Stat260: Bayesian Modeling and Inference **Lecture Date: March 10, 2010**

Bayes Factors, g-priors, and Model Selection for Regression

Lecturer: Michael I. Jordan *Scribe: Tamara Broderick*

The reading for this lecture is Liang et al. (2008).

1 Bayes Factors

We begin by clarifying a point about Bayes factors from the previous lecture. Consider data $x \in \mathcal{X}$ and parameter $\theta \in \Theta$. Take $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ for some partition $\{\Theta_0, \Theta_1\}$ of Θ.

Θ
Θ_0
Θ_1

Figure 1: Cartoon of the partition of Θ space into regions Θ_0 and Θ_1.

Then the Bayes factor BF for comparing models H_0 and H_1 is

$$\mathsf{BF} = \frac{p(x|H_0)}{p(x|H_1)}$$

And we have the following relation between the posterior odds and prior odds.

$$\frac{p(H_0|x)}{p(H_1|x)} = \mathsf{BF} \cdot \frac{p(H_0)}{p(H_1)}$$

Note that the probabilities $p(H_0)$ and $p(H_1)$ are hyperpriors on the models themselves and are not expressible in terms of $\pi(\theta|H_0)$ or $\pi(\theta|H_1)$. Indeed, the overall prior on θ is a mixture distribution where $p(H_0)$ and $p(H_1)$ are the mixture probabilities.

$$\pi(\theta) = p(H_0)\pi(\theta|H_0) + p(H_1)\pi(\theta|H_1)$$

Once we know BF, we can solve for the posterior probability of H_0.

$$\begin{aligned} \frac{p(H_0|x)}{1 - p(H_0|x)} &= \mathsf{BF} \cdot \frac{p(H_0)}{p(H_1)} \\ \Rightarrow p(H_0|x) &= \frac{\mathsf{BF} \cdot \frac{p(H_0)}{p(H_1)}}{1 + \mathsf{BF} \cdot \frac{p(H_0)}{p(H_1)}} \end{aligned}$$

We can see from the last line that an arbitrary constant in the Bayes factor does not factor out when calculating the posterior of H_0. Rather, an arbitrary constant can be used to obtain an arbitrary posterior. Letting the Bayes factor constant decrease to zero yields $p(H_0|x) = 0$ in the limit, and letting the Bayes factor increase to infinity yields $p(H_0|x) = 1$ in the limit.

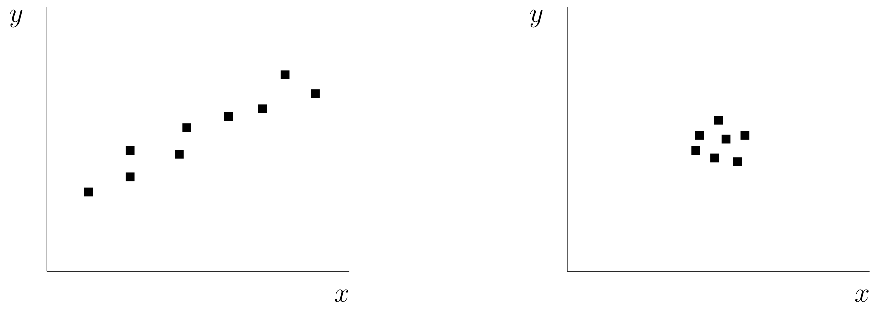

Figure 2: *Left*: In one-dimensional linear regression, sets of points with more scatter along the one-dimensional predictor x are more informative about the slope and intercept. It is easy to roughly predict these parameters from the figure. *Right*: Data with less scatter in the predictor variable are less informative about the slope and intercept.

2 g-priors

We consider a regression model

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

with a Jeffreys prior on σ^2

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

and a prior called a "g-prior" on the vector of regression coefficients β

$$\beta \sim \mathcal{N}(\beta_0, g\sigma^2(X^T X)^{-1}) \tag{3}$$

In Eq. (3), g is a scalar that we will determine later, and the σ^2 factor in the covariance is necessary for conjugacy. Generally in regression models, we make the assumption that the design matrix X is known and fixed, so we are free to use it in our prior.

The particular form in which X appears in Eq. (3) represents a kind of dispersion matrix. For instance, recall that for the usual maximum likelihood estimator $\hat{\beta}$ of β, we have

$$\text{Var}(\hat{\beta}) = (X^T X)^{-1} \cdot \{\text{an estimate of } \sigma^2\}$$

Alternatively, consider a principal component analysis on X (and ignore the response variable y for the moment). The eigenvalues of $X^T X$ give the directions of the new coordinates. Although the g-prior is not a reference prior, we can further motivate the use of X in Eq. (3) by the same intuitive considerations we used for reference priors, namely placing more prior mass in areas of the parameter space where we expect the data to be less informative about the parameters. In the one-dimensional classical regression case (Figure 2), high scatter along the predictor axis corresponds to high leverage. The data in this case are more informative about the intercept and slope (β) than in the case of less scatter and low leverage.

