STAT 5380 - Bayesian Data Analysis Notes

Libao Jin

March 27, 2020

Contents

1 Introduction

1.1 Steps of Bayesian Data Analysis

1.1.1 History

Bayesian methods date to the 1763 paper by the Rev. Thomas Bayes, a minister and amateur mathematician. The area generated interest by Laplace, Gauss, and others in the 19th century, but the Bayesian approach was ignored by early 20th century statisticians. During this period, prominent non-statisticians, Harold Jeffreys (a physicist) and Arthur Bowley (an econometrician), continued to lobby on behalf of Bayesian ideas ("inverse probability"). Around 1950, statisticians such as Savage, Bruno de Finetti, Lindley, Kiefer advocated Bayesian methods as remedies for deficiencies in the Frequentist (classical) approach (Carlin and Louis, 2000, p 6).

1.1.2 Animosity

The Bayesian approach to statistical design and analysis is emerging as an increasingly effective and practical alternative to the frequentist one. Indeed, due to computing advances which enable relevant Bayesian analyses, the philosophical battles between Frequentists and Bayesians that were once common at professional statistical meetings are being replaced by a single, more eclectic approach (Carlin and Louis, 2000, p 6).
Rather than argue the foundations of statistics, we prefer to concentrate on the pragmatic advantages of the Bayesian framework, whose flexibility and generality allow it to cope with complex problems. The central feature of Bayesian inference, the direct quantification of uncertainty, means that there is no impediment in principle to fitting models with many parameters and complicated multilayered probability specifications (p 4).

1.2 Probability as a Measure of Uncertainty

- In Bayesian statistics, probability used to measure uncertainty.
- All probabilities are subjective? (p 12) (Berry, 1996, p 35)

Example 1.1 (Coin Toss)**.** Let the (unknown) parameter denote the true probability of obtaining a 'head' upon flipping the coin (p 12).

(a) Frequency Argument - Probability corresponds to the relative frequency obtained in a long sequence of tosses assumed to be performed identically and independently (equally likely).
(b) Subjective Argument - Probability corresponds to the degree of one's belief (coherence of bets).

Define sample space, parameter space, probability distribution.

- Sample space: Let $X = \begin{cases} 1 & \text{head}, \\ 0 & \text{tail}, \end{cases}$ so that $S_X = \{0, 1\}$.
- Let Θ denote the population probability that $X = 1$ (head) so that $S_\Theta = [0, 1]$.
- Then $X|\theta \sim \text{Ber}(\theta)$ and $\Theta \sim \text{Beta}(a, b)$.

1.3 Bayesian Inference

Let θ be a (unknown) parameter, $y = (y_1, y_2, \ldots, y_n)$ be data (evidence).

1.4 Bayesian versus Frequentist

1.4.1 Bayesian

(a) Parameter of population model is a random variable.
(b) A prior distribution used.
(c) Inference conditional on the observed outcome from the experiment.

1.4.2 Frequentist

(a) Parameter of population model is a fixed constant.
(b) No prior distribution.
(c) Inference based upon hypothetical outcomes that could have occurred.

Definition 1.1 (Likelihood)**.** The plausibility of observing θ given data y

$$L(\theta|y) \propto p(y|\theta) \stackrel{\text{iid}}{\longrightarrow} \prod_{i}^{n} p(y_i|\theta) \quad \text{(sampling distribution)}.$$

Definition 1.2 (Prior Distribution)**.** The (subjective) distribution (information) of θ before observing data $y \to p(\theta)$.

Definition 1.3 (Posterior Distribution)**.** The distribution of θ given the data y

$$\begin{aligned}
\text{posterior density} &= \frac{\text{likelihood} \times \text{prior density}}{\text{marginal density}} \quad \text{(inversion)},\\
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)},\\
\pi(\theta|y) &= \frac{f(y|\theta)\pi(\theta)}{m(y)},\\
[\theta|y] &= \frac{[y|\theta][\theta]}{[y]} \quad \text{(Gelfand and Smith Notation, 1990)}.
\end{aligned}$$

Definition 1.4 (Bayesian Likelihood Principle)**.** All statistical inference is based upon the posterior distribution.

Advantages:

(a) Provides a coherent method of combining piror information with data within a solid decision theoretical framework. Do not need Frequentist ”plug-in” principle for functions of parameter.
(b) Statistical inferences are conditional on the data and are exact, without relying on sample size considerations (large or small).
(c) Provides convenient and interpretable statistical inferences.
(d) Convenient setting for a wide range of models through numerical computational methods.

Disadvantages:

(a) Requires prior specification which can be difficult.
(b) Statistical inference can be heavily influenced by subjectivity through the prior that is used.
(c) Computations can be difficult and time consuming. Simulation based methods are often needed, and these results will neither be exact nor reproducible.

Theorem 1.1 (Bayes Theorem 1)**.**

$$P(G|D) = \frac{P(D|G)P(G)}{P(D|G)P(G) + P(D|G^{\complement})P(G^{\complement})}.$$

2 Single Parameter Models

2.1 Bayesian Inference

Θ is a continuous random variable in parameter space $\mathcal{S}_\theta$.

Definition 2.1.

