Biostatistics 202C: Weeks 1 and 2

Andrew Holbrook

Fall 2022

1 Bayes' theorem

For two events A and B, the definition of conditional probability says

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)},$$

or

$$\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B) = \Pr(B|A) \cdot \Pr(A).$$

It follows that

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}. \tag{1}$$

Now, let $y \sim p(y|\theta)$ and $\theta \sim p(\theta)$. We rewrite (1) as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta} \quad \text{or} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\sum_\Theta p(y|\theta)p(\theta)},$$

where we have used the law of total probability. This extends to $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} p(y|\theta)$:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int_\Theta p(\mathbf{y}|\theta)p(\theta)d\theta} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int_\Theta p(\mathbf{y}|\theta)p(\theta)d\theta} = \frac{p(\theta)\prod_{n=1}^N p(y_n|\theta)}{\int_\Theta p(\theta)\prod_{n=1}^N p(y_n|\theta)d\theta}.$$

In the discrete case, we write

$$p(\theta|\mathbf{y}) = \frac{p(\theta)\prod_{n=1}^N p(y_n|\theta)}{\sum_\Theta p(\theta)\prod_{n=1}^N p(y_n|\theta)}.$$

It is also often useful to model $\mathbf{y} = y_1, \ldots, y_N \overset{\perp}{\sim} p(y_n|z_n, \theta)$ with latent variables $\mathbf{z} = z_1, \ldots, z_N \overset{iid}{\sim} p(z|\theta)$. In this case, we have

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} = \frac{\left(\int_\mathcal{Z} \prod_{n=1}^N p(y_n|z_n, \theta)p(z_n|\theta)d\mathbf{z}\right)p(\theta)}{\int_\Theta \left(\int_\mathcal{Z} \prod_{n=1}^N p(y_n|z_n, \theta)p(z_n|\theta)d\mathbf{z}\right)p(\theta)d\theta}.$$

2 Conjugate priors

Integrals are often difficult, so conjugate priors are sometimes convenient.

- *Conjugacy* refers to the situation when the prior $p(\theta)$ and posterior $p(\theta|\mathbf{y})$ belong to the same distribution (albeit with "updated" parameters).
- When one combines a *conjugate* prior with a specific likelihood, one may obtain the posterior in closed form, no computations necessary.

- Only works for exponential family distributions, e.g., the normal, beta, Bernoulli, gamma and Poisson distributions.

If y follows an exponential family distribution, then

$$p(y|\theta) = h(y)g(\theta)\exp\left(\phi(\theta)^T s(y)\right) .$$

The joint distribution for independent $\mathbf{y} = (y_1, \ldots, y_N)$ is

$$p(\mathbf{y}|\theta) = \left(\prod_{n=1}^{N} h(y_n)\right) g^N(\theta)\exp\left(\phi(\theta)\sum_{n=1}^{N} s(y_n)\right) .$$

$\phi(\theta)$ is the *natural parameter* and $t(\mathbf{y}) = \sum_n s(y_n)$ is the *sufficient statistic*. If we also specify that θ follows an exponential family distribution with prior

$$p(\theta) \propto g(\theta)^\eta \exp\left(\phi(\theta)\cdot\nu\right) .$$

It follows that

$$p(\theta|\mathbf{y}) \propto g^{N+\eta}(\theta)\exp\left(\phi(\theta)\cdot(t(\mathbf{y})+\nu)\right) ,$$

i.e., both the prior and posterior of θ belong to the same exponential family distribution.

