STA 702: Lecture 3

The Normal Model & Prior/Posterior Predictive Distributions

Merlise Clyde

9/8/2022

Outline

- Normal Model

Outline

- Normal Model
- Predictive Distributions

Outline

- Normal Model
- Predictive Distributions
  - Prior Predictive; useful for prior elicitation

Outline

- Normal Model
- Predictive Distributions
  - Prior Predictive; useful for prior elicitation
- Posterior Predictive

Outline

- Normal Model
- Predictive Distributions
  - Prior Predictive; useful for prior elicitation
- Posterior Predictive
  - Predicting/forecasting future events

Outline

- Normal Model
- Predictive Distributions
  - Prior Predictive; useful for prior elicitation
- Posterior Predictive
  - Predicting/forecasting future events
- Comparing Estimators

Normal Model Setup

- Suppose we have independent observations

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

Normal Model Setup

- Suppose we have independent observations

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

where each $y_i \sim \mathsf{N}(\theta, \sigma^2)$ (iid)

Normal Model Setup

- Suppose we have independent observations

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

where each $y_i \sim \mathsf{N}(\theta, \sigma^2)$ (iid)

- We will see that it is more convenient to work with $\tau = 1/\sigma^2$ (precision)

Normal Model Setup

- Suppose we have independent observations

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

where each $y_i \sim \mathsf{N}(\theta, \sigma^2)$ (iid)

- We will see that it is more convenient to work with $\tau = 1/\sigma^2$ (precision)

- reparameterizing the model for the data we have

$$y_i \sim \mathcal{N}(\theta, \tau^{-1})$$

Normal Model Setup

- Suppose we have independent observations

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

where each $y_i \sim \mathsf{N}(\theta, \sigma^2)$ (iid)

- We will see that it is more convenient to work with $\tau = 1/\sigma^2$ (precision)

- reparameterizing the model for the data we have

$$y_i \sim \mathcal{N}(\theta, \tau^{-1})$$

- for simplicity we will treat τ as known initially.

Marginal Distribution

- Recall that the **marginal distribution** is

$$p(y) = p(y_1, \ldots, y_n) = \int_\Theta p(y_1, \ldots, y_n \mid \theta)\pi(\theta)\, d\theta$$

Marginal Distribution

- Recall that the **marginal distribution** is

$$p(y) = p(y_1, \ldots, y_n) = \int_\Theta p(y_1, \ldots, y_n \mid \theta)\pi(\theta)\, d\theta$$

- this is also called the **prior predictive** distribution and is independent of any unknown parameters

- We may care about making predictions before we even see any data.

Marginal Distribution

- Recall that the **marginal distribution** is

$$p(y) = p(y_1, \ldots, y_n) = \int_\Theta p(y_1, \ldots, y_n \mid \theta)\pi(\theta)\, d\theta$$

- this is also called the **prior predictive** distribution and is independent of any unknown parameters

- We may care about making predictions before we even see any data.

- This is often useful as a way to see if the sampling distribution or prior we have chosen is appropriate, after integrating out all unknown parameters.

Marginal Distribution

- Recall that the **marginal distribution** is

$$p(y) = p(y_1, \ldots, y_n) = \int_\Theta p(y_1, \ldots, y_n \mid \theta)\pi(\theta)\, d\theta$$

- this is also called the **prior predictive** distribution and is independent of any unknown parameters

- We may care about making predictions before we even see any data.

- This is often useful as a way to see if the sampling distribution or prior we have chosen is appropriate, after integrating out all unknown parameters.

