Stat260: Bayesian Modeling and Inference **Lecture Date: February 22, 2010**

Reference Priors

Lecturer: Michael I. Jordan *Scribe: Steven Troxler and Wayne Lee*

In this lecture, we assume that $\theta \in \mathbb{R}$; in higher-dimensions, reference priors are defined in a sequential manner based on the parameter of primary interest, and we will address this construction in later lectures.

1 Motivating Reference Priors

Consider an inference scenario in which we have data X coming from a distribution $p(X|\theta)$ depending on a parameter θ, and suppose that $T(X)$ is a sufficient statistic for θ. This implies that $p(X|\theta)$ is in one-to-one correspondence with $p(T|\theta) := p(T(X)|\theta)$.

Our goal is to develop a non-informative prior for θ. One possible way to choose a non-informative prior is via information: we select the prior $\pi(\theta)$ to maximize mutual information between T^k and θ : take $\pi_1(\theta) = \text{argmax}_{p(\theta)} I_{p(\theta)}(\theta, T^k)$, where.

$$I_{p(\theta)}(\theta, T) = \int p(t) \underbrace{\left[\int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta \right]}_{KL(p(\theta|t), p(\theta)} d$$

The inner term in the double integral, $KL(p(\theta|t), p(\theta) := \int p(\theta|t^k) \log \frac{p(\theta|t^k)}{p(\theta)} d\theta$, is the Kullback-Leibler divergence between the posterior and prior when we observe a particular value $T = t$. The mutual information, then, is an average of Kullback-Leibler divergence with respect to the marginal distribution $p(t)$ of T. This idea is clever, but does not quite work as posed. Unfortunately, the problem of maximizing the mutual information between T and θ is often not analytically tractable. We might hope to solve the problem numerically, but this can be difficult.

An alternative is to use asymptotics, which often results in more analytically tractable expressions. To this end, we consider the following hypothetical situation: instead of observing $T(X)$ for just a single experiment, we repeat the experiment k independent (conditional on θ, which remains the same throughout) times, obtaining a vector T^k consisting of k independent copies of T. Instead of maximizing the mutual information just between T and θ, we maximize the information between the vector T^k and θ, obtaining $\pi_k(\theta) = \text{argmax}_{p(\theta)} I_{p(\theta)}(\theta, T^k)$, where.

$$I_{p(\theta)}(\theta, T^k) = \int p(t^k) \left[\int p(\theta|t^k) \log \frac{p(\theta|t^k)}{p(\theta)} d\theta \right] dt^k$$

We can obtain an analytically tractable uninformative prior by then taking $\pi(\theta) = \lim_{k \to \infty} \pi_k(\theta)$, where the 'limit' is in a loose sense that allows for improper priors. Bernardo Bernardo (2005) argues that not only does taking k to infinity give us a convenient way to compute uninformative priors, but it is also philosophically sense the 'right' thing to do. His argument is loosely that, when choosing a prior, we want to not only consider the information we obtain from a particular experiment, but the information we might obtain from many future experiments.

2 Computing Reference Priors and the Bernstein Von Mises Theorem

2.1 Solving the Mutual Information Problem

To find a more convenient form of π_k so that we may apply asymptotic theory, we rewrite $I_{p(\theta)}(\theta, T^k)$ as

$$\begin{aligned} I_{p(\theta)}(\theta, T^k) &= \int p(t^k) \left[\int p(\theta|t^k) \log \frac{p(\theta|t^k)}{p(\theta)} d\theta \right] dt^k \\ &= \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta, \end{aligned}$$

where

$$f_k(\theta) = \exp\{ \int p(t^k|\theta) \log p(\theta|t^k) dt^k \}.$$

Using a functional form of a Lagrangian to include the constraint that $\int p(\theta) = 1$, the problem becomes

$$\pi_k(\theta) = \sup_{p(\theta)} \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} + \lambda \left(\int p(\theta) d\theta - 1 \right) \Rightarrow p^*(\theta) \propto f_k(\theta) \tag{1}$$

This can be solved via methods of calculus of variations, and the solution is $\pi_k(\theta) \propto f_k(\theta)$. Although we will not go through the calculus-of-variations argument here, we will motivate this solution using the discrete case: if T and θ are both discrete, then the problem is of the form

$$\pi = \underset{p}{\operatorname{argmax}} \sum_i p_i \log \left(\frac{q_i}{p_i} \right) + \lambda \left(\sum_i p_i - 1 \right).$$

Taking partial derivatives with respect to p_j, we obtain

$$\begin{aligned} \frac{\partial}{\partial p_j} \left[\sum_i p_i \log \left(\frac{q_i}{p_i} \right) + \lambda (\sum_i p_i - 1) \right] &= \log(q_j/p_i) + p_j (q_j/p_j)^{-1} (-1 \cdot q_j/p_j^2) + \lambda \\ &= -1 - \log p_j + \log q_j + \lambda, \end{aligned}$$

and setting this partial derivative to zero we obtain

$$\log p_i = \log q_i + \lambda - 1 \Rightarrow p_i = q_i e^{\lambda - 1} \Rightarrow \pi = q.$$