2.1 Beta-binomial model

Supposing $y \sim \text{Binomial}(\theta, N)$, one may write the pmf of y

$$p(y|\theta, N) = \binom{N}{y}\cdot\theta^y(1-\theta)^{N-y} \propto (1-\theta)^N \exp\left(y\log\left(\frac{\theta}{1-\theta}\right)\right)$$

It follows that

$$g(\theta) = 1-\theta \quad \text{and} \quad \phi(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

We therefore wish to specify a prior with the form

$$p(\theta) \propto (1-\theta)^\eta \exp\left(\nu\log\left(\frac{\theta}{1-\theta}\right)\right) = (1-\theta)^{\eta-\nu}\theta^\nu$$

This is accomplished by specifying a prior $p(\theta) \equiv \text{beta}(\alpha = \nu+1,\, \beta = \eta-\nu+1)$:

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{N-y}\cdot(1-\theta)^{\eta-\nu}\theta^\nu = (1-\theta)^{(\eta-\nu+N-y)}\theta^{\nu+y} \\ &= (1-\theta)^{(\beta+N-y-1)}\theta^{\alpha+y-1} \end{aligned}$$

Thus, we may write

$$\theta|y \sim \text{beta}(\alpha' := \alpha+y,\, \beta' := \beta+N-y)$$

We can therefore also write

$$\begin{aligned} \text{E}(\theta|y) &= \frac{\alpha'}{\alpha'+\beta'} = \frac{\alpha+y}{\alpha+\beta+N} \\ \text{Var}(\theta|y) &= \frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)} = \frac{(\alpha+y)(\beta+N-y)}{(\alpha+\beta+N)^2(\alpha+\beta+N+1)} . \end{aligned}$$

2.2 Univariate normal, known variance

Suppose we know the variance but not the mean of Gaussian distributed data $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$. Then

$$\begin{aligned} p(\mathbf{y}|\theta, \sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_n (y_n - \theta)^2\right) \propto \exp\left(-\frac{N\theta^2}{2\sigma^2} + \frac{\theta}{\sigma^2}\sum_n y_n\right) \\ &= \exp\left(\boldsymbol{\phi}(\theta)^T \mathbf{t}(\mathbf{y})\right) , \end{aligned}$$

where

$$\boldsymbol{\phi}(\theta) = \begin{bmatrix} -\theta^2/2 \\ \theta \end{bmatrix} , \quad \mathbf{t}(\mathbf{y}) = \frac{1}{\sigma^2}\begin{bmatrix} N \\ \sum_n y_n \end{bmatrix} .$$

The following prior specification therefore satisfies conjugacy:

$$p(\theta) \propto \exp\left(\phi(\theta)^T \boldsymbol{\nu}\right) = \exp\left(-\frac{\theta^2}{2\tau_0^2} + \frac{\mu_0\theta}{\tau_0^2}\right) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right) .$$

Then the posterior takes the form

$$\begin{aligned} p(\theta|\mathbf{y}, \sigma^2) &\propto \exp\left(-\frac{\theta^2}{2\tau_0^2} + \frac{\mu_0\theta}{\tau_0^2}\right) \exp\left(-\frac{N\theta^2}{2\sigma^2} + \frac{\theta}{\sigma^2}\sum_n y_n\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)\theta^2 + \left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right)\theta\right) \\ &= \exp\left(-\frac{1}{2}\frac{\theta^2}{\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}} + \frac{\theta}{\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}}\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\frac{\theta^2}{\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}} + \frac{\theta}{\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}}\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right) - \frac{1}{2}\frac{1}{\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}}\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right)^2\right) . \end{aligned}$$

Evidently, we have

$$\theta|\mathbf{y}, \sigma^2 \sim \mathrm{N}\left(\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right)\left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}, \left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}\right) .$$

2.3 Univariate normal, known mean

Suppose we know the mean but not the variance of Gaussian distributed data $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$:

$$p(\mathbf{y}|\theta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_n (y_n - \theta)^2\right) .$$

Then $\phi(\sigma^2) = \frac{1}{\sigma^2}$, $\mathbf{t}(\mathbf{y}) = -\sum_n (y_n - \theta)^2/2$ and $g(\sigma^2) = (\sigma^2)^{-1/2}$. The following prior therefore satisfies conjugacy:

$$p(\sigma^2) \propto (\sigma^2)^{\eta} \exp\left(\nu\phi(\sigma^2)\right) .$$

Letting $\eta = -\alpha - 1$ and $\nu = -\beta$, we have

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right) \equiv \Gamma^{-1}(\alpha, \beta) .$$

