Review

MTT720 / Bayesian Analysis I

Exam Information

- Final Exam on Dec. 13 (17-20) : need to **register** via SISU.

  - Closed book, bring pencils, erasers, and a calculator
  - A list of pdf/pmf's is given. (Lindley-Smith theorem/law of iterated expectation, etc. will be given if there are any related problems)
  - Be sure to show your work to get to the answers: partial credits are given to the right definitions, (mathematical not verbal) steps, etc.
  - You can write in Finnish for your reasonings.

- Exam covers all the contents dealt. Review the lab problems, lecture examples, and practice problems.

- What I expect you have learned
  - Bayes' Theorem
  - Know how to derive posterior distributions
  - Find conjugate priors. Distinguish noninformative priors.
  - Posterior inference: find posterior mean and variance, posterior probabilities, posterior interval estimates
  - Predictive distributions, etc.

More Information

- Grading : Lab (20%), Take-home task (10%), Final exam (70%)

- Take-home Assignment

  - Need to use R/ BUGS/ JAGS/ STAN for most of the problems.

  - Due 23.12 (a bit flexible but not unlimited!) : course grade will be delayed if submitted late. Inform clearly when you will finish up.

  - For submission, include all output and your own codes with brief comments so that I can reproduce your results when grading.

  - **No collaboration** is allowed: no credit will be given when detected.

  - Feel free to **consult the instructor for help**: Individual guidance will be given for you to proceed further upon requests.

Bayesian statistics

- Based on an idea of subjective probability, it provides a natural, intuitively plausible way to draw inferences for statistical problems by updating previous information based on data observations.

- Bayesian inference specifies probability distributions for the unknown parameters.

  - Anything unknown is a random variable.

  - Probability distributions are assumed for the unknown parameters and for the observations (i.e. both parameters and observations are random quantities).

  - Inferences are based on the prior distribution and the observed data.

Bayes' theorem

- The basic tool of inference is the Bayes' theorem.
- **Bayes' theorem**

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}
\end{aligned}
$$

- Theorem: let A_1,A_2,...,A_n be a set of mutually exclusive and exhaustive events. Then,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

- **Law of Total Probability**:

$$P(B) = P(B, A) + P(B, \sim A) = P(B|A)P(A) + P(B|\sim A)P(\sim A)$$

Bayesian Method for Inference

- Specify the prior distribution $[\theta], f(\theta)$ which expresses our knowledge about θ prior to observing the data.
- Form the Likelihood : $[X|\theta], f(x|\theta)$
  Model a set of observations with a probability distribution (expressed in the form of the likelihood function) with unknown parameter(s)
- Posterior
  Apply Bayes' theorem to derive posterior distribution which expresses all that is known about θ after observing the data: $[\theta|X], f(\theta|x)$
- Posterior Inference
  Derive appropriate inference statements from posterior distribution: e.g. point / interval estimates, probabilities of specified hypotheses.

$$
\begin{aligned}
[\theta|X] &= \frac{[X|\theta][\theta]}{[X]} \\
f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \quad \text{for continuous } \theta
\end{aligned}
$$

Kernel and Normalizing constant in the likelihood function

- In Bayesian statistics spotting kernels of distributions can be very useful in computing posterior distributions.

- For a random variable X with density(mass) function $f(\boldsymbol{x}|\theta)$ if $f(\boldsymbol{x}|\theta)$ can be expressed in the form $cq(\boldsymbol{x}|\theta)$ where c is a constant, not depending upon $\boldsymbol{x}$, then such $q(\boldsymbol{x}|\theta)$ is a <u>kernel</u> of the density $f(\boldsymbol{x}|\theta)$.