- Need to specify a prior for θ on $\mathbb{R}$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$
- Prior variance $\sigma_0^2 = 1/\tau_0$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$
- Prior variance $\sigma_0^2 = 1/\tau_0$
- τ_0 is the prior precision and expresses our certainty about this guess

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$
- Prior variance $\sigma_0^2 = 1/\tau_0$
- τ_0 is the prior precision and expresses our certainty about this guess
- one notion of non-informative is to let $\tau_0 \to 0$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$
- Prior variance $\sigma_0^2 = 1/\tau_0$
- τ_0 is the prior precision and expresses our certainty about this guess
- one notion of non-informative is to let $\tau_0 \to 0$
- better justification is as Jeffreys' prior (uniform measure) **Derive**

$$\pi(\theta) \propto 1$$

Prior for a Normal Mean

- Natural choice is a Normal/Gaussian distribution (Conjugate prior)

$$\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$$

- θ_0 is the prior mean - best guess for θ using information other than $\mathbf{y}$
- Prior variance $\sigma_0^2 = 1/\tau_0$
- τ_0 is the prior precision and expresses our certainty about this guess
- one notion of non-informative is to let $\tau_0 \to 0$
- better justification is as Jeffreys' prior (uniform measure) **Derive**

$$\pi(\theta) \propto 1$$

- parameterization invariant and invariant to shift changes in the data (group invariance)

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Quadratic $\tau_0(\theta - \theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Quadratic $\tau_0(\theta-\theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

1) **Expand** quadratics

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Quadratic $\tau_0(\theta-\theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

1) **Expand** quadratics

2) **Group** terms with θ^2 and θ

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Quadratic $\tau_0(\theta-\theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

1) **Expand** quadratics

2) **Group** terms with θ^2 and θ

3) Read off **posterior precision** and **posterior mean**

Prior Predictive for a Single Case

$$\begin{aligned} p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\ &\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta \end{aligned}$$

Quadratic $\tau_0(\theta-\theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

1) **Expand** quadratics

2) **Group** terms with θ^2 and θ

3) Read off **posterior precision** and **posterior mean**

4) **Complete the square**

Prior Predictive for a Single Case

$$
\begin{aligned}
p(y) &\propto \int_{\mathbb{R}} p(y \mid \theta)\pi(\theta)\, d\theta \\
&\propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\tau(y-\theta)^2\right\} \exp\left\{-\frac{1}{2}\tau_0(\theta-\theta_0)^2\right\} d\theta
\end{aligned}
$$

Quadratic $\tau_0(\theta-\theta_0)^2 = \tau_0\theta^2 - 2\tau_0\theta_0\theta + \tau_0\theta_0^2$

1) **Expand** quadratics

2) **Group** terms with θ^2 and θ

3) Read off **posterior precision** and **posterior mean**

4) **Complete the square**

5) **Integrate** out θ to obtain marginal!

Try it!

$$p(y) \propto \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\left[\tau(y-\theta)^2 + \tau_0(\theta-\theta_0)^2\right\} d\theta$$

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

- posterior mean $\hat{\theta} = \frac{\tau_0}{\tau_0+\tau}\theta_0 + \frac{\tau}{\tau_0+\tau}y$

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

- posterior mean $\hat{\theta} = \frac{\tau_0}{\tau_0+\tau}\theta_0 + \frac{\tau}{\tau_0+\tau}y$
- precision weighted average of prior mean and MLE (based on 1 observation)

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

- posterior mean $\hat{\theta} = \frac{\tau_0}{\tau_0+\tau}\theta_0 + \frac{\tau}{\tau_0+\tau}y$
- precision weighted average of prior mean and MLE (based on 1 observation)
- posterior precision is the sum of prior precision and data precision

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

- posterior mean $\hat{\theta} = \frac{\tau_0}{\tau_0+\tau}\theta_0 + \frac{\tau}{\tau_0+\tau}y$
- precision weighted average of prior mean and MLE (based on 1 observation)
- posterior precision is the sum of prior precision and data precision
- marginal distribution for Y (prior predictive)

$$Y \sim \mathsf{N}\left(\theta_0, \frac{1}{\tau_0} + \frac{1}{\tau}\right) \text{ or } \mathsf{N}(\theta_0, \sigma^2 + \sigma_0^2)$$

