Stat260: Bayesian Modeling and Inference **Lecture Date: March 17, 2010**

g Priors and Hierarchical Models

Lecturer: Michael I. Jordan *Scribe: Mason Liang*

1 Bayes Factor for the g Prior

In the last lecture, we showed that the Bayes factor for the linear regression model under the g prior is given by

$$\mathrm{BF}(M_\gamma, M_N) = \frac{(1+g)^{(n-1+p_\gamma)/2}}{(1+g(1-R_\gamma)^2)^{(n-1)/2}},$$

We showed that g needs to be data dependent and mentioned two approaches for choosing it, empirical Bayes and full Bayes.

1.1 Empirical Bayes

Choose g so that it maximizes the marginal likelihood, $p(y|M_\gamma, g)$. In the previous lecture, we gave an intuitive explanation for why the marginal likelihood penalizes models with more parameters. In the case of linear regression, we find that $\hat{g}_\gamma^{EB} = \max(F_\gamma - 1, 0)$, where

$$F_\gamma = \frac{R_\gamma^2/p_\gamma}{(1-R_\gamma^2)/(n-1-p_\gamma)},$$

and R_γ^2 and p_γ are the coefficient of determination and number of degrees of freedom in the model, respectively. Plugging this back into the expression for the Bayes factor gives:

$$\mathrm{BF}^{EB}(M_\gamma, M_n) = \frac{\left(1+R_\gamma^2(n-1-p_\gamma)/p_\gamma(-R_\gamma^2)\right)^{(n-1-p_\gamma)/2}}{\left(1+(n-1-p_\gamma)R_\gamma^2/p_\gamma\right)^{(n-1)/2}}.$$

As $R_\gamma \to 1$ the Bayes factor diverges, which implies this approach does not suffer from the information paradox. However, Liang et al. (2008) showed that empirical Bayes does not guarantee model selection consistency, a desirable property from the frequentist perspective.

1.2 Full Bayes

A subjective Bayes approach does not really work here, as we don't know how g should behave. An objective Bayes approach would be to put a prior on g that avoids paradoxes *and* gives model selection consistency.

Alternatively, we could try to put a prior on β instead of g. For this to work, Jeffreys suggests that β needs a heavier tailed prior than a normal distribution. The Zellner-Siow prior, which is based on the Cauchy distribution, is an example of this.

To avoid the information paradox, we need the Bayes factor to diverge as $R_\gamma^2 \to 0$. This is equivalent to

$$\int (1+g)^{(n-1-p_\gamma)/2} \pi(g) dg = \infty.$$

One choice of $\pi(g)$ that satisfies this is

$$\pi(g) \propto g^{-3/2} \exp(-n/2g).$$

or $g \sim \text{IG}(\frac{1}{2}, \frac{n}{2})$. This also gives a good prior for β, since $\beta|g \sim$ is normal, so $\pi(\beta) = \int \pi(\beta|g)\pi(g)dg$ will be Cauchy.

Choosing priors in this way also makes computations easier. We can sample g first from an inverse gamma distribution. Conditional on g, β is multivariate normal. If we had tried to sample β directly, we would have to sample from a multivariate Cauchy distribution, which is more expensive computationally. Choosing to sample β in this way is analogous to the concept of disintegration from probability theory.

Since the Bayes factor diverges for $R_\gamma^2 \to 1$,, we do not have the information or Lindley paradoxes. Furthermore, Liang et al. (2008) also prove that this approach gives model consistency.

Note that the inverse gamma prior for g is just an example of a prior that has desirable properties and is not the only one; other priors could avoid paradoxes and have model selection consistency as well as mentioned by Liang et al. (2008).

2 Hierarchical Models

There are many applications for the idea of a hierarchical model. For now, we will focus on a simpler case: random effects models and meta-analysis. Consider the one-way normal random effects ANOVA:

$$y_{ij} = \alpha + \theta_j + \epsilon_{ij},$$

for groups or panels $j = 1, \ldots, J$ each composed of subjects, units, or individuals, $i = 1, \ldots, n_j$. Sometimes, the θ_j are the effects of interest. Alternatively, they could be treated as nuisance parameters. Assuming ϵ_{ij} has a normal distribution, the maximum likelihood estimator for θ_j is

$$\hat{\theta}_j^{MLE} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}.$$

Note that if n_j is small, then we do not get too much information about θ_j. Therefore, if we believe that the θ_j are random effects, we might want to borrow information from estimates of other θ_k's to add precision to our estimate of θ_j. Using this idea in a frequentist framework results in REML (restricted maximum likelihood) estimators.

As Bayesians, however, we give θ_j a prior, i.e., assume that $\theta_j \sim N(\theta, \tau^2)$ in this case. However, we still need to figure out how θ and τ should be chosen. If we were doing an empirical Bayes approach, we could just fix their values. However, to be fully Bayesian, we would need to give priors for these hyperparameters as well. This distribution would presumably involve even more hyperparameters, potentially creating an ever-ascending hierarchy. An infinite hierarchy might make analysis difficult, so typically, you would run sensitivity analyses to see when adding additional layers to the hierarchy stops mattering. If you do not ever seem to reach this point, then you should acknowledge this fact in your analysis.

Returning to our example, we have the following hierarchy:

$$y_{ij}|\theta_j, \sigma^2 \overset{\text{iid}}{\sim} N(\theta_j, \sigma^2) \text{ for } 1 \leq i \leq n_j$$

and

$$\theta_j|\mu,\tau \overset{\text{iid}}{\sim} N(\mu,\tau^2) \text{ for } 1 \leq j \leq N.$$

We want to obtain an estimate of θ_j, for example $\hat{\theta}_j = \mathbb{E}[\theta_j|\mathbf{y}]$. The sufficient statistics of the model are $\bar{y}_{ij} = \frac{1}{n_j}\sum_i y_{ij}$. The model implies that $\bar{y}_{ij} \sim \mathrm{N}(\theta_j, \frac{\sigma^2}{n_j})$. Using the prior

$$p(\mu,\tau) = p(\tau),$$

working through the math gives

$$\theta_j|\mu,\tau,\mathbf{y} \sim \mathrm{N}(\hat{\theta}_j, v_j)$$

with

$$\hat{\theta}_j = \frac{\frac{\bar{y}_{ij}}{\sigma_j^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}} \quad \text{and} \quad v_j = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}.$$

Therefore, our posterior mean is a convex combination of the group mean and the grand mean. We also get that

$$\mu|\tau,\mathbf{y} \sim \mathrm{N}(\hat{\mu}, v_\mu),$$

with

$$\hat{\mu} = \frac{\sum_j \bar{y}_{ij}(\sigma_j^2+\tau^2)^{-1}}{\sum_j(\sigma_j^2+\tau^2)^{-1}} \quad \text{and} \quad v_\mu = \frac{1}{\sum_j(\sigma_j^2+\tau^2)^{-1}}.$$

Note that τ controls how much information we "share" between the different estimates. We still need to specify a prior for τ. The scale invariant prior suggested by Jeffreys, $p(\tau) \propto \frac{1}{\tau}$ will lead to an improper posterior distribution, so it cannot be used. $p(\tau) \propto 1$ gives a proper posterior, and so is a possibility. It seems like there should be a more objective way for choosing a prior for τ, but this is still an open research question. Currently, the approach has been to choose many different priors and see how sensitive the analysis is to the choice of prior.

References

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423.