STAT 206B
Chapter 7: Data Augmentation and Model Choice

Winter 2022

† Data Augmentation

- Data augmentation = adding auxiliary variables.

$$p_y(y \mid \theta) = \int p_{y|V}(y \mid \theta, V) p_V(V \mid \theta) dV$$

⋆⋆ Y is the variable of interest, but $p_y(y \mid \theta)$ not easy to sample from.

⋆⋆ V's are auxiliary variables that cannot be directly observed.

⋆⋆ $p_{y|V}(y \mid \theta, V)$ and $p_V(V \mid \theta)$ are easy to sample from.

- Gibbs sampler computations can often be simplified or convergence accelerated by data augmentation.

- **Example 1** Scale mixtures of normal distributions

Suppose $p(y)$ is a t-distribution with d.f ν, location parameter μ and scale parameter σ^2,

$$p(y) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\sigma^2\pi}\Gamma(\nu/2)} \left\{1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}}$$

⋆⋆ We may directly simulate y from the marginal distribution.

⋆⋆ Alternatively, we utilize the hierarchical structure,

$$p(y) = \int_{\mathbb{R}^+} p_{y|V}(y \mid \mu, V) p_V(V \mid \sigma^2) dV,$$

where $p_{y|V}(y \mid V) = \mathrm{N}(\mu, V)$ and $p_V(V \mid \sigma^2) = \mathrm{IG}(\nu/2, \sigma^2\nu/2) = \text{Inv-}\chi^2(\nu, \sigma^2)$.

- **Example 1** (contd) Consider the following model;

$$\begin{aligned} y_i \mid \nu, \mu, \sigma &\overset{iid}{\sim} t(\nu, \mu, \sigma^2), i = 1, \ldots, n, \\ \pi(\mu, \sigma^2) &\propto 1/\sigma^2, \end{aligned}$$

where degrees of freedom ν is fixed.

⋆⋆ The joint posterior is

$$p(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto \frac{1}{\sigma^2} \prod_{i=1}^{n} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left\{ 1 + \frac{1}{\nu} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\}^{-(\nu+1)/2}.$$

- **Example 1** (contd)

  ⋆⋆ Then the full conditionals are

$$\Rightarrow \quad p(\sigma^2 \mid \mu, y_1, \ldots, y_n) \propto (\sigma^2)^{-1-n/2} \prod_{i=1}^{n} \left\{1 + \frac{1}{\nu}\left(\frac{y_i - \mu}{\sigma}\right)^2\right\}^{-(\nu+1)/2}$$

$$\Rightarrow \quad p(\mu \mid \sigma^2 y_1, \ldots, y_n) \propto \prod_{i=1}^{n} \left\{1 + \frac{1}{\nu}\left(\frac{y_i - \mu}{\sigma}\right)^2\right\}^{-(\nu+1)/2}$$

$\Rightarrow$ Not convenient.

• **Example 1** (contd) We rewrite the model using the normal-scale mixture representation of a t-distribution;

$$\begin{aligned} y_i \mid \mu, V_i &\overset{indep}{\sim} \mathrm{N}(\mu, V_i), i = 1, \ldots, n, \\ V_i \mid \sigma^2 &\overset{iid}{\sim} \text{Inv-}\chi^2(\nu, \sigma^2), \\ \pi(\mu, \sigma^2) &\propto 1/\sigma^2, \end{aligned}$$

where ν is fixed.

⋆⋆ The joint posterior is

$$\begin{aligned} p(\mu, \sigma^2, V_i \mid y_1, \ldots, y_n) &\propto \frac{1}{\sigma^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V_i}} \exp\left\{ -\frac{(y_i - \mu)^2}{2V_i} \right\} \\ &\quad \times \prod_{i=1}^{n} \frac{(\nu\sigma^2/2)^{\nu/2}}{\Gamma(\nu/2)} V_i^{-\nu/2} \exp\left(-\frac{\nu\sigma^2}{2V_i} \right). \end{aligned}$$

- **Example 1 Model 2** (contd)

⋆⋆ Then the full conditionals are

$$p(\mu \mid -) \propto \exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2V_i}\right\}$$

$$\Rightarrow \quad \mu \mid - \sim \mathrm{N}\left((\sum_{i=1}^{n} \frac{1}{V_i})^{-1} \sum_{i} \frac{y_i}{V_i}, (\sum_{i=1}^{n} \frac{1}{V_i})^{-1} \right)$$

$$p(\sigma^2 \mid -) \propto (\sigma^2)^{-1+n\nu/2} \exp\left(-\sum_{i=1}^{n} \frac{\nu\sigma^2}{2V_i}\right)$$

$$\Rightarrow \quad \sigma^2 \mid - \sim \mathrm{Gamma}\left(\frac{n\nu}{2}, \sum_{i=1}^{n} \frac{\nu}{2V_i}\right)$$

- **Example 1 Model 2** (contd)
  - ⋆⋆ (contd) Then the full conditionals are

$$p(V_i \mid -) \propto V_i^{-\nu/2-1/2} \exp\left\{-\frac{(y_i-\mu)^2}{2V_i} - \frac{\nu\sigma^2}{2V_i}\right\}$$

