Applied Probabilistic Programming & Bayesian Machine Learning

§1-5: Choosing Priors

MIT IAP 2017 Max W. Shen

An impressive amount of theory has been developed for choosing theoretically "optimal" priors, notably including Jeffreys' prior for theoretically non-informative priors that are invariant under reparameterization and gives a posterior that maximizes the amount of information used from the data. However, many theoretically motivated priors are "improper", with unintegrable infinite probability densities, and only produce proper posterior distributions under certain circumstances which depend on the data. Checking that the posterior is proper involves some non-trivial analytical work.

This class is primarily motivated by applying probabilistic programming on real-life data, rather than statistical research, so we will present a toolbox of prior distributions that have been found to be useful in practice.

As long as you have enough data (which, for Bayesian machine learning, can be as small as dozens of data points), the model often becomes robust to prior assumptions, instead drawing much of its information from the data. Nevertheless, explicitly testing your model's sensitivity to assumptions baked into the prior is an important part of the modeling process.

The equations for various probability distributions can be daunting at first, but for the purpose of choosing priors, the primary characteristic of importance is simply the *shape of the distribution.*

1 The Normal Distribution

Thus far, we have been using a normal(0, 100) prior for weakly informative priors, because it resembles a uniform distribution in regions of parameter space that are likely to contain the true value (if not, it's suggested you normalize your data to make it scale-free), but isn't completely uniform as to avoid convergence issues arising from non-identifiability.

In general, placing a normal prior with mean 0 on a parameter p and using the posterior mode estimate (the MAP estimate, *maximum a posteriori*) is equivalent to minimizing the empirical risk with L2-regularization on p. For example, in regression, where our parameter p is our weight vector, there exist values for σ and σ_w such that the two following optimization problems have the same objective:

$$\begin{aligned} Y &\sim N(XW, \sigma) \\ W &\sim N(0, \sigma_w) \end{aligned} \tag{1}$$

$$W^* = arg\min_W (Y - XW)^2 + |W|^2 \tag{2}$$

However, the Bayesian approach in equation 1 provides the full posterior distribution rather than a point estimate of the values for W, which are particularly useful for quantifying uncertainty.

The L2-regularization connection can also be seen by noting that the MLE estimate of the location parameter (μ = mean) of a normal distribution given some data is the sample mean, which is also the optimal minimizer of the L2 distance to all the data ($mean(x) = arg\min_s \sum_i (x_i - s)^2$).

2 The Cauchy Distribution and t distribution

The Cauchy distribution is a commonly used weakly-informative prior because it has longer tails than a normal prior. Visually, the Cauchy distribution places less of its probability mass around the mean, instead spreading it out making its shape more similar to a uniform distribution.

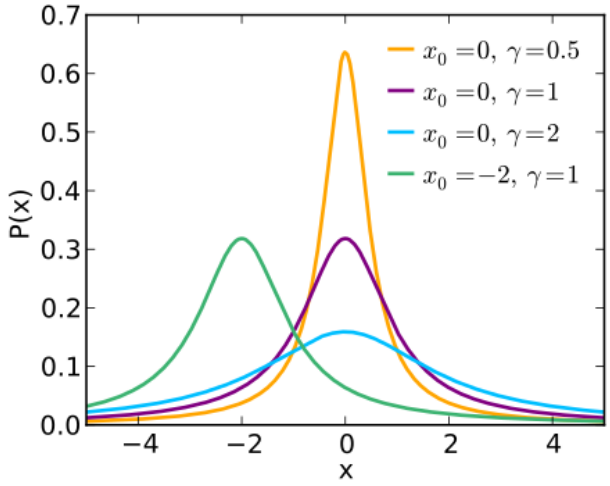

Figure 1: The Cauchy distribution with varying parameters

The blue curve in figure 1 shows the Cauchy distribution with parameters $x_0 = 0$ and $\gamma = 2$, centered at 0 with long tails. All probability density function figures presented below are taken from wikipedia, which sometimes uses alternate names for parameters. Here, x_0 is the location parameter, akin to μ for the normal distribution, and γ is the scale parameter, akin to σ for the normal distribution.

A half-Cauchy is a term that is also commonly used in Bayesian data analysis, and arises when placing a Cauchy prior placed over a parameter constrained to be non-negative such as standard deviation. See figure 2.

