§4.5 Multiple Imputation

1 Introduction

- Assume a parametric model: $y \sim f(y \mid x; \theta)$
- We are interested in making inference about θ.
- In Bayesian approach, we want to make inference about θ from

$$f(\theta \mid x, y) = \frac{\pi(\theta) f(y \mid x, \theta)}{\int \pi(\theta) f(y \mid x, \theta) d\theta}$$

  where $\pi(\theta)$ is a prior distribution, which is assumed to be known for simplicity here.

- The point estimator is

$$\hat{\theta}_n = E\{\theta \mid x, y\} \tag{1}$$

  and its variance estimator is

$$\hat{V}_n = V\{\theta \mid x, y\}. \tag{2}$$

  We may express $\hat{\theta}_n = \hat{\theta}_n(x, y)$ and $\hat{V}_n = \hat{V}_n(x, y)$.

- Now, consider the case when x is always observed and y is subject to missingness. Let $y = (y_{obs}, y_{mis})$ be the (observed, missing) part of the sample.
- We have two approaches of making inference about θ using the observed data (x, y_{obs}):

  1. Direct Bayesian approach: Consider

$$f(\theta \mid x, y_{obs}) = \frac{\pi(\theta) f(y_{obs} \mid x, \theta)}{\int \pi(\theta) f(y_{obs} \mid x, \theta) d\theta}$$

and use

$$\hat{\theta}_r = E\{\theta \mid x, y_{obs}\} \tag{3}$$

and its variance estimator is

$$\hat{V}_r = V\{\theta \mid x, y_{obs}\}. \tag{4}$$

2. Multiple imputation approach:

   (a) For each k, generate $y_{mis}^{*(k)}$ from $f(y_{mis} \mid x, y_{obs})$.

   (b) Apply the k-th imputed values to $\hat{\theta}_n$ in (1) to obtain $\hat{\theta}_n^{*(k)} = \hat{\theta}_n(x, y_{obs}, y_{mis}^{*(k)})$. Also, apply the k-th imputed values to $\hat{V}_n$ in (2) to obtain $\hat{V}_n^{*(k)} = \hat{V}_n(x, y_{obs}, y_{mis}^{*(k)})$.

   (c) Combine the M point estimators to get

   $$\hat{\theta}_{MI} = \frac{1}{M}\sum_{k=1}^{M} \hat{\theta}_n^{*(k)}$$

   as a point estimator of θ.

   (d) The variance estimator of $\hat{\theta}_{MI}$ is

   $$\hat{V}_{MI} = W_M + \left(1 + \frac{1}{M}\right) B_M$$

   where

   $$\begin{aligned} W_M &= \frac{1}{M}\sum_{k=1}^{M} \hat{V}_n^{*(k)} \\ B_M &= \frac{1}{M-1}\sum_{k=1}^{M} \left(\hat{\theta}_n^{*(k)} - \bar{\theta}_{MI}\right)^2. \end{aligned}$$

Comparison

	Bayesian	Frequentist
Model	Posterior distribution $f(\text{latent}, \theta \mid \text{data})$	Prediction model $f(\text{latent} \mid \text{data}, \theta)$
Computation	Data augmentation	EM algorithm
Prediction	I-step	E-step
Parameter update	P-step	M-step
Parameter est'n	Posterior mode	ML estimation
Imputation	Multiple imputation	Fractional imputation
Variance estimation	Rubin's formula	Linearization or Bootstrap

2 Main Result

1. Bayesian Properties (Rubin, 1987): For sufficiently large M, we have

$$
\begin{aligned}
\hat{\theta}_{MI} &= \frac{1}{M}\sum_{k=1}^{M}\hat{\theta}_n^{*(k)} \\
&= \frac{1}{M}\sum_{k=1}^{M}E(\theta \mid x, y_{obs}, y_{mis}^{*(k)}) \\
&\doteq E\{E(\theta \mid x, y_{obs}, Y_{mis}) \mid x, y_{obs}\} \\
&= E(\theta \mid x, y_{obs}),
\end{aligned}
$$

