Introduction to Bayesian Inference

version: 2018-02-28 · 11:38:38

Motivation

- Study by Dr Maya Yampolsky, Université Laval on Multicultural Identity Integration (www.iriscouples.com)
- $n = 524$ subjects
- $p = 22$ questions, e.g.
  - "I identify with one culture more than any other."
  - "Each of my cultural identities is a separate part of who I am."
  - "My cultural identities complement each other."
- *Ordinal Responses:* (7 choices)
  A = Not at all, B = Slightly, ..., F = Mostly, G = Exactly
- **Goal:** Correlation between questions

Motivation

Solution 1:

1. Ignore the fact that e.g., “Slightly Agree” means different things to different people
2. Encode ordinals as numbers: $\mathsf{A} = 1, \ldots, \mathsf{G} = 7$
3. Calculate correlation matrix

But what if $B - A \neq C - B$?

Motivation

Solution 2:

1. Assume that each question corresponds to a *latent* variable x_j
2. Complete data: $\mathbf{x} = (x_1, \ldots, x_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
3. Observed data: $\mathbf{y} = (y_1, \ldots, y_p)$,

$$y_j = \begin{cases} A & x_j < \lambda_{j1} \\ B & \lambda_{1j} \le x_j < \lambda_{j2} \\ \vdots & \\ F & \lambda_{j5} \le x_j \le \lambda_{j6} \\ G & \lambda_{j6} \le x_j \end{cases}$$

4. Calculate $\hat{\mathbf{\Sigma}} = \arg\max_{(\mathbf{\Sigma}, \boldsymbol{\lambda})} \ell(\mathbf{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y})$

The correlation matrix $\mathbf{\Sigma}$ of the latent variable $\mathbf{x}$ is called the polychoric correlation of the observed variable $\mathbf{y}$.

Motivation

Solution 2:

1. Complete data: $\mathbf{x} = (x_1, \dots, x_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$,

   Observed data: $\mathbf{y} = (y_1, \dots, y_p)$, y_j is bin number

2. Calculate $\hat{\boldsymbol{\Sigma}} = \arg\max_{(\boldsymbol{\Sigma},\boldsymbol{\lambda})} \ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y})$

Complete Data Likelihood: $\ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}, \mathbf{X})$ is easy to maximize:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i', \qquad \hat{\boldsymbol{\lambda}}_j \cong \texttt{qnorm}(\bar{\mathbf{Y}}_j, \mathbf{0}, \hat{\boldsymbol{\Sigma}}_{jj})$$

E-Step: Need $E[\mathbf{x} \mid \mathbf{y}]$, where $\mathbf{x}$ is a *multivariate truncated normal*.

Motivation

Solution 2:

1. Complete data: $\mathbf{x} = (x_1, \ldots, x_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$,

   Observed data: $\mathbf{y} = (y_1, \ldots, y_p)$, y_j is bin number

2. Calculate $\hat{\boldsymbol{\Sigma}} = \arg\max_{(\boldsymbol{\Sigma}, \boldsymbol{\lambda})} \ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y})$

Complete Data Likelihood: $\ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}, \mathbf{X})$ is easy to maximize:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i', \qquad \hat{\boldsymbol{\lambda}}_j \cong \mathtt{qnorm}(\bar{\mathbf{Y}}_j, \mathbf{0}, \hat{\boldsymbol{\Sigma}}_{jj})$$

However, a *univariate* truncated normal can easily be *simulated*:

$$z \sim \mathcal{N}(\mu, \sigma^2) \times \mathbb{1}\{L < y < U\}$$
$$\iff z = \mathtt{qnorm}(r, \mu, \sigma^2), \qquad r \sim \mathrm{Unif}(L, U).$$

Motivation

Stochastic EM:

- Step t has $\hat{\boldsymbol{\Sigma}}^{(t)}$, $\hat{\boldsymbol{\lambda}}^{(t)}$, *and* $\mathbf{X}^{(t)} = (\mathbf{X}_1^{(t)}, \dots, \mathbf{X}_p^{(t)})$
- **S-Step:** (Simulation instead of Expectation)
  - Set $\tilde{\mathbf{X}} = \mathbf{X}^{(t)}$
  - For each i, draw $x_{i1} \mid y_{i1}, \tilde{\mathbf{x}}_{i,-1}$ from the corresponding univariate truncated normal, then set $\tilde{\mathbf{X}} = \mathbf{X}_1 \cup \tilde{\mathbf{X}}_{-1}$.
  - For each i, draw $x_{i2} \mid y_{i2}, \tilde{\mathbf{x}}_{i,-2}$ from univariate truncated normal, then set $\tilde{\mathbf{X}} = \mathbf{X}_2 \cup \tilde{\mathbf{X}}_{-2}$
  - Do this for each $j = 1, \dots, p$, then set $\mathbf{X}^{(t+1)} = \tilde{\mathbf{X}}$
- **M-Step**: $(\hat{\boldsymbol{\Sigma}}^{(t+1)}, \hat{\boldsymbol{\lambda}}^{(t+1)}) = \arg\max_{(\boldsymbol{\Sigma}, \boldsymbol{\lambda})} \ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}, \mathbf{X}^{(t+1)})$.

Motivation

Stochastic EM:

- Step t has $\hat{\boldsymbol{\Sigma}}^{(t)}$, $\hat{\boldsymbol{\lambda}}^{(t)}$, *and* $\mathbf{X}^{(t)} = (\mathbf{X}_1^{(t)}, \ldots, \mathbf{X}_p^{(t)})$
- **S-Step:** (Simulation instead of Expectation)
  - Sequentially update each element of $\mathbf{X}$ by drawing from univariate truncated normal conditioning on everything else
- **M-Step**: $(\hat{\boldsymbol{\Sigma}}^{(t+1)}, \hat{\boldsymbol{\lambda}}^{(t+1)}) = \arg\max_{(\boldsymbol{\Sigma},\boldsymbol{\lambda})} \ell(\boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}, \mathbf{X}^{(t+1)})$.

- Does not converge to a single $\hat{\boldsymbol{\Sigma}}$, instead produces a Markov chain.
- $\hat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{t=1}^{M} \hat{\boldsymbol{\Sigma}}^{(t)}$ is a *consistent* estimator of $\boldsymbol{\Sigma}$, but less efficient that MLE
- Now we'll see how to do it as efficiently, more generally, but we have to dance with the devil...

Recap of Frequentist Inference

- **Model:**

$$\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p).$$

- **Likelihood:**

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}).$$

For calculations, often more useful to work with the **loglikelihood**:

$$\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}).$$

Recap of Frequentist Inference

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$
- **Point Estimate:** Maximum likelihood estimator (MLE)

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \mathbf{y})$$

  Asymptotically as $n \to \infty$, we have

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})),$$

  where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the (expected) Fisher Information:

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}_0 \mid \mathbf{y})\right] = -\int \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}_0 \mid \mathbf{y}) \cdot p(\mathbf{y} \mid \boldsymbol{\theta}_0) \, \mathrm{d}\mathbf{y},$$

  and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$ is the true parameter value.

