STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 21 — Prior distributions

This lecture is a discussion of some topics on the interpretation and use of Bayesian priors and their influence on posterior inference.

21.1 Conjugate priors and improper priors

Last lecture, we saw two examples of conjugate priors:

1. If $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, then a conjugate prior for λ is $\text{Gamma}(\alpha, \beta)$, and the corresponding posterior given $X_1 = x_1, \ldots, X_n = x_n$ is $\text{Gamma}(s + \alpha, n + \beta)$ where $s = x_1 + \ldots + x_n$. A Bayesian estimate of λ is the posterior mean
$$\hat{\lambda} = \frac{s + \alpha}{n + \beta} = \frac{n}{n + \beta} \cdot \frac{s}{n} + \frac{\beta}{n + \beta} \cdot \frac{\alpha}{\beta}.$$

2. If $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$, then a conjugate prior for p is $\text{Beta}(\alpha, \beta)$, and the corresponding posterior given $X_1 = x_1, \ldots, X_n = x_n$ is $\text{Beta}(s + \alpha, n - s + \beta)$.[1] A Bayesian estimate of p is the posterior mean
$$\hat{p} = \frac{s + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta} \cdot \frac{s}{n} + \frac{\alpha + \beta}{n + \alpha + \beta} \cdot \frac{\alpha}{\alpha + \beta}.$$

In addition to being mathematically convenient, conjugate priors oftentimes have intuitive interpretations: In example 1 above, the posterior mean behaves as if we observed, a priori, β additional count observations that sum to α. β may be interpreted as an effective prior sample size and α/β as a prior mean, and the posterior mean is a weighted average of the prior mean and the data mean. In example 2 above, the posterior mean behaves as if we observed, a priori, α additional heads and β additional tails. $\alpha+\beta$ is an effective prior sample size, $\alpha/(\alpha + \beta)$ is a prior mean, and the posterior mean is again a weighted average of the prior mean and the data mean. These interpretations may serve as a guide for choosing the prior parameters α and β.

Sometimes it is convenient to use the formalism of Bayesian inference, but with an "uninformative prior" that does not actually impose prior knowledge, so that the resulting analysis is more objective. In both examples above, the priors are "uninformative" for the posterior mean when α and β are small. We may take this idea to the limit by considering $\alpha = \beta = 0$. As the PDF of the Gamma distribution is proportional to $x^{\alpha-1}e^{-\beta x}$ on $(0, \infty)$, the "PDF" for $\alpha = \beta = 0$ may be considered to be
$$f(x) \propto x^{-1}.$$

[1] We assumed $\alpha = \beta$ last lecture so that the prior is centered around 1/2, but the same calculation of the posterior distribution holds when $\alpha \neq \beta$.

Similarly, as the PDF of the Beta distribution is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on $(0,1)$, the "PDF" for $\alpha=\beta=0$ may be considered to be

$$f(x) \propto x^{-1}(1-x)^{-1}.$$

These are *not* real probability distributions: There is no such distribution as Gamma$(0,0)$, and $f(x) \propto x^{-1}$ does not actually describe a valid PDF on $(0,\infty)$, because $\int x^{-1}dx = \infty$ so that it is impossible to choose a normalizing constant to make this PDF integrate to 1. Similarly, there is no such distribution as Beta$(0,0)$, and $f(x) \propto x^{-1}(1-x)^{-1}$ does not describe a valid PDF on $(0,1)$. These types of priors are called **improper priors**.

Nonetheless, we may formally carry out Bayesian analysis using improper priors, and this oftentimes yields valid posterior distributions: In the Poisson example, we obtain the posterior PDF

$$f_{\Lambda|X}(\lambda|x_1,\ldots,x_n) \propto f_{X|\Lambda}(x_1,\ldots,x_n|\lambda)f_\Lambda(\lambda) \propto \lambda^s e^{-n\lambda} \times \lambda^{-1} = \lambda^{s-1}e^{-n\lambda},$$

which is the PDF of Gamma(s,n). In the Bernoulli example, we obtain the posterior PDF

$$f_{P|X}(p|x_1,\ldots,x_n) \propto f_{X|P}(x_1,\ldots,x_n|p)f_P(p) \propto p^s(1-p)^{n-s} \times p^{-1}(1-p)^{-1} = p^{s-1}(1-p)^{n-s-1},$$

which is the PDF of Beta$(s,n-s)$. These posterior distributions *are* real probability distributions (as long as $s>0$ in the Poisson example and $s, n-s>0$ in the Bernoulli example), and may be thought of as approximations to the posterior distributions that we would have obtained if we used proper priors with small but positive values of α and β.