- Bayes Theorem: $p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}$.
- Sample Distribution: $p(\mathbf{y}|\theta)$.
- Likelihood: $L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)$.
- Prior Distribution: $p(\theta)$.
- Prior Predictive Distribution: $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)\, d\theta$.
- Posterior Predictive Distribution: $p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|\theta)p(\theta|\mathbf{y})\, d\theta$, where $\tilde{y}$ is the future value, $p(\tilde{y}|\theta)$ is the sampling distribution, and $p(\theta|\mathbf{y})$ is the posterior distribution.

2.2 Summarizing Posterior Inference

By the Bayesian Likelihood Principle, all inference is to be based upon the posterior distribution.

2.2.1 Bayes Estimate $\hat{\theta}_B(\mathbf{y})$

The Bayesian estimate is a specific numerical summary of *location* with respect to the posterior distribution.

- Frequentist. Estimator: $\hat{\Theta}_{ML}(\mathbf{Y})$; Estimator: $\theta_{ML}(\mathbf{y})$.
- Bayesian.
  - $\text{mean}(\Theta|\mathbf{y}) = E(\Theta|\mathbf{y}) = \int \theta p(\theta|\mathbf{y})\, d\theta$.
  - $\text{median}(\Theta|\mathbf{y}) = \theta_{0.5}(\mathbf{y})$ where $\int_{-\infty}^{\theta_{0.5}(\mathbf{y})} p(\theta|\mathbf{y})\, d\theta = 0.5$.
  - $\text{mode}(\Theta|\mathbf{y}) = \theta_*(\mathbf{y})$ where $p(\theta_*(\mathbf{y})|\mathbf{y}) = \max_\theta p(\theta|\mathbf{y})$.

2.2.2 Bayes Uncertainty $u_B(\mathbf{y})$

The Bayesian measure of uncertainty is a specific numerical summary of *variability* for the posterior distribution.

- Frequentist: Standard error is the standard deviation of the estimator, that is,

$$\text{se}(\hat{\theta}) = \text{sd}(\hat{\Theta}) = \text{sd}(\hat{\Theta}(\mathbf{y})).$$

- Bayesian:

$$\begin{aligned} \text{sd}(\Theta|\mathbf{y}) &= \text{Var}(\Theta|\mathbf{y})^{1/2} = [E(\Theta^2|\mathbf{y}) - E(\Theta|\mathbf{y})^2]^{1/2}, \\ iqr(\Theta|\mathbf{y}) &= \theta_{0.75} - \theta_{0.25} \quad \text{(interquartile range)} \end{aligned}$$

2.2.3 Credible Interval $B_{1-\alpha}(\mathbf{y})$

Posterior probability that Θ is in $B_{1-\alpha}(\mathbf{y})$ is $1-\alpha$. Credible Interval (CI) $100(1-\alpha)\%$ of the values of Θ given $\mathbf{y}$.

- Frequentist: $1-\alpha$ confidence interval: $\hat{\Theta}_{ML} \pm \text{multiplier} \times \text{se}(\hat{\theta})$, e.g., the multiplier can be $z_{1-\alpha/2}$.
- Bayesian: $B_{1-\alpha}(\mathbf{y}) = [L(\mathbf{y}), U(\mathbf{y})]$ is *any* interval satisfying

$$1-\alpha = P(L(\mathbf{y}) \le \Theta \le U(\mathbf{y})|\mathbf{y}) = \int_{L(\mathbf{y})}^{U(\mathbf{y})} p(\theta|\mathbf{y})\, d\theta.$$

  - Equal Tails Percentile Interval. $B_{1-\alpha}(\mathbf{y}) = [\theta_{\alpha/2}(\mathbf{y}), \theta_{1-\alpha/2}(\mathbf{y})]$.
  - Highest Posterior Density (HPD) Set. $B_{1-\alpha}(\mathbf{y}) = \{\theta : p(\theta|\mathbf{y} > \lambda)\}$ with λ as $P(\Theta \in B_{1-\alpha}(\mathbf{y})|\mathbf{y}) = \int_{\{\theta: p(\theta|\mathbf{y}) > \lambda} p(\theta|\mathbf{y})\, d\theta = 1-\alpha$.
  - HPD CI consists of values of Θ with *highest posterior density*. HPD set is not an interval unless posterior PDF is unimodal.

2.2.4 Posterior Probability $p_B^0(\mathbf{y})$

Bayesian "p-value" is posterior probability of H_0.

- Frequentist: p-value $= p(W(\mathbf{Y}) \ge W(\mathbf{y})|\theta \in H_0)$, where H_0 is true, where $S_\theta = H_0 \sqcup H_1$.
- Bayesian: $p_B^0(\mathbf{y}) = P(\Theta \in H_0|\mathbf{y}) = \int_{\theta \in H_0} p(\theta|\mathbf{y})\, d\theta = 1 - P(\Theta \in H_1|\mathbf{y})$.

Table 1: Approximate Converted Jeffrey Measure of Evidence

	slight	positive	substantial	decisive
evidence against	$[0.25, 0.50)$	$[0.10, 0.25)$	$[0.01, 0.10)$	$[0.00, 0.01)$

2.2.5 Bayesian Interpretation

Example 2.1. Assume the population of weights of high school students is normally distributed with unknown mean and known standard deviation $\sigma = 20$ pounds. A random sample of n students from this population has sample mean 150 pounds. Suppose the prior distribution for Θ is normal with mean $\mu = 180$ and standard deviation $\tau = 40$.

