3.6 Data Augmentation

Prediction (=Imputation)

- Goal: We wish to generate $\mathbf{y}_{mis}$ given the observed data $(\mathbf{y}_{obs}, \boldsymbol{\delta})$.
- Problem: The prediction model depends on unknown parameter

$$p(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta) = \frac{f(\mathbf{y}, \boldsymbol{\delta}; \eta)}{\int f(\mathbf{y}, \boldsymbol{\delta}; \eta) d\mathbf{y}_{mis}}.$$

- Remedy: Two different approaches
  1. Bayesian approach: generate $\mathbf{y}^*_{mis}$ from

$$f\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}\right) = \int p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right) p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) d\eta \tag{1}$$

  2. Frequentist approach: generate $\mathbf{y}^*_{mis,i}$ from $f\left(\mathbf{y}_{mis,i} \mid \mathbf{y}_{obs,i}, \boldsymbol{\delta}; \hat{\eta}\right)$, where $\hat{\eta}$ is a consistent estimator of η.

Bayesian approach to prediction (= imputation)

- Goal: We wish to generate $\mathbf{y}_{mis}$ from (1).
- Idea: Note that

$$\int p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right) p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) d\eta = E\left\{p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}\right\},$$

  where the expectation is wrt the posterior distribution with density $p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta})$.
- Thus, the following two-step method can be used for Bayesian imputation.
  1. Generate $\boldsymbol{\eta}^*$ from $p(\boldsymbol{\eta} \mid \mathbf{y}_{obs}, \boldsymbol{\delta})$.
  2. Given $\boldsymbol{\eta}^*$ obtained from Step 1, generate $\mathbf{y}_{mis}^*$ from $p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \boldsymbol{\eta}^*\right)$.
- Problem: How to generate $\boldsymbol{\eta}^*$ from $p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta})$?

Remark

- Posterior distribution

$$
\begin{aligned}
p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) &= \frac{f(\mathbf{y}_{obs}, \boldsymbol{\delta} \mid \eta)\pi(\eta)}{\int f(\mathbf{y}_{obs}, \boldsymbol{\delta} \mid \eta)\pi(\eta)d\eta} \\
&= \frac{\int f(\mathbf{y}, \boldsymbol{\delta} \mid \eta)\pi(\eta)d\mathbf{y}_{mis}}{\int f(\mathbf{y}_{obs}, \boldsymbol{\delta} \mid \eta)\pi(\eta)d\eta} \\
&= \int p(\eta, \mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta})d\mathbf{y}_{mis}
\end{aligned}
$$

- Predictive distribution

$$
p(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) = \int p(\eta, \mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta})d\eta
$$

Gibbs sampling

Idea: Sample from conditional distributions

Given $X^{(t)} = \left(X_1^{(t)}, X_2^{(t)}, \cdots, X_p^{(t)}\right)$, draw $X^{(t+1)}$ by sampling from the full conditionals of f,

$$\begin{aligned}
X_1^{(t+1)} &\sim P\left(X_1 \mid X_2^{(t)}, X_3^{(t)}, \cdots, X_p^{(t)}\right) \\
X_2^{(t+1)} &\sim P\left(X_2 \mid X_1^{(t+1)}, X_3^{(t)}, \cdots, X_p^{(t)}\right) \\
&\vdots \\
X_p^{(t+1)} &\sim P\left(X_p \mid X_1^{(t+1)}, X_2^{(t+1)}, \cdots, X_{p-1}^{(t+1)}\right).
\end{aligned}$$

Important questions to ask

1. Only the so-called *full-conditional* distributions $X_i \mid X_{-i}$ are used in the Gibbs sampler.
   - Do the full conditionals fully specify the joint distribution?
2. The sequence $\left(X^{(0)}, X^{(1)}, \cdots\right)$ is a Markov chain.
   - Is the target distribution $f(x_1, \cdots, x_p)$ the invariant distribution of this Markov chain?
   - Will the Markov chain converge to this distribution?

The Hammersley-Clifford theorem

Definition (Positivity condition)

A distribution with density $f(x_1, \cdots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the *positivity condition* if $f(x_1, \cdots, x_p) > 0$ for all $x_1, \cdots, x_p$ with $f_{X_i}(x_i) > 0$.

Theorem

Let $(X_1, \cdots, X_p)$ satisfy the positivity condition and have joint density $f(x_1, \cdots, x_p)$. Then for all $(\zeta_1, \cdots, \zeta_p) \in supp(f)$

$$f(x_1, \cdots, x_p) \propto \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j \mid x_1, \cdots, x_{j-1}, \zeta_{j+1}, \cdots, \zeta_p)}{f_{X_j|X_{-j}}(\zeta_j \mid x_1, \cdots, x_{j-1}, \zeta_{j+1}, \cdots, \zeta_p)}$$

Note: The theorem does not guarantee the existence of a joint distribution for every set of full conditionals!