- The constant c is called a <u>normalizing constant</u> with the fact

$$\int_S f(\boldsymbol{x}|\theta)d\boldsymbol{x} = \int_S cq(\boldsymbol{x}|\theta)d\boldsymbol{x} = 1 \quad \Rightarrow 1/c = \int_S q(\boldsymbol{x}|\theta)d\boldsymbol{x}.$$

Choice of Priors

- Informative prior distributions reflect specific information about the parameter of interest.
  - priors based on subjective opinion should be chosen with care in practice
- Noninformative priors: 'reference priors' (reference for prior sensitivity) vague (or diffuse) prior, flat prior
- Improper priors: not a valid probability distribution - $\int_{\Theta} f(\theta)d\theta = \infty$. It can be used as long as it induces a proper posterior distribution
- Conjugate prior produces a posterior distribution (along with the data model) that has the same functional form as the prior (but with new, updated parameter values).
  - easy to understand the respective contributions of the prior and the data information to the posterior

Examples of conjugacy:

Beta prior-binomial data, Gamma prior-Poisson data, Normal prior-Normal data

Example: Binomial data & Beta conjugate prior

Likelihood : $X|\theta \sim$ Binomial (n, θ) $\quad f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

Prior : $\theta \sim$ Beta (α,β) $\quad f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)f(\theta) \propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} : \text{kernel of Beta}(x+\alpha, n-x+\beta) \end{aligned}$$

Thus, the posterior distribution is $\theta|X \sim \text{Beta}(x+\alpha, n-x+\beta)$

Suppose that only individual binary values are observed for data:
$Y_1, ..., Y_n|\theta \sim$ Bern(θ) where $\sum_{i=1}^n Y_i = X$,

Then, $L(\theta : \boldsymbol{y}) = \prod \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} = \theta^x(1-\theta)^{n-x}$

$\Rightarrow \theta|\boldsymbol{Y} \sim \text{Beta}(x+\alpha, n-x+\beta)$ **Likelihood Principle**

Example: Poisson - Gamma conjugacy

Likelihood: $\boldsymbol{X}|\lambda \sim$ Poisson(λ) $\qquad p(\boldsymbol{x}|\lambda) = \frac{1}{\prod x_i!}\lambda^{\sum x_i}e^{-n\lambda}$

Prior : $\lambda \sim$ Gamma(α, β) $\qquad p(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta}$

$$\begin{aligned} p(\lambda|\boldsymbol{x}) \propto f(\boldsymbol{x}|\lambda)f(\lambda) &= \lambda^{\sum x_i}\exp(-n\lambda)\ \lambda^{\alpha-1}\exp(-\lambda/\beta) \\ &= \lambda^{\sum x_i+\alpha-1}\exp(-(n+\frac{1}{\beta})\lambda) \end{aligned}$$

: kernel of Gamma$(\sum x_i + \alpha, \dfrac{1}{n+1/\beta})$

$\Rightarrow$ Posterior: $\lambda|\boldsymbol{X} \sim$ Gamma$(n\overline{X}+\alpha, \frac{1}{n+1/\beta})$

Posterior Mean : $E[\ \lambda|\boldsymbol{X}] = \dfrac{\sum x_i + \alpha}{n+\frac{1}{\beta}} = \dfrac{n\frac{\sum x_i}{n} + \frac{\alpha\beta}{\beta}}{n+\frac{1}{\beta}}$

: weighted average of $\overline{X}$ and prior mean $E[\lambda] = \alpha\beta$

Posterior Inference

1. Point Estimation
Posterior mean ($E[\theta|\boldsymbol{X}]$), Posterior variance($Var[\theta|\boldsymbol{X}]$), Mode, etc.

2. Interval Estimation : Bayesian credible intervals
Interpretation: there is 100$(1-\alpha)$% probability that θ lies in such an interval.
i) Bayesian **equal-tailed** intervals / quantile intervals
$(\theta_L,\ \theta_U)$ is a 100$(1-\alpha)$% equal-tailed interval for θ when

$$P(\theta < \theta_L|\boldsymbol{X}) = \frac{\alpha}{2} = P(\theta > \theta_U|\boldsymbol{X})$$

e.g. 95% equal-tailed interval is an interval : $[q_{0.025}, q_{0.975}]$
where $q_{0.025}$ and $q_{0.975}]$ are the quantiles of the posterior distribution.

ii) **Highest posterior density (HPD)** intervals (or Highest density region)
: posterior density for every point in this set is higher than the posterior density for any point outside of this set.
More meaningful for multimodal distributions

