# Unit 3: Simple Models

# Course Units

- Introduction to Bayesian Statistics
- Prior Distributions
- Simple Models
- Bayesian Asymptotics
- Hierarchical Modeling
- Bayesian Computation
- Model Assessment and Comparison
- Regression Modeling
- Bayesian Nonparametrics
- Survival Analysis and Missing Data
- Clinical Trials and Bayesian Design

# Outline of the Unit



Motivation: models are simple but serve as good practice for understanding the concepts and as building blocks for more complicated models, particularly in doing MCMC, which involves sequentially drawing parameter values from their conditional distributions, pretending the other parameters are known.

# Basic models

- univariate outcomes:
  - binomial
  - normal, variance known
  - normal, mean known
  - normal, mean and variance unknown
- multivariate outcomes
  - multinomial
  - multivariate normal

# Binomial model

$$Y \sim \text{Binomial}(n, \theta)$$
$$\theta \sim \text{Beta}(\alpha, \beta)$$

We saw that a-posteriori:

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

$$\begin{aligned} &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{y} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &\equiv \text{Beta}(y+\alpha, n-y+\beta) \\ \therefore \theta|y &\sim \text{Beta}(y+\alpha, n-y+\beta) \end{aligned}$$

# Shrinkage

$$\begin{aligned} \text{if } \theta|y &\sim \text{Beta}(y+\alpha, n-y+\beta) \\ E(\theta \mid y) &= \frac{\alpha+y}{\alpha+\beta+n} \\ &= B\left(\frac{\alpha}{\alpha+\beta}\right) + (1-B)\frac{y}{n} \end{aligned}$$

is a shrinkage factor defined as:

$$B = \frac{\alpha+\beta}{\alpha+\beta+n}$$

- if $(\alpha+\beta) \to 0$, $B \to 0$, then $E(\theta \mid y) = \bar{y}$.
- if $(\alpha+\beta) \to \infty$, $B \to 1$, then $E(\theta \mid y) = \frac{\alpha}{\alpha+\beta}$.

# Normal model (variance known)

$$y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2) \quad \sigma^2 \text{ known}$$
$$\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$$

We saw that a-posteriori:

$$\theta|y \sim \mathcal{N}\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left[\frac{\theta_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right], \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$$

# Shrinkage

Now

$$E(\theta|y) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left[\frac{\theta_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right] = B\theta_0 + (1-B)\bar{y}$$

$$B = \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \qquad 0 \leq B \leq 1$$

$$E(\theta|y) = B\theta_0 + (1-B)\bar{y}$$

The data, $\bar{y}$, are shrunk toward the prior, $\theta_0$, by a factor, $B$, reflecting the relative precisions.

- if $\sigma_0^2 \to \infty$, $B \to 0$, $E(\theta|y) = \bar{y}$
- if $\sigma_0^2 \to 0$, $B \to 1$, $E(\theta|y) = \theta_0$

$$\begin{aligned} V(\theta|y) &= \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \\ &= B\sigma_0^2 \quad 0 \leq B \leq 1 \end{aligned}$$

$\therefore$ posterior variance $\leq$ prior variance.

# Filtering

Can also write the posterior mean as:

$$E(\theta|y) = \underbrace{\theta_0}_{\text{prior mean}} + \underbrace{(\bar{y} - \theta_0)}_{\text{deviation of data mean from its marginal mean}} \underbrace{(1 - B)}_{\text{Kalman Filter}}$$

$\therefore$ posterior mean is obtained by updating the prior mean by *filtering* the deviation of $\bar{y}$ from its marginal mean, $\theta_0 = E(\bar{Y}) = E_{\theta|\theta_0}(E_{Y|\theta}(\bar{Y}|\theta))$

# Normal model (mean known)

Normal Likelihood, $\sigma^2$ unknown, $\theta$ known

$$y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$$

Let $\tau = \frac{1}{\sigma^2}$ = precision

Suppose $\tau \sim \mathcal{G}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$

---

Aside: Recall $\theta \sim \mathcal{G}(\alpha, \beta)$

then $P(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$

$\theta > 0$

$\alpha > 0$ (shape), $\beta > 0$ (inverse scale=rate)

$E(\theta) = \frac{\alpha}{\beta}$; $V(\theta) = \frac{\alpha}{\beta^2}$

Also parameterized as the scale ($1/\beta$). Default in R is to use the 'rate' as the 2nd parameter. (be careful)

---

# Normal Model (mean known), cont'd

$$P(\tau|y,\theta) \propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\sum_i (y_i-\theta)^2\right\} \times$$
$$\tau^{\frac{\delta_0}{2}-1} \exp\left(\frac{-\tau\gamma_0}{2}\right)$$
$$= \tau^{\frac{n+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}\left(\gamma_0 + \sum_i (y_i-\theta)^2\right)\right\}$$

Recognizing the kernel of a gamma:

$\therefore$ a-posteriori,

$$\tau|y,\theta \sim \mathcal{G}\left(\frac{n+\delta_0}{2}, \frac{\gamma_0 + \sum_i (y_i-\theta)^2}{2}\right)$$

# Inverse Chi-Square and Inverse Gamma

- **Inverse Chi-Square Distribution.**
  - If $W \sim \chi^2_\nu$ then $\theta = \frac{1}{W} \sim \frac{1}{\chi^2_\nu}$ , so we say $\theta \sim$ Inv-$\chi^2_\nu$
  - $P(\theta) = \dfrac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{-(\frac{\nu}{2}+1)} e^{-\frac{1}{2\theta}}$ $\quad \theta > 0$, $\nu > 0$
- **Inverse Gamma Distribution.**
  - If $W \sim \mathcal{G}(\alpha, \beta)$ then $\theta = \frac{1}{W} \sim$ Inv-gamma$(\alpha, \beta)$ $(\mathcal{IG}(\alpha, \beta))$
  - $P(\theta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\frac{\beta}{\theta}}$ $\quad \theta > 0$, $\alpha > 0,\ \beta > 0$
  - As with the gamma, this is sometimes parameterized in terms of $1/\beta$.
  - This is the conjugate prior for the Normal variance.

# Scaled Inverse Chi-Square

- **Scaled Inverse Chi-Square Distribution.**

$$\theta \sim \text{Inv} - \chi^2(\nu, s^2) \text{ if}$$

$$P(\theta) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} s^{\nu}\, \theta^{-\left(\frac{\nu}{2}+1\right)} e^{\frac{-\nu s^2}{2\theta}} \qquad \begin{array}{l} \theta > 0 \\ \nu > 0 \\ s^2 > 0 \end{array}$$

$$E(\theta) = \frac{\nu}{\nu - 2} s^2; \qquad \nu > 2$$

$$V(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4; \qquad \nu > 4$$

$$\text{Inv-}\chi^2(\nu, s^2) \equiv \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu s^2}{2}\right)$$

Note: I tend to consider distributions as gamma or inverse gamma and don't tend to think about the parameterization as chi-square or inverse-chi-square.

# Normal-Inverse Chi-Square

- **Normal Inverse Chi-Square Distribution.**

  Denoted $\mathcal{N}\text{-Inv-}\chi^2\left(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2\right)$

  $$P(\mu, \sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu_0+1}{2}+1\right)} \exp\left[\frac{-1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right)\right]$$

  for $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

  $\mu_0$ = location of $\mu$
  $\frac{\sigma_0^2}{\kappa_0}$ = scale of $\mu$ (since $\sigma_0^2$ is scale of $\sigma^2$)
  $\kappa_0$ = prior sample size
  $\nu_0$ = degrees of freedom for $\sigma^2$
  $\sigma_0^2$ = scale of $\sigma^2$

  This is the conjugate joint prior for the univariate normal model.

# Normal-Inverse Chi-Square (2)

Motivation:

# Student-$t$

- **Student-$t$ Distribution.**
  Definition: $\theta \sim t_\nu(\mu, \sigma^2)$ if

$$P(\theta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\sigma}\left[1 + \frac{1}{\nu}\frac{(\theta-\mu)^2}{\sigma^2}\right]^{-\left(\frac{\nu+1}{2}\right)}$$

  $-\infty < \theta < \infty$
  $\nu =$ degrees of freedom; $\mu =$ location; $\sigma^2 =$ scale
  Standard $t$: $z \sim t_\nu \Rightarrow$

$$P(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left[1 + \frac{z^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)}$$

# Multivariate Student-$t$

- **Multivariate Student-$t$ Distribution.**

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d) \quad d \times 1 \text{ random vector}$$

Definition: $\theta \sim t_\nu(\mu, \Sigma)$ if

$$P(\theta) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{d}{2}}} |\Sigma|^{-\frac{1}{2}} \times \left[1 + \frac{1}{\nu}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right]^{-\left(\frac{\nu+d}{2}\right)}$$

$d$-dimensional multivariate (Student)-$t$ distribution with degrees of freedom $\nu$, location $\mu$, and symmetric positive definite scale $\Sigma_{d\times d}$ .

