be expressed by subjective priors and be justified in the Subjective Bayesian framework.

Note 6. Some tools used to specify non-informative (objective) priors often break the Bayesian paradigm, and may produce unreasonable results, e.g. violation of the likelihood principle.

Note 7. In most applications it is almost impossible to specify non-informative priors representing exactly total ignorance about the problem at hand. They should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing. More realistically, non-informative priors about specific aspects of the problem or features of the statistical model (e.g., transformations, rotations) can be derived, and justified in the Subjective and Objective Bayes framework.

2 Laplace non-informative priors

Definition 8. The principle of insufficient prior states that we do not have any reason to think that one value of the unknown quantity is more likely than any other, so we should use a Uniform prior.

Definition 9. Laplace prior $\Pi(\theta)$ for $\theta \in \Theta$ is specified such as

$$\mathrm{d}\Pi(\theta) \propto \underbrace{1}_{\propto \pi(\theta)} \mathrm{d}\theta \quad \text{with pdf/pmf} \quad \pi(\theta) \propto 1 \tag{1}$$

and builds upon the principle of insufficient prior.

Remark 10. Laplace prior (1) places the same degree of believe at each value of θ if θ discrete, and at each interval $\mathrm{d}\theta$ of the same length if θ is continuous.

Remark 11. Laplace priors (1) can be improper. In such cases the properness condition has to be checked. Laplace priors are improper when Θ is unbounded parametric space but are proper when Θ is bounded.

Example 12. Consider the Bayesian model

$$\begin{cases} y_i|\mu & \sim \mathrm{N}(\mu, 1),\ \forall i = 1, ..., n \\ \mu & \sim \Pi(\mu) \end{cases}$$

Laplace prior $\mathrm{d}\Pi(\mu) \propto 1\mathrm{d}\mu$ is an improper prior since $\int_\Theta \mathrm{d}\Pi(\mu) = \int_{\mathbb{R}} \pi(\mu)\mathrm{d}\mu = \int_{\mathbb{R}} 1\mathrm{d}\mu = +\infty$. However, it can be used as a prior because it satisfies the properness condition

$$\int_{\mathbb{R}} \prod_{i=1}^{n} \mathrm{N}(y_i|\mu, 1)1\mathrm{d}\mu \propto 2^{-\frac{n}{2}} (\pi)^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2}\left(\sum y_i^2\right) + \frac{1}{2}\left(\sum y_i\right)^2 \right) < \infty$$

and hence it leads to a well defined posterior probability distribution.

Remark 13. Laplace prior (1) are not invariant under transformations. Assume Laplace (non-informative) prior for θ, with pdf $\pi(\theta) \propto 1$. Consider a random quantity ψ where $\psi = g(\theta)$, such that $g : \Theta \to \Psi$ is an 1-1 transformation. Then it is

$$\pi_\psi(\psi) \propto \pi_\theta\left(g^{-1}(\psi)\right) \left| \frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi) \right|$$

which is not necessarily flat, and hence not necessarily non-informative. This is strange because it means that we a priori know nothing about θ but we a priori know something about $\psi = g(\theta)$...

Example 14. Consider an experiment with sampling distribution

$$y_i|\theta \sim \mathrm{Br}(\theta),\ \forall i = 1, ..., n$$

The Laplace (non-informative) prior for success frequency $\theta \in [0, 1]$ is $\mathrm{d}\Pi(\theta) \propto 1\mathrm{d}\theta$ and hence $\theta \sim \mathrm{Un}(0, 1)$ which is a proper prior. This implies that odds $\psi = \frac{\theta}{1-\theta}$ have prior

$$\pi_\psi(\psi) \propto \pi_\theta\left(\frac{\psi}{1+\psi}\right) \left| \frac{\mathrm{d}}{\mathrm{d}\psi} \frac{\psi}{1+\psi} \right| \propto \frac{1}{(1+\psi)^2}$$

which is informative. So I a priori know nothing about the frequency θ but I know something about the odds ψ...

Note 15. Those using Laplace priors argue that You should parametrize the likelihood $f(y|\theta)$ according to a desired parameterization (e.g., success frequency θ), assign a Laplace prior, and stick with it ignoring reparametrizations...

