Statistical Methods for Analysis with Missing Data

Lecture 9: Gibbs sampling, ignorability under Bayesian inference, data augmentation

Mauricio Sadinle

Department of Biostatistics
UNIVERSITY of WASHINGTON

Previous Lecture

Introduction to Bayesian inference:

- Alternative framework for deriving inferences from data
- Philosophical motivation: inclusion of prior belief or knowledge, uncertainty quantification in terms of distributions for parameters
- Practical motivation: convenient in some problems, might lead to good frequentist performance
- Complex problems become computationally involved – posterior distribution needs to be approximated (e.g., Gibbs sampling)

Today's Lecture

- Gibbs sampling to sample from complex distributions, including posterior distributions
- Bayesian inference with missing data, the concept of *ignorability*
- Data augmentation to handle missing data in the Bayesian framework

Outline

Gibbs Sampling

Bayesian Inference with Missing Data Under Ignorability

Data Augmentation

Gibbs Sampling

- Consider a distribution with density

$$p(z_1, z_2, \ldots, z_k)$$

- Say you want to sample from it but you don't know how
- Say the conditionals are easy to sample from, e.g., each

$$p(z_1 \mid z_2, z_3, \ldots, z_k)$$

$$p(z_2 \mid z_1, z_3, \ldots, z_k)$$

$$\vdots$$

$$p(z_k \mid z_1, z_2, \ldots, z_{k-1})$$

corresponds to a known and commonly used distribution

Gibbs Sampling

- Consider a distribution with density

$$p(z_1, z_2, \ldots, z_k)$$

- Say you want to sample from it but you don't know how
- Say the conditionals are easy to sample from, e.g., each

$$p(z_1 \mid z_2, z_3, \ldots, z_k)$$

$$p(z_2 \mid z_1, z_3, \ldots, z_k)$$

$$\vdots$$

$$p(z_k \mid z_1, z_2, \ldots, z_{k-1})$$

corresponds to a known and commonly used distribution

Gibbs Sampling

- Fix initial values $(z_2^{(0)}, z_3^{(0)}, \ldots, z_k^{(0)})$
- At iteration t, draw

$$z_1^{(t)} \sim p(z_1 \mid z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$z_2^{(t)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$\vdots$$

$$z_k^{(t)} \sim p(z_k \mid z_1^{(t)}, z_2^{(t)}, \ldots, z_{k-1}^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(z_1^{(t)}, z_2^{(t)}, \ldots, z_k^{(t)}) \sim p(z_1, z_2, \ldots, z_k)$$

- To learn the theory behind this you'll need to take a course on Bayesian statistics (or just learn it on your own![1])

[1] https://doi.org/10.1080/00031305.1992.10475878

Gibbs Sampling

- Fix initial values $(z_2^{(0)}, z_3^{(0)}, \ldots, z_k^{(0)})$
- At iteration t, draw

$$z_1^{(t)} \sim p(z_1 \mid z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$z_2^{(t)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$\vdots$$

$$z_k^{(t)} \sim p(z_k \mid z_1^{(t)}, z_2^{(t)}, \ldots, z_{k-1}^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(z_1^{(t)}, z_2^{(t)}, \ldots, z_k^{(t)}) \sim p(z_1, z_2, \ldots, z_k)$$

- To learn the theory behind this you'll need to take a course on Bayesian statistics (or just learn it on your own![1])

[1] https://doi.org/10.1080/00031305.1992.10475878

Gibbs Sampling

- Fix initial values $(z_2^{(0)}, z_3^{(0)}, \ldots, z_k^{(0)})$
- At iteration t, draw

$$z_1^{(t)} \sim p(z_1 \mid z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$z_2^{(t)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$\vdots$$

$$z_k^{(t)} \sim p(z_k \mid z_1^{(t)}, z_2^{(t)}, \ldots, z_{k-1}^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(z_1^{(t)}, z_2^{(t)}, \ldots, z_k^{(t)}) \sim p(z_1, z_2, \ldots, z_k)$$

- To learn the theory behind this you'll need to take a course on Bayesian statistics (or just learn it on your own![1])

[1] https://doi.org/10.1080/00031305.1992.10475878

Gibbs Sampling

- Fix initial values $(z_2^{(0)}, z_3^{(0)}, \dots, z_k^{(0)})$
- At iteration t, draw

$$z_1^{(t)} \sim p(z_1 \mid z_2^{(t-1)}, z_3^{(t-1)}, \dots, z_k^{(t-1)})$$

$$z_2^{(t)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t-1)}, \dots, z_k^{(t-1)})$$

$$\vdots$$

$$z_k^{(t)} \sim p(z_k \mid z_1^{(t)}, z_2^{(t)}, \dots, z_{k-1}^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(z_1^{(t)}, z_2^{(t)}, \dots, z_k^{(t)}) \sim p(z_1, z_2, \dots, z_k)$$

- To learn the theory behind this you'll need to take a course on Bayesian statistics (or just learn it on your own![1])

[1] https://doi.org/10.1080/00031305.1992.10475878

Gibbs Sampling

- Fix initial values $(z_2^{(0)}, z_3^{(0)}, \ldots, z_k^{(0)})$
- At iteration t, draw

$$z_1^{(t)} \sim p(z_1 \mid z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$z_2^{(t)} \sim p(z_2 \mid z_1^{(t)}, z_3^{(t-1)}, \ldots, z_k^{(t-1)})$$

$$\vdots$$

$$z_k^{(t)} \sim p(z_k \mid z_1^{(t)}, z_2^{(t)}, \ldots, z_{k-1}^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(z_1^{(t)}, z_2^{(t)}, \ldots, z_k^{(t)}) \sim p(z_1, z_2, \ldots, z_k)$$

- To learn the theory behind this you'll need to take a course on Bayesian statistics (or just learn it on your own![1])