Then the posterior satisfies

$$p(\sigma^2|\mathbf{y},\theta) \propto (\sigma^2)^{-\alpha-N/2-1} \exp\left(-\frac{\beta}{\sigma^2} + \frac{\sum_n (y_n-\theta)^2}{2\sigma^2}\right)$$
$$\text{or,} \quad \sigma^2|\mathbf{y},\theta \sim \Gamma^{-1}\left(\alpha' = \alpha + \frac{N}{2}, \beta' = \beta + \frac{\sum_n (y_n-\theta)^2}{2}\right)$$

Thus, the posterior mean is

$$\mathrm{E}(\sigma^2|\mathbf{y},\theta) = \frac{\beta'}{\alpha'-1} = \frac{\beta + \sum_n (y_n-\theta)^2/2}{\alpha + N/2 - 1}$$

2.4 Gamma-Poisson

Suppose we observe data $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} \mathrm{Pois}(\theta)$:

$$p(\mathbf{y}|\theta) \propto \prod_n e^{-\theta}\theta^{y_n} = e^{-N\theta}\theta^{\sum_n y_n}.$$

Specifying a $\Gamma(\alpha,\beta)$ prior

$$p(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta},$$

we obtain

$$\theta|\mathbf{y} \sim \Gamma(\alpha' = \alpha + \sum_n y_n\,,\ \beta' = N + \beta).$$

Thus, the posterior mean is

$$\mathrm{E}(\theta|\mathbf{y}) = \frac{\alpha'}{\beta'} = \frac{\alpha + \sum_n y_n}{N+\beta} = \frac{N}{N+\beta}\frac{\sum_n y_n}{N} + \frac{\beta}{N+\beta}\frac{\alpha}{\beta}.$$

2.5 Multivariate normal, known mean

Suppose $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathrm{N}_D(\boldsymbol{\theta}, \boldsymbol{\Sigma})$:

$$p(\mathbf{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{1}{2}\sum_n (\mathbf{y}_n-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_n-\boldsymbol{\theta})\right) = |\boldsymbol{\Sigma}|^{-N/2}e^{-\frac{N}{2}\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})}.$$

If we specify an inverse-Wishart prior $\boldsymbol{\Sigma} \sim \mathrm{W}^{-1}(\boldsymbol{\Psi},\nu)$ for $\nu > D-1$, we have

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+D+1)/2}e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})}.$$

It follows that the posterior density satisfies

$$p(\boldsymbol{\Sigma}|\mathbf{Y},\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}|^{-(N+\nu+D+1)/2}e^{-\frac{1}{2}\mathrm{tr}\left((\boldsymbol{\Psi}+N\widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}\right)},$$

and

$$\boldsymbol{\Sigma}|\mathbf{Y},\boldsymbol{\theta} \sim \mathrm{W}^{-1}(\boldsymbol{\Psi}' = \boldsymbol{\Psi} + N\widehat{\boldsymbol{\Sigma}}, \nu' = N+\nu).$$

For $\nu > D+1$, the posterior mean is then

$$\mathrm{E}(\boldsymbol{\Sigma}|\mathbf{Y},\boldsymbol{\theta}) = \frac{\boldsymbol{\Psi}'}{\nu'-D-1} = \frac{\boldsymbol{\Psi}+N\widehat{\boldsymbol{\Sigma}}}{N+\nu-D-1}.$$

3 Mathematically interesting priors

The Jeffreys and reference priors both formalize the idea of an 'uninformative' prior in different ways.