Likelihood of Normal data

$X_1, ..., X_n$ iid N(μ, τ) $\quad \tau > 0$

$$p(\boldsymbol{x}|\mu,\tau) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x_i-\mu)^2}{2\tau}\right)$$

$$\propto \tau^{-\frac{n}{2}} \exp\left(-\frac{\sum x_i^2 - 2\mu\sum x_i + n\mu^2}{2\tau}\right) \propto \tau^{-\frac{n}{2}} \exp\left(-\frac{n\mu^2 - 2\mu n\overline{x}}{2\tau}\right)$$

Or,

$$\begin{aligned}
p(\boldsymbol{x}|\mu,\tau) &= \left[\frac{1}{\sqrt{2\pi\tau}}\right]^n \exp\left(-\frac{1}{2\tau}[\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu)^2]\right) \\
&= \left[\frac{1}{\sqrt{2\pi\tau}}\right]^n \exp\left(-\frac{\sum(x_i-\overline{x})^2}{2\tau}\right) \exp\left(-\frac{n(\mu-\overline{x})^2}{2\tau}\right) \\
&\propto \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau}[n(\mu-\overline{x})^2 + S]\right), \qquad S = \sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&\propto \tau^{-\frac{n}{2}} \exp\left(-\frac{n(\mu-\overline{x})^2}{2\tau}\right)
\end{aligned}$$

Normal Samples with one unknown parameter

- Case 1. μ is unknown with no prior information, but τ is known
  Take a flat prior for μ : $p(\mu) = 1$ (Jeffreys' prior)
  Posterior : $\mu|\boldsymbol{X} \sim N(\overline{X}, \frac{\tau}{n})$
- Case 2. Conjugate prior for $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ & Variance (τ) known
  Posterior: Normal(μ^*, σ^{2*}) where
  Posterior mean : $\mu^* = \dfrac{n\bar{x}/\tau + \mu_0/\sigma_0^2}{n/\tau + 1/\sigma_0^2}$ variance : $\sigma^{2*} = \dfrac{1}{n/\tau + 1/\sigma_0^2}$
- Case 3. $\mathrm{N}(\mu, \tau)$: μ is known & τ is unknown
  Conjugate prior for τ : $\tau \sim \mathrm{IG}(a, b)$
  Posterior: $\tau|\boldsymbol{X} \sim$ inverse Gamma $\left(a + \frac{n}{2}, \frac{1}{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} + \frac{1}{b}}\right)$

Case 2. Conjugate prior for μ & Variance (τ) known

Likelihood : $\boldsymbol{X}|\mu \sim$ Normal(μ, τ)
Prior: $\mu \sim$ Normal(μ_0, σ_0^2)

$$\begin{aligned} p(\mu|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\mu)p(\mu) \propto \exp\left(-\frac{n}{2\tau}(\mu-\overline{x})^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\ &= \exp\left(-\frac{1}{2}\left[\frac{n}{\tau}(\mu-\overline{x})^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right]\right) = \exp\left(-\frac{1}{2}Q\right) \end{aligned}$$

Complete the square for μ, i.e. find μ^*, σ^{2*} such that the inside of the exponential function has the form $\frac{1}{\sigma^{2*}}(\mu-\mu^*)^2$,

$$\begin{aligned} Q &\propto \left(\frac{n}{\tau}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{n}{\tau}\overline{x}+\frac{1}{\sigma_0^2}\mu_0\right)\mu + \ldots \\ &= \left(\frac{n}{\tau}+\frac{1}{\sigma_0^2}\right)\left[\mu - \frac{n\bar{x}/\tau+\mu_0/\sigma_0^2}{n/\tau+1/\sigma_0^2}\right]^2 + R \end{aligned}$$

Posterior: $\mu|\boldsymbol{X} \sim$ Normal(μ^*, σ^{2*}) with $\mu^* = \frac{n\bar{x}/\tau+\mu_0/\sigma_0^2}{n/\tau+1/\sigma_0^2}$, $\sigma^{2*} = \frac{1}{n/\tau+1/\sigma_0^2}$

