Daeyoung Lim

International Finance Seminar

Fall Semester, 2016

1. Consider the following linear regression model:

$$\sigma^2 \sim \text{InvGam}\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right) \tag{1}$$

$$\beta \sim \mathcal{N}(\beta_0, B_0) \tag{2}$$

$$Y \mid \beta, \sigma^2 \sim \mathcal{N}\left(X\beta, \sigma^2 I_T\right) \tag{3}$$

with $Y = T \times 1$ vector and $X = T \times k$ matrix. Then, derive the full conditional distribution of β and σ^2.

Solution: It is well-known that the priors in this model are conditionally conjugate, which means the full conditionals are of the same parametric families as their prior distributions. Therefore, we can compute them as follows.

$$L(\beta, \sigma^2 \mid Y)\pi(\beta)\,\pi\left(\sigma^2\right) \propto \mathcal{N}\left(Y \mid X\beta, \sigma^2 I_T\right) \cdot \mathcal{N}(\beta \mid \beta_0, B_0) \cdot \text{InvGam}\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right) \tag{4}$$

$$\propto \left(\sigma^2\right)^T \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \times \exp\left(-\frac{1}{2}(\beta - \beta_0)' B_0^{-1}(\beta - \beta_0)\right) \tag{5}$$

$$\times \left(\sigma^2\right)^{-(\alpha_0/2+1)} \exp\left(-\frac{\delta_0}{2}\frac{1}{\sigma^2}\right) \tag{6}$$

Therefore, ignoring all the terms that don't have β, we get

$$\pi\left(\beta \mid Y, \sigma^2\right) \propto \exp\left(-\frac{1}{2}\left(\beta'\left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)\beta - 2\left(\frac{1}{\sigma^2}X'y + B_0^{-1}\beta_0\right)'\beta\right)\right) \tag{7}$$

This is exactly the kernel of a normal distribution. Thus, $\beta \mid Y, \sigma^2 \sim \mathcal{N}(\beta_T, \Sigma_T)$ where

$$\Sigma_T = \left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)^{-1} \tag{8}$$

$$\beta_T = \Sigma_T\left(\frac{1}{\sigma^2}X'y + B_0^{-1}\beta_0\right) \tag{9}$$

Likewise,

$$\pi\left(\sigma^2 \mid Y, \beta\right) \propto \left(\sigma^2\right)^{-((T+\alpha_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}\left(\delta_0 + (y - X\beta)'(y - X\beta)\right)\right) \tag{10}$$

Therefore,

$$\sigma^2 \mid Y, \beta \sim \text{InvGam}\left(\frac{\alpha_0 + T}{2}, \frac{1}{2}\left(\delta_0 + (y - X\beta)'(y - X\beta)\right)\right) \tag{11}$$

2. (Greenberg, chapter 2) Consider the uniform distribution with density function $f(y_i \mid \theta) = \theta^{-1}, 0 \leq y_i \leq \theta$, and θ is unknown.

   (a) Show that the Pareto distribution, Pareto(a, k)

   $$\pi(\theta) = \begin{cases} ak^a \theta^{-(a+1)}, & \text{if } \theta \geq k \text{ and } a > 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

   is a conjugate prior distribution for the uniform distribution assuming that $a > 1$. Hint: If $a > 1$, then $\mathbf{E}(\theta) = ak/(a-1)$.

   (b) Show that $\widehat{\theta} = \max(y_1, y_2, \ldots, y_n)$ is the MLE of θ, where the y_i is the random variable from $f(y_i \mid \theta)$.

   (c) Find the posterior distribution of θ and the expected value.

Solution:

(a) Conjugacy implies that the prior and posterior distributions are in the same parametric distributional family. Therefore, we simply need to show that the posterior distribution is also a Pareto distribution.

$$L(\theta \mid y_1, \ldots, y_n) \propto \pi(\theta) \prod_{i=1}^{n} f(y_i \mid \theta) \mathbf{I}_{(y_{(n)}, \infty)}(\theta) \cdot \mathbf{I}_{(k, \infty)}(\theta) \tag{13}$$

$$\propto \theta^{-(\alpha+1)} \cdot \theta^{-1} \cdot \mathbf{I}_{(y_{(n)} \vee k, \infty)}(\theta) \tag{14}$$

$$\propto \theta^{-(\alpha+2)} \mathbf{I}_{(y_{(n)} \vee k, \infty)}(\theta) \tag{15}$$

Therefore,

$$\theta \mid y_1, \ldots, y_n \sim \text{Pareto}(\alpha + 1, y_{(n)} \vee k) \tag{16}$$

where $\vee$ denotes the maximum operator and $y_{(n)}$ is the maximum order statistic.

