

3: Introduction to multiparameter models

Taylor

University of Virginia

Introduction

We discuss a few examples of models with more than one parameter.

A noninformative prior with a normal likelihood

Consider a normal likelihood

$$\begin{aligned} p(y \mid \mu, \sigma^2) &\propto (\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right] \\ &= (\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_i ([y_i - \bar{y}] + [\bar{y} - \mu])^2 \right] \\ &= (\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \left\{ \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 + 0 \right\} \right] \\ &= (\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \{ (n-1)s^2 + n(\bar{y} - \mu)^2 \} \right] \end{aligned}$$

and the noninformative, improper prior $p(\mu, \sigma^2) \propto \sigma^{-2}$. Clearly

$$p(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{1}{2\sigma^2} \{ (n-1)s^2 + n(\bar{y} - \mu)^2 \} \right]$$

A noninformative prior with a normal likelihood

Suppose instead that σ^2 is a nuisance parameter, and we're only interested in μ . Then, we want the marginal posterior.

Let $z = \frac{1}{2\sigma^2} \{(n-1)s^2 + n(\bar{y} - \mu)^2\} = \frac{A}{2\sigma^2}$. Then

$$\begin{aligned} p(\mu | y) &\propto \int_0^\infty (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{1}{2\sigma^2} \{(n-1)s^2 + n(\bar{y} - \mu)^2\} \right] d\sigma^2 \\ &= \int_\infty^0 (A/2)^{-(n+2)/2} z^{(n+2)/2} \exp[-z] (-A/2) z^{-2} dz \\ &= (A/2)^{-n/2} \underbrace{\int_0^\infty z^{n/2-1} \exp[-z] dz}_{\Gamma(n/2)} \end{aligned}$$

A noninformative prior with a normal likelihood

So

$$\begin{aligned}p(\mu|y) &\propto (A/2)^{-n/2} \\&\propto A^{-n/2} \\&\propto A^{-n/2}[(n-1)s^2]^{n/2} \\&\propto \left(1 + \frac{(\bar{y} - \mu)^2}{(n-1)s^2/n}\right)^{-n/2}\end{aligned}$$

$$\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n)$$

A noninformative prior with a normal likelihood

Suppose that μ is a nuisance parameter, and we're only interested in σ^2 . Then, we want the marginal posterior:

$$\begin{aligned} p(\sigma^2 \mid y) &\propto \int (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{1}{2\sigma^2} \{ (n-1)s^2 + n(\bar{y} - \mu)^2 \} \right] d\mu \\ &= (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{(n-1)}{2\sigma^2} s^2 \right] \int \exp \left[-\frac{1}{2\sigma^2} n(\mu - \bar{y})^2 \right] d\mu \\ &\propto (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{(n-1)}{2\sigma^2} s^2 \right] (\sigma^2)^{1/2} \\ &= (\sigma^2)^{-[(n-1)/2+1]} \exp \left[-\frac{(n-1)s^2}{2\sigma^2} \right] \end{aligned}$$

$$\sigma^2 \mid y \sim \text{Inv-Gamma} \left(\frac{n-1}{2}, \frac{(n-1)s^2}{2} \right)$$

A noninformative prior with a normal likelihood

Recall the joint posterior:

$$p(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{1}{2\sigma^2} \{ (n-1)s^2 + n(\bar{y} - \mu)^2 \} \right]$$

Clearly:

$$p(\mu \mid \sigma^2, y) \propto \exp \left[-\frac{n}{2\sigma^2} (\bar{y} - \mu)^2 \right]$$

A noninformative prior with a normal likelihood

Recall the joint posterior:

$$p(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-(n+2)/2} \exp \left[-\frac{1}{2\sigma^2} \{ (n-1)s^2 + n(\bar{y} - \mu)^2 \} \right]$$

Clearly:

$$p(\mu \mid \sigma^2, y) \propto \exp \left[-\frac{n}{2\sigma^2} (\bar{y} - \mu)^2 \right]$$

We also have $p(\sigma^2 \mid y)$ from the last slide. This means that we can figure out the normalizing constants for the joint posterior if we multiply these two known densities together:

$$p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y).$$

Sometimes this is called a **normal-inverse-gamma** distribution.