Predictive Distribution

Want to predict a **future observation** y given that we have observed $\boldsymbol{X} = (X_1, ..., X_n)$. Assume that $X_1, ..., X_n$ are independent given θ. Bayesian inference about y is based on its posterior predictive distribution:

$$
\begin{aligned}
p(y|\boldsymbol{x}) &= \int p(y, \theta|\boldsymbol{x}) d\theta = \int p(y|\theta, \boldsymbol{x}) p(\theta|\boldsymbol{x}) d\theta = \int \underbrace{p(y|\theta)}_{lkhd} \underbrace{p(\theta|\boldsymbol{x})}_{posterior} d\theta \\
&= \sum_\theta p(y|\theta) p(\theta|\boldsymbol{x}) \quad \text{for discrete cases}
\end{aligned}
$$

The mean and variance of a predictive distribution can be obtained using standard formulae:

$$
\begin{aligned}
E(Y) &= E_\Theta[E(Y|\theta)] \\
\mathrm{Var}(Y) &= E_\Theta[\mathrm{Var}(Y|\theta)] + \mathrm{Var}_\Theta[E(Y|\theta)]
\end{aligned}
$$

$\Rightarrow E(Y|\boldsymbol{X}), Var(Y|\boldsymbol{X})$ can be otained in the same pattern.

Example : Posterior predictive distribution

Predicting t successes in m future observations $\Leftrightarrow t|\theta \sim$ Binomial(m, θ)

Data: $X \sim$ Binom$(n, \theta\ )$ with a conjugate prior $\theta \sim$ Beta(α, β)
$\Rightarrow$ Posterior: $\theta|X \sim$ Beta$(\alpha + s, n + \beta - s)$

$$
\begin{aligned}
p(t|x) &= \int p(t|\theta)p(\theta|x)d\theta \\
&= \int_0^1 \binom{m}{t} \theta^t (1-\theta)^{m-t} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+s)\Gamma(n+\beta-s)} \theta^{\alpha+s-1}(1-\theta)^{n+\beta-s-1} d\theta \\
&= \binom{m}{t} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+s)\Gamma(n+\beta-s)} \frac{\Gamma(t+\alpha+s)\Gamma(m-t+n+\beta-s)}{\Gamma(m+n+\alpha+\beta)}
\end{aligned}
$$

Frequentists' approach
i) First, estimate θ by $\hat{\theta}$ (MLE, unbiased estimator, etc.)
ii) Substitute θ by $\hat{\theta}$. (e.g. Binom(m, θ) by Binom$(m, \hat{\theta})$)
iii) Predict y based on the distribution with the estimate of θ plugged in.
(e.g. Binom$(m, \hat{\theta})$) - ignores uncertainty of $\hat{\theta}$

Example : Discrete posterior predictive function

	θ	$p(\theta)$	likelihood	prior $\times$ lkhd	posterior
'great'	1/2	0.2	0.1048	0.02096	0.5008
'good'	1/4	0.5	0.0413	0.02065	0.4934
'poor'	1/8	0.3	0.0008	0.00024	0.0058
		1		0.04185	1

$$\begin{aligned}
p(y|x=10) &= \sum_{\theta=\frac{1}{2},\frac{1}{4},\frac{1}{8}} p(y|\theta)p(\theta|x=10) \\
&= \left(\frac{12^y e^{-12}}{y!} \times 0.5008\right) + \left(\frac{6^y e^{-6}}{y!} \times 0.4934\right) + \left(\frac{3^y e^{-3}}{y!} \times 0.0058\right) \\
&= \frac{(12^y e^{-12})0.5008 + (6^y e^{-6})0.4934 + (3^y e^{-3})0.0058}{y!}
\end{aligned}$$

For example,

$$\begin{aligned}
P(y=10|x=10) &= \frac{(12^{10} e^{-12})0.5008 + (6^{10} e^{-6})0.4934 + (3^{10} e^{-3})0.0058}{10!} \\
&= 0.073
\end{aligned}$$

Mixtures of conjugacy

The family of mixtures of conjugates can approximate any prior distribution to any required level of accuracy.

