Stat260: Bayesian Modeling and Inference **Lecture Date: March 1, 2010**

Reference Priors, Nuisance Parameters, and Multiple Regression

Lecturer: Michael I. Jordan *Scribe: Dave Golland*

1 Recap of Reference Priors

Recall that, given a likelihood $p(x|\theta)$, the reference prior is the specific uninformative prior that maximizes the divergence between the prior and posterior:

$$\pi_{ref}(\theta) = \arg\max_{p(\theta)} I(\theta, T) \tag{1}$$

where T is a sufficient statistic of the data and I is the expected KL-divergence between the posterior and the prior:

$$I(\theta, T) = \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt \tag{2}$$

$$= \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta)p(t)} d\theta dt \tag{3}$$

In a previous lecture we showed that for a given $p(x|\theta)$, where θ is one-dimensional, the reference prior is identical to the Jeffrey's prior, $\pi_J(\theta)$. Specifically, if θ is one-dimensional,

$$\pi_{ref}(\theta) = \pi_J(\theta) \tag{4}$$

$$\propto \sqrt{-\mathbb{E}\left[\frac{d^2 \log p(X|\theta)}{d\theta^2}\right]} \tag{5}$$

2 Nuisance Parameters

During statistical modeling, when the parameter is multidimensional or there are multiple parameters, it is often the case that we are interested in only a subset of the parameters (or components of a multidimensional parameter). To handle such situations, we employ the machinery of reference priors with nuisance parameters. Often, this machinery simplifies computations because the steps in the problem can be reduced to one-dimension, which means we can simply compute Jeffrey's prior.

Consider a likelihood of the form:

$$p(x|\theta, \lambda) \tag{6}$$

where θ is the parameter of interest and λ is the nuisance parameter.

We would like to find a joint prior $\pi(\theta, \lambda)$ that captures our unequal interests in the parameters. The general procedure for handling nuisance parameters is:

1. Condition on θ.

2. Holding θ fixed, find $\pi(\lambda|\theta)$ using the standard procedure for reference priors. (If λ is one-dimensional, simply compute the Jeffrey's prior of $p(x|\lambda,\theta)$ assuming θ is a constant.)
3. If $\pi(\lambda|\theta)$ is proper, integrate out λ to find: $p(x|\theta) = \int p(x|\lambda,\theta)\pi(\lambda|\theta)d\lambda$.
4. Based on $p(x|\theta)$, find $\pi(\theta)$ using the standard procedure for reference priors. (If θ is one-dimensional, simply compute the Jeffrey's prior of $p(x|\theta)$.)
5. $\pi(\theta,\lambda) = \pi(\lambda|\theta)\pi(\theta)$

Remark 1. For more than two parameters, we order the parameters in decreasing order of interest and repeatedly apply the above procedure.

3 Asymptotics

(θ,λ) are asymptotically normal (AN), with variance $V(\hat{\theta_n},\hat{\lambda_n})/n$.

Where:

$$V(\theta,\lambda) = \begin{pmatrix} V_{\theta\theta}(\theta,\lambda) & V_{\theta\lambda}(\theta,\lambda) \\ V_{\theta\lambda}(\theta,\lambda) & V_{\lambda\lambda}(\theta,\lambda) \end{pmatrix} \tag{7}$$

Now we're in the world of Gaussians.

Let $H(\theta,\lambda) = V^{-1}(\theta,\lambda)$.

We have: $\pi(\lambda|\theta) \propto h_{\lambda\lambda}(\hat{\theta}_n,\hat{\lambda}_n)$, where $h_{\lambda\lambda}$ is the lower, right hand corner of the inverse matrix.

And $\pi(\theta) \propto \exp\left(\int \pi(\lambda|\theta)\log V_{\theta\theta}^{-1/2}(\theta,\lambda)d\lambda\right)$, where $V_{\theta\theta}^{-1/2}(\theta,\lambda)$ is the marginal variance.

Example 2. Univariate normal.

μ is the parameter of interest.

σ is the nuisance parameter.

Since it is a univariate normal, the likelihood is: $N(x|\mu,\sigma)$. We have a mechanical procedure to get a prior. In general the first step is to calculate the asymptotic covariance, but here the likelihood is already normal, so we don't even need asymptotics.