# The $t$ as a scale mixture

A $t$-distribution can be obtained as a scale mixture of Normals. Let $\tau = \sigma^{-2}$.

$$y|\tau \sim \mathcal{N}_d(\mu, \tau^{-1}\Sigma) \quad \tau \sim \mathcal{G}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$$

Then the marginal distribution of $y$ is:

$$\begin{aligned} P(y) &= \int_0^\infty P(y|\tau)P(\tau)d\tau \\ &= \int_0^\infty (2\pi)^{-\frac{d}{2}}\tau^{\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\tau^{\frac{\delta_0}{2}-1}\exp\left\{-\frac{\tau\gamma_0}{2}\right\} \\ &\times \exp\left\{-\frac{\tau}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}d\tau \end{aligned}$$

$$\propto \int_0^\infty \underbrace{\tau^{\frac{\delta_0+d}{2}-1}\exp\left\{-\frac{\tau}{2}\left[\gamma_0 + (y-\mu)^T\Sigma^{-1}(y-\mu)\right]\right\}}_{\mathcal{G}\left(\frac{\delta_0+d}{2}, \frac{\gamma_0+(y-\mu)^T\Sigma^{-1}(y-\mu)}{2}\right)} d\tau$$

$$\propto \left[\frac{\gamma_0}{2} + \frac{(y-\mu)^T\Sigma^{-1}(y-\mu)}{2}\right]^{-\left(\frac{\delta_0+d}{2}\right)}$$

$$\propto \left[1 + \frac{(y-\mu)^T\Sigma^{-1}(y-\mu)}{\gamma_0}\right]^{-\left(\frac{\delta_0+d}{2}\right)}$$

$$\Rightarrow y \sim t_{\delta_0}(\mu, \frac{\gamma_0}{\delta_0}\Sigma).$$

# Distributional Theory (summary)

**You should now be familiar with (i.e., recognize the kernel):**

1. Normal Distribution
2. Beta Distribution
3. Gamma Distribution (and $\chi^2$ distribution)
4. Inverse Gamma Distribution (and Scaled Inverse Chi-Square Distribution)
5. Student-t Distribution
6. Multivariate Student-t Distribution
7. Normal-Inverse Chi-Square Distibution

- Note: in terms of recognizing gamma, inverse gamma, chi-square, etc., the key is to remember to look for the random variable raised to a power and in the exponential.
- Let GCSR Appendix A or CL Appendix A be your bible.

# Joint and Marginal Distributions

Joint Posterior Distribution:

$$P(\theta_1, \theta_2|y) \propto P(y|\theta_1, \theta_2)P(\theta_1, \theta_2)$$

Marginal Posterior Distribution:

$$P(\theta_1|y) = \int_{-\infty}^{\infty} P(\theta_1, \theta_2|y)d\theta_2$$

$$\text{or} = \int_{-\infty}^{\infty} P(\theta_1|\theta_2, y)P(\theta_2|y)d\theta_2$$

$\therefore$ Posterior for $\theta$ is a mixture of the conditional posterior distribution given $\theta_2$ with weights $= P(\theta_2|y)$. So if $\theta_1$ is a mean parameter and $\theta_2$ a variance term, our inference for the mean accounts for our uncertainty in the variance term.

- Note that we refer to $P(\theta_1|\theta_2, y)$ as the conditional posterior distribution and $P(\theta_2|y)$ as the marginal posterior distribution.
- Part of the elegance of Bayesian statistics comes from its intuitive and simple inferential basis in distributional properties: joint, marginal, and conditional distributions.

# A simple sampling strategy

Practical strategy for evaluation of multi-parameter models:

$$P(\theta_1|y) = \int_{\Theta_2} P(\theta_1|\theta_2, y)P(\theta_2|y)d\theta_2$$

- Draw $\theta_2$ from $P(\theta_2 \mid y)$.
- Given $\theta_2$, draw $\theta_1$ from $P(\theta_1 \mid \theta_2, y)$.

# Nuisance parameters

Generally: $\theta = (\theta_1, \theta_2)$

$\theta_1$: primary interest; $\theta_2$: nuisance parameter

- The Bayesian method: Bayes integrates over nuisance parameters, averaging over the uncertainty in them. The marginal posterior for the parameter of interest is a mixture of the conditional posterior over the marginal posterior of the nuisance parameter(s). The latter serves as a weighting function, acknowledging the uncertainty in $\theta_2$.
- From a decision-theoretic perspective, inference about $\theta_1$ should be based on a loss function that is a function only of $\theta_1$. Therefore, we want to work with the marginal posterior for $\theta_1$. The Bayesian approach to nuisance parameters follows naturally.

# Nuisance parameters (cont'd)

- Remember that nuisance parameters are introduced to simplify the specification of the sampling model, allowing conditional independences. Otherwise we would just work with the prior and likelihood solely in terms of $\theta_1$. The downside is that we need to specify a prior for $\theta_2$ once it is introduced into the model.
- In classical statistics, much effort needs to be made to deal with nuisance parameters; e.g., finding pivots, profile likelihoods, etc.
  - An example of this is the difference between the unbiased estimate of a variance component, e.g., $\frac{\sum e_i^2}{n-p}$ vs. the MLE: $\frac{\sum e_i^2}{n}$. The whole mess arises because we don't have a natural way of accounting for the fact that we've estimated the mean. We'll see that using the posterior mean (but not the posterior mode) accounts for the uncertainty in estimating the other parameters and therefore is inflated towards the unbiased estimate compared to the MLE.
- We'll now see all this in the normal example.

# Normal model (mean, variance unknown)

Data: $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ $\mu$ and $\sigma^2$ unknown

Prior: $\left.\begin{array}{l} \mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{array}\right\}$ Conjugate Prior

$\Rightarrow \mathcal{N}\text{-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2)$

Question: What is the joint posterior distribution of $(\mu, \sigma^2)$?

Answer: $\mathcal{N}\text{-Inv-}\chi^2(?, ?, ?, ?)$.

Solution:

$$P(\mu, \sigma^2|y) \propto P(y|\mu, \sigma^2)P(\mu, \sigma^2)$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{\sum_i (y_i - \bar{y})^2}_{(n-1)s_y^2} + n(\mu - \bar{y})^2\right]\right\} \times$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+1}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2\right]\right\}$$

# General normal model

$$\left(\tfrac{1}{\sigma^2}\right)^{\frac{n+\nu_0+1}{2}+1} \times \exp\left(-\tfrac{1}{2\sigma^2}\left(\underbrace{(n-1)s_y^2+\nu_0\sigma_0^2}_{\text{known}}+\boxed{n(\mu-\bar{y})^2+\kappa_0(\mu-\mu_0)^2}\right)\right)$$

Now $n(\mu-\bar{y})^2+\kappa_0(\mu-\mu_0)^2$ can be written as:

$$(n+\kappa_0)(\mu-\mu_n)^2+\frac{n\cdot\kappa_0}{(n+\kappa_0)}(\mu_0-\bar{y})^2$$

$$\text{where}\quad \mu_n=\frac{\kappa_0\mu_0+n\bar{y}}{n+\kappa_0}$$

So

$$P(\mu,\sigma^2|y)\propto$$

$$\left(\sigma^2\right)^{-\left(\frac{n+\nu_0+1}{2}+1\right)}\times$$

$$\exp\left\{-\tfrac{1}{2\sigma^2}\left[\tfrac{n\cdot\kappa_0}{(n+\kappa_0)}(\mu_0-\bar{y})^2+\nu_0\sigma_0^2+(n-1)s_y^2+(n+\kappa_0)(\mu-\mu_n)^2\right.\right.$$

Recall that a $\mathcal{N}$-Inv-$\chi^2\left(\mu_0,\frac{\sigma_0^2}{\kappa_0};\nu_0,\sigma_0^2\right)$ has density proportional to:

$$(\sigma^2)^{-\left(\frac{\nu_0+1}{2}+1\right)}\exp\left[\frac{-1}{2\sigma^2}\left(\nu_0\sigma_0^2+\kappa_0\left(\mu-\mu_0\right)^2\right)\right]$$

# General normal model

Let

$$\begin{aligned}
\mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{n+\kappa_0} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s_y^2 + \frac{n\cdot\kappa_0}{n+\kappa_0}(\mu_0 - \bar{y})^2\right]
\end{aligned}$$

$\uparrow$ prior SS; $\uparrow$ sample SS; $\uparrow$ discrepancy between prior mean and sample mean

Thus,

$$\mu, \sigma^2 | y \sim \mathcal{N}\text{-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2)$$

$$P(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\right)\right]$$

# General normal model

Question: What is the marginal posterior distribution of $\mu$?

$$
\begin{aligned}
P(\mu \mid y) &= \int_0^\infty P(\mu, \sigma^2 | y) d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \times \\
&\quad \exp\left\{ -\frac{1}{2\sigma^2}\left[\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\right]\right\} d\sigma^2 \\
&= \int_0^\infty \mathcal{F}_{\sigma^2}(?,?) d\sigma^2 \\
&\propto \left[\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\right]^{-\left(\frac{\nu_n+1}{2}\right)} \\
&\propto \left[1 + \frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right]^{-\left(\frac{\nu_n+1}{2}\right)}
\end{aligned}
$$

# General normal model

$$\therefore \mu \mid y \sim t_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right)$$
$$\nu_n = \nu_0 + n$$
$$\kappa_n = \kappa_0 + n$$

Note how Bayes has naturally accounted for the uncertainty in $\sigma^2$ through the $t$ posterior.
So that:

$$E(\mu|y) = \mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}; \quad \nu_n > 1$$
$$V(\mu|y) = \left(\frac{\nu_n}{\nu_n - 2}\right)\frac{\sigma_n^2}{\kappa_n}; \quad \nu_n > 2.$$

# General normal model

Question: What is the conditional posterior distribution of $\mu$ given $\sigma^2$?

$$P(\mu|\sigma^2, y) \propto P(\mu|\sigma^2)P(y|\mu, \sigma^2)$$

i.e., proportional to joint posterior with $\sigma^2$ held constant.

$$P(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\right)\right]$$

$$P(\mu \mid \sigma^2, y) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\kappa_n(\mu-\mu_n)^2\right)\right]$$

$$\mu \mid \sigma^2, y \sim \mathcal{N}\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$

Note that $\mu_n$ does not depend on $\sigma^2$.