Recap of Frequentist Inference

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$
- **MLE:** $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \mathbf{y}) \approx \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})), \qquad \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta} \mid \mathbf{y})\right].$
- **Confidence Interval:**
  - For each θ_i, want a pair of random variables $L = L(\mathbf{y})$ and $U = U(\mathbf{y})$ such that $\Pr(L < \theta_i < U) = 95\%$.
  - *Observed Fisher Information:* $\hat{\boldsymbol{\mathcal{I}}} = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\hat{\boldsymbol{\theta}} \mid \mathbf{y}) \xrightarrow{n} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

    $\implies \hat{\theta}_i \approx \mathcal{N}(\theta_i, [\hat{\boldsymbol{\mathcal{I}}}^{-1}]_{ii})$

    $\implies$ (approximate) 95% CI for θ_i:

$$\hat{\theta}_i \pm 1.96 \times \mathrm{se}(\hat{\theta}_i), \qquad \mathrm{se}(\hat{\theta}_i) = \sqrt{[\hat{\boldsymbol{\mathcal{I}}}^{-1}]_{ii}}.$$

Recap of Frequentist Inference

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$
- **MLE:** $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \mathbf{y}) \approx \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}))$
- **Hypothesis Testing:**

  1. $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$
  2. *Test statistic:* $T = T(\mathbf{y})$, large values of T are evidence against H_0
  3. *p-value:*

     $$p_v = \Pr(T > T_{\text{obs}} \mid H_0),$$

     where $T_{\text{obs}} = T(\mathbf{y}_{\text{obs}})$ is calculated for current dataset, and $T = T(\mathbf{y})$ is for a new dataset.

     - p_v is probability of observing more evidence against H_0 in new data than current data, given that H_0 is true.
     - Typically $p(T \mid H_0)$ doesn't exist, only $p(T \mid \boldsymbol{\theta})$. So often use an asymptotic p-value

     $$p_v \approx \Pr(T > T_{\text{obs}} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0), \qquad \hat{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \ell(\boldsymbol{\theta} \mid \mathbf{y}).$$

Bayesian Inference

- **Model:** $\mathbf{y} = (y_1, \dots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$
- **Likelihood:** $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \prod_{i=1}^n f(y_i \mid \boldsymbol{\theta})$
- **Prior Distribution:** $\pi(\boldsymbol{\theta})$
- **Posterior Distribution:**

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})} \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \cdot \pi(\boldsymbol{\theta})$$

IGNORE everything that doesn't depend on $\boldsymbol{\theta}$.

I.e., if $g(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid \mathbf{y})$, then

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = Z^{-1} g(\boldsymbol{\theta}), \qquad Z = \int g(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},$$

where Z is the *normalizing constant*.

Bayesian Inference

Model: $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$

- **Prior Distribution:** $\pi(\boldsymbol{\theta})$
- **Posterior Distribution:** $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \cdot \pi(\boldsymbol{\theta})$
- **Point Estimate:** $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta} \mid \mathbf{y}]$
- **Interval Estimate:** (L, U) such that $\Pr(L < \theta_i < U \mid \mathbf{y}) = 95\%$

  No asymptotics, and conditioned on this $\mathbf{y}$
- **Hypothesis Testing:** For *nondegenerate* $H_0 : \theta_j \in \Theta_{j0}$,

  simply calculate $\Pr(H_0 \mid \mathbf{y}) = \Pr(\theta_j \in \Theta_{j0} \mid \mathbf{y})$!

Bayesian Inference

Model: $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$

- **Prior Distribution:** $\pi(\boldsymbol{\theta})$
- **Posterior Distribution:** $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \cdot \pi(\boldsymbol{\theta})$
- **Point Estimate:** $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta} \mid \mathbf{y}]$
- **Interval Estimate:** (L, U) such that $\Pr(L < \theta_i < U \mid \mathbf{y}) = 95\%$

  No asymptotics, and conditioned on this $\mathbf{y}$
- **Hypothesis Testing:** For *sharp* $H_0 : \theta_j = \theta_{j0}$,
  - Test statistic: $T = T(\mathbf{y}) \sim f(T \mid \boldsymbol{\theta})$
  - *Posterior p-value:*

$$\Pr(T > T_{\text{obs}} \mid \mathbf{y}_{\text{obs}}, H_0) = \int \Pr(T > T_{\text{obs}} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}, \theta_j = \theta_{j0}) \, \mathrm{d}\boldsymbol{\theta}.$$

  No asymptotics!

Example I

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$
- **Likelihood:**

$$\ell(\mu \mid \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 = -\frac{n}{2}(\bar{y} - \mu)^2,$$

  where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- **Prior Specification: ALWAYS** in this order:
  1. What prior information do we have about μ?
  2. What would make calculations simple?

Example I

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$
- **Likelihood:** $\ell(\mu \mid \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2 = -\frac{n}{2}(\bar{y} - \mu)^2$
- **Prior Specification:** **ALWAYS** in this order:
  1. What prior information do we have about μ?
  2. What would make calculations simple?

  In this case, a convenient choice is $\mu \sim \mathcal{N}(\lambda, \tau^2)$, since

$$\begin{aligned}\log p(\mu \mid \mathbf{y}) &= \ell(\mu \mid \mathbf{y}) + \log \pi(\mu) \\ &= -\frac{n(\bar{y} - \mu)^2}{2} - \frac{(\lambda - \mu)^2}{2\tau^2} = -\frac{(\mu - B\lambda - (1 - B)\bar{y})^2}{2(1 - B)/n},\end{aligned}$$

  where $B = \frac{1}{n}/(\frac{1}{n} + \tau^2) \in (0, 1)$ is called the *shrinkage factor*.

$$\implies \qquad \mu \mid \mathbf{y} \sim \mathcal{N}\left(B\lambda + (1 - B)\bar{y}, \frac{1 - B}{n}\right).$$

Example I

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$
- **Likelihood:** $\ell(\mu \mid \mathbf{y}) = -\frac{n}{2}(\bar{y} - \mu)^2$ **Prior:** $\mu \sim \mathcal{N}(\lambda, \tau^2)$
- **Posterior:** $\mu \mid \mathbf{y} \sim \mathcal{N}\left(B\lambda + (1-B)\bar{y}, \frac{1-B}{n}\right)$, $B = \frac{1}{n}/(\frac{1}{n} + \tau^2)$.