(b) We take the logarithm of the likelihood and use the first-order condition to find the maximum.

$$\ell\left(\theta \mid \{y_i\}_{i=1}^{n}\right) \propto -n \log \theta \cdot \mathbf{I}_{(y_{(n)}, \infty)}(\theta) \tag{17}$$

$$\frac{d}{d\theta} \ell(\theta) = -\frac{n}{\theta} \tag{18}$$

Because the first-derivative is never zero, we need only use the monotone property of the likelihood function. Since it is monotonically decreasing, the maximum value is obtained when

θ is smallest, i.e., $\theta \in (y_{(n)}, \infty)$. Therefore,

$$\widehat{\theta}^{\text{MLE}} = \inf\left\{\theta \mid y_{(n)} < \theta\right\} = y_{(n)} \tag{19}$$

(c) We have already found the posterior of θ: Pareto$(\alpha + 1, y_{(n)} \vee k)$. If we let $w = y_{(n)} \vee k$, by definition, the expected value is

$$\mathbf{E}\left(\theta \mid \{y_i\}_{i=1}^n\right) = \int_w^\infty (\alpha+1) w^{\alpha+1} \theta^{-(\alpha+1)}\, d\theta \tag{20}$$

$$= \frac{\alpha+1}{\alpha} w, \quad \text{where } w > 0 \tag{21}$$

3. (Greenberg, chapter 2) The density function of the exponential distribution is

$$f(y_i \mid \theta) = \theta e^{-\theta y_i}, \quad \theta > 0,\ y_i > 0, \tag{22}$$

and let $y_1, y_2, \ldots, y_n$ be a random sample from the distribution.

(a) Show that the gamma distribution Ga(α, β) is a conjugate prior distribution for the exponential distribution. Hint: The density of Ga(α, β) is

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \tag{23}$$

(b) Show that $1/\overline{y}$ is the MLE for θ where $\overline{y}$ is the sample mean of the observations.

(c) Write the mean of the posterior distribution as a weighted average of the mean of the prior distribution and the MLE.

(d) What happens to the weight on the prior mean as n becomes large?

Solution:

(a) Again, we need to show that the posterior is again a gamma distribution.

$$\pi\left(\theta \mid y_1, \ldots, y_n\right) \propto \pi(\theta) \prod_{i=1}^n f(y_i \mid \theta) \tag{24}$$

$$\propto \theta^{n+\alpha-1} \exp\left(-\left(\beta + \sum_{i=1}^n y_i\right)\theta\right) \tag{25}$$

$$\sim \text{Ga}\left(\alpha + n, \beta + \sum_{i=1}^n y_i\right) \tag{26}$$

It is easily recognized that the kernel of the posterior is again a gamma distribution.

(b) Using the logarithm and the first-order condition,

$$\ell(\theta \mid y_1, \ldots, y_n) \propto n \log \theta - \theta \sum_{i=1}^{n} y_i \tag{27}$$

$$\frac{d}{d\theta} \ell\left(\theta \mid y_1, \ldots, y_n\right) = \frac{n}{\theta} - \sum_{i=1}^{n} y_i = 0 \tag{28}$$

$$\widehat{\theta}^{\mathrm{MLE}} = n \Bigg/ \left(\sum_{i=1}^{n} y_i \right) \tag{29}$$

$$= \frac{1}{\overline{y}} \tag{30}$$

(c) For a gamma distribution with parameters α and β, the expected value is α/β. Thus, the posterior expected value is

$$\mathbf{E}\left(\theta \mid y_1, \ldots, y_n\right) = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} y_i} \tag{31}$$

The posterior mean can be decomposed in to a weighted average between the prior mean and the MLE as follows:

$$\frac{\alpha + n}{\beta + \sum_{i=1}^{n} y_i} = \frac{\beta}{\beta + \sum_{i=1}^{n} y_i} \times \frac{\alpha}{\beta} + \frac{\sum_{i=1}^{n} y_i}{\beta + \sum_{i=1}^{n} y_i} \times \frac{n}{\sum_{i=1}^{n} y_i} \tag{32}$$

(d) Taking the limit, the weight on the prior mean tends to infinity,

$$\frac{\beta}{\beta + \sum_{i=1}^{n} y_i} \xrightarrow{n \to \infty} \infty \tag{33}$$

because $\sum_{i=1}^{n} y_i \to \infty$. Furthermore, the weight on the MLE converges to 1 with which we can conclude that the more we data, the closer the posterior mean gets to MLE. In short, with a large amount of data, the posterior mean is not very different from MLE.

