

Section 9: Bayesian Inference (cont.) & Decision Theory

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30–9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM–12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30–11:30 AM in Cabot D-Hall (Emily).

2 Bayesian Model Choice

Idea. How do we determine which model is most supported by the data? We can use the Bayes factor.

Definition 1 (Bayes factor). Suppose Bill uses likelihood $f(y | \theta)$ and prior $f(\theta | \text{Bill})$, and Jose uses $g(y | \lambda)$ and $g(\lambda | \text{Jose})$. The **Bayes factor** is

$$\text{BF} = \frac{f(y | \text{Bill})}{g(y | \text{Jose})} = \frac{\int f(y | \theta, \text{Bill}) f(\theta | \text{Bill}) d\theta}{\int g(y | \lambda, \text{Jose}) g(\lambda | \text{Jose}) d\lambda}.$$

It converts prior odds to posterior odds:

$$\frac{P(\text{Bill} | y)}{P(\text{Jose} | y)} = \frac{P(\text{Bill})}{P(\text{Jose})} \times \text{BF}.$$

- **Candidate's formula.** For any fixed value t ,

$$f(y) = \frac{f(y | \theta = t) \pi(\theta = t)}{\pi(\theta = t | y)}.$$

Choose t to make the right-hand side easy to compute (e.g., $t = \hat{\theta}_{\text{MAP}}$ or $t = \mu_0$).

- **Bayesian vs. frequentist testing.** For a composite null $H_0 : \theta \leq 0$, the posterior probability $P(\theta \leq 0 | y)$ is close to the frequentist one-sided p-value when the prior is diffuse. For a *simple* null $H_0 : \theta = \theta_0$, the posterior probability is 0 under any continuous prior—Bayesians must put positive point mass at θ_0 (Cromwell's rule).

• **Lindley’s paradox.** For large n , the Bayes factor penalizes complex models more heavily than a frequentist test does—the log Bayes factor is the BIC. A t -statistic leading to rejection at $\alpha = 0.05$ may still favor the simpler model under the Bayes factor.

3 Posterior Predictive Distribution

Definition 2 (Posterior predictive). Given observed data y , the **posterior predictive distribution** of a future observation \tilde{Y} is

$$f(\tilde{y} | y) = \int_{\theta \in \Theta} f(\tilde{y} | y, \theta) \pi(\theta | y) d\theta.$$

This averages the sampling distribution over remaining uncertainty about θ (i.e., LOTP with Extra Conditioning).

- This can be extended to include covariates \mathbf{X} : $f(\tilde{y} | \mathbf{x}, \vec{y}) = \int_{\theta \in \Theta} f(\tilde{y} | \mathbf{x}, \vec{y}, \theta) \pi(\theta | \mathbf{x}, \vec{y}) d\theta$.
- By Adam’s law: $\mathbb{E}[\tilde{Y} | y] = \mathbb{E}_{\theta|y}[\mathbb{E}(\tilde{Y} | \theta)] = \mathbb{E}[\theta | y]$ (in the Normal case).
- By Eve’s law: $\text{Var}(\tilde{Y} | y) = \underbrace{\mathbb{E}[\text{Var}(\tilde{Y} | \theta)]}_{\text{aleatoric}} + \underbrace{\text{Var}(\mathbb{E}[\tilde{Y} | \theta])}_{\text{epistemic}} = \sigma^2 + \tau_n^2$ (in Normal-Normal).
- **Simulation strategy.** Draw $\theta^{[b]} \sim \pi(\theta | y)$, then $\tilde{y}^{[b]} \sim f(\tilde{y} | \theta^{[b]})$. The empirical distribution of $\{\tilde{y}^{[b]}\}$ approximates the posterior predictive.
- **Gamma-Poisson predictive.** If $\lambda | y \sim \text{Gamma}(r_0 + S, b_0 + n)$, then

$$\tilde{Y} | y \sim \text{NBin}\left(r_0 + S, \frac{b_0 + n}{b_0 + n + 1}\right).$$

The predictive is more dispersed than $\text{Pois}(\hat{\lambda})$ because it propagates uncertainty in λ .

