ECON5529 Bayesian Theory

Ellis Scharfenaker
Fall 2017, Lecture 4

Multiparameter Models

- For most practical models we are interested in more than a single unknown parameter. In this case we are interested in sampling from a higher dimensional joint distribution of all unknowns.
- In some cases all parameters will be of interest to us, but in others they may merely be a *nuisance*. In this case, the ultimate aim is to obtain the *marginal* posterior distribution of the particular parameters of interest.
- In general, multiparameter models pose a more challenging optimization problem.

Marginal Distributions

- The posterior distribution in a Bayesian analysis describes a joint distribution over all of the parameters of a model.
- For example, in the linear regression model,

$$\begin{aligned} y &= \beta_0 + \beta_1\, x + \epsilon \\ \epsilon &= y - (\beta_0 + \beta_1\, x) \end{aligned} \tag{1}$$

- We typically use the normal distribution to model the error terms making the likelihood:

$$p\left[\epsilon \,\middle|\, \sigma^2, \beta_0, \beta_1\right] = \frac{1}{\sqrt{2\,\pi\sigma^2}}\, e^{-\frac{\epsilon^2}{2\,\sigma^2}} = \frac{1}{\sqrt{2\,\pi\sigma^2}}\, e^{-\frac{(y-(\beta_0+\beta_1\,x))^2}{2\,\sigma^2}} \tag{2}$$

- The parameters are the intercept β_0, the slope β_1, and σ, the standard deviation by which we scale the errors.
- But we are often much more interested in what the data analysis tells us about some parameters than about others.
- In these cases the less interesting parameters are called *nuisance parameters*. In the regression example, the error standard deviation σ is often regarded as such a nuisance parameter.
- What we really want to know is the posterior distribution of β_0 and β_1 properly averaging out over all the possibilities for σ.
- In principle it is easy enough to see that we would like to know the marginal distribution for the parameter(s) of interest after integrating out over the posterior probability of the nuisance parameters, for example:

$$p[\beta_0, \beta_1 \mid x] = \int p[\beta_1, \beta_0, \sigma \mid x]\, d\sigma \tag{3}$$

Nuisance Parameters

- Suppose that our multiparameter model contains two unknown or unobserved quantities $\theta = \{\theta_1, \theta_2\}$ and that we are only really interested in making inferences with θ_1. In such a case θ_2 is the *nuisance parameter.*
- In this situation we are interested in the conditional distribution of the parameter of interest given the data, $p[\theta_1 \mid x]$, but face the possible dependence of θ_1 on θ_2 so that $p[\theta_1 \mid x, \theta_2] \neq p[\theta_1 \mid x]$
- To obtain the conditional distribution of interest ($p[\theta_1 \mid x]$) apply the product rule to the joint posterior for the unknown parameter:

$$p[\theta_1, \theta_2 \mid x] = p[\theta_1 \mid \theta_2, x]\, p[\theta_2 \mid x]\, p[x] \tag{4}$$

- But note the $p[x]$ cancels out when normalizing.

$$p[\theta_1, \theta_2 \mid x] = p[\theta_1 \mid \theta_2, x]\, p[\theta_2 \mid x] \tag{5}$$

- Averaging over the nuisance parameter θ_2 we properly eliminate it from the posterior distribution:

$$p[\theta_1 \mid x] = \int_{\theta_2} p[\theta_1 \mid \theta_2, x]\, p[\theta_2 \mid x]\, d\theta_2 \tag{6}$$