A mixture of the distributions $\pi_j(\theta)$ with weights $a_j (j = 1, 2, ..., m)$ has probability (density) function

$$\pi(\theta) = \sum_{j=1}^{m} a_j \pi_j(\theta) \text{ where } \sum_{j=1}^{m} a_j = 1$$

For the individual prior $\pi_j(\theta)$, the (individual) posterior density $p_j(\theta|\boldsymbol{x})$ can be derived using the Bayes theorem.

$$p_j(\theta|\boldsymbol{x}) = \frac{\pi_j(\theta) p(\boldsymbol{x}|\theta)}{\int \pi_j(\theta) p(\boldsymbol{x}|\theta) d\theta} \equiv \frac{\pi_j(\theta) p(\boldsymbol{x}|\theta)}{c_j} \quad c_j: \text{ normalizing constant}$$

Then, the posterior is a mixture of the respective posterior distributions under each prior.

$$p(\theta|\boldsymbol{x}) = \sum_{j=1}^{m} a_j^* p_j(\theta|\boldsymbol{x}) \text{ where } a_j^* = \frac{a_j c_j}{\sum_{j=1}^{m} a_j c_j}$$

Normal samples with conjugate NIC prior

Case 4. $N(\mu, \tau)$: **Both μ and τ are unknown**

Prior : $(\mu, \tau) \sim$ NIC (p, q, m, v), conjugate family of joint prior for μ and τ
Posterior : $(\mu, \tau)|X \sim$ NIC(p_1, q_1, m_1, v_1)

$$\begin{aligned} p_1 &= p + n, \quad m_1 = \frac{v^{-1}m + n\bar{x}}{v^{-1} + n}, \quad v_1 = (v^{-1} + n)^{-1} \\ q_1 &= q + S + (v + n^{-1})^{-1}(\bar{x} - m)^2 \end{aligned}$$

Marginal inference about μ: $\mu|\boldsymbol{X} \sim t_{p_1}(m_1, q_1 v_1/p_1)$

$$\begin{aligned} E[\mu|\boldsymbol{X}] &= m_1 = \frac{v^{-1}m + n\bar{x}}{v^{-1} + n} \\ Var[\mu|\boldsymbol{X}] &= \frac{q_1 v_1}{p_1 - 2} \end{aligned}$$

Marginal inference about τ: $\tau|\boldsymbol{X} \sim IC(p_1, q_1)$

Normal Samples with two unknown parameters

Case 5. Normal(μ, τ): μ, τ are unknown.
Take a noninformative prior: $p(\mu, \tau) \propto \frac{1}{\tau} \times 1 \equiv$ NIC $(-1, 0, m, \infty)$

$$
\begin{aligned}
p(\mu, \tau|\boldsymbol{x}) &\propto \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau}[\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu)^2]\right) \times \frac{1}{\tau} \\
&\propto \tau^{-\frac{n}{2}-1} \exp\left(-\frac{1}{2\tau}[n(\mu - \overline{x})^2 + S]\right), \qquad S = \sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned}
$$

$\Rightarrow \mu|\tau, \boldsymbol{X} \sim N(\overline{X}, \tau/n)$

Marginal posterior of μ : $\mu|\boldsymbol{X} \sim t_{n-1}(\overline{X}, S/n(n-1))$
Marginal posterior of τ : $\tau|\boldsymbol{X} \sim$ IC$(n-1, S)$

$p(\mu|\boldsymbol{x}) = \int_0^\infty p(\mu, \tau|\boldsymbol{x})d\tau$ by integrating τ out of joint posterior
$p(\tau|\boldsymbol{x}) = \int_{-\infty}^\infty p(\mu, \tau|\boldsymbol{x})d\mu$ by integrating μ out of joint posterior

Two Normal Samples

Likelihood: $Y_{1i}|\mu_1, \tau_1 \sim$ Normal$(\mu_1, \tau_1) \quad i = 1, ..., n_1$
$Y_{2i}|\mu_2, \tau_2 \sim$ Normal$(\mu_2, \tau_2) \quad i = 1, ..., n_2$