- Estimation. The Bayes estimate of the population mean weight in pounds Θ of these students is 150.73 ($n = 10$) which corresponds to the posterior mean (mean = median = mode since the posterior distribution is normal).
- Uncertainty. The uncertainty on the population mean weight in pounds Θ of these students is 6.25 ($n = 10$) which corresponds to the posterior standard deviation. Use IQR, use `qnorm`.
- Interval. The 95% of the population mean weight in pounds Θ fall within $[138.49, 162.98]$ ($n = 10$) based upon the posterior distribution. (This credible interval is the equal tails interval and the HPD interval since the posterior distribution is normal i.e., symmetric and unimodal).
- Bayesian p-value. The Bayesian p-value for the population mean weight in pounds Θ being less than Θ_0 pounds is $qnorm(\theta_0, \mu_*, \sigma_*)$ which corresponds to the posterior probability of being less than θ_0. $H_0 : \Theta \le \theta_0$, $H_1 : \Theta > \theta_0$.

2.2.6 Weakly Informative Prior Distribution

- Prior distribution is key to Bayesian inference. Determination of prior is most important and most difficult step.
- Prior distribution is a tool for *approximating* information on Θ.
- Population Interpretation - prior is the population of values for Θ before observing data.
- Subjective Knowledge - express personal knowledge, uncertainty.
- Use plots and summaries to match prior to info.
- Conjugate Family. A class of prior distributions $\mathcal{P}$ is a conjugate family for a family of sampling distributions $\mathcal{F}$ provided the posterior distribution is a member of $\mathcal{P}$ for all $p \in \mathcal{F}$.

$$p(\theta|\mathbf{y}) \in \mathcal{P}, \forall p(\mathbf{y}|\theta) \in \mathcal{F}, p(\theta) \in \mathcal{P}, \theta \in S_\theta, \mathbf{y} \in \mathcal{S}_{\mathbf{y}}.$$

  - Calculations for conjugate (convenience) priors ten to be a much easier problem as we only need to update parameters in the family $\mathcal{P}$.

$$p(\theta|\boldsymbol{\phi}) \to p(\theta|\mathbf{y}, \boldsymbol{\phi}_*), p \in \mathcal{P},$$

    where $p(\theta|\boldsymbol{\phi})$ is prior, $p(\theta|\mathbf{y}, \boldsymbol{\phi}_*)$ is posterior, update $\boldsymbol{\phi}$ to $\boldsymbol{\phi}_*$.
  - Single Parameter Conjugate Families.

Table 2: Single Parameter Conjugate Families

parameter	sampling dist. $f(\mathbf{y}\|\theta)$	prior dist. $p(\theta)$	posterior dist. $p(\theta\|\mathbf{y})$	ref
mean	Normal(θ, σ^2)	Normal(μ, τ^2)	Normal(μ_*, τ_*^2)	class, p40
rate	Gamma(a, θ)	Gamma(a, b)	Gamma(a_*, b_*)	hw1
mean	Poisson(θ)	Gamma(a, b)	Gamma(a_*, b_*)	p44 / ch7, 5660
probability	Binomial(n, θ)	Beta(a, b)	Beta(a_*, b_*)	p35 / hw5, 5660
probability	NegBinomial(r, θ)	Beta(a, b)	Beta(a_*, b_*)	p35
precision	Normal$(\mu, 1/\theta)$	Gamma(a, b)	Gamma(a_*, b_*)	

- The *influence* of the prior distribution can be negligible moderate, or enormous, that is,, amount of information the prior has about θ.
- Non-formative (Reference, Vague, Flat, Diffuse, Objective). A prior distribution that has minimal role in posterior distribution.
- Sensitivity Analysis - assess influence of prior choice on posterior.
- Improper Prior. $\int p(\theta)\, d\theta = \infty \to \int p(\theta|\mathbf{y})\, d\theta \overset{?}{<} \infty$.
- Invariant Prior. Location $p(\theta) \propto 1$, scale $p(\theta) \propto \frac{1}{\theta}$ (p.54).
- Jeffery Prior. $p(\theta) \propto [J^{\mathbf{y}|\theta}(\theta)]^{1/2} = [-E(l''(\theta|\mathbf{Y}))]^{1/2}$.
- Weakly Information. Proper prior with weak information.

3 Multi-parameter Models

3.1 Multiple Parameters

- Consider $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$ where Θ_1 is the parameter of *interest* and Θ_2 could be a *nuisance* parameter.
  - Joint posterior PDF: $p(\boldsymbol{\theta}|\mathbf{y}) = p[(\theta_1, \theta_2|\mathbf{y})] \propto p(\mathbf{y}|(\theta_1, \theta_2))p(\theta_1, \theta_2)$.

Table 3: caption

type	roman (scalar, vector, matrix)	Greek (scalar, vector, matrix)
random	X, $\mathbf{X}$, $\mathbb{X}$	Θ, $\mathbf{\Theta}$, $\overline{\mathbf{\Theta}}$
fixed	x, $\mathbf{x}$, X	θ, $\boldsymbol{\theta}$, Θ

- Marginal posterior PDF: $p(\theta_1|\mathbf{y}) = \int p[(\theta_1, \theta_2)|\mathbf{y}]\, d\theta_2 = \int p(\theta_1|\theta_2, \mathbf{y})p(\theta_2, \mathbf{y})/p(\mathbf{y})\, d\theta_2 = \int p(\theta_1|\theta_2, \mathbf{y})p(\theta_2|\mathbf{y})\, d\theta_2$.