1. $\log p(\mu \mid \mathbf{y}) = -\frac{1}{2}[n(\bar{y} - \mu)^2 + \tau^{-2}(\lambda - \mu)^2] = \ell(\mu \mid \mathbf{y}, \tilde{\mathbf{y}})$,

   where $\tilde{\mathbf{y}}$ consists of τ^{-2} additional data points with mean λ.

   $\Longrightarrow$ Think of the prior as adding "fake" data to the data you already have.

2. As $\tau \to \infty$, posterior converges to $\mu \mid \mathbf{y} \sim \mathcal{N}(\bar{y}, \frac{1}{n})$.

   Gives exactly same point and interval estimate as Frequentist inference.

   But as $\tau \to \infty$ we have $\pi(\mu) \propto 1$ which is not a PDF...

General Case: Exponential Families

- **Model:** $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n) \overset{\text{iid}}{\sim} \exp\left\{\mathbf{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right\} \cdot h(\mathbf{y})$
- **Likelihood:**
$$\ell(\boldsymbol{\eta} \mid \mathbf{Y}) = \sum_{i=1}^{n} \left[\mathbf{T}_i'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right]$$
$$= n\left[\bar{\mathbf{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right], \qquad \bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}_i$$
- **Conjugate Prior:**
$$\pi(\boldsymbol{\eta}) = g(\boldsymbol{\eta} \mid \mathbf{T}_0, \nu_0)$$
$$\propto \exp\left\{\nu_0\left[\mathbf{T}_0{}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right]\right\}$$
- **Posterior Distribution:** Has same form as the prior:
$$\log p(\boldsymbol{\eta} \mid \mathbf{Y}) = n\left[\bar{\mathbf{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right] + \nu_0\left[\mathbf{T}_0{}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right]$$
$$\implies \qquad \boldsymbol{\eta} \mid \mathbf{Y} \sim g\left(\boldsymbol{\eta} \mid \tfrac{n}{n+\nu_0}\bar{\mathbf{T}} + \tfrac{\nu_0}{n+\nu_0}\mathbf{T}_0, n + \nu_0\right)$$

General Case: Exponential Families

- **Model:** $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \overset{\text{iid}}{\sim} \exp\left\{\mathbf{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right\} \cdot h(\mathbf{y})$
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \mathbf{Y}) = n\left[\bar{\mathbf{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right], \qquad \bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}_i$
- **Conjugate Prior:** $\pi(\boldsymbol{\eta}) = g(\boldsymbol{\eta} \mid \mathbf{T}_0, \nu_0) \propto \exp\left\{\nu_0\left[\mathbf{T}_0{}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right]\right\}$
- **Posterior Distribution:**

$$\boldsymbol{\eta} \mid \mathbf{Y} \sim g\left(\boldsymbol{\eta} \mid \tfrac{n}{n+\nu_0}\bar{\mathbf{T}} + \tfrac{\nu_0}{n+\nu_0}\mathbf{T}_0, n + \nu_0\right)$$

  - **Interpretation:** The conjugate prior family is not unique, but the one above is proportional to the likelihood.

    In this case, the prior is as if we'd observed ν_0 additional observations with average sufficient statistic $\mathbf{T}_0$.

    An example of a conjugate prior not proportional to $\mathcal{L}(\boldsymbol{\eta} \mid \mathbf{Y})$: mixture of above priors, i.e.,

$$\pi(\boldsymbol{\eta}) = \rho \cdot g(\boldsymbol{\eta} \mid \mathbf{T}_1, \nu_1) + (1 - \rho) \cdot g(\boldsymbol{\eta} \mid \mathbf{T}_2, \nu_2).$$

General Case: Exponential Families

- **Model:** $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \overset{\text{iid}}{\sim} \exp\left\{\mathbf{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right\} \cdot h(\mathbf{y})$
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \mathbf{Y}) = n\left[\bar{\mathbf{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right], \qquad \bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^n \mathbf{T}_i$
- **Conjugate Prior:** $\pi(\boldsymbol{\eta}) = g(\boldsymbol{\eta} \mid \mathbf{T}_0, \nu_0) \propto \exp\left\{\nu_0\left[\mathbf{T}_0{}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right]\right\}$
- **Posterior Distribution:**

$$\boldsymbol{\eta} \mid \mathbf{Y} \sim g\left(\boldsymbol{\eta} \mid \tfrac{n}{n+\nu_0}\bar{\mathbf{T}} + \tfrac{\nu_0}{n+\nu_0}\mathbf{T}_0, n + \nu_0\right)$$

  - **Improper Priors:** As $\nu_0 \to 0$ we get $\pi(\boldsymbol{\eta}) \propto 1$, and thus $p(\boldsymbol{\eta} \mid \mathbf{Y}) \propto \mathcal{L}(\boldsymbol{\eta} \mid \mathbf{Y})$.

    However, $\pi(\boldsymbol{\eta}) \propto 1$ typically doesn't integrate to 1, so are we allowed to use this as a prior?

    OK as long as $\int \mathcal{L}(\boldsymbol{\eta} \mid \mathbf{Y})\pi(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} < \infty$. This is because the posterior is

$$p(\boldsymbol{\eta} \mid \mathbf{Y}) = \frac{\mathcal{L}(\boldsymbol{\eta} \mid \mathbf{Y})\pi(\boldsymbol{\eta})}{\int \mathcal{L}(\boldsymbol{\eta} \mid \mathbf{y})\pi(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}},$$

    so get a valid distribution as long as denominator is finite.

Example I (Continued)

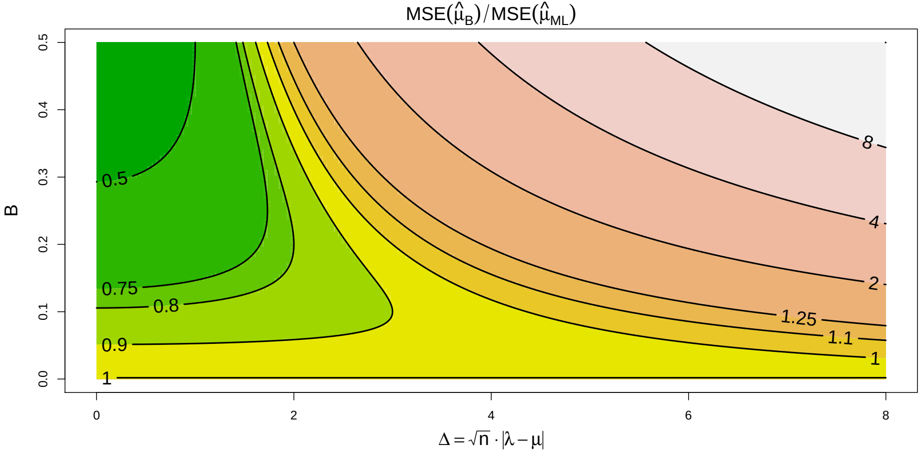
- **Model:** $\mathbf{y} = (y_1, \dots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$
- **Likelihood:** $\ell(\mu \mid \mathbf{y}) = -\frac{n}{2}(\bar{y} - \mu)^2$ **Prior:** $\mu \sim \mathcal{N}(\lambda, \tau^2)$
- **Posterior:** $\mu \mid \mathbf{y} \sim \mathcal{N}\left(B\lambda + (1 - B)\bar{y}, \frac{1-B}{n}\right)$, $B = (\frac{1}{n})/(\frac{1}{n} + \tau^2)$
- **Comparison:** $\hat{\mu}_{\text{ML}} = \bar{y}$ vs. $\hat{\mu}_{\text{B}} = E[\mu \mid \mathbf{y}] = B\lambda + (1 - B)\bar{y}$.
  - Metric: mean square error