Concept Checker 1. Let $Y_1, \dots, Y_n | \mu \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$, with σ^2 , μ_0 , and τ_0^2 known. We want $Y_{n+1} | Y_1, \dots, Y_n$. First, is $Y_{n+1} \perp\!\!\!\perp (Y_1, \dots, Y_n)$? Second, what is the distribution of $(Y_1, \dots, Y_{n+1}, \mu)$? Third, what does that imply about the distribution of $Y_{n+1} | Y_1, \dots, Y_n$?

Solution

First, $Y_{n+1} \not\perp\!\!\!\perp (Y_1, \dots, Y_n)$ because Y_1, \dots, Y_n give information on the unknown mean μ , which in turn gives information about Y_{n+1} . However, $(Y_{n+1} \perp\!\!\!\perp (Y_1, \dots, Y_n)) | \mu$. Second, $(Y_1, \dots, Y_{n+1}, \mu) \sim \text{MVN}(\vec{\mu}^*, \Sigma^*)$ (i.e., it is Multivariate Normal with some mean vector $\vec{\mu}^*$ and some variance-covariance matrix Σ^*) since it is a vector of Normals. Third, this implies $Y_{n+1} | Y_1, \dots, Y_n \sim \mathcal{N}(\mu_{n+1}, \sigma_{n+1}^2)$ (i.e., it is Normal with some mean μ_{n+1} and variance σ_{n+1}^2) as a property of the Multivariate Normal.

Concept Checker 2. With the same setup above, find the distribution of $Y_{n+1} | Y_1, \dots, Y_n$.

Solution

Since $Y_{n+1} \mid Y_1, \dots, Y_n \sim \mathcal{N}(\mu_{n+1}, \sigma_{n+1}^2)$, let's find μ_{n+1} and σ_{n+1}^2 . To clean the notation, let $\vec{Y}_n = (Y_1, \dots, Y_n)$. Note this doesn't include Y_{n+1} .

$$\begin{aligned}
 \mu_{n+1} &= \mathbb{E}[Y_{n+1} \mid \vec{Y}_n] && \text{by definition} \\
 &= \mathbb{E} \left[\mathbb{E}[Y_{n+1} \mid \vec{Y}_n, \mu] \mid \vec{Y}_n \right] && \text{by Adam's w/ Extra Cond.} \\
 &= \mathbb{E} \left[\mathbb{E}[Y_{n+1} \mid \mu] \mid \vec{Y}_n \right] && \text{by cond. independence} \\
 &= \mathbb{E} \left[\mu \mid \vec{Y}_n \right] && \text{since } Y_{n+1} \mid \mu \sim \mathcal{N}(\mu, \sigma^2) \\
 &= (1 - b_n)\bar{Y}_n + b_n\mu_0 && \text{by Normal-Normal Conj.}
 \end{aligned}$$

Note $b_n = \frac{\sigma^2}{\sigma^2 + n\tau_0^2}$. To clean up notation

$$\begin{aligned}
 \sigma_{n+1}^2 &= \text{Var}[Y_{n+1} \mid \vec{Y}_n] && \text{by definition} \\
 &= \mathbb{E} \left[\text{Var}[Y_{n+1} \mid \vec{Y}_n, \mu] \mid \vec{Y}_n \right] + \text{Var} \left[\mathbb{E}[Y_{n+1} \mid \vec{Y}_n, \mu] \mid \vec{Y}_n \right] && \text{by Eve's w/ Extra Cond.} \\
 &= \mathbb{E} \left[\text{Var}[Y_{n+1} \mid \mu] \mid \vec{Y}_n \right] + \text{Var} \left[\mathbb{E}[Y_{n+1} \mid \mu] \mid \vec{Y}_n \right] && \text{by cond. independence} \\
 &= \mathbb{E} \left[\sigma^2 \mid \vec{Y}_n \right] + \text{Var} \left[\mu \mid \vec{Y}_n \right] && \text{since } Y_{n+1} \mid \mu \sim \mathcal{N}(\mu, \sigma^2) \\
 &= \sigma^2 + \text{Var} \left[\mu \mid \vec{Y}_n \right] && \text{by linearity} \\
 &= \sigma^2 + \tau_n^2 && \text{by Normal-Normal Conj.}
 \end{aligned}$$

Note $\tau_n^{-2} = n\sigma^{-2} + \tau_0^{-2}$.