- Require the *joint* posterior distribution of all unknowns.
- Then integrate over unknowns that are not of interest (nuisance parameters) to obtain the *marginal* posterior distributions.
- Direct Simulation requisite posterior known.

(a) $p[(\theta_1, \theta_2)|\mathbf{y}]$ known: draw samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$ from joint posterior distribution.
(b) $p(\theta_1|\theta_2, \mathbf{y})$, $p(\theta_2|\mathbf{y})$ known: draw θ_{2j} from marginal posterior distribution, and then draw θ_{1j} from conditional posterior distribution of θ_1 given θ_2.

- Compute summaries on parameters of interest $\theta_{11}, \theta_{12}, \ldots, \theta_{1B}$ and ignore draws from the nuisance parameters $\theta_{21}, \theta_{22}, \ldots, \theta_{2B}$.
- Predictive posterior simulated by obtaining $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_B$ where $\tilde{y}_j$ is computed from the sampling distribution conditional on $\boldsymbol{\theta}_j$.

Example 3.1 (Cork Data)**.** Rao (1948) present data consisting of weights of bark deposits (cg) obtained from cork borings (holes drilled in the dark) in four directions (N, E, S, W) for a single sample of 28 trees. Interpretation:

- Interval. The values for the population mean of the average of the cork deposits are between 41.85 and 53.93 cg with posterior probability 0.95.

3.1.1 Likelihood

Textbook: p.84 Given that the likelihood

$$L((\mu, \phi)|\overline{x}, s^2) = \phi^{-n/2} \exp(-\frac{1}{2\phi}\sum_{i=1}^{n}(x_i - \mu)^2) \implies l((\mu, \phi)|\overline{x}, s^2) = -\frac{n}{2}\ln\phi - \frac{1}{2\phi}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then the partial derivatives of $l((\mu,\phi)|\overline{x},s^2)$ with respect to μ and ϕ can be obtained as below

$$\begin{aligned}
\frac{\partial l((\mu,\phi)|\overline{x},s^2)}{\partial\mu} &= \frac{1}{\phi}\sum_{i=1}^{n}(x_i-\mu)\\
\frac{\partial^2 l((\mu,\phi)|\overline{x},s^2)}{\partial\mu^2} &= -\frac{n}{\phi}\\
\frac{\partial^2 l((\mu,\phi)|\overline{x},s^2)}{\partial\phi\partial\mu} &= -\frac{1}{\phi^2}\sum_{i=1}^{n}(x_i-\mu)\\
\frac{\partial l((\mu,\phi)|\overline{x},s^2)}{\partial\phi} &= -\frac{n}{2\phi}+\frac{1}{2\phi^2}\sum_{i=1}^{n}(x_i-\mu)^2\\
\frac{\partial^2 l((\mu,\phi)|\overline{x},s^2)}{\partial\phi^2} &= \frac{n}{2\phi^2}-\frac{1}{\phi^3}\sum_{i=1}^{n}(x_i-\mu)^2\\
\frac{\partial^2 l((\mu,\phi)|\overline{x},s^2)}{\partial\mu\partial\phi} &= -\frac{1}{\phi^2}\sum_{i=1}^{n}(x_i-\mu).
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
I^{(\overline{x},s^2)|\mu,\phi}(\mu,\phi) &= -\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ln p(\overline{x},s^2|\boldsymbol{\theta})\right\}\\
&= -\begin{bmatrix}\partial l((\mu,\phi)|\overline{x},s^2)/\partial\mu^2 & \partial l((\mu,\phi)|\overline{x},s^2)/\partial\phi\partial\mu\\ \partial l((\mu,\phi)|\overline{x},s^2)/\partial\mu\partial\phi & \partial l((\mu,\phi)|\overline{x},s^2)/\partial\phi^2\end{bmatrix}\\
&= \begin{bmatrix} n/\phi & \frac{1}{\phi^2}\sum_{i=1}^{n}(x_i-\mu)\\ \frac{1}{\phi^2}\sum_{i=1}^{n}(x_i-\mu) & -\frac{n}{2\phi^2}+\frac{1}{\phi^3}\sum_{i=1}^{n}(x_i-\mu)^2\end{bmatrix}\\
&= \begin{bmatrix} n/\phi & 0\\ 0 & \frac{n}{2\phi^2}-\frac{n\phi}{\phi^3}\end{bmatrix}\\
&= \begin{bmatrix} n/\phi & 0\\ 0 & -\frac{n}{2\phi^2}\end{bmatrix}
\end{aligned}$$

3.2 Normal Model with Non-informative Prior

Let X_i denote the bark deposit for tree i averaged over the four directions. Assume $X_i|\mu,\phi \,\text{iid}\, \mathcal{N}(\mu,\phi)$ for $i=1,2,\ldots,n$ where $\boldsymbol{\Theta}=(\mathcal{M},\Phi)$, $\Phi=\Sigma^2$. Then

$$p(x_i|\mu,\phi)=\frac{1}{\sqrt{2\pi\phi}}\exp\left[-\frac{1}{2\phi}(x_i-\mu)^2\right].$$

(a) Obtain the likelihood in terms of the sufficient statistic.

$$\begin{aligned}
L(\boldsymbol{\theta}|\mathbf{x}) &= L[(\mu,\phi)|\mathbf{x}]\\
&= \phi^{-n/2}\exp\left[-\frac{1}{2\phi}\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2\right]\\
&= \phi^{-n/2}\exp\left\{-\frac{1}{2\phi}[(n-1)s^2+n(\overline{x}-\mu)^2]\right\}\\
&= L[(\mu,\phi)|\overline{x},s^2],
\end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ and $2\sum_{i=1}^n (x_i - \overline{x})(\overline{x} - \mu) = 0$.