which is equal to $\hat{\theta}_r$ in (3). Also, for sufficiently large M,

$$
\begin{aligned}
\hat{V}_{MI} &= W_M + B_M \\
&= \frac{1}{M}\sum_{k=1}^{M}\hat{V}_n^{*(k)} + \frac{1}{M-1}\sum_{k=1}^{M}\left(\hat{\theta}_n^{*(k)} - \bar{\theta}_{MI}\right)^2 \\
&\doteq E\{V(\theta \mid x, y_{obs}, Y_{mis}) \mid x, y_{obs}\} + V\{E(\theta \mid x, y_{obs}, Y_{mis}) \mid x, y_{obs}\},
\end{aligned}
$$

which is equal to $\hat{V}_r$ in (4)

2. Frequentist Properties (Wang and Robins, 1998)

Assume that, under complete data, $\hat{\theta}_n$ is the MLE of θ and $\hat{V}_n = \{I(\hat{\theta}_n)\}^{-1}$ is asymptotically unbiased for $V(\hat{\theta}_n)$. Under the existence of missing data, $\hat{\theta}_{MI}$ is asymptotically equivalent to the MLE of θ and $\hat{V}_{MI}$ is approximately unbiased for $V(\hat{\theta}_{MI})$. That is,

$$
\hat{\theta}_{MI} \cong \hat{\theta}_{MLE} \tag{5}
$$

and

$$
E\{\hat{V}_{MI}\} \cong V(\hat{\theta}_{MI}) \tag{6}
$$

for sufficiently large M and n. (See Appendix A for a sketched proof.)

3 Computation

- Gibbs sampling
  - Geman and Geman (1984): the "Gibbs sampler" for Bayesian image reconstruction
  - Tanner and Wong (1987): data augmentation for Bayesian inference in generic missing-data problems
  - Gelfand and Smith (1990): simulation of marginal distributions by repeated draws from conditionals

- Idea for Gibbs sampling: Sample from conditional distributions

  Given $Z^{(t)} = \left(Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_J^{(t)}\right)$, draw $Z^{(t+1)}$ by sampling from the full conditionals of f,

$$\begin{aligned}
Z_1^{(t+1)} &\sim P\left(Z_1 \mid Z_2^{(t)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
Z_2^{(t+1)} &\sim P\left(Z_2 \mid Z_1^{(t)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
&\vdots \\
Z_J^{(t+1)} &\sim P\left(Z_J \mid Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_{J-1}^{(t)}\right).
\end{aligned}$$

  Under mild regularity conditions, $P\left(Z^{(t)}\right) \to f$ as $t \to \infty$.

- Data augmentation: Application of the Gibbs sampling to missing data problem

$$\begin{aligned}
y &= \text{observed data} \\
z &= \text{missing data} \\
\theta &= \text{model parameters}
\end{aligned}$$

  Predictive distribution:

$$P(z \mid y) = \int P(z \mid y, \theta)\, dP(\theta \mid y)$$

  Posterior distribution:

$$P(\theta \mid y) = \int P(y \mid y, z)\, dP(z \mid y)$$

- Algorithm: Iterative method of data augmentation

  I-step: Draw
$$z^{(t+1)} \sim P\left(z \mid y, \theta^{(t)}\right)$$
  P-step: Draw
$$\theta^{(t+1)} \sim P\left(\theta \mid y, z^{(t+1)}\right).$$

- 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 z,
$$z^{(t)}, z^{(2t)}, \cdots, z^{(mt)}$$

- Parameter simulation (Bayesian approach)

$$\theta_1 = \text{ component of function of } \theta \text{ of interest}$$

  Collect iterates of θ_1 from data augmentation
$$\theta_1^{(t+1)}, \theta_1^{(t+2)}, \cdots, \theta_1^{(t+N)},$$
  where t is large enough to ensure stationarity and N is the Monte Carlo sample size.

