Stat260: Bayesian Modeling and Inference **Lecture Date: March 15, 2010**

g-priors for Linear Regression

Lecturer: Michael I. Jordan *Scribe: Andrew H. Chan*

1 Linear regression and g-priors

In the last lecture, we mentioned the use of g-priors for linear regression in a Bayesian framework. In this lecture we continue to discuss several of the issues that arise when using g-priors.

As a recapitulation of the problem setup, recall that there is a set of data

$$y = X\beta + \epsilon,$$

where X is the **design matrix**, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and $\beta \sim \mathcal{N}(\beta_0, g\sigma(X^T X)^{-1})$. The prior on σ^2 is the Jeffreys prior, $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, and usually, β_0 is taken to be 0 for simplification purposes. The appeal of the method is that there is only one free parameter g for all linear regression. Furthermore, the simplicity of the g-prior model generally leads to easily obtained analytical results. However, we still face the problem of selecting g or a prior for g, and this lecture provides an overview of the issues that come up.

1.1 Marginal likelihood and Bayes factors

To compare different models to each other, we need to compute the marginal likelihoods of the given models, and then take their ratio to obtain the Bayes factor. Fortunately, for the g-prior, the marginal likelihood for a model M_γ is easily computed as

$$p(y|M_\gamma, g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n-1}{2}} n^{\frac{1}{2}}} \|y - \bar{y}\|^{-(n-1)} \frac{(1+g)^{\frac{n-1-p_\gamma}{2}}}{(1+g(1-R_\gamma^2))^{\frac{n-1}{2}}}, \tag{1}$$

where γ is an indicator of which covariates are used in the model, $\bar{y}$ is the grand mean of the data, R_γ^2 is the coefficient of determination, and p_γ is the numbers of degrees of freedom for the given model. Recall from the last lecture that the coefficient of determination is

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

where $\hat{\beta} = (X^T X)^{-1} X^T y$ is the MLE of β.

To find the Bayes factor of a given model with respect to the null model M_N, where in the null model, $p_\gamma = 0$ and $R_\gamma^2 = 0$, we take the ratio of their marginal likelihoods. The constants in (1) cancel, and we are left with

$$\frac{(1+g)^{\frac{n-1-p_\gamma}{2}}}{(1+g(1-R_\gamma^2))^{\frac{n-1}{2}}}.$$

With this in hand, we can compare any model to the null model, and thus, we can compare any pair of models by using

$$\mathrm{BF}(M_\gamma, M_{\gamma'}) = \frac{\mathrm{BF}(M_\gamma, M_N)}{\mathrm{BF}(M_{\gamma'}, M_N)}.$$

Before moving on to the question of how g should be selected in the model, we consider the reason why the marginal likelihood is used to compare models against each other.

1.2 Why the marginal likelihood?

If one were to plot the "classical" likelihood with respect to the "complexity" of the model, one would find that as the complexity of the model increased, so would the likelihood, as illustrated in Figure 1. In this setting, the likelihood is obtained after estimating the parameters using some statistical method (e.g. maximum likelihood), and then plotting the value of $p(x|\theta)$. However, the problem of using this method of comparison is that more "complex" models will always do better than simpler models. Because more complex models have more parameters, whatever can be done in a simple model can be done in a complex model, and thus, a more complex model will lead to a better fit. The downside is that this leads to **overfitting**, which is particularly damaging in the setting of *prediction*. In frequentist statistics, many methods have been developed to penalize the likelihood with respect to increasing complexity of the model. However, these methods are typically only justified after proposing the idea and then performing a series of analyses to show that it has desirable properties, rather than being well-motivated from the start.

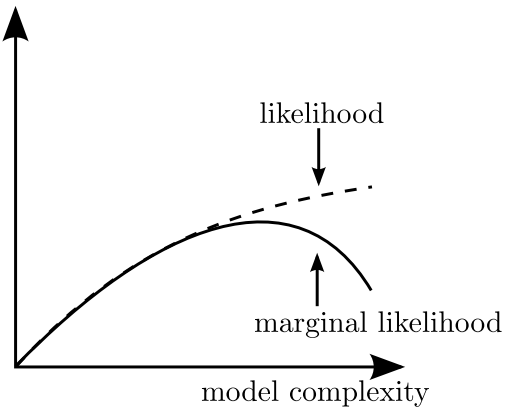

Figure 1: The dashed line represents the likelihood and the solid line is the marginal likelihood. With increasing model complexity, the likelihood continually increases, whereas the marginal likelihood has a maximum at some point, after which it begins to decrease.