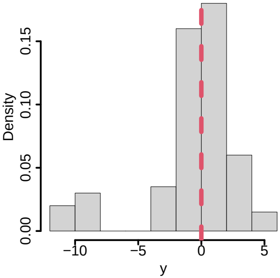
$$\Rightarrow \quad V_i \mid - \overset{indep}{\sim} \mathrm{IG}\left(\frac{\nu+1}{2}, \frac{(y_i-\mu)^2+\nu\sigma^2}{2}\right)$$

  - ⋆⋆ It is straightforward to perform the Gibbs sampler on V, μ and σ^2 in the augmented model.
  - ⋆⋆ More importantly, the simulations for μ and σ^2 under the augmented model represent the posterior distribution of μ and σ^2 under the original t model.

- Simulated data for **Example 1**

★★ Simulate data

$$y_i \overset{iid}{\sim} \mathrm{N}(0,4), i = 1, \ldots, 90, \quad \texttt{good obs}$$
$$y_i \overset{iid}{\sim} \mathrm{N}(-10,1), i = 91, \ldots, 100. \quad \texttt{bad obs}$$



- Simulated data for **Example 1** (contd)

Consider the following models

⋆⋆ Model A:

$$\begin{aligned} y_i \mid \mu, V_i &\overset{indep}{\sim} \mathrm{N}(\mu, V_i), i = 1, \ldots, n, \\ V_i \mid \sigma^2 &\overset{iid}{\sim} \text{Inv-}\chi^2(\nu, \sigma^2), \\ \pi(\mu, \sigma^2) &\propto 1/\sigma^2, \end{aligned}$$

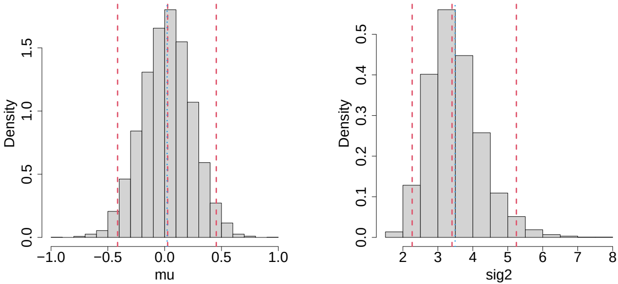
where $\nu = 3$ is fixed.

⋆⋆ Model B:

$$\begin{aligned} y_i \mid \mu, \sigma^2 &\overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2), i = 1, \ldots, n, \\ \pi(\mu, \sigma^2) &\propto 1/\sigma^2. \end{aligned}$$

- Model A:

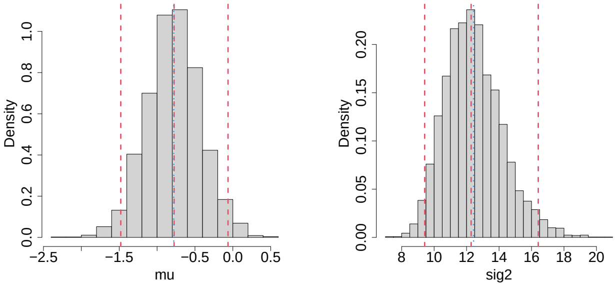
  - ⋆⋆ post. mean $\hat{\mu} = 0.022$ with 95% CI $(-0.414, 0.454)$
  - ⋆⋆ post. mean $\hat{\sigma}^2 = 3.493$ with 95% CI $(2.061, 4.578)$.



- Model B:

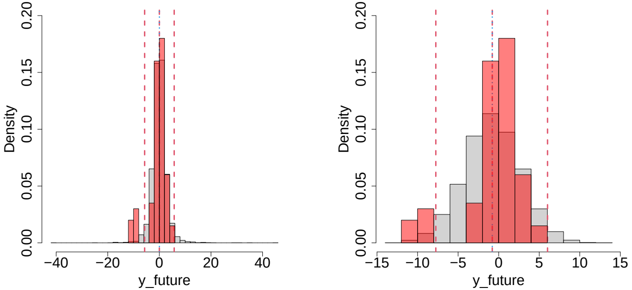
  ★★ post. mean $\hat{\mu} = -0.78$ with 95% CI $(-1.479, -0.064)$

  ★★ post. mean $\hat{\sigma}^2 = 12.429$ with 95% CI $(9.417, 16.415)$.