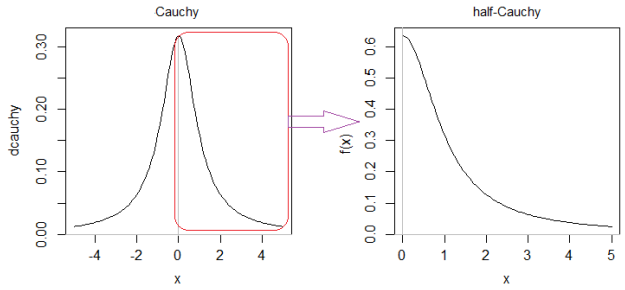

Figure 2: The Half-Cauchy for Non-Negative Parameters

Distribution	Notation	Parameters	Density function	Mean, variance, and mode
t	$\theta \sim t_\nu(\mu, \sigma^2)$ $p(\theta) = t_\nu(\theta \mid \mu, \sigma^2)$ t_ν is short for $t_\nu(0, 1)$	degrees of freedom $\nu > 0$ location μ scale $\sigma > 0$	$p(\theta) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}(1 + \frac{1}{\nu}(\frac{\theta-\mu}{\sigma})^2)^{-(\nu+1)/2}$	$\mathrm{E}(\theta) = \mu$, for $\nu > 1$ $\mathrm{var}(\theta) = \frac{\nu}{\nu-2}\sigma^2$, for $\nu > 2$ $\mathrm{mode}(\theta) = \mu$
Normal	$\theta \sim \mathrm{N}(\mu, \sigma^2)$ $p(\theta) = \mathrm{N}(\theta \mid \mu, \sigma^2)$	location μ scale $\sigma > 0$	$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)^2\right)$	$\mathrm{E}(\theta) = \mu$ $\mathrm{var}(\theta) = \sigma^2$ $\mathrm{mode}(\theta) = \mu$

Figure 3: t-distribution and Normal distribution

The t-distribution contains the normal distribution and Cauchy distribution as special cases.

The t-distribution shares a location parameter μ and non-negative scale parameter σ in common with the normal distribution, but has an additional parameter $\nu > 0$ for degrees of freedom. When $\nu = \infty$, the t-distribution is equivalent to the normal distribution, and when $\nu = 1$, the t-distribution is equivalent to a Cauchy distribution.

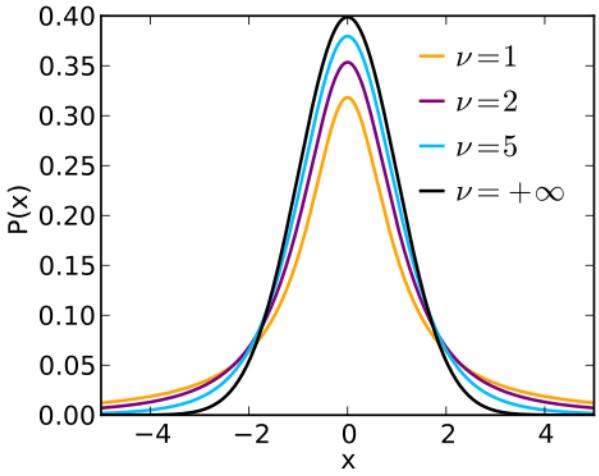

Figure 4: The t-distribution at varying ν

Figure 4 shows the relationship between the normal distribution, the t-distribution, and the Cauchy distribution, where $\nu = +\infty$ corresponds to the normal distribution, $\nu = 1$ corresponds to a Cauchy distribution, and the t-distribution allows an intermediate between the two extremes by setting ν to some positive value. Effectively, decreasing ν makes the prior less informative, and increasing ν makes the prior more informative.

3 The Double Exponential Distribution (Laplace Distribution)

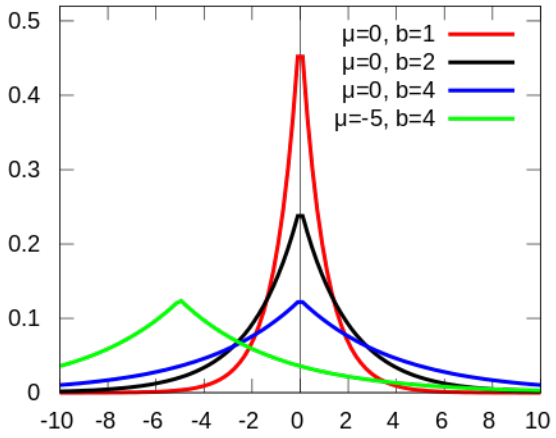

Figure 5: The Double Exponential (Laplace) distribution

Distribution	Notation	Parameters	Density function	Mean, variance, and mode
Laplace (double-exponential)	$\theta \sim \text{Laplace}(\mu, \sigma)$ $p(\theta) = \text{Laplace}(\theta \mid \mu, \sigma)$	location μ scale $\sigma > 0$	$p(\theta) = \frac{1}{2\sigma} \exp\left(-\frac{\lvert x-\mu \rvert}{\sigma}\right)$	$\text{E}(\theta) = \mu$ $\text{var}(\theta) = 2\sigma^2$ $\text{mode}(\theta) = \mu$

Figure 6: The Double Exponential (Laplace) distribution

Placing a double exponential prior over a parameter p, and using the posterior mode estimate (the MAP estimate, *maximum a posteriori*), is equivalent to applying L1-regularization to p. In the empirical risk minimization (ERM) setup, L1-regularization induces sparsity on parameters, driving them to exactly 0.

$$\begin{aligned} Y &\sim N(XW, \sigma) \\ W &\sim \text{double-exponential}(0, \sigma_w) \end{aligned} \tag{3}$$

$$W^* = arg\min_W (Y - XW)^2 + |W| \tag{4}$$

However, L1-regularization via the double exponential prior is not *really* Bayesian since it requires summarizing the posterior distribution using a point estimate - the posterior mode.