The Fisher information matrix is:

$$I(\mu,\sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix} \tag{8}$$

$$V^{-1}(\mu,\sigma) = \begin{pmatrix} \sigma^{2} & 0 \\ 0 & \frac{1}{2}\sigma^{2} \end{pmatrix} \tag{9}$$

From the square root of the (2,2) entry, we have: $\pi(\sigma|\mu) \propto \sigma^{-1}$. Since μ does not appear in the matrix, we see that $\pi(\mu) \propto$ constant. Hence, $\pi(\mu,\sigma) \propto \sigma^{-1}$. Compare this result to the multivariate Jeffreys prior: $\propto \sigma^{-2}$.

It turns out we get the same prior when μ is the nuisance parameter. To see that this does not happen in general, consider the following example.

Example 3. Univariate normal.

$\phi = \mu/\sigma$ is the parameter of interest (this is also sometimes referred to as the coefficient of variation).

σ is the nuisance parameter.

After churning through the math, we get:

$$I(\phi, \sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix} \tag{10}$$

The prior that is induced from this procedure on μ and σ is: $\pi(\mu, \sigma) \propto \left(1 + \frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2\right)^{-1/2} \sigma^{-2}$.

Remark 4. This prior is not equal to the prior we got from the previous exercise: σ^{-1}. Hence, reference priors with nuisance parameters are not invariant under reparameterization.

4 Experimental Design Matters for Reference Priors

Experimental design refers to the method used to collect the data. The likelihood principle refers to the concept that all the information carried in a sample is contained in the likelihood function. The objective Bayesian believes that determining the prior should be incorporated in the experimental design, thereby violating the likelihood principle. The following examples illustrate the objective Bayesian view on the likelihood principle and experimental design.

Example 5. Consider the scenario in which we toss a coin m times and observe r heads. The likelihood for the data is binomial:

$$p(x|\theta) \propto \binom{m}{r} \theta^r (1-\theta)^{m-r} \tag{11}$$

The Jeffreys prior in this case is: $\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$. The corresponding posterior is:

$$\pi(\theta|x) = \text{Beta}\left(\theta \,\middle|\, r + \frac{1}{2}, m - r + \frac{1}{2}\right) \tag{12}$$

By contrast, consider the scenario in which we toss a coin until we see r heads, and end up tossing it m times in total. The likelihood for this second scenario is the negative binomial:

$$p(x|\theta) \propto \binom{m-1}{r-1} \theta^r (1-\theta)^{m-r} \tag{13}$$

The difference between the two scenarios is captured in the constant term: $\binom{m-1}{r-1}$ vs. $\binom{m}{r}$. This term is considered to be exclusively part of the experimental design since it does not affect the shape of the likelihood.

In the negative binomial case, the Jeffreys prior becomes $\pi_J(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$. The corresponding posterior is:

$$\pi(\theta|x) = \text{Beta}\left(\theta \,\middle|\, r, m - r + \frac{1}{2}\right) \tag{14}$$

Note, the first parameter of the Beta posterior in the negative binomial case (r) differs from that in the binomial case ($r + 1/2$). Unlike in the binomial, where r can be 0, in the negative binomial, $r \neq 0$. Hence, the posterior will always be proper.

Remark 6. From this example, we see that reference priors are responsive to experimental design, thus they violate the likelihood principle.

Example 7. Product of Normal Means

$$p(x, y|\alpha, \beta) = \prod_i N(x_i|\alpha, 1) \prod_i N(y_i|\beta, 1) \tag{15}$$

$\phi = \alpha\beta$ is the parameter of interest.

$\lambda = \alpha/\beta$ is the nuisance parameter.

The product of means (parameter of interest) can intuitively be interpreted as the area of a rectangle that we want to infer from noisy samples.

The joint reference prior turns out to be:

$$\pi(\phi, \lambda) \propto \phi^{-1/2}\lambda^{-1}\left(\frac{\lambda}{n} + \frac{1}{n\lambda}\right)^{1/2} \tag{16}$$

Remark 8. Notice that the prior depends on n, the sample size. The objective Bayesian views the selection of the prior as part of the experimental design and therefore is willing to allow the prior to depend on the sample side. However, as a believer in the likelihood principle, the subjective Bayesian has philosophical concerns with allowing the sample size appear in the expression for the prior.

According to Berger and Bernardo,[1] the above value for $\pi(\phi, \lambda)$ is a satisfactory prior.

Example 9. Multivariate Normal

Consider the multivariate normal distribution parameterized by mean and precision with likelihood:

$$N_p(x|\mu, \tau I_{p\times p}) \tag{17}$$

where μ is the mean, τ is the precision (inverse of covariance), and I is the identity matrix. The subscript p indicates that we have a p-dimensional normal distribution.

