Non-informative prior distributions

When there is no prior information/knowledge available for one or more parameters, or one wants to "let the data speak for themselves", it can be desirable to have a prior distribution that is "guaranteed" to not affect the posterior analysis.

↳ Then we use what we call "non-informative" (or "reference") prior distributions

Proper and improper distributions

Consider normal data, with known variance σ^2 and mean θ with a $N(\mu_0, \tau_0^2)$ prior. If $\tau_0^2 \to \infty$, then $p(\theta) \approx$ constant for $\theta \in (-\infty, \infty)$

and $p(\theta | y) \approx N(\theta | \bar{y}, \sigma^2/n)$ (prop. to the likelihood)

In this case, $p(\theta)$ is not a true distribution, since $\int_{-\infty}^{\infty} p(\theta) d\theta = \infty$, and we call it an <u>improper prior distribution</u>. A proper prior distribution has a density that integrates to 1 (if it integrates to a positive, finite value, then it is called an unnormalised, proper density, and can be normalised to integrate to 1)

- In the example above the improper prior distr. leads to a proper posterior distr. for θ, but this is not always the case!

sep 29-12:16

· Consider normal data with known mean, and $p(\sigma^2) = \text{Inv}\chi^2(\nu_0, \sigma_0^2)$

Letting $\nu_0 \to 0$ yields $p(\sigma^2) \to \propto \frac{1}{\sigma^2}$ and

$$p(\sigma^2 | y) \approx \text{Inv}\chi^2\left(n, \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2\right)$$

Proper posterior, but improper prior: $\int_0^\infty \frac{1}{\sigma^2} d\sigma^2 = \infty$

<u>NB</u> Example:

Binomial data with n successes in n trials, θ is the proportion of successes. If we $p(\theta) = \text{Beta}(a,b)$, let $a \to 0$, $b \to 0$, then

$p(\theta) \propto \theta^{-1} \cdot (1-\theta)^{-1}$ – improper prior

$p(\theta | y) \propto \theta^{n-1}(1-\theta)^{-1}$ – an improper posterior distribution!

$$\int_0^1 p(\theta | y) d\theta = \infty$$

- An improper prior distribution is only an approximation which is sometimes convenient but should be used with care. If it results in an improper posterior, it cannot be used!

sep 29-12:32

Finding appropriate non-informative prior distribution

Challenging: no clear choice

↳ E.g. you think a flat/uniform prior is appropriate, but if it is flat for one parametrization, it might not be for another

Transformation of variables

Suppose $p_\theta(\theta)$ is the (prior) density for the parameter θ. Consider the transformed variable $\phi = h(\theta)$, then the corresponding (prior) density for ϕ is

$$p_\phi(\phi) = p_\theta(\theta) \cdot \left| \frac{d\theta}{d\phi} \right| = p_\theta(\theta) \cdot |h'(\theta)|^{-1}$$

Example: Consider the parameter σ^2, with prior $p_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}$ (Improper)

Consider the alternative parametrization $\phi = \log \sigma = h(\sigma)$

Then $p_\phi(\phi) \propto \frac{1}{\sigma^2} \cdot \left| \frac{d \log \sigma}{d \sigma^2} \right|^{-1} \propto 1$

Hence, the prior is $\propto \frac{1}{\sigma^2}$ for σ^2, while it is flat/uniform for $\log \sigma$.
One parametrization is not "more correct" than the other.

sep 29-12:43

Jeffrey's principle

The choice of prior should be invariant to transformation, so that all choices of parametrization give the same model/results.

The Jeffrey's prior for the scalar parameter θ has this property:

$$p(\theta) \propto [J(\theta)]^{1/2}, \quad \text{where } J(\theta) \text{ is the \underline{Fisher information} for } \theta$$

$$J(\theta) = E\left[\left(\frac{d \log p(y|\theta)}{d\theta} \right)^2 \Big| \theta \right] = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} \Big| \theta \right]$$

Can be extended to multiparameter models, but with more controversial results.

Proof of invariance

Consider $\phi = h(\theta)$. We have $p(\theta) = [J(\theta)]^{1/2}$

Now $J(\phi) = -E\left[\frac{d^2 \log p(y|\phi)}{d\phi^2} \Big| \phi \right] = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} \left(|h'(\theta)|^{-1} \right)^2 \Big| \phi \right] = \left| \frac{d\theta}{d\phi} \right|^2 \cdot J(\theta)$

and hence $(J(\phi))^{1/2} = J(\phi) \cdot \left| \frac{d\theta}{d\phi} \right|$

sep 29-12:54

Example of Jeffrey's prior

n y_i iid $N(\theta, \sigma^2)$, σ^2 known, want to find the Jeffrey's prior for θ:

$$p(y|\theta) \propto \exp\{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2\}$$, $\log p(y|\theta) = -\frac{n}{2\sigma^2}(\theta-\bar{y})^2$

$$\frac{d \log p(y|\theta)}{d\theta} = -\frac{2n}{2\sigma^2}(\theta-\bar{y})$$

$$\frac{d^2 \log p(y|\theta)}{d\theta^2} = -\text{ constant w.r.t } \theta$$

The Jeffrey's prior for θ is $\propto 1$

sep 29-13:16

Non-informative prior for the normal data model with unknown mean, unknown variance:

$$y_i \sim N(\mu, \sigma^2) \quad , i = 1, \dots, n$$

Likelihood function:

$$p(y|\mu,\sigma^2) \propto \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right)\}$$

