

Bayesian Statistics

Debdeep Pati
Florida State University

September 8, 2016

Complicated Models: The linear model

- ▶ $Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$
- ▶ $\epsilon \sim N(0, \sigma^2 I_n)$.
- ▶ What are the unknowns? (β and σ)
- ▶ Usually $\beta \sim N(\beta_0, \Sigma_0)$ and $\sigma^{-2} \sim \text{IG}(a, b)$
- ▶ We want to find $\beta, \sigma^2 \mid Y$
- ▶ But easier to find $\beta \mid \sigma^2, Y$ and $\sigma^2 \mid \beta, Y$
- ▶ $\beta \mid \sigma^2, Y$ is a Normal distribution and $\sigma^2 \mid \beta, Y$ is an IG distribution.
- ▶ How to use $\beta \mid \sigma^2, Y$ and $\sigma^2 \mid \beta, Y$ to sample from $\beta, \sigma^2 \mid Y$?

- ▶ Likelihood:

$$L(\mathbf{y}; \mathbf{x}, \beta, \tau) = \prod_{i=1}^n (2\pi\tau^{-1})^{1/2} \exp\{-\tau/2(y_i - \mathbf{x}'_i\beta)^2\}, \text{ where } \tau = \sigma^{-2}$$

- ▶ $\pi(\beta, \sigma^2) = \mathbf{N}_p(\beta; \beta_0, \Sigma_0)\text{Ga}(\tau; a_\tau, b_\tau)$.
- ▶ The hyperparameters β_0, Σ_0 quantify our state of knowledge about the regression parameters β prior to observing the data from the current study
- ▶ In particular, β_0 is our best guess for β before looking at the current data & Σ_0 expresses uncertainty in this guess

- ▶ The prior for the error precision follows the gamma density

$$\pi(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau-1} \exp(-b_\tau \tau)$$

which has expectation $E(\tau) = a_\tau/b_\tau$ and $V(\tau) = a_\tau/b_\tau^2$.

- ▶ Hyperparameters a_τ, b_τ are chosen to express knowledge about τ .

- ▶ After specifying the prior, we update the prior to incorporate information in the likelihood using Bayes rule.
- ▶ This updating process yields the posterior distribution:

$$\pi(\beta, \tau | \mathbf{y}, \mathbf{x}) = \frac{\pi(\beta, \tau)L(\mathbf{y}; \mathbf{x}, \beta, \tau)}{\int \pi(\beta, \tau)L(\mathbf{y}; \mathbf{x}, \beta, \tau)d\beta d\tau} = \frac{\pi(\beta, \tau)L(\mathbf{y}; \mathbf{x}, \beta, \tau)}{\pi(\mathbf{y}; \mathbf{x})}$$

where $\pi(\mathbf{y}; \mathbf{x})$ is the marginal likelihood of the data (obtained by integrating the likelihood across the prior for the parameters)

The linear model

- ▶ The conditional posterior for the regression coefficients can be derived as follows: $\pi(\beta \mid \mathbf{y}, \mathbf{x}, \tau)$
- ▶ Let $\mathbf{X}' = [x_1, \dots, x_n]$.

$$\begin{aligned}\pi(\beta \mid \mathbf{y}, \mathbf{x}, \tau) &\propto \pi(\beta)L(\mathbf{y}; \mathbf{x}, \beta, \tau) \\ &\propto \exp\left\{-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1}(\beta - \beta_0)\right\} \\ &\times \exp\left\{-\frac{1}{2} \sum_{i=1}^n \tau (y_i - x_i' \beta)^2\right\} \\ &\propto \exp\left[-\frac{1}{2}\left\{\beta'(\Sigma_0^{-1} + \tau \sum_{i=1}^n x_i x_i')\beta - 2\beta'(\beta_0 + \tau \sum_{i=1}^n x_i y_i)\right\}\right] \\ &\propto N_p(\beta; \hat{\beta}, \hat{\Sigma}_\beta),\end{aligned}$$

The linear model

- ▶ Thus, the posterior distribution of β given τ is multivariate normal.
- ▶ The posterior mean is

$$\hat{\beta} = E(\beta | \tau, \mathbf{y}, \mathbf{x}) = \hat{\Sigma}_{\beta}(\Sigma_0^{-1}\beta_0 + \tau X'y)$$

- ▶ The posterior variance is

$$\hat{\Sigma}_{\beta} = V(\beta | \tau, \mathbf{y}, \mathbf{x}) = (\Sigma_0^{-1} + \tau X'X)^{-1}$$

- ▶ Note that in the limiting case as the prior variance increases, $\hat{\beta} \rightarrow (X'X)^{-1}X'y$, which is simply the least squares estimator or MLE
- ▶ Hence, the posterior mean is shrunk back towards the prior mean β_0 to a degree dependent on the prior variance.

- ▶ We can similarly derive the posterior distribution of τ :

$$\pi(\tau \mid \mathbf{y}, \mathbf{x}, \beta) \propto \text{Ga}(\tau; a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i' \beta)^2)$$