3 Jeffreys' priors

Note 16. Let $y = (y_1, ..., y_n)$ observables drawn from a sampling distribution $F(y|\theta)$ with density $f(y|\theta)$. I use this notation hereafter. We aim to specify a prior $\Pi(\theta)$ with density $\pi(\theta)$ in the Bayesian model

$$\begin{cases} y|\theta & \sim F(y|\theta) \\ \theta & \sim \Pi(\theta) \end{cases} \tag{2}$$

so that $\Pi(\theta)$ can be invariant to $1-1$ transformations. Precisely, this is an invariance to a 1-1 transformations, and it is reasonable in certain type of applications (mentioned in the classroom).

Definition 17. The Jeffreys' prior distribution $\Pi(\theta)$ of the unknown parameter $\theta \in \Theta = \mathbb{R}^k$ $k \geq 1$ is defined such as

$$\mathrm{d}\Pi(\theta) \propto \underbrace{\sqrt{\det(\mathscr{I}(\theta))}}_{\propto \pi(\theta)} \mathrm{d}\theta \quad \text{with pdf/pmf} \quad \pi(\theta) \propto \sqrt{\mathscr{I}(\theta)},$$

where $\mathscr{I}(\theta)$ is the Fisher Information.

Definition 18. Let $y = (y_1, ..., y_n)$ observables drawn from a sampling distribution $F(y|\theta)$ with density $f(y|\theta)$ labeled by parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $f(y|\theta)$ denote the likelihood.

Fisher Information $\mathscr{I}(\theta)$ is a $k \times k$ matrix defined as

$$\mathscr{I}(\theta) = \mathrm{E}_{F(\cdot|\theta)} \left(\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \right] \right)$$

where

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y|\theta) = \left[\tfrac{\mathrm{d}}{\mathrm{d}\theta_1} \log(f(y|\theta)), \quad ... \quad \tfrac{\mathrm{d}}{\mathrm{d}\theta_j} \log(f(y|\theta)) \quad , ... \quad \tfrac{\mathrm{d}}{\mathrm{d}\theta_d} \log(f(y|\theta)) \right] \in \mathbb{R}^k$$

So the (i,j) element of Fisher Information $\mathscr{I}(\theta)$ is

$$[\mathscr{I}(\theta)]_{i,j} = \mathrm{E}_{F(\cdot|\theta)} \left(\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log(f(y|\theta)) \right] \left[\frac{\mathrm{d}}{\mathrm{d}\theta_j} \log(f(y|\theta)) \right] \right)$$

Note 19. In the univariate case $\theta \in \Theta \subseteq \mathbb{R}$, it is

$$\mathscr{I}(\theta) = \mathrm{E}_{F(\cdot|\theta)} \left(\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta) \right)^2 \right) \tag{3}$$

Fact 20. *Some properties of Fisher information $\mathscr{I}(\theta)$:*

1. *Under regularity conditions, when $\theta \in \Theta = \mathbb{R}^k$, $\mathscr{I}(\theta)$ simplifies to*

$$[\mathscr{I}(\theta)]_{i,j} = -E_{F(\cdot|\theta)} \left(\frac{d^2}{d\theta_i d\theta_j} \log(f(y|\theta)) \,|\, \theta \right) = -\int_{\mathcal{X}} \frac{d^2}{d\theta_i d\theta_j} \log(f(y|\theta)) dF(y|\theta).$$

   and when $\theta \in \Theta = \mathbb{R}$ (univariate), $\mathscr{I}(\theta)$ simplifies to

$$\mathscr{I}(\theta) = -E_{F(\cdot|\theta)} \left(\frac{d^2}{d\theta^2} \log(f(y|\theta)) \,|\, \theta \right) = -\int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log(f(y|\theta)) dF(y|\theta),$$

2. *Let $y = (y_1, ..., y_n)$ be observables frown iid from distribution $F(\cdot|\theta)$. Let $\mathscr{I}_n(\theta)$ be the associated Fisher information based on n observables. Then $\mathscr{I}_n(\theta) = n\mathscr{I}_1(\theta)$.*

3. *Let $g : \Theta \to \Psi$ with $\psi = g(\theta)$ be a $1-1$ transformation. Then the Fisher information is*

$$\mathscr{I}(\psi) = J^\top \mathscr{I}(\theta) J, \qquad \textit{and} \qquad \det(\mathscr{I}(\psi)) = \det(\mathscr{I}(\theta)) \det(J)^2$$

where $J = \frac{d\theta}{d\psi}$ is the Jacobian of the transformation $\psi = g(\theta)$ whose (i,j) element is $[J]_{i,j} = \frac{\partial \theta_i}{\partial \psi_i}$.