4 Decision Theory

Definition 3 (Risk function). For an estimator $\hat{\theta} = T(Y)$, the **risk function** is the expected loss:

$$\text{Risk}(\theta) = \mathbb{E}_\theta \left[\text{Loss}(\theta, \hat{\theta}) \right] = \int \text{Loss}(\theta, T(y)) f_{Y;\theta}(y) dy.$$

Definition 4 (Admissibility). An estimator $\hat{\theta}$ is **inadmissible** if there exists another estimator with risk \leq that of $\hat{\theta}$ for all θ , with strict inequality for at least one θ . It is **admissible** if no such dominating estimator exists.

5 Hierarchical Models & Stein's Paradox

Idea. When data come from multiple groups—classrooms within a district, schools within a state—a hierarchical model places a prior on the group-level parameters. This induces *partial pooling*: groups with few observations are pulled toward the overall mean, while groups with many observations stay close to their own sample mean.

Definition 5 (Two-level Gaussian hierarchical model). For $j = 1, \dots, K$:

$$Y_j \mid \mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \quad \mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2).$$

Hyperparameters $(\sigma, \gamma, \lambda_0)$ known. The posterior is

$$\mu_j \mid y \sim \mathcal{N}(m_j, \lambda_K^2), \quad m_j = \lambda_K^2(\lambda_0^{-2}\gamma + \sigma^{-2}y_j), \quad \lambda_K^{-2} = \lambda_0^{-2} + \sigma^{-2}.$$

Marginally, $Y_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2)$.

- The posterior mean m_j is a precision-weighted average of γ (population mean) and y_j (individual observation). Groups with large σ^2 or few observations shrink more toward γ .

Definition 6 (Three-level Gaussian hierarchical model). For $j = 1, \dots, K$:

$$Y_j \mid \mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \quad \mu_j \mid \gamma \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2), \quad \gamma \sim \mathcal{N}(g_0, \tau_0^2).$$

The posteriors are $\gamma \mid y \sim \mathcal{N}(g_K, \tau_K^2)$ and

$$g_K = \frac{\tau_0^{-2}g_0 + (\sigma^2 + \lambda_0^2)^{-1}K\bar{y}}{\tau_0^{-2} + K(\sigma^2 + \lambda_0^2)^{-1}}, \quad \tau_K^{-2} = \tau_0^{-2} + K(\sigma^2 + \lambda_0^2)^{-1}.$$

5.1 Conditional independence in hierarchical models

Unconditionally, Y_1 and Y_2 are *not* independent—they share information about μ . Given μ_1, μ_2 , they *are* conditionally independent, because all of μ 's information is now fixed. This is the key structural feature of hierarchical models and is easiest to see by drawing the data-generating graph $\mu \rightarrow \mu_j \rightarrow Y_j$.

Concept Checker 3. You are given a mysterious die that may or may not be loaded (i.e., let p be the probability of rolling 6, and it is unknown whether $p > \frac{1}{6}$). You observe n rolls that are all 6 (i.e., $\vec{Y}_n = \vec{6}$)! Your friend will roll next, and denote this value as Y_{n+1} . Is $Y_{n+1} \perp\!\!\!\perp \vec{Y}_n$? What if, magically, we condition on $p = \frac{1}{6}$?

Solution

$Y_{n+1} \not\perp\!\!\!\perp \vec{Y}_n$ because Y_1, \dots, Y_n all being 6 gives information on the unknown probability p , which in turn gives information about Y_{n+1} . However, $(Y_{n+1} \perp\!\!\!\perp \vec{Y}_n) \mid p$. It is possible those n rolls were just very lucky!

5.2 Stein's Paradox

Theorem 1 (Stein, 1956). Let $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$ independently for $j = 1, \dots, K$ with $K \geq 3$ and σ^2 known. Under total squared error loss $\sum_j (\mu_j - \hat{\mu}_j)^2$, the MLE $\hat{\mu} = Y$ is *inadmissible*.

Theorem 2 (James-Stein estimator). Let $S = \sum_j Y_j^2$. The estimator

$$\hat{\mu}_j^{JS} = \left(1 - \frac{(K-2)\sigma^2}{S}\right) Y_j$$

has strictly lower risk than Y for all $\mu \in \mathbb{R}^K$. The risk of Y is $K\sigma^2$; the risk of $\hat{\mu}^{JS}$ is $[K - (K-2)^2\sigma^2\mathbb{E}(1/S)]\sigma^2$.