Results

Posterior for θ based on a single observation (Conjugate family)

$$\theta \mid y \sim \mathsf{N}\left(\hat{\theta}, \frac{1}{\tau_0 + \tau}\right)$$

- posterior mean $\hat{\theta} = \frac{\tau_0}{\tau_0+\tau}\theta_0 + \frac{\tau}{\tau_0+\tau} y$
- precision weighted average of prior mean and MLE (based on 1 observation)
- posterior precision is the sum of prior precision and data precision
- marginal distribution for Y (prior predictive)

$$Y \sim \mathsf{N}\left(\theta_0, \frac{1}{\tau_0} + \frac{1}{\tau}\right) \text{ or } \mathsf{N}(\theta_0, \sigma^2 + \sigma_0^2)$$

- two sources of variability: variability from the model for the data and prior variability

Prior Predictive

- useful to think about observable quantities when choosing the prior

Prior Predictive

- useful to think about observable quantities when choosing the prior
- sample directly from the prior predictive and assess whether the samples are consistent with our prior knowledge

Prior Predictive

- useful to think about observable quantities when choosing the prior
- sample directly from the prior predictive and assess whether the samples are consistent with our prior knowledge
- if not, go back and modify the prior & repeat

Prior Predictive

- useful to think about observable quantities when choosing the prior
- sample directly from the prior predictive and assess whether the samples are consistent with our prior knowledge
- if not, go back and modify the prior & repeat
- sequential substitution sampling (repeat T times)

  1) draw $\theta^{(t)} \sim \pi(\theta)$

  2) draw $y^{(t)} \sim p(y \mid \theta^{(t)})$

Prior Predictive

- useful to think about observable quantities when choosing the prior

- sample directly from the prior predictive and assess whether the samples are consistent with our prior knowledge

- if not, go back and modify the prior & repeat

- sequential substitution sampling (repeat T times)

  1) draw $\theta^{(t)} \sim \pi(\theta)$

  2) draw $y^{(t)} \sim p(y \mid \theta^{(t)})$

- takes into account uncertain about θ and variability in Y!

Posterior Updating

- Sequential updating using the previous result as our prior!

Posterior Updating

- Sequential updating using the previous result as our prior!
- New prior after seeing 1 observation is

$$\mathsf{N}(\theta_1, 1/\tau_1)$$

Posterior Updating

- Sequential updating using the previous result as our prior!
- New prior after seeing 1 observation is

$$\mathsf{N}(\theta_1, 1/\tau_1)$$

- prior mean weighted average

$$\theta_1 \equiv \frac{\tau_0\theta_0 + \tau y_1}{\tau_0 + \tau_1}$$

Posterior Updating

- Sequential updating using the previous result as our prior!
- New prior after seeing 1 observation is

$$\mathsf{N}(\theta_1, 1/\tau_1)$$

- prior mean weighted average

$$\theta_1 \equiv \frac{\tau_0 \theta_0 + \tau y_1}{\tau_0 + \tau_1}$$

- prior precision after 1 observation

$$\tau_1 \equiv \tau_0 + \tau$$

Posterior Updating

- Sequential updating using the previous result as our prior!
- New prior after seeing 1 observation is

$$\mathsf{N}(\theta_1, 1/\tau_1)$$

- prior mean weighted average

$$\theta_1 \equiv \frac{\tau_0 \theta_0 + \tau y_1}{\tau_0 + \tau_1}$$

- prior precision after 1 observation

$$\tau_1 \equiv \tau_0 + \tau$$

- prior variance is now $\sigma_1^2 = 1/\tau_1$

Posterior Predictive for y_2 given y_1

- Conditional $p(y_2 \mid y_1) = p(y_2, y_1)/p(y_1)$ (Hard way!)

Posterior Predictive for y_2 given y_1

- Conditional $p(y_2 \mid y_1) = p(y_2, y_1)/p(y_1)$ (Hard way!)