$$\text{MSE}(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = (\underbrace{E[\hat{\mu}] - \mu}_{\text{Bias}(\hat{\mu})})^2 + \text{var}(\hat{\mu})$$

  - $\text{MSE}(\hat{\mu}_{\text{ML}}) = 1/n$, $\text{MSE}(\hat{\mu}_{\text{B}}) = B^2(\lambda - \mu)^2 + (1 - B)^2/n$.
  - Plot $\text{MSE}(\hat{\mu}_{\text{B}})/\text{MSE}(\hat{\mu}_{\text{ML}})$ as a function of $\Delta = n^{1/2}|\lambda - \mu|$ and B.

Example I (Continued)

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$
- **Likelihood:** $\ell(\mu \mid \mathbf{y}) = -\frac{n}{2}(\bar{y} - \mu)^2$ **Prior:** $\mu \sim \mathcal{N}(\lambda, \tau^2)$
- **Posterior:** $\mu \mid \mathbf{y} \sim \mathcal{N}\left(B\lambda + (1 - B)\bar{y}, \frac{1-B}{n}\right)$, $B = (\frac{1}{n})/(\frac{1}{n} + \tau^2)$



Example I

Summary:

- Many statistical models have conjugate priors, which one can think of as adding fake data to the data we have already observed.
- Priors don't need to integrate to 1, as long as the posterior does. This can be useful to avoid thinking too much about what prior to use, i.e., simply use $\pi(\boldsymbol{\theta}) \propto 1$.

Example II

- **Model:** $\mathbf{y} = (y_1, \dots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
- **Likelihood:**

$$\mathcal{L}(\sigma^2 \mid \mathbf{y}) \propto \exp\left\{-\frac{n}{2}\log\sigma^2 - \frac{S^2/2}{\sigma^2}\right\}, \qquad S = \sum_{i=1}^{n} y_i^2.$$

- **Conjugate Prior:**

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta)$$
$$\iff \pi(\sigma^2) \propto \exp\left\{-(\alpha+1)\log\sigma^2 - \frac{\beta}{\sigma^2}\right\}$$

- **Posterior Distribution:**

$$\sigma^2 \mid \mathbf{y} \sim \text{Inv-Gamma}\left(\tfrac{n}{2} + \alpha, \tfrac{S}{2} + \beta\right)$$

Example II

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
- **Likelihood:** $\ell(\sigma^2 \mid \mathbf{y}) = -\frac{1}{2}\left(S/\sigma^2 + n \log \sigma^2\right), \qquad S = \sum_{i=1}^n y_i^2.$
- **Conjugate Prior:** $\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta) \iff \pi(\sigma^2) \propto (1/\sigma^2)^{\alpha+1} e^{-\beta/\sigma^2}$
- **Posterior Distribution:**

$$\sigma^2 \mid \mathbf{y} \sim \text{Inv-Gamma}\left(\tfrac{n}{2} + \alpha, \tfrac{S}{2} + \beta\right) \qquad \Longrightarrow \qquad \hat{\sigma}^2_{\mathrm{B}} = E[\sigma^2 \mid \mathbf{y}] = \frac{\frac{S}{2} + \beta}{\frac{n}{2} + \alpha - 1}$$

Prior	(α, β)	$\pi(\sigma^2)$	$\hat{\sigma}^2_{\mathrm{B}}$
Flat	$(-1, 0)$	$\propto 1$	$S/(n-4)$
MLE-matching	$(1, 0)$	$\propto 1/\sigma^4$	S/n $(= \hat{\sigma}^2_{\mathrm{ML}})$

Example II

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
- **Likelihood:** $\ell(\sigma^2 \mid \mathbf{y}) = -\frac{1}{2}\left(S/\sigma^2 + n \log \sigma^2\right), \qquad S = \sum_{i=1}^n y_i^2.$
- **Maximum Likelihood Estimate:**
  - *For variance* σ^2: $\hat{\sigma}^2_{\mathrm{ML}} = S/n$
  - *For precision* $\tau^2 = 1/\sigma^2$: $\hat{\tau}^2_{\mathrm{ML}} = n/S = 1/\hat{\sigma}^2_{\mathrm{ML}}$.
- **Invariance Principle:** For given $\ell(\boldsymbol{\theta} \mid \mathbf{y})$, if $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ and g is a bijection, then can reparametrize the model via $\ell(\boldsymbol{\eta} \mid \mathbf{y}) = \ell(\boldsymbol{\theta} = g^{-1}(\boldsymbol{\eta}) \mid \mathbf{y})$, such that

$$\begin{aligned} &\max_{\boldsymbol{\eta}} \ell(\boldsymbol{\eta} \mid \mathbf{y}) \le \ell(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} \mid \mathbf{y}) = \ell(\boldsymbol{\eta} = g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) \mid \mathbf{y}) \\ \implies & \hat{\boldsymbol{\eta}}_{\mathrm{ML}} = g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}). \end{aligned}$$

Example II

- **Model:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
- **Conjugate Prior:** $\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta) \iff \pi(\sigma^2) \propto (1/\sigma^2)^{\alpha+1} e^{-\beta/\sigma^2}$
- **Posterior Distribution:** $\sigma^2 \mid \mathbf{y} \sim \text{Inv-Gamma}\left(\frac{n}{2} + \alpha, \frac{S}{2} + \beta\right)$
- **Bayesian Estimate:**
  - For *variance* σ^2: (MLE is $\hat{\sigma}^2_{\text{ML}} = S/n$)

    $\hat{\sigma}^2_{\text{B}} = E[\sigma^2 \mid \mathbf{y}] = (\frac{S}{2} + \beta)/(\frac{n}{2} + \alpha - 1)$