[1] https://doi.org/10.1080/00031305.1992.10475878

Example: Bhattacharyya's Distribution

Consider real-valued random variables X and Y having a joint distribution with density[2]

$$p_{X,Y}(x,y) = \exp\left\{ \begin{bmatrix}1, x, x^2\end{bmatrix} \begin{bmatrix} m_{00}, m_{01}, m_{02} \\ m_{10}, m_{11}, m_{12} \\ m_{20}, m_{21}, m_{22} \end{bmatrix} \begin{bmatrix} 1 \\ y \\ y^2 \end{bmatrix} \right\},$$

where either

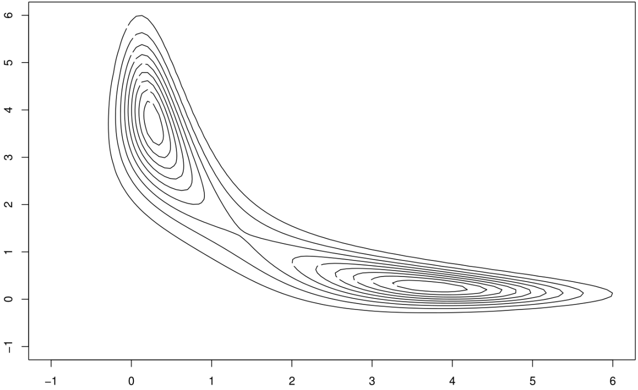
(a) $m_{22} = m_{21} = m_{12} = 0;\ m_{20}, m_{02} < 0;\ m_{11}^2 < 4m_{20}m_{02};$

(b) $m_{22} < 0,\ 4m_{22}m_{02} > m_{12}^2,\ 4m_{22}m_{20} > m_{21}^2.$

m_{00} is determined by the other m_{ij}'s so that $p_{X,Y}$ integrates to 1.

[2] Distribution credited to Anil Kumar Bhattacharyya, who was a professor at the Indian Statistical Institute. See, e.g., https://projecteuclid.org/download/pdf_1/euclid.ss/1009213728

Example: Bhattacharyya's Distribution



Example: Bhattacharyya's Distribution

From $p_{X,Y}(x,y)$ it is easy to see that

$$p_{X|y}(x|y) \propto \frac{1}{\sigma_X(y)} \exp\left\{-\frac{[x-\mu_X(y)]^2}{2\sigma_X^2(y)}\right\},$$

where

$$\mu_X(y) = -\frac{m_{10} + m_{11}y + m_{12}y^2}{2(m_{20} + m_{21}y + m_{22}y^2)},$$

and

$$\sigma_X^2(y) = -\frac{1}{2(m_{20} + m_{21}y + m_{22}y^2)}$$

Example: Bhattacharyya's Distribution

And analogously, it is easy to see that

$$p_{Y|x}(y|x) \propto \frac{1}{\sigma_Y(x)} \exp\left\{-\frac{[y-\mu_Y(x)]^2}{2\sigma_Y^2(x)}\right\},$$

where

$$\mu_Y(x) = -\frac{m_{01}+m_{11}x+m_{21}x^2}{2(m_{02}+m_{12}x+m_{22}x^2)},$$

and

$$\sigma_Y^2(x) = -\frac{1}{2(m_{02}+m_{12}x+m_{22}x^2)}$$

Example: Bhattacharyya's Distribution

- In fact, Bhattacharyya's distribution characterizes *all* bivariate distributions with normal conditionals[3]
- Gibbs sampler to draw from $p_{X,Y}$ is easy to implement! (R session 3)

[3] Arnold, Castillo and Sarabia (Statistical Science, 2001): `https://projecteuclid.org/download/pdf_1/euclid.ss/1009213728`

Gibbs Sampling for Bayesian Inference

For Bayesian inference we work with the posterior

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

- This expression might not be available in closed form
- Computing functionals of interest $E[f(\theta) \mid \mathbf{z}]$ might be complicated
- Idea: sample from $p(\theta \mid \mathbf{z})$ and evaluate functionals of interest via Monte Carlo, i.e., draw $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)} \sim p(\theta \mid \mathbf{z})$ and approximate

$$E[f(\theta) \mid \mathbf{z}] \approx \frac{1}{m}\sum_{t=1}^{m} f(\theta^{(t)})$$

- Problem: we might not know how to sample from $p(\theta \mid \mathbf{z})$

Gibbs Sampling for Bayesian Inference

For Bayesian inference we work with the posterior

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

- This expression might not be available in closed form
- Computing functionals of interest $E[f(\theta) \mid \mathbf{z}]$ might be complicated
- Idea: sample from $p(\theta \mid \mathbf{z})$ and evaluate functionals of interest via Monte Carlo, i.e., draw $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)} \sim p(\theta \mid \mathbf{z})$ and approximate

$$E[f(\theta) \mid \mathbf{z}] \approx \frac{1}{m}\sum_{t=1}^{m} f(\theta^{(t)})$$

- Problem: we might not know how to sample from $p(\theta \mid \mathbf{z})$

Gibbs Sampling for Bayesian Inference

For Bayesian inference we work with the posterior

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

- This expression might not be available in closed form
- Computing functionals of interest $E[f(\theta) \mid \mathbf{z}]$ might be complicated
- Idea: sample from $p(\theta \mid \mathbf{z})$ and evaluate functionals of interest via Monte Carlo, i.e., draw $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)} \sim p(\theta \mid \mathbf{z})$ and approximate

$$E[f(\theta) \mid \mathbf{z}] \approx \frac{1}{m}\sum_{t=1}^{m} f(\theta^{(t)})$$

- Problem: we might not know how to sample from $p(\theta \mid \mathbf{z})$

Gibbs Sampling for Bayesian Inference

- Say $\theta = (\theta_1, \dots, \theta_d)$
- Say you can sample from each of the conditionals

$$p(\theta_1 \mid \theta_2, \dots, \theta_d, \mathbf{z})$$
$$\vdots$$
$$p(\theta_d \mid \theta_1, \dots, \theta_{d-1}, \mathbf{z})$$

- Then a Gibbs sampler can be implemented to obtain draws

$$\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_d^{(t)}) \sim p(\theta \mid \mathbf{z}), \quad t = 1, \dots, m$$

and approximate

$$E[f(\theta) \mid \mathbf{z}] \approx \frac{1}{m} \sum_{t=1}^{m} f(\theta^{(t)})$$