- This result shows that the marginal posterior distribution ($p[\theta_1 \mid x]$) is a mixture of the conditional posterior distributions given θ_2, ($p[\theta_1 \mid \theta_2, x]$) weighted by $p[\theta_2 \mid x]$, the possible values of θ_2.
- The weights, $p[\theta_2 \mid x]$, are the *posterior density* $p[\theta_2 \mid x] \propto p[x \mid \theta_2]\, p[\theta_2]$ of θ_2.
- Rarely do we evaluate the integral (6) explicitly as posterior distributions can be obtained by marginal and conditional simulation.
- Simulating an integral like (6) requires a **sampling strategy**. In this case it is quite straightforward.
  - (1) Draw a sample $\tilde{\theta}_2$ from its marginal posterior distribution $p[\theta_2 \mid x]$
  - (2) Sample $\tilde{\theta}_1$ from its conditional posterior distribution, $p\left[\theta_1 \mid \tilde{\theta}_2, x\right]$, using the $\tilde{\theta}_2$ draw from step 1.
- In this way the integral (6) is computed indirectly and much of the remainder of the course will rely on posterior sampling methods such as this.

Example: Normal Data with Conjugate prior

- To illustrate multiparameter modeling and posterior marginalization of nuisance parameters let's take the rather typical example of estimating the mean of a population from a sample.
- The data in this model is a list $x = \{x_1, x_2, \ldots, x_n\}$ of n independent observations from a univariate normal distribution $x \sim N\left[\mu, \sigma^2\right]$.
- One choice for a noninformative prior for μ and σ^2 assuming prior independence is Jeffreys' prior, which in this case is $p\left[\mu, \sigma^2\right] \propto \sigma^{-2}$
- The **joint posterior** is then equal to the likelihood multiplied by the factor σ^{-2}

$$\begin{aligned} p\left[\mu, \sigma^2 \mid x\right] &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \sigma^{-2} \\ &\propto \sigma^{-n-2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= \sigma^{-n-2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\tilde{x})^2+n(\tilde{x}-\mu)^2\right)} \\ &= \sigma^{-n-2} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\tilde{x}-\mu)^2\right)} \end{aligned} \tag{7}$$

- Where $\tilde{x} = \frac{1}{n} \sum_i x_i$ is the sample mean and $s^2 = \frac{1}{n-1} \Sigma_i (x_i - \tilde{x})^2$ is the sample variance. But remember we are interested in the the conditional posterior distribution of μ after eliminating σ^2.

$$p[\mu \mid x] = \int p\left[\mu, \sigma^2 \,\middle|\, x\right] d\sigma^2 = \int p\left[\mu \,\middle|\, \sigma^2, x\right] p\left[\sigma^2 \,\middle|\, x\right] d\sigma^2 \tag{8}$$

- To calculate the conditional posterior distribution of μ we first consider the conditional posterior density $p\left[\mu \,\middle|\, \sigma^2, x\right]$ and then the marginal posterior distribution $p\left[\sigma^2 \,\middle|\, x\right]$.
- For the mean of a normal distribution with known variance and a conjugate prior distribution we can use the results derived from the previous lecture $p\left[\mu \,\middle|\, \sigma^2, x\right] \sim \mathcal{N}\left[\hat{\mu}, \hat{\tau}^2\right]$ with:

$$\hat{m} = \frac{\frac{1}{\tau^2} m + \frac{1}{\sigma^2} \tilde{x}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}, \ \hat{\tau}^2 = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2}\right)^{-1}$$

- For the marginal distribution $p\left[\sigma^2 \,\middle|\, x\right]$ we need to average over the joint distribution

$$\begin{aligned} p\left[\sigma^2 \,\middle|\, x\right] &\propto \int \sigma^{-n-2} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\tilde{x}-\mu)^2\right)} d\mu \\ &= \sigma^{-n-2} e^{-\frac{1}{2\sigma^2}(n-1)s^2} \int e^{-\frac{1}{2\sigma^2} n (\tilde{x}-\mu)^2} d\mu \end{aligned} \tag{9}$$