Posterior

With this generative model (Eqs. 1, 2, 3) in hand, we can calculate the full posterior.

$$\begin{aligned}\pi(\beta,\sigma^2|y,X) \propto (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp\{ &-\frac{1}{2\sigma^2}(y-X\hat{\beta})^T(y-X\hat{\beta}) \\ &-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^T X^T X(\beta-\hat{\beta}) \\ &-\frac{1}{2\sigma^2}(\beta-\beta_0)^T X^T X(\beta-\beta_0)\} \end{aligned} \tag{4}$$

Recall that the maximum likelihood estimator $\hat{\beta}$ satisfies

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

and note that the first term in the sum is proportional to the classical residual s^2.

$$s^2 = (y-X\hat{\beta})^T(y-X\hat{\beta})$$

From Eq. (4), we obtain a Gaussian posterior on β and an inverse gamma posterior on σ^2.

$$\beta|\sigma^2, y, X \overset{d}{=} \mathcal{N}\left(\frac{g}{g+1}\left(\frac{\beta_0}{g}+\hat{\beta}\right), \frac{\sigma^2 g}{g+1}(X^T X)^{-1}\right) \tag{5}$$

$$\sigma^2|y, X \overset{d}{=} \mathrm{IG}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}(\hat{\beta}-\beta_0)X^T X(\hat{\beta}-\beta_0)\right) \tag{6}$$

Posterior means

From Eqs. (5, 6), we can find each parameter's posterior mean. First,

$$\begin{aligned}\mathbb{E}(\beta|y,X) &= \mathbb{E}\left(\mathbb{E}(\beta|\sigma^2,y,X)|y,X\right) \\ & \text{by the tower rule} \\ &= \mathbb{E}\left(\frac{g}{g+1}\left(\frac{\beta_0}{g}+\hat{\beta}\right)\middle| y,X\right) \\ &= \frac{g}{g+1}\left(\frac{\beta_0}{g}+\hat{\beta}\right)\end{aligned}$$

Letting g increase to infinity, we recover the frequentist estimate of β: $\mathbb{E}(\beta|y,X) = \hat{\beta}$. Moreover, as g increases to infinity, we approach a flat prior on β. We have previously seen that such a prior can be useful for estimation but becomes problematic when performing hypothesis testing.

Next, using the fact that the mean of an $IG(a,b)$-distributed random variable is $b/(a-1)$ for $a>1$, we have

$$\mathbb{E}(\sigma^2|y,X) = \frac{s^2 + \frac{1}{g+1}(\hat{\beta}-\beta_0)^T X^T X(\hat{\beta}-\beta_0)}{n-2}$$

Letting $g\to\infty$ yields $\mathbb{E}(\sigma^2|y,X) = s^2/(n-2)$. Compare this result to the classical frequentist unbiased estimator $s^2/(n-p)$ for σ^2. Here, p is the number of degrees of freedom, or equivalently the number of components of the β vector. One might worry that $s^2/(n-2)$ is overly conservative for $p>2$, but note that the σ^2-estimate does not provide the error bars on β for, e.g., confidence intervals. We will see how to calculate confidence intervals for β next.

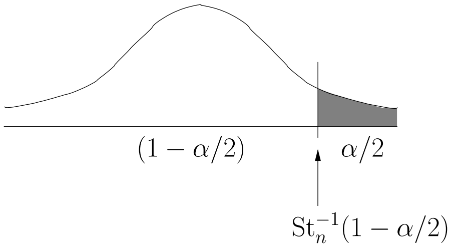

Figure 3: Illustration of the quantile function St_n^{-1}. Mass $\alpha/2$ of the Student's t-density appears to the right of $\mathrm{St}_n^{-1}(1-\alpha/2)$ and is shown in gray. The remaining mass is on the left.

Highest posterior density region for β_i

In order to find the HPD region for β_i, we first calculate the posterior of β_i with σ^2 integrated out.

$$\beta_i | y, X \overset{d}{=} \mathrm{St}(T_i, K_{ii}, n),$$

where St represents Student's t-distribution, which has three parameters: a mean, a spread, and a number of degrees of freedom. In our case, we further have

$$\begin{aligned} T_i &:= \frac{g}{g+1}\left(\frac{\beta_{0i}}{g} + \hat{\beta}_i\right) \\ K_{ii} &:= \frac{g}{g+1}\left(s^2 + \frac{(\hat{\beta}-\beta_0)^T X^T X(\hat{\beta}-\beta_0)}{g+1}\right)\frac{\omega_{ii}}{n} \\ \omega_{ii} &:= (X^T X)^{-1}_{ii} \end{aligned}$$

Here, β_{0i} and $\hat{\beta}_i$ are the i^{th} components of β_0 and $\hat{\beta}$, respectively. It follows that

$$\frac{\beta_i - T_i}{\sqrt{K_{ii}}} \overset{d}{=} \mathrm{St}_n,$$

where St_n is the standard Student's t-distribution (zero mean, unit variance) with n degrees of freedom. It follows that, if we let St^{-1} be the quantile function for the standard Student's t-distribution with n degrees of freedom (Figure 3), the HPD interval at level α is

$$T_i \pm \sqrt{K_{ii}}\mathrm{St}_n^{-1}\left(1 - \frac{\alpha}{2}\right)$$

How to set g

We've already seen that letting $g \to \infty$ is not an option if we want to perform hypothesis testing. Moreover, letting $g \to 0$ takes the posterior to the prior distribution, also reducing our ability to perform inference. Some other options for choosing g include using BIC, empirical Bayes, and full Bayes.

