This concludes our introduction to conjugate priors. Conjugate priors are a matter of convenience, easy to implement and as such widely used in software implementations. They have some nice properties, in particular they are optimal asymptotically. They are often used in applications, when one lacks prior knowledge. Using conjugate priors, only needs to assess the prior parameters.

2 Jeffreys priors

Though conjugate priors are computationally nice, objective Bayesians instead prefer priors which do not strongly influence the posterior distribution. Such a prior is called an *uninformative prior*.

This is a hard problem, and a number of things we might try are not appropriate. The historical approach, followed by Laplace and Bayes, was to assign flat priors. This prior seems reasonably uninformative. We do not know where the actual value lies in the parameter space, so we might as well consider all values equiprobable. This prior however is not invariant. Consider for example a binomial distribution $X \sim \text{Binom}(n, \theta)$ in which we want to put a prior on θ. We know that θ lies between 0 and 1. The flat prior on θ is the uniform distribution: $\pi(\theta) = 1$. Since θ lies between 0 and 1, we can use a new parametrization using the log-odds ratio: $\rho = \log \frac{\theta}{1-\theta}$. This is a perfectly valid parametrization, and a natural one if we want to map θ to the full scale of the reals. Under this parametrization the prior distribution $\pi(\rho)$ is not flat anymore. This example shows a prior that is uninformative in one parametrization, but becomes informative through a change of variables.

This becomes more problematic in higher dimensions: the uniform prior in large dimension does not integrate anymore. In addition, the flat prior becomes very informative: it tells that most of the probability mass lies at $+\infty$, far from the origin. If instead one considers a high-dimensional Gaussian distribution $X \sim \mathcal{N}(0, 1)$, most of the mass is concentrated in a (high dimensional) unit sphere centered at the origin.

Faced with these issues, we see that flat priors and uninformative priors raise mathematical and philosophical problems. These examples show that finding prior distributions that have a minimal impact as possible on the data raises deep practical issues.

We first consider some special cases in one dimension, then consider the general case.

2.1 Examples

2.1.1 The example of an uninformative location prior

Consider the case where we have a location parameter: a probability distribution over a variable X of density $f(X - \theta)$ where θ is a *location parameter* that we endow with a prior. A candidate for a prior would be $\pi(\theta) \propto 1$. If θ lies in an interval, we can consider the uniform distribution as a prior estimate. If θ can take any value in $\mathbb{R}$, the flat prior is not a probability density because it does not integrate. Such a prior is called an *improper prior*. It expresses our state of ignorance (hence the flat prior) and can be defined as the limit of a proper prior.

2.1.2 The example of an uninformative scaling prior

Consider a density factor θ:

$$f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right) \tag{11}$$

The $\frac{1}{\theta}$ in the front ensures that f_θ still integrates to 1. The prior on θ should be invariant to rescaling by any arbitrary positive constant, i.e.:

$$\pi(\theta) = \frac{1}{c}\pi\left(\frac{\theta}{c}\right) \tag{12}$$

for all $c > 0$. This means if we rescale our variable, the prior will not change, that is, the prior does not give any information when we rescale the variable. The previous relation is a functional equation that admits a single solution for π (up to a scaling factor):

$$\pi(\theta) \propto \frac{1}{\theta} \tag{13}$$

Note how it is an improper prior because it does not have a finite integral. This is the uninformative scale prior.

We now look for a space in which this prior transforms into a flat prior. Consider the change of variable $\rho = \log\theta$ (or equivalently $\theta = e^\rho$). Then the probability density function in the new parametrization is:

$$\begin{aligned} \pi(\rho) &= \pi(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\rho}\right| \\ &\propto e^{-\rho}e^{\rho} = 1 \end{aligned}$$

Thus our scale invariant prior is actually a flat prior in the log scale. It treats equally any order of magnitude.

This prior has another derivation based on the (proper) conjugate prior of the variance of the Gaussian. We saw that the conjugate prior for the variance of the Gaussian is the inverse gamma:

$$p\left(\sigma^2|\alpha,\beta\right) \propto \left(\sigma^2\right)^{-(\alpha+1)} e^{-\beta/\sigma^2} \tag{14}$$

which is parametrized by two parameters α and β. The parameter α can be interpreted as the number of observations and β is some inverse concentration of the parameter in a certain region. It then makes sense to take α and β to the 0 limit, as if we had no prior information. When we do that we obtain:

$$p\left(\sigma^2|\alpha,\beta\right) \propto \frac{1}{\sigma^2} \tag{15}$$

which is the same prior as the one we derived.