Justification (p=3)

Note that

$$f(x_1, x_2, x_3) = f(x_3 \mid x_1, x_2) f(x_1, x_2). \tag{2}$$

Now, for any fixed number $(\zeta_1, \zeta_2, \zeta_3) \in \text{supp}(f)$, we have

$$f(x_1, x_2, \zeta_3) = f(\zeta_3 \mid x_1, x_2) f(x_1, x_2)$$

which implies

$$f(x_1, x_2) = \frac{f(x_1, x_2, \zeta_3)}{f(\zeta_3 \mid x_1, x_2)}.$$

Thus, (2) is changed to

$$f(x_1, x_2, x_3) = f(x_1, x_2, \zeta_3) \frac{f(x_3 \mid x_1, x_2)}{f(\zeta_3 \mid x_1, x_2)}. \tag{3}$$

Applying the same argument for obtaining (3), we have

$$f(x_1, x_2, \zeta_3) = f(x_1, \zeta_2, \zeta_3) \frac{f(x_2 \mid x_1, \zeta_3)}{f(\zeta_2 \mid x_1, \zeta_3)} \tag{4}$$

and

$$f(x_1, \zeta_2, \zeta_3) = f(\zeta_1, \zeta_2, \zeta_3) \frac{f(x_1 \mid \zeta_2, \zeta_3)}{f(\zeta_1 \mid \zeta_2, \zeta_3)}. \tag{5}$$

Combining the three results, we obtain

$$f(x_1, x_2, x_3) = f(\zeta_1, \zeta_2, \zeta_3) \frac{f(x_1 \mid \zeta_2, \zeta_3)}{f(\zeta_1 \mid \zeta_2, \zeta_3)} \frac{f(x_2 \mid x_1, \zeta_3)}{f(\zeta_2 \mid x_1, \zeta_3)} \frac{f(x_3 \mid x_1, x_2)}{f(\zeta_3 \mid x_1, x_2)}.$$

This completes the proof for Hammersly-Clifford theorem for $p = 3$.

Example

- Consider the following model

$$\begin{aligned} X_1 \mid X_2 &\sim Exp(\lambda X_2) \\ X_2 \mid X_1 &\sim Exp(\lambda X_1) \end{aligned}$$

- Trying to apply the Hammersley-Clifford theorem, we obtain

$$f(x_1, x_2) \propto \frac{f_{X_1|X_2}(x_1 \mid \zeta_2) \cdot f_{X_2|X_1}(x_2 \mid x_1)}{f_{X_1|X_2}(\zeta_1 \mid \zeta_2) \cdot f_{X_2|X_1}(\zeta_2 \mid x_1)} \propto \exp(-\lambda x_1 x_2)$$

- Joint density cannot be normalized.

$$\int \int \exp(-\lambda x_1 x_2)\, dx_1\, dx_2 = \infty$$

- There is no joint density with the above full conditionals.

Convergence Properties

Main results

1. The joint distribution $f(x_1, \cdots, x_p)$ is indeed the invariant distribution of the Markov chain $(X^{(0)}, X^{(1)}, \cdots)$ generated by the Gibbs sampler.
2. If the joint distribution $f(x_1, \cdots, x_p)$ satisfies the positivity condition, the Gibbs sampler yields an irreducible, recurrent Markov chain.
3. If the Markov chain generated by the Gibbs sampler is irreducible and recurrent (which is the case when the positivity condition holds), then for any integrable function h

$$\lim_n \frac{1}{n} \sum_{t=1}^{n} h\left(X^{(t)}\right) = E_f\{h(X)\}$$

with probability one, for almost every starting value $X^{(0)}$.

Example

- Consider

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right]$$

- Associated full conditional

$$\begin{aligned} X_1 \mid (X_2 = x_2) &\sim N\left[\mu_1 + (\sigma_{12}/\sigma_{22})(x_2 - \mu_2), \sigma_{11} - (\sigma_{12})^2/\sigma_{22}\right] \\ X_2 \mid (X_1 = x_1) &\sim N\left[\mu_2 + (\sigma_{12}/\sigma_{11})(x_1 - \mu_1), \sigma_{22} - (\sigma_{12})^2/\sigma_{11}\right] \end{aligned}$$

- Gibbs sampler consists of iterating for $t = 1, 2, \cdots$
  1. Draw $X_1^{(t)} \sim N\left[\mu_1 + (\sigma_{12}/\sigma_{22})(X_2^{(t-1)} - \mu_2), \sigma_{11} - (\sigma_{12})^2/\sigma_{22}\right]$
  2. Draw $X_2^{(t)} \sim N\left[\mu_2 + (\sigma_{12}/\sigma_{11})(X_1^{(t)} - \mu_1), \sigma_{22} - (\sigma_{12})^2/\sigma_{11}\right]$

Remark

- $X^{(t-1)}$ and $X^{(t)}$ are dependent and typically positively correlated
- The amount of correlation increases with the dependency (correlation) of the components $(X_1^{(t)}, \cdots, X_p^{(t)})$.
- Consequence: a sample of size n from a Gibbs sampler can contain less information than an i.i.d. sample of size n, especially when the correlation between $X^{(t-1)}$ and $X^{(t)}$ is large.