- Predictive distribution

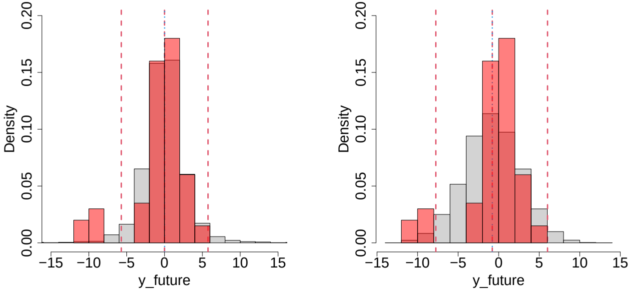
  - ⋆⋆ Model A: post. pred. mean $\hat{y}^{\text{NEW}} = 0.009$ with 95% posterior predictive interval $(-5.710, 5.747)$
  - ⋆⋆ Model B: post. pred. mean $\hat{y}^{\text{NEW}} = -0.789$ with 95% posterior predictive interval $(-7.750, 6.026)$



- Predictive distribution (contd)

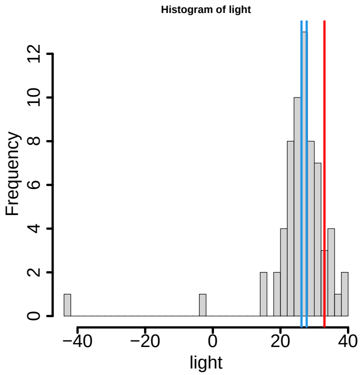
  ⋆⋆ Model A: post. pred. mean $\hat{y}^{\text{NEW}} = 0.009$ with 95% posterior predictive interval $(-5.710, 5.747)$

  ⋆⋆ Model B: post. pred. mean $\hat{y}^{\text{NEW}} = -0.789$ with 95% posterior predictive interval $(-7.750, 6.026)$



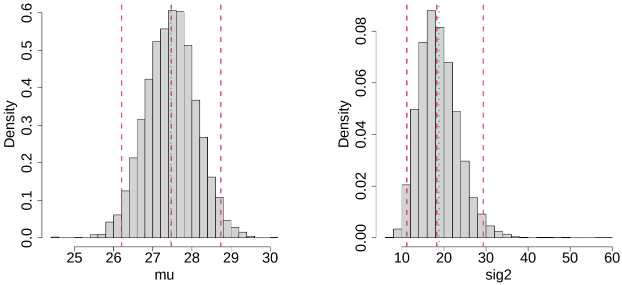
† Example (revisit): Estimating the speed of light (BDA p 66)

- Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters. He made 66 measurements. Consider the problem of estimating the speed of light.



† Example: Estimating the speed of light (contd)

- Use a t-model; Posterior summary of θ and σ^2



(a) θ (b) σ^2

† Example: Estimating the speed of light (contd)

- Use a t-model; Posterior summary of θ and σ^2

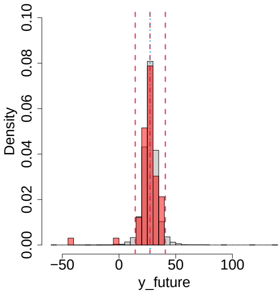
```
> print(round(quantile(SAVE_MCMC_sam$mu[inf_sam], prob=c(0.025, 0.5, 0.975)), 3))
  2.5%    50%  97.5%
26.203 27.472 28.740
> print(round(mean(SAVE_MCMC_sam$mu[inf_sam]), 3))
[1] 27.463
>
> print(round(quantile(SAVE_MCMC_sam$sig2[inf_sam], prob=c(0.025, 0.5, 0.975)), 3))
  2.5%    50%  97.5%
11.169 18.298 29.337
> print(round(mean(SAVE_MCMC_sam$sig2[inf_sam]), 3))
[1] 18.801
>
```

(a) *t*-model

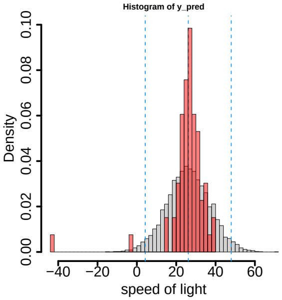
```
> ### summaries of the margianl posterior of theta
> post_m_th <- mean(my_th)
> post_sd_th <- sd(my_th)
> ci_th <- quantile(my_th, prob=c(0.025, 0.975))
> post_m_th
[1] 26.30754
> post_sd_th
[1] 1.355212
> ci_th
    2.5%    97.5%
23.66675 29.01357
>
> ### summaries of the margianl posterior of sig2
> post_m_sig2 <- mean(my_sig2)
> post_sd_sig2 <- sd(my_sig2)
> ci_sig2 <- quantile(my_sig2, prob=c(0.025, 0.975))
> post_m_sig2
[1] 119.0088
> post_sd_sig2
[1] 21.49393
> ci_sig2
     2.5%     97.5%
 84.55515 167.76078
```

† Example: Estimating the speed of light (contd)

- Use a t-model; Summary of the posterior predictive distribution of unobserved y



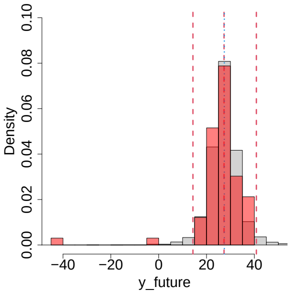
(a) t-model



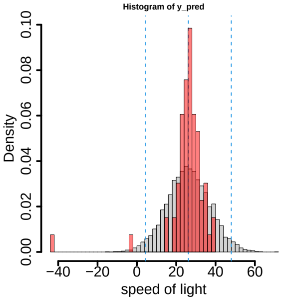
(b) normal

† Example: Estimating the speed of light (contd)

- Use a t-model; Summary of the posterior predictive distribution of unobserved y