Gibbs Sampling for Bayesian Inference

- Say $\theta = (\theta_1, \ldots, \theta_d)$

- Say you can sample from each of the conditionals

$$p(\theta_1 \mid \theta_2, \ldots, \theta_d, \mathbf{z})$$
$$\vdots$$
$$p(\theta_d \mid \theta_1, \ldots, \theta_{d-1}, \mathbf{z})$$

- Then a Gibbs sampler can be implemented to obtain draws

$$\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_d^{(t)}) \sim p(\theta \mid \mathbf{z}), \quad t = 1, \ldots, m$$

and approximate

$$E[f(\theta) \mid \mathbf{z}] \approx \frac{1}{m} \sum_{t=1}^{m} f(\theta^{(t)})$$

Example of Gibbs Sampling for Bayesian Inference

Consider the changepoint detection problem presented by Carlin, Gelfand and Smith (1992)[4]

- The data are counts generated over discrete time as

$$X_s \sim \text{Poisson}(\mu), \text{ if } s = 1, \ldots, \tau$$
$$X_s \sim \text{Poisson}(\lambda), \text{ if } s = \tau + 1, \ldots, T$$

  where τ is unknown

- The vector of parameters is $\theta = (\mu, \lambda, \tau)$
- The likelihood function is given by

$$L(\mu, \lambda, \tau \mid x_1, \ldots, x_T) = \prod_{s \leq \tau} \frac{\mu^{x_s} e^{-\mu}}{x_s!} \prod_{\tau < s \leq T} \frac{\lambda^{x_s} e^{-\lambda}}{x_s!}$$

[4] `www.jstor.org/stable/2347570`

Example of Gibbs Sampling for Bayesian Inference

Consider the changepoint detection problem presented by Carlin, Gelfand and Smith (1992)[4]

- The data are counts generated over discrete time as

$$X_s \sim \text{Poisson}(\mu), \text{ if } s = 1, \ldots, \tau$$
$$X_s \sim \text{Poisson}(\lambda), \text{ if } s = \tau + 1, \ldots, T$$

  where τ is unknown

- The vector of parameters is $\theta = (\mu, \lambda, \tau)$

- The likelihood function is given by

$$L(\mu, \lambda, \tau \mid x_1, \ldots, x_T) = \prod_{s \leq \tau} \frac{\mu^{x_s} e^{-\mu}}{x_s!} \prod_{\tau < s \leq T} \frac{\lambda^{x_s} e^{-\lambda}}{x_s!}$$

[4] www.jstor.org/stable/2347570

Example of Gibbs Sampling for Bayesian Inference

- Consider the independent priors
  - $\mu \sim \text{Gamma}(a_1, b_1)$
  - $\lambda \sim \text{Gamma}(a_2, b_2)$
  - $\tau \sim \text{Uniform}(\{1, \ldots, T\})$

- Leading to the posterior (HW3)

$$
\begin{aligned}
p(\mu, \lambda, \tau \mid x_1, \ldots, x_T) \propto\ & \mu^{a_1 + \sum_{s \le \tau} x_s - 1} e^{-\mu(\tau + b_1)} \\
& \times \lambda^{a_2 + \sum_{\tau < s \le T} x_s - 1} e^{-\lambda(T - \tau + b_2)}
\end{aligned}
$$

- Jointly sampling μ, λ, τ doesn't seem to be easy

Example of Gibbs Sampling for Bayesian Inference

- Consider the independent priors
  - $\mu \sim \text{Gamma}(a_1, b_1)$
  - $\lambda \sim \text{Gamma}(a_2, b_2)$
  - $\tau \sim \text{Uniform}(\{1, \ldots, T\})$
- Leading to the posterior (HW3)

$$p(\mu, \lambda, \tau \mid x_1, \ldots, x_T) \propto \mu^{a_1 + \sum_{s \le \tau} x_s - 1} e^{-\mu(\tau + b_1)} \times \lambda^{a_2 + \sum_{\tau < s \le T} x_s - 1} e^{-\lambda(T - \tau + b_2)}$$

- Jointly sampling μ, λ, τ doesn't seem to be easy

Example of Gibbs Sampling for Bayesian Inference

- Consider the independent priors
  - $\mu \sim \text{Gamma}(a_1, b_1)$
  - $\lambda \sim \text{Gamma}(a_2, b_2)$
  - $\tau \sim \text{Uniform}(\{1, \dots, T\})$
- Leading to the posterior (HW3)

$$
\begin{aligned}
p(\mu, \lambda, \tau \mid x_1, \dots, x_T) \propto\ & \mu^{a_1 + \sum_{s \le \tau} x_s - 1} e^{-\mu(\tau + b_1)} \\
& \times \lambda^{a_2 + \sum_{\tau < s \le T} x_s - 1} e^{-\lambda(T - \tau + b_2)}
\end{aligned}
$$

- Jointly sampling μ, λ, τ doesn't seem to be easy

Example of Gibbs Sampling for Bayesian Inference

However, the posterior conditionals are easy to sample from

$$\mu \mid \lambda, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_1 + \sum_{s \le \tau} x_s, \tau + b_1)$$

$$\lambda \mid \mu, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_2 + \sum_{\tau < s \le T} x_s, T - \tau + b_2)$$

$$\tau \mid \mu, \lambda, x_1, \ldots, x_T \sim \text{Categorical}(q_1, \ldots, q_T)$$

$$\text{where } q_t \propto L(\mu, \lambda, \tau = t \mid x_1, \ldots, x_T)$$

$$\propto e^{(\lambda - \mu)t + (\log \mu - \log \lambda) \sum_{s \le t} x_s}$$

HW3: confirm that these are indeed the correct conditionals, and implement the corresponding Gibbs sampler