A topic known as "Bayesian variable selection" can induce sparsity in parameters by learning a mixture model over inclusion/exclusion of parameters. This is closely related to "spike and slab" priors. This will be discussed in a later section.

The MLE estimate of the location parameter (μ) of a Laplace distribution given some data is the sample median, which is also the minimizer of the L1 distance to all the data ($median(x) = arg\min_s \sum_i |x_i - s|$).

Exercise 1

Try a t-distribution prior, a Cauchy prior, and a double-exponential prior in Stan. In each case, fully specify the parameters of the priors instead of estimating it from data. For example, the t-distribution is parameterized by μ, σ, and ν. Incorporate these values into your model directly.

Exercise 2

For the double-exponential prior, set a prior over its scale parameter. Run inference to find the posterior distribution on the scale parameter.

We now have a distribution on the scale parameter, which we can draw from to get varying degrees of sparsity on our weight vector.

Simulate 1000 weight vectors from the double-exponential distribution. Zero out weight values that are within [-0.01, 0.01] or some other threshold around 0. From your 1000 simulated weight vectors, make a histogram of the number of zero elements.

This is one benefit of a fully Bayesian framework, allowing an extension to typical L1-regularization. Our model can reflect uncertainty in the number of weights that are set exactly to 0.

4 Conjugate Priors

The above priors can be interpreted via their shape to induce sparsity, regularize towards 0, or act as weakly-informative priors since they resemble uniform distributions.

Placing a conjugate prior over parameters enables the encoding of prior beliefs in a way that can be directly interpreted as having seen additional data.

As an example, consider the following model:

$$\begin{aligned} y &\sim binomial(n, q) \\ q &\sim beta(a, b) \end{aligned} \tag{5}$$

The beta distribution is the conjugate prior to the binomial distribution. Informally, this means that our prior distribution for q, the success rate, is a beta distribution, and after seeing data, our posterior distribution for q is also a beta distribution. This enables efficient analytical inference. In particular, before efficient approximate inference methods were developed (which form the backbone of probabilistic programming languages), conjugacy was a crucial tool for enabling inference to be possible at all.

Let the data y contain x successes and $n - x$ failures. Then:

$$\begin{aligned} p(q) &= beta(a, b) \\ p(q|y) &= beta(a + x, b + n - x) \end{aligned} \tag{6}$$

Here, the parameters a, b in the beta-distribution prior can be interpreted as having seen a successes and b failures before seeing the current data y. This can be thought of as "hallucinating data" or

having "pseudo-observations". It turns out this interpretation is valid for all conjugate priors - the conjugate prior parameters can be exactly interpreted as additional data.

Similarly, under the ordinary regression setting with $E[Y] \sim XW$ and normally distributed errors, the act of placing normal priors over the values of the weights has an equivalent correspondence to pseudo-observations. This reflects the fact that the normal distribution is a conjugate prior to itself.

Often, it can be difficult to pin down the exact effect of certain prior assumptions in Bayesian models. However, with conjugate priors, we can set a, b to reflect prior knowledge with the precise interpretation of having seen additional data. Unfortunately, this isn't always useful in practice, since prior knowledge generally has qualitative aspects as well as quantitative aspects.

Exercise 3

Implement the model described by equation 5. Try different values for a, b and see how they impact your posterior on q.

5 Priors for Covariance Matrices

The Inverse Wishart is the conjugate prior to the multivariate normal covariance matrix, and the LKJ distribution (developed in 2009!) is a distribution over positive-definite symmetric matrices with unit diagonals - that is, correlation matrices.

For more information, refer to Bayesian Data Analysis 3 by Andrew Gelman.

6 Nonparametric Priors

Nonparametric priors arise from a large subfield of Bayesian statistics, unsurprisingly known as nonparametric Bayesian statistics.

Nonparametric Bayes will be explored in greater depth in later sections, but we note here that nonparametric versions of parametric distributions (such as the Dirichlet distribution and Normal distribution) exist, known as the Dirichlet *process* and Gaussian *process*. A closely related process to the Dirichlet process is distinctively known as the Chinese Restaurant process.

One practical application of nonparametric Bayes is clustering where the number of clusters is learned directly from the data, rather than being specified beforehand. This is not possible with parametric Bayes, the domain we have been exploring thus far.

7 Bibliography and Additional Resources

Andrew Gelman has compiled some recommendations for priors in a huge variety of situations here: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Priors are discussed in Chapter 14 and 20 in Bayesian Data Analysis 3.

The Stan reference is available here: http://mc-stan.org/documentation/