Stat210A: Theoretical Statistics **Lecture Date: October 7, 2014**

More on Bayesian Priors, Asymptotics

Lecturer: Michael I. Jordan *Scribe: Kellie Ottoboni*

1 Objective Bayesian Priors

The goal is to choose a prior that doesn't have a big influence on the posterior; most of the information to update the prior should come from the likelihood.

1.1 Examples

- Suppose θ is a location parameter, i.e. X has density $f(x-\theta), \theta \in \mathbf{R}$. If we put a prior π on θ, then we'd hope that any shift of $\pi(\theta)$ would give us the same information as the original distribution. This location invariance can be written

$$\pi(\theta) = \pi(\theta + c), \forall c \in \mathbf{R}$$

  Then $\pi(\theta) \propto 1$.

  Remark 1. $\pi(\theta) \propto 1$ doesn't integrate to 1 so it can't be a density. This is called an "improper prior". This is okay as long as $p(x|\theta)\pi(\theta)$ integrates to something finite and produces a proper posterior distribution. In practice, one needs to show this in order to justify using an improper prior.

- Suppose θ is a scale parameter, i.e. $p(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta})$. The prior $\pi(\theta)$ should be scale invariant or "scale free", i.e.

$$\pi(\theta) = \frac{1}{c}\pi(\frac{\theta}{c}) \forall c > 0$$

  Thus we should let $\pi(\theta) \propto 1/\theta$.

  Remark 2. This prior puts more weight on small values of θ. If we wanted a flat prior, we could use the transformation $\rho = \log\theta$. Then $\tilde{\pi}(\rho) = \pi(\theta)\frac{d\theta}{d\rho} = \frac{1}{e^\rho}e^\rho = 1$.

1.2 Another heuristic for finding objective priors

We often interpret the hyperparameters of a conjugate prior distribution as some data we knew about before we observed the actual data. Instead, we may want the prior to reflect our ignorance due to having not seen any data yet. We'd do this by taking limits of the hyperparameters.

For example, consider the Gaussian scale parameter σ^2. In lecture 11 we saw that the conjugate prior is $\sigma^2 \sim IG(\alpha, \beta)$ so $p(\sigma^2|\alpha,\beta) \propto (\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}$. The hyperparameter α can be thought of as the number of previous data points seen and β can be thought of as the spread of the prior data. Ignorance about prior

data would imply we had no prior data points and infinite spread, so we'd take $\alpha \to 0$ and $\beta \to \infty$ to get the ignorance prior

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

It's left as an exercise to find the ignorance prior for the Gaussian location parameter μ. Recall from lecture 11 that the conjugate prior is $N(\mu_0, \tau^2)$. The ignorance prior will be flat.

2 Jeffreys' Priors

2.1 Definition

Given a family $p(x|\theta), \theta \in \Omega$, the Jeffreys' prior for θ is $\pi_J(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information for θ.

$\pi_J(\theta)$ is invariant to transformations of the parameter.

Proof. Recall $I(\theta) = -E_\theta\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$. Consider the change of variables $\phi = h(\theta)$. Then

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

Note that the second term in the sum inside the expectation vanishes because the expected value of the score function is always 0.
The result follows because $\tilde{\pi}(\phi) \propto \sqrt{I(\phi)} = \sqrt{I(\theta)}|\frac{d\theta}{d\phi}| = \pi(\theta)|\frac{d\theta}{d\phi}|$. □

2.2 Example: Gaussian

Fix σ^2. Then the Fisher information for μ is $I(\mu) = 1$, so $\pi_J(\mu) \propto 1$.
Fix μ. Then the Fisher information for σ^2 is $I(\sigma^2) = \frac{1}{\sigma^4}$, so $\pi_J(\sigma^2) \propto \frac{1}{\sigma^2}$.
Notice that the Jeffreys' priors here are the same as the location/scale priors we derived earlier.