• **Bayesian interpretation.** Under $\mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\mu^2)$, the posterior mean is $(1-b)Y_j$ with $b = \sigma^2/(\sigma^2 + \sigma_\mu^2)$. Replacing σ_μ^2 with its unbiased estimator $S/K - \sigma^2$ and correcting for bias recovers the James-Stein estimator exactly. The shrinkage feels natural from the Bayesian perspective, but Stein's inadmissibility result is purely frequentist.

• ☠: The MLE is inadmissible for $K \geq 3$, but is admissible for $K \leq 2$. This shocked the statistical world in 1956.

Concept Checker 4 (Hierarchical Normal model (adapted from Leo Vanciu, STAT 220)).

For $j = 1, \dots, J$, suppose $\bar{y}_j \mid \theta_j \stackrel{\text{ind.}}{\sim} \mathcal{N}(\theta_j, \sigma_j^2)$ with known σ_j^2 , and

$$\theta_j \mid \mu, \tau \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \tau^2), \quad \pi(\mu, \tau) \propto 1.$$

1. By Normal-Normal conjugacy, derive $\theta_j \mid \mu, \tau, \bar{y}_j$. Write the posterior mean as a weighted average $\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j)\mu$ and identify λ_j . What happens as $\tau^2 \rightarrow \infty$? As $\tau^2 \rightarrow 0$?
2. Give an intuitive explanation of why groups with larger σ_j^2 are shrunk more toward μ .
3. Write down $p(\tilde{y}_j \mid y)$ for a future observation from *existing* group j , and $p(\tilde{y}_{\text{new}} \mid y)$ for a future observation from a *new* exchangeable group.

Solution

1. By Normal-Normal conjugacy with likelihood precision $1/\sigma_j^2$ and prior precision $1/\tau^2$:

$$\theta_j \mid \mu, \tau, \bar{y}_j \sim \mathcal{N}(\hat{\theta}_j, V_j), \quad V_j = \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1}, \quad \hat{\theta}_j = V_j \left(\frac{\bar{y}_j}{\sigma_j^2} + \frac{\mu}{\tau^2}\right).$$

Writing $\lambda_j = \tau^2/(\tau^2 + \sigma_j^2)$, this simplifies to $\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j)\mu$.

As $\tau^2 \rightarrow \infty$: $\lambda_j \rightarrow 1$, so $\hat{\theta}_j \approx \bar{y}_j$ (no pooling—the prior on θ_j is so diffuse it contributes nothing). As $\tau^2 \rightarrow 0$: $\lambda_j \rightarrow 0$, so $\hat{\theta}_j \approx \mu$ (complete pooling—all group means are forced toward the population mean).

2. Groups with larger σ_j^2 have noisier data, so \bar{y}_j is a less reliable estimate of θ_j . The model compensates by downweighting the data and leaning more on the population mean μ . Formally, $\lambda_j = \tau^2/(\tau^2 + \sigma_j^2)$ decreases as σ_j^2 increases, so $(1 - \lambda_j)$ (the weight on μ) increases.

3. For an *existing* group j , we condition on the group's own posterior:

$$p(\tilde{y}_j | y) = \int p(\tilde{y}_j | \theta_j) p(\theta_j, \mu, \tau | y) d\theta_j d\mu d\tau.$$

For a *new* exchangeable group, we first draw a fresh $\tilde{\theta}$ from the population distribution:

$$p(\tilde{y}_{\text{new}} | y) = \int p(\tilde{y}_{\text{new}} | \tilde{\theta}) p(\tilde{\theta} | \mu, \tau) p(\mu, \tau | y) d\tilde{\theta} d\mu d\tau.$$

The existing-group predictive uses the posterior for θ_j already updated by group j 's data; the new-group predictive draws $\tilde{\theta}$ fresh from the hyperprior, so it is more uncertain—it must account for the full between-group variability τ^2 on top of within-group variance σ^2 .

6 Practice Problems

Problem 1 (Coverage of posterior intervals). Consider any parametric model with scalar parameter θ .

- Prove that if θ is drawn from the prior and $Y | \theta$ from the data model, a Bayesian 50% credible interval contains the true θ with probability exactly 50%.
- Suppose $\theta \sim \mathcal{N}(0, 4)$ and $Y | \theta \sim \mathcal{N}(\theta, 1)$. The true value is $\theta_0 = 1$. What is the exact frequentist coverage of the posterior 50% interval?