# General normal model

Question: What is the marginal posterior distribution of $\sigma^2$?

$$
\begin{aligned}
P(\sigma^2|y) &\propto \int_{-\infty}^{\infty} (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\right)\right] d\mu \\
&\propto (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2\right)\right] \times \\
&\quad \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(\kappa_n(\mu-\mu_n)^2\right)\right] d\mu \\
&\propto (\sigma^2)^{-\left(\frac{\nu_n}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2\right)\right] \\
&\therefore \sigma^2|y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right).
\end{aligned}
$$

# General normal model

Summary:

$$
\begin{aligned}
y_1, \ldots, y_n &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \\
\mu|\sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0) \\
\sigma^2 &\sim \text{ Inv-}\chi^2(\nu_0, \sigma_0^2) \\
\Rightarrow \mu, \sigma^2 &\sim \mathcal{N}\text{-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \delta_0^2)
\end{aligned}
$$

Then

- $\mu, \sigma^2|y \sim \mathcal{N}\text{-Inv-}\chi^2(\mu_n, \frac{\sigma_n^2}{\kappa_n}; \nu_n, \sigma_n^2)$ where

$$
\begin{aligned}
\mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{n + \kappa_0} \\
\sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s_y^2 + \frac{n \cdot \kappa_0}{n + \kappa_0}(\mu_0 - \bar{y})^2\right] \\
\nu_n &= \nu_0 + n \\
\kappa_n &= n + \kappa_0
\end{aligned}
$$

- $\mu|\sigma^2, y \sim \mathcal{N}\left(\mu_n, \frac{\sigma^2}{n + \kappa_0}\right)$.
- $\sigma^2|y \sim \text{ Inv-}\chi^2(\nu_n, \sigma_n^2)$.
- $\mu|y \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$

# Predictive distribution

Derive the **Posterior Predictive Distribution** for $\tilde{y}$ if $\tilde{y} \sim N(\mu, \sigma^2)$ independently of $y_1, \ldots, y_n$.
We want $P(\tilde{y}|y)$ which is proportional to:

$$\begin{aligned}
\propto & \int_0^\infty \int_{-\infty}^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(\tilde{y}-\mu)^2\right] \times \\
& (\sigma^2)^{-\left(\frac{\nu_n+1}{2}+1\right)} \exp\left[\frac{-1}{2\sigma^2}\left(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\right)\right] d\mu d\sigma^2 \\
\propto & \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_n+2}{2}+1\right)} \exp\left[-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right] \times \\
& \left\{\int_{-\infty}^\infty \exp\left[\frac{-1}{2\sigma^2}\left(\kappa_n(\mu-\mu_n)^2 + (\tilde{y}-\mu)^2\right)\right] d\mu\right\} d\sigma^2
\end{aligned}$$

# Predictive distribution (cont'd)

The integral over $\mu$ involves the kernel of a Normal distribution with variance $\frac{\sigma^2}{\kappa_n+1}$ and a quadratic term involving $\tilde{y}$. This leaves:

$$
\begin{aligned}
& \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_n+2}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 + \frac{\kappa_n}{\kappa_n+1}(\mu_n-\tilde{y})^2\right)\right] \sigma d\sigma^2 \\
= & \int_0^\infty \text{Inv-}\chi^2\left(\sigma^2; \nu_n+1, \frac{1}{\nu_n+1}\left(\nu_n\sigma_n^2 + \frac{\kappa_n}{\kappa_n+1}(\mu_n-\tilde{y})^2\right)\right) d\sigma^2 \\
\propto & \left[1 + \frac{1}{\nu_n}\left(\frac{\kappa_n}{\kappa_n+1}\right)\frac{(\mu_n-\tilde{y})^2}{\sigma_n^2}\right]^{-\left(\frac{\nu_n+1}{2}\right)} \\
\therefore & \ \tilde{y} \mid y \sim t_{\nu_n}\left(\mu_n, \left(1+\frac{1}{\kappa_n}\right)\sigma_n^2\right).
\end{aligned}
$$

# Predictive distribution (cont'd)

Question: What is the posterior predictive distribution for $\tilde{y}(m \times 1)$?

Suppose now that $\tilde{y}(m \times 1)$ is a random vector and is taken to be independent of $y$

$$\tilde{y} \sim \mathcal{N}(\mu 1_m, \sigma^2 I_m)$$

Derive the posterior predictive distribution for $\tilde{y}$.

$$\left[\text{Still assuming } \begin{array}{r} \mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 \sim \text{ Inv-}\chi^2(\nu_0, \sigma_0^2) \end{array}\right]$$

$$\begin{aligned} P(\tilde{y}|y) &\propto \int_0^\infty \int_{-\infty}^\infty \left(\tfrac{1}{\sigma^2}\right)^{\frac{m}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{j=1}^m (\tilde{y}_j - \mu)^2\right] \\ &\times \left(\tfrac{1}{\sigma^2}\right)^{\frac{\nu_n+1}{2}+1} \exp\left[-\tfrac{1}{2\sigma^2}(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2)\right] d\mu d\sigma^2 \\ &= \int_0^\infty \int_{-\infty}^\infty \left(\tfrac{1}{\sigma^2}\right)^{\frac{m}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\left((m-1)s_{\tilde{y}}^2 + m\left(\mu - \bar{\tilde{y}}\right)^2\right)\right] \\ &\times \left(\tfrac{1}{\sigma^2}\right)^{\frac{\nu_n+1}{2}+1} \exp\left[-\tfrac{1}{2\sigma^2}(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2)\right] d\mu d\sigma^2 \end{aligned}$$

# Predictive distribution (cont'd)

Now, the exponent, apart from $-\frac{1}{2\sigma^2}$, can be written as:

$$m\left(\mu-\bar{\tilde{y}}\right)^2+\kappa_n(\mu-\mu_n)^2+(m-1)s_{\tilde{y}}^2+\nu_n\sigma_n^2$$

which, apart from $-\frac{1}{2\sigma^2}$, is:

$$=(m+\kappa_n)(\mu-\mu^*)^2+\frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}}-\mu_n\right)^2+(m-1)s_{\tilde{y}}^2+\nu_n\sigma_n^2$$

where $\mu^*=\dfrac{\kappa_n\mu_n+m\bar{\tilde{y}}}{\kappa_n+m}$. Thus, the posterior predictive distribution is proportional to:

$$\int_0^\infty\left(\frac{1}{\sigma^2}\right)^{\frac{m+\nu_n+1}{2}+1}$$
$$\times\ \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2+(m-1)s_{\tilde{y}}^2+\frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}}-\mu_n\right)^2\right)\right]$$
$$\times\ \left\{\int_{-\infty}^\infty\exp\left[-\frac{1}{2\sigma^2}(m+n\kappa_n)(\mu-\mu^*)^2\right]d\mu\right\}d\sigma^2.$$

# Predictive distribution (cont'd)

Integration with respect to $\mu$ yields:

$$\begin{aligned}
&\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{m+\nu_n+1}{2}+1}(\sigma)\times \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2+(m-1)s_{\tilde{y}}^2+\frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}}-\mu_n\right)^2\right)\right]d\sigma^2 \\
&= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{m+\nu_n}{2}+1}\times \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2+(m-1)s_{\tilde{y}}^2+\frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}}-\mu_n\right)^2\right)\right]d\sigma^2
\end{aligned}$$

Next,

$$\sigma^2 \sim \text{ Inv-}\chi^2\left(m+\nu_n, \frac{1}{m+\nu_n}\left(\nu_n\sigma_n^2+(m-1)s_{\tilde{y}}^2+\frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}}-\mu_n\right)^2\right)\right)$$

# Predictive distribution (cont'd)

So that integration with respect to $\sigma^2$ yields:

$$P(\tilde{y}|y) \propto \left[\nu_n\sigma_n^2 + (m-1)s_{\tilde{y}}^2 + \frac{m\cdot\kappa_n}{m+\kappa_n}\left(\bar{\tilde{y}} - \mu_n\right)^2\right]^{-\left(\frac{\nu_n+m}{2}\right)}$$

$$= \left[\nu_n\sigma_n^2 + Q(\tilde{y})\right]^{-\left(\frac{\nu_n+m}{2}\right)}$$

Now, $\bar{\tilde{y}} - \mu_n = \frac{1}{m}\sum_{j=1}^{m}(\tilde{y}_j - \mu_n) = \frac{1}{m}(\tilde{y} - \mu_n 1_m)^T 1_m$

and

$$(m-1)s_{\tilde{y}}^2 = \sum_{j=1}^{m}\left(\tilde{y}_j - \bar{\tilde{y}}\right)^2 = \tilde{y}^T\left(I_m - \frac{1}{m}1_m 1_m^T\right)\tilde{y}$$

$$= (\tilde{y} - \mu_n 1_m)^T\left[I_m - \frac{1}{m}1_m 1_m^T\right](\tilde{y} - \mu_n 1_m)$$

One can show easily that the last two quantities are equal.