In the univariate case where $\theta \in \Theta \subseteq \mathbb{R}$ $\psi \in \Psi \subseteq \mathbb{R}$ it is

$$\mathscr{I}(\psi) = \mathscr{I}(\theta) \left(\frac{d\theta}{d\psi} \right)^2$$

Note 21. The rational of Jeffreys' prior is that Fisher information is widely accepted as an indicator of the amount of information brought by the statistics model (or the observation) about θ. By (3), we can intuit that Fisher information $\mathscr{I}(\theta)$ can evaluate the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$ through the expected slope of $\log(f(y|\theta))$. Hence, the values of θ for which $\mathscr{I}(\theta)$ is larger should be more likely for the prior distribution.

- Therefore, to favor the values of θ for which $\mathscr{I}(\theta)$ is large is equivalent to minimizing the influence of the prior distribution and is therefore as noninformative as possible.

Theorem 22. *Jeffreys' priors are invariant under 1-1 transformations. Let Jeffreys' prior on θ with density $\pi(\theta) \propto \sqrt{\mathscr{I}(\theta)}$ associated to sampling pdf/pmf $f(y|\theta)$, and Jeffreys' prior on ψ with density $p(\psi) \propto \sqrt{\mathscr{I}(\psi)}$ associated to sampling pdf/pmf $f(y|\psi)$, where $\psi = g(\theta)$ and $g : \Theta \to \Psi$ is a 1-1 transformation. Then $p(\psi) \propto \pi(\theta) \left| \det \left(\frac{\partial \theta}{\partial \psi} \right) \right|$.*

Proof. It is $\mathscr{I}(\psi) = \left(\frac{\partial \theta}{\partial \psi} \right)^\top \mathscr{I}(\theta) \frac{\partial \theta}{\partial \psi}$, because

$$\begin{aligned}
\mathscr{I}(\psi) &= \mathrm{E}_{y \sim F(\cdot|\psi)} \left(\left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\psi)) \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\psi)) \right] \right) = \int \left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\psi)) \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\psi)) \right] \mathrm{d}F(y|\psi) \\
&= \int \left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\theta)) \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\psi} \log(f(y|\theta)) \right] \mathrm{d}F(y|\theta) = \int \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \frac{\partial \theta}{\partial \psi} \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \frac{\partial \theta}{\partial \psi} \right] \mathrm{d}F(y|\theta) \\
&= \left(\frac{\partial \theta}{\partial \psi} \right)^\top \int \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y|\theta)) \right] \mathrm{d}F(y|\theta) \left(\frac{\partial \theta}{\partial \psi} \right) \\
&= \left(\frac{\partial \theta}{\partial \psi} \right)^\top \mathrm{E}_{y \sim F(\cdot|\theta)} \left(\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right)^2 \left(\frac{\partial \theta}{\partial \psi} \right) = \left(\frac{\partial \theta}{\partial \psi} \right)^\top \mathscr{I}(\theta) \left(\frac{\partial \theta}{\partial \psi} \right)
\end{aligned}$$

Then

$$\pi_\psi(\psi) \propto \sqrt{\mathscr{I}(\psi)} = \sqrt{\left(\frac{\partial \theta}{\partial \psi} \right)^\top \mathscr{I}(\theta) \frac{\partial \theta}{\partial \psi}} \propto \sqrt{\mathscr{I}(\theta)} \left| \det \left(\frac{\partial \theta}{\partial \psi} \right) \right| \propto \pi_\theta(\theta) \left| \det \left(\frac{\partial \theta}{\partial \psi} \right) \right|$$

□

Example 23. Consider observable r drawn from a Binomial sampling distribution $r|\theta \sim \mathrm{Bn}(n, \theta)$ with pdf $\mathrm{Br}(r|\theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r}$ and mean $\mathrm{E}_{\mathrm{Bn}(n,\theta)}(r) = n\theta$. Find the Jeffreys prior for θ. Compute the posterior of θ given r.