  - $\bar{\theta}_1 = N^{-1} \sum_{k=1}^{N} \theta_1^{(t+k)}$ estimates the posterior mean $E(\theta \mid y)$.
  - $N^{-1} \sum_{k=1}^{N} \left(\theta_1^{(t+k)} - \theta_1\right)^2$ estimates the posterior variance $V(\theta_1 \mid y)$.
  - The 2.5th and 97.5th percentiles of $\theta_1^{(t+1)}, \theta_1^{(t+2)}, \cdots, \theta_1^{(t+N)}$ estimate the endpoints of a 95% equal-tailed Bayesian interval for θ_1.

4 Examples

4.1 Example 1(Univariate Normal distribution)

- Let $y_1, \cdots, y_n$ be IID observations from $N(\mu, \sigma^2)$ and only the first r elements are observed and the remaining $n-r$ elements are missing. Assume that the response mechanism is ignorable.

- Bayesian imputation: the j-th posterior values of (μ, σ^2) are generated from

$$\sigma^{*(j)2} \mid \mathbf{y}_r \sim r\hat{\sigma}_r^2/\chi_{r-1}^2 \tag{7}$$

and

$$\mu^{*(j)} \mid (\mathbf{y}_r, \sigma^{*(j)2}) \sim N\left(\bar{y}_r, r^{-1}\sigma^{*(j)2}\right) \tag{8}$$

where $\mathbf{y}_r = (y_1, \cdots, y_r)$, $\bar{y}_r = r^{-1}\sum_{i=1}^r y_i$, and $\hat{\sigma}_r^2 = r^{-1}\sum_{i=1}^r (y_i - \bar{y}_r)^2$. Given the posterior sample $(\mu^{*(j)}, \sigma^{*(j)2})$, the imputed values are generated from

$$y_i^{*(j)} \mid \left(\mathbf{y}_r, \mu^{*(j)}, \sigma^{*(j)2}\right) \sim N\left(\mu^{*(j)}, \sigma^{*(j)2}\right) \tag{9}$$

independently for $i = r+1, \cdots, n$.

- Let $\theta = E(Y)$ be the parameter of interest and the MI estimator of θ can be expressed as

$$\hat{\theta}_{MI} = \frac{1}{M}\sum_{j=1}^M \hat{\theta}_I^{(j)}$$

where

$$\hat{\theta}_I^{(j)} = \frac{1}{n}\left\{\sum_{i=1}^r y_i + \sum_{i=r+1}^n y_i^{*(j)}\right\}.$$

Then,

$$\hat{\theta}_{MI} = \bar{y}_r + \frac{n-r}{nM}\sum_{j=1}^M \left(\mu^{*(j)} - \bar{y}_r\right) + \frac{1}{nM}\sum_{i=r+1}^n\sum_{j=1}^M \left(y_i^{*(j)} - \mu^{*(j)}\right). \tag{10}$$

Asymptotically, the first term has mean μ and variance $r^{-1}\sigma^2$, the second term has mean zero and variance $(1-r/n)^2\sigma^2/(mr)$, the third term has mean zero and variance $\sigma^2(n-r)/(n^2m)$, and the three terms are mutually independent. Thus, the variance of $\hat{\theta}_{MI}$ is

$$V\left(\hat{\theta}_{MI}\right) = \frac{1}{r}\sigma^2 + \frac{1}{M}\left(\frac{n-r}{n}\right)^2\left(\frac{1}{r}\sigma^2 + \frac{1}{n-r}\sigma^2\right). \tag{11}$$

- For variance estimation, note that

$$
\begin{aligned}
V(y_i^{*(j)}) &= V(\bar{y}_r) + V(\mu^{*(j)} - \bar{y}_r) + V(y_i^{*(j)} - \mu^{*(j)}) \\
&= \frac{1}{r}\sigma^2 + \frac{1}{r}\sigma^2 \left(\frac{r+1}{r-1}\right) + \sigma^2 \left(\frac{r+1}{r-1}\right) \\
&\cong \sigma^2.
\end{aligned}
$$