21.2 Normal approximation for large n

For any fixed α,β in the above examples, as $n\to\infty$, the influence of the prior diminishes and the posterior mean becomes close to the MLE s/n. This is true more generally for parametric models satisfying mild regularity conditions, and in fact the posterior distribution is approximately a normal distribution centered at the MLE $\hat{\theta}$ with variance $\frac{1}{nI(\hat{\theta})}$ for large n, where $I(\theta)$ is the Fisher information. We sketch the argument for why this occurs:

Consider Bayesian inference applied with the prior $f_\Theta(\theta)$, for a parametric model $f_{X|\Theta}(x|\theta)$. Let $X_1,\ldots,X_n \overset{IID}{\sim} f_{X|\Theta}(x|\theta)$, and let

$$l(\theta) = \sum_{i=1}^n \log f_{X|\Theta}(x_i|\theta)$$

be the usual log-likelihood. Then the posterior distribution of Θ is given by

$$f_{\Theta|X}(\theta|x_1,\ldots,x_n) \propto f_{X|\Theta}(x_1,\ldots,x_n|\theta)f_\Theta(\theta) = \exp(l(\theta))f_\Theta(\theta).$$

Applying a second-order Taylor expansion of $l(\theta)$ around the MLE $\theta=\hat{\theta}$,

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}) + (\theta-\hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 l''(\hat{\theta}) \\ &\approx l(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^2 \cdot nI(\hat{\theta}), \end{aligned}$$

where the second equality follows because $l'(\hat{\theta}) = 0$ if $\hat{\theta}$ is the MLE, and $l''(\hat{\theta}) \approx -nI(\hat{\theta})$ for large n. Since $\hat{\theta}$ is a function of the data $x_1, \ldots, x_n$ and doesn't depend on θ, we may absorb $\exp(l(\hat{\theta}))$ into the proportionality constant to obtain

$$f_{\Theta|X}(\theta|x_1, \ldots, x_n) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right) f_\Theta(\theta).$$

For large n, the value of $\exp(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta}))$ is small unless θ is within order $1/\sqrt{n}$ distance from $\hat{\theta}$. In this region of θ, the prior $f_\Theta(\theta)$ is approximately constant and equal to $f_\Theta(\hat{\theta})$. Absorbing this constant also into the proportionality factor in $\propto$, we finally arrive at

$$f_{\Theta|X}(\theta|x_1, \ldots, x_n) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right).$$

This describes a normal distribution for Θ with mean $\hat{\theta}$ and variance $\frac{1}{nI(\hat{\theta})}$.

To summarize, the posterior mean of Θ is, for large n, approximately the MLE $\hat{\theta}$. Furthermore, a $100(1-\alpha)\%$ Bayesian credible interval is approximately given by $\hat{\theta} \pm z(\alpha/2)/\sqrt{nI(\hat{\theta})}$, which is exactly the $100(1-\alpha)\%$ Wald confidence interval for θ. In this sense, frequentist and Bayesian methods yield similar inferences for large n.