Solution. The likelihood is $f(r|\theta) = \mathrm{Br}(r|\theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r}$ So

$$\begin{aligned}
\log(f(r|\theta)) &= \log \binom{n}{r} + r \log(\theta) + (n-r) \log(1-\theta) \Longrightarrow \\
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(f(r|\theta)) &= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \Longrightarrow \\
\mathscr{I}(\theta) &= -\mathrm{E}_{\mathrm{Bn}(n,\theta)} \left(-\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \right) = \frac{n}{\theta(1-\theta)} \Longrightarrow \\
\text{Jeffreys' prior is } \quad \pi^{(\mathrm{Jef};\mathrm{Bn})}(\theta) &\propto \frac{1}{\theta^{1/2}(1-\theta)^{1/2}} \propto \mathrm{Be}(\theta|0.5, 0.5) \qquad (4)
\end{aligned}$$

The posterior is

$$\pi^{(\text{Jef;Bn})}(\theta|r) \propto \text{Bn}(r|n,\theta)\pi^{(\text{Jef;Bn})}(\theta) \propto \theta^{r-1/2}(1-\theta)^{n-r-1/2}$$

Example 24. Consider observable n drawn from a Negative binomial $n|\theta \sim \text{Nb}(r,\theta)$ with pdf $\text{Nb}(n|r,\theta) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}$and mean $\text{E}_{\text{Nb}(r,\theta)}(n) = r/\theta$. Find the Jeffreys prior for θ. Compute posterior of θ given n.

Solution. So

$$\log(f(r|\theta)) = \log\binom{n-1}{r-1} + r\log(\theta) + (n-r)\log(1-\theta) \Longrightarrow$$

$$\frac{\text{d}^2}{\text{d}\theta^2}\log(f(r|\theta)) = -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \Longrightarrow$$

$$\mathscr{I}(\theta) = -\text{E}_{\text{Nb}(r,\theta)}\left(-\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2}\right) = \frac{r}{\theta^2(1-\theta)} \Longrightarrow$$

$$\text{Jeffreys' prior is} \quad \pi^{(\text{Jef;Nb})}(\theta) \propto \frac{1}{\theta^1(1-\theta)^{1/2}} \tag{5}$$

The posterior is $\pi^{(\text{Jef;Nb})}(\theta|n) \propto f(r|\theta)\pi^{(\text{Jef;Nb})}(\theta) \propto \theta^{r-1}(1-\theta)^{n-r-1/2}$.

Remark 25. Jeffreys' prior can violate the Likelihood Principle. This is because its derivation depends on the form of the specific experiment via Fisher information which can differ for two different experiments even though they have proportional likelihoods. E.g.; For two experiments with propositional likelihoods $f_1(\theta|y_1) \propto f_2(\theta|y_2) \propto L(\theta)$, Jeffreys' priors are $\pi^{(\text{jp},1)}(\theta) \propto \mathscr{I}_1(\theta)$, and $\pi^{(\text{jp},2)}(\theta) \propto \mathscr{I}_2(\theta)$, so

$$\pi^{(\text{exp }1)}(\theta|y_1) = \frac{L(\theta)\sqrt{\mathscr{I}_1(\theta)}}{\int L(\theta)\sqrt{\mathscr{I}_1(\theta)}\text{d}\theta}; \quad \pi^{(\text{exp }2)}(\theta|y_2) = \frac{L(\theta)\sqrt{\mathscr{I}_2(\theta)}}{\int L(\theta)\sqrt{\mathscr{I}_2(\theta)}\text{d}\theta}; \quad \overset{\mathscr{I}_1(\theta)\neq\mathscr{I}_2(\theta)}{\Longrightarrow} \quad \pi^{(\text{exp }1)}(\theta|y_1) \neq \pi^{(\text{exp }2)}(\theta|y_2)$$

In the Examples 23 and 24, even though the two likelihoods are equal up to a multiplicative constant, i.e.

$$\text{Bn}(r|n,\theta) \propto \text{Nb}(n|r,\theta) \propto \theta^r(1-\theta)^{n-r} \tag{6}$$

Jeffreys' priors led to different posteriors $\pi^{(\text{Jef;Br})}(\theta|r) \neq \pi^{(\text{Jef;Nb})}(\theta|n)$.

Example 26. Let $y \in \mathbb{R}$ be an observable. Consider the statistical model

$$y|\mu,\sigma^2 \overset{\text{iid}}{\sim} \text{N}(\mu,\sigma^2) \quad \text{where} \quad (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty).$$

1. Specify the Jeffreys' prior for $\theta = \mu$, when σ is known.
2. Specify the Jeffreys' prior for $\theta = \sigma$, when μ is known.
3. Specify the Jeffreys' prior for $\theta = (\mu,\sigma)$.