- Use latent variable representation

$$p(y_2 \mid y_1) = \int_\Theta \frac{p(y_2, \mid \theta) p(y_1 \mid \theta) \pi(\theta) \, d\theta}{p(y_1)}$$

Posterior Predictive for y_2 given y_1

- Conditional $p(y_2 \mid y_1) = p(y_2, y_1)/p(y_1)$ (Hard way!)

- Use latent variable representation

$$p(y_2 \mid y_1) = \int_\Theta \frac{p(y_2, \mid \theta) p(y_1 \mid \theta) \pi(\theta) \, d\theta}{p(y_1)}$$

- simplify to previous problem and use results

$$p(y_2 \mid y_1) = \int_\Theta p(y_2 \mid \theta) \pi(\theta \mid y_1) \, d\theta$$

Posterior Predictive for y_2 given y_1

- Conditional $p(y_2 \mid y_1) = p(y_2, y_1)/p(y_1)$ (Hard way!)

- Use latent variable representation

$$p(y_2 \mid y_1) = \int_\Theta \frac{p(y_2, \mid \theta) p(y_1 \mid \theta) \pi(\theta) \, d\theta}{p(y_1)}$$

- simplify to previous problem and use results

$$p(y_2 \mid y_1) = \int_\Theta p(y_2 \mid \theta) \pi(\theta \mid y_1) \, d\theta$$

- (Posterior) Predictive

$$y_2 \mid y_1 \sim \mathsf{N}(\theta_1, \sigma^2 + \sigma_1^2)$$

Iterated Expectations

Based on expressions we have an exponential of a quadratic in y_2 so know that distribution is Gaussian

Iterated Expectations

Based on expressions we have an exponential of a quadratic in y_2 so know that distribution is Gaussian

- Find the mean and variance using iterated expectations:

Iterated Expectations

Based on expressions we have an exponential of a quadratic in y_2 so know that distribution is Gaussian

- Find the mean and variance using iterated expectations:

- mean

$$\mathsf{E}[Y_2 \mid y_1] = \mathsf{E}_{\theta \mid y_1}[\mathsf{E}_{Y_2 \mid y_1, \theta}[Y_2 \mid y_1, \theta] \mid y_1]$$

Variance via Iterated Expectations

$\text{Var}[Y_2 \mid y_1] =$

$$\text{E}_{\theta \mid y_1}[\text{Var}_{Y_2 \mid y_1, \theta}[Y_2 \mid y_1, \theta] \mid y_1] + \text{Var}_{\theta \mid y_1}[\text{E}_{Y_2 \mid y_1, \theta}[Y_2 \mid y_1, \theta] \mid y_1]$$

Updated Posterior for θ

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(y_1 \mid \theta) \pi(\theta)$$

Updated Posterior for θ

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(y_1 \mid \theta) \pi(\theta)$$

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(\theta \mid y_1)$$

Updated Posterior for θ

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(y_1 \mid \theta) \pi(\theta)$$

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(\theta \mid y_1)$$

Apply previous updating rules

Updated Posterior for θ

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(y_1 \mid \theta) \pi(\theta)$$

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(\theta \mid y_1)$$

Apply previous updating rules

- new posterior mean

$$\theta_2 = \frac{\tau_1 \theta_1 + \tau y_2}{\tau_1 + \tau} = \frac{\tau_0 \theta_0 + 2\tau \bar{y}}{\tau_0 + 2\tau}$$

Updated Posterior for θ

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(y_1 \mid \theta) \pi(\theta)$$

$$p(\theta \mid y_1, y_2) \propto p(y_2 \mid \theta) p(\theta \mid y_1)$$

Apply previous updating rules

- new posterior mean

$$\theta_2 = \frac{\tau_1 \theta_1 + \tau y_2}{\tau_1 + \tau} = \frac{\tau_0 \theta_0 + 2\tau \bar{y}}{\tau_0 + 2\tau}$$

- new precision

$$\tau_2 = \tau_1 + \tau = \tau_0 + 2\tau$$

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

- Shrinkage of the MLE to the prior mean

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

- Shrinkage of the MLE to the prior mean
- More accurate estimation of θ as $n \to \infty$ (reducible error)