A noninformative prior with a normal likelihood

After we have figured out the joint posterior, we may be interested in predicting new observations with the **posterior predictive distribution**:

$$p(\tilde{y} | y) = \iint p(\tilde{y} | \mu, \sigma^2) p(\mu, \sigma^2 | y) d\mu d\sigma^2.$$

A noninformative prior with a normal likelihood

After we have figured out the joint posterior, we may be interested in predicting new observations with the **posterior predictive distribution**:

$$p(\tilde{y} | y) = \iint p(\tilde{y} | \mu, \sigma^2) p(\mu, \sigma^2 | y) d\mu d\sigma^2.$$

We can simulate \tilde{y}_i as follows:

- 1 draw $\sigma_i^2 | y \sim p(\sigma^2 | y)$
- 2 draw $\mu_i | \sigma_i^2, y \sim p(\mu | \sigma_i^2, y)$
- 3 draw $\tilde{y}_i | \mu_i, \sigma_i^2 \sim p(\tilde{y} | \mu_i, \sigma_i^2)$

A noninformative prior with a normal likelihood

After we have figured out the joint posterior, we may be interested in predicting new observations with the **posterior predictive distribution**:

$$p(\tilde{y} | y) = \iint p(\tilde{y} | \mu, \sigma^2) p(\mu, \sigma^2 | y) d\mu d\sigma^2.$$

It's a homework question to show that

$$\tilde{y} | y \sim t_{n-1} \left(\bar{y}, s^2 \left(1 + \frac{1}{n} \right) \right)$$

A noninformative prior with a normal likelihood

Let's get some practice simulating predictions, which will come in handy when we are dealing with more complicated scenarios where a closed-form posterior predictive distribution isn't available. We can simulate each \tilde{y}_i as follows:

Sampling Strategy

For $i = 1, 2, \dots$

- 1 draw $\sigma_i^2 \mid y \sim p(\sigma^2 \mid y)$
- 2 draw $\mu_i \mid \sigma_i^2, y \sim p(\mu \mid \sigma_i^2, y)$
- 3 draw $\tilde{y}_i \mid \mu_i, \sigma_i^2 \sim p(\tilde{y} \mid \mu_i, \sigma_i^2)$

Each triple

$$(\tilde{y}_i, \mu_i, \sigma_i^2) \sim p(\tilde{y}, \mu, \sigma^2 \mid y) = p(\tilde{y} \mid \mu, \sigma^2) p(\mu \mid \sigma^2 \mid y) p(\sigma^2 \mid y).$$

$$\text{So } \tilde{y}_i \sim p(\tilde{y} \mid y) = \iint p(\tilde{y} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid y) d\mu d\sigma^2$$

Tip 1: If the joint is easier to sample from

If you simulate $(\tilde{y}^i, \theta_1^i, \theta_2^i)_{i=1}^n \sim p(\tilde{y}, \theta_1, \theta_2 \mid y)$, then ignoring pieces of each sample is analogous to sampling from the marginal:

$$n^{-1} \sum_{i=1}^n h(\tilde{y}^i) \rightarrow E_{\tilde{y}, \theta_1, \theta_2}[h(\tilde{y}^i)] = E_{\tilde{y}}[h(\tilde{y}^i)]$$

Tip 2: if the “top” factor of a joint is tractable

If $p(\tilde{y}, \theta_1, \theta_2 \mid y) = p(\tilde{y} \mid \theta_1, \theta_2, y)p(\theta_1, \theta_2 \mid y)$, then