# Predictive distribution (cont'd)

Recalling $\kappa_n = n + \kappa_0$, then $Q(\tilde{y})$ can be written as:

$$\frac{m(n+\kappa_0)}{m+n+\kappa_0}\frac{1}{m}(\tilde{y}-\mu_n 1_m)^T 1_m 1_m^T (\tilde{y}-\mu_n 1_m) + (y-\mu_n 1_m)^T \left[I_m - \frac{1}{m}1_m 1_m^T\right](\tilde{y}-\mu_n 1_m)$$

$$= (\tilde{y}-\mu_n 1_m)^T \underbrace{\left[I_m - 1_m\, 1_m^T \left(\frac{1}{m} - \frac{n+\kappa_0}{m(m+n+\kappa_0)}\right)\right]}_{M^{-1}} (\tilde{y}-\mu_n 1_m)$$

where

$$M^{-1} = I_m - 1_m\, 1_m^T \left(\frac{1}{m+n+\kappa_0}\right)$$

$$\therefore\ P(\tilde{y}|y) \propto \left[1 + \frac{(\tilde{y}-\mu_n 1_m)^T M^{-1}(\tilde{y}-\mu_n 1_m)}{\nu_n \sigma_n^2}\right]^{-\left(\frac{\nu_n+m}{2}\right)}$$

$$\therefore\ \tilde{y}|y \sim t_{\nu_n}(\mu_n 1_m, \sigma_n^2 M)$$

$$\left(\text{ Show } M = I_m + \frac{1}{(n+\kappa_0)}1_m 1_m^T\right)$$

# Predictive distribution (cont'd)

This is the posterior predictive distribution for $m$ observations from a multivariate Student-$t$.

$$
\begin{aligned}
E(\tilde{y}|y) &= \mu_n 1_m \qquad \text{for } \nu_n > 1; \text{ i.e., } n + \nu_0 > 1 \\
V(\tilde{y}|y) &= \left(\frac{\nu_n}{\nu_n - 2}\right) \sigma_n^2 M \quad \text{for } \nu_n > 2 \\
V(\tilde{y}_j|y) &= \left(\frac{\nu_n}{\nu_n - 2}\right) \sigma_n^2 \left(1 + \frac{1}{\kappa_n}\right).
\end{aligned}
$$

# Normal Data with a Non-Informative Prior

- What is the non-informative prior for $(\mu, \sigma^2)$? Assuming prior independence then Jeffreys' Rule yields $P(\mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)$

$$\therefore \text{ Likelihood : } y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$
$$\text{Prior : } \quad P(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

- What is the joint posterior distribution for $(\mu, \sigma^2)$?

$$P(\mu, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$
$$= \left(\tfrac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left[-\tfrac{1}{2\sigma^2}\left(n(\mu - \bar{y})^2 + (n-1)s_y^2\right)\right]$$

- What is the conditional posterior for $\mu|\sigma^2, y$?

$$P(\mu|\sigma^2, y) \propto \exp\left[-\frac{1}{2\sigma^2}n(\mu - \bar{y})^2\right]$$

$$\text{i.e.,} P(\mu|\sigma^2, y) = \mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

# Normal, non-informative prior

Marginal Posterior Distribution of $\sigma^2$

$$P(\sigma^2|y) \propto \int_{-\infty}^{\infty} \left(\tfrac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \underbrace{\exp\left[-\frac{1}{2\sigma^2}\left(n(\mu-\bar{y})^2+(n-1)s_y^2\right)\right]}_{\mu\sim\mathcal{N}(\bar{y},\sigma^2/n)} d\mu$$

$$\text{so that} \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \sigma \exp\left[-\frac{1}{2\sigma^2}(n-1)s_y^2\right]$$

$$= \left(\frac{1}{\sigma^2}\right)^{\frac{n-1}{2}+1} \exp\left\{-\frac{(n-1)s_y^2}{2\sigma^2}\right\}$$

$$\therefore \ \sigma^2|y \sim \text{Inv-}\chi^2(n-1, s_y^2)$$

# Normal, non-informative prior

Marginal Posterior Distribution of $\mu$:

$$P(\mu|y) \propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left[-\frac{1}{2\sigma^2}\left(n(\mu-\bar{y})^2+(n-1)s_y^2\right)\right] d\sigma^2$$

We see that
$\sigma^2 \sim \text{Inv}-\chi^2\left(n, \frac{1}{n}\left(n(\mu-\bar{y})^2+(n-1)s_y^2\right)\right)$ so that integrating over $\sigma^2$ involves introducing the term $(s^2)^{-\nu/2}$ where $\nu$ and $s^2$ are the parameters of the inverse chi-square.

$$\begin{aligned} P(\mu|y) &\propto \left[(n-1)s_y^2+n(\mu-\bar{y})^2\right]^{-\frac{n}{2}} \\ &= \left[1+\frac{n(\mu-\bar{y})^2}{(n-1)s_y^2}\right]^{-\frac{n}{2}} \end{aligned}$$

i.e., $\mu|y \sim t_{n-1}(\bar{y}, s_y^2/n)$
or $\frac{\mu-\bar{y}}{s_y/\sqrt{n}}|y \sim t_{n-1}$ Look familiar?

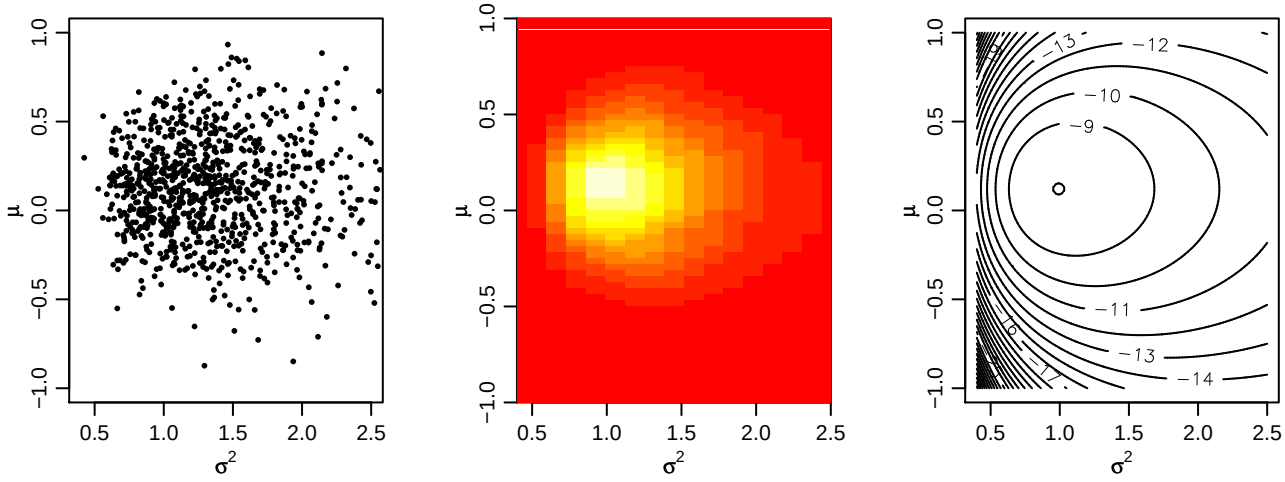
# Summary: Normal likelihood

1. $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$ and $\sigma^2 \sim$ Inv-$\chi^2(\nu_0, \sigma_0^2)$
   - A priori, $\mu$ and $\sigma^2$ are dependent (but they are uncorrelated)
   - A posteriori, $\mu$ and $\sigma^2$ are dependent (but they are uncorrelated)
2. $P(\mu, \sigma^2) \propto \sigma^{-2}$
   - A priori, $\mu$ and $\sigma^2$ are independent (and uncorrelated)
   - A posteriori, $\mu$ and $\sigma^2$ are dependent (and uncorrelated)
3. GCSR Section 3.4 discusses a semi-conjugate prior in which the prior for $\mu$ does not depend on $\sigma^2$.

# Dependence in the prior and posterior

There is an important general point here. Prior independence does not mean that quantities will be independent in the posterior.

- Posterior dependence in the non-informative prior for the normal model:




- Question: does prior dependence imply posterior dependence?
  - What about asymptotically?

# Brief Introduction to MCMC

MCMC=Markov chain Monte Carlo:

- Monte Carlo because we will rely on random draws from the (approximate) distribution
- Markov chain because samples will be drawn as a Markov chain

MCMC methods set up a Markov chain whose stationary distribution is the posterior.
We'll start with the simplest MCMC procedure, and the one that revolutionized Bayesian statistics in the early 1990s.

# The Gibbs sampler

Gibbs sampling: Suppose we have a collection of $k$ random variables denoted by $\theta = (\theta_1, \ldots, \theta_k)$. We assume that the **full conditional distributions**

$$\{P(\theta_i \mid \theta_{j:j\neq i}, y),\ i = 1, \ldots k\}$$

are available for sampling. Here, "available" means that samples may be generated by some method.
We do **not** require the one dimensional conditional distributions $P(\theta_i \mid \theta_{j:j\neq i}, y)$, $i = 1, \ldots, k$ to have a closed form, but we only need to be able to write them up to a normalizing constant.
Under mild conditions (see Besag, 1974) (we will investigate these further later), the one-dimensional conditional distributions **uniquely** determine the full joint distribution $P(\theta_1, \ldots, \theta_k|y)$, and hence all marginal distributions $P(\theta_i|y)$, $i = 1, \ldots, k$.
The algorithm proceeds as follows.