Example of Gibbs Sampling for Bayesian Inference

However, the posterior conditionals are easy to sample from

$$\mu \mid \lambda, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_1 + \sum_{s \le \tau} x_s, \tau + b_1)$$

$$\lambda \mid \mu, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_2 + \sum_{\tau < s \le T} x_s, T - \tau + b_2)$$

$$\tau \mid \mu, \lambda, x_1, \ldots, x_T \sim \text{Categorical}(q_1, \ldots, q_T)$$

$$\text{where } q_t \propto L(\mu, \lambda, \tau = t \mid x_1, \ldots, x_T) \propto e^{(\lambda - \mu)t + (\log \mu - \log \lambda) \sum_{s \le t} x_s}$$

HW3: confirm that these are indeed the correct conditionals, and implement the corresponding Gibbs sampler

Example of Gibbs Sampling for Bayesian Inference

However, the posterior conditionals are easy to sample from

$$\mu \mid \lambda, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_1 + \sum_{s \le \tau} x_s, \tau + b_1)$$

$$\lambda \mid \mu, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_2 + \sum_{\tau < s \le T} x_s, T - \tau + b_2)$$

$$\tau \mid \mu, \lambda, x_1, \ldots, x_T \sim \text{Categorical}(q_1, \ldots, q_T)$$

$$\text{where } q_t \propto L(\mu, \lambda, \tau = t \mid x_1, \ldots, x_T)$$

$$\propto e^{(\lambda - \mu)t + (\log \mu - \log \lambda) \sum_{s \le t} x_s}$$

HW3: confirm that these are indeed the correct conditionals, and implement the corresponding Gibbs sampler

Example of Gibbs Sampling for Bayesian Inference

However, the posterior conditionals are easy to sample from

$$\mu \mid \lambda, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_1 + \sum_{s \le \tau} x_s, \tau + b_1)$$

$$\lambda \mid \mu, \tau, x_1, \ldots, x_T \sim \text{Gamma}(a_2 + \sum_{\tau < s \le T} x_s, T - \tau + b_2)$$

$$\tau \mid \mu, \lambda, x_1, \ldots, x_T \sim \text{Categorical}(q_1, \ldots, q_T)$$

$$\text{where } q_t \propto L(\mu, \lambda, \tau = t \mid x_1, \ldots, x_T)$$

$$\propto e^{(\lambda - \mu)t + (\log \mu - \log \lambda) \sum_{s \le t} x_s}$$

HW3: confirm that these are indeed the correct conditionals, and

implement the corresponding Gibbs sampler

Practical Considerations for Gibbs Sampling

- *Starting point*: initial value $\theta^{(0)}$ should ideally be chosen in a high probability region of the posterior, but this is not always easy
- *Burn-in period*: what if your $\theta^{(0)}$ was far from the high probability region?: run the sampler for m iterations, discard the initial $m_0 < m$
- *Trace plots*: to choose m and m_0 you can plot each entry of $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ versus the iteration number t: keep the draws after the "chain has converged"
- We'll cover these and other diagnostics in R session 3

Practical Considerations for Gibbs Sampling

- *Starting point*: initial value $\theta^{(0)}$ should ideally be chosen in a high probability region of the posterior, but this is not always easy
- *Burn-in period*: what if your $\theta^{(0)}$ was far from the high probability region?: run the sampler for m iterations, discard the initial $m_0 < m$
- *Trace plots*: to choose m and m_0 you can plot each entry of $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ versus the iteration number t: keep the draws after the "chain has converged"
- We'll cover these and other diagnostics in R session 3

Practical Considerations for Gibbs Sampling

- *Starting point*: initial value $\theta^{(0)}$ should ideally be chosen in a high probability region of the posterior, but this is not always easy
- *Burn-in period*: what if your $\theta^{(0)}$ was far from the high probability region?: run the sampler for m iterations, discard the initial $m_0 < m$
- *Trace plots*: to choose m and m_0 you can plot each entry of $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ versus the iteration number t: keep the draws after the "chain has converged"
- We'll cover these and other diagnostics in R session 3

Practical Considerations for Gibbs Sampling

- *Starting point*: initial value $\theta^{(0)}$ should ideally be chosen in a high probability region of the posterior, but this is not always easy
- *Burn-in period*: what if your $\theta^{(0)}$ was far from the high probability region?: run the sampler for m iterations, discard the initial $m_0 < m$
- *Trace plots*: to choose m and m_0 you can plot each entry of $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ versus the iteration number t: keep the draws after the "chain has converged"
- We'll cover these and other diagnostics in R session 3

Outline

Gibbs Sampling

Bayesian Inference with Missing Data Under Ignorability

Data Augmentation

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

- Under a Bayesian approach, in general we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta)\ dz_{i(\bar{r}_i)}$$

- Under a Bayesian approach, in general we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

Missing Data and Bayes Under MAR

- Remember: for computing MLEs, life is easier under *ignorability* (MAR + separability)

- Is it the same for Bayesian inference?

- MAR + separability lead to the observed-data likelihood function

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi)} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta)\, dz_{i(\bar{r}_i)}\right]}_{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})}$$

- Under a Bayesian approach we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi),$$

but typically only θ is of interest, while ψ is a nuisance

Missing Data and Bayes Under MAR

- Remember: for computing MLEs, life is easier under *ignorability* (MAR + separability)
- Is it the same for Bayesian inference?
- MAR + separability lead to the observed-data likelihood function

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{p(\mathbf{r}|\mathbf{z}_{(\mathbf{r})},\psi)} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta)\ dz_{i(\bar{r}_i)}\right]}_{L_{obs}(\theta|\mathbf{z}_{(\mathbf{r})})}$$

- Under a Bayesian approach we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi),$$

but typically only θ is of interest, while ψ is a nuisance

Missing Data and Bayes Under MAR

- Remember: for computing MLEs, life is easier under *ignorability* (MAR + separability)

- Is it the same for Bayesian inference?