Solution

(a) Let $I = \mathbf{1}(\theta \in C(Y))$. By construction of the 50% credible interval, $P(I = 1 | Y) = 0.5$. By Adam's law:

$$P(I = 1) = \mathbb{E}[I] = \mathbb{E}[\mathbb{E}[I | Y]] = \mathbb{E}[0.5] = 0.5. \quad \square$$

(b) With $\tau_0^2 = 4$, $\sigma^2 = 1$: $\tau_1^{-2} = 5/4$, so $\tau_1^2 = 4/5$ and $\mu_1 = 4Y/5$. The 50% central credible interval is $C(Y) = [4Y/5 + \tau_1 \Phi^{-1}(0.25), 4Y/5 + \tau_1 \Phi^{-1}(0.75)]$. Coverage at $\theta_0 = 1$: $1 \in C(Y)$ iff $|4Y/5 - 1| \leq \sqrt{4/5} \cdot 0.6745 \approx 0.603$. Since $Y | \theta_0 = 1 \sim \mathcal{N}(1, 1)$, we have $4Y/5 | \theta_0 = 1 \sim \mathcal{N}(4/5, 16/25)$ and

$$P\left(\left|\frac{4Y}{5} - 1\right| \leq 0.603 \mid \theta_0 = 1\right) \approx 0.535.$$

Coverage is $\approx 53.5\%$, higher than 50% because the prior pulls the interval toward 0, making it more likely to cover the positive true value $\theta_0 = 1$.

Problem 2 (Hierarchical model for SAT scores). Randomized control trials on a tutoring program are run in $J = 7$ schools. Let y_j be the estimated average treatment effect with known standard error σ_j :

School j	y_j	σ_j
1	3	8
2	16	6
3	-6	12
4	-7	6
5	4	8
6	8	25
7	-1	5

Assume $Y_j | \theta_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_j, \sigma_j^2)$, $\theta_j | \mu, \lambda \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \lambda^2)$, $\mu | \lambda \sim \mathcal{N}(0, 6^2)$, with λ known.

- Find $\theta_j | \mu, Y, \lambda$. Explain why $\mathbb{E}[\theta_j | \mu, Y, \lambda]$ makes sense.
- Find $\mu | Y, \lambda$. Explain why $\mathbb{E}[\mu | Y, \lambda]$ makes sense.
- Find $\theta_j | Y, \lambda$. Explain why $\mathbb{E}[\theta_j | Y, \lambda]$ makes sense.
- Plot $\mathbb{E}[\theta_j | Y, \lambda]$ vs. $\lambda \in [0, 40]$ for all 7 schools on one graph (different colors and line types). Hint: `matplot` in R.
- Describe a principled way to infer λ if it is unknown.

Solution

(a) By Normal-Normal conjugacy:

$$\theta_j | Y_j, \mu, \lambda \sim \mathcal{N}((1 - b_j)Y_j + b_j\mu, b_j\lambda^2), \quad b_j = \frac{\sigma_j^2}{\sigma_j^2 + \lambda^2}.$$

The posterior mean is a weighted average of the data Y_j and prior mean μ : more weight on Y_j when σ_j^2 is small (precise school estimate) or λ^2 is large (diffuse prior on θ_j).

(b) Since $(Y_1, \dots, Y_J, \theta_1, \dots, \theta_J, \mu)$ is jointly MVN, by Adam's and Eve's laws $Y_j | \mu, \lambda \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma_j^2 + \lambda^2)$. Completing the square in the log-posterior gives $\mu | Y, \lambda \sim \mathcal{N}(m_P, s_P^2)$ where

$$s_P^2 = \frac{1}{6^{-2} + \sum_j (\sigma_j^2 + \lambda^2)^{-1}}, \quad m_P = s_P^2 \sum_j \frac{Y_j}{\sigma_j^2 + \lambda^2}.$$

This is a precision-weighted average of the y_j 's and the prior mean 0, weighting more informative schools (smaller σ_j^2) more heavily.