Non-informative prior

$$p(\mu) \propto 1 \quad , \quad p(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

The joint posterior distribution is:

$$p(\mu, \sigma^2|y) \propto \sigma^{-(n+2)} \exp\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right)\}$$

Marginal posterior for σ^2:

$$p(\sigma^2|y) = \int_{-\infty}^{\infty} p(\mu,\sigma^2|y)\, d\mu = \text{Inv-}\chi^2\left(n-1, \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$$

sep 29-13:23

Marginal posterior distr. for μ:

$$p(\mu \mid y) = \int_0^\infty p(\mu, \sigma^2 \mid y)\, d\sigma^2 = \dots = t_{n-1}\left(\bar{y}, \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n \cdot (n-1)}\right)$$

Weakly informative priors

- Prior distributions that are intended to be only weakly informative, but that are proper distributions
- Sometimes, some (weak) information is needed to regularize the posterior distribution
- It may be just including knowledge on the parameter space, for example $p(\theta) \propto 1$ for $\theta \in [1, 1000]$
- Always important: Have non-zero prior density for all possible values of θ!

sep 29-13:29

Multivariate normal model

Example

Model the joint distribution of the height and weight of 15 women.

Data $y_i = (y_{i1}, y_{i2})^T$, $i = 1, \dots, 15$

y_{i1}: Height of woman i (cm)

y_{i2}: Weight of woman i (kg)

Assumed sampling distr.:

$$\underset{2\times1}{y_i} \sim N(\underset{2\times1}{\mu}, \underset{2\times2}{\Sigma})$$

, unknown μ and Σ. Need a prior (μ, Σ)

General model for n iid obs $\underset{d\times1}{y_1}, \dots, \underset{d\times1}{y_n}$, each of length d

$$\underset{d\times1}{y_i} \sim N(\underset{d\times1}{\mu}, \underset{d\times d}{\Sigma})$$

Σ — Symmetric and positive definite covariance matrix

The likelihood is:

$$p(y_1, \dots, y_n \mid \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right\}$$

sep 29-13:46

Conjugate prior (generalization of the univariate case):

The multivariate version of the scaled inverse χ^2 distribution is called the inverse Wishart distribution, with degrees of freedom parameter ν and scale matrix Λ (symmetric, positive definite)

$\nu > d-1$ for a proper distribution

The conjugate prior distribution for (μ, Σ) is the normal-inverse-Wishart with hyperparameters $(\mu_0, \Lambda_0/\kappa_0 ; \nu_0, \Lambda_0)$

$$\underset{d\times d}{\Sigma} \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$$

$$\underset{d\times 1}{\mu} \mid \Sigma \sim N(\underset{d\times 1}{\mu_0}, \Sigma/\kappa_0)$$

$\nu_0 > d-1$ for a proper prior distr.

with density

$$p(\mu, \Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\{-\tfrac{1}{2}\underset{\text{trace}}{tr}(\Lambda_0 \Sigma^{-1}) - \tfrac{\kappa_0}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0)\}$$

sep 29-13:53

Now, calculating $p(\mu, \Sigma \mid y) \propto p(\mu, \Sigma) \cdot p(y \mid \mu, \Sigma)$ results in a posterior density of the same family (Normal-Inverse-Wishart) with parameters (a multivariate generalization of the univariate results):

$$\underset{d\times 1}{\mu_n} = \frac{\kappa_0}{\kappa_0+n}\underset{d\times 1}{\mu_0} + \frac{n}{\kappa_0+n}\underset{d\times 1}{\bar{y}}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\underset{d\times d}{\Lambda_n} = \underset{d\times d}{\Lambda_0} + \underset{d\times d}{S} + \frac{\kappa_0 n}{\kappa_0+n}\underset{d\times d}{(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T}$$

where S is the sum of squares about the sample mean:

$$\underset{d\times d}{S} = \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^T$$

Further generalizations of the univariate results:

$$\mu \mid y_1, \dots, y_n \sim \text{Multivariate } t_{\nu_n-d+1}\left(\mu_n, \frac{\Lambda_n}{\kappa_n(\nu_n-d+1)}\right)$$

$$\Sigma \mid y_1, \dots, y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$$

$$\mu \mid \Sigma, y_1, \dots, y_n \sim N(\mu_n, \Sigma/\kappa_n)$$

sep 29-14:17

Example of a (improper) non-informative prior distribution

Letting $\kappa_0 \to 0$, $\nu_0 \to -1$, $|\Lambda_0| \to 0$, the above prior becomes

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

This is the multivariate Jeffreys prior for this model.

If $n > d$, this results in a proper posterior distribution given by the formulas for the proper normal-inverse-Wishart prior:

$$\mu \mid y_1, \dots, y_n \sim t_{n-d}(\bar{y}, S/(n(n-d)))$$

$$\Sigma \mid y_1, \dots, y_n \sim \text{Inv-Wishart}_{n-1}(S^{-1})$$

$$\mu \mid \Sigma, y \sim N(\bar{y}, \Sigma/n)$$

Height/weight example continued

Proper prior: $\nu_0 = 2$, $\Lambda_0 = I$, $\mu_0 = (160, 65)^T$, $\kappa_0 = 0.1$

sep 29-14:26