Biostat 2065
Analysis of Incomplete Data

Gong Tang
Dept of Biostatistics
University of Pittsburgh

October 25, 2005
1. Data augmentation

With an ignorable missing-data mechanism, bayesian inference is based on the posterior distribution

\[ p(\theta \mid Y_{\text{obs}}) \propto p(\theta) f(Y_{\text{obs}}; \theta), \]

where \( p(\theta) \) is the prior distribution; \( f(Y_{\text{obs}}; \theta) \) is the observed likelihood function. A random sample are drawn from the posterior distribution of \( \theta \) and the sample property can be used to make inference on \( \theta \).

However, the posterior distribution \( p(\theta \mid Y_{\text{obs}}) \) may be very complicated or difficult to draw from. A method called data augmentation was proposed to avoid such difficulty. This method iteratively simulates random samples of the missing values and model parameter given the observed data with each iteration consisting of an imputation (I) step and a posterior (P) step.

Start with an initial draw \( \theta^{(0)} \) from an approximation to the posterior distribution of \( \theta \). Given a value of \( \theta^{(t)} \) of \( \theta \) drawn at iteration \( t \):

I Step: Draw \( Y_{mis}^{(t+1)} \) with density \( p(Y_{mis} \mid Y_{obs}, \theta^{(t)}) \);

P Step: Draw \( \theta^{(t+1)} \) with density \( p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}) \).

This iterative procedure will eventually yield a draw from the joint distribution of \( (Y_{mis}, \theta \mid Y_{obs}) \), as \( t \to \infty \).
Example 10.2. The one-parameter multinomial model.
2. The Gibbs’ sampler

Sometimes it is not easy to even draw from the distribution of \((Y_{mis} \mid Y_{obs}, \theta)\). Gibbs’ sampler is an iterative procedure where draw from a complicated joint distribution is factored into sequential draws from easily obtained conditional distributions.

For example, the target is the joint distribution of \(X_1, X_2, \ldots, X_p\) where the joint distribution is hard to compute but the conditional distributions \(p(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)\) are relatively easy to compute, \(j = 1, \ldots, p\). In Gibbs’ sampler, initial values \(\{x_1^{(0)}, \ldots, x_p^{(0)}\}\) are chosen in some way. Then with current values \(\{x_1^{(t)}, \ldots, x_p^{(t)}\}\) at iteration \(t\), draw sequentially from the following conditional distributions:

\[
\begin{align*}
x_1^{(t+1)} &\sim p(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)}), \\
x_2^{(t+1)} &\sim p(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}), \\
x_3^{(t+1)} &\sim p(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_p^{(t)}), \\
\vdots \\
x_p^{(t+1)} &\sim p(x_p \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}),
\end{align*}
\]

Under quite general conditions, the sequence \(x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})\) converges to a draw from the joint distribution.

Usually several Gibbs’ samplers need to be run in order to generate a random sample of the joint distribution.
Example 10.3. A multivariate normal regression model with incomplete data. Model:

\[ y_i \sim N_K(X_i\beta, \Sigma), \quad \text{with prior } p(\beta, \Sigma) \propto |\Sigma|^{-(K+1)/2}. \]

Each iteration of the Gibbs’ sampler consists of three steps: (I) for imputing random draws of \( Y_{mis} \); (CP1) for conditionally drawing \( \beta \) given \( Y \) and \( \Sigma \); (CP2) for conditionally drawing \( \Sigma \) given \( Y \) and \( \beta \). Let \((Y_{mis}^{(d,t)}, \beta^{(d,t)}, \Sigma^{(d,t)})\) be draws of the missing data and model parameters at iteration \( t \) of the \( d \)th Gibbs’ sampler. Then

I Step: draw from the conditional distribution of \( Y_{mis} \) given \((Y_{obs}, \beta^{(d,t)}, \Sigma^{(d,t)})\), a multivariate normal distribution.

CP1 Step: Draw \( \beta^{(d,t+1)} \) from a multivariate normal distribution with mean

\[
\hat{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^{n} X_i^T (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^{n} X_i^T (\Sigma^{(d,t)})^{-1} Y_i^{(d,t)} \right\},
\]

and variance-covariance matrix \( \left\{ \sum_{i=1}^{n} X_i^T (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1} \).