(b) Assume a vague prior $p(\mu, \phi) = p(\mu|\phi)p(\phi) \propto (1)\left(\frac{1}{\phi}\right)$.

(c) Show $(\mathcal{M}, \Phi)|\overline{x}, s^2 \sim \mathcal{N}(\overline{x}, \Phi/n) \times \text{InvGamma}(\frac{n-1}{2}, \frac{n-1}{2}s^2)$.

$$\begin{aligned}
p(\mu, \phi|\overline{x}, s^2) &\propto L(\mu, \phi|\overline{x}, s^2)p(\mu, \phi) \\
&\propto \phi^{-n/2-1} \exp\left\{-\frac{1}{2\phi}[(n-1)s^2 + n(\overline{x} - \mu)^2]\right\} \\
&= \phi^{-1/2} \exp\left[-\frac{n}{2\phi}(\overline{x} - \mu)^2\right] \phi^{-n/2-1/2} \exp\left[-\frac{1}{2\phi}(n-1)s^2\right] \\
&= \left\{\frac{1}{\phi^{1/2}} \exp\left[-\frac{1}{2\phi/n}(\overline{x} - \mu)^2\right]\right\} \left\{\frac{1}{\phi^{(n/2-1/2)+1}} \exp\left[-\frac{(n-1)}{2\phi}s^2\right]\right\}. \\
&\propto p(\mu|\phi, \overline{x}, s^2)p(\phi|\overline{x}, s^2),
\end{aligned}$$

where $p(\mu|\phi, \overline{x}, s^2)$ is the density function of $\mathcal{N}(\overline{x}, \phi/n)$ and $p(\phi|\overline{x}, s^2)$ is the density function of $\text{InvGamma}[(n-1)/2, (n-1)s^2/2]$. Note that the density function $p(\theta)$ of inverse Gamma is:

$$p(\theta) = \text{InvGamma}(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}, \theta > 0.$$

(d) Show that $\mathcal{M}|\overline{x}, s^2 \sim t_{n-1}(\overline{x}, s^2/n)$.

$$\begin{aligned}
p(\mu|\overline{x}, s^2) &= \int_0^\infty p(\mu, \phi|\overline{x}, s^2)\, d\phi \\
&\propto \int_0^\infty \phi^{-n/2-1} \exp\left\{-\frac{1}{2\phi}[(n-1)s^2 + n(\overline{x} - \mu)^2]\right\} d\phi \\
&= \int_0^\infty \phi^{-(n/2+1)} \exp\left\{-\frac{[(n-1)s^2 + n(\overline{x} - \mu)^2]}{2}\frac{1}{\phi}\right\} d\phi \\
&= \int_0^\infty \phi^{-(\alpha+1)}e^{-\beta/\phi}\, d\phi \\
&= \frac{\Gamma(\alpha)}{\beta^\alpha}\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}\phi^{-(\alpha+1)}e^{-\beta/\phi}\, d\phi \\
&= \Gamma(\alpha)\beta^{-\alpha} \cdot 1 \\
&= \Gamma(\alpha)\left\{\frac{[(n-1)s^2 + n(\overline{x} - \mu)^2]}{2}\right\}^{-n/2} \\
&\propto [(n-1)s^2 + n(\overline{x} - \mu)^2]^{-n/2} \\
&\propto \left[1 + \frac{n(\overline{x} - \mu)^2}{(n-1)s^2}\right]^{-(n-1+1)/2} \\
&\propto \left[1 + \frac{1}{n-1}\frac{(\overline{x} - \mu)^2}{s^2/n}\right]^{-(n-1+1)/2} \\
&\propto \left[1 + \frac{1}{\nu}\frac{(\overline{x} - \mu)^2}{s^2/n}\right]^{-(\nu+1)/2}
\end{aligned}$$

where $\alpha = n/2$, $\beta = [(n-1)s^2 + n(\overline{x} - \mu)^2]/2$, $\nu = n - 1$. Note that the density function of

$\theta \sim t_\nu(\mu, \sigma^2)$ is

$$p(\theta) = t_\nu(\theta|\mu, \sigma^2) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma}} \left[1 + \frac{1}{\nu}\left(\frac{\theta-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2}.$$

That implies

$$\mathcal{M}|\overline{x}, s^2 \sim t_{n-1}(\overline{x}, s^2/n) \implies \frac{\mathcal{M}|\overline{x}, s^2 - \overline{x}}{s/\sqrt{n}} \sim t_{n-1}(0,1).$$

4 Multivariate Normal Model with Non-informative Prior

Let Y_{ij} denote bark deposit for tree i and direction j. Then $\mathbf{Y}_i|\boldsymbol{\mu}, \Sigma$ i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma), i = 1, 2, \ldots, n$ where $\boldsymbol{\Theta} = (\mathcal{M}, \Sigma)$. Then

$$p(\mathbf{y}_i|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} \det \Sigma^{-1/2} \exp\left[-\frac{(\mathbf{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\mu})}{2}\right].$$

(a) Obtain the likelihood in terms of the sufficient statistic.
(b) Assume vague prior $p(\boldsymbol{\mu}, \Sigma) = p(\boldsymbol{\mu}|\Sigma)p(\Sigma) \propto |\Sigma|^{-(p+1)/2}$.
(c) Show $\mathcal{M}, \Sigma|overline\mathbf{y}, T \sim \mathcal{N}_p(\overline{\mathbf{y}}, \Sigma/n) \cdot \text{InvWishart}_{n-1}(T^{-1})$, where $T = (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T$.
(d) Show that $\mathcal{M}|\overline{\mathbf{y}, T} \sim t_{n-p}(\overline{\mathbf{y}}, T/[n(n-p)])$.