    $\implies$ MLE-matching prior is $\pi(\sigma^2) \propto 1/\sigma^4$
  - For *precision* $\tau^2 = 1/\sigma^2$: (MLE is $\hat{\tau}^2_{\text{ML}} = n/S$)

    $\tau^2 \mid \mathbf{y} \sim \text{Gamma}(\frac{n}{2} + \alpha, \frac{S}{2} + \beta) \quad \implies \quad \hat{\tau}^2_{\text{B}} = E[\tau^2 \mid \mathbf{y}] = \dfrac{\frac{n}{2} + \alpha}{\frac{S}{2} + \beta}$

    $\implies$ MLE-matching prior is: $\pi(\sigma^2) \propto 1/\sigma^2$

Example II

Summary:

- Bayesian inference **cannot be made invariant** to the choice of prior.
  - **Change-of-Variables Formula:** If $\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$ and $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ is a bijection, then prior on $\boldsymbol{\eta}$ scale is

$$\pi(\boldsymbol{\eta}) = f(g^{-1}(\boldsymbol{\eta})) \times \left| \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}} g^{-1}(\boldsymbol{\eta}) \right|.$$

$\implies$ No "completely uninformative" prior for every parameter transformation, since

$$\pi(\boldsymbol{\theta}) \propto 1 \qquad \implies \qquad \pi(\boldsymbol{\eta}) \propto \left| \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}} g^{-1}(\boldsymbol{\eta}) \right|.$$

Example II

Summary:

- Bayesian inference **cannot be made invariant** to the choice of prior.

  No "completely uninformative" prior for every parameter transformation: if $\boldsymbol{\eta} = g(\boldsymbol{\theta})$, then

  $$\pi(\boldsymbol{\theta}) \propto 1 \qquad \Longrightarrow \qquad \pi(\boldsymbol{\eta}) \propto \left| \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}} g^{-1}(\boldsymbol{\eta}) \right|.$$

- **Folk theorem:** For any choice of prior $\pi(\boldsymbol{\theta})$ and fixed sample size n, there exists some $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ such that $\hat{\boldsymbol{\eta}}_{\mathrm{B}} = E[\boldsymbol{\eta} \mid \mathbf{y}]$ is arbitrarily far from $\hat{\boldsymbol{\eta}}_{\mathrm{ML}}$.

- **Asymptotic theory:** For any choice of prior $\pi(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$, as $n \to \infty$ we have

  $$\boldsymbol{\theta} \mid \mathbf{y} \to \mathcal{N}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \hat{\boldsymbol{\mathcal{I}}}).$$

  $\Longrightarrow$ Bayesian and Frequentist inference are asymptotically equivalent.

Decision Theory

- **Goal:** Compare various estimators $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_k(\mathbf{y})$ of $\boldsymbol{\theta}$.
- **Loss Function:** $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq 0$ and $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0 \iff \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. (Most common one is $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$.)
- **Risk:** Expected loss as a function of true parameter $\boldsymbol{\theta}$:

$$R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = E[L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \int L(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) \cdot p(\mathbf{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\mathbf{y}.$$

- **Admissibility:** $\hat{\boldsymbol{\theta}}_1$ is an inadmissible estimator if exists $\hat{\boldsymbol{\theta}}_2$ such that

$$R(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{\theta}) \preceq R(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{\theta}) \quad \forall \, \boldsymbol{\theta},$$

  i.e., the risk of $\hat{\boldsymbol{\theta}}_2$ is never greater than that of $\hat{\boldsymbol{\theta}}_1$, and for *at least* one value of $\boldsymbol{\theta}$ it is lower. Otherwise, $\hat{\boldsymbol{\theta}}_1$ is admissible, i.e., isn't strictly dominated by another estimator.

Decision Theory

- **Goal:** Compare various estimators $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_k(\mathbf{y})$ of $\boldsymbol{\theta}$.
- **Loss Function:** $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \ge 0$ and $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0 \iff \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.
- **Risk:** $R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = E[L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \int L(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) \cdot p(\mathbf{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\mathbf{y}$.
- **Admissibility:** $\hat{\boldsymbol{\theta}}_1$ is inadmissible if exists $\hat{\boldsymbol{\theta}}_2$ such that $R(\hat{\boldsymbol{\theta}}_2, \boldsymbol{\theta}) \preceq R(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta})$.
- **Bayes Rule:** For given prior $\pi(\boldsymbol{\theta})$, the Bayes rule minimizes the expected loss conditioned on the data:

$$\hat{\boldsymbol{\theta}}_{\mathrm{BR}} = \arg\min_{\tilde{\boldsymbol{\theta}}} E[L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mid \mathbf{y}] = \arg\min_{\tilde{\boldsymbol{\theta}}} \int L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

  - **Point Estimate:** For $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ we have $\hat{\boldsymbol{\theta}}_{\mathrm{BR}} = E[\boldsymbol{\theta} \mid \mathbf{y}]$.
  - **Credible Interval:** For $\tau = g(\boldsymbol{\theta})$ and

$$L(\hat{\tau}, \tau) = (\hat{\tau} - \tau) \cdot (\alpha - \delta\{\hat{\tau} - \tau < 0\}),$$

    we have $\hat{\tau}_{\mathrm{BR}} = F^{-1}_{\tau \mid \mathbf{y}}(\alpha \mid \mathbf{y})$, the α-level quantile of $p(\tau \mid \mathbf{y})$.

Decision Theory

- **Goal:** Compare various estimators $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_k(\mathbf{y})$ of $\boldsymbol{\theta}$.
- **Loss Function:** $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq 0$ and $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0 \iff \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.
- **Risk:** $R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = E[L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \int L(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) \cdot p(\mathbf{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\mathbf{y}$.
- **Admissibility:** $\hat{\boldsymbol{\theta}}_1$ is inadmissible if exists $\hat{\boldsymbol{\theta}}_2$ such that $R(\hat{\boldsymbol{\theta}}_2, \boldsymbol{\theta}) \preceq R(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta})$.
- **Bayes Rule:** $\hat{\boldsymbol{\theta}}_{\mathrm{BR}} = \arg\min_{\tilde{\boldsymbol{\theta}}} E[L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mid \mathbf{y}]$.
- **Theorem:** If $\pi(\boldsymbol{\theta})$ is proper, then $\hat{\boldsymbol{\theta}}_{\mathrm{BR}}$ is admissible. Moreover, any admissible $\hat{\boldsymbol{\theta}}$ is the Bayes rule for some proper or improper prior. (However, not all Bayes rules from improper priors are admissible.)

$\implies$ Only estimators which have a Bayesian interpretation can be admissible.

Bayesian vs. Frequentist?

Some **bad words**:

- Bayesian inference is *subjective*
- Frequentist inference is *ad-hoc*

Don't **be** Bayesian or Frequentist – **use** Bayesian or Frequentist methods depending on the problem.