Data Augmentation

Idea: Application of the Gibbs sampling to missing data problem

$$
\begin{aligned}
(\mathbf{y}_{\rm obs}, \boldsymbol{\delta}) &= \text{observed data} \\
(\mathbf{y}, \boldsymbol{\delta}) &= \text{complete data} \\
\eta = (\theta, \phi) &= \text{model parameters}
\end{aligned}
$$

Predictive distribution:

$$
P(\mathbf{y}_{\rm mis} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}) = \int P(\mathbf{y}_{mis} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}, \eta)\, dP(\eta \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta})
$$

Posterior distribution:

$$
\begin{aligned}
P(\eta \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}) &= \int P(\eta \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}, \mathbf{y}_{mis})\, dP(\mathbf{y}_{mis} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}) \\
&= \int P(\eta \mid \boldsymbol{\delta}, \mathbf{y})\, dP(\mathbf{y} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta})
\end{aligned}
$$

Data Augmentation

Algorithm: Iterative method of data augmentation

1. I-step: Draw
$$\mathbf{y}_{\text{mis}}^{(t)} \sim P\left(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}, \boldsymbol{\eta}^{(t)}\right)$$
2. P-step: Draw
$$\boldsymbol{\eta}^{(t+1)} \sim P\left(\boldsymbol{\eta} \mid \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(t)}, \boldsymbol{\delta}\right).$$

Remark

- Data augmentation (DA) is similar in sprit to EM algorithm. The I-step corresponds to E-step of the EM algorithm.
- The parameter update steps (P-step vs M-step) are different. In the EM algorithm, the parameters are updated deterministically. In the DA algorithm, the parameters are updated stochastically.
- The uncertainty in the parameter estimation is automatically captured in the Bayesian framework.
- In the monotone missing patters, the iterative algorithm is not necessary. That is, two-step method is enough.

Example 3.19

$Y_i \sim Bernoulli(p)$, $i = 1, 2, \cdots, r$, with prior $p \sim Beta(\alpha, \beta)$, ($\alpha, \beta$: given). How to generate $Y_i^{*(t)}, t = r+1, r+2, \cdots, n$?

1. Method 1: (Noniterative method)
   1. Generate p^* from $P(p \mid Y_1, Y_2, \cdots, Y_r)$. (Note that the observed posterior distribution is $Beta\left(\alpha + \sum_{i=1}^r y_i, \beta + r - \sum_{i=1}^r y_i\right)$.)
   2. Generate Y_i^* from $P(Y_i \mid p^*)$.
2. Method 2: (Iterative method using DA)
   1. I-step : $Y_i^* \sim Bernoulli(p^*)$
   2. P-step : $p^* \sim Beta\left(\alpha + \sum_{i=1}^n y_i^*, \beta + n - \sum_{i=1}^n y_i^*\right)$

Posterior Distribution (under no missign data)

- Likelihood

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

- Prior

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

- Posterior

$$P(p \mid y_1, \cdots, y_n) \propto p^{\sum_{i=1}^{n} y_i+\alpha-1}(1-p)^{n-\sum_{i=1}^{n} y_i+\beta-1}$$

The posterior distribution is $Beta(\alpha^*, \beta^*)$ where $\alpha^* = \sum_{i=1}^{n} y_i + \alpha$ and $\beta^* = n - \sum_{i=1}^{n} y_i + \beta$.

Equivalence of the two methods

- Note that, for $i > r$,

$$
\begin{aligned}
E\left(y_i^{*(t+1)} \mid Y^{*(t)}\right) &= E\left(\theta^{*(t)} \mid Y^{*(t)}\right) \\
&= \frac{\alpha + \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} y_i^{*(t)}}{\alpha + \beta + n} \\
&= \frac{\alpha + \sum_{i=1}^{r} y_i}{\alpha + \beta + r} \\
&\quad + \lambda \left(\frac{\sum_{i=r+1}^{n} y_i^{*(t)}}{n - r} - \frac{\alpha + \sum_{i=1}^{r} y_i}{\alpha + \beta + r} \right)
\end{aligned}
$$

where $Y^{*(t)} = (y_1^{*(t)}, \cdots, y_n^{*(t)})$ and $\lambda = (n - r)/(\alpha + \beta + n)$.