4.1 Bayesian Asymptotics

- Observed Information: $I^{\mathbf{y}|\theta}(\theta) = -\frac{d^2}{d\theta^2} \ln p(\mathbf{y}|\theta)$.
- Fisher Information: $J^{\mathbf{y}|\theta}(\theta) = -E\left[\frac{d^2}{d\theta^2} \ln p(\mathbf{Y}|\theta)\right]$.

4.1.1 Frequency Evaluations

Asymptotic Normality of MLE. Assume X_i iid for $i = 1, 2, \ldots, n$ where n is large. If MLE exists, then $\hat{\Theta}_{\text{ML}}|\theta \sim \mathcal{N}(\theta, [J^{\mathbf{y}|\theta}(\theta)]^{-1})$.

4.1.2 Normal Approximation to Posterior Distribution

Asymptotic Normality of Posterior Mode. Assume X_i iid for $i = 1, 2, \ldots, n$ where n is large. Then $\Theta|\mathbf{y} \sim \mathcal{N}(\tilde{\theta}, [I^{\theta|\mathbf{y}}(\tilde{\theta})]^{-1})$.

$$\begin{aligned}
p(\theta|\mathbf{y}) &= \exp[\ln p(\theta|\mathbf{y})] \\
&\approx \exp\left[\ln p(\tilde{\theta}|\mathbf{y}) + (\theta - \tilde{\theta})\frac{d \ln p(\theta|\mathbf{y})}{d\theta}\bigg|_{\theta=\tilde{\theta}} + \frac{(\theta-\tilde{\theta})^2}{2}\frac{d^2 \ln p(\theta|\mathbf{y})}{d\theta^2}\bigg|_{\theta=\tilde{\theta}}\right]. \\
&\propto \exp\left[\frac{(\theta-\tilde{\theta})^2}{2}\frac{d^2 \ln p(\theta|\mathbf{y})}{d\theta^2}\bigg|_{\theta=\tilde{\theta}}\right] \\
&= \exp\left[-\frac{(\theta-\tilde{\theta})^2}{2} I^{\theta|\mathbf{y}}(\tilde{\theta})\right].
\end{aligned}$$

Bayesian CLT. Assume X_i iid for $i = 1, 2, \ldots, n$ and large n. Then $\Theta|\mathbf{y} \sim \mathcal{N}(\hat{\theta}, [J^{\mathbf{y}|\theta}(\hat{\theta})]^{-1})$ where $\hat{\theta} = \hat{\theta}_{\text{ML}}$ exists, $p(\theta) \approx 1$. Note $\tilde{\theta} = \hat{\theta}$ when $p(\theta) \approx 1$. Note

$$
\begin{aligned}
I^{\theta|\mathbf{y}}(\theta) &\propto -\frac{d^2}{d\theta^2} \ln\left[p(\mathbf{y}|\theta)p(\theta)\right] \\
&= -\frac{d^2}{d\theta^2} \ln\left[p(\theta)\prod_{i=1}^n p(y_i|\theta)\right] \\
&= -\frac{d^2}{d\theta^2}\left[\sum_{i=1}^n \ln p(y_i|\theta) + \ln p(\theta)\right] \\
&= -\frac{d^2}{d\theta^2}\left[\sum_{i=1}^n \ln p(y_i|\theta) + \ln p(\theta)\right] \\
&= -\frac{d^2}{d\theta^2}\left[n\frac{1}{n}\sum_{i=1}^n \ln p(y_i|\theta) + \ln p(\theta)\right] \\
&\approx nE[I^{x_i|\theta}(\theta)] + 0 \\
&= nJ^{x_i|\theta}(\theta) \\
&= J^{\mathbf{y}|\theta}(\theta).
\end{aligned}
$$

- Can be extended to $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, but approximations more accurate for conditional, marginal distribution.
- Prior distribution plays little role for large n.
- Approximate distributions lead to same summaries in large samples for the Frequentist (ML), Bayesian approaches.

4.1.3 Implicit of Bayesian CLT

(a) Bayesian inferential results the "same" as for maximum likelihood.
(b) Prior distribution plays little role.
(c) Can be a convenient tool for checking results (even when n is not very large).

5 Hierarchical Model

Example 5.1. Intelligence test scores recorded for a particular child over 7 years given by $105, 127, 115, 130, 110, 135$ Interest is the true mean IQ at year 7. The variance of IQ scores is assumed to be 100 for each year, and the variance of the true mean IQ score for each year is assumed to be 225. The IQ test is adjusted for age.