Solution. It is

$$\log f(y|\theta) = \log(\text{N}(y|\mu,\sigma^2)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}$$

1. It is

$$\frac{\text{d}}{\text{d}\mu}\log(\text{N}(y|\mu,\sigma^2)) = \frac{(y-\mu)}{\sigma^2}$$

$$\frac{\text{d}^2}{\text{d}\mu^2}\log(\text{N}(y|\mu,\sigma^2)) = \frac{\partial}{\partial\mu}\frac{(y-\mu)}{\sigma^2} = -\frac{1}{\sigma^2}$$

$$\mathscr{I}(\mu) = -\text{E}_{y\sim\text{N}(\mu,\sigma^2)}\left(\frac{\text{d}^2}{\text{d}\mu^2}\log\left(\text{N}(y|\mu,\sigma^2)\right)\right) = \frac{1}{\sigma^2}$$

and hence $\pi^{(\text{JP})}(\mu) \propto \sqrt{\mathscr{I}(\mu)} \propto 1$

2. It is

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\log(\mathrm{N}(y|\mu,\sigma^2)) = -\frac{1}{\sigma}+\frac{(y-\mu)^2}{\sigma^3}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\log(\mathrm{N}(y|\mu,\sigma^2)) = \frac{1}{\sigma^2}-3\frac{(y-\mu)^2}{\sigma^4} = \frac{1}{\sigma^2}-3\frac{1}{\sigma^2}\left(\frac{y-\mu}{\sigma^2}\right)^2$$

So

$$\mathscr{I}(\sigma) \quad = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{\partial^2}{\partial\theta^2}\log(\mathrm{N}(y|\mu,\sigma^2))\right) = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{1}{\sigma^2}-3\frac{1}{\sigma^2}\left(\frac{y-\mu}{\sigma^2}\right)^2\right) \quad \propto \frac{1}{\sigma^2}$$

and hence $\pi^{(\mathrm{JP})}(\sigma)\propto\sqrt{\mathscr{I}(\sigma)}\propto\frac{1}{\sigma}$

3. it is

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu\mathrm{d}\sigma}\log f(y|\theta) = \frac{\mathrm{d}^2}{\mathrm{d}\mu\mathrm{d}\sigma}\log\left(\mathrm{N}(y|\mu,\sigma^2)\right)) = -2\frac{(y-\mu)}{\sigma^3}$$

So

$$\mathscr{I}(\mu,\sigma) = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(\mathrm{N}(y|\mu,\sigma^2))\right) = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\begin{bmatrix}-\frac{1}{\sigma^2} & -2\frac{(y-\mu)}{\sigma^3}\\ -2\frac{(y-\mu)}{\sigma^3} & \frac{1}{\sigma^2}-3\frac{(y-\mu)^2}{\sigma^4}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & 2\frac{1}{\sigma^2}\end{bmatrix}$$

and $\pi^{(\mathrm{JP})}(\mu,\sigma)\propto\sqrt{\det(\mathscr{I}(\mu,\sigma))}\propto\frac{1}{\sigma^2}$.

Example 27. 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 $y_i|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}(\phi_i^\top\beta,\sigma^2)$ for $i=1,...,n$ with unknown $(\beta,\sigma^2)\in\mathbb{R}^d\times\mathbb{R}_+$. Namely, the sampling distribution in vector form is

$$y|\beta,\sigma^2 \sim \mathrm{N}_n(\Phi\beta, I\sigma^2)$$

where $y=(y_1,...y_n)$, and Φ is the design matrix. Here β is d-dimensional. Find the Jeffreys' priors for (β,σ^2).

Hint: Recall your AMV: $\frac{\mathrm{d}}{\mathrm{d}x}x^\top Ax = 2Ax$, $\frac{\mathrm{d}}{\mathrm{d}x}(c+Ax) = A$, and $\frac{\mathrm{d}}{\mathrm{d}x}(A(x))^\top = \left(\frac{\mathrm{d}}{\mathrm{d}x}A(x)\right)^\top$.

Hint: If $y|\beta,\sigma^2\sim\mathrm{N}_n(\Phi\beta,I\sigma^2)$, then $\mathrm{E}_{y|\beta,\sigma^2\sim\mathrm{N}_n(\Phi\beta,I\sigma^2)}\left((y-\Phi\beta)^\top(y-\Phi\beta)\right) = n\sigma^2$.