4. (Greenberg, chapter 3) Compute the predictive distribution for y_{n+1} if the y_i have independent normal distributions $\mathcal{N}(\mu, 1)$, where the prior distribution for μ is $\mathcal{N}(\mu_0, \sigma_0^2)$.

Solution: First, the posterior distribution of μ is

$$\pi\left(\mu \mid y_1, \ldots, y_n\right) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right) \times \exp\left(-\frac{1}{2\sigma_0^2}\left(\mu-\mu_0\right)^2\right) \tag{34}$$

$$\propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2+n\left(\overline{y}-\mu\right)^2\right)-\frac{1}{2\sigma_0^2}\left(\mu-\mu_0\right)^2\right) \tag{35}$$

$$\propto \exp\left(-\frac{n}{2}\left(\mu^2-2\overline{y}\mu\right)-\frac{1}{2\sigma_0^2}\left(\mu^2-2\mu_0\mu\right)\right) \tag{36}$$

$$\propto \exp\left(-\frac{1}{2}\left(\left(n+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(n\overline{y}+\frac{1}{\sigma_0^2}\mu_0\right)\mu\right)\right) \tag{37}$$

Therefore, $\mu \mid y_1, \ldots, y_n \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$ where

$$\sigma_n^2 = \left(n+\frac{1}{\sigma_0^2}\right)^{-1} \tag{38}$$

$$\mu_n = \sigma_n^2\left(n\overline{y}+\frac{\mu_0}{\sigma_0^2}\right) \tag{39}$$

By definition of the predictive distribution,

$$p\left(y_{n+1} \mid y_n, \ldots, y_1\right) = \int_{-\infty}^{\infty} p(y_{n+1} \mid \mu) \cdot \pi(\mu \mid y_1, \ldots, y_n)\, d\mu \tag{40}$$

$$= \frac{e^{y_{n+1}^2+\mu_n^2/\sigma_n^2}}{2\pi\sigma_n}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\left(1+\frac{1}{\sigma_n^2}\right)\mu^2-2\left(y_{n+1}+\mu_n\right)\mu\right)\right) \tag{41}$$

$$= \frac{\exp\left(y_{n+1}^2+\mu_n^2/\sigma_n^2\right)}{\sqrt{2\pi\sigma_n^2}} \cdot \left(1+\frac{1}{\sigma_n^2}\right)^{-1} \tag{42}$$

5. Show that the median of the posterior distribution minimizes the absolute loss function.

Solution: Let X be a random variable with a distribution function F which is a legitimate measure and let $\mathbf{E}$ denote the expectation with respect to the measure F. Then, if we let m denote the median, we can assume that F is zero to the left of an arbitrary constant A and one to the right of B.

$$\phi = \mathbf{E}\left|X-m\right| \tag{43}$$

$$= \int_A^m (m-x)\, dF + \int_m^B (x-m)\, dF \tag{44}$$

By interchanging the differentiation and integral (since we need to differentiate so as to get the

minimum—first-order condition),

$$\frac{d\phi}{dm} = \int_A^m 1\, dF - \int_m^B 1\, dF = 0 \tag{45}$$

We should solve

$$\int_A^m 1\, dF = \int_m^B 1\, dF \tag{46}$$

which is essentially comparing two probability measures, $\Pr(X < m) = \Pr(X > m)$. Since the probability below a constant m and the probability above it are identical, by definition, m is the median that minimizes the absolute loss function. By the same logic, the posterior density must be a Radon-Nikodym derivative of some absolutely continuous measure P with respect to either the Lebesgue measure or the counting measure. Then, the posterior median must be the solution to

$$\int_A^m 1\, dP = \int_m^B 1\, dP \tag{47}$$

which translates to $\Pr(\theta < m \mid X) = \Pr(\theta > m \mid X)$. Therefore, the posterior median minimizes the absolute loss function $\mathbf{E}\left(|\theta - m| \mid X\right)$.

6. A vector of observations at time t, $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{k,t})'$ has a multinomial distribution, a priori, with $\sum_{i=1}^k x_{i,t} = N$ and probabilities $p = (p_1, p_2, \ldots, p_k)$. p_i is the probability of $x_{i,t}$ and $\sum_{i=1}^k p_i = 1$. Then, its joint probability of x_t is given by

$$f(x_t \mid p) = \frac{N!}{\prod_{i=1}^k x_{i,t}!} \prod_{i=1}^k p_i^{x_{i,t}}, \qquad 0 \le x_{i,t} \le N. \tag{48}$$

Note that x_t is identically and independently distributed for $t = 1, 2, \ldots, T$. Then, we assume that the joint prior distribution for p is a Dirichlet distribution with parameter $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ and its density is given by

$$\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}, \qquad 0 < p_i < 1, \;\; \alpha_i > 0, \;\; i = 1, 2, \ldots, k \tag{49}$$

Derive the posterior distribution of p.