- The integral to be evaluated is just a simple normal integral

$$\texttt{Integrate}\left[e^{-\frac{1}{2\sigma^2} n (x-\mu)^2}, \{\mu, -\infty, \infty\}, \texttt{Assumptions} \rightarrow \left\{\texttt{Re}\left[\frac{n}{\sigma^2}\right] \geq 0\right\}\right]$$

$$\frac{\sqrt{2\pi}}{\sqrt{\frac{n}{\sigma^2}}}$$

$$\begin{aligned} p\left[\sigma^2 \,\middle|\, x\right] &\propto \sigma^{-n-2} \sqrt{\frac{2\pi\sigma^2}{n}} e^{-\frac{1}{2\sigma^2}(n-1)s^2} \\ &\propto \left(\sigma^2\right)^{\frac{-(n+1)}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \end{aligned} \tag{10}$$

- In this case $\sigma^2 \sim \mathcal{I}nv.\chi^2[n-1, s^2]$.
- For $\theta \sim \mathcal{I}nv\text{-}\chi^2$, the density function is:

$$p[\theta] = \frac{2^{-\nu/2}}{\Gamma[\nu/2]} \theta^{-\frac{\nu}{2}+1} e^{-\frac{\nu s^2}{2\theta}} \tag{11}$$

- The Inv-χ^2 distribution is the same as a $\mathcal{IG}\left[\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right]$.

Sampling Strategy

- (1) Draw a single σ^2 sample from $p\left[\sigma^2 \,\middle|\, x\right] = \mathcal{I}nv.\chi^2[n-1, s^2]$.
- (2) Given the σ^2 draw μ from $p\left[\mu \,\middle|\, \sigma^2, x\right] = \mathcal{N}\left[\hat{\mu}, \hat{\tau}^2\right]$
- (3) Repeat

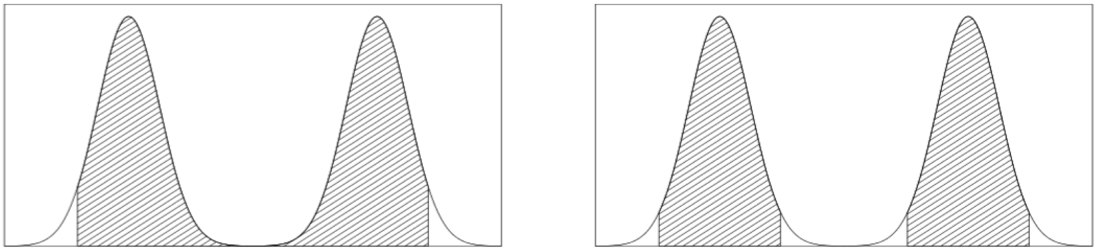
In R:

Reporting the Posterior

- Generally the first stage of investigating the posterior probability of the parameters of a model is to find the maximum posterior probability value of the parameters.
- A local maximum of a probability distribution is often called the *mode*, so the maximum posterior probability parameters are sometimes called the *modal parameters*.
- Other commonly used summaries of posteriors are the mean, and median of the distribution; variation is commonly summarized by the standard deviation, the interquartile range, and other quantiles.

High posterior probability sets or intervals

- If posterior probabilities were always normal, it would suffice to put “error bars” around parameter estimates representing two standard deviations of the posterior probability distribution.
- It turns out that posterior probabilities for relatively small numbers of data points, such as economists often work with, can deviate significantly from normal distributions.
- Hence, in addition to point summaries, it is always important to report ***posterior uncertainty***. Our usual approach is to present a central interval of posterior probability.
- The ***highest posterior density*** (HPD) region is the set of values that contains $100\,(1 - \alpha)$ % of the posterior probability.
- The HPD indicates which points of a distribution are most credible, and which cover most of the distribution.
- The *highest posterior density* is identical to a central posterior interval if the posterior distribution is unimodal and symmetric. However, if the posterior is skewed or multimodal, the HPD can be quite different from a standard confidence interval.