- Y_i is the IQ at year i for $i = 1, 2, \ldots, n (= 7)$ (1 observation per group).
- $\{\theta_i\}$ is the true mean IQ for year i (heterogeneity across years).
- σ^2 is the varianceof IQ for year i (homogeneity across years), $\sigma^2 = 100$ in our case.
- μ is the true mean IQ of child for population of all tests.
- τ^2 is the variance of IQ of child for population of all tests.
- Θ_i (population mean IQ for year i) is assumed to vary from year i to year i' because of changes to the individual or changes in the test (Berger, 1985).

Sample distribution: $Y_i|\theta_i \overset{iid}{\sim} \mathcal{N}(\theta_i, \sigma^2)$.

(a) Stage 1: $\Theta_i|\mu \overset{iid}{\sim} \mathcal{N}(\mu, \tau^2)$ for $i = 1, 2, \ldots, n$.
(b) Stage 2: $\mathcal{M} \sim 1$ (vague improper prior)

(a) Show that $Y_i|\mu \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$. Identify the distribution of $\overline{Y}|\mu$.

$$
\begin{aligned}
p(y_i|\mu) &= \int_{-\infty}^{\infty} p(y_i|\theta_i)p(\theta_i|\mu)\, d\theta_i \\
&\propto \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(y_i-\theta_i)^2 - \frac{1}{2\tau^2}(\theta_i-\mu)^2\right] d\theta_i \\
&= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(y_i^2+\theta_i^2-2y_i\theta_i) - \frac{1}{2\tau^2}(\theta_i^2+\mu^2-2\theta_i\mu)\right] d\theta_i \\
&= \exp\left(-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(\theta_i^2-2y_i\theta_i) - \frac{1}{2\tau^2}(\theta_i^2-2\theta_i\mu)\right] d\theta_i \\
&= \exp\left(-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty} \exp\left[-\frac{\tau^2}{2\sigma^2\tau^2}(\theta_i^2-2y_i\theta_i) - \frac{\sigma^2}{2\sigma^2\tau^2}(\theta_i^2-2\theta_i\mu)\right] d\theta_i \\
&= \exp\left(-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[\tau^2(\theta_i^2-2y_i\theta_i) + \sigma^2(\theta_i^2-2\theta_i\mu)\right]\right\} d\theta_i \\
&= \exp\left(-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[(\tau^2+\sigma^2)\theta_i^2 - 2(\tau^2 y_i+\sigma^2\mu)\theta_i\right]\right\} d\theta_i \\
&= \exp\left(-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\theta_i^2 - 2\frac{\tau^2 y_i+\sigma^2\mu}{\tau^2+\sigma^2}\theta_i\right]\right\} d\theta_i \\
&= \exp\left[-\frac{y_i^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2} + \frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\frac{(\tau^2 y_i+\sigma^2\mu)^2}{(\tau^2+\sigma^2)^2}\right]\int_{-\infty}^{\infty} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left(\theta_i - \frac{\tau^2 y_i+\sigma^2\mu}{\tau^2+\sigma^2}\right)^2\right\} d\theta_i \\
&\propto \exp\left[-\left(\frac{y_i^2\tau^2+\mu^2\sigma^2}{2\sigma^2\tau^2}\right) + \frac{1}{2\sigma^2\tau^2}\frac{(\tau^2 y_i+\sigma^2\mu)^2}{\tau^2+\sigma^2}\right] \\
&= \exp\left[-\frac{(y_i^2\tau^2+\mu^2\sigma^2)(\tau^2+\sigma^2)}{2\sigma^2\tau^2(\tau^2+\sigma^2)} + \frac{(\tau^2 y_i)^2+(\sigma^2\mu)^2+2\tau^2 y_i\sigma^2\mu}{2\sigma^2\tau^2(\tau^2+\sigma^2)}\right] \\
&= \exp\left[-\frac{y_i^2\tau^2\sigma^2+\mu^2\sigma^2\tau^2-2\sigma^2\tau^2 y_i\mu}{2\sigma^2\tau^2(\tau^2+\sigma^2)}\right] \\
&= \exp\left[-\frac{(y_i-\mu)^2}{2(\tau^2+\sigma^2)}\right].
\end{aligned}
$$

(b) Show that $\mathcal{M}|\overline{y} \sim \mathcal{N}(\overline{y}, \sigma_\mu^2) = \mathcal{N}[\overline{y}, (\sigma^2+\tau^2)/n]$. What is the mode?

$$
\begin{aligned}
p(\mu|\overline{y}) &\propto p(\overline{y}|\mu)p(\mu) \\
&\propto \exp\left[-\frac{(\overline{y}-\mu)^2}{2(\tau^2+\sigma^2)/n}\right] \times [1] \\
&\implies \mathcal{M}|\overline{y} \sim \mathcal{N}[\overline{y}, (\sigma^2+\tau^2)/n] \\
&\implies \tilde{y} = \text{mode}(\mathcal{M}|\overline{y}) = \overline{y}.
\end{aligned}
$$

(c) (a) Justify that $\Theta_i|y_i,\mu \sim \mathcal{N}(\mu_*, \sigma_*^2) = \mathcal{N}\left(\frac{\tau^2 y_i+\sigma^2\mu}{\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right)$.