Writing

$$
\begin{aligned}
\hat{V}_I^{(j)}(\hat{\theta}) &= n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \left\{ \tilde{y}_i^{*(j)} - \frac{1}{n} \sum_{k=1}^{n} \tilde{y}_k^{*(j)} \right\}^2 \\
&= n^{-1}(n-1)^{-1} \left\{ \sum_{i=1}^{n} \left(\tilde{y}_i^{*(j)} - \mu \right)^2 - n \left(\frac{1}{n} \sum_{k=1}^{n} \tilde{y}_k^{*(j)} - \mu \right)^2 \right\}
\end{aligned}
$$

where $\tilde{y}_i^* = \delta_i y_i + (1-\delta_i) y_i^{*(j)}$, we have

$$
\begin{aligned}
E\left\{ \hat{V}_I^{(j)}(\hat{\theta}) \right\} &= n^{-1}(n-1)^{-1} \left\{ \sum_{i=1}^{n} E\left(\tilde{y}_i^{*(j)} - \mu \right)^2 - nV\left(\frac{1}{n} \sum_{k=1}^{n} \tilde{y}_k^{*(j)} \right) \right\} \\
&\cong n^{-1}(n-1)^{-1} \left[n\sigma^2 - n \left\{ \frac{1}{r}\sigma^2 + \left(\frac{n-r}{n} \right)^2 \left(\frac{1}{r}\sigma^2 + \frac{1}{n-r}\sigma^2 \right) \right\} \right] \\
&\cong n^{-1}\sigma^2
\end{aligned}
$$

which shows that $E(W_M) \cong V(\hat{\theta}_n)$. Also,

$$
\begin{aligned}
E(B_m) &= V\left(\hat{\theta}_I^{*(1)} \right) - Cov\left(\hat{\theta}_I^{*(1)}, \hat{\theta}_I^{*(2)} \right) \\
&= V\left\{ \frac{n-r}{n} \left(\mu^{*(1)} - \bar{y}_r \right) + \frac{1}{n} \sum_{i=r+1}^{n} \left(y_i^{*(1)} - \mu^{*(1)} \right) \right\} \\
&\cong \left(\frac{n-r}{n} \right)^2 \left(\frac{1}{r} + \frac{1}{n-r} \right) \sigma^2 \\
&= \left(\frac{1}{r} - \frac{1}{n} \right) \sigma^2.
\end{aligned}
$$

Thus, Rubin's variance estimator satisfies

$$
E\left\{ \hat{V}_{MI}(\hat{\theta}_{MI}) \right\} \cong \frac{1}{r}\sigma^2 + \frac{1}{M} \left(\frac{n-r}{n} \right)^2 \left(\frac{1}{r}\sigma^2 + \frac{1}{n-r}\sigma^2 \right) \cong V\left(\hat{\theta}_{MI} \right),
$$

which is consistent with the general result in (6).

4.2 Example 2 (Censored regression model, or Tobit model)

- Model

$$z_i = x_i'\beta + \epsilon_i \qquad \epsilon_i \sim N\left(0, \sigma^2\right)$$
$$y_i = \begin{cases} z_i & \text{if } z_i \geq c_i \\ c_i & \text{if } z_i < c_i. \end{cases}$$

- Data augmentation

1. I-step: Given $\theta^{(t)} = \left(\beta^{(t)}, \sigma^{(t)}\right)$, generate the imputed value (for $\delta_i = 0$) from

$$z_i^{(t+1)} = x_i'\beta^{(t)} + \epsilon_i^{(t)}$$

with

$$\epsilon_i^{(t)} \sim \frac{\phi\left(s\right)}{\Phi\left[\left(c_i - x_i'\beta^{(t)}\right)/\sigma^{(t)}\right]}$$

If $\delta_i = 1$, then $z_i^{(t+1)} = z_i$.

2. P-step: Given $\mathbf{z}^{(t+1)} = \left(z_1^{(t+1)}, z_2^{(t+1)}, \cdots, z_n^{(t+1)}\right)$, generate $\theta^{(t+1)}$ from

$$\theta^{(t+1)} \sim P\left(\theta \mid \mathbf{z}^{(t+1)}\right).$$

That is, generate

$$\sigma^{2(t+1)} \mid \mathbf{z}^{(t+1)} \sim \left(n-p\right)\hat{\sigma}^{2(t+1)}/\chi^2_{n-p}$$