Case 1. τ_1 **and** τ_2 **are assumed to be known**
Can take independent reference priors for μ_1 and μ_2: $p(\mu_1, \mu_2) = 1$

$$p(\boldsymbol{y}|\mu_1, \mu_2, \tau) \propto \tau_1^{-\frac{n_1}{2}} \exp\left(-\frac{1}{2\tau_1}[n_1(\mu_1 - \overline{y}_1)^2 + S_1]\right)$$
$$\times \tau_2^{-\frac{n_2}{2}} \exp\left(-\frac{1}{2\tau_2}[n_2(\mu_2 - \overline{y}_2)^2 + S_2]\right)$$
$$\Rightarrow p(\mu_1, \mu_2|\boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\mu_1|\boldsymbol{y}_1)p(\mu_2|\boldsymbol{y}_2)$$

Posterior:

$$\mu_1|\boldsymbol{Y} \sim N(\overline{Y}_1, \frac{\tau_1}{n_1}) \quad \text{independent of} \quad \mu_2|\boldsymbol{Y} \sim N(\overline{Y}_2, \frac{\tau_2}{n_2})$$
$$\Rightarrow \mu_1 - \mu_2|\boldsymbol{Y} \sim N(\overline{Y}_1 - \overline{Y}_2, \frac{\tau_1}{n_1} + \frac{\tau_2}{n_2})$$

Case 2. Take **independent priors uniform in** μ_1, μ_2, τ: $p(\mu_1, \mu_2, \tau) = \frac{1}{\tau}$
Assume $\tau_1 = \tau_2 \equiv \tau$

Joint posterior: $$p(\mu_1, \mu_2, \tau | Y) \propto \underbrace{\tau^{-\frac{n_1+n2+2}{2}} \exp\left(-\frac{1}{2\tau}(S_1 + S_2)\right)}_{IC(n_1+n_2, S_1+S_2)}$$

$$\times \underbrace{\exp\left(-\frac{1}{2\tau}[n_1(\mu_1 - \overline{Y}_1)^2]\right)}_{N(\overline{Y}_1, \frac{\tau}{n_1})} \underbrace{\exp\left(-\frac{1}{2\tau}[n_2(\mu_2 - \overline{Y}_2)^2]\right)}_{N(\overline{Y}_2, \frac{\tau}{n_2})}$$

$$\Rightarrow \mu_1 - \mu_2 | \tau, \boldsymbol{Y} \sim N(\overline{Y}_1 - \overline{Y}_2, \tau\left(\frac{1}{n_1} + \frac{1}{n_2}\right))$$

Initegrating τ out

$$\Rightarrow \mu_1 - \mu_2 | Y \sim t_{n_1+n_2-2}(\overline{Y}_1 - \overline{Y}_2, \frac{S_1 + S_2}{n_1 + n_2 - 2}(1/n_1 + 1/n_2))$$

Case. 3 $\tau_1 \neq \tau_2$ unknown

Take conjugate NIC priors independently for (μ_1, τ_1) and (μ_2, τ_2)

Prior: $(\mu_i, \tau_i) \sim \text{NIC}(p_i, q_i, m_i, v_i) \quad i = 1, 2$
Posterior: $(\mu_i, \tau_i)|\boldsymbol{Y} \sim \text{NIC}(p_i^*, q_i^*, m_i^*, v_i^*)$ independently

$$
\begin{aligned}
p_i^* &= p_1 + n_1, \qquad\qquad q_i^* = q_1 + S_1 + (v_1 + n_1^{-1})^{-1}(\overline{Y}_1 - m_1)^2 \\
m_i^* &= \frac{v_1^{-1} m_1 + n_1 \overline{Y}_1}{v_1^{-1} + n_1}, \quad v_i^* = (v_1^{-1} + n_1)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\text{Marginally, } \mu_1|\boldsymbol{Y}_1 &\sim t_{p_1^*}(m_1^*, q_1^* v_1^* / p_1^*) \\
\mu_2|\boldsymbol{Y}_2 &\sim t_{p_2^*}(m_2^*, q_2^* v_2^* / p_2^*)
\end{aligned}
$$

Direct simulation is easier to get samples from $\mu_1 - \mu_2|\boldsymbol{Y}$.

