Multi-parametric Models

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Previously

- Last week, we discussed the basics of Bayesian inference in the simple, single parameter setting
  - Deriving the posterior
  - Summarizing the posterior
  - Impact of the prior distribution
- Extending to the multi-parameter setting is simple, in principle, but does involve some additional steps

Deriving the posterior

- Consider a model with two parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, both unknown.
- Deriving the posterior in the multi-parameter setting is no different to the single parameter setting:

$$p(\theta_1, \theta_2|\mathbf{y}) \propto f(\mathbf{y}|\theta_1, \theta_2)\pi(\theta_1, \theta_2)$$

- Now we must specify a joint prior for the two parameters.
- However, we often have little data about either parameter, let alone the covariance between the two parameters!
- Two easy ways of specifying a joint prior:
  - Assume independence a *priori*:

$$\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$$

  - Assume a hierarchical prior:

$$\pi(\theta_1, \theta_2) = \pi(\theta_1|\theta_2)\pi(\theta_2)$$

Summarizing a multi-parameter posterior

- We have our joint posterior ... now what?
- We are typically interested in one or a subset of parameters, while the other parameters are nuisance parameters that still must be interested.
  - i.e. We are typically interested in the mean but must estimate the variance for inference
- Moving forward, assume:
  - θ_1 is our parameter of interest
  - θ_2 is a nuisance parameter

Marginal posterior for parameter of interest

- Inference on the parameter of interest is based on its marginal posterior:

$$
\begin{aligned}
p(\theta_1|\mathbf{y}) &= \int p(\theta_1, \theta_2|\mathbf{y}) d\theta_2 \\
&= \int p(\theta_1|\theta_2, \mathbf{y}) p(\theta_2|\mathbf{y}) d\theta_2
\end{aligned}
$$

- The marginal posterior $p(\theta_1|\mathbf{y})$ is a mixture of the conditional posterior weighted by the marginal posterior density of θ_2.

  $\Rightarrow$ The marginal posterior averages the conditional posterior over all possible value of θ_2

- We often do not need to explicitly integrate out θ_2 but approximate the marginal posteriors numerically!

Illustration via the normal distribution

- We investigate the posterior of a normal-distributed dataset in the following scenarios:
  - Unknown mean, and known variance
  - Unknown mean, and unknown variance
- **Likelihood:** Let $y_1, y_2, \ldots, y_n$ be iid normal random variables with mean μ and variance σ^2. The resulting likelihood is

$$f(\mathbf{y}|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

When σ^2 is known

- In this case, we are only required to place a prior on μ.
- Consider a "non-informative", flat prior:

$$\pi(\mu) = 1$$

This is an example of an improper pior.

- We know that $\bar{\mathbf{y}}$ is the sufficient statistic for μ. Recall that

$$p(\theta|\mathbf{y}) \propto g(T(x)|\theta) \cdot \pi(\theta)$$

- Therefore, the posterior of μ can be derived by

$$\begin{aligned} p(\mu|\mathbf{y}, \sigma^2) &\propto g(\bar{\mathbf{y}}|\mu, \sigma^2) \cdot \pi(\mu) \\ &= N(\bar{\mathbf{y}}|\mu, \sigma^2/2) \cdot 1 \propto \exp\left\{-\frac{1}{2\sigma^2/n}(\mu - \bar{\mathbf{y}})^2\right\} \end{aligned}$$

- That is, the posterior for μ with a flat prior is $N(\bar{\mathbf{y}}, \sigma^2/n)$.

  $\Rightarrow$ Same as what we would expect in a frequentist analysis!

What if σ^2 is unknown?

- Now we must specify a joint prior for μ and σ^2
- Consider the improper prior:

$$\pi(\mu) = 1; \pi(\sigma^2) \propto 1/\sigma^2, \quad \text{or equivalently} \quad \pi(\mu, \sigma^2) \propto 1/\sigma^2$$

- The joint prior is uniform in $(\mu, \log \sigma)$.

Deriving the joint posterior

- The likelihood can be re-written as

$$\begin{aligned} f(\mathbf{y}|\mu,\sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\bar{y})^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

  where $s^2 = \sum_{i=1}^n (y_i-\bar{y})^2/(n-1)$.

- $\bar{y}$ and s^2 are sufficient statistics.
- The joint posterior is therefore

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

Marginal posterior of μ

- Our goal is to complete inference on μ
- Therefore, we need the marginal posterior of μ

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

- This implies that $\frac{\mu-\bar{y}}{s/\sqrt{n}}$ follows a t-distribution with $n-1$ degree of freedom.

$\Rightarrow$ With the non-informative prior, the infrence results of the Bayesian method are the same as the classical case!

Marginal posterior of μ

- Luckily, we were able to integrate out σ^2 analytically and obtain a closed-form marginal posterior of μ.
- However, we are not always so lucky ...
- An alternative approach is to factorize the joint posterior

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

- This requires two components:
  - $p(\mu|\sigma^2, \mathbf{y})$
  - $p(\sigma^2|\mathbf{y})$

Marginal posterior of μ

- From our previous derivation, we know the conditional posterior of μ

$$\mu|\sigma^2, \mathbf{y} \sim N(\bar{\mathbf{y}}, \sigma^2/n)$$

- To find the marginal posterior of σ^2, we must complete the integral:

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

Therefore, $\sigma^2|\mathbf{y} \sim Inv - Gamma(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$.