21.3 Prior distributions and average MSE

Last lecture we introduced the prior distribution for Θ as something that encodes our prior belief about its value. A different (but related) interpretation and motivation for the prior comes from the following considerations:

Let's return to the frequentist setting where we assume that there is a true parameter θ for a parametric model $\{f(x|\theta) : \theta \in \Omega\}$. Suppose we have two estimators for θ based on data $X_1, \ldots, X_n \sim f(x|\theta)$: $\hat{\theta}_1$ and $\hat{\theta}_2$. Which estimator is "better"? Without appealing to asymptotic (large n) arguments, one answer to this question is to compare their mean-squared-errors:

$$\text{MSE}_1(\theta) = \mathbb{E}_\theta[(\hat{\theta}_1 - \theta)^2] = \text{Variance of } \hat{\theta}_1 + (\text{Bias of } \hat{\theta}_1)^2$$
$$\text{MSE}_2(\theta) = \mathbb{E}_\theta[(\hat{\theta}_2 - \theta)^2] = \text{Variance of } \hat{\theta}_2 + (\text{Bias of } \hat{\theta}_2)^2$$

The estimator with smaller MSE is "better".

Unfortunately, the problem with this approach is that the MSEs might depend on the true parameter θ (hence why we have written MSE_1 and MSE_2 as functions of θ in the above), and neither may be uniformly better than the other. For example, suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$. Let $\hat{\theta}_1 = \bar{X}$; this is unbiased with variance $\frac{1}{n}$, so its MSE is $\frac{1}{n}$. Let $\hat{\theta}_2 \equiv 0$ be the constant estimator that always estimates θ by 0. This has bias $-\theta$ and variance 0, so its MSE is θ^2. If the true parameter θ happens to be close to 0—more specifically, if $|\theta|$ is less than $1/\sqrt{n}$—then $\hat{\theta}_2$ is "better", and otherwise $\hat{\theta}_1$ is "better".

To resolve this ambiguity, we might consider a weighted average MSE,

$$\int \text{MSE}(\theta) w(\theta) d\theta,$$

where $w(\theta)$ is a weight function over the parameter space such that $\int_\Omega w(\theta)d\theta = 1$, and find the estimator that minimizes this weighted average. This weighted average MSE is called the **Bayes risk**. Writing the expectation in the definition of the MSE as an integral, and letting $\mathbf{x}$ denote the data and $f(\mathbf{x}|\theta)$ denote the PDF of the data, we may write the Bayes risk of an estimator $\hat{\theta}$ as

$$\int \left(\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x}|\theta) d\mathbf{x} \right) w(\theta) d\theta.$$

Exchanging the order of integration, this is

$$\int \left(\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x}|\theta) w(\theta) d\theta \right) d\mathbf{x}.$$

In order to minimize the Bayes risk, for each possible value $\mathbf{x}$ of the observed data, $\hat{\theta}(\mathbf{x})$ should be defined so as to minimize

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x}|\theta) w(\theta) d\theta.$$

Let us now interpret $w(\theta)$ as a prior $f_\Theta(\theta)$ for the parameter Θ, and $f(\mathbf{x}|\theta)$ as the likelihood $f_{X|\Theta}(\mathbf{x}|\theta)$ given $\Theta = \theta$. Then

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x}|\theta) w(\theta) d\theta = \int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{X,\Theta}(\mathbf{x}, \theta) d\theta = f_X(\mathbf{x}) \int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{\Theta|X}(\theta|\mathbf{x}) d\theta.$$

So given the observed data $\mathbf{x}$, $\hat{\theta}(\mathbf{x})$ should be defined to minimize

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{\Theta|X}(\theta|\mathbf{x}) d\theta = \mathbb{E}[(\hat{\theta}(\mathbf{x}) - \Theta)^2],$$

where the expectation is with respect to the posterior distribution of Θ for the fixed and observed value of $\mathbf{x}$. For any random variable Y, $\mathbb{E}[(c - Y)^2]$ is minimized over c when $c = \mathbb{E}[Y]$—hence the minimizer $\hat{\theta}(\mathbf{x})$ of the above is exactly the posterior mean of Θ. We have thus arrived at the following conclusion:

The posterior mean of Θ for the prior $f_\Theta(\theta)$ is the estimator that minimizes the average mean-squared-error $\int \text{MSE}(\theta) f_\Theta(\theta) d\theta$.

Thus a Bayesian prior may be interpreted as the weighting of parameter values for which we wish to minimize the weighted-average mean-squared-error.