2.2 Jeffreys priors

Uninformative priors we have seen so far are appealing because they are flat priors in some meaningful parametrization. Jeffreys priors are a generalization of these ideas, and can deliver a broad range of priors that incorporates these special cases. They are quite reasonable in one dimension. They are based on a principle of invariance: one should be able to apply these priors to certain situations, apply a change of variable, and still get the same answer. Suppose we are provided with some model and some data, i.e. with a likelihood function $p(x|\theta)$. One should be able to manipulate the likelihood and get a prior on θ, from the likelihood only. Note how this approach goes contrary to the subjective Bayesian frame of mind, in which one first chooses a prior on then θ and then applies it to the likelihood to derive the posterior.

The answer in the one-dimensional case is:

$$\pi_{\mathrm{J}}(\theta) \propto \mathbf{I}(\theta)^{1/2} \tag{16}$$

in which $\mathbf{I}$ is the *Fisher information*, defined when θ is unidimensional by the second derivative of the log likelihood:

$$\mathbf{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\mathrm{d}^2 \log p(X|\theta)}{\mathrm{d}\theta^2}\right] \tag{17}$$

This is an integral over the values of X with keeping θ fixed (the expectation in the frequentist sense). Using the maximum likelihood principle, the best parameter θ cancels the first derivative of the log likelihood, and the second derivative gives the curvature of the likelihood around the MLE.

$\frac{\mathrm{d}^2 \log p(X|\theta)}{\mathrm{d}\theta^2}$ is a random variable over X and $\mathbb{E}_\theta$ denotes the fact that we are integrating with respect to the distribution f_θ indexed by the (fixed) variable θ: $f_\theta(X) = p(X|\theta)$. The Fisher information is locally concave around the MLE, globally concave for the exponential family, but not globally concave for all distributions. This is not the case for example for mixture models.

We check that it works for a case we already saw: the Gaussian with fixed variance $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ for which we want to get prior on the location parameter μ. The likelihood is:

$$p(X|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}(X-\mu)^2\right) \tag{18}$$

thus when we take the second derivative with respect to μ:

$$\frac{\mathrm{d}^2 \log p(X|\mu)}{\mathrm{d}\mu^2} = -\frac{1}{\sigma^2} \tag{19}$$

which is a constant with respect to the random variable X and μ, so when we take the expectation, we get the flat prior we obtained before:

$$\mathbf{I}(\mu) \propto 1 \tag{20}$$

Now that we saw the answer, here are some explanations as to where the result comes from. Let us define a new parameter $\phi = h(\theta)$ as a reparametrization. If we calculate π_{J} with respect to the variable θ and then transform variables, this will give a prior π on ϕ by the change of variable formula. The question is thus to check if this prior $\pi(\phi)$ is indeed the Jeffreys prior $\pi_{\mathrm{J}}(\phi)$ that we would have computed in the first place by using the variable ϕ. We apply Jeffreys' principle in the ϕ space by using the chain rule and reexpress in terms of θ:

$$\begin{aligned}
\mathbf{I}(\phi) &= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(X|\phi)}{\mathrm{d}\phi^2}\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(X|\theta)}{\mathrm{d}\theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + \frac{\mathrm{d}\log p(X|\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(X|\theta)}{\mathrm{d}\theta^2}\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + \mathbb{E}\left[\frac{\mathrm{d}\log p(X|\theta)}{\mathrm{d}\theta}\right]\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}
\end{aligned}$$

The previous formulas are simply in application of the chain rule. We know:

$$\mathbb{E}\left[\frac{\mathrm{d}\log p(X|\theta)}{\mathrm{d}\theta}\right] = 0 \tag{21}$$

One way to see this fact is to use the total probability:

$$\forall\theta, \int p(X|\theta)\, dX = 1 \tag{22}$$

Assuming sufficient regularity, when we take the derivative with respect to θ:

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int p\left(X|\theta\right)dX \\ &= \int \frac{\mathrm{d}p\left(X|\theta\right)}{\mathrm{d}\theta}\frac{p\left(X|\theta\right)}{p\left(X|\theta\right)}\mathrm{d}X \\ &= \int \left[\frac{\mathrm{d}p\left(X|\theta\right)}{\mathrm{d}\theta}\frac{1}{p\left(X|\theta\right)}\right]p\left(X|\theta\right)\mathrm{d}X \\ &= \int \left[\frac{\mathrm{d}\log p\left(X|\theta\right)}{\mathrm{d}\theta}\right]p\left(X|\theta\right)\mathrm{d}X \\ &= \mathbb{E}\left[\frac{\mathrm{d}\log p\left(X|\theta\right)}{\mathrm{d}\theta}\right] \end{aligned}$$

This is formally proved using the dominated convergence theorem on distributions. Using this result, taking the expectation over X with θ fixed is equivalent to take the expectation with ϕ fixed, so we get:

$$\mathbf{I}\left(\phi\right) = \mathbf{I}\left(\theta\right)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 \tag{23}$$

and by taking the square root:

$$\sqrt{\mathbf{I}\left(\phi\right)} = \sqrt{\mathbf{I}\left(\theta\right)}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right| \tag{24}$$

By the change of variable formula, this shows that the Jeffreys prior $\pi_{\mathrm{J}}\left(\theta\right) = \sqrt{\mathbf{I}\left(\theta\right)}$ is invariant to a change of variable.