and

$$\beta^{(t+1)} \mid \mathbf{z}^{(t+1)}, \sigma^{2(t+1)} \sim N\left[\hat{\beta}_n^{(t+1)}, \left(\sum_{i=1}^n x_i x_i'\right)^{-1}\sigma^{2(t+1)}\right],$$

where

$$\hat{\beta}_n^{(t+1)} = \left(\sum_{i=1}^n x_i x_i'\right)^{-1}\sum_{i=1}^n x_i z_i^{(t+1)}$$

and

$$\hat{\sigma}^{2(t+1)} = \left(n-p\right)^{-1}\mathbf{z}^{(t+1)\prime}\left(I - P_x\right)\mathbf{z}^{(t+1)}.$$

4.3 Example 3 (Bayesian bootstrap)

- Nonparametric approach to Bayesian imputation
- First proposed by Rubin (1981).
- Assume that an element of the population takes one of the values $d_1, \cdots, d_K$ with probability $p_1, \cdots, p_K$, respectively. That is, we assume

$$P(Y = d_k) = p_k, \quad \sum_{k=1}^{K} p_k = 1. \tag{12}$$

- Let $y_1, \cdots, y_n$ be an IID sample from (12) and let n_k be the number of y_i equal to d_k. The parameter is a vector of probabilities $\mathbf{p} = (p_1, \cdots, p_K)$, such that $\sum_{i=1}^{K} p_i = 1$. In this case, the population mean $\theta = E(Y)$ can be expressed as $\theta = \sum_{i=1}^{K} p_i d_i$ and we only need to estimate $\mathbf{p}$.
- If the improper Dirichlet prior with density proportional to $\prod_{k=1}^{K} p_k^{-1}$ is placed on the vector $\mathbf{p}$, then the posterior distribution of $\mathbf{p}$ is proportional to

$$\prod_{k=1}^{K} p_k^{n_k - 1}$$

  which is a Dirichlet distribution with parameter $(n_1, \cdots, n_K)$. This posterior distribution can be simulated using $n-1$ independent uniform random numbers. Let $u_1, \cdots, u_{n-1}$ be IID $U(0,1)$, and let $g_i = u_{(i)} - u_{(i-1)}, i = 1, 2, \cdots, n-1$ where $u_{(k)}$ is the k-th order statistic of $u_1, \cdots, u_{n-1}$ with $u_{(0)} = 0$ and $u_{(n)} = 1$. Partition the $g_1, \cdots, g_n$ into K collections, with the k-th one having n_k elements, and let p_k be the sum of the g_i in the k-th collection. Then, the realized value of $p_1, \cdots, p_k$ follows a $(K-1)$-variate Dirichlet distribution with parameter $(n_1, \cdots, n_K)$. In particular, if $K = n$, then $(g_1, \cdots, g_n)$ is the vector of probabilities to attach to the data values $y_1, \cdots, y_n$ in that Bayesian bootstrap replication.
- To implement Rubin's Bayesian bootstrap to multiple imputation, assume that the first r elements are observed and the remaining $n-r$ elements are missing. The imputed values can be generated with the following steps:

[Step 1] From $\mathbf{y}_r = (y_1, \cdots, y_r)$, generate $\mathbf{p}_r^* = (p_1^*, \cdots, p_r^*)$ from the posterior distribution using the Bayesian bootstrap as follows.

1. Generate $u_1, \cdots, u_{r-1}$ independently from $U(0,1)$ and sort them to get $0 = u_{(0)} < u_{(1)} < \cdots < u_{(r-1)} < u_{(r)} = 1$.
2. Compute $p_i^* = u_{(i)} - u_{(i-1)},\ i = 1, 2, \cdots, r-1$ and $p_r^* = 1 - \sum_{i=1}^{r-1} p_i^*$.

[Step 2] Select the imputed value of y_i by

$$y_i^* = \begin{cases} y_1 & \text{with probability } p_1^* \\ \cdots & \cdots \\ y_r & \text{with probability } p_r^* \end{cases}$$

independently for each $i = r+1, \cdots, n$.