(a) t-model



(b) normal

† Example: Estimating the speed of light (contd)

- Use a t-model; Summary of the posterior predictive distribution of unobserved y

```
> print(round(quantile(y_pred, prob=c(0.025, 0.5, 0.975)), 3))
  2.5%    50%  97.5%
14.321 27.317 40.842
> print(round(mean(y_pred), 3))
[1] 27.441
```

(a) *t*-model

```
> ###%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
> #%%%## preditive distribution
> y_pred <- rnorm(length(my_th), my_th, sqrt(my_sig2))
> mean(y_pred)
[1] 26.24271
> sd(y_pred)
[1] 11.01951
> quantile(y_pred, prob=c(0.025, 0.975))
     2.5%     97.5%
 4.343315 47.816016
>
```

(b) normal

- **Example 1** (contd) More examples?

  ⋆⋆ $y \mid \theta \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Be}(\alpha, \beta)$ (Beta-Binomial Mixture) where θ is an auxiliary variable.

  $\Rightarrow y \mid \alpha, \beta \sim \text{Beta-Binom}(n, \alpha, \beta)$

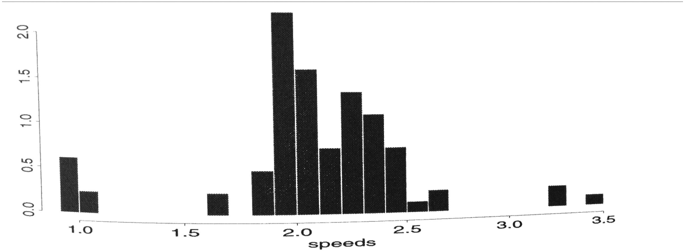
  See also **Example 6.3.4**.

  ⋆⋆ $y \mid \theta \sim \text{Poi}(\theta)$ and $\theta \sim \text{Gamma}(r, \frac{1-p}{p})$ (Gamma-Poisson Mixture) where θ is an auxiliary variable.

  $\Rightarrow y \mid r, p \sim \text{Neg-Binom}(r, p)$ where r: # of failures and p: success probability.

• **Example 7.1.2** (I changed a bit, especially notation) The dataset consists in 82 observations of galaxy velocities.

⋆⋆ Histogram of the galaxy dataset of Roeder (1992)



⋆⋆ For astrophysical reasons, the distribution of this dataset can be represented as a mixture of normal distributions. Suppose the number of components is k (fixed).

• **Example 7.1.2** (contd) Recall a mixture model with k components:

$$y_j \stackrel{iid}{\sim} \sum_{\ell=1}^{k} p_\ell \mathrm{N}(\mu_\ell, \sigma^2), \quad j = 1, \ldots, J(= 82).$$

The mixture model can be represented as follows;

⋆⋆ We introduce auxiliary variables $\lambda_j \in \{1, \ldots, k\}$.

⋆⋆ We assume $p(\lambda_j = \ell) = p_\ell$, independence between λ_j.

⋆⋆ Given λ_j, we write the distribution of y_j

$$\Rightarrow \quad y_j \mid \boldsymbol{\mu}, \sigma^2, \lambda_j = \ell \sim \mathrm{N}(\mu_\ell, \sigma^2).$$

- **Example 7.1.2** (contd) Let's develop the model further.
  - ⋆⋆ The likelihood

$$y_j \mid \lambda_j, \mu, \sigma \sim \mathrm{N}(\mu_{\lambda_j}, \sigma^2).$$

  - ⋆⋆ (prior) Let $p(\lambda_j = \ell \mid p) = p_\ell$, independence between λ_j.
  - ⋆⋆ (prior) Let $p = (p_1, \ldots, p_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ with fixed α.
  - ⋆⋆ (prior) Let $\mu_\ell \overset{iid}{\sim} \mathrm{N}(\bar{\mu}, \tau^2)$ with fixed $\bar{\mu}$ and τ^2 and $\sigma^2 \sim \mathrm{IG}(a, b)$ with fixed a and b.
  - ⇒ We have random parameters $\boldsymbol{\theta} = (\{\lambda_j\}_{j=1}^n, p, \{\mu_\ell\}_{\ell=1}^k, \sigma^2)$.
  - ⇒ Without λ_j, the likelihood evaluation becomes messy. But the likelihood evaluation conditional on λ_j is so simple! We will simulate $\boldsymbol{\theta}$ through MCMC.

- **Example 7.1.2** (contd) We first write the joint posterior of $\boldsymbol{\theta}$.

$$
\begin{aligned}
\pi(\boldsymbol{\theta} \mid y) \quad & \propto \quad \prod_{\ell=1}^{k} \pi(\mu_\ell)\ \pi(\sigma^2)\ \pi(p) \prod_{j=1}^{J} \pi(\lambda_j)\ p(y_j \mid \lambda_j, \mu, \sigma) \\
& \propto \quad \underbrace{\exp\left\{-\sum_{\ell=1}^{k} \frac{(\mu_\ell - \bar{\mu})^2}{2\tau^2}\right\}}_{\pi(\mu_\ell)} \underbrace{(\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right)}_{\pi(\sigma^2)} \\
& \qquad \times \underbrace{\prod_{\ell=1}^{k} p_\ell^{\alpha_\ell - 1}}_{\pi(p)} \underbrace{\prod_{j=1}^{J} p_{\lambda_j}}_{\pi(\lambda_j)} \underbrace{\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_j - \mu_{\lambda_j})^2}{2\sigma^2}\right\}}_{p(y_j \mid \lambda_j, \mu, \sigma)}
\end{aligned}
$$

⋆⋆ We use the Gibbs sampler to simulate θ. We first drive the full conditionals.