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

- Shrinkage of the MLE to the prior mean
- More accurate estimation of θ as $n \rightarrow \infty$ (reducible error)
- Cannot reduce the error for prediction Y_{n+1} due to σ^2

After n observations

Posterior for θ

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

- Shrinkage of the MLE to the prior mean
- More accurate estimation of θ as $n \to \infty$ (reducible error)
- Cannot reduce the error for prediction Y_{n+1} due to σ^2
- predictive distribution for a next observation given *everything* we know - prior and likelihood

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)
- Prior predictive $\mathsf{N}(\theta_0, \sigma_0^2 + \sigma^2)$ (not proper in the limit)

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)
- Prior predictive $\mathsf{N}(\theta_0, \sigma_0^2 + \sigma^2)$ (not proper in the limit)
- Posterior for θ (formal posterior)

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)
- Prior predictive $\mathsf{N}(\theta_0, \sigma_0^2 + \sigma^2)$ (not proper in the limit)
- Posterior for θ (formal posterior)

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

$$\to \qquad \theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\bar{y}, \frac{1}{n\tau}\right)$$

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)
- Prior predictive $\textsf{N}(\theta_0, \sigma_0^2 + \sigma^2)$ (not proper in the limit)
- Posterior for θ (formal posterior)

$$\theta \mid y_1, \ldots, y_n \sim \textsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

$$\to \qquad \theta \mid y_1, \ldots, y_n \sim \textsf{N}\left(\bar{y}, \frac{1}{n\tau}\right)$$

- Recovers the MLE as the posterior mode!

Results with Jeffreys' Prior

- What if $\tau_0 \to 0$? (or $\sigma_0^2 \to \infty$)
- Prior predictive $\mathsf{N}(\theta_0, \sigma_0^2 + \sigma^2)$ (not proper in the limit)
- Posterior for θ (formal posterior)

$$\theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0 \theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

$$\to \qquad \theta \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\bar{y}, \frac{1}{n\tau}\right)$$

- Recovers the MLE as the posterior mode!
- Posterior variance of $\theta = \sigma^2/n$ (same as variance of the MLE)

Posterior Predictive Distribution

Posterior predictive distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

Posterior Predictive Distribution

Posterior predictive distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \textsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

Under Jeffreys' prior

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \textsf{N}\left(\bar{y}, \sigma^2(1 + \frac{1}{n})\right)$$

Posterior Predictive Distribution

Posterior predictive distribution for Y_{n+1}

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\frac{\tau_0\theta_0 + n\tau\bar{y}}{\tau_0 + n\tau}, \frac{1}{\tau} + \frac{1}{\tau_0 + n\tau}\right)$$

Under Jeffreys' prior

$$Y_{n+1} \mid y_1, \ldots, y_n \sim \mathsf{N}\left(\bar{y}, \sigma^2(1 + \frac{1}{n})\right)$$

Captures extra uncertainty due to not knowing θ (compared to plug-in approach where we plug in MLE in sampling model!

Comparing Estimators

Expected loss (from frequentist perspective) of using Bayes Estimator

Comparing Estimators

Expected loss (from frequentist perspective) of using Bayes Estimator

- Posterior mean is optimal under squared error loss (min Bayes Risk) [also absolute error loss]

Comparing Estimators

Expected loss (from frequentist perspective) of using Bayes Estimator

- Posterior mean is optimal under squared error loss (min Bayes Risk) [also absolute error loss]

Compute Mean Square Error (or Expected Average Loss)

$$\mathsf{E}_{\bar{y}\mid\theta}\left[\left(\hat{\theta}-\theta\right)^2\mid\theta\right]$$

$$= \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

Comparing Estimators

Expected loss (from frequentist perspective) of using Bayes Estimator

- Posterior mean is optimal under squared error loss (min Bayes Risk) [also absolute error loss]