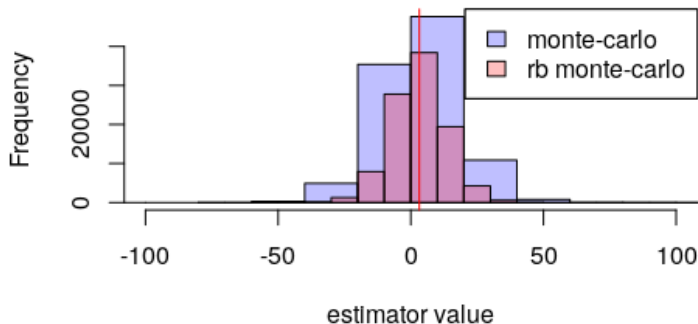
$$\begin{aligned} n^{-1} \sum_{i=1}^n E[h(\tilde{y}, \theta_1^i, \theta_2^i) \mid \theta_1^i, \theta_2^i, y] &\rightarrow E(E[h(\tilde{y}, \theta_1, \theta_2) \mid \theta_1, \theta_2, y]) \\ &= E[h(\tilde{y}, \theta_1, \theta_2) \mid y] \end{aligned}$$

If you can derive expectations of $p(\tilde{y} \mid \theta_1, \theta_2, y)$, and you can sample from the other piece, then this **Rao-Blackwellization** or **marginalization** strategy can be a useful variance reduction technique.

A comparison in R

See 3.r for details:

Monte Carlo: Naive versus RB



Another multiparameter example of conjugacy: Dirichlet-multinomial

Let $y = (y_1, y_2, \dots, y_k)$ be a vector of counts. Let $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ be the probabilities of any trial resulting in each of the k outcomes. We assume that there is a known total count (which means $\sum_i y_i = n$) and that the only possible outcomes are these k outcomes $\sum_i \theta_i = 1$.

The likelihood is a **multinomial** (aka a **categorical**) distribution

$$p(y \mid \theta) \propto \prod_{i=1}^k \theta_i^{y_i},$$

and the prior is a Dirichlet distribution

$$p(\theta) \propto \prod_{i=1}^k \theta_i^{\alpha_i - 1}.$$

The chosen hyper-parameters have a very nice interpretation of counts!

Multivariate Normal Observations

Let each observation y_i follow a multivariate normal distribution. The likelihood $p(y_1, \dots, y_n \mid \mu, \Sigma)$ is usefully written with a few properties of the trace operator:

$$\begin{aligned} &\propto \det(\Sigma)^{-n/2} \exp \left(-\frac{1}{2} \sum_i (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \right) \\ &= \det(\Sigma)^{-n/2} \exp \left[-\frac{1}{2} \sum_i \text{tr} \{ (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \} \right] \\ &= \det(\Sigma)^{-n/2} \exp \left[-\frac{1}{2} \sum_i \text{tr} \{ \Sigma^{-1} (y_i - \mu) (y_i - \mu)' \} \right] \\ &= \det(\Sigma)^{-n/2} \exp \left[-\frac{1}{2} \text{tr} \left\{ \Sigma^{-1} \overbrace{\sum_i (y_i - \mu) (y_i - \mu)'}^{S_0} \right\} \right] \end{aligned}$$

Multivariate Normal Observations with known covariance matrix

A conjugate prior for $p(y \mid \mu) \propto \det(\Sigma)^{-n/2} \exp \left[-\frac{1}{2} \text{tr} \left\{ \overbrace{\Sigma^{-1}}^{\text{known}} S_0 \right\} \right]$ is

$$p(\mu \mid \mu_0, \Lambda_0) = \det(\Lambda_0)^{-1/2} \exp [(\mu - \mu_0)' \Lambda_0^{-1} (\mu - \mu_0)]$$

This makes the posterior distribution (homework question exercise 3.13) normal with mean and precision

$$\mu_n = (\Lambda_0 + n\Sigma^{-1})^{-1} (\Lambda_0^{-1} \mu_0 + n\Sigma^{-1} \bar{y})$$

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.$$

Multivariate Normal Observations with unknown covariance matrix

When all of the elements of Σ are unknown, we need a prior for that as well. This prior must put zero mass on matrices that aren't positive definite or aren't symmetric.