However, the posterior mean and variance can be found using
$E[\delta|\boldsymbol{Y}] = E[\mu_1|\boldsymbol{Y}] + E[\mu_2|\boldsymbol{Y}]$ and $Var[\delta|\boldsymbol{Y}] = Var[\mu_1|\boldsymbol{Y}] + Var[\mu_2|\boldsymbol{Y}]$

Linear Regression: Bayesian Inference

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \tau I)$$

Likelihood: $\boldsymbol{Y}_{n\times 1}|\boldsymbol{\beta}, \tau \sim N(\boldsymbol{X}_{n\times k}\boldsymbol{\beta}_{k\times 1}, \tau I)$
$p(\boldsymbol{Y}|\boldsymbol{\beta}, \tau) = (2\pi\tau)^{-n/2} \exp\left(-\frac{1}{2\tau}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$

Prior: $\boldsymbol{\beta}, \tau \sim$ Multivariate NIC $(p, q, \boldsymbol{m}, V)$

$$p(\boldsymbol{\beta}, \tau) \propto \tau^{-(p+k+2)/2} \exp(-\frac{1}{2\tau}\{q + (\boldsymbol{\beta} - \boldsymbol{m})^T V^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\})$$

Posterior:
$$\begin{aligned} \boldsymbol{\beta}, \tau|\boldsymbol{Y} &\sim NIC(p^*, q^*, \boldsymbol{m}^*, V^*) \\ \boldsymbol{\beta}|\boldsymbol{Y} &\sim t_{p^*}(\boldsymbol{m}^*, \frac{q^*}{p^*}V^*) \end{aligned}$$

$$\begin{aligned} p^* &= p + n \quad q^* = q + S + (\hat{\boldsymbol{\beta}} - \boldsymbol{m})^T(V + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{m}) \\ \boldsymbol{m}^* &= V^*(V^{-1}\boldsymbol{m} + (\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}}) \quad V^* = (V^{-1} + (\boldsymbol{X}^T\boldsymbol{X}))^{-1} \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad S = (Y - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$

Simple regression

Observations Y_i's are independent given (a, b, τ) :

$$Y_i | a, b, \tau \sim N(a + bx_i, \tau)$$

This is a linear model in which $k = 2$ and

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}$$

For conjugate NIC $(p, q, \boldsymbol{m}, V)$ prior,

$$\boldsymbol{m} = \begin{bmatrix} E(a) \\ E(b) \end{bmatrix} \quad Var(\boldsymbol{\beta}) = E(\tau)V = \begin{bmatrix} Var(a) & \mathsf{Cov}(a, b) \\ \mathsf{Cov}(a, b) & Var(b) \end{bmatrix}$$

Note $V = \frac{1}{E(\tau)} Var(\boldsymbol{\beta})$.

Two normal samples in linear model formulation

Data : Y_{ij}, for $i = 1, 2$ and $j = 1, 2, ..., n_i$
Assume that variances for two samples are **equal** : $\tau_1 = \tau_2 \equiv \tau$.

Prior: $(\boldsymbol{\beta}, \tau)^T = (\mu_1, \mu_2, \tau) \sim$ Multivariate NIC $(p, q, \boldsymbol{m}, V)$

Joint Posterior: $(\mu_1, \mu_2, \tau)^T | \boldsymbol{Y} \sim$ NIC $(p^*, q^*, \boldsymbol{m}^*, V^*)$

Let $\delta = \mu_1 - \mu_2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \equiv A\boldsymbol{\beta}$
By Theorem,

$$(\delta, \tau)|\boldsymbol{Y} \sim NIC(p^*, q^*, A\boldsymbol{m}^*, AV^*A^T)$$

$$\delta|\boldsymbol{Y} \sim t_{p^*}(A\boldsymbol{m}^*, AV^*A^T)$$

where $A\boldsymbol{m}^* = m_1^* - m_2^*$

Other Multiparameter Models

- **Theorem (Lindely and Smith)**

  Likelihood: $\boldsymbol{Y}|\boldsymbol{\theta_1} \sim N(A_1\boldsymbol{\theta_1}, C_1)$ where A_1, C_1 is known.

  Prior: $\boldsymbol{\theta_1} \sim N(A_2\boldsymbol{\theta_2}, C_2)$ where $A_2, \boldsymbol{\theta_2}, C_2$ is known.

