STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 3B: Normal Models

UNIVERSITY OF TORONTO

Normal Models

- Important in many statistical modeling problems
- Often useful as approximation or a component in more complicated models
- We will treat separately cases with known variance and known mean.

Natural Conjugate Priors

If $\mathcal{F}$ is a class of sampling distributions $\pi(y \mid \theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ is

$$\pi(\theta \mid y) \in \mathcal{P}, \text{ for all } \pi(\cdot \mid \theta) \in \mathcal{F} \text{ and } \pi(\cdot) \in \mathcal{P}.$$

Is beta distribution conjugate for binomial distribution?

What distributions are conjugate for normal distribution?

The natural conjugate prior families: $\mathcal{P}$ is the set of all densities having the same functional form as the likelihood.

Normal Model: Unknown Mean, Known Variance

- Suppose $x_i|\mu \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known. Let $x = (x_1, \ldots, x_n)$.
- What is the likelihood?

$$\pi(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$
$$\propto \exp\left[-(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

- Expanding the quadratic term in the exponent, we see that $\pi(x_1, \ldots, x_n|\mu, \sigma^2)$ depends on $x_1, \ldots, x_n$ through:

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 - 2\frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i + n\frac{\mu^2}{\sigma^2}$$

-
- The sufficient statistic for the normal distribution is $\{\sum x_i^2, \sum x_i\}$.
- Knowing the values of these quantities is equivalent to knowing $\frac{1}{n} \sum x_i = \overline{x}$ and $\frac{1}{n-1} \sum (x_i - \overline{x})^2 = s^2$.
- As a result, inference in the context of normal models can be broken down into inference about the mean and the variance – i.e., μ and σ^2.

Normal Model: Unknown Mean, Known Variance

- What is the natural conjugate prior?
- Goal: pick a prior that has the same functional form as the likelihood, then derive the posterior and evaluate whether it too will have the same functional form.
- In other words, we know that for any (conditional) prior distribution $\pi(\mu|\sigma^2)$, the posterior distribution will look like:

$$\begin{aligned} \pi(\mu|x_1, ...x_n, \sigma^2) &\propto \pi(\mu|\sigma^2) \times \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right] \\ &\propto \pi(\mu|\sigma^2) \times \exp\left[c_1(\mu-c_2)^2\right] \end{aligned}$$

- Hence, if $\pi(\mu|\sigma^2)$ is to be conjugate, it must include terms like $e^{c_1(\mu-c_2)^2}$.
- And it must be a density on $\mathbb{R}$.
- That would suggest a normal density for $\pi(\mu|\sigma^2)$.
- Conjecture: If $\pi(\mu|\sigma^2)$ is normal, and $x_1, ...x_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\pi(\mu|\sigma^2, x_1, ...x_n)$ is also normal.
- Let's see if we can show this.

Normal Model: Unknown Mean, Known Variance

- Let $\mu \sim \mathrm{N}(\mu_0, \tau_0^2)$. Then,

$$\begin{aligned} \pi(\mu|x_1, ...x_n, \sigma^2) &\propto \pi(\mu|\sigma^2) f(x_1, ...x_n|\mu, \sigma^2) \\ &\propto \exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right] \end{aligned}$$

- Adding the terms in the exponents and ignoring the $-1/2$, we obtain:

$$\frac{1}{\tau_o^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}(\sum x_i^2 - 2\mu\sum x_i + n\mu^2) = a\mu^2 - 2b\mu + c$$

where

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum x_i}{\sigma^2} \quad c = c(\mu_0, \tau_0^2, \sigma^2, x_1, ..., x_n)$$

- Now let's see if $\pi(\mu|x_1, ...x_n, \sigma^2)$ takes the form of a normal density.

Unknown mean, known variance

$$
\begin{aligned}
\pi(\mu|\sigma^2, x_1, ...x_n) &\propto \exp\left[-\frac{1}{2}(a\mu^2 - 2b\mu)\right] \\
&\propto \exp\left[-\frac{1}{2}a(\mu^2 - 2b\mu/a + b^2/a^2) + \frac{1}{2}b^2/a\right] \\
&\propto \exp\left[-\frac{1}{2}a(\mu - b/a)^2\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{\mu - b/a}{1/a^{1/2}}\right)^2\right]
\end{aligned}
$$

- This function has exactly the same shape as a normal density curve, with $1/a$ playing the role of the standard deviation and b/a playing the role of the mean. Since probability distributions are determined by their shape, this means that $\pi(\mu|\sigma^2, x_1, ..., x_n)$ is indeed a normal density.
- We refer to the mean and variance of this density as u_1 and τ_1^2 where

$$\tau_1^2 = 1/a = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

- And where

$$\mu_1 = b/a = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

Precision

- Posterior variance precision: The formula for $\frac{1}{\tau_1^2}$ is

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- That is, the prior inverse variance is combined with the inverse of the data variance.
- Inverse variance is oten referred to as the precision. We will frequently parameterize Bayesian models in terms of their precision.