"Strive for simplicity. Stubbornly resist complexity in your approach."

– *Rob Tibshirani, inventor of LASSO*

Example: When NOT to Use Bayes

- **Model:** $\mathbf{y} = (y_1, \ldots, y_{100}) \overset{\text{iid}}{\sim} F(y)$.
- **Goal:** Estimate $\tau = F^{-1}(.25)$, the 25% quantile of $F(y)$.
- **Frequentist Inference:**
  - *Point Estimate:* $\hat{\tau} = y_{(25)}$, the corresponding order statistic.
  - *Interval Estimate:* For any $F(y)$ and $0 < p < 1$, let $X = \#\{y_i : y_i < F^{-1}(p)\}$. Then $X \sim \text{Binomial}(100, p)$, and

$$\Pr(y_{(a)} < F^{-1}(p) < y_{(b)}) = \sum_{i=a}^{b-1} \binom{100}{i} p^i (1-p)^{100-i}.$$

$\implies$ 95% CI: $(y_{(17)}, y_{(34)})$

- **Bayesian Point/Interval Estimates??**

Example: When to Use Bayes

- **Data:** $K = 8$ schools and their test scores:

School	1	2	3	4	5	6	7	8
x	28	8	-3	7	-1	1	18	12
σ	15	10	16	11	9	11	10	18

- **Goal:** Rank the schools based on μ_i, the "true" score for each school.
- **Parameter Inference:** Consider the following two extremes:

  1. *Individual means*: $\hat{\mu}_i = x_i$.
  2. *Common mean*: $\hat{\mu}_i \equiv \sum_{j=1}^{K} w_j \cdot x_j, \qquad w_j = \sigma_j^{-2} / (\sum_{k=1}^{K} \sigma_k^{-2})$.

     (This is the MLE of model $x_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu, \sigma_i^2)$.)

  Neither are good for ranking (1 has high uncertainty, 2 makes all schools equal).

  A third alternative is to compromise between the two.

Example: When to Use Bayes

- **Data:** $K = 8$ schools and their test scores.
- **Goal:** Rank the schools based on μ_i, the "true" score for each school.
- **Parameter Inference:** Consider the following hierarchical model:

$$x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2).$$

The parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ are called random effects.

  - **Posterior distribution** of μ_i (though nothing Bayesian yet):

$$\mu_i \mid \mathbf{x} \overset{\text{ind}}{\sim} \mathcal{N}\big(B_i\lambda + (1 - B_i)x_i, (1 - B_i)\sigma_i^2\big), \qquad B_i = \sigma_i^{-2}/(\sigma_i^{-2} + \tau^{-2}).$$

  Thus we have the two extremes:

  1. *Individual means*: $\tau = \infty \implies E[\mu_i \mid \mathbf{x}] = x_i$
  2. *Common mean*: $\tau = 0 \implies E[\mu_i \mid \mathbf{x}] = \lambda$

  Moreover, for any $0 < \tau < \infty$ we can compromise between the two (i.e., partial pooling).

Hierarchical Modeling: Frequentist Approach

- **Hierarchical Model:** $x_i \mid \mu_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2).$
- **Marginal Data Distribution:** $x_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\lambda, \sigma_i^2 + \tau^2).$
- **Profile Likelihood:**

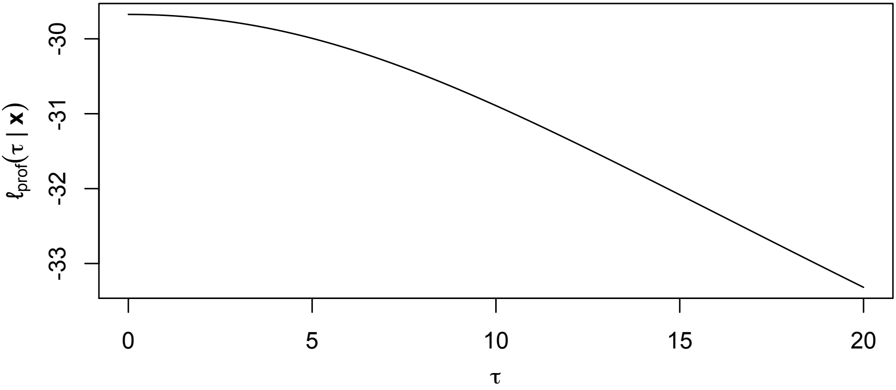
$$\hat{\lambda}_\tau = \arg\max_\lambda \ell(\lambda, \tau \mid \mathbf{x}) = \frac{\sum_{i=1}^K x_i/(\sigma_i^2 + \tau^2)}{\sum_{j=1}^K 1/(\sigma_j^2 + \tau^2)}$$

$$\ell_{\text{prof}}(\tau \mid \mathbf{x}) = \ell(\lambda = \hat{\lambda}_\tau, \tau \mid \mathbf{x}) = -\frac{1}{2}\sum_{i=1}^K \left[\frac{(x_i - \hat{\lambda}_\tau)^2}{\sigma_i^2 + \tau^2} + \log(\sigma_i^2 + \tau^2)\right]$$

$\Longrightarrow$ 2-d optimization reduces to 1-d.

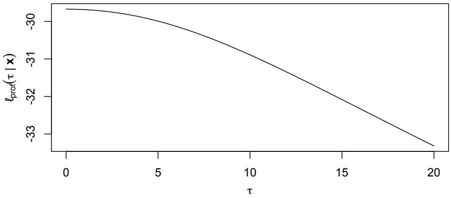
Hierarchical Modeling: Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Profile likelihood:**



Hierarchical Modeling: Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Profile likelihood:**



$\implies \hat{\tau}_{\text{ML}} = 0.$

- **Random-Effects Posterior:**

$$\mu_i \mid \mathbf{x} \overset{\text{ind}}{\sim} \mathcal{N}\big(B_i\lambda + (1 - B_i)x_i, (1 - B_i)\sigma_i^2\big), \quad B_i = \sigma_i^2/(\sigma_i^2 + \tau^2).$$

  - *Naive CI for μ_i:* $[\hat{B}_i\hat{\lambda} + (1 - \hat{B}_i)x_i] \pm 1.96 \times \sigma_i\sqrt{1 - \hat{B}_i}, \qquad \hat{\lambda} = \hat{\lambda}_{\hat{\tau}}$

    $\hat{B}_i = B_i(\hat{\tau}).$

  - Ridiculous CI $\hat{\lambda} \pm 1.96 \times 0$ with plugin $\hat{\tau} = \hat{\tau}_{\text{ML}}$.

Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Random-Effects Posterior:**

$$\mu_i \mid \mathbf{x} \overset{\text{ind}}{\sim} \mathcal{N}\big(B_i\lambda + (1-B_i)x_i, (1-B_i)\sigma_i^2\big), \quad B_i = \sigma_i^2/(\sigma_i^2+\tau^2).$$

  - *Naive CI for μ_i:* $\hat{\lambda} \pm 0$
  - *Bootstrap CI for μ_i:*

1. Generate bootstrap datasets $\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(M)}, \quad \tilde{\mathbf{x}}^{(m)} = (\tilde{x}_1^{(m)}, \ldots, \tilde{x}_K^{(m)})$

Parametric: $\tilde{x}_i^{(m)} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\hat{\lambda}, \hat{\tau}^2)$

Nonparametric: $(\tilde{x}_i^{(m)}, \tilde{\sigma}_i^{(m)})$ resampled from $(x_1, \sigma_1), \ldots (x_K, \sigma_K)$

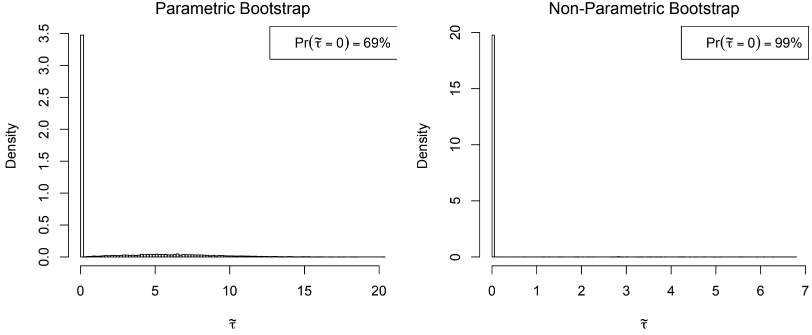
2. Calculate $(\tilde{\lambda}^{(m)}, \tilde{\tau}^{(m)}) = \arg\max \ell(\lambda, \tau \mid \mathbf{x}^{(m)})$ and

$$\tilde{\mu}_i^{(m)} = E[\mu_i \mid \tilde{\mathbf{x}}^{(m)}, \tilde{\lambda}^{(m)}, \tilde{\tau}^{(m)}] = \tilde{B}_i^{(m)}\tilde{\lambda}^{(m)} + (1-\tilde{B}_i^{(m)})\tilde{x}_i^{(m)}$$

3. Basic Bootstrap 95% CI: $(\hat{\mu}_i + \tilde{L}_i, \hat{\mu}_i + \tilde{U}_i)$, where $(\tilde{L}_i, \tilde{U}_i)$ are the the 2.5% and 97.5% sample quantiles of $\tilde{T}_i^{(1)}, \ldots, \tilde{T}_i^{(M)}$, where $\tilde{T}_i^{(m)} = \hat{\mu}_i - \tilde{\mu}_i^{(m)}$.

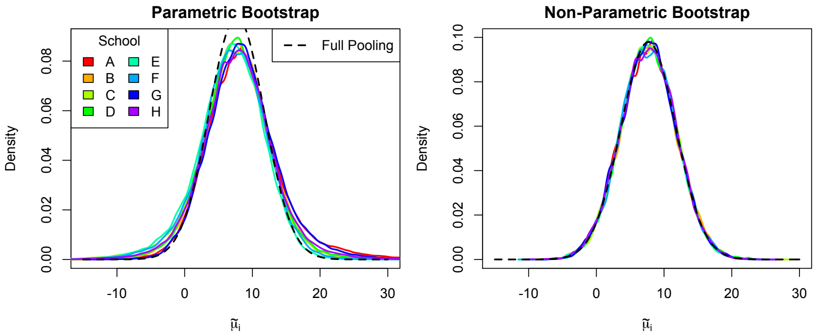
Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Bootstrap distribution of $\tilde{\tau}$:**



Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Bootstrap distribution of $\tilde{\mu}_\mathbf{i}$:**



Frequentist Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Random-Effects Estimate:**
  - Naive, Bootstrap-P, Bootstrap-NP: $\hat{\mu}_i \approx \hat{\lambda}$, i.e., full pooling
  - Penalize $\ell_{\text{prof}}(\tau \mid \mathbf{x})$ away from $\tau = 0$? If so, how? (e.g., **R** package `lme4`)

Bayesian Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Prior:** If $\pi(\lambda, \tau) = \pi(\tau)$, then

$$
\begin{aligned}
\log p(\lambda, \tau \mid \mathbf{x}) &= \ell(\lambda, \tau \mid \mathbf{x}) + \log \pi(\tau) \\
&= -\frac{1}{2} \sum_{i=1}^{K} \left[\frac{(x_i - \lambda)^2}{\sigma_i^2 + \tau^2} + \log(\sigma_i^2 + \tau^2) \right] + \log \pi(\tau) \\
&= -\frac{1}{2} \left[\frac{(\lambda - \lambda_\tau)^2}{\sigma_\tau^2} + \log(\sigma_\tau^2) \right] + \ell_{\text{prof}}(\tau \mid \mathbf{x}) + \log(\sigma_\tau) + \log \pi(\tau),
\end{aligned}
$$

where $\lambda_\tau = \hat{\lambda}_\tau$ (the conditional MLE) and $\sigma_\tau^2 = 1 / \sum_{i=1}^{K} (\sigma_i^2 + \tau^2)^{-1}$.

$$
\implies \qquad \lambda \mid \tau, \mathbf{x} \sim \mathcal{N}(\lambda_\tau, \sigma_\tau^2)
$$

$$
\log p(\tau \mid \mathbf{x}) = \ell_{\text{prof}}(\tau \mid \mathbf{x}) + \log(\sigma_\tau) + \log \pi(\tau)
$$

Bayesian equivalent of profile likelihood is integrating some parameters out

Bayesian Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Posterior:** If $\pi(\lambda, \tau) = \pi(\tau)$,

$$\lambda \mid \tau, \mathbf{x} \sim \mathcal{N}(\lambda_\tau, \sigma_\tau^2)$$
$$\log p(\tau \mid \mathbf{x}) = \ell_{\text{prof}}(\tau \mid \mathbf{x}) + \log(\sigma_\tau) + \log \pi(\tau)$$

- **Possible priors:**

  1. $\pi(\tau) \propto 1$

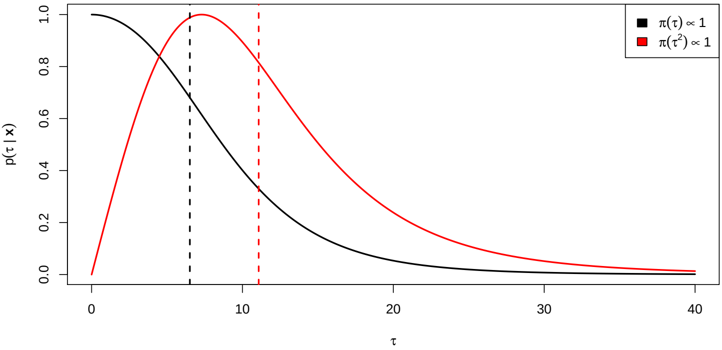
  2. $\pi(\tau^2) \propto 1 \implies \pi(\tau) \propto \tau$