# Gibbs Sampling Algorithm

- (0) Suppose we have a set of **arbitrary** starting values $\{\theta_1^0, \ldots, \theta_k^0\}$.
- (1) Draw $\theta_1^1$ from $P(\theta_1 \mid \theta_2^0, \ldots, \theta_k^0, y)$
- (2) Draw $\theta_2^1$ from $P(\theta_2 \mid \theta_1^1, \theta_3^0, \ldots, \theta_k^0, y)$
- $\vdots$
- (k) Draw $\theta_k^1$ from $P(\theta_k \mid \theta_1^1, \ldots, \theta_{k-1}^1, y)$

This completes one iteration of the Gibbs sampler. Thus after 1 iteration, we have $(\theta_1^1, \ldots, \theta_k^1)$. After $T$ such iterations, we would obtain $(\theta_1^T, \ldots, \theta_k^T)$.
We are interested in sampling from the joint posterior distribution $P(\theta_1, \ldots, \theta_k \mid y)$. The Gibbs sampler requires draws only from each of the univariate conditional distributions $P(\theta_i \mid \theta_{j:j \neq i}, y)$.

# Example 1

## Example 1: Normal Likelihood

Let $y \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $n = 10$ and $y = \{78, 66, 65, 63, 60, 60, 58, 56, 52, 50\}$ so that $\bar{y} = 60.8$ and $s^2 = 63.5$. . Devise a Gibbs sampler to sample from the joint posterior distribution. Recall from earlier in the unit the conditional posterior distributions for $\mu$ and $\sigma^2$. Using $P(\mu, \sigma^2) \propto \sigma^{-2}$, we get the simple forms;

$$\mu \mid y, \sigma^2 \sim \mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

$$\sigma^2 \mid y, \mu \sim \mathcal{IG}\left(\frac{n}{2}, \frac{\sum_i (y_i - \mu)^2}{2}\right)$$

Sampling scheme:

- Start with arbitrary $(\mu^{(0)}, \sigma^{2\,(0)})$
- Sample $\mu^{(1)}$ from $P(\mu|\sigma^{2(0)}, y)$
- Sample $\sigma^{2\,(1)}$from $P(\sigma^2|\mu^{(1)}, y)$
- Repeat until we have $T$ iterations, sampling from $P(\mu|\sigma^{2(t-1)}, y)$ and $P(\sigma^2|\mu^{(t)}, y)$, saving each sample.

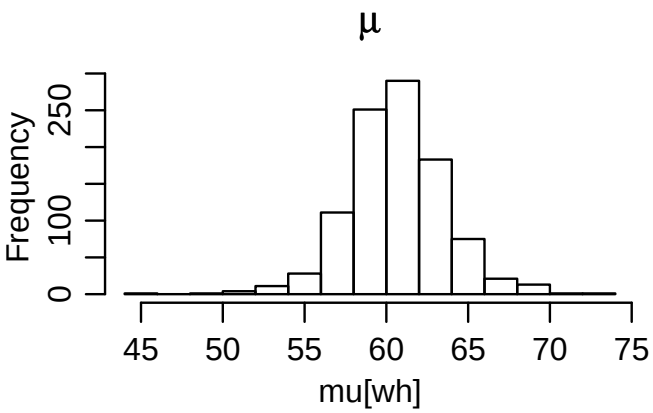
# Code example:

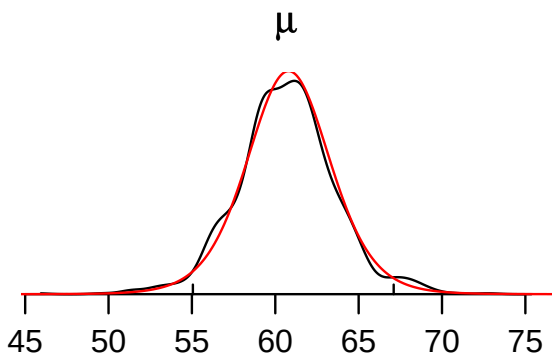
In this simple case, we do know the exact marginal posterior distributions of the parameters, specifically:

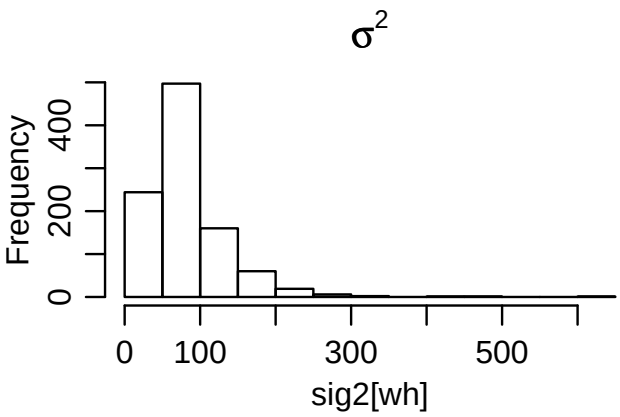
$$\sigma^2|y \sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{n-1}{2}s^2\right) \quad \text{and}$$

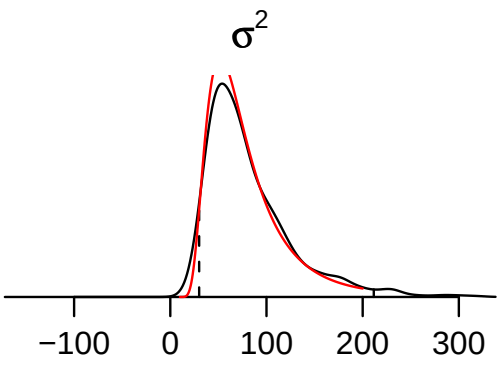
$$\mu|y \sim t_{n-1}\left(\bar{y}, \frac{s^2}{n}\right)$$

To simulate in R:


μ
Frequency
0
100
250
45
50
55
60
65
70
75
mu[wh]


μ
45
50
55
60
65
70
75


σ2
Frequency
0
200
400
0
100
300
500
sig2[wh]


σ2
−100
0
100
200
300

# Metropolis algorithm

Suppose we cannot sample from some of the conditional distributions directly.
The Metropolis algorithm is a popular way to deal with this and is often used within Gibbs sampling to deal with messy univariate conditionals.
Let $J_t(\theta^*|\theta^{t-1})$ be a density such that $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$ (e.g., $\theta_a \sim \mathcal{N}(\theta_b, c \cdot I)$). The function $J_t(\cdot|\cdot)$ is called a **candidate, jumping,** or **proposal** density. We generate values as follows.

1. Draw $\theta^* \sim J_t(\theta^*|\theta^{t-1})$, where $\theta^{t-1}$ is the current parameter value (current state of the chain).
2. Compute the acceptance ratio
$$r = \frac{P(\theta^*|y)}{P(\theta^{t-1}|y)} = \frac{P(y|\theta^*)P(\theta^*)}{P(y|\theta^{t-1})P(\theta^{t-1})}$$
3. Set
$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise (Rejection)} \end{cases}$$

# Comments on Metropolis

- Note that even when a proposal is not accepted, this counts as an iteration and this most recent value is repeated as part of the sample.
- Because Metropolis always accepts when the posterior density of the proposal is higher than the density of the current value, but also accepts with probability less than one when the new density is lower, one can see that it is a stochastic version of an optimization algorithm.
    - The ability to move to places in the parameter space with lower density is critical in exploring the posterior distribution.
- Normal distributions or $t$ distributions are common choices for the proposal distribution.
- Possible strategy for non-negative parameters?

# Multinomial Distribution

Revisit Binomial

$$\begin{aligned} y &\sim \text{Bin}(n, \theta) \\ \theta &\sim \text{Beta}(\alpha, \beta) \\ \Rightarrow \; P(\theta|y) &\propto \theta^{y+\alpha}(1-\theta)^{n-y+\beta} \end{aligned}$$

$\therefore \quad \theta|y \sim \text{Beta}(y+\alpha, n-y+\beta)$

Suppose there are $k$, rather than 2, possible classifications of an outcome. Then

$$P(y|\theta) = \frac{n!}{\prod_{j=1}^{k} y_j!}\theta_1^{y_1}, \theta_2^{y_2}, \ldots, \theta_k^{y_k} \qquad \theta_j \in [0,1] \;\; \forall j \qquad \sum_j y_j = n \quad \sum_j \theta_j = 1$$

and $y_j \in \{0, 1, 2, \ldots, n\}$

$\Rightarrow \quad y \sim \text{Multin}(n; \theta_1, \theta_2, \ldots, \theta_k)$

# Multinomial Distribution (2)

Conjugate prior distribution: Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$
$(\mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_k))$

$$P(\theta) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \qquad \alpha_j \geq 0$$

$$\alpha_0 \equiv \sum_{j=1}^{k} \alpha_j$$

$$\sum \theta_j = 1 \qquad 0 \leq \theta_j \leq 1$$

$$E(\theta_j) = \frac{\alpha_j}{\alpha_0} \qquad V(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$$

(Multivariate generalization of the Beta distribution)

Posterior Distribution

$$\begin{aligned} \text{if} \qquad & y \sim \text{Multin}(n, \theta_1, \ldots, \theta_k) \\ & \theta \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_k) \\ \text{then } & \theta|y \sim \text{???} \end{aligned}$$

$\therefore \quad E(\theta_j|y) = \frac{\alpha_j + y_j}{\alpha_0 + \sum y_i} =$ weighted average of prior mean $(\frac{\alpha_j}{\alpha_0})$ and MLE $(\frac{y_j}{\sum y_i})$.