2.2 The Bernstein Von Mises Theorem and an Asymptotic Solution

We have now reduced the problem to computing

$$f_k(\theta) = \exp \left\{ \int p(t^k|\theta) \log p(\theta|t^k) dt^k \right\}. \tag{2}$$

It is possible to obtain an analytical solution for the limit as $k \to \infty$ using the fact that $p(\theta|t^k)$ is asymptotically Gaussian, concentrated at the 'true' value θ_0 i.e. the value of θ such that $T_j^k \sim_{iid} p(t|\theta_0)$. This fact, which ensures similar behavior of Bayesian posteriors and frequentists sampling distributions as the sample size tends to infinity, is a consequence of the Bernstein Von Mises Theorem, sometimes called the "Bayesian Central Limit Theorem:"

Theorem 1. *Assume regularity conditions on the model which ensure asymptotic normality of an asymptotically efficient (in the frequentist sense) estimator $\tilde{\theta}_k$ and also assume that the prior satisfies regularity assumptions, in particular that it is near θ_0. If T^k denotes a vector of iid components T_j^k drawn from the distribution of $T|\theta_0$, then*

$$\left\| p(\theta|t^k) - N(\tilde{\theta}_k, I_k^{-1}(\theta_0)) \right\| \to 0, \tag{3}$$

where $I_k(\theta_0)$ denotes the Fisher information at θ_0 and the convergence denotes convergence in probability. Here, $\|\cdot\|$ denotes the total variation distance.

In general, any asymptotically efficient estimator $\tilde{\theta}_k$ is also asymptotically sufficient, so we may replace t_k in (3) with $\tilde{\theta}_k$, obtaining

$$\left\| p(\theta|\tilde{\theta}_k) - N\left(\tilde{\theta}_k, I_k^{-1}(\theta_0)\right) \right\| \to 0. \tag{4}$$

Also, using the density of a multivariate normal, we know that if $y \sim N\left(\tilde{\theta}_k, I_k^{-1}(\theta_0)\right)$, then

$$p(y) = \sqrt{I_k(\theta_0)} \exp\left(-\frac{I_k(\theta_0)}{2} (y - \tilde{\theta}_k)^2 \right)$$

By independence, we know that $I_k^{-1}(\theta_0)) = 1/k \cdot I^{-1}(\theta_0))$ with $I^{-1}(\theta_0)) := I_1^{-1}(\theta_0))$, so the preceding expression, combined with limit in (4), and the fact that $\tilde{\theta}_k$ is consistent, leads to the following approximate representation for large k :

$$p(\theta|\tilde{\theta}_k) \propto k^{\frac{1}{2}} I^{\frac{1}{2}}(\tilde{\theta}_k) \exp\left(-\frac{k^{-1} I(\tilde{\theta}_k)}{2} (\theta - \tilde{\theta}_k)^2 \right)$$

The remainder of the argument will be somewhat loose; for rigorous arguments of our final result (that a one-dimensional reference prior is a Jeffreys prior), see Bernardo's review paper Bernardo (2005).

Suppose now that $\tilde{\theta}_k$ results from k independent draws of X, where X has distribution of $X|\theta_0$ for a particular θ_0. Then, because $\tilde{\theta}_k$ is consistent, $\tilde{\theta}_k \to_p \theta_0$ and under regularity conditions, also $I(\tilde{\theta}_k) \to I(\theta_0)$. Hence,

$$\begin{aligned} p(\theta_0|\tilde{\theta}_k) &\propto k^{\frac{1}{2}} I^{\frac{1}{2}}(\tilde{\theta}_k) \exp\left(-\frac{k^{-1} I(\tilde{\theta}_k)}{2} (\theta_0 - \tilde{\theta}_k)^2 \right) \\ &\approx k^{\frac{1}{2}} I^{\frac{1}{2}}(\theta_0) \exp\left(-\frac{k^{-1} I(\theta_0)}{2} (\theta_0 - \theta_0)^2 \right) \\ &= k^{\frac{1}{2}} I^{\frac{1}{2}}(\theta_0). \end{aligned}$$

Returning now to equation (2), since the inner integral is an expectation with respect to $p(t^k|\theta)$ so that the preceding theory applies so long as there exists some asymptotically normal efficient estimator $\tilde{\theta}_k = \tilde{\theta}(t^k)$, we obtain

$$f_k(\theta) \approx \exp\left\{ \int p(t^k|\theta) \log\left(I^{\frac{1}{2}}(\theta) \right) dt^k \right\}.$$

Since the term inside the integral does not depend on t^k, and it is being integrated against a density, as $k \to \infty$ we have

$$f_k(\theta) \approx I^{\frac{1}{2}}(\theta_0).$$

In other words, when there is an asymptotically normal asymptotically efficient estimator $\tilde{\theta}_k$, the Jeffreys prior is a reference prior in one dimension!