CP2 Step: Draw \( \Sigma^{(d,t+1)} \) from a inverse Wishart distribution with scale matrix

\[
\Sigma^{(t+1)} = n^{-1} \sum_{i=1}^{n} (Y_i^{(d,t)} - X_i\beta^{(d,t+1)})(Y_i^{(d,t)} - X_i\beta^{(d,t+1)})^T.
\]

and degrees of freedom \( n \).
3. Assessing convergence of iterative simulations

If the DA or Gibbs’ sampler are not iterated long enough, the simulated distribution may not represent the posterior distribution well.

It’s recommended to simulated $D > 1$ sequences with starting values dispersed throughout the parameter space and compare variation between and within simulated sequences. Iterate until these two variations are roughly equal.

For example, the model parameter is a scalar and $D$ parallel sequences have been drawn $T$ times: \( \{ \psi^{d,t}, d = 1, \ldots, D; t = 1, \ldots, T. \} \). Then compute

\[
B = \frac{T}{D - 1} \sum_{d=1}^{D} (\bar{\psi}_d. - \bar{\psi}..)^2 \quad \text{and} \quad \bar{V} = \frac{1}{D} \sum_{d=1}^{D} s^2_d, \]

where \( \bar{\psi}_d. = \frac{1}{T} \sum_{t=1}^{T} \psi_{d,t}, \quad \bar{\psi}.. = \frac{1}{D} \sum_{d=1}^{D} \bar{\psi}_d. \), \( s^2_d = \frac{1}{T - 1} \sum_{t=1}^{T} (\psi_{d,t} - \bar{\psi}_d.)^2 \)

The marginal posterior variance \( \text{var}(\psi \mid Y_{\text{obs}}) \) can be estimated by

\[
\widehat{\text{var}}^{+}(\psi \mid Y_{\text{obs}}) = \frac{T - 1}{T} \bar{V} + \frac{1}{T} B.
\]

The monitoring of convergence can be done by checking whether the following value is close to 1:

\[
\sqrt{\hat{R}} = \sqrt{\widehat{\text{var}}^{+}(\psi \mid Y_{\text{obs}}) / \bar{V}}.
\]
4. Multiple imputation

Previous procedures usually require at least modest number of draws from the joint posterior distribution in order to make inference.

In cases where inference from the complete-data posterior distribution is based on multivariate normality (or the multivariate \( t \)), posterior moments of model parameter \( \theta \) can be reliably estimated from small number of draws of the missing data, i.e., even when \( D < 10 \), given the fraction of missing information is not too large.

Based on

\[
p(\theta \mid Y_{obs}) = \int p(\theta, Y_{mis} \mid Y_{obs}) \, dY_{mis} \\
= \int p(\theta \mid Y_{mis}, Y_{obs}) p(Y_{mis}, \mid Y_{obs}) \, dY_{mis}.
\]

After drawing the missing values, \( Y_{mis}^{(d)} \), from their joint posterior distribution, then \( \theta \) can be drawn from the complete-data posterior distribution \( p(\theta \mid Y_{obs}, Y_{mis}^{(d)}) \). The posterior distribution of \( \theta \) can be approximated by

\[
p(\theta \mid Y_{obs}) \approx \frac{1}{D} \sum_{d=1}^{D} p(\theta \mid Y_{mis}^{(d)}, Y_{obs}),
\]

and

Biostat 2065: Analysis of Incomplete Data, 2005
\[
E(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^{D} E(\theta \mid Y_{\text{mis}}^{(d)}, Y_{\text{obs}}) = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d = \bar{\theta},
\]

\[
\text{Var}(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^{D} V_d + \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2 = \bar{V} + B,
\]

where \( V_d = \text{var}(\theta \mid Y_{\text{mis}}^{(d)}, Y_{\text{obs}}) \).

An improved approximation is:

\[
\text{VarVar}(\theta \mid Y_{\text{obs}}) \approx \bar{V} + (1 + D^{-1})B.
\]

The ratio of estimated between-imputation to total variance,

\[
\hat{\gamma}_d = \frac{(1 + D^{-1})B}{\bar{V} + (1 + D^{-1})B},
\]

estimates the fraction of missing information.