Solution. Let's set $\xi=\sigma^2$ to simplify notation... The log likelihood is

$$\log\left(f(y|\beta,\xi)\right) = -\frac{n}{2}\log(\xi)-\frac{1}{2}\frac{1}{\xi}(y-\Phi\beta)^\top(y-\Phi\beta)$$

Let's compute the derivatives

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\log\left(f(y|\beta,\xi)\right) = -\frac{n}{2}\frac{1}{\xi}+\frac{1}{2}\frac{1}{\xi^2}(y-\Phi\beta)^\top(y-\Phi\beta)$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta,\xi)\right) = \frac{\mathrm{d}}{\mathrm{d}\xi}\left(-\frac{n}{2}\frac{1}{\xi}+\frac{1}{2}\frac{1}{\xi^2}(y-\Phi\beta)^\top(y-\Phi\beta)\right) = \frac{n}{2}\frac{1}{\xi^2}-\frac{1}{\xi^3}(y-\Phi\beta)^\top(y-\Phi\beta)$$
$$\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta,\xi)\right) = -\frac{1}{\xi}\Phi^\top(y-\Phi\beta)$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta,\xi)\right) = \frac{\mathrm{d}}{\mathrm{d}\beta}\left(-\frac{1}{\xi}\Phi^\top(y-\Phi\beta)\right) = -\frac{1}{\xi}\Phi^\top\Phi$$
$$\frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta,\xi)\right) = \frac{\mathrm{d}}{\mathrm{d}\xi}\left(-\frac{1}{\xi}\Phi^\top(y-\Phi\beta)\right) = \frac{1}{\xi^2}\Phi^\top(y-\Phi\beta)$$

Lets compute the components of the Fisher information. The sampling distribution of y is $y|\beta,\xi \sim \mathrm{N}(\Phi\beta, I\sigma^2)$ with $\mathrm{E}(y|\beta,\xi) = \Phi\beta$ and $\mathrm{Var}(y|\beta,\xi) = I\sigma^2$. So

$$\begin{aligned}
\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta,\xi)\right)\right) &= \mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{n}{2}\frac{1}{\xi^2} - \frac{1}{\xi^3}\left(y-\Phi\beta\right)^\top\left(y-\Phi\beta\right)\right) = +\frac{n}{2}\frac{1}{\xi^2} - \frac{1}{\xi^3}n\xi = -\frac{n}{2}\frac{1}{\xi^2} \\
\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta,\xi)\right)\right) &= -\frac{1}{\xi}\Phi^\top\Phi \\
\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta,\xi)\right)\right) &= \mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{1}{\xi^2}\Phi^\top\left(y-\Phi\beta\right)\right) = \frac{1}{\xi^2}\Phi^\top\left(\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}(y) - \Phi\beta\right) = 0
\end{aligned}$$

Then

$$\begin{aligned}
\mathscr{I} &= -\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(y|\theta))\right) = -\mathrm{E}_{\mathrm{N}(\Phi\beta,I\sigma^2)}\left(\begin{bmatrix} \frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta,\xi)\right) & \frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta,\xi)\right) \\ \frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta,\xi)\right) & \frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta,\xi)\right) \end{bmatrix}\right) \\
&= \begin{bmatrix} \frac{1}{\xi}\Phi^\top\Phi & 0 \\ 0 & \frac{n}{2}\frac{1}{\xi^2} \end{bmatrix}
\end{aligned}$$

So the Jeffreys prior for (β,ξ) has density such as

$$\pi^{(\mathrm{JP})}(\beta,\xi) \propto \sqrt{\det\left(\mathscr{I}\right)} = \sqrt{\det\left(\frac{1}{\xi}\Phi^\top\Phi\right)\det\left(\frac{n}{2}\frac{1}{\xi^2}\right)} = \left(\frac{1}{\xi}\right)^{\frac{d}{2}+1}\sqrt{\det\left(\Phi^\top\Phi\right)\det\left(\frac{n}{2}\right)} \propto \left(\frac{1}{\xi}\right)^{\frac{d}{2}+1}$$

Namely $\pi^{(\mathrm{JP})}(\beta,\sigma^2) = \left(\sigma^2\right)^{-\frac{d}{2}-1}$.