3.1 The Jeffreys prior

Suppose $y \sim p(y|\theta)$. Then the Jeffreys prior is defined as

$$p(\theta) \propto \sqrt{|\mathcal{I}(\theta)|}$$

for

$$\mathcal{I}(\theta) = \mathrm{E}_{y|\theta}\left[\left(\frac{d}{d\theta}\log p(y|\theta)\right)^2\right] = -\mathrm{E}_{y|\theta}\left[\frac{d^2}{d\theta^2}\log p(y|\theta)\right].$$

Note that the Jeffreys prior may be proper or improper depending on the integrability of $\sqrt{\mathcal{I}(\theta)}$

3.1.1 Invariance under reparameterization

In addition to the model $y \sim p_\theta(y|\theta)$, we have a reparameterized model $y \sim p_\eta(y|\eta)$, where $\eta = h(\theta)$ and h is a monotone, differentiable function. In a multivariate setting, we require that h be a diffeomorphism. Let $\mathcal{I}_\theta(\theta)$ and $\mathcal{I}_\eta(\eta)$ be the parameterizations' respective Fisher informations. Under regularity conditions, we have

$$\begin{aligned}
\mathcal{I}_\theta(\theta) &= \int \left(\frac{d}{d\theta}\log p_\theta(y|\theta)\right)^2 p_\theta(y|\theta)dy \qquad (2)\\
&= \int \left(\frac{d}{d\theta}\log p_\eta(y|h(\theta))\right)^2 p_\eta(y|h(\theta))dy\\
&= \int \left(\frac{d}{d\eta}\log p_\eta(y|h(\theta))\big|_{\eta=h(\theta)} h'(\theta)\right)^2 p_\eta(y|h(\theta))dy\\
&= h'(\theta)^2\mathcal{I}_\eta(\eta)\,.
\end{aligned}$$

It is often said that the Jeffreys prior is parameterization invariant, but it might be better to say that it is consistent with the change of variables formula, i.e., a Jeffreys prior remains a Jeffreys prior after change of variables. This follows immediately from (2):

$$p_\theta(\theta) \propto \sqrt{\mathcal{I}_\theta(\theta)} = \sqrt{h'(\theta)^2\mathcal{I}_\eta(\eta)} \propto p_\eta(\eta)|h'(\theta)|\,,$$

which we can rewrite as

$$p_\eta(\eta) \propto p_\theta(h^{-1}(\eta))\left|\frac{d}{d\eta}h^{-1}(\eta)\right|.$$

3.1.2 Mean of Gaussian distribution

Consider $y \sim \mathrm{N}(\mu, \sigma^2)$ and fix the variance σ^2. Then

$$\begin{aligned}
p(\mu) &\propto \mathrm{E}_{y|\mu}\left[\left(\frac{d}{d\mu}\log p(y|\mu)\right)^2\right]^{1/2} = \mathrm{E}_{y|\mu}\left[\left(\frac{y-\mu}{\sigma^2}\right)^2\right]^{1/2}\\
&= \frac{1}{\sigma^2}\mathrm{E}_{y|\mu}\left[(y-\mu)^2\right]^{1/2} = \frac{1}{\sigma} \propto 1\,.
\end{aligned}$$

Evidently, the Jeffreys prior is improper in this situation, but can be interpreted as a conjugate prior in the limit as $\sigma_0^2 \to \infty$.

3.1.3 Standard deviation of Gaussian distribution

Consider $y \sim \mathrm{N}(\mu, \sigma^2)$ and fix the mean μ. Then

$$\begin{aligned}
p(\sigma) &\propto \mathrm{E}_{y|\sigma}\left[\left(\frac{d}{d\sigma}\log p(y|\sigma)\right)^2\right]^{1/2} = \mathrm{E}_{y|\sigma}\left[\left(\frac{(y-\mu)^2-\sigma^2}{\sigma^3}\right)^2\right]^{1/2}\\
&= \frac{1}{\sigma^3}\mathrm{E}_{y|\sigma}\left[(y-\mu)^4 - 2(y-\mu)^2\sigma^2 + \sigma^4\right]^{1/2} = \frac{\sqrt{3\sigma^4 - 2\sigma^4 + \sigma^4}}{\sigma^3} \propto \frac{1}{\sigma}\,.
\end{aligned}$$

Thus, the Jeffreys prior in this situation amounts to a power law and is improper.