After some computation, we have:

$$H(\mu, \tau) = \begin{pmatrix} \tau I & 0 \\ 0 & \frac{pn}{2\tau^2} \end{pmatrix} \tag{18}$$

The prior is:

$$\pi(\mu_1, \ldots, \mu_p, \tau) \propto \tau^{-1} \tag{19}$$

where μ_i is the i^{th} component of the mean vector. Notice, the expression for the prior does not depend on μ, hence we have a flat prior on μ.

Since τ is the precision, we have $\sigma = \frac{1}{\tau}$, where σ is the covariance. Using the expression for σ we perform a change of variables and get:

$$\pi(\mu_1, \ldots, \mu_p, \tau) = \sigma^{-1} \tag{20}$$

However, Stein's paradox implies that we should not always use this prior in multivariate normal. The prior we use should depend on the parameter of interest. For instance, if $||\mu||$ is the parameter of interest, then

[1] Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. Annals of Statistics 37, pg. 905-938.

we should not get a flat prior over μ. In fact, if we were to follow the reference prior procedure with $||\mu||$ as the parameter of interest, we would not find a flat prior. In other words, reference priors resolve the Stein paradox.

Example 10. Correlation Coefficient

Let ρ be the correlation coefficient for a bivariate normal distribution. After some work we find that:

$$\pi(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2) \propto (1 - \rho^2)\sigma_1^{-1}\sigma_2^{-1} \tag{21}$$

where ρ is the parameter of interest, and the rest of the parameters appear in order of decreasing interest (increasing nuisance): $\mu_1, \mu_2, \sigma_1, \sigma_2$.

Notice that the expression for the prior does not depend on μ. Furthermore, the product of $\sigma_1^{-1}\sigma_2^{-1}$ shows that the prior depends on the product of the Jeffrey's priors for these quantities.

Remark 11. Intuitively, it is satisfying to see that the strength of the prior increases as the correlation coefficient decreases because it is consistent with what we expect from an uninformative prior. When the data is uncorrelated (correlation coefficient is low), then there is less redundancy in the data. In other words, the data carries more information when the value of the correlation coefficient is small. This means that it is safe for the prior to put mass on small values of the correlation coefficient because if the prior is wrong, it will be overwhelmed by the data. The opposite is true when the correlation coefficient is large; the data is redundant, and therefore the prior will have a large effect on the posterior. Hence, to stay uninformative the prior puts little mass on large values of ρ. These behaviors are favorable since reference priors are meant to be uninformative priors.

5 Multivariate Regression

We are given a set of data:

$$\{(x_i, y_i)\}_{i=1}^n \tag{22}$$

where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

We form the design matrix:

$$X = \begin{pmatrix} -x_1^\top - \\ -x_2^\top - \\ \vdots \\ -x_n^\top - \end{pmatrix} \tag{23}$$

Let:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{24}$$

The likelihood is:

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I) \tag{25}$$

That is, each of the y_i is independent, but not identically distributed. They are generated from normal distributions with different means.

5.1 Frequentist View

The frequentist looks for the maximum likelihood estimate (MLE):

$$\hat{\beta}_{MLE} = (X^\top X)^{-1} X^\top y \tag{26}$$

It turns out that $\hat{\beta}_{MLE}$ also is the least squares estimate, the value of β that minimizes the sum of squared residuals: $\left(y - X\hat{\beta}\right)^\top \left(y - X\hat{\beta}\right)$.

Frequentist confidence intervals are:

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 w_{ii}}} \tag{27}$$

where:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\left(y - X\hat{\beta}\right)^\top \left(y - X\hat{\beta}\right)$$
$$w_{ii} = (X^\top X)^{-1}_{ii}$$

$\hat{\sigma}$ is a heuristic estimator for the σ

It turns out that $\hat{\beta}$ is an unbiased estimator. The frequentists consider the variability of the estimator over multiple training sets:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1} \tag{28}$$

assuming σ^2 is known.

5.1.1 ANOVA

An important instance of multiple regression is analysis of variance (ANOVA). In ANOVA, X is an indicator vector, but everything else is unchanged.

5.2 Bayesian View

Again, the likelihood for the data is:

$$p(y|\beta, \sigma^2, X) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right) \tag{29}$$

The Bayesian considers conjugate priors: $\beta|\sigma^2, X$. Since β appears quadratically in the likelihood, the conjugate prior is of the form:

$$\beta|\sigma^2, X \sim N(\beta_0, \sigma^2 M^{-1}) \tag{30}$$
$$\sigma^2|X \sim IG(a, b) \tag{31}$$

where IG is an inverse gamma distribution.

It turns out that the conjugate prior is often too informative, so we will talk about g-priors.