Compute Mean Square Error (or Expected Average Loss)

$$\mathsf{E}_{\bar{y}\mid\theta}\left[\left(\hat{\theta}-\theta\right)^2 \mid \theta\right]$$

$$= \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

- For the MLE $\bar{Y}$ this is just the variance of $\bar{Y}$ or σ^2/n

MSE for Bayes

$$\mathsf{E}_{\bar{y}|\theta}\left[\left(\hat{\theta}-\theta\right)^2 \mid \theta\right] = \mathsf{MSE} = \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

- Bias of Bayes Estimate

$$\mathsf{E}_{\bar{Y}|\theta}\left[\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n}\right] = \frac{\tau_0(\theta_0 - \theta)}{\tau_0 + \tau n}$$

MSE for Bayes

$$\mathsf{E}_{\bar{y}|\theta}\left[\left(\hat{\theta}-\theta\right)^2 \mid \theta\right] = \mathsf{MSE} = \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

- Bias of Bayes Estimate

$$\mathsf{E}_{\bar{Y}|\theta}\left[\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n}\right] = \frac{\tau_0(\theta_0 - \theta)}{\tau_0 + \tau n}$$

- Variance

$$\mathsf{Var}\left(\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n} - \theta \mid \theta\right) = \frac{\tau n}{(\tau_0 + \tau n)^2}$$

MSE for Bayes

$$\mathsf{E}_{\bar{y}|\theta}\left[\left(\hat{\theta}-\theta\right)^2 \mid \theta\right] = \mathsf{MSE} = \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

- Bias of Bayes Estimate

$$\mathsf{E}_{\bar{Y}|\theta}\left[\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n}\right] = \frac{\tau_0(\theta_0 - \theta)}{\tau_0 + \tau n}$$

- Variance

$$\mathsf{Var}\left(\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n} - \theta \mid \theta\right) = \frac{\tau n}{(\tau_0 + \tau n)^2}$$

(Frequentist) expected Loss when truth is θ

$$\mathsf{MSE} = \frac{\tau_0^2(\theta - \theta_0)^2 + \tau n}{(\tau_0 + \tau n)^2}$$

MSE for Bayes

$$\mathsf{E}_{\bar{y}\mid\theta}\left[\left(\hat{\theta}-\theta\right)^2 \mid \theta\right] = \mathsf{MSE} = \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$$

- Bias of Bayes Estimate

$$\mathsf{E}_{\bar{Y}\mid\theta}\left[\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n}\right] = \frac{\tau_0(\theta_0 - \theta)}{\tau_0 + \tau n}$$