- MAR + separability lead to the observed-data likelihood function

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{p(\mathbf{r}|\mathbf{z}_{(\mathbf{r})}, \psi)} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta)\ dz_{i(\bar{r}_i)}\right]}_{L_{obs}(\theta|\mathbf{z}_{(\mathbf{r})})}$$

- Under a Bayesian approach we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi),$$

but typically only θ is of interest, while ψ is a nuisance

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\int\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\iint L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\int\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori

- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\iint L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \qquad \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta \ \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\int\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\int\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Missing Data and Bayes Under MAR

- Under a Bayesian approach, nuisance parameters are integrated over

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &= \int p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) d\psi \\
&= \frac{\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\psi}{\int\int L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi) d\theta d\psi} \\
&\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) \int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\theta, \psi) d\psi d\theta}
\end{aligned}
$$

- If additionally, $\theta \perp\!\!\!\perp \psi$ a priori

$$
\begin{aligned}
p(\theta \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) &\overset{\text{MAR}}{=} \frac{L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)}{\int L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) d\theta} \frac{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi}{\int p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \psi) p(\psi) d\psi} \\
&\overset{\text{MAR}}{\propto} L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta) \\
&\overset{\text{MAR}}{\propto} p(\theta \mid \mathbf{z}_{(\mathbf{r})})
\end{aligned}
$$

- Therefore, *ignorability* for Bayesian inference requires MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Even then, how to obtain or sample from $p(\theta \mid \mathbf{z}_{(\mathbf{r})})$?

Outline

Data Augmentation

Main idea, say:

- We want to sample from posterior

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta),$$

  but this is difficult

- It is easy to sample from

$$p(\theta \mid y, x) \propto p(x, y \mid \theta)p(\theta)$$

  for some unobserved x

- It is easy to sample from

$$p(x \mid y, \theta)$$

Data Augmentation

Main idea, say:

- We want to sample from posterior

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta),$$

  but this is difficult

- It is easy to sample from

$$p(\theta \mid y, x) \propto p(x, y \mid \theta)p(\theta)$$

  for some unobserved x

- It is easy to sample from

$$p(x \mid y, \theta)$$

Data Augmentation

Main idea, say:

- We want to sample from posterior

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta),$$

  but this is difficult

- It is easy to sample from

$$p(\theta \mid y, x) \propto p(x, y \mid \theta)p(\theta)$$

  for some unobserved x

- It is easy to sample from

$$p(x \mid y, \theta)$$

Data Augmentation

- Start from $x^{(0)}$ (or from $\theta^{(0)}$ and switch the steps)
- At iteration t, draw

$$\theta^{(t)} \sim p(\theta \mid y, x^{(t-1)})$$
$$x^{(t)} \sim p(x \mid y, \theta^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(x^{(t)}, \theta^{(t)}) \sim p(x, \theta \mid y)$$

  and

$$\theta^{(t)} \sim p(\theta \mid y)$$

- Generally applicable, not only to missing data problems!
- Looks very much like an application of Gibbs sampling, what's special?

Data Augmentation

- Start from $x^{(0)}$ (or from $\theta^{(0)}$ and switch the steps)

- At iteration t, draw

$$\theta^{(t)} \sim p(\theta \mid y, x^{(t-1)})$$
$$x^{(t)} \sim p(x \mid y, \theta^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(x^{(t)}, \theta^{(t)}) \sim p(x, \theta \mid y)$$

  and

$$\theta^{(t)} \sim p(\theta \mid y)$$

- Generally applicable, not only to missing data problems!

- Looks very much like an application of Gibbs sampling, what's special?

Data Augmentation

- Start from $x^{(0)}$ (or from $\theta^{(0)}$ and switch the steps)

- At iteration t, draw

$$\theta^{(t)} \sim p(\theta \mid y, x^{(t-1)})$$

$$x^{(t)} \sim p(x \mid y, \theta^{(t)})$$

- There exists t_0 such that for $t > t_0$ it is guaranteed that

$$(x^{(t)}, \theta^{(t)}) \sim p(x, \theta \mid y)$$

  and

$$\theta^{(t)} \sim p(\theta \mid y)$$

- Generally applicable, not only to missing data problems!

- Looks very much like an application of Gibbs sampling, what's special?

A Timeline for Gibbs Sampling and Data Augmentation

- 1977: Dempster, Laird & Rubin popularize the EM algorithm
- 1984: Geman & Geman introduce the Gibbs sampler for image processing – ignored in the statistics literature
- 1987: Tanner & Wong introduce the original Data Augmentation paper, as the Bayesian analog of the EM algorithm – not quite what we presented as DA above
- 1990: Gelfand & Smith introduce the Gibbs sampler in the statistics literature and show its connection with DA – nowadays we see DA as an application of Gibbs sampling
- 2001: van Dyk & Meng publish "The Art of Data Augmentation" as a comprehensive view of different DA-type algorithms

A Timeline for Gibbs Sampling and Data Augmentation

- 1977: Dempster, Laird & Rubin popularize the EM algorithm
- 1984: Geman & Geman introduce the Gibbs sampler for image processing – ignored in the statistics literature
- 1987: Tanner & Wong introduce the original Data Augmentation paper, as the Bayesian analog of the EM algorithm – not quite what we presented as DA above
- 1990: Gelfand & Smith introduce the Gibbs sampler in the statistics literature and show its connection with DA – nowadays we see DA as an application of Gibbs sampling
- 2001: van Dyk & Meng publish "The Art of Data Augmentation" as a comprehensive view of different DA-type algorithms

A Timeline for Gibbs Sampling and Data Augmentation

- 1977: Dempster, Laird & Rubin popularize the EM algorithm
- 1984: Geman & Geman introduce the Gibbs sampler for image processing – ignored in the statistics literature
- 1987: Tanner & Wong introduce the original Data Augmentation paper, as the Bayesian analog of the EM algorithm – not quite what we presented as DA above
- 1990: Gelfand & Smith introduce the Gibbs sampler in the statistics literature and show its connection with DA – nowadays we see DA as an application of Gibbs sampling
- 2001: van Dyk & Meng publish "The Art of Data Augmentation" as a comprehensive view of different DA-type algorithms

