

# Bayesian Linear Models

Sudipto Banerjee<sup>1</sup> and Andrew O. Finley<sup>2</sup>

<sup>1</sup> Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

<sup>2</sup> Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

March 3, 2010

- Ingredients of a linear model include an  $n \times 1$  response vector  $\mathbf{y} = (y_1, \dots, y_n)^T$  and an  $n \times p$  design matrix (e.g. including regressors)  $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ , assumed to have been observed without error. The linear model:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

- The linear model is the most fundamental of all serious statistical models encompassing:
  - ANOVA:  $\mathbf{y}$  is continuous,  $\mathbf{x}_i$ 's are categorical
  - REGRESSION:  $\mathbf{y}$  is continuous,  $\mathbf{x}_i$ 's are continuous
  - ANCOVA:  $\mathbf{y}$  is continuous, some  $\mathbf{x}_i$ 's are continuous, some categorical.
- Unknown parameters include the regression parameters  $\boldsymbol{\beta}$  and the variance  $\sigma^2$ . We assume  $X$  is observed without error and all inference is conditional on  $X$ .

- The classical unbiased estimates of the regression parameter  $\beta$  and  $\sigma^2$  are

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

$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{y} - X\hat{\beta})^T (\mathbf{y} - X\hat{\beta}).$$

- The above estimate of  $\beta$  is also a least-squares estimate. The *predicted* value of  $\mathbf{y}$  is given by

$$\hat{\mathbf{y}} = X\hat{\beta} = P_X \mathbf{y} \text{ where } P_X = X(X^T X)^{-1} X^T.$$

- $P_X$  is called the *projector* of  $X$ . It projects any vector to the space spanned by the columns of  $X$ .
- The model residual is estimated as:

$$\hat{\mathbf{e}} = (\mathbf{y} - X\hat{\beta})^T (\mathbf{y} - X\hat{\beta}) = \mathbf{y}^T (I - P_X) \mathbf{y}.$$

- For Bayesian analysis, we will need to specify priors for the unknown regression parameters  $\beta$  and the variance  $\sigma^2$ .
- Consider independent flat priors on  $\beta$  and  $\log \sigma^2$ :

$$p(\beta) \propto 1; p(\log(\sigma^2)) \propto 1 \text{ or equivalently } p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?
- It turns out that even if the priors are *improper* (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

- With a flat prior on  $\beta$  we obtain, after some algebra, the *conditional posterior* distribution:

$$p(\beta | \sigma^2, \mathbf{y}) = N(\beta | (X^T X)^{-1} X^T \mathbf{y}, \sigma^2 (X^T X)^{-1}).$$

- The conditional posterior distribution of  $\beta$  would have been the desired posterior distribution had  $\sigma^2$  been known.
- Since that is not the case, we need to obtain the *marginal posterior* distribution by integrating out  $\sigma^2$  as:

$$p(\beta | \mathbf{y}) = \int p(\beta | \sigma^2, \mathbf{y}) p(\sigma^2 | \mathbf{y}) d\sigma^2$$

- Can we solve this integration using composition sampling?  
YES: if we can generate samples from  $p(\sigma^2 | \mathbf{y})!$

- So, we need to find the marginal posterior distribution of  $\sigma^2$ . With the choice of the flat prior we obtain:

$$\begin{aligned} p(\sigma^2 | \mathbf{y}) &\propto \frac{1}{(\sigma^2)^{(n-p)/2+1}} \exp\left(-\frac{(n-p)s^2}{2\sigma^2}\right) \\ &= IG\left(\sigma^2 \mid \frac{n-p}{2}, \frac{(n-p)s^2}{2}\right), \end{aligned}$$

where  $s^2 = \hat{\sigma}^2 = \frac{1}{n-p} \mathbf{y}^T (I - P_X) \mathbf{y}$ .

- This is known as an *inverted Gamma* distribution (also called a *scaled chi-square* distribution)  $IG(\sigma^2 \mid (n-p)/2, (n-p)s^2/2)$ .
- In other words:  $[(n-p)s^2/\sigma^2 \mid \mathbf{y}] \sim \chi_{n-p}^2$  (with  $n-p$  degrees of freedom). A striking similarity with the classical result: The distribution of  $\hat{\sigma}^2$  is also characterized as  $(n-p)s^2/\sigma^2$  following a chi-square distribution.