Combining Information: Posterior Precision

- The (conditional) posterior parameters τ_1^2 and μ_1 combine the prior parameters τ_0^2 and μ_0 with terms from the data.
- We already saw that
$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$
- Now, let
  - $\tilde{\sigma}^2 = 1/\sigma^2$. This is sampling precision – how "close" the x_i's are to μ, as opposed to how dispersed they are.
  - $\tilde{\tau}_0^2 = 1/\tau_0^2$. This is prior precision. How sharp my prior beliefs are.
  - $\tilde{\tau}_1^2 = 1/\tau_1^2$. This is posterior precision. How sharp my posterior beliefs are.
- It is convenient to think about precision as the quantity of information on an additive scale. For the normal model, we can now express posterior precision as:
$$\tilde{\tau}_1^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$$
- And so in that sense: posterior information = prior informaton + data information

Combining Information: Posterior Mean

- We can now write the posterior mean in terms of prior and posterior precision as follows:

$$\mu_1 = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\overline{x}$$

- The posterior mean is a weighted average of the prior mean and the sample mean, as it was in the beta binomial model.
- The weight on the sample mean is n/σ^2, the sampling precision of the sample mean.
- The weight on the prior mean is $1/\tau_0^2$, the prior precision.
- Example: If the prior mean were based on k_0 prior observations from the same (or a similar) population as $X_1, \ldots, X_n$ (i.e., $k_0 = \alpha + \beta$ in the binomial model), then we might want to set $\tau_0^2 = \sigma_0^2/k_0$, the variance of the mean of the prior observations. In this case, the formula for the posterior mean reduces to:

$$\mu_1 = \frac{k_0}{k_0 + n}\mu_0 + \frac{n}{k_0 + n}\overline{x}$$

Normal Approximation to the Beta/Binomial

- Suppose that our prior is beta with $\alpha = 2, \beta = 3$. The prior mean is 2/5 = 0.4.
- We observe 5 heads and 5 tails from a coin toss – i.e., data drawn from a binomial distribution. The sample mean is 0.5.
- Our posterior distribution is beta with $\alpha = 7, \beta = 8$. The posterior mean is $7/15 = 0.47$.
- Now let $\mu_0 = \alpha/(\alpha + \beta)$ and $\sigma_0^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. And let $\mu = np$ and $\sigma^2 = np(1-p)$
- What is τ_1^2? What is μ_1?

Normal Predictions

- Suppose we want to predict $\widetilde{Y}$ from the population after observing $Y_1 = y = 1, ..., Y_n = y_n$.
- Note that $\widetilde{Y} \sim \mathrm{N}(\mu, \sigma^2) \leftrightarrow \widetilde{Y} = \mu + \widetilde{\epsilon}, \quad \widetilde{\epsilon} \sim \mathrm{N}(0, \sigma^2)$
- Saying that $\widetilde{Y}$ is normal with mean μ is the same as saying $\widetilde{Y}$ is equal to μ plus some mean-zero normally distributed noise.
- Using this, we can compute posterior mean and variance of $\widetilde{Y}$:

$$\begin{aligned}
\mathrm{E}[\widetilde{Y}|y, \sigma^2] &= \mathrm{E}[\mu + \widetilde{\epsilon}|y, \sigma^2] \\
&= \mathrm{E}[\mu|y, \sigma^2] + \mathrm{E}[\widetilde{\epsilon}|y, \sigma^2] \\
&= \mu_1 + 0 = \mu_1 \\
\mathrm{Var}(\widetilde{Y}|y, \sigma^2) &= \mathrm{Var}(\mu + \widetilde{\epsilon}|y, \sigma^2) \\
&= \mathrm{Var}(\mu|y, \sigma^2) + \mathrm{Var}(\widetilde{\epsilon}|y, \sigma^2) \\
&= \tau_1^2 + \sigma^2
\end{aligned}$$

- Recall that the sum of iid normal RVs is also normal.
- Hence, since both μ and $\widetilde{\epsilon}$, conditional on y and σ^2, are normal, so is $\widetilde{Y} = \mu + \widetilde{\epsilon}$.
- The predictive distribution is therefore

$$\widetilde{Y}|\sigma^2, y \sim \mathrm{N}(\mu_1, \tau_1^2 + \sigma^2)$$

Normal Predictions

- Our uncertainty about a new sample $\widetilde{Y}$ is therefore a function of our uncertainty about the center of the population as well as how variable the population is. As $n \to \infty$ we become more confident about where μ is. But certainty about μ does not reduce the sampling variability, and so our uncertainty about $\widetilde{Y}$ never goes below σ^2.

Normal Predictions

- Grogan and Wirth (1981) provide data on the wing length in millimeters of nine members of a species of midge. From these nine measurements we wish to make inferences on the population mean μ. Studies from other populations suggest that wing lengths are typically around 1.9 mm, and so we set $\mu_0 = 1.9$.
- Suppose we set $\tau_0 = 0.95$.
- The mean of the observations is $\overline{y} = 1.804$ and $s^2 = 0.017$.
- Find μ_1 and τ_1^2 and $\pi(\mu|y, s^2)$
- Find a 95% credibility interval for μ based on $\pi(\mu|y, s^2)$.
- This example assumes that $s^2 = \sigma^2$. We get a more accurate representation of uncertainty we will need to develop a model where σ^2 is also unknown.