- Rubin and Schenker (1986) proposed an approximation of this Bayesian bootstrap method, called the approximate Bayesian boostrap (ABB) method, which provides an alternative approach of generating imputed values from the empirical distribution. The ABB method can be described as follows:

[Step 1] From $\mathbf{y}_r = (y_1, \cdots, y_r)$, generate a donor set $\mathbf{y}_r^* = (y_1^*, \cdots, y_r^*)$ by bootstrapping. That is, we select

$$y_i^* = \begin{cases} y_1 & \text{with probability } 1/r \\ \cdots & \cdots \\ y_r & \text{with probability } 1/r \end{cases}$$

independently for each $i = 1, \cdots, r$.

[Step 2] From the donor set $\mathbf{y}_r^* = (y_1^*, \cdots, y_r^*)$, select an imputed value of y_i by

$$y_i^{**} = \begin{cases} y_1^* & \text{with probability } 1/r \\ \cdots & \cdots \\ y_r^* & \text{with probability } 1/r \end{cases}$$

independently for each $i = r+1, \cdots, n$.

Reference

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **62**, 721-741.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, **9**, 130-134.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366-374.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association* **82**, 528-540.

Wang, N. and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedure, *Biometrika*, **85**, 935-948.

Appendix

A. Proof of (5) and (6)

Let $S_{com}(\theta) = S(\theta; x, y)$ be the score function of θ under complete response. The MLE under complete response, denoted by $\hat{\theta}_n$, is asymptotically equivalent to

$$\hat{\theta}_n \cong \theta + \mathcal{I}_{com}^{-1} S_{com}(\theta),$$

where $\mathcal{I}_{com} = E\{-\partial S_{com}(\theta)/\partial\theta'\}$. Thus,

$$\begin{aligned}
\hat{\theta}_{MI} &= \frac{1}{M}\sum_{k=1}^{M} \hat{\theta}_n^{*(k)} \\
&\doteq \theta + \mathcal{I}_{com}^{-1}\frac{1}{M}\sum_{k=1}^{M} S(\theta; x, y_{obs}, y_{mis}^{*(k)}) \\
&= \theta + \mathcal{I}_{com}^{-1}\frac{1}{M}\sum_{k=1}^{M} E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta^{*(k)}\} \\
&+ \mathcal{I}_{com}^{-1}\frac{1}{M}\sum_{k=1}^{M}\left[S(\theta; x, y_{obs}, y_{mis}^{*(k)}) - E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta^{*(k)}\}\right],
\end{aligned}$$

where $y_{mis}^{*(k)} \sim f(y_{mis} \mid x, y_{obs}; \theta^{*(k)})$ and $\theta^{*(k)} \sim p(\theta \mid x, y_{obs})$. Under some regularity conditions, the posterior distribution converges to a normal distribution with mean $\hat{\theta}_{MLE}$ and variance $I_{obs}^{-1} = V(\hat{\theta}_{MLE})$. (This is often called Bernstein–von Mises theorem.) Thus, we can apply Taylor linearization on $S(\theta; x, y_{obs}, y_{mis}^{*(k)})$ with respect to $\theta^{*(k)}$ around the true θ to get

$$\begin{aligned}
E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta^{*(k)}\} &\cong E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta\} + \mathcal{I}_{mis}(\theta^{*(k)} - \theta) \\
&= S_{obs}(\theta) + \mathcal{I}_{mis}(\hat{\theta}_{MLE} - \theta) + \mathcal{I}_{mis}(\theta^{*(k)} - \hat{\theta}_{MLE}),
\end{aligned} \tag{13}$$

where $\mathcal{I}_{mis}$ is the information matrix associated with $f(y_{mis} \mid x, y_{obs}; \theta)$. Since $\hat{\theta}_{MLE}$ is the solution to $S_{obs}(\theta) = 0$, we have

$$\hat{\theta}_{MLE} \cong \theta + \mathcal{I}_{obs}^{-1} S_{obs}(\theta)$$

and (13) further simplifies to

$$\begin{aligned}
E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta^{*(k)}\} &= S_{obs}(\theta) + \mathcal{I}_{mis}\mathcal{I}_{obs}^{-1} S_{obs}(\theta) + \mathcal{I}_{mis}(\theta^{*(k)} - \hat{\theta}_{MLE}) \\
&= \mathcal{I}_{com}\mathcal{I}_{obs}^{-1} S_{obs}(\theta) + \mathcal{I}_{mis}(\theta^{*(k)} - \hat{\theta}_{MLE}).
\end{aligned} \tag{14}$$