In the Bayesian framework, marginal likelihoods have a natural built-in penalty for more complex models. At a certain point, the marginal likelihood will begin to decrease with increasing complexity, and hence, does not directly suffer from the overfitting problem that occurs when considering only likelihoods. The intuition for why the marginal likelihood begins to decrease is that as the complexity of the model increases, the prior will be spread out more thinly across both the "good" models and the "bad" models. Because the marginal likelihood is the likelihood integrated with respect to the prior, spreading the prior across too many models will place too little prior mass on the good models, and as a result, cause the marginal likelihood to decrease.

Another way to look at it is as follows. Suppose X is the data, θ are the parameters and M is the model. The posterior is then

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}. \tag{2}$$

Solving for the marginal likelihood $p(x|M)$ in (2), we obtain

$$p(x|M) = \frac{p(x|\theta, M)p(\theta|M)}{p(\theta|x, M)}. \tag{3}$$

Taking the log of (3) results in

$$\log p(x|M) = \underbrace{\log p(x|\theta, M)}_{\text{log likelihood}} + \underbrace{\log p(\theta|M) - \log p(\theta|x, M)}_{\text{penalty}}. \tag{4}$$

This is true for any choice of θ, and in particular, the maximum likelihood estimate of θ. The $\log p(x|\theta, M)$ term in (4) is the log likelihood, and only increases with increasing model complexity. On the other hand, the $\log p(\theta|M) - \log p(\theta|x, M)$ term can be viewed as a "penalty" that penalizes against complex models. The overall sign of this penalty is negative because $p(\theta|x, M)$, the posterior, is generally larger than $p(\theta|M)$, the prior, assuming that given the data, the posterior "sharpens" up with respect to the prior. Hence, the penalty term balances out the increase in likelihood as the model complexity increases.

2 Fixed value for g

We still need to find a prior for g, and so we first consider using a fixed g. The first idea might be to let g go to infinity since this gives a posterior expectation equal to the maximum likelihood estimate. However, as $g \to \infty$, the Bayes factor $\text{BF}(M_\gamma, M_N) \to 0$, and thus, the null model would always be preferred to any other model. This is generally a bad idea, and is called the *Lindley paradox*. In another attempt we might consider letting R^2_γ go to 1. R^2_γ, the coefficient of determination, is a measure of how well the data fits the regression. Intuitively, letting R^2_γ go to 1 considers what happens as we see data sets that are more and more appropriate for the model. However, in letting R^2_γ go to 1, we find that

$$\text{BF}(M_\gamma, M_N) \to (1+g)^{\frac{n-1-p_\gamma}{2}}, \tag{5}$$

which shows that the Bayes factor converges to a constant. One would expect the Bayes factor to go to infinity as we consider data sets that fit the model better and better, but here we find that it instead tops off at a constant, which is considered a problem and is essentially a symptom of a deeper issue. This problem is called the *information paradox*. In the end, we conclude that a fixed g is a bad idea, and g should, in fact, depend on the data. We will eventually construct a hierarchical model to handle the selection of g, but first we give a brief overview of hierarchical models.

3 Hierarchical models

Suppose we have some data x, parameters θ, and a hyperparameter λ, such that the distribution of x given θ is $p(x|\theta)$ and the distribution of θ given λ is $p(\theta|\lambda)$. In the regression setting, y is the data, β is the parameter, and g is the hyperparameter. The first question we must answer is how to set λ, the hyperparameter. We will consider two ways to do so, the empirical Bayes method and the fully Bayesian approach.

3.1 Empirical Bayes

The empirical Bayes method is to choose the value of λ that maximizes $p(x|\lambda)$; i.e.

$$\hat{\lambda}_{EB} = \underset{\lambda}{\text{argmax}}\, p(x|\lambda).$$

This is essentially a maximization problem over λ of

$$p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)d\theta.$$

Having chosen $\hat{\lambda}_{EB}$, we can use the model $p(x|\theta)$ with the prior $p(\theta|\hat{\lambda}_{EB})$ (where the prior is now fixed using $\hat{\lambda}_{EB}$) to do "Bayesian" inference. However, a major concern is that the data is used twice: once for the hyperparameter and another time in the likelihood $p(x|\theta)$. This typically leads to over-confidence in the posterior because the data is used multiple times. Indeed, empirical Bayes is not Bayesian. Rather, it's frequentist, i.e. it needs frequentist theory to justify it (and that theory is available).