3.1.4 Mean of Poisson distribution

If $y \sim \text{Pois}(y|\lambda)$, then

$$p(\lambda) \propto \text{E}_{y|\lambda}\left[\left(\frac{d}{d\lambda}\log p(y|\lambda)\right)^2\right]^{1/2} = \text{E}_{y|\lambda}\left[\left(\frac{y-\lambda}{\lambda}\right)^2\right]^{1/2} = \sqrt{\frac{\lambda}{\lambda^2}} = \frac{1}{\sqrt{\lambda}}\,.$$

3.2 Reference prior

Again consider $y \sim p(y|\theta)$ and suppose $t(y)$ is a sufficient statistic for θ. Formally, this means that $I(\theta, y) = I(\theta, t(y))$, where $I(\theta, t)$ is the mutual information

$$I(\theta, t) = \int\int p(\tilde{\theta}, \tilde{t}) \log\left(\frac{p(\tilde{\theta}, \tilde{t})}{p(\tilde{\theta})p(\tilde{t})}\right) d\tilde{\theta}d\tilde{t}\,.$$

We want to find a prior $p(\theta)$ that maximizes the K-L divergence between itself and the posterior $p(\theta|y)$, where the K-L divergence is

$$D_{KL}(p(\theta|y)||p(\theta)) = \int p(\tilde{\theta}|y) \log\left(\frac{p(\tilde{\theta}|y)}{p(\tilde{\theta})}\right) d\tilde{\theta}.$$

Taking the expectation of this quantity with respect to y gives

$$\begin{aligned}\text{E}_y\left(D_{KL}(p(\theta|y)||p(\theta))\right) &= \int p(\tilde{y}) \int p(\tilde{\theta}|\tilde{y}) \log\left(\frac{p(\tilde{\theta}|\tilde{y})}{p(\tilde{\theta})}\right) d\tilde{\theta}d\tilde{y} \\ &= \int\int p(\tilde{\theta}, \tilde{y}) \log\left(\frac{p(\tilde{\theta}, \tilde{y})}{p(\tilde{\theta})p(\tilde{y})}\right) d\tilde{\theta}d\tilde{y} = I(\theta, y) = I(\theta, t)\,.\end{aligned}$$

Then choosing a reference prior amounts to finding

$$p_1(\theta) = \arg\max_{p(\theta)} I(\theta, t)\,.$$

Asymptotics help to obtain tractable solutions to this problem: we consider instead the problem of finding

$$p_N(\theta) = \arg\max_{p(\theta)} I(\theta, t_N)\,,$$

where t_k is a sufficient statistic derived from $y_1, \ldots, y_N \sim p(y|\theta)$. To obtain such a solution, we may rewrite

$$\begin{aligned}I(\theta, t_N) &= \int p(\tilde{t}_N) \int p(\tilde{\theta}|\tilde{t}_N) \log\left(\frac{p(\tilde{\theta}|\tilde{t}_N)}{p(\tilde{\theta})}\right) d\tilde{\theta}d\tilde{t}_N \\ &= \int p(\tilde{\theta}) \int p(\tilde{t}_N) \frac{p(\tilde{\theta}|\tilde{t}_N)}{p(\tilde{\theta})} \log\left(\frac{p(\tilde{\theta}|\tilde{t}_N)}{p(\tilde{\theta})}\right) d\tilde{t}_N d\tilde{\theta} \\ &= \int p(\tilde{\theta}) \int p(\tilde{t}_N|\tilde{\theta}) \log\left(\frac{p(\tilde{\theta}|\tilde{t}_N)}{p(\tilde{\theta})}\right) d\tilde{t}_N d\tilde{\theta} \\ &= \int p(\tilde{\theta}) \left(\int p(\tilde{t}_N|\tilde{\theta}) \log\left(p(\tilde{\theta}|\tilde{t}_N)\right) d\tilde{t}_N - \log(p(\tilde{\theta}))\right) d\tilde{\theta} \\ &= \int p(\tilde{\theta}) \log\left(\frac{q_N(\tilde{\theta})}{p(\tilde{\theta})}\right) d\tilde{\theta}\end{aligned}$$

for

$$q_N(\theta) = \exp\left(\int p(\tilde{t}_N|\theta) \log\left(p(\theta|\tilde{t}_N)\right) d\tilde{t}_N\right)\,.$$