2.3 Example: Binomial

Let $X \sim \text{Binom}(n, \theta)$. The binomial likelihood is $p(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. Then the log-likelihood and its derivatives are

$$\log p(x|\theta) = \log \binom{n}{x} + x \log \theta + (n-x)\log(1-\theta)$$
$$\frac{\partial \log p(x|\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$
$$\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$

The Fisher information is $-E(\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}) = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$. Then the Jeffreys' prior is

$$\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

Remark 3. Notice that this is a $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ distribution. Compare this to the ignorance prior, which is $\text{Beta}(0,0)$, and the flat prior $\text{Beta}(1,1)$. The Jeffreys' prior has a nice intuitive interpretation in this case. Since the variance of the binomial distribution is $n\theta(1-\theta)$, the Jeffreys' prior is proportional to the inverse of the variance, so there is more mass at the extremes where the variance is small. Conversely, when the likelihood has larger variance and doesn't give us reliable information about θ, the Jeffreys' prior puts less mass there.

3 Reference Priors

Beyond heuristics and invariance principles, how do we incorporate the idea of ignorance into our choice of priors? We can think of ignorance as minimizing how much the change in distribution from prior to posterior depends on the prior distribution, i.e. we want the ratio $\frac{p(\theta|x)}{\pi(\theta)}$ to be large. One way to measure this is by taking the expectation of the log of this ratio. Define the functional

$$J(\pi) := \int p(x) \int p(\theta|x) \log \frac{p(\theta|x)}{\pi(\theta)} d\theta dx$$

The reference prior maximizes this:

$$\pi_R(\theta) := \text{argmax}_\pi J(\pi)$$

Remark 4. Notice that $\int p(\theta|x) \log \frac{p(\theta|x)}{\pi(\theta)} d\theta$ is the Kullback-Leibler divergence of the prior and the posterior. Then $J(\pi)$ is the expected Kullback-Leibler divergence over X; this is the mutual information of θ and X. The reference prior maximizes the mutual information.

When the parameter is one-dimensional, the reference prior is the same as the Jeffreys' prior. In higher dimensions, Jeffreys' priors don't work so well.

4 Asymptotics

We'd like to study what happens as $n \to \infty$. Asymptotic behavior is what justifies using maximum likelihood principles; MLE is usually not the best estimator when n is small. We'll use asymptotics to study when

MLE does turn out to be a good estimator.

4.1 Basic probability

- Convergence in probability: $Y_n \xrightarrow{P} Y$ if for all $\epsilon > 0$, $P(|Y_n - Y| \geq \epsilon) \to 0$.
- Almost sure convergence: For fixed $\omega \in \Omega$, does the sequence $Y_1(\omega), Y_2(\omega), \ldots$ converge to $Y(\omega)$? If the set of ω such that $Y_n(\omega) \to Y(\omega)$ has probability 1, then $Y_n \xrightarrow{\text{a.s.}} Y$ or "with probability 1".
- Markov's inequality: If $X \geq 0$, then $P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$.

  Proof. Notice that $X \geq \epsilon 1_{X \geq \epsilon}$. Take expectations of both sides. □
- Chebyshev's inequality: $P(|X| \geq a) \leq \frac{E(X^2)}{a^2}$. Prove this by using Markov's inequality on X^2.

4.2 Additional results

Theorem 5 (keener's Proposition 8.5, p. 130)**.** *If f is continuous at c and if $Y_n \xrightarrow{P} c$, then $f(Y_n) \xrightarrow{P} f(c)$.*

The idea is that continuous functions preserve convergence in probability.

Definition 6. The sequence δ_n of estimators is consistent for $g(\theta)$ if $\delta_n \xrightarrow{P} g(\theta)$, where convergence in probability is with respect to the probability measure P_θ on X.

Sufficient conditions for consistency are that $\text{Bias}(\delta_n) \to 0$ and $\text{Var}(\delta_n) \to 0$. We can prove this using the bias-variance decomposition of squared error loss, because convergence in mean square error implies convergence in probability.