A popular option is the **inverse Wishart** distribution, which is analagous to the inverse-Gamma distribution. It has a degrees of freedom parameter: ν_0 . And it has a scale matrix parameter Λ_0 .

Multivariate Normal Observations with unknown covariance matrix

When all of the elements of Σ are unknown, we need a prior for that as well. This prior must put zero mass on matrices that aren't positive definite or aren't symmetric.

A popular option is the **inverse Wishart** distribution, which is analagous to the inverse-Gamma distribution. It has a degrees of freedom parameter: ν_0 . And it has a scale matrix parameter Λ_0 .

If $\Sigma \in \mathbb{R}^{d \times d}$, we will write

$$\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$$

and we can write (something proportional to) the density as

$$p(\Sigma) \propto \det(\Sigma)^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\text{tr}[\Lambda_0 \Sigma^{-1}]\right)$$

Multivariate Normal Observations with unknown covariance matrix

The following is a conjugate prior

$$\begin{aligned} p(\mu \mid \Sigma)p(\Sigma) &= N(\mu_0, \Sigma/\kappa_0) \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1}) \\ &\propto \left[\det(\Sigma)^{-1/2} \exp \left(-\frac{\kappa_0}{2} (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0) \right) \right] \times \\ &\quad \left[\det(\Sigma)^{-(\nu_0+d+1)/2} \exp \left(-\frac{1}{2} \text{tr} [\Lambda_0 \Sigma^{-1}] \right) \right] \end{aligned}$$

Multivariate Normal Observations with unknown covariance matrix

Here's the posterior:

$$\begin{aligned} p(\mu, \Sigma \mid y) &\propto p(y \mid \mu, \Sigma) p(\mu \mid \Sigma) p(\Sigma) \\ &\propto \det(\Sigma)^{-n/2} \exp \left[-\frac{1}{2} \text{tr} \left\{ \Sigma^{-1} \sum_i (\mu - y_i)(\mu - y_i)' \right\} \right] \times \\ &\quad \det(\Sigma)^{-1/2} \exp \left(-\frac{\kappa_0}{2} (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0) \right) \times \\ &\quad \det(\Sigma)^{-(\nu_0 + d + 1)/2} \exp \left(-\frac{1}{2} \text{tr} [\Lambda_0 \Sigma^{-1}] \right) \\ &= \dots \end{aligned}$$

Multivariate Normal Observations with unknown covariance matrix

It helps to recognize $p(\mu \mid \Sigma, y)$ first, and then $p(\Sigma \mid y)$. Here is negative twice the log of the exponent:

$$\begin{aligned} & \text{tr} \left\{ \Sigma^{-1} \sum_i (\mu - y_i)(\mu - y_i)' + \kappa_0 (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0) \right\} + c_1 \\ &= \sum_i (\mu - y_i)' \Sigma^{-1} (\mu - y_i) + \kappa_0 (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0) + c_1 \\ &= n\mu' \Sigma^{-1} \mu - 2n\mu' \Sigma^{-1} \bar{y} + \kappa_0 \mu' \Sigma^{-1} \mu - 2\kappa_0 \mu' \Sigma^{-1} \mu_0 + c_2 \\ &= \mu' [(\Sigma/n)^{-1} + (\Sigma/\kappa_0)^{-1}] \mu - 2\mu' [(\Sigma/n)^{-1} \bar{y} + (\Sigma/\kappa_0)^{-1} \mu_0] + c_2 \\ &= (\mu - \mu_n)' B (\mu - \mu_n) + c_3 \end{aligned}$$

where $B = (\Sigma/n)^{-1} + (\Sigma/\kappa_0)^{-1}$ and $\mu_n = B^{-1} [(\Sigma/n)^{-1} \bar{y} + (\Sigma/\kappa_0)^{-1} \mu_0]$