- Now we are ready to carry out composition sampling from  $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$  as follows:

- Draw  $M$  samples from  $p(\sigma^2 | \mathbf{y})$ :

$$\sigma^{2(j)} \sim IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}(n-p)\right), j = 1, \dots, M$$

- For  $j = 1, \dots, M$ , draw from  $p(\boldsymbol{\beta} | \sigma^{2(j)}, \mathbf{y})$ :

$$\boldsymbol{\beta}^{(j)} \sim N\left((X^T X)^{-1} X^T \mathbf{y}, \sigma^{2(j)} (X^T X)^{-1}\right)$$

- The resulting samples  $\{\boldsymbol{\beta}^{(j)}, \sigma^{2(j)}\}_{j=1}^M$  represent  $M$  samples from  $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ .
- $\{\boldsymbol{\beta}^{(j)}\}_{j=1}^M$  are samples from the marginal posterior distribution  $p(\boldsymbol{\beta} | \mathbf{y})$ . This is a *multivariate t* density:

$$p(\boldsymbol{\beta} | \mathbf{y}) = \frac{\Gamma(n/2)}{(\pi(n-p))^{p/2} \Gamma((n-p)/2) |s^2(X^T X)^{-1}|} \left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (X^T X) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(n-p)s^2}\right]^{-n/2}$$

- The marginal distribution of each individual regression parameter  $\beta_j$  is a non-central univariate  $t_{n-p}$  distribution. In fact,

$$\frac{\beta_j - \hat{\beta}_j}{s\sqrt{(X^T X)^{-1}_{jj}}} \sim t_{n-p}.$$

The 95% credible intervals for each  $\beta_j$  are constructed from the quantiles of the  $t$ -distribution. The credible intervals exactly coincide with the 95% classical confidence intervals, but the interpretation is direct: the probability of  $\beta_j$  falling in that interval, given the observed data, is 0.95.

- Note: an intercept only linear model reduces to the simple univariate  $N(\bar{y} | \mu, \sigma^2/n)$  likelihood, for which the marginal posterior of  $\mu$  is:

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}.$$



- Suppose we have observed the new predictors  $\tilde{X}$ , and we wish to predict the outcome  $\tilde{\mathbf{y}}$ . We specify  $p(\tilde{\mathbf{y}}, \mathbf{y} | \boldsymbol{\theta})$  to be a normal distribution:

$$\begin{pmatrix} \mathbf{y} \\ \tilde{\mathbf{y}} \end{pmatrix} \sim N \left( \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \boldsymbol{\beta}, \sigma^2 I \right)$$

- Note  $p(\tilde{\mathbf{y}} | \mathbf{y}, \boldsymbol{\beta}, \sigma^2) = p(\tilde{\mathbf{y}} | \boldsymbol{\beta}, \sigma^2) = N(\tilde{\mathbf{y}} | \tilde{X}\boldsymbol{\beta}, \sigma^2 I)$ .
- The *posterior predictive* distribution:

$$\begin{aligned} p(\tilde{\mathbf{y}} | \mathbf{y}) &= \int p(\tilde{\mathbf{y}} | \mathbf{y}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int p(\tilde{\mathbf{y}} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) d\boldsymbol{\beta} d\sigma^2. \end{aligned}$$

- By now we are comfortable evaluating such integrals:
  - First obtain:  $(\boldsymbol{\beta}^{(j)}, \sigma^{2(j)}) \sim p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ ,  $j = 1, \dots, M$
  - Next draw:  $\tilde{\mathbf{y}}^{(j)} \sim N(\tilde{X}\boldsymbol{\beta}^{(j)}, \sigma^{2(j)} I)$ .

- Suppose that  $\theta = (\theta_1, \theta_2)$  and we seek the posterior distribution  $p(\theta_1, \theta_2 | \mathbf{y})$ .
- For many interesting hierarchical models, we have access to *full conditional distributions*  $p(\theta_1 | \theta_2, \mathbf{y})$  and  $p(\theta_2 | \theta_1, \mathbf{y})$ .
- The *Gibbs sampler* proposes the following sampling scheme. Set starting values  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$  For  $j = 1, \dots, M$ 
  - Draw  $\theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j-1)}, \mathbf{y})$
  - Draw  $\theta_2^{(j)} \sim p(\theta_2 | \theta_1^{(j)}, \mathbf{y})$
- This constructs a *Markov Chain* and, after an initial “burn-in” period when the chains are trying to find their way, the above algorithm guarantees that  $\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=M_0+1}^M$  will be samples from  $p(\theta_1, \theta_2 | \mathbf{y})$ , where  $M_0$  is the burn-in period..

- More generally, if  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$  are the parameters in our model, we provide a set of initial values  $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_p^{(0)})$  and then performs the  $j$ -th iteration, say for  $j = 1, \dots, M$ , by updating successively from the *full conditional* distributions:

$$\boldsymbol{\theta}_1^{(j)} \sim p(\boldsymbol{\theta}_1^{(j)} \mid \boldsymbol{\theta}_2^{(j-1)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

$$\boldsymbol{\theta}_2^{(j)} \sim p(\boldsymbol{\theta}_2^{(j)} \mid \boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

...