- **Example 7.1.2** (contd) the full conditionals

We use S_ℓ to denote the set of y_j having λ_j, $S_\ell = \{j : \lambda_j = \ell, j = 1, \ldots, J\}$. Also, let $\bar{y}_\ell = \frac{\sum_{j \in S_\ell} y_j}{|S_\ell|}$.

⋆⋆ μ_ℓ, $\ell = 1, \ldots, k$.

$$p(\mu_\ell \mid \lambda, \sigma^2, y) \propto \exp\left\{ -\frac{(\mu_\ell - \bar{\mu})^2}{2\tau^2} - \sum_{j \in S_\ell} \frac{(y_j - \mu_\ell)^2}{2\sigma^2} \right\}.$$

$$\Rightarrow \mu_\ell \mid \lambda, \sigma^2, y \sim \mathrm{N}\left((\tfrac{1}{\tau^2} + \tfrac{|S_\ell|}{\sigma^2})^{-1} (\tfrac{\bar{\mu}}{\tau^2} + \tfrac{\bar{y}_\ell}{\sigma^2/|S_\ell|}), (\tfrac{1}{\tau^2} + \tfrac{|S_\ell|}{\sigma^2})^{-1} \right).$$

⋆⋆ σ^2

$$p(\sigma^2 \mid \lambda, \mu, y) \propto (\sigma^2)^{-a-1} \exp(-\frac{b}{\sigma^2}) (\sigma^2)^{-J/2} \exp\left\{ -\sum_{j=1}^{J} \frac{(y_j - \mu_{\lambda_j})^2}{2\sigma^2} \right\}.$$

$$\Rightarrow \mu_\ell \mid \lambda, \sigma^2, y \sim \mathrm{IG}\left(a + \tfrac{J}{2}, b + \textstyle\sum_{j=1}^{J} \tfrac{(y_j - \mu_{\lambda_j})^2}{2} \right).$$

- **Example 7.1.2** (contd) the full conditionals

⋆⋆ $p = (p_1, \ldots, p_k)$

$$p(p \mid \lambda) \propto \prod_{\ell=1}^{k} p_\ell^{\alpha_\ell - 1} \prod_{\ell=1}^{k} p_\ell^{|S_\ell|}.$$

$\Rightarrow$ $p \mid \lambda \sim \text{Dir}(\alpha_1 + |S_1|, \ldots, \alpha_k + |S_k|)$.

⋆⋆ λ_j, $j = 1, \ldots, J$

$$p(\lambda_j = \ell \mid \mu, \sigma^2, y) \propto p_\ell \exp\left\{-\frac{(y_j - \mu_\ell)^2}{2\sigma^2}\right\}.$$

$\Rightarrow$ No standard form. So we sample on the grid of $(1, \ldots, k)$.

- **Example 7.1.2** (contd) Hyperparameters

  ⋆⋆ $k = 4$

  ⋆⋆ $\bar{\mu} = 2.08$ and $\tau^2 = 10$

  ⋆⋆ $a = 1$ and $b = 0.01$

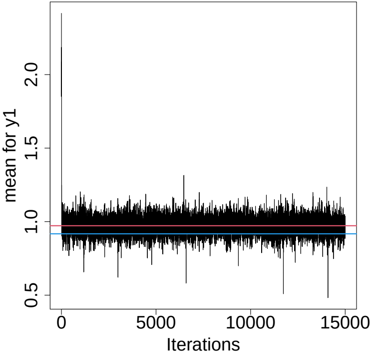
  ⋆⋆ $\alpha_1 = \ldots = \alpha_k = 1$

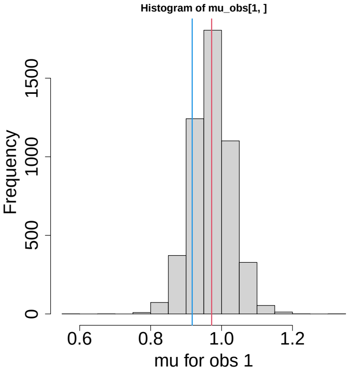
∗ run MCMC

```
n_sam <- 15000
n_burn_in <- 5000
n_thin <- 2
```

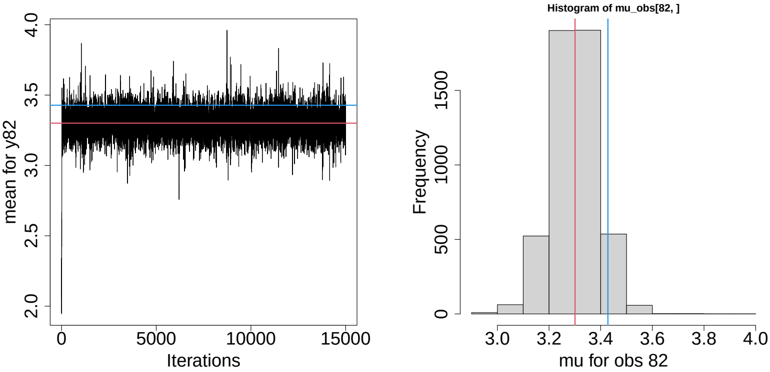
For details, see my code (posted on the course webpage).