A Timeline for Gibbs Sampling and Data Augmentation

- 1977: Dempster, Laird & Rubin popularize the EM algorithm
- 1984: Geman & Geman introduce the Gibbs sampler for image processing – ignored in the statistics literature
- 1987: Tanner & Wong introduce the original Data Augmentation paper, as the Bayesian analog of the EM algorithm – not quite what we presented as DA above
- 1990: Gelfand & Smith introduce the Gibbs sampler in the statistics literature and show its connection with DA – nowadays we see DA as an application of Gibbs sampling
- 2001: van Dyk & Meng publish "The Art of Data Augmentation" as a comprehensive view of different DA-type algorithms

A Timeline for Gibbs Sampling and Data Augmentation

- 1977: Dempster, Laird & Rubin popularize the EM algorithm
- 1984: Geman & Geman introduce the Gibbs sampler for image processing – ignored in the statistics literature
- 1987: Tanner & Wong introduce the original Data Augmentation paper, as the Bayesian analog of the EM algorithm – not quite what we presented as DA above
- 1990: Gelfand & Smith introduce the Gibbs sampler in the statistics literature and show its connection with DA – nowadays we see DA as an application of Gibbs sampling
- 2001: van Dyk & Meng publish "The Art of Data Augmentation" as a comprehensive view of different DA-type algorithms

DA for Handling Missing Data in Bayesian Inference

- Consider the full-data likelihood

$$L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta, \psi \mid \mathbf{z}, \mathbf{r}) \propto L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) p(\theta, \psi)$$

- Say you can sample from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, r_i, \theta, \psi) \propto p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

  for $i = 1, \ldots, n$

- If this is the case, you can iteratively sample from these to run a DA algorithm!

DA for Handling Missing Data in Bayesian Inference

- Consider the full-data likelihood

$$L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta, \psi \mid \mathbf{z}, \mathbf{r}) \propto L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) p(\theta, \psi)$$

- Say you can sample from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, r_i, \theta, \psi) \propto p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

for $i = 1, \ldots, n$

- If this is the case, you can iteratively sample from these to run a DA algorithm!

DA for Handling Missing Data in Bayesian Inference

- Consider the full-data likelihood

$$L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta, \psi \mid \mathbf{z}, \mathbf{r}) \propto L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) p(\theta, \psi)$$

- Say you can sample from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, r_i, \theta, \psi) \propto p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

  for $i = 1, \ldots, n$

- If this is the case, you can iteratively sample from these to run a DA algorithm!

DA for Handling Missing Data in Bayesian Inference

- Consider the full-data likelihood

$$L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta, \psi \mid \mathbf{z}, \mathbf{r}) \propto L(\theta, \psi \mid \mathbf{z}, \mathbf{r}) p(\theta, \psi)$$

- Say you can sample from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, r_i, \theta, \psi) \propto p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

  for $i = 1, \ldots, n$

- If this is the case, you can iteratively sample from these to run a DA algorithm!

DA for Handling Missing Data in Bayesian Inference

- Typically, sampling from $p(\theta, \psi \mid \mathbf{z}, \mathbf{r})$ is not easy
- Say $\theta = (\theta_1, \ldots, \theta_{d_1})$ and $\psi = (\psi_1, \ldots, \psi_{d_2})$
- Instead of sampling (θ, ψ) jointly, we might have to sample sequentially from the conditionals

$$
\begin{gathered}
p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \psi, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi_1 \mid \psi_2, \ldots, \psi_{d_2}, \theta, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\psi_{d_2} \mid \psi_1, \ldots, \psi_{d_2 - 1}, \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

- Or, we might be able to sample from

$$
\begin{gathered}
p(\theta \mid \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi \mid \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

DA for Handling Missing Data in Bayesian Inference

- Typically, sampling from $p(\theta, \psi \mid \mathbf{z}, \mathbf{r})$ is not easy
- Say $\theta = (\theta_1, \ldots, \theta_{d_1})$ and $\psi = (\psi_1, \ldots, \psi_{d_2})$
- Instead of sampling (θ, ψ) jointly, we might have to sample sequentially from the conditionals

$$
\begin{gathered}
p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \psi, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1-1}, \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi_1 \mid \psi_2, \ldots, \psi_{d_2}, \theta, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\psi_{d_2} \mid \psi_1, \ldots, \psi_{d_2-1}, \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

- Or, we might be able to sample from

$$
\begin{gathered}
p(\theta \mid \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi \mid \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

DA for Handling Missing Data in Bayesian Inference

- Typically, sampling from $p(\theta, \psi \mid \mathbf{z}, \mathbf{r})$ is not easy
- Say $\theta = (\theta_1, \ldots, \theta_{d_1})$ and $\psi = (\psi_1, \ldots, \psi_{d_2})$
- Instead of sampling (θ, ψ) jointly, we might have to sample sequentially from the conditionals

$$
\begin{gathered}
p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \psi, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1-1}, \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi_1 \mid \psi_2, \ldots, \psi_{d_2}, \theta, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\psi_{d_2} \mid \psi_1, \ldots, \psi_{d_2-1}, \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

- Or, we might be able to sample from

$$
\begin{gathered}
p(\theta \mid \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi \mid \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

DA for Handling Missing Data in Bayesian Inference

- Typically, sampling from $p(\theta, \psi \mid \mathbf{z}, \mathbf{r})$ is not easy
- Say $\theta = (\theta_1, \ldots, \theta_{d_1})$ and $\psi = (\psi_1, \ldots, \psi_{d_2})$
- Instead of sampling (θ, ψ) jointly, we might have to sample sequentially from the conditionals

$$
\begin{gathered}
p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \psi, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1-1}, \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi_1 \mid \psi_2, \ldots, \psi_{d_2}, \theta, \mathbf{z}, \mathbf{r}) \\
\vdots \\
p(\psi_{d_2} \mid \psi_1, \ldots, \psi_{d_2-1}, \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

- Or, we might be able to sample from

$$
\begin{gathered}
p(\theta \mid \psi, \mathbf{z}, \mathbf{r}) \\
p(\psi \mid \theta, \mathbf{z}, \mathbf{r})
\end{gathered}
$$

DA for Bayesian Inference Under Ignorability

Even under ignorability, the integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things

- Consider the full-data likelihood for the study variables only

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^{n} p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta \mid \mathbf{z}) \propto L(\theta \mid \mathbf{z})p(\theta)$$

or you can sample from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \mathbf{z})$$
$$\vdots$$
$$p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \mathbf{z})$$

- Say you can sample for $i = 1, \ldots, n$ from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, \theta)$$

- Then you can implement a DA algorithm under ignorability!