BIC The Bayesian Information Criterion (BIC) is an approximation to the marginal likelihood obtained by a Laplace expansion. We can use the BIC to determine a value for g, but the resulting model is not necessarily consistent (e.g. for the model selection in Section 3).

Empirical Bayes Empirical Bayes works by drawing a line in the hierarchical description of our model.

$$\begin{array}{l} y|\beta,\sigma^2 \\ \beta|\sigma^2, g \\ \sigma^2 \\ \hline g \end{array}$$

Then we take a Bayesian approach to estimation above the line and a frequentist approach below the line. While a full frequentist would consider a profile likelihood for g, empirical Bayes methodology prescribes integrating out other parameters and maximizing the marginal likelihood. Thus, the empirical Bayes estimate for g is $\hat{g}_{EB}$:

$$\hat{g}_{EB} = \underset{g}{\operatorname{argmax}}\, p(y|X, g)$$

Inference under the prior with g set to $\hat{g}_{EB}$ is, in general, consistent.

We can calculate the marginal likelihood for our case as follows. Integrating out β and σ^2 yields

$$p(y|X, g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n+1}{2}} n^{1/2}} ||y - \bar{y}||^{-(n-1)} \frac{(1+g)^{\frac{n-1-p}{2}}}{(1+g(1-R^2))^{\frac{n-1}{2}}}$$

In the formula above, $\bar{y}$ is the grand mean, $||\cdot||$ is a vector norm, and R^2 is the usual coefficient of determination.

$$R^2 = 1 - \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{(y-\bar{y})^T(y-\bar{y})}$$

Full Bayes A full Bayesian puts a prior distribution on g. Not only is a hierarchy of this type consistent, but the convergence will have the best rate (of the three methods considered here) in the frequentist sense.

3 Model Selection for Regression

Let γ be an indicator vector encoding which components of β are used in a model. When p is the dimensionality of β, we might have, for example,

$$\gamma = \overbrace{(0, 0, 1, 1, 0, 1, \ldots, 0)}^{p \text{ elements}}$$

Further define X_γ to be the matrix that has only those columns of X corresponding to ones in the γ vector, and let β_γ be the vector that has only those components of β corresponding to ones in γ. Then there are 2^p models M_γ of the form

$$y = X_\gamma \beta_\gamma + \epsilon$$

"Model selection" refers to the problem of picking one of these models. That is, we wish to choose γ based on data (X, y). The problem is like hypothesis testing but with many simultaneous hypotheses.

The issue we encounter is that 2^p can be very large. We need a way to select γ based on priors on the parameters in each model, but with potentially very many models, assigning separate priors for each model may be infeasible. We consider instead a Bayesian approach utilizing nested models. Recall that, when we wish to use the Bayes factor, we *may* use improper priors when the same parameter appears in both models.

Then we might, for instance, use a g-prior on the remaining parameters. This pairwise concept extends to model selection by comparing all models to the null model or the full model. That is,

$$\gamma_{null} = (0, \dots, 0)$$
$$\gamma_{full} = (1, \dots, 1)$$

For some base model M_b, we can calculate for each model M_γ,

$$\mathsf{BF}(M_\gamma, M_b) = \frac{p(y|M_\gamma)}{p(y|M_b)}$$

Then for any two models M_γ and $M_{\gamma'}$, we have

$$\mathsf{BF}(M_\gamma, M_{\gamma'}) = \frac{\mathsf{BF}(M_\gamma, M_b)}{\mathsf{BF}(M_{\gamma'}, M_b)}$$

If the base model is null, then the only factor in common between any model and M_{null} is σ^2 (and often the intercept α). We can place a Jeffreys prior separately on each of these parameters. In particular, α is a scale parameter, so

$$\pi(\sigma^2, \alpha) \propto \frac{1}{\sigma^2}$$

For the case where the base model is the full model, see Liang et al. (2008) for more information.

How to set g, continued

Choosing g-priors for the remaining components of β yields a closed-form expression for $\mathsf{BF}(M_\gamma, M'_\gamma)$ via

$$\mathsf{BF}(M_\gamma, M_{null}) = \frac{(1+g)^{(n-p_\gamma-1)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}}$$

Here, R^2_γ is the coefficient of determination for the model M_γ, and p_γ is the remaining number of parameters of β_γ.

We can see directly now that if we send $g \to \infty$, then $\mathsf{BF}(M_\gamma, M_{null}) \to 0$. That is, for any value of n, the null model is always preferred in the limit $g \to \infty$. This behavior is an instance of "Lindley's paradox."

On the other hand, consider choosing a fixed g, e.g. by using BIC, and letting $n \to \infty$. Then $\mathsf{BF}(M_\gamma, M) \to \infty$. That is, no matter the data, the alternative to the null is always preferred in the limit $n \to \infty$.

We will see a resolution to this quandary in the next lecture.

References

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.