4 Limiting posterior distributions

Note 28. One way to derive a posterior in the absence of prior information, is to specify a non-informative (possibly improper) prior, check the properness condition, and compute the posterior by the Bayes theorem; i.e.

$$\pi(\theta|y) = \frac{f(y|\theta)\pi^{(\mathrm{impr.})}(\theta)}{\int f(y|\theta)\pi^{(\mathrm{impr.})}(\theta)\mathrm{d}\theta}$$

Note 29. An alternative way to derive a posterior distribution in the absence of prior information, is to compute it as a limit of a posterior updated from a proper prior. Namely:

1. specify a proper prior with a specific parametric form $\pi(\theta|\tau) \propto \tilde{\pi}(\theta|\tau)$ (e.g., conjugate prior with hypoerparameter τ), and compute the kernel $\tilde{\pi}(\theta|\tau)$ of its pdf/pmf as $\pi(\theta|\tau) \propto \tilde{\pi}(\theta|\tau)$
2. find values τ' for the prior hyper parameters such that the limit

   $$\tilde{\pi}'(\theta) = \lim_{\tau\to\tau'}\tilde{\pi}(\theta|\tau), \quad \text{when} \quad \tau \to \tau'$$

   can be considered as the kernel of a non-informative prior. E.g., $\tilde{\pi}'(\theta)$ can be a Jeffreys prior.
3. compute the posterior distribution $\pi(\theta|y,\tau)$ updated from the proper prior $\pi(\theta|\tau)$, by Bayes theorem

   $$\pi(\theta|y,\tau) = \frac{f(y|\theta)\tilde{\pi}(\theta|\tau)}{\int f(y|\theta)\tilde{\pi}(\theta|\tau)\mathrm{d}\theta}$$

4. compute the limiting posterior as

   $$\pi'(\theta|y) = \lim_{\tau\to\tau'}\pi(\theta|y,\tau)$$

The limiting posterior $\pi'(\theta|y)$ is derived based on no prior information,

Note 30. A convenient way is to specify a conjugate prior $\pi(\theta|\tau)$ and let the τ approach values that reduce the strength of the prior information.

Example 31. Let $\pi^{(\text{CP})}(\theta|\tau)$ denote the conjugate prior, $\pi^{(\text{JP})}(\theta|\tau)$ denote the Jeffreys' prior, and $\pi^{(\text{LP})}(\theta|\tau)$ denote the Laplace prior. For Bernoulli statistical model $y_i \overset{\text{iid}}{\sim} \text{Br}(\theta),\ \forall i = 1, ..., n$, it is

Conjugate prior: $\theta|\tau = (a, b) \sim \text{Be}(a, b)$ has pdf kernel $\pi^{(\text{CP})}(\theta|a, b) \propto \tilde{\pi}^{(\text{CP})}(\theta|a, b) = \theta^{a-1}(1-\theta)^{b-1}$. By Bayes theorem, the posterior is

$$\pi^{(\text{CP})}(\theta|y, a, b) = \text{Be}\left(\sum_{i=1}^{n} y_i + a, n - \sum_{i=1}^{n} y_i + b\right)$$

Laplace prior: At point $(a', b') = (1, 1)$, the asymptotic non-informative prior is

$$\pi(\theta) = \lim_{(a,b)\to(1,1)} \tilde{\pi}^{(\text{CP})}(\theta|a, b) = 1$$

that is $\theta \sim \text{U}(0, 1)$ a Laplace prior. Then the asymptotic posterior is

$$\pi(\theta|y) = \lim_{(a,b)\to(1,1)} \pi^{(\text{CP})}(\theta|y, a, b) = \text{Be}(\sum_{i=1}^{n} y_i + 1, n - \sum_{i=1}^{n} y_i + 1)$$

Jeffreys prior: At point $(a', b') = (0.5, 0.5)$, the asymptotic non-informative prior is

$$\pi(\theta) = \lim_{(a,b)\to(0.5,0.5)} \tilde{\pi}^{(\text{CP})}(\theta|a, b) = \theta^{-0.5}(1-\theta)^{-0.5}$$

which is the Jeffreys prior and improper prior. Then the asymptotic posterior is

$$\pi(\theta|y) = \lim_{(a,b)\to(0.5,0.5)} \pi^{(\text{CP})}(\theta|y, a, b) = \text{Be}(\sum_{i=1}^{n} y_i + 0.5, n - \sum_{i=1}^{n} y_i + 0.5)$$

which is a valid probability distribution.

5 Practice

Question 32. *For practice try to address the Exercises 54, 55, and 31, from the Exercise sheet. You can try the Exercise 59 from the Exercise sheet. which is related to Regression.*