Thus, combining all terms together, we have

$$\hat{\theta}_n^{*(k)} = \hat{\theta}_{MLE} + \mathcal{I}_{com}^{-1}\mathcal{I}_{mis}(\theta^{*(k)} - \hat{\theta}_{MLE}) + \mathcal{I}_{com}^{-1}\left\{S^{*(k)} - E(S \mid x, y_{obs}; \theta^{*(k)})\right\} \quad (15)$$

where $S^{*(k)} = S(\theta; x, y_{obs}, y_{mis}^{*(k)})$. Therefore,

$$\begin{aligned}
\hat{\theta}_{MI} &= \hat{\theta}_{MLE} + \mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\frac{1}{M}\sum_{k=1}^{M}(\theta^{*(k)} - \hat{\theta}_{MLE}) \\
&+ \mathcal{I}_{com}^{-1}\frac{1}{M}\sum_{k=1}^{M}\left[S(\theta; x, y_{obs}, y_{mis}^{*(k)}) - E\{S(\theta; x, y_{obs}, Y_{mis}) \mid x, y_{obs}; \theta^{*(k)}\}\right] \\
&= \hat{\theta}_{MLE} + \mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\left\{M^{-1}\sum_{k=1}^{M}(\theta^{*(k)} - \hat{\theta}_{MLE})\right\} \\
&\quad +\mathcal{I}_{com}^{-1}M^{-1}\sum_{k=1}^{M}\left\{S^{*(k)}(\theta) - E(S \mid x, y_{obs}; \theta^{*(k)})\right\}.
\end{aligned}$$

Note that the second term reflects the variability due to generating θ^* and the third therm reflects the variability due to generating $y_{mis}^{*(k)}$ from $f(y_{mis} \mid x, y_{obs}; \theta^{*(k)})$. The three terms are independent and so we obtain

$$V\{\hat{\theta}_{MI}\} = V(\hat{\theta}_{MLE}) + \frac{1}{M}\mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\mathcal{I}_{obs}^{-1}\mathcal{I}_{mis}\mathcal{I}_{com}^{-1} + \frac{1}{M}\mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\mathcal{I}_{com}^{-1}. \quad (16)$$

The last two terms are negligible for large M.

To prove (6), note first that $E(V_n) = V(\hat{\theta}_n) = \mathcal{I}_{com}^{-1}$ and so we have $E(W_M) \cong I_{com}^{-1}$. Now, inserting (15) into

$$B_M = \frac{1}{M-1}\sum_{k=1}^{M}(\hat{\theta}_n^{*(k)} - \hat{\theta}_{MI})^{\otimes 2},$$

we have

$$\begin{aligned}
E(B_M) &= V(\hat{\theta}_n^{*(1)}) - Cov(\hat{\theta}_n^{*(1)}, \hat{\theta}_n^{*(2)}) \\
&= \mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\mathcal{I}_{obs}^{-1}\mathcal{I}_{mis}\mathcal{I}_{com}^{-1} + \mathcal{I}_{com}^{-1}\mathcal{I}_{mis}\mathcal{I}_{com}^{-1} \\
&= \mathcal{I}_{obs}^{-1} - \mathcal{I}_{com}^{-1},
\end{aligned}$$

where the last equality follows from

$$(A + BCB')^{-1} = A^{-1} - A^{-1}BCB'A^{-1} + A^{-1}BCB'(A + BCB')^{-1}BCB'A^{-1}$$

with $A = \mathcal{I}_{com}$, $B = I$, and $C = -\mathcal{I}_{mis}$. Therefore, $E(\hat{V}_{MI}) \cong \mathcal{I}_{obs}^{-1} = V(\hat{\theta}_{MLE}) \cong V(\hat{\theta}_{MI})$.