- **Example 7.1.2** (contd) $y_1 = 0.9172$ (blue, smallest), posterior mean for y_1=0.9716 (red).

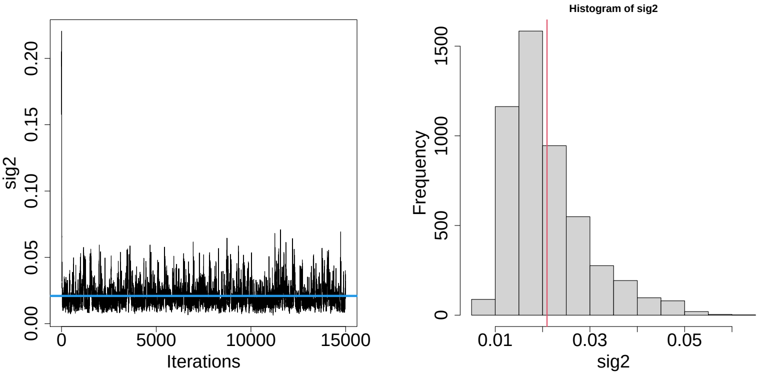




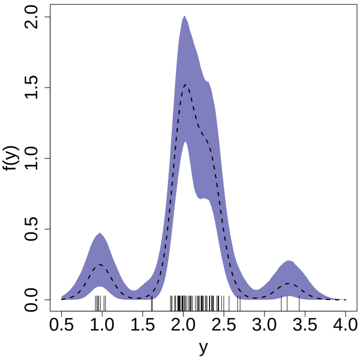
- **Example 7.1.2** (contd) $y_{82} = 3.4279$ (blue, largest), posterior mean for y_{82}=3.30 (red).



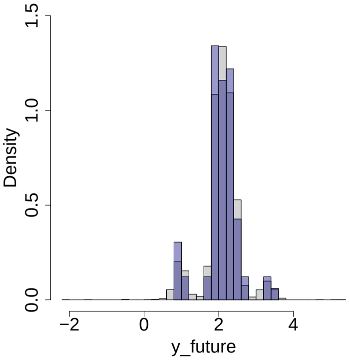
- **Example 7.1.2** (contd) σ^2



- **Example 7.1.2** (contd) σ^2



(a) $\hat{f}(y)$



(b) $p(y^{\text{new}} \mid \boldsymbol{y})$

† Model Choice - CR 7

- Suppose several models are in competition,

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \ldots, p\}.$$

- Model choice can be considered a special case of testing.

- The problem is not so simple since *while no model is true, several models may be appropriate*.

- **Example 7.1.1** Consider the data set relating the monthly unemployment rate with the monthly number of accidents in Michigan from 1978 to 1987. We may consider the following two models for the number of accidents N in a given month,

$$\begin{aligned}
\mathcal{M}_1 : N &\sim \text{Poi}(\lambda), \lambda > 0. \\
\mathcal{M}_2 : N &\sim \text{NB}(m, p), m > 0 \text{ and } p \in [0, 1].
\end{aligned}$$

• **Example 7.1.2:** The dataset consists in 82 observations of galaxy velocities. For astrophysical reasons, the distribution of this dataset can be represented as a mixture of normal distributions whose number of components k is unknown.

$$\mathcal{M}_i : y_j \overset{iid}{\sim} \sum_{\ell=1}^{i} p_{\ell i} \mathrm{N}(\mu_{\ell i}, \sigma_{\ell i}^2), \quad j = 1, \ldots, 82.$$

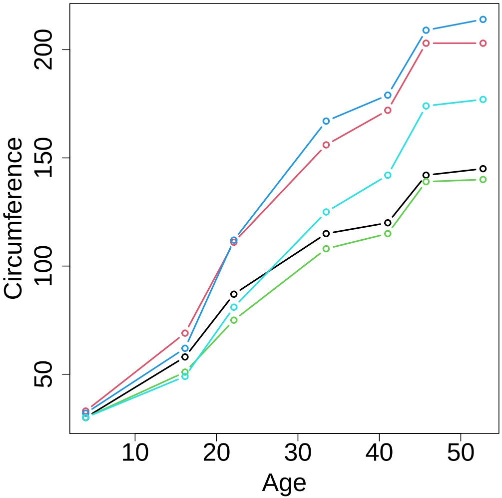
Here i varies between 1 and some arbitrary upper bound.

⋆⋆ Note that a k component model is a submodel of a $(k + p)$ component mixture by letting the the p remaining components have weights 0.