# Example

Pre-Election Polling (GCSR, pg. 83)
Updated for McCain-Obama 2008:

- Gallup 10/31/08-11/02/08 poll of 3050 adults in the U.S. about support for the 2008 presidential candidates: (McCain, Obama, Other, No Opinion, Undecided)
  - Suppose the respondents were sampled randomly and among likely voters we find

$$y = [1281, 1617, 152]$$

$$\text{then } P(y) = \left( \frac{3050!}{1281! \cdot 1617! \cdot 152!} \right) \theta_1^{1281} \, \theta_2^{1617} \, \theta_3^{152}$$

↑ Probability of voting for McCain

Question: Does Obama have more support than McCain in the population?
Want to examine

$$P(\theta_2 - \theta_1 > 0 | y).$$

Take $\theta = (\theta_1, \theta_2, \theta_3) \sim \mathcal{D}(1, 1, 1)$

$$E(\theta_i) = \frac{1}{3} \qquad V(\theta) = \frac{2}{9 \times 4} = \frac{1}{18}$$

Posterior Distribution for $\theta$ $\theta|y \sim \mathcal{D}(1282, 1618, 153)$

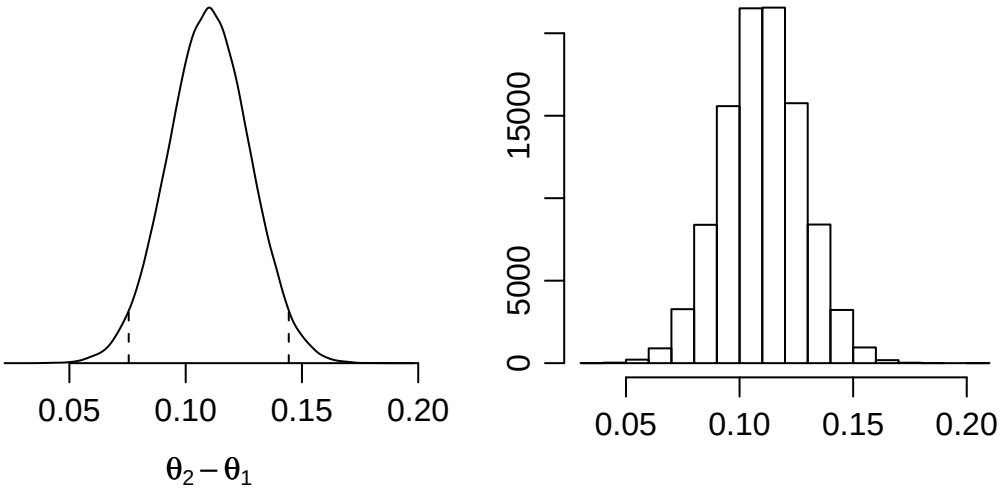
Can calculate $P(\theta_2 - \theta_1 > 0|y)$ using simulation.

(to sample from a Dirichlet, draw $x_i \sim \mathcal{G}(\alpha_i, 1)$)

Let $\theta_j = \frac{x_j}{\sum_i x_i}$.

(or use rdirichlet() from gtools library)

Then create a vector, $\theta_2 - \theta_1$, and compute posterior quantities




Posterior for the difference in support of Obama and McCain

Aside: Specifying the hyperparameters.
Suppose $k = 3$ and your prior means are:

$$E(\theta_1) = 0.10$$
$$E(\theta_2) = 0.30$$
$$E(\theta_3) = 0.60$$

The corresponding Dirichlet hyperparameters are:

$$E(\theta_1) = \frac{\alpha_1}{\alpha_0} = 0.10 \Rightarrow \alpha_1 = 0.10\alpha_0$$
$$\alpha_2 = 0.30\alpha_0$$
$$\alpha_3 = 0.60\alpha_0$$

$\therefore$ Need to specify $\sum_{j=1}^{3} \alpha_j = \alpha_0$ in order to get $\alpha_1$, $\alpha_2$, $\alpha_3$.
$\alpha_0 =$ "strength of prior information"

Why? If we increase $\alpha_0$, then we decrease the prior variance, e.g.,

$$V(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$$

$\therefore$ Can specify a prior Dirichlet distribution by

1. specifying $E(\theta_j)$
2. overall measure of strength of information

# Example: Predictive Distribution

Question: What is the predictive distribution for the actual election?
Final election results: Obama 69,456,897 (52.9%) vs. McCain 59,934,786 (45.7%).
Solution: Recall Beta-Binomial case $\Rightarrow$ posterior predictive distribution is Beta-Binomial.
$y \sim$ Beta-Bin$(n, \alpha, \beta)$ if

$$P(y|\alpha,\beta) = \frac{\Gamma(n+1)}{\Gamma(y+1)\,\Gamma(n-y+1)}\,\frac{\Gamma(\alpha+y)\,\Gamma(n+\beta-y)}{\Gamma(\alpha+\beta+n)}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}$$

for $n$ a positive integer; $\alpha > 0$, $\beta > 0$ and $y = 0, 1, 2, \ldots, n$.

$$E(y) = \frac{n\alpha}{(\alpha+\beta)}; \quad V(y) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

This is a mixture of binomials over a beta mixing distribution for the population proportions.
Result is the analogous one for the Dirichlet-multinomial

$$P(y|\alpha) = \frac{n!\Gamma(\alpha_0)}{\Gamma(n+\alpha_0)}\prod_i \frac{\Gamma(\alpha_i+y_i)}{\Gamma(\alpha_i)y_i!}$$

# Properties of the Dirichlet Distribution

- Dirichlet is a $(k-1)$ dimensional distribution because $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j \Rightarrow$ parameter space for $\theta$ is a $(k-1)$ dimensional simplex.
- Moments. Let

$$S = \left\{ \theta : \theta_j \geq 0,\ j = 1, 2, \ldots, k;\ \sum_{j=1}^{k} \theta_j = 1 \right\}$$

$$\begin{aligned}
E(\theta_1) &= \int_S \theta_1 \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1} d\theta \\
&= \frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)} \int_S \underbrace{\theta_1^{(\alpha_1+1)-1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1} \cdots \theta_k^{\alpha_k - 1}}\, d\theta \\
&= \frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)} \quad \frac{\Gamma(\alpha_1+1)\,\Gamma(\alpha_2)\,\Gamma(\alpha_3) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k+1)} \\
&= \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k + 1)} \cdot \frac{\Gamma(\alpha_1 + 1)}{\Gamma(\alpha_1)} \\
&= \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_k} \quad \text{using the fact that } \Gamma(t+1) = t\Gamma(t)
\end{aligned}$$

# Mode of the Dirichlet

- Mode. To find the mode, maximize the log density subject to the constraint $g(\theta) = \sum_{j=1}^{k} \theta_j - 1 = 0$.

  This can be accomplished by using Lagrange multipliers.

  i.e., $\log P(\theta) =$ constant $+ \sum_{j=1}^{k} (\alpha_j - 1) \log(\theta_j)$.

  We want to optimize $\log P(\theta) - \lambda g(\theta) = F(\theta, \lambda)$.

$$\frac{\partial F(\theta, \lambda)}{\partial \theta_j} = \frac{(\alpha_j - 1)}{\theta_j} - \lambda \quad j = 1, \ldots, k$$

$$\frac{\partial F(\theta, \lambda)}{\partial \lambda} = g(\theta) = \sum_{j=1}^{k} \theta_j - 1$$

# Mode of Dirichlet (2)

$$\frac{\partial F(\theta,\lambda)}{\partial \theta_j} = \frac{(\alpha-1)}{\theta_j} - \lambda \quad j = 1,\ldots,k$$

$$\frac{\partial F(\theta,\lambda)}{\partial \lambda} = g(\theta) = \sum_{j=1}^{k} \theta_j - 1$$

Setting the derivatives equal to zero, yields

$$\tilde{\theta}_j = \frac{\alpha_j - 1}{\lambda} \quad j = 1,2,\ldots,k$$

and because

$$\sum_{j=1}^{k} \tilde{\theta}_j = \sum_{j=1}^{k} \frac{(\alpha_j - 1)}{\lambda} = 1 \Rightarrow \lambda = \sum_{j=1}^{k} (\alpha_j - 1)$$

$\therefore$ the mode is $\tilde{\theta} = \left(\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k\right)$ where

$$\tilde{\theta}_i = \frac{\alpha_i - 1}{\sum(\alpha_j - 1)}.$$

Note: If all the $\alpha_j$'s are exactly 1, then the density is uniform over $S$. (If all $\alpha_j$'s $< 1$, then $\tilde{\theta}$ is a minimum).

# Properties of the Dirichlet (1)

- Relationship to the Gamma. Suppose $Y_1, \ldots, Y_k$ are independent standard gamma random variables with shape parameters $\alpha_1, \ldots, \alpha_k$ respectively.

  i.e., $Y_j \sim \mathcal{G}(\alpha_j, 1)$

  If we take
  $\theta_1 = \frac{Y_1}{\sum_{j=1}^{k} Y_j}, \ \theta_2 = \frac{Y_2}{\sum_{j=1}^{k} Y_j}, \ \ldots, \ \theta_k = \frac{Y_k}{\sum_{j=1}^{k} Y_j}$, then
  $\theta = (\theta_1, \ldots, \theta_k)$ has a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_k)$.
- Collapsing Rule. Suppose
  $\theta = (\theta_1, \ldots, \theta_k) \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Then the vector
  $$\theta^* = (\theta_1 + \theta_2, \theta_3, \ldots, \theta_k)$$
  is also Dirichlet with parameter
  $\alpha^* = (\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_k)$.