1.2 Limitations of Jeffreys priors

Jeffreys priors work well for single parameter models, but not for models with multidimensional parameters. By analogy with the one-dimensional case, one might construct a naive Jeffreys prior as the joint density:

$$\pi_J(\theta) = |I(\theta)|^{1/2}$$

where the Fisher information *matrix* is given by:

$$I(\theta)_{ij} = -\mathbb{E}_\theta \left[\frac{\partial^2 \log p(X|\theta)}{\partial\theta_i \partial\theta_j} \right]$$

Let's see what happens when we apply a Jeffreys prior for θ to a multivariate Gaussian location model. Suppose $X \sim N(\theta, I)$ for some p-dimensional random vector X, and we are interested in performing inference on $||\theta||^2$. In this case the Jeffreys prior for θ is flat. It turns out that the posterior has the form of a noncentral χ^2 distribution with p degrees of freedom. The posterior mean given one observation of X is $\mathbb{E}(||\theta||^2|X) = ||X||^2 + p$. This is not a good estimate because it adds p to the square of the norm of X whereas we might normally want to shrink our estimate towards zero. By contrast, the minimum variance frequentist estimate of $||\theta||^2$ is $||X||^2 - p$.

Intuitively, a multidimensional flat prior carries a lot of information about the expected value of a parameter. Since most of the mass of a flat prior distribution is in a shell at infinite distance, it says that we expect the value of θ to lie at some extreme distance from the origin, which causes our estimate of the norm to be pushed further away from zero.

Example 2. Consider a naive Jeffreys prior for a two-parameter Gaussian: $X \sim N(\mu, \sigma^2)$, and let $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. We take derivatives to compute the Fisher information matrix:

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \begin{pmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^2} \\ \frac{2(X-\mu)}{\sigma^2} & \frac{3}{\sigma^4}(X-\mu)^2 - \frac{1}{\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix} \end{aligned}$$

since $\mathbb{E}_\theta(X-\mu) = 0$ and $\mathbb{E}_\theta(X-\mu)^2 = \sigma^2$. Therefore

$$\pi_J(\theta) = |I(\theta)|^{1/2} \propto \frac{1}{\sigma^2}.$$

Unfortunately, this prior turns out to have poor convergence properties.

Jeffreys himself proposed using the prior $\pi_J(\theta) \propto \frac{1}{\sigma}$, which is a product of the separate priors for μ and σ. This prior is better motivated and gives better results as well. It also turns out to be the same as the reference prior, which we will discuss next.

2 Reference Priors

Reference priors were proposed by Jose Bernardo in a 1979 paper, and further developed by Jim Berger and others from the 1980's through the present. They are credited with bringing about an "objective Bayesian renaissance"; an annual conference is now devoted to the objective Bayesian approach.

The idea behind reference priors is to formalize what exactly we mean by an "uninformative prior": it is a function that maximizes some measure of distance or divergence between the posterior and prior, as data observations are made. Any of several possible divergence measures can be chosen, for example the Kullback-Leibler divergence or the Hellinger distance. By maximizing the divergence, we allow the data to have the maximum effect on the posterior estimates.

For one dimensional parameters, it will turn out that reference priors and Jeffreys priors are equivalent. For multidimensional parameters, they differ.

One might ask, how can we choose a prior to maximize the divergence between the posterior and prior, without having seen the data first? Reference priors handle this by taking the *expectation* of the divergence, given a model distribution for the data. This sounds superficially like a frequentist approach - basing inference on "imagined" data. But once the prior is chosen based on some model, inference proceeds in a standard Bayesian fashion. (This contrasts with the frequentist approach, which continues to deal with imagined data even after seeing the real data!)

2.1 Reference priors and mutual information

Consider an inference problem in which we have data X parameterized by Θ, with sufficient statistic $T = T(X)$. We want to find a reference prior $p(\theta)$ that maximizes its K-L divergence from the posterior $p(\theta|t)$, averaged over the distribution of T. This K-L divergence is

$$\int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta$$

Its expectation over the distribution of T can be written:

$$\begin{aligned} I(\Theta, T) &= \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt \\ &= \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta)p(t)} d\theta dt \end{aligned}$$

This may be recognized as the mutual information between θ and t. Therefore, choosing a reference prior involves finding $p^*(\theta)$ that maximizes the mutual information:

$$p^*(\theta) = \arg\max_{p(\theta)} I(\Theta, T) \tag{3}$$

We note that defining reference priors in terms of mutual information implies that they are invariant under reparameterization, since the mutual information itself is invariant.

Solving equation (3) is a problem in the calculus of variations. In the next lecture we'll derive reference priors for a variety of common situations.