• **Example 7.1.3 (Model Selection):** For 5 orange tress, the growth of tree i is measured through the circumferences y_{it} at different times T_t, resulting in the data of Table 7.1.1.

	tree number				
time	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

- **Example 7.1.3 (Model Selection):**



• **Example 7.1.3** (contd): The models under scrutiny are $(i = 1, \ldots, 5, t = 1, \ldots, 7)$

$$
\begin{aligned}
\mathcal{M}_1 : y_{it} &\sim \mathrm{N}(\beta_{10} + b_{1i}, \sigma_1^2), \\
\mathcal{M}_2 : y_{it} &\sim \mathrm{N}(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2), \\
\mathcal{M}_3 : y_{it} &\sim \mathrm{N}\left(\frac{\beta_{30}}{1 + \beta_{31} \exp(\beta_{32} T_t)}, \sigma_3^2\right), \\
\mathcal{M}_4 : y_{it} &\sim \mathrm{N}\left(\frac{\beta_{40} + b_{4i}}{1 + \beta_{41} \exp(\beta_{42} T_t)}, \sigma_4^2\right),
\end{aligned}
$$

where the b_{ji}'s are random effects, distributed as $\mathrm{N}(0, \tau^2)$.

† Prior modeling for model choice: Testing problem

- Recall

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \ldots, p\}.$$

- Assign probability p_i to the models $\mathcal{M}_i$, $i \in I$.
- Given $\mathcal{M}_i$, we define priors $\pi_i(\theta_i)$, $\theta_i \in \Theta_i$.
- Compute the posterior probability of $\mathcal{M}_i$,

$$p(\mathcal{M}_i \mid x) = \frac{p_i m_i(x)}{\sum_j p_j m_j(x)} = \frac{p_i \int_{\Theta_i} f_i(x \mid \theta_i)\pi_i(\theta_i) d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x \mid \theta_j)\pi_j(\theta_j) d\theta_j}.$$

- Determine the model with the largest $p(\mathcal{M}_i \mid x)$.

† Some difficulties: Testing problem

- Require the construction of (π_i, p_i) for each $i \in I$.
- Cannot use improper priors for π_i.
- If some models are embedded into others, $\mathcal{M}_{i_0} \subset \mathcal{M}_{i_1}$, then there should be some coherence in the choice of π_{i_0} and π_{i_1}.

⋆⋆ **Example 7.1.3** (contd): Compare $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\begin{aligned} \mathcal{M}_1 : y_{it} &\sim \mathrm{N}(\beta_{10} + b_{1i}, \sigma_1^2), \\ \mathcal{M}_2 : y_{it} &\sim \mathrm{N}(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2). \end{aligned}$$

† Bayes factors (CR 7.2.2)

- Recall

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \ldots, p\}.$$

- Bayes factors

$$\begin{aligned} B_{12} &= \frac{P(\mathcal{M}_1 \mid x)}{P(\mathcal{M}_2 \mid x)} \Big/ \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \\ &= \frac{m_1(x)}{m_2(x)} = \frac{\int_{\Theta_1} f_1(x \mid \theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_2} f_2(x \mid \theta_2)\pi_2(\theta_2)d\theta_2}. \end{aligned}$$

- The model ordering is transitive; $B_{ij} = B_{ik}B_{kj}$ for $(\mathcal{M}_i, \mathcal{M}_j)$.

- Improper priors cannot be used.

† Bayesian Deviance (CR 7.2.4)

$$\text{Deviance } D(\theta) = -2\log(f(x \mid \theta)).$$

- An important role in statistical model comparison
- Proportional to MSE, $1/n \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$ if the model is normal with constant variance.
- It favors higher dimensional models. $\Rightarrow$ Introduce a penalized deviance.
- For more, also see Chapter 6 of Bayesian Analysis.

† Deviance Information Criterion (DIC)

$$\begin{aligned} \mathrm{DIC} &= \mathrm{E}[D(\theta) \mid x] + p_D \\ &= \mathrm{E}[D(\theta) \mid x] + \{\mathrm{E}[D(\theta) \mid x] - D(\mathrm{E}[\theta \mid x])\} \\ &= 2\mathrm{E}[D(\theta) \mid x] - D(\mathrm{E}[\theta \mid x]). \end{aligned}$$

⋆⋆ $\mathrm{E}[D(\theta) \mid x]$: a measure of fit.

⋆⋆ p_D: a measure of model complexity (also called the effective number of parameters)

- Suggested as a criterion of model fit when the goal is to pick a model with best out-of-sample predictive power.
- Bayesian alternative to AIC and BIC.
- Allow for improper priors
- The smaller the value of DIC, the better the model

† Deviance Information Criterion (DIC) - contd

- $\mathrm{DIC} = 2\mathrm{E}[D(\theta) \mid x] - D(\mathrm{E}[\theta \mid x])$, where $D(\theta) = -2\log(f(x \mid \theta))$.