Equivalence of the two methods (Cont'd)

- Writing
$$E\left(y_i^{*(t+1)} \mid Y^{*(t)}\right) = a_0 + \lambda(a^{(t)} - a_0)$$
where $a_0 = (\alpha + \sum_{i=1}^{r} y_i)/(\alpha + \beta + r)$ and $a^{(t)} = (\sum_{i=r+1}^{n} y_i^{*(t)})/(n-r)$, we can obtain
$$E\left(y_i^{*(t+1)} \mid Y^{*(1)}\right) = a_0 + \lambda^t \left(a^{(1)} - a_0\right).$$
- Thus, as $\lambda < 1$,
$$\lim_{t\to\infty} E\left(y_i^{*(t+1)} \mid y_1, y_2 \cdots, y_r\right) = a_0 = \frac{\alpha + \sum_{i=1}^{r} y_i}{\alpha + \beta + r},$$
which can also be obtained directly from Method 1.

Two uses of data augmentation

- Parameter simulation: collect and summarize a sequence of dependent draws of θ,
$$\theta^{(t+1)}, \theta^{(t+2)}, \cdots, \theta^{(t+N)},$$
where t is large enough to ensure stationarity.
- Multiple imputation: collect independent draws of $\mathbf{y}$,
$$\mathbf{y}^{*(t)}, \mathbf{y}^{*(2t)}, \cdots, \mathbf{y}^{*(mt)}$$

Example: Gaussian Mixture Model

- Model: Gaussian Mixture Model

$$f(y) = \sum_{g=1}^{G} \pi_g \phi(y; \mu_g, \sigma^2)$$

where $\sum_{g=1}^{G} \pi_g = 1$ and $\phi(y; \mu, \sigma^2)$ is the density of $N(\mu, \sigma^2)$ distribution. That is

$$\phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

- We assume that G and σ^2 are known.
- Goal: Want to make Bayesian inference for $\boldsymbol{\theta} = (\pi_1, \cdots, \pi_G, \mu_1, \cdots, \mu_G)$

Example: GMM

- The mixture can be explained using a vector of latent variables $\mathbf{Z} = (Z_1, \cdots, Z_G)$:

$$f(y) = \sum_{g=1}^{G} P(Z_g = 1) f(y \mid Z_g = 1).$$

- Prior distribution for $(\pi_1, \cdots, \pi_G)$: Dirichlet $(\alpha_1, \cdots, \alpha_G)$

$$f(\pi_1, \cdots, \pi_G) = \frac{\Gamma(\sum_g \alpha_g)}{\prod_g \Gamma(\alpha_g)} \prod_g \pi_g{}^{\alpha_g - 1}$$

- Prior distribution for $(\mu_1, \cdots, \mu_G)$:

$$f(\mu_g) \propto \exp\{-(\mu_g - \mu_0)^2 / (2\sigma_0^2)\}$$

Example: GMM

The joint distribution of the augmented system is

$$
\begin{aligned}
& f(y_1, \cdots, y_n, \mathbf{z}_1, \cdots, \mathbf{z}_n, \mu_1, \cdots, \mu_G, \pi_1, \cdots, \pi_G) \\
\propto \quad & \left(\prod_g {\pi_g}^{\alpha_g - 1}\right) \cdot \left(\prod_{g=1}^{G} \exp\{-(\mu_g - \mu_0)^2/(2\sigma_0^2)\}\right) \\
& \times \left[\prod_{i=1}^{n} \prod_g \left\{\pi_g \exp(-(y_i - \mu_g)^2/(2\sigma^2))\right\}^{z_{ig}}\right]
\end{aligned}
$$

Full conditionals (1)

- We can show that

$$Pr(z_{ig} = 1 \mid \text{others}) = \frac{\pi_g N(y_i \mid \mu_g, \sigma^2)}{\sum_{g=1}^{G} \pi_g N(y_i \mid \mu_g, \sigma^2)}$$

for $g = 1, 2, \ldots, G$.

Full conditionals (2)

- We can show that

$$(\pi_1, \ldots, \pi_G) \mid \text{others} \quad \sim \quad \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_G + n_G)$$

where $n_g = \sum_{i=1}^{n} z_{ig}$.

Full conditionals (3)

- We can show that

$$\mu_g \mid \text{others} \overset{ind}{\sim} N(\hat{\mu}_g, \hat{\sigma}_g^2)$$

where

$$\begin{aligned} \hat{\mu}_g &= \frac{(n_g/\sigma^2)\bar{y}_g + (1/\sigma_0^2)\mu_0}{n_g/\sigma^2 + 1/\sigma_0^2} \\ \hat{\sigma}_g^2 &= \left(n_g/\sigma^2 + 1/\sigma_0^2\right)^{-1}. \end{aligned}$$