3 Example: Reference Prior for Exponential Distribution

Let $X_i \sim \text{Exp}(\theta)$ be iid. A sufficient statistic for θ is $\bar{x} := \frac{\sum X_i}{n}$, and the maximum likelihood estimator is $\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$. We could use the Jeffreys prior directly, but let us instead work through some of the work above for this specific case.

Set $X = (X_1, \ldots, X_n)$. Then $p(x|\theta) = \theta^n \exp(-n\bar{x}\theta)$. Since the Bernstein Von Mises theorem ensures that the posterior is the same as $n \to \infty$ regardless of the prior we use, we may just take the prior to be flat for convenience. Hence, asymptotically, we have

$$p(\theta|\hat{\theta}_{ML}) \propto \theta^n \exp(-n\theta/\hat{\theta}_{ML}),$$

and since $\hat{\theta}_{ML}$ is consistent, i.e. $\hat{\theta}_{ML} \to \theta_0$ when $X \sim Exp(\theta_0)$, we obtain

$$\pi_n(\theta) = \left(\frac{n}{\hat{\theta}_{ML}}\right)^{n+1} \frac{1}{\Gamma(k+1)} \theta^n \exp\left(-\frac{n\theta}{\hat{\theta}_{ML}}\right) |_{\hat{\theta}_{ML}=\theta} \propto \frac{1}{\theta},$$

where we evaluated at $\hat{\theta}_{ML} = \theta$ because, in the definition of f_n, we are integrating with respect to $p(x|\theta)$, so that the consistency applies.

4 Invariance to Transformations

We have already showed that reference priors are the same as Jeffreys priors under regularity conditions, so they are invariant to transformations in that setting, but it also follows directly from the definition that they are invariant to transformation even in the absence of such regularity conditions. Specifically, the reference prior was defined in terms of mutual information, but mutual information is transformation-invariant. That is,

$$I(\theta, T^k) = \int p(t^k) \int p(\theta|t^k) \log \frac{p(\theta|t^k)}{p(\theta)} d\theta dt^k = \int p(t^k) \int p(\phi|t^k) \log \frac{p(\phi|t^k)}{p(\phi)} d\phi dt^k.$$

The equality holds because, when we do the changes of variables $p(\phi) = p(\theta(\phi)) \left|\frac{d\theta}{d\phi}\right|$ and $p(\phi|t^k) = p(\theta(\phi)|t^k) \left|\frac{d\theta}{d\phi}\right|$, the Jacobian terms inside the logarithm cancel, so that the logarithms in the integrals are equal. The term $\left|\frac{d\theta}{d\phi}\right|$ in $p(\phi) = p(\theta(\phi)) \left|\frac{d\theta}{d\phi}\right|$, on the other hand, is exactly what we obtain if we do a change of variables from θ to ϕ in the inner integral on the left hand side, obtaining

$$\int p(\theta|t^k) \log \frac{p(\theta|t^k)}{p(\theta)} d\theta = \int p(\theta(\phi)|t^k) \log \frac{p(\theta(\phi)|t^k))}{p(\theta)} \left|\frac{d\theta}{d\phi}\right| d\phi.$$

Hence, mutual information is transformation-invariant, and so are reference priors.

5 Example: Location and Scale Families

5.1 Location Families

For a given density f (which we identify with the induced distribution) define a class of measures

$$m_1 = \{f(x - \mu) : x \in \mathbb{R}, \mu \in \mathbb{R}\}.$$

For a particular μ and random variable $X \sim f(x-\mu)$, let $Y = X + \alpha$, and $\theta = \mu + \alpha$. If π denotes the reference prior under the parametrization in our definition of m_1 above, then if $f'(y) = f(y - \alpha - \mu)$, the similarly constructed family

$$m_1' = \{f'(y-\theta) : y \in \mathbb{R}, \theta \in \mathbb{R}\}$$

is actually equal to m_1, reparametrized by the transformation $\theta = \mu + \alpha$. Since we know reference priors are transformation-invariant and the Jacobian of the transformation is equal to 1, $\pi'(\theta) = \pi(\mu)$. But since the two families are identical, we also have $\pi'(\theta) = \pi(\mu + \alpha)$, and hence $\pi(\mu + \alpha) = \pi(\mu)$. In other words, reference priors for one-dimensional location families are flat.

5.2 Scale Families

Define the family

$$m_2 = \left\{ \frac{1}{\sigma} f(\frac{X}{\sigma}) : x > 0, \sigma > 0 \right\}.$$

Taking $y = \log x$ and $\phi = \log \sigma$, define an equivalent reparametrized family

$$m_2' = \{f(\exp(y - \phi)) : y \in \mathbb{R}, \phi \in \mathbb{R}\}.$$

In m_2', ϕ is a location family, so $\pi'(\phi)$ is flat by the work in Section 5.1. But since $\pi'(\phi) = \sigma\pi(\sigma)$ by a change of variables and transformation-invariance, we therefore obtain $\pi(\sigma) \propto \frac{1}{\sigma}$.

References

Bernardo, J. (2005). Reference analysis. *Handbook of Statistics*, 25:17–60.