(the generic  $k^{th}$  element)

$$\boldsymbol{\theta}_k^{(j)} \sim p(\boldsymbol{\theta}_k^{(j)} \mid \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{k-1}^{(j)}, \boldsymbol{\theta}_{k+1}^{(j)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

...

$$\boldsymbol{\theta}_p^{(j)} \sim p(\boldsymbol{\theta}_p^{(j)} \mid \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{p-1}^{(j)}, \mathbf{y})$$

- Example: Consider the linear model. Suppose we set  $p(\sigma^2) = IG(\sigma^2 | a, b)$  and  $p(\boldsymbol{\beta}) \propto 1$ .
- The full conditional distributions are:

$$p(\boldsymbol{\beta} | \mathbf{y}, \sigma^2) = N(\boldsymbol{\beta} | (X^T X)^{-1} X^T \mathbf{y}, \sigma^2 (X^T X)^{-1})$$

$$p(\sigma^2 | \mathbf{y}, \boldsymbol{\beta}) = IG\left(\sigma^2 | a + n/2, b + \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right).$$

- Thus, the Gibbs sampler will initialize  $(\boldsymbol{\beta}^{(0)}, \sigma^{2(0)})$  and draw, for  $j = 1, \dots, M$ :
  - Draw  $\boldsymbol{\beta}^{(j)} \sim N((X^T X)^{-1} X^T \mathbf{y}, \sigma^{2(j-1)} (X^T X)^{-1})$
  - Draw  $\sigma^{2(j)} \sim IG\left(a + n/2, b + \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta}^{(j)})^T(\mathbf{y} - X\boldsymbol{\beta}^{(j)})\right)$

- In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful - and often easier to code - algorithm is the Metropolis-Hastings (MH) algorithm.
- This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.

- The Metropolis-Hastings algorithm: Start with a initial value for  $\theta = \theta^{(0)}$ . Select a *candidate* or *proposal* distribution from which to propose a value of  $\theta$  at the  $j$ -th iteration:  $\theta^{(j)} \sim q(\theta^{(j-1)}, \nu)$ . For example,  $q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu)$  with  $\nu$  fixed.

- Compute

$$r = \frac{p(\theta^* | \mathbf{y})q(\theta^{(j-1)} | \theta^*, \nu)}{p(\theta^{(j-1)} | \mathbf{y})q(\theta^* | \theta^{(j-1)}, \nu)}$$

- If  $r \geq 1$  then set  $\theta^{(j)} = \theta^*$ . If  $r \leq 1$  then draw  $U \sim (0, 1)$ . If  $U \leq r$  then  $\theta^{(j)} = \theta^*$ . Otherwise,  $\theta^{(j)} = \theta^{(j-1)}$ .
- Repeat for  $j = 1, \dots, M$ . This yields  $\theta^{(1)}, \dots, \theta^{(M)}$ , which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio  $r$  of the sampler through the iterations. Rough recommendations: for vector updates  $r \approx 20\%$ , for scalar updates  $r \approx 40\%$ . This can be controlled by “tuning”  $\nu$ .
- Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.

- Example: For the linear model, our parameters are  $(\beta, \sigma^2)$ . We write  $\theta = (\beta, \log(\sigma^2))$  and, at the  $j$ -th iteration, propose  $\theta^* \sim N(\theta^{(j-1)}, \Sigma)$ . The log transformation on  $\sigma^2$  ensures that all components of  $\theta$  have support on the entire real line and can have meaningful proposed values from the multivariate normal. But we need to transform our prior to  $p(\beta, \log(\sigma^2))$ .
- Let  $z = \log(\sigma^2)$  and assume  $p(\beta, z) = p(\beta)p(z)$ . Let us derive  $p(z)$ . **REMEMBER:** we need to adjust for the jacobian. Then  $p(z) = p(\sigma^2)|d\sigma^2/dz| = p(e^z)e^z$ . The jacobian here is  $e^z = \sigma^2$ .
- Let  $p(\beta) = 1$  and an  $p(\sigma^2) = IG(\sigma^2 | a, b)$ . Then log-posterior is:

$$-(a + n/2 + 1)z + z - \frac{1}{e^z} \left\{ b + \frac{1}{2}(Y - X\beta)^T(Y - X\beta) \right\}.$$

- A symmetric proposal distribution, say  $q(\theta^* | \theta^{(j-1)}, \Sigma) = N(\theta^{(j-1)}, \Sigma)$ , cancels out in  $r$ . In practice it is better to compute  $\log(r)$ :  $\log(r) = \log(p(\theta^* | \mathbf{y})) - \log(p(\theta^{(j-1)} | \mathbf{y}))$ . For the proposal,  $N(\theta^{(j-1)}, \Sigma)$ ,  $\Sigma$  is a  $d \times d$  variance-covariance matrix, and  $d = \dim(\theta) = p + 1$ .
- If  $\log r \geq 0$  then set  $\theta^{(j)} = \theta^*$ . If  $\log r \leq 0$  then draw  $U \sim (0, 1)$ . If  $U \leq r$  (or  $\log U \leq \log r$ ) then  $\theta^{(j)} = \theta^*$ . Otherwise,  $\theta^{(j)} = \theta^{(j-1)}$ .
- Repeat the above procedure for  $j = 1, \dots, M$  to obtain samples  $\theta^{(1)}, \dots, \theta^{(M)}$ .