DA for Bayesian Inference Under Ignorability

Even under ignorability, the integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things

- Consider the full-data likelihood for the study variables only

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^{n} p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta \mid \mathbf{z}) \propto L(\theta \mid \mathbf{z}) p(\theta)$$

  or you can sample from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \mathbf{z})$$

$$\vdots$$

$$p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \mathbf{z})$$

- Say you can sample for $i = 1, \ldots, n$ from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, \theta)$$

- Then you can implement a DA algorithm under ignorability!

DA for Bayesian Inference Under Ignorability

Even under ignorability, the integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things

- Consider the full-data likelihood for the study variables only

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^{n} p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta \mid \mathbf{z}) \propto L(\theta \mid \mathbf{z}) p(\theta)$$

  or you can sample from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \mathbf{z})$$
$$\vdots$$
$$p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \mathbf{z})$$

- Say you can sample for $i = 1, \ldots, n$ from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, \theta)$$

- Then you can implement a DA algorithm under ignorability!

DA for Bayesian Inference Under Ignorability

Even under ignorability, the integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things

- Consider the full-data likelihood for the study variables only

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^{n} p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta \mid \mathbf{z}) \propto L(\theta \mid \mathbf{z}) p(\theta)$$

  or you can sample from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \mathbf{z})$$
$$\vdots$$
$$p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \mathbf{z})$$

- Say you can sample for $i = 1, \ldots, n$ from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, \theta)$$

- Then you can implement a DA algorithm under ignorability!

DA for Bayesian Inference Under Ignorability

Even under ignorability, the integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things

- Consider the full-data likelihood for the study variables only

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^{n} p(z_i \mid \theta)$$

- Say you can sample from

$$p(\theta \mid \mathbf{z}) \propto L(\theta \mid \mathbf{z}) p(\theta)$$

  or you can sample from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_{d_1}, \mathbf{z})$$
$$\vdots$$
$$p(\theta_{d_1} \mid \theta_1, \ldots, \theta_{d_1 - 1}, \mathbf{z})$$

- Say you can sample for $i = 1, \ldots, n$ from

$$p(z_{i(\bar{r}_i)} \mid z_{i(r_i)}, \theta)$$

- Then you can implement a DA algorithm under ignorability!

Example: Multinomial Data, Dirichlet Prior

Continuing our example from the previous classes:

- Let $Z_i = (Z_{i1}, Z_{i2})$, $\quad Z_{i1}, Z_{i2} \in \{1, 2\}$, $\quad Z_i$'s are i.i.d.,

$$p(Z_{i1} = k, Z_{i2} = l \mid \theta) = \pi_{kl}$$

- $\theta = (\dots, \pi_{kl}, \dots)$, $\qquad W_{ikl} = I(Z_{i1} = k, Z_{i2} = l)$

- The likelihood of the study variables is

$$L(\theta \mid \mathbf{z}) = \prod_i \left[\prod_{k,l} \pi_{kl}^{W_{ikl}} \right] = \prod_{k,l} \pi_{kl}^{n_{kl}}$$

  where $n_{kl} = \sum_i W_{ikl}, \quad k, l \in \{1, 2\}$

- Say $\theta = (\dots, \pi_{kl}, \dots) \sim \text{Dirichlet}(\alpha)$, $\quad \alpha = (\dots, \alpha_{kl}, \dots)$

- Therefore, $\theta \mid \mathbf{z} \sim \text{Dirichlet}(\alpha')$, $\quad \alpha' = (\dots, \alpha_{kl} + n_{kl}, \dots)$

Example: Multinomial Data, Dirichlet Prior

However, we have missing data (we'll assume ignorability)

- Let $R_i = (R_{i1}, R_{i2})$, $R_{i1}, R_{i2} \in \{0, 1\}$, R_i's are i.i.d.
- In HW2, you show that the observed-data likelihood for the study variables can be written as

$$L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) = \prod_i \pi_{z_{i1}z_{i2}}^{I(r_i=11)} \pi_{z_{i1}+}^{I(r_i=10)} \pi_{+z_{i2}}^{I(r_i=01)}$$

- A quick inspection shows that Bayesian inference with $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ becomes complicated
- However, notice that the distribution of $Z_{(\bar{r})} \mid z_{(r)}, \theta$ is easy to derive!

$$\text{For } r = 01,\ Z_1 \mid z_2, \theta \sim \text{Categorical}[\pi_{+z_2}^{-1}(\pi_{1,z_2}, \pi_{2,z_2})]$$
$$\text{For } r = 10,\ Z_2 \mid z_1, \theta \sim \text{Categorical}[\pi_{z_1+}^{-1}(\pi_{z_1,1}, \pi_{z_1,2})]$$
$$\text{For } r = 00,\ Z \mid \theta \sim \text{Categorical}[(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})]$$

For $r = 11$, there's nothing to sample!

Example: Multinomial Data, Dirichlet Prior

However, we have missing data (we'll assume ignorability)

- Let $R_i = (R_{i1}, R_{i2})$, $\quad R_{i1}, R_{i2} \in \{0, 1\}$, $\quad R_i$'s are i.i.d.
- In HW2, you show that the observed-data likelihood for the study variables can be written as

$$L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) = \prod_i \pi_{z_{i1} z_{i2}}^{I(r_i=11)} \pi_{z_{i1}+}^{I(r_i=10)} \pi_{+z_{i2}}^{I(r_i=01)}$$

- A quick inspection shows that Bayesian inference with $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ becomes complicated
- However, notice that the distribution of $Z_{(\bar{r})} \mid z_{(r)}, \theta$ is easy to derive!

$$\begin{aligned}
&\text{For } r = 01,\ Z_1 \mid z_2, \theta \sim \text{Categorical}[\pi_{+z_2}^{-1}(\pi_{1,z_2}, \pi_{2,z_2})] \\
&\text{For } r = 10,\ Z_2 \mid z_1, \theta \sim \text{Categorical}[\pi_{z_1+}^{-1}(\pi_{z_1,1}, \pi_{z_1,2})] \\
&\text{For } r = 00,\ Z \mid \theta \sim \text{Categorical}[(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})] \\
&\text{For } r = 11,\ \text{there's nothing to sample!}
\end{aligned}$$

Example: Multinomial Data, Dirichlet Prior

Therefore, implementing a DA algorithm is very straightforward!

