Non-Informative Priors:

A non-informative prior (an objective prior) is a prior which is somehow automatic, reflecting the lack of any initial knowledge about the parameter. It may have no probabilistic interpretation and so does not have to be a valid probability distribution. Non-informative priors can be used when little no reliable information is available.

(1) Uniform Priors:

Definition: The uniform prior or flat prior is the prior $p(\theta) \propto 1$.

This is the obvious choise of lack of information; every value being equally likely. Under this prior,

$$P(\vartheta|x) = \frac{L(\vartheta, x)}{\int_{\Theta} L(\vartheta, x)\, d\vartheta}$$

which is well defined as long as

$$\int_{\Theta} L(\vartheta, x)\, d\vartheta < \infty.$$

Example: Let $X \sim \mathrm{Exp}(\vartheta)$ and $P(\vartheta) = 1$

The marginal likelihood is $\int_0^\infty e^{-\vartheta x} \vartheta\, d\vartheta$

which is finite for all $x > 0$, so

the posterior is well-defined.

Properties: Let $\eta = \log \vartheta$. Then the

prior for η is $\quad \hookrightarrow \vartheta = e^{\eta}$

$$\tilde{P}(\eta) = P(\vartheta(\eta)) \frac{d\vartheta}{d\eta} = \frac{d\vartheta}{d\eta} = 1 \cdot e^{\eta} \neq 1$$

After reparametrisation the prior is not flat anymore. In fact, as a prior in η, $\tilde{p}$ is very informative (large values are more likely than small ones)

(2) Jeffrey's Prior:

We need a prior which doesnot depend on the parametrisation.

Definition: In one-dimensional case Jeffrey's prior is given by:

$$p(\vartheta) \propto \left(I(\vartheta)\right)^{1/2}$$

where $I(\vartheta) = E_\vartheta\left(-\frac{\partial^2}{\partial\theta^2} \ell(\vartheta, x)\right)$ is the Fisher Inf.

Remark: If $\vartheta = g(\psi)$ for some one-to-one differentiable function g then the reparametrised prior is

$$\tilde{p}(\psi) \propto p(g(\psi)) \cdot |g'(\psi)|.$$

Recall, that $I(\psi) = (g'(\psi))^2 I_\vartheta$

so $\sqrt{I_\psi} = \sqrt{I_\vartheta}\,|g'(\psi)|$.

Hence, $\tilde{p}(\psi) \propto \sqrt{I_\psi}$

⚠ Jeffrey's prior is invariant under reparametrisation.

Definition: The k-dimensional Jeffrey's prior is given by

$$p(\vartheta) \propto |I_\vartheta|^{1/2}$$

where $|I_\vartheta| = \det I_\vartheta$, I_ϑ is the Fisher Information matrix, so under the standard regularity assumptions

$$(I_\vartheta)_{ij} = -\mathbb{E}_\vartheta\left(\frac{\partial^2}{\partial\vartheta_i\,\partial\vartheta_j}\ell(\vartheta, x)\right)$$

Example: Suppose $X \sim \mathrm{Poi}(\vartheta)$ so that $f(x,\vartheta) = \dfrac{e^{-\vartheta} \cdot \vartheta^x}{x!}$ for $x = 0, 1, 2, \dots$

The jeffrey's prior is $\left(\log f(x;\vartheta) = x \log \vartheta - \vartheta - \log x! \right)$

$$P(\vartheta) \propto \left(I_X(\vartheta) \right)^{1/2}$$

$$= \left(\mathbb{E} \left(\frac{X}{\vartheta} - 1 \right)^2 \right)^{1/2}$$

$$= \left(\sum_{x=0}^{\infty} f(x,\vartheta) \left(\frac{x-\vartheta}{\vartheta} \right)^2 \right)^{1/2}$$

$$= \left(e^{-\vartheta} \sum_{x=0}^{\infty} \frac{\vartheta^x}{x!} \left(\frac{x^2}{\vartheta^2} - \frac{2x}{\vartheta} + 1 \right) \right)$$

$$= \left(\frac{1}{\vartheta^2} \, \mathbb{E} \left(\mathbb{E} (X - \vartheta)^2 \right) \right)^{1/2}$$

$$= \vartheta^{-1/2} .$$

__Note__: This is an improper prior.

(3) Maximum Entropy Prior:

This prior is inspired by information theory.

Definition: The entropy of a pdf p is defined as

$$H(p) = -\int_{\Theta} p(\theta) \log p(\theta) \, d\theta.$$

A maximum entropy probability distribution has entropy that is at least as great as that of all other members of a specified class of probability distributions. This ensures the least biased or the most non-informative choice, assuming no other knowledge is available.

This approach minimizes prior assumption or information

Example: Suppose we know the mean and variance of ϑ. Then the maximum entropy distribution is the Gaussian distribution.

$$P(\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\vartheta-\mu)^2/(2\sigma^2)}$$

Note: This is because Gaussian distribution has the highest entropy among all distributions on $\mathbb{R}$ with a given mean μ and variance σ^2.