# Properties of the Dirichlet (2)

- Conditioning Rule. If we take a subvector of a Dirichlet and make it sum to one, then we get a Dirichlet. That is, if $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ then

$$\theta^* = \left( \frac{\theta_1}{\sum_{j=1}^{i} \theta_j}, \frac{\theta_2}{\sum_{j=1}^{i} \theta_j}, \ldots, \frac{\theta_i}{\sum_{j=1}^{i} \theta_j} \right) \quad i < k$$

  then $\theta^* \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_i)$.
  - Therefore, if $\theta_j =$ probability of falling into cell "$j$" of a contingency table, $j = 1, 2, \ldots, k$ then $\theta_j / \sum_{p=1}^{i} \theta_p$ is the conditional probability of falling into cell "$j$" given you are in cell 1 or cell 2 or ... or cell $i$.

# Choosing the hyperparameters

Non-Informative Priors. In the case of little prior information, take the $\alpha_j$'s equal to some small constant "$c$". Proposals:

1. $c = 0$: This is an improper prior distribution that leads to an improper posterior distribution if any of the cell counts are 0.
2. $c = 1/2$: This is Jeffreys' prior.
3. $c = 1$ : This prior is uniform on the cell probabilities over the simplex $S$. Posterior mode under this prior corresponds to the MLE.
4. $c = 3/2$ One commonly-used *ad hoc* procedure for smoothing a sparse contigency table (e.g., a table with a lot of zero counts) is to add "1/2" of an observation to each cell and then use maximum likelihood on the "augmented" table. This procedure corresponds to estimating $\theta$ by its posterior mode under a Dirichlet prior with $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 3/2$.

Aside: A Dirichlet prior with all the $\alpha_j$'s identical has the effect of smoothing the data towards a uniform table, e.g., one with

$$\theta_1 = \theta_2 = \cdots = \theta_k$$

# Multivariate Normal

Recall if $\underbrace{y}_{d\times 1}|\mu,\Sigma \sim \mathcal{N}_d(\mu,\Sigma)$ where $\mu$ is a $d\times 1$ vector and $\Sigma$ is a $d\times d$ symmetric positive definite matrix

then

$$P(y|\mu,\Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right).$$

Suppose $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}_d(\mu,\Sigma)$, then

$$P(y_1, y_2, \ldots, y_n|\mu,\Sigma)$$

$$\propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right]$$

$$= |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\,\text{tr}\,\left(\Sigma^{-1}S_0\right)\right)$$

$$\text{where } S_0 = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T$$

that is, $S_0$ = sum of squares matrix.

# Conjugate Prior for Precision Matrix

A **Wishart Distribution** is the natural generalization of the scaled $\chi^2$ distribution (or gamma). Recall,

1. if $z_1, z_2, \ldots, z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \Rightarrow \sum_{i=1}^{n} z_i^2 \sim \chi^2_n$.
2. if $y_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ then $\sum_{i=1}^{n} \left(\frac{1}{\sigma^2} y_i^2\right) \sim \chi^2_n$ or $\Sigma y_i^2 \sim \sigma^2 \chi^2_n$.

The Wishart generalizes this result to vectors.
Let $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}_d(0_d, \Sigma)$and let

$$y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \vdots & & & \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{bmatrix}$$

be a $(n \times d)$ data matrix with $n \geq d$.

Then $y^T y =$ the sum of squares and cross-product matrix (SSCP) is defined as:

$$\underset{(d \times d)}{y^T y} = \begin{bmatrix} \sum_i y_{i1}^2 & \sum_i y_{i1} y_{i2} & \cdots & \sum_i y_{i1} y_{id} \\ \sum_i y_{i2} y_{i1} & \sum_i y_{i2}^2 & \cdots & \sum_i y_{i2} y_{id} \\ \vdots & & \ddots & \\ \sum_i y_{id} y_{i1} & \sum_i y_{id} y_{i2} & \cdots & \sum_i y_{id}^2 \end{bmatrix} = \sum_{i=1}^{n} y_i y_i^T$$

then $y^T y \sim \text{Wishart}_n(\Sigma)$ $(\mathcal{W}_n(\Sigma))$ with degrees of freedom $n$ and scale $\Sigma$.

# Wishart Distribution

## Properties of a Wishart

- If $W \sim \mathcal{W}_n(\Sigma)$, then $E(W) = n\Sigma$
- If $W \sim \mathcal{W}_n(\Sigma)$, then $B^T W B \sim \mathcal{W}_n(B^T \Sigma B)$
  - If $B = \Sigma^{-1/2}$ (a matrix): Let $\Sigma^{\frac{1}{2}}$ be a square root matrix for $\Sigma$ such that $\left(\Sigma^{1/2}\right)^T \Sigma^{1/2} = \Sigma$.
  - (Typically we would take $\Sigma^{1/2}$ to be the upper-triangular Cholesky factor).
  - Let $\Sigma^{-1/2}$ be the inverse of $\Sigma^{1/2}$ s.t. $\Sigma^{1/2}\Sigma^{-1/2} = I$.
  - Then $\left(\Sigma^{-1/2}\right)^T W \Sigma^{-1/2} \sim \mathcal{W}_n(I)$ which is the standard Wishart distribution.
- Kernel of this density is:

$$P(W) \propto |W|^{\frac{(n-d-1)}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}W\right)\right\}$$

# Wishart Distribution (2)

- The Wishart involves $d(d+1)/2$ random variables
- The diagonal elements of a Wishart are scaled $\chi^2$ random variables: $W_{ii} \sim \Sigma_{ii}\chi^2_n$
- Additivity: if $W_j \sim \mathcal{W}(n_j, \Sigma)$ then

$$\sum_{j=1}^{k} W_j \sim \mathcal{W}(\sum_{j=1}^{k} n_j, \Sigma)$$

# Inverse-Wishart Distribution

Bayesians will typically work with the Inverse-Wishart rather than the Wishart itself. Now, if $W \sim \text{Wishart}_n(\Sigma)$, then $V = W^{-1}$ is said to have an Inverse-Wishart ($\mathcal{IW}$) distribution with degrees of freedom $= n$ and scale factor $\Sigma$.

The Kernel of the Inverse-Wishart density is:

$$P(V) \propto |V|^{-\left(\frac{n+d+1}{2}\right)} \exp\left\{-\frac{1}{2}\text{tr}\left(\Sigma V^{-1}\right)\right\}$$

$\Rightarrow V \sim \mathcal{IW}_n(\Sigma^{-1})$. $E(V) = (n-d-1)^{-1}\Sigma$.
This is the conjugate prior for covariance matrix in the MVN Distribution.

# Joint Conjugate Prior for Multivariate Normal

Normal-Inverse-Wishart arises from

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \kappa_0^{-1}\Sigma)$$
$$\Sigma \sim \mathcal{IW}_{\nu_0}(\Lambda_0^{-1})$$

where $\kappa_0$ and $\nu_0$ are scalars. Think of:

$\mu_0 =$ prior estimate of $\mu$

$\kappa_0 =$ number of prior 'observations' on which $\mu_0$ is based

$\frac{\Lambda_0}{\nu_0} =$ prior estimate of $\Sigma$

$\nu_0 =$ number of prior degrees of freedom

# Back to the MVN

Case I Unknown Mean (Known $\Sigma$). Let $y$ be $d$-dimensional

$$\begin{aligned} y &\overset{\text{iid}}{\sim} \mathcal{N}_d(\mu, \Sigma) \\ \mu &\sim \mathcal{N}_d(\mu_0, \Lambda_0) \\ \Rightarrow \mu|y, \Sigma &\sim \mathcal{N}_d(\mu_n, \Sigma_n) \end{aligned}$$

where

$$\begin{aligned} \Sigma_n &= \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1} \\ \mu_n &= \Sigma_n\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}\right) \end{aligned}$$

# MVN (cont'd)

Case II Unknown Mean and Variance

1. Joint conjugate prior $\Rightarrow \mathcal{N}-\mathcal{IW}$
   $(\mu_0, \Lambda_0/\kappa_0, \nu_0, \Lambda_0)$.
   The posterior is of course of the same form,
   $P(\mu, \Sigma|y) = \mathcal{N}-\mathcal{IW}(\mu_n, \Lambda_n/\kappa_n, \nu_n, \Lambda_n)$ where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T \\
S &= \sum_i (y_i-\bar{y})(y_i-\bar{y})^T
\end{aligned}$$

2. Multivariate Jeffreys' prior $\Rightarrow P(\mu, \Sigma) \propto |\Sigma|^{-\frac{d+1}{2}}$
   This is the limit of the conjugate prior as
   $\kappa_0 \to 0$, $\nu_0 \to -1$, $|\Lambda_0| \to 0$.