- Variance

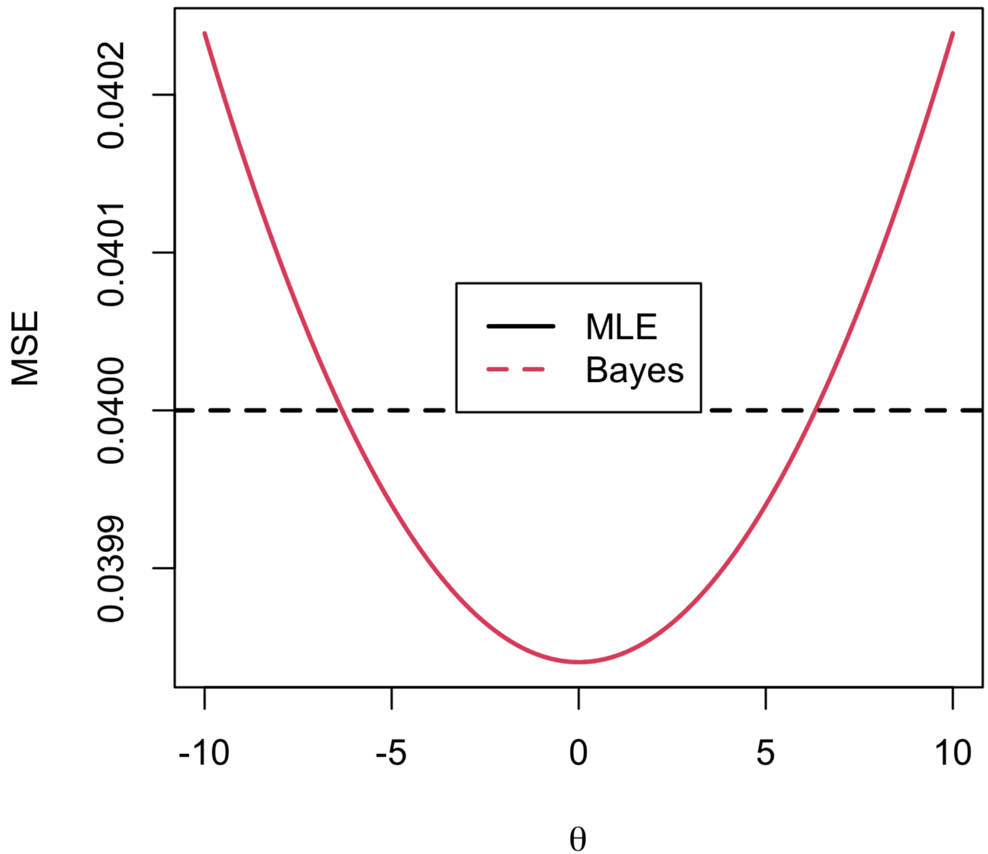
$$\mathsf{Var}\left(\frac{\tau_0\theta_0 + \tau n\bar{Y}}{\tau_0 + \tau n} - \theta \mid \theta\right) = \frac{\tau n}{(\tau_0 + \tau n)^2}$$

(Frequentist) expected Loss when truth is θ

$$\mathsf{MSE} = \frac{\tau_0^2(\theta - \theta_0)^2 + \tau n}{(\tau_0 + \tau n)^2}$$

Behavior ?

Plot



Updating with n Observations

- We can use the $\mathcal{L}(\theta)$ based on n observations and repeat completing the square with the original prior $\theta \sim \mathsf{N}(\theta_0, 1/\tau_0)$

Likelihood Function

- The likelihood for θ is proportional to the sampling model

$$p(y \mid \theta, \tau) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \tau^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\tau(y_i - \theta)^2\right\}$$

Likelihood Function

- The likelihood for θ is proportional to the sampling model

$$p(y \mid \theta, \tau) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \tau^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \tau (y_i - \theta)^2 \right\}$$

Rewrite in terms of sufficient statistics!

Simplification

$$
\begin{aligned}
\mathcal{L}(\theta) &\propto \tau^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\tau \sum_{i=1}^{n} (y_i - \theta)^2 \right\} \\
&\propto \tau^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\tau \sum_{i=1}^{n} \left[(y_i - \bar{y}) - (\theta - \bar{y})\right]^2 \right\} \\
\\
&\propto \tau^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\tau \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\theta - \bar{y})^2 \right] \right\} \\
&\propto \tau^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\tau \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\theta - \bar{y})^2 \right] \right\} \\
&\propto \tau^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\tau s^2 (n-1) \right\} \ \exp\left\{ -\frac{1}{2}\tau n (\theta - \bar{y})^2 \right\}. \\
&\propto \exp\left\{ -\frac{1}{2}\tau n (\theta - \bar{y})^2 \right\}
\end{aligned}
$$

Exercises for Practice

Try this

1) Use $\mathcal{L}(\theta)$ based on n observations and $\pi(\theta)$ to find $\pi(\theta \mid y_1, \ldots, y_n)$ based on the sufficient statistics

Exercises for Practice

Try this

1) Use $\mathcal{L}(\theta)$ based on n observations and $\pi(\theta)$ to find $\pi(\theta \mid y_1, \ldots, y_n)$ based on the sufficient statistics

2) Use $\pi(\theta \mid y_1, \ldots, y_n)$ to find the posterior predictive distribution for Y_{n+1}