Solution: Multiplying the likelihood function and the prior,

$$\pi\left(p \mid \mathbf{x}\right) \propto \pi\left(p\right) \prod_{t=1}^{T} f(x_t \mid p) \tag{50}$$

$$\propto \prod_{i=1}^{k} p_k^{\alpha_i - 1} \prod_{t=1}^{T} \prod_{i=1}^{k} p_i^{x_{i,k}} \tag{51}$$

$$\propto \prod_{i=1}^{k} p_i^{\sum_{t=1}^{T} x_{i,k} + \alpha_i - 1} \tag{52}$$

This is the kernel of a Dirichlet distribution. Thus,

$$p \mid \mathbf{x} \sim \text{Dir}\left(\sum_{t=1}^{T} x_{1,t} + \alpha_1, \ldots, \sum_{t=1}^{T} x_{k,t} + \alpha_k\right) \tag{53}$$

Since the posterior belongs to the same parametric family as the prior, it is conjugate.

7. Show that the mode of the posterior distribution minimizes the following loss function

$$L_3(\widehat{\theta} \mid \theta) = \mathbf{I}\left(\left|\widehat{\theta} - \theta\right| > b\right) \tag{54}$$

for any small $b > 0$.

Solution: The expected zero-one loss is computed as follows.

$$\mathbf{E}\left(\mathbf{I}\left(\left|\theta - \widehat{\theta}\right| > b\right) \,\middle|\, X\right) = \int_{-\infty}^{\infty} \mathbf{I}\left(\left|\theta - \widehat{\theta}\right| > b\right) p\left(\theta \mid X\right)\, d\theta \tag{55}$$

$$= \int_{-\infty}^{\infty} \left(1 - \mathbf{I}\left(\left|\theta - \widehat{\theta}\right| \leq b\right)\right) p\left(\theta \mid X\right)\, d\theta \tag{56}$$

$$= 1 - \int_{\left|\theta - \widehat{\theta}\right| \leq b} p\left(\theta \mid X\right)\, d\theta \tag{57}$$

$$= 1 - \Pr\left(\left|\theta - \widehat{\theta}\right| \leq b \,\middle|\, X\right) \tag{58}$$

$$= 1 - \Pr\left(\widehat{\theta} - b \leq \theta \leq \widehat{\theta} + b \,\middle|\, X\right) \tag{59}$$

With a small arbitrary constant b, the posterior probability measure on the neighborhood of $\widehat{\theta}$ of distance b is maximized when $\widehat{\theta}$ is the posterior mode. This is equivalent to saying 1 minus the posterior probability measure on the neighborhood of $\widehat{\theta}$ of distance b is minimized when $\widehat{\theta}$ is the posterior mode. Thus, the posterior mode minimizes the expected zero-one loss.

8. Explain the following terminologies.

   (1) Prior distribution

   (2) Model

   (3) Posterior distribution

   (4) Likelihood

   (5) Marginal likelihood

   (6) Posterior predictive distribution

   (7) Posterior predictive density

   (8) Predictive likelihood

   (9) Posterior probability of models

   (10) Full conditional distribution

   (11) Gibbs sampling

Solution:

(1) A prior distribution is the distribution that encodes the belief that one has about the parameter before observing data.

(2) A model is essentially a representation which describes the data generating process that the researcher believes in.

(3) The posterior distribution is the updated distribution of the parameter after having incoporated all the observed data.

(4) The likelihood is a function of the parameter given data. Although the likelihood is the same as the joint probability of the data, the distinction should be made between the two in that we only discuss *probability* before we observe the data whereas the likelihood is discussed after we have obtained the data.

(5) A marginal likelihood is mathematically the expression in which the parameter has been integrated out. Intuitively, it is the distribution of the data that involves the uncertainty of the parameter. It also has an interpretation in terms of the predictive distribution.

(6) The posterior predictive distribution is the distribution of the first observation, y_{T+1} in the future given the observations up to time T, $y_1, \ldots, y_T$.

(7) The posterior predictive density is the distribution of future H observations,$y_{T+1}, \ldots, y_{T+H}$ given the observations up to time T, $y_1, \ldots, y_T$.

(8) The predictive likelihood is

(9) The posterior probability of models is the probability of postulated models, $M_1, M_2, \ldots$ after having observed data. Normally, the prior for each model is assumed to be equal across all models.

(10) The full conditional distribution is the distribution of a chosen parameter given everything else.

(11) Gibbs sampling is an MCMC algorithm that can be used if every parameter is conditionally conjugate, which is a special case of the Metropolis-Hastings algorithm.