Bayesian Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Posterior:** If $\pi(\lambda, \tau) = \pi(\tau)$,

$$\lambda \mid \tau, \mathbf{x} \sim \mathcal{N}(\lambda_\tau, \sigma_\tau^2)$$

$$\log p(\tau \mid \mathbf{x}) = \ell_{\text{prof}}(\tau \mid \mathbf{x}) + \log(\sigma_\tau) + \log \pi(\tau)$$



Bayesian Approach

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Prior:** $\pi(\lambda, \tau^2) \propto 1$
- **Inference for $\boldsymbol{\mu}$:**

$$p(\boldsymbol{\mu} \mid \mathbf{x}) = \int \underbrace{p(\boldsymbol{\mu} \mid \lambda, \tau, \mathbf{x})}_{\mathcal{N}(\mathbf{B}\lambda + (1-\mathbf{B})\mathbf{x}, (1-\mathbf{B})\boldsymbol{\sigma}^2)} \times \underbrace{p(\lambda \mid \tau, \mathbf{x})}_{\mathcal{N}(\lambda_\tau, \sigma_\tau^2)} \times p(\tau \mid \mathbf{x}) \quad \mathrm{d}\lambda\, \mathrm{d}\tau$$

Monte Carlo method:

1. $\tau^{(m)} \overset{\text{iid}}{\sim} p(\tau \mid \mathbf{x})$ (1-d grid sampling)

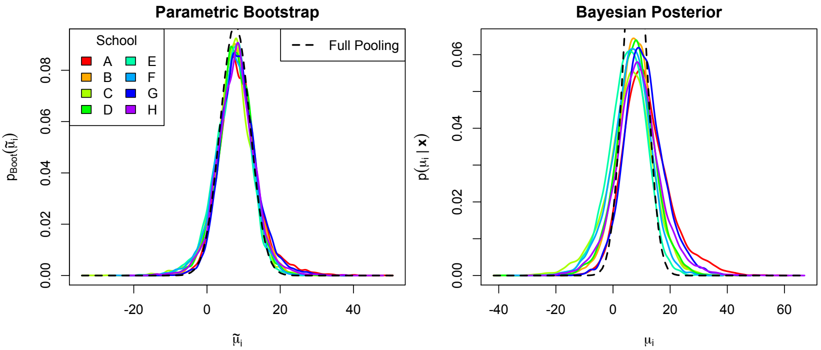
2. $\lambda^{(m)} \mid \boldsymbol{\tau}^{(1:M)} \overset{\text{ind}}{\sim} \mathcal{N}(\lambda_{\tau^{(m)}}, \sigma^2_{\tau^{(m)}})$

3. $\boldsymbol{\mu}^{(m)} \mid \boldsymbol{\lambda}^{(1:M)}, \boldsymbol{\tau}^{(1:M)} \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{B}^{(m)}\lambda^{(m)} + (1 - \mathbf{B}^{(m)})\mathbf{x}, \mathrm{diag}\{(1 - \mathbf{B}^{(m)})\boldsymbol{\sigma}^2\})$

This produces M iid draws from $p(\boldsymbol{\mu}, \lambda, \tau \mid \mathbf{x})$.

Bayesian Approach

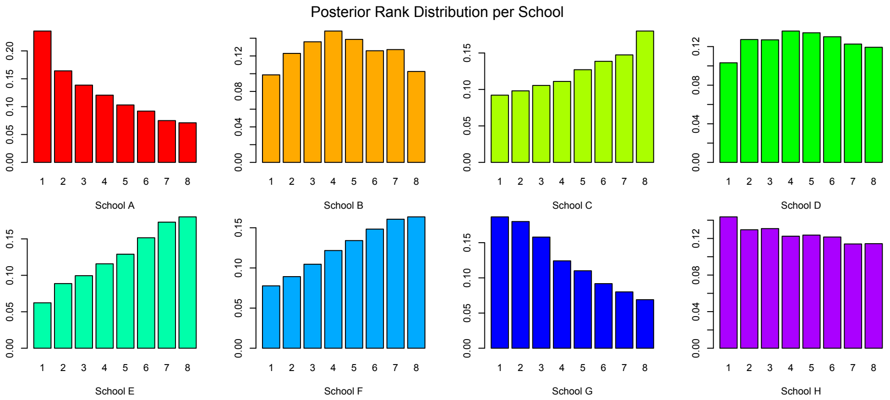
- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Inference on $\boldsymbol{\mu_i}$:**



$\Longrightarrow$ Bayesian inference reports more of a difference between the schools.

Quantity of Interest

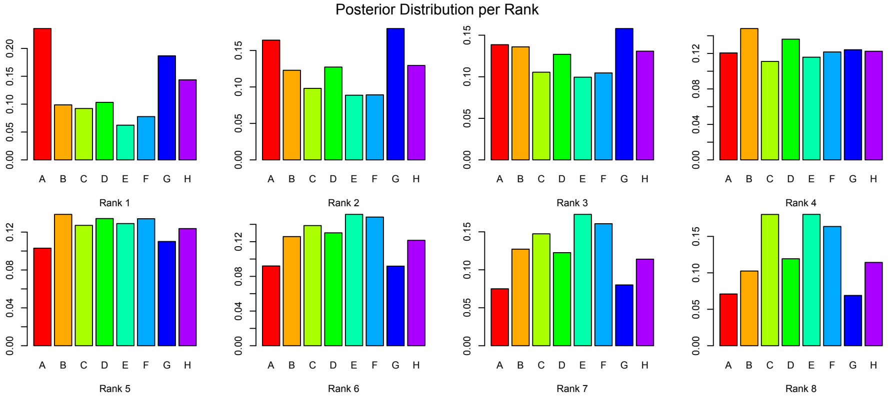
- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Inference on rankings:**



So $\Pr(\text{School A} = \text{Rank } 1 \mid \mathbf{x}) = 25\%$, $\Pr(\text{School A} = \text{Rank } 8 \mid \mathbf{x}) = 8\%$, etc.

Quantity of Interest

- **Hierarchical model:** $x_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \tau^2)$
- **Inference on rankings:**



So $\Pr(\text{Rank } 1 = \text{School A} \mid \mathbf{x}) = 25\%$, $\Pr(\text{Rank } 1 = \text{School E} \mid \mathbf{x}) = 8\%$, etc.