Multinomial Model

- The multinomial model is a generalization of the binomial model for categorical outcomes where the possible qualitative indicators are greater than two.
- In the multinomial model the data is a list of counts of the number of observations of each outcome $x = \{x_1, \ldots, x_k\}$ where x_i is the number of observations for the i^{th} outcome category. The probability of a particular outcome is θ_i where $\sum \theta_i = 1$.
- The likelihood of the data is

$$p[x \mid \theta] \propto \prod_{i=1}^{k} \theta_i^{x_i} = \theta_1^{\sum x_1} \ldots \theta_k^{\sum x_k} \qquad (12)$$

- The conjugate prior for the multinomial model is called the *Dirichlet prior.* Just as the multinomial distribution is a multivariate generalization of the binomial distribution, the Dirichlet distribution is a multivariate generalization of the Beta distribution and is defined as:

  $$\theta \sim \mathcal{D}[\alpha], \quad p[\theta] \propto \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

- Similar to the beta distribution we get a uniform prior on all θ_j when $\alpha_j = 1 \; \forall \; j$.
- The posterior distribution for the unknown probabilities θ_j using the conjugate Dirichlet prior is:

  $$p[\theta \mid x] \propto \prod_{i=1}^{k} \theta_i^{x_i} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} = \prod_{i=1}^{k} \theta_i^{(x_i + \alpha_i - 1)} \propto \mathcal{D}[x + \alpha]$$

Example

- During the 2016 presidential election an NYT/CBS News Poll surveyed $n = 1433$ individuals about. The results were:

	Registered Voters
Clinton	659
Trump	588
Undecided / Other	129

- The difference $\theta_1 - \theta_2$ is the difference between Clinton vs. Trump support and represents the margin of Clinton over Trump.
- If the prior over $(\theta_1, \theta_2, \theta_3)$ is $\mathcal{D}[\theta_1, \theta_2, \theta_3]$ then the posterior is

  $$\mathcal{D}[\alpha_1 + x_1, \alpha_2 + x_2, \alpha_3 + x_3] \propto \prod_{i=1}^{3} \theta_i^{(x_i + \alpha_i - 1)} = \theta_1^{(x_1 + \alpha_1 - 1)} \theta_2^{(x_2 + \alpha_2 - 1)} \theta_3^{(x_3 + \alpha_3 - 1)}$$

- Note that most of the mass of this distribution is on positive values, indicating that there is strong evidence that the proportion of voters for Clinton exceeds the proportion for Trump.
- Now suppose we wish to predict the total number of electoral votes (EV_0) obtained by Clinton. Let $\theta_{C\,j}$ and $\theta_{T\,j}$ denote the unknown proportion of voters respectively for Clinton and Trump in the j^{th} state.
- The number of electoral votes for Clinton (using the indicator function) is:

  $$\text{EV}_0 = \sum_{j=1}^{51} \text{EV}_j \, I\left[\theta_{C\,j} > \theta_{T\,j}\right]$$

- Let $q_{C\,j}$and $q_{T\,j}$ denote the observed data of sample proportions of voters for Clinton and Trump in the j^{th} state.
- Further, assume that each poll is based on a sample of 500 voters.
- We can assign a uniform prior on the vector of proportions, $(\theta_{C\,1}, \theta_{T\,1}), \; \ldots, (\theta_{C\,51}, \theta_{T\,51})$, i.e. for sate j $\left(0_{C\,j}, 100_{T\,j}\right)$ has the same probability as $\left(100_{C\,j}, 0_{T\,j}\right)$.
- Thus, the vectors $(\theta_{C\,1}, \theta_{T\,1}), \; \ldots, (\theta_{C\,51}, \theta_{T\,51})$ have independent posterior distributions, where the proportions favoring the candidates in the j^{th} state, $\left(\theta_{C\,j}, \theta_{T\,j}, 1 - \theta_{C\,j} - \theta_{T\,j}\right)$, have a Dirichlet distribution.
- It would be possible to improve our prediction by using more than a single poll in each state as is commonly done by Nate Silver on FiveThirtyEight.com