- Choose starting point $\theta^{(0)}$
- Iteratively do
  (a) For $i = 1, \ldots, n$, sample
  $$Z_{i(\bar{r}_i)}^{(t)} \sim p(z_{(\bar{r}_i)} \mid z_{i(r_i)}, \theta^{(t-1)})$$
  and define $z_i^{(t)} = \text{“}(z_{i(r_i)}, Z_{i(\bar{r}_i)}^{(t)})\text{”}$[5]

  (b) Sample $\theta^{(t)} \mid \mathbf{z}^{(t)} \sim \text{Dirichlet}(\alpha^{(t)})$, where $\alpha^{(t)} = (\ldots, \alpha_{kl} + n_{kl}^{(t)}, \ldots)$ where $\mathbf{z}^{(t)} = \{z_i^{(t)}\}_{i=1}^n$ and each $n_{kl}^{(t)}$ is computed from $\mathbf{z}^{(t)}$

HW3: note that part (b) only uses $n_{kl}^{(t)}$ from part (a). Can you find a way of simplifying part (a) so that we don't need to sample each $z_i^{(t)}$ individually but still obtain each $n_{kl}^{(t)}$?

[5]We don't really mean "put $z_{i(r_i)}$ on the left and $Z_{i(\bar{r}_i)}^{(t)}$ on the right," but rather, keep the observed entries of z_i fixed at $z_{i(r_i)}$ and fill its missing entries with $Z_{i(\bar{r}_i)}^{(t)}$

Example: Multinomial Data, Dirichlet Prior

Therefore, implementing a DA algorithm is very straightforward!

- Choose starting point $\theta^{(0)}$
- Iteratively do
  (a) For $i = 1, \ldots, n$, sample

  $$Z_{i(\bar{r}_i)}^{(t)} \sim p(z_{(\bar{r}_i)} \mid z_{i(r_i)}, \theta^{(t-1)})$$

  and define $z_i^{(t)} = \text{"}(z_{i(r_i)}, Z_{i(\bar{r}_i)}^{(t)})\text{"}$[5]

  (b) Sample $\theta^{(t)} \mid \mathbf{z}^{(t)} \sim \text{Dirichlet}(\alpha^{(t)})$, where $\alpha^{(t)} = (\ldots, \alpha_{kl} + n_{kl}^{(t)}, \ldots)$ where $\mathbf{z}^{(t)} = \{z_i^{(t)}\}_{i=1}^n$ and each $n_{kl}^{(t)}$ is computed from $\mathbf{z}^{(t)}$

HW3: note that part (b) only uses $n_{kl}^{(t)}$ from part (a). Can you find a way of simplifying part (a) so that we don't need to sample each $z_i^{(t)}$ individually but still obtain each $n_{kl}^{(t)}$?

[5] We don't really mean "put $z_{i(r_i)}$ on the left and $Z_{i(\bar{r}_i)}^{(t)}$ on the right," but rather, keep the observed entries of z_i fixed at $z_{i(r_i)}$ and fill its missing entries with $Z_{i(\bar{r}_i)}^{(t)}$

Example: Multinomial Data, Dirichlet Prior

Therefore, implementing a DA algorithm is very straightforward!

- Choose starting point $\theta^{(0)}$
- Iteratively do
  (a) For $i = 1, \dots, n$, sample

$$Z_{i(\bar{r}_i)}^{(t)} \sim p(z_{(\bar{r}_i)} \mid z_{i(r_i)}, \theta^{(t-1)})$$

  and define $z_i^{(t)} = "(z_{i(r_i)}, Z_{i(\bar{r}_i)}^{(t)})"$[5]

  (b) Sample $\theta^{(t)} \mid \mathbf{z}^{(t)} \sim \text{Dirichlet}(\alpha^{(t)})$, where $\alpha^{(t)} = (\dots, \alpha_{kl} + n_{kl}^{(t)}, \dots)$ where $\mathbf{z}^{(t)} = \{z_i^{(t)}\}_{i=1}^n$ and each $n_{kl}^{(t)}$ is computed from $\mathbf{z}^{(t)}$

HW3: note that part (b) only uses $n_{kl}^{(t)}$ from part (a). Can you find a way of simplifying part (a) so that we don't need to sample each $z_i^{(t)}$ individually but still obtain each $n_{kl}^{(t)}$?

[5] We don't really mean "put $z_{i(r_i)}$ on the left and $Z_{i(\bar{r}_i)}^{(t)}$ on the right," but rather, keep the observed entries of z_i fixed at $z_{i(r_i)}$ and fill its missing entries with $Z_{i(\bar{r}_i)}^{(t)}$

Summary

Main take-aways from today's lecture:

- Gibbs sampling to sample from complex distributions via sequential sampling from conditionals – commonly applied to sampling from posterior distributions
- Ignorability for Bayesian inference: MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Data augmentation to handle missing data in Bayesian inference – it can be straightforward for some problems, but more generally it needs additional Gibbs steps

Next lecture:

- Multiple imputation (finally!)
- Multiple imputation by chained equations

Summary

Main take-aways from today's lecture:

- Gibbs sampling to sample from complex distributions via sequential sampling from conditionals – commonly applied to sampling from posterior distributions
- Ignorability for Bayesian inference: MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori
- Data augmentation to handle missing data in Bayesian inference – it can be straightforward for some problems, but more generally it needs additional Gibbs steps

Next lecture:

- Multiple imputation (finally!)
- Multiple imputation by chained equations