Constraints
- $\int p(\vartheta)\, d\vartheta = 1$
- $\int \vartheta p(\vartheta)\, d\vartheta = \mu$
- $\int (\vartheta-\mu)^2 p(\vartheta)\, d\vartheta = \sigma^2$

We want to maximize entropy:

$$H(p) = -\int_{-\infty}^{+\infty} p(\vartheta) \log p(\vartheta)\, d\vartheta$$

subject to constraints.

Use lagrangian multipliers

$$\mathcal{L}(p) = -\int p(\vartheta) \log p(\vartheta) \, d\vartheta + \lambda_1 \int p(\vartheta) \, d\vartheta - 1$$

$$+ \lambda_2 \left(\int \vartheta \, p(\vartheta) \, d\vartheta - \mu \right)$$

$$+ \lambda_3 \left(\int (\vartheta - \mu)^2 \, p(\vartheta) \, d\vartheta - \sigma^2 \right)$$

$$\Rightarrow \frac{\partial L}{\partial p(\vartheta)} = -\log p(\vartheta) - 1 + \lambda_1 + \lambda_2 \vartheta + \lambda_3 (\vartheta - \mu)^2 = 0$$

Solve for $p(\vartheta)$

$$\Rightarrow p(\vartheta) = \exp\left(\lambda_2 \vartheta + \lambda_3 (\vartheta - \mu)^2 + \lambda_1 - 1 \right)$$

$$\Rightarrow p(\vartheta) = C \exp\left(\lambda_2 \vartheta + \lambda_3 (\vartheta - \mu)^2 \right)$$

if $\lambda_3 < 0$ then the exponent is concave $p(\vartheta)$ integrates to 1

choosing $\lambda_2 = 0$ $\lambda_3 = -\frac{1}{2\sigma^2}$

$$p(\vartheta) = c \cdot \exp\left(-\frac{(\vartheta-\mu)^2}{2\sigma^2}\right)$$

we normalize

$$\int_{-\infty}^{+\infty} c \cdot \exp\left(-\frac{(\vartheta-\mu)^2}{2\sigma^2}\right) d\vartheta = 1$$

we solve for c

this gives $c = 1/\sqrt{2\pi\sigma^2}$.

Theorem: Let

$$p(\vartheta) = \exp\left(\sum_{i=1}^{K} \lambda_i T_i(\vartheta) - B(\lambda)\right), \forall \vartheta \in \Theta.$$

be a probability density function

and suppose that

$$\int T_i(x)\, p(\theta)\, dx = t_i \quad \text{for } i=1,\dots,k \quad (*)$$

Then p uniquely maximizes $H(p)$ among all densities satisfying the constraint.

<u>Proof</u>: Let Π be the class of distributions satisfying $(*)$.

Recall that for 2 distributions $\Pi_1 << \Pi_2$ the Kullback-Leibler divergence $KL(\Pi_1 \| \Pi_2)$ is defined through:

$$KL(\Pi_1 \| \Pi_2) = \int \Pi_1(dx) \log\left(\frac{d\Pi_1}{d\Pi_2}\right)$$

where $\frac{d\Pi_1}{d\Pi_2}$ is the radon Nikodym derivative. If Π_1 is not absolute continuous w.r.t Π_2 we set $KL(\Pi_1 \| \Pi_2) = +\infty$. It is a simple application of Jensen's inequality to check that $KL(\Pi_1 \| \Pi_2) \geq 0$

Let p' be an element on Π

$$H(p') = -\int p'(x) \log p'(x)\, dx$$

$$= -\int p'(x) \log\left(\frac{p'(x)}{p(x)}\right) dx$$

$$- \int p'(x) \log p(x)\, dx$$

$$= -KL(p' \| p) - \int p'(x) \log p(x)\, dx$$

$$= -KL(p' \| p) - \int p'(x)\left(\sum_{i=1}^{p} \lambda_i T_i(x)\right) dx + B(\lambda)$$

and since $p' \in \Pi$, it satisfies the same moment constraints as p

So,

$$\int p' \log p = \sum_{i=1}^{k} \lambda_i t_i - B(\lambda)$$

$$= \int \pi \log \pi$$

Therefore

$$H(p') = -KL(p' \| p) + H(p)$$

$$\Rightarrow \quad H(p) - H(p') = KL(p' \| p) \geq 0$$

Equality holds only when

$p' = p$, so p is uniquely maximizes the entropy under the given constraint (*) $\square$

<u>Note</u>: Theorem confirms that the maximum entropy distribution under moment constraint is the exponential family distribution.

<u>Example</u>: In the Gaussian example:

$\mathbb{E}(T_1(\vartheta)) = \mu \qquad \mathbb{E}(T_2(\vartheta)) = \sigma^2$

where $T_1(\vartheta) = \vartheta \qquad T_2(\vartheta) = (\vartheta - \mu)^2$

By the previous theorem the maximum entropy prior is of the form:

$$p(\vartheta) \propto \exp\left(\lambda_1 \vartheta + \lambda_2 (\vartheta-\mu)^2\right)$$

The 2 constraints then imply that $\lambda_1 = 0$ $\lambda_2 = -\dfrac{1}{2\sigma^2}$.