3.2.1 Solution using Lagrange multipliers

In the discrete case, we use Lagrange multipliers and solve

$$p_* = \arg\max_p \sum_i p_i \log \frac{q_i}{p_i} + \lambda \left(\sum_i p_i - 1 \right)$$

by obtaining partial derivatives

$$\frac{\partial}{\partial p_j} \left(\sum_i p_i \log \frac{q_i}{p_i} + \lambda \left(\sum_i p_i - 1 \right) \right) = \log(q_j/p_j) + p_j \frac{p_j}{q_j} \left(\frac{-q_j}{p_j^2} \right) + \lambda$$

and setting

$$\log q_j - \log p_j - 1 + \lambda = 0 \,.$$

From here, we obtain

$$p_j = q_j e^{\lambda - 1} \,.$$

The original term is the negated K-L divergence, which obtains its maximum when $p_j = q_j$, a solution we obtain by setting $\lambda = 0$. Similarly, for the continuous example, one may use the calculus of variations to solve

$$p_N(\theta) = \sup_{p(\theta)} \int p(\tilde{\theta}) \log \left(\frac{q_N(\tilde{\theta})}{p(\tilde{\theta})} \right) d\tilde{\theta} + \lambda \left(\int p(\tilde{\theta}) d\tilde{\theta} - 1 \right)$$

and obtain

$$p_N(\theta) = q_N(\theta) = \exp \left(\int p(\tilde{t}_N|\theta) \log \left(p(\theta|\tilde{t}_N) \right) d\tilde{t}_N \right) \,.$$

This form is intractable, but we can use an asymptotic approximation of the term $p(\theta|t_N)$.

3.2.2 1D reference priors are Jeffreys priors (handwavy)

The Bernstein-von Mises theorem may be thought of as a Bayesian central limit theorem. For $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} p(y|\theta_0)$, the theorem roughly says that $\theta|\mathbf{y}$ is asymptotically Gaussian with mean θ_0.

Theorem 1. *Assume regularity conditions conducive to the asymptotic normality of the MLE $\hat{\theta}_N$ and that $p(\theta)$ is continuous and positive in an open set surrounding θ_0. Finally, let t_N be a sequence of independent random variables following the distributions $t_N|\theta_0$ indexed by $N \in \mathbb{N}$. It follows that*

$$\| p(\theta|t_N) - N(\hat{\theta}_N, \mathcal{I}_N^{-1}(\theta_0)) \|_{TV} \xrightarrow{p} 0 \,.$$

By the asymptotic sufficiency of the MLE, we may also write

$$\| p(\theta|\hat{\theta}_N) - \mathrm{N}(\hat{\theta}_N, \mathcal{I}_N^{-1}(\theta_0)) \|_{TV} \xrightarrow{p} 0 \,.$$

Using the fact that $\mathcal{I}_N(\theta) = N\mathcal{I}(\theta)$, we have the asymptotic approximation

$$p(\theta|\hat{\theta}_N) \dot{\propto} N^{1/2} \mathcal{I}^{1/2}(\theta_0) \exp \left(-\frac{N}{2} \mathcal{I}(\theta_0)(\theta - \hat{\theta}_N)^2 \right) \,.$$

Evaluating this form at the truth results in the equation

$$p(\theta_0|\hat{\theta}_N) \dot{\propto} N^{1/2} \mathcal{I}^{1/2}(\theta_0) \exp \left(-\frac{N}{2} \mathcal{I}(\theta_0)(\theta_0 - \hat{\theta}_N)^2 \right) \,,$$

and consistency of the MLE results in the approximation

$$p(\theta_0|\hat{\theta}_N) \dot{\propto} N^{1/2} \mathcal{I}^{1/2}(\theta_0) \,.$$