(c) By Adam's and Eve's laws applied to the results of (a) and (b):

$$\mathbb{E}[\theta_j | Y, \lambda] = (1 - b_j)Y_j + b_j m_P, \quad \text{Var}(\theta_j | Y, \lambda) = b_j \lambda^2 + b_j^2 s_P^2.$$

This replaces the unknown μ from (a) with its posterior estimate m_P . The extra $b_j^2 s_P^2$ term in the variance reflects uncertainty about μ itself.

(d) R code:

```

y    <- c(3, 16, -6, -7, 4, 8, -1)
sig  <- c(8, 6, 12, 6, 8, 25, 5)
sig2 <- sig^2
sp2  <- 1 / (1/36 + sum(1/sig2))
mp   <- sp2 * sum(y / sig2)
X    <- matrix(0, 40, 7)
for (i in 1:40) {
  lam2 <- (i - 1)^2
  b    <- sig2 / (lam2 + sig2)
  X[i,] <- (1 - b) * y + b * mp
}
matplot(0:39, X, type = "l", lty = 1:7, col = 1:7,
        xlab = "lambda", ylab = "Posterior mean",
        main = "E[theta_j | Y, lambda] vs lambda")
legend("topright", legend = paste("School", 1:7),
      lty = 1:7, col = 1:7, cex = 0.7)

```

When $\lambda = 0$: complete pooling, all estimates equal m_P . As $\lambda \rightarrow \infty$: $b_j \rightarrow 0$, so $\mathbb{E}[\theta_j | Y, \lambda] \rightarrow Y_j$ (no pooling).

(e) Treat λ as unknown and form the marginal likelihood $f(y_1, \dots, y_J | \lambda)$, which is tractable here (a product of Normal densities after integrating out θ_j and μ). Introduce a prior $\pi(\lambda)$ and apply Bayes' theorem to obtain a posterior $\pi(\lambda | y)$.

Problem 3 (James-Stein and batting averages). A sabermetrician estimates batting averages μ_1, \dots, μ_k of $k > 3$ players. Let Y_j be the proportion of hits in n at-bats. Assume $Y_j | \mu_j \stackrel{\text{ind.}}{\sim} \mathcal{N}(\mu_j, V)$ with $\mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_\mu^2)$ a priori. Use total squared error loss $C(\mu, \hat{\mu}) = \sum_j (\mu_j - \hat{\mu}_j)^2$.

- Find $\hat{\mu}_{\text{MLE}}$ and the risk $r_{\text{MLE}}(\mu)$.
- The Bayes estimator has j th component $\mathbb{E}[\mu_j | Y] = b\mu_0 + (1 - b)Y_j$ where $b = V/(V + \sigma_\mu^2)$. Find $r_{\text{Bayes}}(\mu)$. Is it always smaller than $r_{\text{MLE}}(\mu)$?
- Explain intuitively how $\hat{\mu}_{\text{Bayes}}$ relates to regression toward the mean.
- Now suppose μ_0, σ_μ^2 are unknown. Let $S = \sum_j (Y_j - \bar{Y})^2$. Show $\hat{b} = (k - 3)V/S$ is unbiased for b . The James-Stein estimator is $\hat{\mu}_{j,\text{JS}} = \hat{b}\bar{Y} + (1 - \hat{b})Y_j$.
- Using the Efron-Morris baseball dataset (18 players, first $n = 45$ at-bats, $V = \bar{y}(1 - \bar{y})/45$): compute $\hat{\mu}_{\text{JS}}$ for each player. What percentage beat the MLE? Compare total squared error losses.

Solution

(a) The likelihood $L(\mu) \propto \prod_i \exp\{-(y_i - \mu_i)^2/(2V)\}$ is maximized at $\hat{\mu}_{\text{MLE}} = (Y_1, \dots, Y_k)$. The risk is constant: $r_{\text{MLE}}(\mu) = \sum_i \mathbb{E}[(Y_i - \mu_i)^2 | \mu_i] = kV$.

(b) Expanding the squared error:

$$\begin{aligned} r_{\text{Bayes}}(\mu) &= \sum_i \mathbb{E}[(b(\mu_0 - \mu_i) + (1 - b)(Y_i - \mu_i))^2 \mid \mu_i] \\ &= b^2 \sum_i (\mu_0 - \mu_i)^2 + k(1 - b)^2 V. \end{aligned}$$

When $\mu_i = \mu_0$ for all i : $r_{\text{Bayes}} = k(1 - b)^2 V < kV = r_{\text{MLE}}$ since $0 < b < 1$. But when $\|\mu - \mu_0 \mathbf{1}\|$ is large, $r_{\text{Bayes}} > r_{\text{MLE}}$. So it depends on μ —the Bayes estimator is not uniformly better.