For example, the James-Stein estimator is an empirical Bayes procedure [EM73]. This estimator is used in the setting where there is one draw of data x from a normal distribution, $x \sim \mathcal{N}(\theta, \sigma^2 I)$, where σ^2 is known. With only one draw of data, the maximum likelihood estimate for the mean is simply $\hat{\theta}_{ML} = x$. However, James and Stein [JS61] showed that in three or more dimensions, the estimate $\hat{\theta}_{ML}$ is inadmissible with respect to the least squares loss function, $l = \|\theta - \delta(x)\|^2$, where θ is the true mean and $\delta(x)$ is an estimate of the mean as a function of the data x. That is, there exists an estimator that is better than $\hat{\theta}_{ML}$ for all θ. They showed this by devising their own estimator $\hat{\theta}_{JS}$,

$$\hat{\theta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||x||_2^2}\right) x,$$

where p is the number of dimensions.

This estimator dominates $\hat{\theta}_{ML}$ for all θ in settings with three or more dimensions. Efron and Morris showed that the James-Stein estimator can be interpreted as a two-level Gaussian model, where the hyperparameter is obtained by applying empirical Bayes.

Empirical Bayesianism has some properties that make it especially useful in practice. For example, in large hierarchies, it is helpful to stop growing the hierarchy and plug in some numbers for the hyperparameters. These hyperparameters can be obtained by finding a maximum, which typically involves taking a derivative. This can often be easier than integrating throughout the entire hierarchy.

3.2 Fully Bayesian

An alternative to the empirical Bayes method is the fully Bayesian approach. We still have $p(x|\theta)$ and $p(\theta|\lambda)$, but instead of fixing λ to some estimate, we put a prior $p(\lambda)$ on λ. In this case, any inference depends on integrating over λ, generally using the posterior $p(\lambda|x)$. This approach is more conservative than the empirical Bayesian approach (i.e. it gives mores spread on the estimate of λ), and it also doesn't suffer from the issue of using the data twice. This approach can have good frequentist properties, but the choice of the prior $p(\lambda)$ matters, and that choice is essential to any frequentist analysis.

3.3 Other approaches

In addition to empirical Bayesianism and full Bayesianism, there are other approaches, such as **cross-validation**. In cross-validation, a fraction of the data is left out as "test" data, and the model is trained on the remaining data. The trained model is then used on the test data, and the parameter that generates the best likelihoods is chosen as the inferred parameter. One problem with cross-validation is that part of the data must be left out of the training set, which might not be practical if there is already very little data available.

4 Prior for g

Returning to the issue of selecting a prior for g in linear regression, we have a list of essential properties that the prior should have. The first is that it suffer from no paradoxes, such as the *Lindley paradox* and the *information paradox* described previously. The second is that it has *model selection consistency*, which is typically a frequentist property. However, applied here, this requirement will have both Bayesian and frequentist properties. For the model selection procedure to be consistent, the probability of the true model given the data must approach 1 as the number of data points n goes to infinity. That is, when M_λ is the true model,

$$p(M_\lambda|y, X) \xrightarrow{a.s.} 1 \text{ as } n \to \infty, \tag{6}$$

where $a.s.$ is for almost-sure convergence.

The Bayesian aspect of the statement in (6) is the fact that the model is assigned a probability. The frequentist aspect is that the convergence considers what happens when y is generated over multiple draws of the data, which is a frequentist notion.

If we use the empirical Bayes approach to select g, we do not encounter any paradoxes. However, this method does not have model selection consistency. Using the fully Bayesian approach, we do not encounter any paradoxes, and we also achieve model selection consistency. In the next lecture, we will cover two priors for g that have these appealing properties, the Zellner-Siow prior and the hyper-g prior.

References

[JS61] W. James, C. M. Stein, "Estimation with quadratic loss," *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

[EM73] B. Efron, C. Morris, "Stein's Estimation Rule and its Competitors – an Empirical Bayes Approach," *Journal of the American Statistical Association*, Vol. 68, No. 341, 1973.