Plugging this into the formula for $p_N(\theta)$, we have

$$\begin{aligned} p_N(\theta) = q_N(\theta) &= \exp\left(\int p(\tilde{t}_N|\theta) \log\left(p(\theta|\tilde{t}_N)\right) d\tilde{t}_N\right) \\ &= \exp\left(\int p(\tilde{t}_N|\theta) \log\left(N^{1/2}\mathcal{I}^{1/2}(\theta)\right) d\tilde{t}_N\right) \\ &= N^{1/2}\mathcal{I}^{1/2}(\theta) \exp\left(\int p(\tilde{t}_N|\theta) d\tilde{t}_N\right) \\ &= N^{1/2}\mathcal{I}^{1/2}(\theta) \propto \mathcal{I}^{1/2}(\theta)\,. \end{aligned}$$

3.2.3 Exponential distribution

Consider $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} \exp(\theta)$. Then the Fisher information is

$$\begin{aligned} \mathcal{I}(\theta) &= -\mathrm{E}_{y|\theta}\left(\frac{d^2}{d\theta^2} \log p(y|\theta)\right) = -\mathrm{E}_{y|\theta}\left(\frac{d^2}{d\theta^2} \left(\log(\theta) - y\theta\right)\right) \\ &= -\mathrm{E}_{y|\theta}\left(\frac{d}{d\theta}\left(\frac{1}{\theta} - y\right)\right) = \frac{1}{\theta^2}\,, \end{aligned}$$

and the Jeffreys prior is

$$p(\theta) \propto \sqrt{\mathcal{I}(\theta)} = \frac{1}{\theta}\,.$$

The likelihood is

$$p(\mathbf{y}|\theta) \propto \theta^N e^{-\theta \sum_n y_n}\,,$$

and the MLE $\hat{\theta}_N = \bar{y}_N^{-1}$is obtained by equating the log-likelihood derivative with 0:

$$0 = \frac{N}{\theta} - \sum_n y_n$$

Because of lack of dependence of the posterior on the prior under Bernstein-von Mises, we are free to assume a flat prior within the derivation and obtain

$$p(\theta|\hat{\theta}_N) \propto \theta^N e^{-N\theta/\hat{\theta}_N}\,.$$

Finally, we note that $\sum_n y_n \sim \Gamma(N, \theta)$, $1/\sum_n y_n \sim \Gamma^{-1}(N, \theta)$ and $N/\sum_n y_n \sim \Gamma^{-1}(N, N\theta)$, or

$$p(\hat{\theta}_N|\theta) \propto \theta^N \hat{\theta}_N^{-(N+1)} e^{-N\theta/\hat{\theta}_N}\,.$$

$$\begin{aligned} p_N(\theta) &= \exp\left(\int p(\hat{\theta}_N|\theta) \log(p(\theta|\hat{\theta}_N)) d\hat{\theta}_N\right) \\ &\propto \exp\left(\int \theta^N \hat{\theta}_N^{-(N+1)} e^{-N\theta/\hat{\theta}_N} \log\left(\theta^N e^{-N\theta/\hat{\theta}_N}\right) d\hat{\theta}_N\right) \\ &= \exp\left(\int \theta^N \hat{\theta}_N^{-(N+1)} e^{-N\theta/\hat{\theta}_N} \left(N\log(\theta) - N\theta/\hat{\theta}_N\right) d\hat{\theta}_N\right) \\ &= \frac{1}{\theta} \exp\left(-N\theta \int \theta^N \hat{\theta}_N^{-(N+2)} e^{-N\theta/\hat{\theta}_N} d\hat{\theta}_N\right) \\ &= \frac{1}{\theta} \exp\left(-N\theta\, \mathrm{E}_{\hat{\theta}_N|\theta}\left(\frac{1}{\hat{\theta}_N}\right)\right) \propto \frac{1}{\theta}\,, \end{aligned}$$

where we use the fact that the expected value of the reciprocal of an inverse-Wishart is α/β, or, in this case, $1/\theta$.