(c) $\hat{\mu}_{\text{Bayes}}$ shrinks the MLE toward the prior mean μ_0 . A player with an unusually high batting average in the first n at-bats likely benefited from both skill and luck; regression toward the mean predicts their true average is closer to the population mean. The Bayes estimator captures this by blending μ_0 with the data, rather than taking the MLE at face value. When μ_0 is unknown, data from other players informs its estimate, creating indirect connections between individual estimates.

(d) Since $Y_j \mid \mu_0, V, \sigma_\mu^2 \sim \mathcal{N}(\mu_0, V + \sigma_\mu^2)$, the sample variance satisfies

$$S \sim (V + \sigma_\mu^2) \chi^2(k - 1) \sim \text{Gamma}\left(\frac{k-1}{2}, \frac{1}{2(V + \sigma_\mu^2)}\right).$$

Set $a = (k - 1)/2$, $\lambda_0 = 1/[2(V + \sigma_\mu^2)]$. By LOTUS and pattern-matching to the Gamma PDF:

$$\mathbb{E}[\hat{b}] = (k-3)V \cdot \mathbb{E}[1/S] = (k-3)V \cdot \frac{\lambda_0}{a-1} = (k-3)V \cdot \frac{1}{2(V + \sigma_\mu^2)} \cdot \frac{2}{k-3} = \frac{V}{V + \sigma_\mu^2} = b. \quad \square$$

(e) R code:

```
baseball_data <- read.csv("battingaverages.csv")
mu.true <- baseball_data$mu
Y <- baseball_data$hits1 / 45
k <- length(Y)
V <- mean(Y) * (1 - mean(Y)) / 45
S <- sum((Y - mean(Y))^2)
b.hat <- (k - 3) * V / S
mu.js <- b.hat * mean(Y) + (1 - b.hat) * Y
mean(abs(mu.js - mu.true) <= abs(Y - mu.true))
sum((mu.js - mu.true)^2)
sum((Y - mu.true)^2)
```

JS beats the MLE for 88.9% of players (16 out of 18). Total squared error: JS is 0.0214 vs. the MLE's 0.0755—roughly $3.5\times$ improvement, a striking practical illustration of Stein's paradox.

Summary Table

Formula or idea	Description or name
$f(y \text{Bill})/g(y \text{Jose})$	Bayes factor for model comparison
Post. odds = prior odds \times BF	Bayes factor converts prior to posterior odds
$f(y) = f(y \theta = t)\pi(t)/\pi(t y)$	Candidate's formula
$P(\theta \in \Theta_0 y)$	Posterior probability of null
Simple null needs positive prior mass at θ_0	Cromwell's rule for hypothesis testing
BF log \approx BIC; penalizes complexity for large n	Lindley's paradox
$f(\tilde{y} y) = \int f(\tilde{y} \theta)\pi(\theta y) d\theta$	Posterior predictive
$\mathbb{E}[\tilde{Y} y] = \mathbb{E}[\theta y]$	Posterior predictive mean (Adam's law)
$\text{Var}(\tilde{Y} y) = \sigma^2 + \tau_n^2$	Aleatoric + epistemic variance (Eve's law)
$\text{Risk}(\theta) = \mathbb{E}_\theta[\text{Loss}(\theta, \hat{\theta})]$	Risk function
Inadmissible: dominated by another estimator	Admissibility
$Y_j \mu_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \mu_j \gamma \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2)$	Two-level hierarchical model
$m_j = \lambda_K^2(\lambda_0^{-2}\gamma + \sigma^{-2}y_j)$	Partial pooling posterior mean
Unconditionally dependent; conditionally independent given μ_j	Hierarchical dependence structure
MLE inadmissible for $K \geq 3$ under squared loss	Stein's paradox
$(1 - (K - 2)\sigma^2/S)Y_j$	James-Stein estimator
$b = \sigma^2/(\sigma^2 + \sigma_\mu^2)$; posterior mean = $(1 - b)Y_j + b\mu_0$	Bayesian shrinkage and JS connection