- This result, again, parallels the classical inference:
  $\Rightarrow$ $(n-1)s^2/\sigma^2$ follows a chi-square distribution!

Sampling-based approximation of the posterior

- We can obtain the marginal posterior of μ by analytically deriving

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

  but we rarely need to work with this. Instead, we can use a simpler sampling based mechanism to approximate the posterior distribution.

- The sampling based mechanism is conducted as follows:
  For each $i = 1, \ldots, M$,
  1. draw $\sigma^2_{(i)} \sim p(\sigma^2|\mathbf{y}) = IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$
  2. draw $\mu_{(i)} \sim p(\mu|\sigma^2, \mathbf{y}) = N\left(\bar{\mathbf{y}}, \sigma^2_{(i)}/n\right)$.

- The resulting paired samples $\left(\mu_{(i)}, \sigma^2_{(i)}\right)_{i=1}^{M}$ are precisely from the $p(\mu, \sigma^2|\mathbf{y})$.

- The distribution of $\left(\mu_{(i)}\right)_{i=1}^{M}$ approximates the $p(\mu|\mathbf{y})$; and the distribution of $\left(\sigma^2_{(i)}\right)_{i=1}^{M}$ approximates $p(\sigma^2|\mathbf{y})$

Prediction using the sampling algorithm

- To predict a "future" observation y^*, we need the posterior predictive distribution, $p(y^*|\mathbf{y})$.

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

- Luckily, the integral is analytically tractable, and the posterior predictive $p(y^*|\mathbf{y})$ is again a t distribution

$$\frac{y^* - \bar{y}}{s\sqrt{1 + 1/n}} \sim t_{n-1}$$

- Yet another way: sampling-based approximation
  1. draw $\left(\mu_{(i)}, \sigma^2_{(i)}\right)_{i=1}^{M}$ from the joint posterior $p(\mu, \sigma^2|\mathbf{y})$ as discussed above
  2. draw $y^*_{(i)}$ from $N(y^*|\mu_{(i)}, \sigma^2_{(i)})$ for each $i = 1, \ldots, M$

  The resulting samples $\left(y^*_{(i)}\right)_{i=1}^{M}$ represents the posterior predictive distribution.

Implications

- Estimating the nuisance parameter σ^2 changes the posterior distribution of μ
- The marginal posterior of μ will have heavier tails than when the variance is known
- The marginal posterior of μ can be considered as mixture of normals (conditioning on σ^2), and thus can be approximated numerically by sequential sampling
- This is another example of when the Bayesian analysis will result in similar inference to a standard frequentist analysis using non-informative priors.

An Example

- Design: A randomized clinical trial was conducted to evaluate the impact of nicotine reduction on cigarette use and dependence (Donny et al., NEJM, 2015)
  - Participants randomized to one of seven conditions
  - Primary endpoint was total cigarettes smoked per day (CPD) at week 6
  - We will focus on estimation of CPD in lowest nicotine content group
  - CPD approximately follows a normal distribution

- Data: For a total of $n = 109$ participants in the lowest nicotine group, we obtain the sufficient statistics for their CPD at week 6:

$$\bar{\mathbf{y}} = 15.4; \quad s = 7.6$$

When variance is known

- First, let's assume that the variance is known and equal to 7.6^2.
- In this case, the posterior distribution of the mean is

$$\mu|\mathbf{y} \sim N\left(15.4, \frac{7.6}{\sqrt{109}} = 0.73\right)$$

- Summarizing the posterior:
  - Posterior mean, median, and mode: 15.4
  - 95% credible interval: $(13.97, 16.83)$

When variance is unknown

- Now, assume the variance is unknown and $s^2 = 7.6^2$ is an estimate of the variance.
- In this case $\frac{\mu - \bar{y}}{s/\sqrt{n}} = \frac{\mu - 15.4}{7.6/\sqrt{109}}$ follows a t-distribution with $n - 1 = 108$ degrees of freedom, which has the following summaries:
  - Posterior mean, median, and mode: 0
  - Posterior variance: $\frac{n-1}{n-3} = 1.02$
  - 95% credible interval: $(-1.98, 1.98)$
- Summaries for μ can be found via transformation:
  - Posterior mean, median, and mode: 15.4
  - 95% credible interval: $(13.95, 16.85)$

Comparison

- Point estimates are the same: 15.4
- Credible intervals are slightly wider when variance is unknown

$$(13.97, 16.83) \quad vs \quad (13.95, 16.85)$$

- As we known, the difference between a t-distribution and a normal distribution with df $= 108$ is minimal; the difference would be more with a smaller sample size.

Lab Excercise

- Use the sampling-based approach to approximate the posterior distribution, setting $M = 1000$
- Plot the approximate marginal posterior of μ
- Obtain the point estimate and 95% credible interval of μ from the approximate posteriors
- Compare to the values obtained analytically
- What if increasing $M = 10,000$?
- Obtain a 95% prediction interval for a new observation y^*.