3.2.4 Invariance under transformations

Again, we define the reference prior as

$$p_\theta^*(\theta) = \arg\max_{p_\theta(\theta)} I(\theta, t)\,,$$

where t is a sufficient statistic and I is the mutual information. Let $\eta = h(\theta)$ for h monotonic and continuously differentiable. Similarly, define the reference prior

$$p_\eta^*(\eta) = \arg\max_{p_\eta(\eta)} I(\eta, t)\,.$$

We wish to show that

$$p_\eta^*(\eta) = p_\theta^*(h^{-1}(\eta)) \left| \frac{d}{d\eta} h^{-1}(\eta) \right| .$$

This can be demonstrated by noting that

$$\begin{aligned} I(\eta, t) &= \int p(\tilde{t}) \int p_\eta(\tilde{\eta}|\tilde{t}) \log\left(\frac{p_\eta(\tilde{\eta}|\tilde{t})}{p_\eta(\tilde{\eta})}\right) d\tilde{\eta}\, d\tilde{t} \\ &= \int p(\tilde{t}) \int \frac{p_\theta(h^{-1}(\tilde{\eta})|\tilde{t})}{|h'(\tilde{\theta})|} \log\left(\frac{p_\theta(h^{-1}(\tilde{\eta})|\tilde{t})|h'(\tilde{\theta})|}{p_\theta(h^{-1}(\tilde{\eta}))|h'(\tilde{\theta})|}\right) d\tilde{\eta}\, d\tilde{t} \\ &= \int p(\tilde{t}) \int p_\theta(\tilde{\theta}|\tilde{t}) \log\left(\frac{p_\theta(\tilde{\theta}|\tilde{t})}{p_\theta(\tilde{\theta})}\right) d\tilde{\theta}\, d\tilde{t} = I(\theta, t)\,, \end{aligned}$$

and hence the prior $p_\theta(\theta)$ maximizes $I(\theta, t)$ if and only if $p_\eta(\eta) = p_\theta(h^{-1}(\eta)) \left|\frac{d}{d\eta} h^{-1}(\eta)\right|$ maximizes $I(\eta, t)$.

3.2.5 Location family

For a given density p define the location family of distributions

$$\{p(y - \theta)\,,\ y, \theta \in \mathbb{R}\}\ .$$

Let $y \sim p(y-\theta)$ and define $x = y+a$ and $\eta = \theta+a$. Evidently, $x \sim p(x-\eta)$. Equality of reparameterized reference priors $p_\theta(\theta) = p_\eta(\eta)$ follows from the transformation invariance of the reference prior and the fact that $|h'(\theta)| = 1$. We may write this fact thus:

$$p_\eta(\eta) = p_\theta(\eta - a)$$

But these are reference priors for the exact same model, i.e.,

$$\{p(y - \theta)\,,\ y, \theta \in \mathbb{R}\} = \{p(x - \eta)\,,\ x, \eta \in \mathbb{R}\}\ ,$$

so $p_\eta = p_\theta$ and

$$p_\eta(\eta) = p_\eta(\eta - a)\,.$$

We conclude that the reference prior for the location of a location family of distributions is flat.

3.2.6 Scale family

Define the scale family of distributions

$$\left\{\frac{1}{\theta} p\left(\frac{y}{\theta}\right)\,,\ y, \theta \in \mathbb{R}^+\right\}$$

and the transformations $x = \log y$ and $\eta = \log \theta$. These result in a location family

$$\left\{p\left(\exp\left(x - \eta\right)\right)\,,\ x, \eta \in \mathbb{R}^+\right\}\ ,$$

which has a flat reference prior $p_\eta(\eta)$. But $p_\eta(\eta) = p_\theta(e^\eta)e^\eta = p_\theta(\theta)\theta$. It follows that $p_\theta(\theta) \propto \frac{1}{\theta}$.