- Given MCMC sample of $\theta^{(\ell)}$, we estimate DIC

$$
\begin{aligned}
\mathrm{DIC} &\approx 2\hat{D}(\theta) - D(\hat{\theta}) \\
&= \frac{2}{m}\sum_{\ell=1}^{m} D(\theta^{(\ell)}) - D(\hat{\theta}),
\end{aligned}
$$

where $\hat{\theta}$ is a point estimate for θ such as the mean of the posterior simulations.

- **Example** ((PH Chapter 9) Oxygen uptake: Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimen on oxygen uptake.

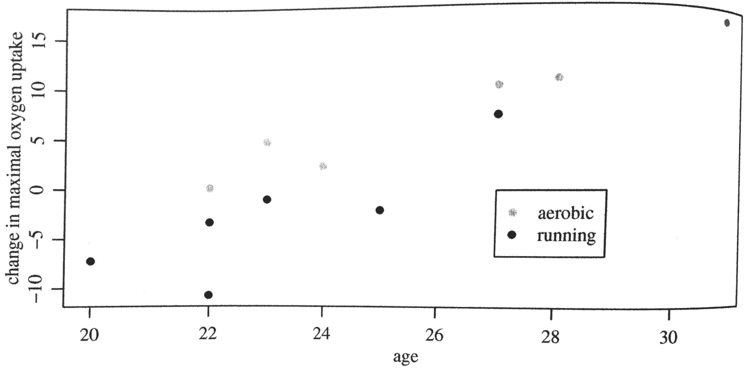
  ⋆⋆ Six are randomly assigned to a 12-week flat-terrain running program, and the remaining six to a 12-week step aerobics program.

  ⋆⋆ The maximum oxygen uptake of each subject was measured

  ⋆⋆ Age is expected to affect the change in maximal uptake.

  ⋆⋆ Goal: want to understand how a subject's change in maximal oxygen uptake may depend on the programs.

- **Example** Oxygen uptake (contd)



- **Example** Oxygen uptake (contd)

Consider the following covariates

⋆⋆ $x_{i,1} = 0$ if subject i is on the running program, 1 if on aerobic.

⋆⋆ $x_{i,2} =$ age of subject i

⋆⋆ $x_{i,3} = x_{i,1} \times x_{i,2}$: interaction effects

- **Example** Oxygen uptake (contd)

Consider four regression model;

⋆⋆ Model 1:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Model 2:

$$Y_i = \beta_0 + \beta_2 x_{i,2} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_2)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Model 3:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

- **Example** Oxygen uptake (contd)

Consider four regression model;

⋆⋆ Model 4:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Under each model, we assume

$$\pi(\boldsymbol{\beta}, \sigma^2) = \mathrm{N}_p(\boldsymbol{\beta}_0, \Sigma_0)\mathrm{IG}(\nu/2, s_0^2/2),$$

where p denotes the number of unknown covariates. Let $\boldsymbol{\beta}_0$, Σ_0, ν and s_0^2 fixed (HW#3-Q10(b)).

- **Example** Oxygen uptake (contd)

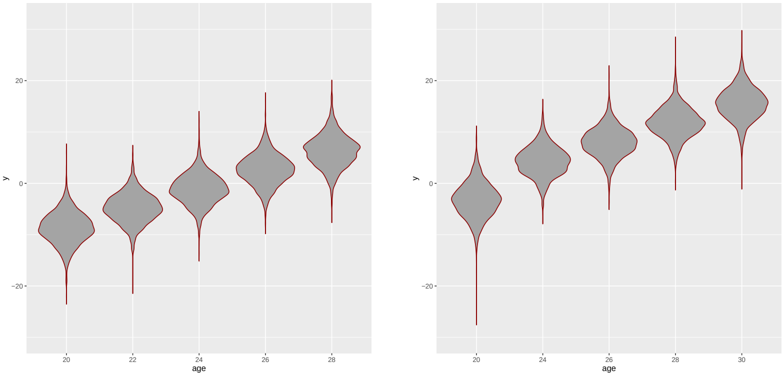
  ⋆⋆ Posterior mean estimates of the parameters;

Model	β_0	β_1	β_2	β_3	σ^2	BIC
M1	-2.78	10.34			35.24	233.42
M2	-52.76		2.25		13.04	197.14
M3	-46.22	5.43	1.88		7.34	174.06
M4	-50.56	12.52	2.06	-0.289	7.86	175.79

  ⋆⋆ Under M3, the 95% CIs are (-59.39, -32.36), (1.95, 8.97), and (1.29, 2.45) for β_0, β_1 and β_3, respectively, and (3.135, 16.75) for σ^2

- **Example** Oxygen uptake (contd)

  ★★ Posterior predictive distributions under M3



- **Example** Oxygen uptake (contd)

Consider a regression model;

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i.$$

⋆⋆ We suspect some of the regression coefficients are potentially equal to zero.

⋆⋆ Consider a mixture prior for β_j, $j = 1, \ldots, 3$;

$$\pi(\beta_j \mid p_j, \bar{\beta}_j, \tau_j^2) = p_j 1(\beta = 0) + (1 - p_j)\mathrm{N}(\bar{\beta}_j, \tau_j^2).$$

We may further consider priors for p_j, $\bar{\beta}_j$ and τ_j^2

⋆⋆ Specify priors for β_0 and σ^2.