Multivariate Normal Observations with unknown covariance matrix

Clearly

$$B = (\Sigma/n)^{-1} + (\Sigma/\kappa_0)^{-1} = \Sigma^{-1}(n + \kappa_0)$$

and

$$\mu_n = B^{-1} [(\Sigma/n)^{-1}\bar{y} + (\Sigma/\kappa_0)^{-1}\mu_0] = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

Multivariate Normal Observations with unknown covariance matrix

Back to neg. twice the log-exponent of the *entire* posterior (can't ignore Σ anymore so keep track of c_1, c_2, c_3)

$$\begin{aligned} & (\mu - \mu_n)' B (\mu - \mu_n) - \mu_n' B \mu_n + \text{tr} \left[\Lambda_0 \Sigma^{-1} + \sum_i y_i' \Sigma^{-1} y_i + \kappa_0 \mu_0 \mu_0' \Sigma^{-1} \right] \\ &= (\mu - \mu_n)' B (\mu - \mu_n) - \mu_n' B \mu_n + \text{tr} \left[\left(\Lambda_0 + \sum_i y_i y_i' + \kappa_0 \mu_0 \mu_0' \right) \Sigma^{-1} \right] \\ &= (\mu - \mu_n)' B (\mu - \mu_n) + \\ & \quad \text{tr} \left[\underbrace{\left(\Lambda_0 + \sum_i y_i y_i' + \kappa_0 \mu_0 \mu_0' - (n + \kappa_0) \mu_n \mu_n' \right)}_{\text{hw is to show that this equals } \Lambda_n} \Sigma^{-1} \right] \end{aligned}$$

A few notes on example 3.7

- 1 It's a logistic regression model with two parameters: slope and intercept
- 2 Groups: $i = 1, 2, 3, 4$
- 3 For each group, sample size n_i is known
- 4 For each group, y_i is a count (tumors, deaths, etc.)
- 5 For each group, x_i is a dose (continuous amount of treatment for each group)

A few notes on example 3.7

For each group:

$$y_i \mid \alpha, \beta \sim \text{Binomial}(n_i, \text{invlogit}(\alpha + \beta x_i))$$

We can write $\theta_i = \text{invlogit}(\alpha + \beta x_i)$ to make it cleaner, but note that this isn't introducing more parameters. The likelihood is

$$p(y \mid \alpha, \beta) = \prod_{i=1}^4 \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i}$$

A few notes on example 3.7

For each group:

$$y_i \mid \alpha, \beta \sim \text{Binomial}(n_i, \text{invlogit}(\alpha + \beta x_i)) \quad (1)$$

- 1 The **dose-response** is the relationship between x_i and θ_i (which is assumed the same for each group i).
- 2 **LD-50** is the unknown quantity $-\alpha/\beta$. It only makes sense when $\beta > 0$, and it is the value of x_i that yields $\theta_i = .5$ (plug it into eqn (1) above). Sometimes scientists are more interested in estimating this than they are in estimating individual parameters.

A few notes on example 3.7

We assume an improper prior: $p(\alpha, \beta) \propto 1$. So

$$\begin{aligned} & \log p(\alpha, \beta \mid y) \\ &= c + \log p(y \mid \alpha, \beta) + \log p(\alpha, \beta) \\ &= c + \log \prod_{i=1}^4 \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i} \\ &= c + \log \prod_{i=1}^4 \left(\frac{\theta_i}{(1 - \theta_i)} \right)^{y_i} (1 - \theta_i)^{n_i} \\ &= c + \sum_{i=1}^4 y_i (\alpha + \beta x_i) - n_i \log[1 + \exp(\alpha + \beta x_i)] \end{aligned}$$

- ① One of the authors has provided R code: https://github.com/avehtari/BDA_R_demos/tree/master/demos_ch3