Recall joint conjugate prior distributions:

| Univariate Case | Multivariate Case |
|---|---|
| $y \sim \mathcal{N}(\mu, \sigma^2)$ | $y \sim \mathcal{N}(\mu, \Sigma)$ |
| $\mu \vert \sigma^2 \sim \mathcal{N}\left(\mu_0, \sigma^2 \kappa_0^{-1}\right)$ | $\mu \vert \Sigma \sim \mathcal{N}\left(\mu_0, \kappa_0^{-1} \Sigma\right)$ |
| $\sigma^2 \sim$ Inv-$\chi^2\left(\nu_0, \sigma_0^2\right)$ | $\Sigma \sim \mathcal{IW}_{\nu_0}\left(\Lambda^{-1}\right)$ |

In both cases, we take the prior (co)variance for the location parameter to be proportional to the data (co)variance (e.g., $\sigma^2$ or $\Sigma$). This implies that our prior knowledge about $\mu$ is equivalent to observing a fixed # of observations from the data distribution.
Note that in some cases, this is not ideal, particularly if our chosen $\mu_0$ is far from the actual $\mu$. Also, the Wishart prior has structural properties that are not ideal in some cases. Often we may want to avoid the conjugate prior.

# Linear regression model

Let $y = X\beta + \epsilon \qquad (X \text{ is full rank})$

$y = n \times 1$ vector of observations
$X = n \times k$ matrix of explanatory variables (rank $= k$)
$\epsilon = n \times 1$ vector of iid $\mathcal{N}(0, \sigma^2)$ errors
$\beta = k \times 1$ vector of unknown regression coefficients

$\therefore y|X, \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$

Likelihood for the Regression Model:

$$P(y|X, \beta, \sigma^2) \propto \left|\sigma^2 I_n\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y - X\beta)^T(\sigma^2 I_n)^{-1}(y - X\beta)\right\}$$
$$= (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

# Non-informative Prior

- Non-informative prior for $\beta$ and $\sigma^2$: $P(\beta, \sigma^2) \propto \sigma^{-2}$, i.e., uniform on $(\beta, \log \sigma)$.
- Derive the joint posterior:
  1. Let $\widehat{\beta} = (X^T X)^{-1} X^T y$
     $$\widehat{y} = X\widehat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\text{Hat Matrix}} y = Hy$$
     $$\widehat{\epsilon} = y - \widehat{y} = [I - H]y$$
  2. Rewrite $(y - X\beta)^T (y - X\beta) \Rightarrow$
     $$\begin{aligned} &= (y - \widehat{y} + \widehat{y} - X\beta)^T (y - \widehat{y} + \widehat{y} - X\beta) \\ &= (y - \widehat{y})^T (y - \widehat{y}) + 2(y - \widehat{y})^T (\widehat{y} - X\beta) + \\ &\quad (\widehat{y} - X\beta)^T (\widehat{y} - X\beta) \\ &= \hat{\epsilon}^T \widehat{\epsilon} + 2\widehat{\epsilon}^T (\widehat{y} - X\beta) + (\widehat{y} - X\beta)^T (\widehat{y} - X\beta) \end{aligned}$$

# Posterior distribution

But $\widehat{\epsilon}^T(\widehat{y} - X\beta) = 0$ because $\widehat{\epsilon}$ is orthogonal to $\widehat{y} - X\beta$

$$\begin{aligned}\therefore\ (y - X\beta)^T(y - X\beta) &= \widehat{\epsilon}^T\widehat{\epsilon} + (\widehat{y} - X\beta)^T(\widehat{y} - X\beta) \\ &= \widehat{\epsilon}^T\widehat{\epsilon} + (\widehat{\beta} - \beta)^T(X^TX)(\widehat{\beta} - \beta)\end{aligned}$$

where

$$\widehat{\beta} = (X^TX)^{-1}X^Ty$$

Thus the joint posterior can be written as:

$$P(\beta, \sigma^2|y, X) \propto (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \times \exp\left[-\frac{1}{2\sigma^2}\left(\widehat{\epsilon}^T\widehat{\epsilon} + (\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta})\right)\right]$$

(proper if $n > k$; e.g., at least as many data points as parameters $k$ and rank of $X = k$)

# Conditional posterior of $\beta$

What is the conditional posterior, $P(\beta|\sigma^2, y, X)$?

# Marginal posterior for $\sigma^2$

Note that $E(\sigma^2|y,X) = \frac{n-k}{n-k-2}\frac{\widehat{\epsilon}^T\widehat{\epsilon}}{n-k} = \frac{\widehat{\epsilon}^T\widehat{\epsilon}}{n-k-2}$. We see that Bayes, unlike ML, has accounted for the number of regression coefficients that must be estimated.

# Marginal posterior of $\beta$

Recall the conditional for $\beta$ and the marginal for $\sigma^2$.
We know that the marginal should be a multivariate-$t$.

$$P(\beta|x,y) \propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \times \exp\left\{-\frac{1}{2\sigma^2}\left(\widehat{\epsilon}^T\widehat{\epsilon} + \left(\widehat{\beta}-\beta\right)^T (X^TX)\left(\widehat{\beta}-\beta\right)\right)\right\} d\sigma^2$$

Now, we recognize this as involving

$$\sigma^2|\beta,y,X \sim \text{Inv-}\chi^2\left(n, \frac{\widehat{\epsilon}^T\widehat{\epsilon}+\left(\widehat{\beta}-\beta\right)^T(X^TX)\left(\widehat{\beta}-\beta\right)}{n}\right)$$

$$\therefore \; P\left(\beta|X,y\right) \propto \left(s^2\right)^{-n/2}$$
$$= \left[\widehat{\epsilon}^T\widehat{\epsilon} + \left(\widehat{\beta}-\beta\right)^T\left(X^TX\right)\left(\widehat{\beta}-\beta\right)\right]^{-\frac{n}{2}}$$

Now $\widehat{\epsilon}^T\widehat{\epsilon} = y^T\left[I-H\right]y = (n-k)s_\epsilon^2$, which does not involve $\beta$, so dividing by $(n-k)s_\epsilon^2$ and ignoring the $(n-k)s_\epsilon^2$ that appears outside the term raised to the power, we have

$$P\left(\beta|X,y\right) \propto \left[1 + \frac{\left(\widehat{\beta}-\beta\right)^T\left(X^TX\right)\left(\widehat{\beta}-\beta\right)}{(n-k)s_\epsilon^2}\right]^{-\left(\frac{(n-k)+k}{2}\right)}$$

which is a multivariate Student-$t$. That is:

$$\beta|X,y \sim t_{n-k}\left(\widehat{\beta}, s_\epsilon^2(X^TX)^{-1}\right).$$

# More on regression

- The predictive distribution for $\tilde{Y}(m \times 1)$ under $P(\beta, \sigma^2) \propto \sigma^{-2}$ is (see GCSR)

$$P(\tilde{Y}|\tilde{X}, Y) = t_{n-k}\left(\tilde{X}\hat{\beta}, s_\epsilon^2(I + \tilde{X}(X^T X)^{-1}\tilde{X}^T)\right)$$

- When the prior distribution for $\beta$ is not uniform, Bayesian regression takes the form of ridge regression, shrinking the estimate of $\beta$ towards the prior mean.
- As you saw in the first problem set, Jeffreys' prior for the regression model is $P(\beta, \sigma^2) \propto (\sigma^2)^{-(\frac{k}{2}+1)}$

# A computational trick

- Prior information can be treated as additional 'data points' as follows.
  - Suppose $\beta_j \sim \mathcal{N}(\beta_{j0}, \sigma_j^2)$. Now let's consider the prior as a distribution for $\beta_{j0}$ instead of $\beta_j$, which suggests that we treat $\beta_{j0}$ as an additional observation, appended to $y$ with covariate vector, $(0, \ldots, 1, \ldots 0)$, with a '1' in the jth position, and with error variance of $\sigma_j^2$. Then one can do a weighted regression with the added 'data point', which introduces heteroscedasticity.
  - See GCSR p. 383 for more details.

We'll come back to regression in the context of high-dimensional data where we'll talk about shrinkage priors, variable selection, etc., as well as in more general regression settings.

# Informative Prior Elicitation for the Linear Model

$$Y = X\beta + \epsilon \ ,$$
$$\epsilon \sim \mathcal{N}_n(0, \sigma^2 I) \ ,$$

where $Y$ is $n \times 1$, $X$ is $n \times k$, $\beta$ is $k \times 1$, and $\epsilon$ is $n \times 1$. The joint conjugate informative prior specification for $(\beta, \sigma^2)$ is

$$\beta \mid \sigma^2 \sim \mathcal{N}_k(\beta_0, \sigma^2 \Sigma_0) \ ,$$
$$\sigma^2 \sim \text{Inv-}\chi^2\left(\kappa_0, \sigma_0^2\right) \quad .$$

How do we specify $(\beta_0, \Sigma_0)$ and $(\kappa_0, \sigma_0^2)$ in practice?

1. The investigator has previous experience with such data, and thus obtains "estimates" of $(\beta_0, \Sigma_0, \kappa_0, \sigma_0^2)$ from literature or previous similar studies.
2. The investigator has raw data from a previous similar study with which to elicit $(\beta_0, \Sigma_0, \kappa_0, \sigma_0^2)$ .

# Using historical data

Suppose $D_0 = (n_0, Y_0, X_0) =$ *historical or raw data*. Take

$$\beta_0 = (X_0^T X_0)^{-1} X_0^T Y_0 \ , \quad X_0 \text{ is } n_0 \times k$$
$$\Sigma_0 = a_0 (X_0^T X_0)^{-1} \ , \quad \text{where } a_0 \text{ is a specified scalar}$$
$$\kappa_0 = n_0$$
$$\sigma_0^2 = s_{\epsilon,0}^2.$$

One may want to downweight the influence of the $n_0$ observations relative to the new observations by choosing $\kappa_0 < n_0$ and $a_0 > 1$.