$$\begin{aligned}
p(\theta_i|y_i,\mu) &\propto p(y_i|\theta_i,\mu)p(\theta_i|\mu)\\
&= p(y_i|\theta_i)p(\theta_i|\mu)\\
&\propto \exp\left[-\frac{(y_i-\theta_i)^2}{2\sigma^2} - \frac{(\theta_i-\mu)^2}{2\tau^2}\right]\\
\implies \Theta_i|y_i,\mu &\sim \mathcal{N}(\mu_*, \sigma_*^2) = \mathcal{N}\left(\frac{\tau^2 y_i+\sigma^2\mu}{\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right).
\end{aligned}$$

(b) Apply the Empirical Bayes approach to approximate the posterior distribution.

$$\Theta_i|y_i,\tilde{\mu} \sim \mathcal{N}\left(\frac{\tau^2 y_i+\sigma^2\tilde{\mu}}{\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right) \sim \mathcal{N}\left(\frac{\tau^2 y_i+\sigma^2\overline{y}}{\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right).$$

(d) Show $\Theta_i|\mathbf{y} \sim \mathcal{N}(\overline{\theta}_i, v_i)$ where $\overline{\theta}_i = \frac{\tau^2 y_i+\sigma^2\overline{y}}{\tau^2+\sigma^2}$, $v_i = \frac{\sigma^2(\sigma^2/n+\tau^2)}{\tau^2+\sigma^2}$.

$$\begin{aligned}
p(\theta_i|\mathbf{y}) &= \int p(\theta_i|y_i,\mu)p(\mu|\overline{y})\,d\mu\\
&\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_*^2}(\theta_i-\mu_*)^2 - \frac{1}{2\sigma_\mu^2}(\mu-\overline{y})^2}\,d\mu\\
&\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma_*^2}(\theta_i-\mu_*)^2 - \frac{1}{2\sigma_\mu^2}\left[\frac{(\tau^2+\sigma^2)\mu_* - \tau^2 y_i}{\sigma^2} - \overline{y}\right]^2\right\}d\mu_*
\end{aligned}$$

Then

$$\Theta_i|\mathbf{y} \sim \mathcal{N}(\mu/\phi,)$$

$$v_i = = \frac{\sigma^4/n + \sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Note that $\overline{\theta}_i = \frac{\tau^2 y_i+\sigma^2\overline{y}}{\tau^2+\sigma^2}$ is the posterior mean, median, mode of normal.

$$\overline{\theta}_i = \frac{\tau^2 y_i+\sigma^2\overline{y}}{\tau^2+\sigma^2} = \frac{\sigma^2}{\tau^2+\sigma^2}\overline{y} + \frac{\tau^2}{\tau^2+\sigma^2}y_i = \overline{y} + (1-b)(y_i-\overline{y})$$

where $\overline{y}$ is the prior mean, y_i is the observed data, $0 < b = \sigma^2/(\tau^2+\sigma^2) < 1$ is the shrinkage factor.

6 Regression Models

7 Bayesian Analysis of Classical Regression

(a) Define the Multiple Linear Regression (MLR) model.

$$\begin{aligned}
y_i &= \mathbf{x}_i\boldsymbol{\beta} + e_i\\
\mathbf{y} &= X\boldsymbol{\beta} + \mathbf{e}
\end{aligned}$$

where $E_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$, $\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I)$, $\mathbf{Y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, X is called the design or model matrix, $\mathbf{e}$ is unbiased variable.

(b) Obtain the likelihood for the MLR model.

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, X) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right].$$

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, X) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, X) = (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right].$$

(c) Assume an improper uniform prior on $(\mathbf{B}, \log\Sigma)$. Show that $\mathbf{B}|\sigma^2, \mathbf{y}, X \sim \mathcal{N}_p(\hat{\beta}, \sigma^2(X^TX)^{-1})$, $\Sigma^2|\mathbf{y}, X \sim \text{InvGamma}[(n-p)/2, (n-p)s^2/2]$.

$$p(\boldsymbol{\beta}, \sigma^2) \propto (1)\frac{1}{\sigma^2} = \sigma^{-2}.$$

THen

$$\begin{aligned}
&p(\boldsymbol{\beta}, \sigma^2|y, X) \propto L(\boldsymbol{\beta}, \sigma^2|y, X)p(\boldsymbol{\beta}, \sigma^2)\\
&\quad\propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right]\\
&\quad= (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]\\
&\quad= (\sigma^2)^{-n/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]^T\left[(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]\right\}\\
&\quad= (\sigma^2)^{-n/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) + 2(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]\right\}\\
&\quad= (\sigma^2)^{-p/2-1} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]\\
&\quad(\sigma^2)^{-(n-p)/2-1} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right]\\
&\quad= p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, X)p(\sigma^2|\mathbf{y}, X)
\end{aligned}$$

where $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, X)$ is conditional posterior PDF of $\mathbf{B}|\Sigma^2$ and $p(\sigma^2|\mathbf{y}, X)$ is marginal posterior PDF of Σ^2.

Example 7.1 (Radon, p195, 378). (a) Cafully formulate the MLR model for this problem. Hypothesis: no intercept and no interaction.

$$\ln y_i = y_i^t = \beta_2 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i,$$

where

$$x_{1i} = \begin{cases} 1 & \text{basememt observation } i \\ 0 & \text{first floor observation } i \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{Blue Earth observation } i \\ 0 & \text{Otherwise observation } i \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{Clay observation } i \\ 0 & \text{Otherwise observation } i \end{cases}$$

$$x_{4i} = \begin{cases} 1 & \text{Goodhue observation } i \\ 0 & \text{Otherwise observation } i \end{cases}$$