  $\Rightarrow$ Posterior: $\boldsymbol{\theta_1}|\boldsymbol{Y} \sim N(B\boldsymbol{b}, B)$ where

  $$B = (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1} \quad b = A_1^T C_1^{-1}\boldsymbol{Y} + C_2^{-1} A_2 \boldsymbol{\theta_2}$$

  - Marginal distribution of $\boldsymbol{Y} \sim N(A_1 A_2 \boldsymbol{\theta_2}, C_1 + A_1 C_2 A_1^T)$

- Multinomial data : $\boldsymbol{Y}|\boldsymbol{\theta} \sim$ Multinomial$(n, \boldsymbol{\theta})$ i.e. $p(\boldsymbol{y}|\boldsymbol{\theta}) \propto \theta_1^{y_1} \cdots \theta_k^{y_k}$
  Prior: $\boldsymbol{\theta} \sim$ Dirichlet $D(\alpha_1, ..., \alpha_k)$

  $$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha_1 + ... + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \quad \sum_j \theta_j = 1$$

  $\Rightarrow$ Posterior: $\boldsymbol{\theta}|\boldsymbol{Y} \sim$ Dirichlet $D(\alpha_1 + y_1, ..., \alpha_k + y_k)$

Hierarchical Models

Hierarchical modeling is used when information is available on several different levels of observational units.

$$\begin{cases} \boldsymbol{x}|\theta & \text{Likelihood} \\ \theta|\phi & \text{prior} \\ \phi & \text{Hyperprior} \end{cases} \qquad \begin{cases} \boldsymbol{x}|\boldsymbol{\theta} & \text{Likelihood} \\ \boldsymbol{\theta}|\boldsymbol{\phi} & \text{prior} \\ \boldsymbol{\phi}|\boldsymbol{\lambda} & \text{Hyperprior} \\ \boldsymbol{\lambda}|\boldsymbol{\xi} & \text{Hyper-Hyper prior} \\ \boldsymbol{\xi} & \vdots \end{cases}$$

The joint posterior is :

$$\begin{aligned} p(\theta, \phi|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)p(\theta|\phi)p(\phi) \\ p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\xi}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\xi})p(\boldsymbol{\xi}) \end{aligned}$$

The marginal posterior distributions :

$$p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)p(\theta) \qquad p(\phi|\boldsymbol{x}) \propto p(\phi)p(\boldsymbol{x}|\phi)$$

Example : Hierarchical models

Likelihood: $Y_j|\theta_j \sim \text{Binom}(n_j, \theta_j) \quad j = 1, ..., J$

Prior: $\theta_j|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$ Hyperprior: $p(\alpha, \beta) \propto \frac{1}{(\alpha+\beta)^{5/2}}$

a) Joint posterior: $p(\boldsymbol{\theta}, \alpha, \beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha, \beta)p(\alpha, \beta)$

$$\propto \quad \frac{1}{(\alpha+\beta)^{5/2}} \prod_{j=1}^{J} \left[\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\right]$$

b) Marginal conditional density:

$$\Rightarrow p(\boldsymbol{\theta}|\alpha, \beta, \boldsymbol{y}) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}$$

c) Marginal posterior (up to a proportionality constant):

$$p(\alpha, \beta|\boldsymbol{y}) \quad = \quad \frac{p(\alpha, \beta, \boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\alpha, \beta, \boldsymbol{y})} \equiv \frac{\text{density from a)}}{\text{density from b)}}$$

$$\propto \quad \frac{1}{(\alpha+\beta)^{5/2}} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

Bayesian vs. Frequentist Inference

- Frequentists are disturbed by the dependence of the posterior results on the subjective prior distribution
- Bayesians say that the prior distribution is not the only subjective element in an analysis. The assumptions about the sampling distributions are also subjective.
- Whose probability distribution should be used?
  When there are enough data, a good Bayesian analysis and a good frequentist analysis will tend to agree.
  If the results are sensitive to prior information, a Bayesian analyst is obligated to report this sensitivity and to present different results obtained from a wide range of prior information.
- Bayesians can often handle problems the frequentist approach cannot. Bayesians often apply frequentist techniques but with a Bayesian interpretation. Most untrained people interpret results in the Bayesian way more easily. (Often the Bayesian answer is what the decision maker really wants to hear.)