σ_0^2 arbitrary

$$f(\mathbf{b} \mid \sigma^2 = \sigma_0^2, \mathbf{y})$$

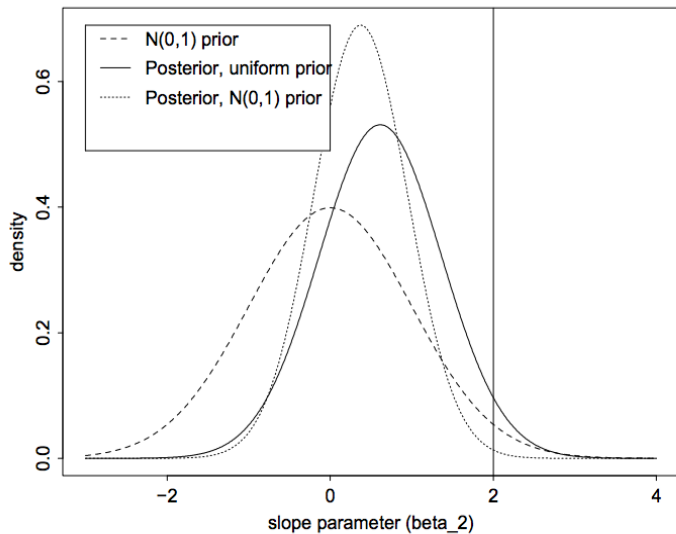
$$f(\sigma^2 \mid \mathbf{b} = \mathbf{b}_0, \mathbf{y})$$

- ▶ Suppose for example that we have a simple linear regression model

$$y_i = \beta_0 + \beta_1 \text{dose}_i + \epsilon_i, \epsilon_i \sim N(0, 1)$$

- ▶ We simulate data under the true model: $\beta = (-1, 2)$, $n = 25$, $\text{dose}_i \sim \text{Uniform}(0, 1)$
- ▶ We consider priors $\pi(\beta) \propto 1$ & $\pi(\beta) = N(0, I_2)$.

Posterior



Some comments

- ▶ For a uniform prior on β posterior is centered on the least squares estimator (specific to normal linear models)
- ▶ For an informative prior, posterior mean is shrunk back towards prior mean and posterior variance decreases
- ▶ As sample size increases, the contribution of the prior is swamped out by the likelihood
- ▶ Hence, as $n \rightarrow \infty$, the posterior will be centered on the MLE regardless of the prior & frequentist/Bayes inferences will be similar
- ▶ However, for finite samples, there can be substantial differences
- ▶ Choosing a $N(0, I)$ prior results in a type of shrinkage estimator

- ▶ Since the result of Stein (1955) and James & Stein (1960) (MLE is inadmissible for $p \geq 3$), shrinkage estimators have been very popular
- ▶ Choosing a $N(0, \kappa I_p)$ prior for β , results in a ridge regression (Hoerl and Kennard, 1970) estimator
- ▶ Hence, priors for the regression parameters having diagonal covariance are commonly referred to as ridge regression priors.
- ▶ For a recent article on shrinkage estimators and properties, refer to Maruyama & Strawderman (2005, Annals of Statistics 33, 1753-1770).

Latent variable for binary response models

- ▶ Y_i binary response x_i predictors, $i = 1, \dots, n$.
- ▶ Probit Model:

$$P(y_i = 1 \mid x_i, \beta) = \Phi(x_i' \beta)$$

- ▶ In toxicology studies, dose is the explanatory variable and there exists a latent variable V denoting the minimum level of dose needed to produce a response (i.e., tolerance)
- ▶ Under the second formulation, $y_i = 1$ if $x_i' \beta > v_i$
- ▶ It follows that $P(Y = 1 \mid x_i) = P(V \leq x_i' \beta)$.
- ▶ Note that the shape of the dose-response curve is determined by the distribution function of V
- ▶ If $V \sim N(0, 1)$, then $P(Y = 1 \mid x_i) = \Phi(x_i' \beta)$

Example: Modeling the risk of preterm birth

- ▶ Let $y_i = 1$ if preterm birth and $y_i = 0$ if full-term birth
 - ▶ $x_i = (1, dde_i, x_{i3}, \dots, x_{i7})'$
 - ▶ x_{i3}, \dots, x_{i7} represent possible confounders
 - ▶ $\beta_1 =$ intercept
 - ▶ $\beta_2 =$ dde slope

Example: Modeling the risk of preterm birth

- ▶ Prior: $\pi(\beta) = N(\beta_0, \Sigma_\beta)$
- ▶ Likelihood:

$$L(y; \beta, \mathbf{x}) = \prod_{i=1}^n \Phi(x'_i \beta)^{y_i} \{1 - \Phi(x'_i \beta)\}^{1-y_i}$$

- ▶ Posterior: $\pi(\beta | y, \mathbf{x}) \propto \pi(\beta)L(y; \beta, \mathbf{x})$
- ▶ No closed form available for the normalizing constant.

Example: Modeling the risk of preterm birth

- ▶ Full conditional posterior distributions needed for Gibbs sampling are not automatically available
- ▶ However, we can rely on a very useful data augmentation trick proposed by [Albert and Chib \(1993\)](#):
- ▶ Augment observed data $\{y_i, x_i\}$ with latent z_i .
- ▶ Probit model can be expressed in hierarchical form as follows:

$$y_i = 1(z_i > 0), z_i \sim N(x_i' \beta, 1)$$

- ▶ Marginalizing out z_i , we obtain $P(y_i = 1 \mid x_i, \beta) = \Phi(x_i' \beta)$.

- ▶ Gibbs sampling relies on alternately sampling from full conditional posterior distributions of unknown parameters
- ▶ After data augmentation, unknowns include latent data $\{z_i\}$ and regression parameters β
- ▶ Full conditional posterior distributions:
 - ▶ $\pi(z_i | \mathbf{y}, \mathbf{x}, \beta) = N(x_i' \beta)$ truncated below by zero if $y_i = 1$ and above by zero if $y_i = 0$.
 - ▶ $\pi(\beta | \mathbf{z}, \mathbf{y}, \mathbf{x}) = N_p(\hat{\beta}, \hat{\Sigma}_\beta)$, $\hat{\Sigma}_\beta = (\Sigma_\beta^{-1} + X'X)^{-1}$, $\hat{\beta} = \hat{\Sigma}_\beta(\Sigma_\beta^{-1}\beta_0 + X'z)$.