CHAPTER 10

Bayes and Multiple Imputation

10.1. BAYESIAN ITERATIVE SIMULATION METHODS

10.1.1. Data Augmentation

When sample sizes are small, a useful alternative approach to ML is to add a prior distribution for the parameters and compute the posterior distribution of the parameters of interest. We have already been introduced to this approach in Section 6.1.4 with complete data, and with incomplete data through Example 7.3 and Section 7.4.4 in the special case of multivariate normal data with a monotone missing-data pattern.

The posterior distribution for a model with an ignorable missing-data mechanism is:

$$p(\theta|Y_{\text{obs}}, M) \equiv p(\theta|Y_{\text{obs}}) = \text{constant} \times p(\theta) \times f(Y_{\text{obs}}|\theta), \tag{10.1}$$

where $p(\theta)$ is the prior distribution and $f(Y_{\text{obs}}|\theta)$ is the density of the observed data. In the examples of Chapter 7, simulation from the posterior distribution could be accomplished without iteration. Specifically, the likelihood was factored into complete-data components,

$$\text{L}(\phi|Y_{\text{obs}}) = \prod_{q=1}^{Q} \text{L}_q(\phi_q|Y_{\text{obs}}),$$

and, assuming that the parameters $\phi_1, \ldots, \phi_Q$ were also *a priori* independent, the posterior distribution factored in an analogous way, with $\phi_1, \ldots, \phi_Q$ *a posteriori* independent. Consequently, draws $\phi^{(d)} = (\phi_1^{(d)}, \ldots, \phi_Q^{(d)})$ could be obtained directly from the factored complete-data posterior distribution. Draws of θ were then obtained as $\theta^{(d)} = \theta(\phi^{(d)})$, where $\theta(\phi)$ is the inverse transformation from ϕ to θ. With more general patterns of missing data or parameters ϕ_j that are not *a priori* independent, this method does not work. As with ML estimation with a general pattern of missing values, Bayes simulation requires iteration.

Data augmentation (Tanner and Wong, 1987)[1] is an iterative method of simulating the posterior distribution of θ that combines features of the EM algorithm and multiple imputation. It can be thought of as a small-sample refinement of the EM algorithm using simulation, with the imputation (or I) step corresponding to the E step and the posterior (or P) step corresponding to the M step. Start with an initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of θ. Given a value $\theta^{(t)}$ of θ drawn at iteration t:

I Step: Draw $Y_{\text{mis}}^{(t+1)}$ with density $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$;
P Step: Draw $\theta^{(t+1)}$ with density $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

The procedure is motivated by the fact that the distributions in these two steps are often much easier to draw from than either of the posterior distributions $p(Y_{\text{mis}}|Y_{\text{obs}})$ and $p(\theta|Y_{\text{obs}})$, or the joint posterior distribution $p(\theta, Y_{\text{mis}}|Y_{\text{obs}})$. The iterative procedure can be shown eventually to yield a draw from the joint posterior distribution of Y_{mis}, θ given Y_{obs}, in the sense that as t tends to infinity, this sequence converges to a draw from the joint distribution of (θ, Y_{mis}) given Y_{obs}.

EXAMPLE 10.1. *Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missing Data (Example 8.3 continued).* Example 8.3 described the EM algorithm for a bivariate normal sample, with one group of units having Y_1 observed but Y_2 missing, a second group of units having both Y_1 and Y_2 observed, and the third group of units having Y_2 observed but Y_1 missing (see Figure 8.1). We now consider the DA algorithm for this example.

Each iteration t consists of an I step and a P step. The I step of DA is similar to the E step, except that each missing value is replaced by a draw from its conditional distribution given the observed data and the current values of the parameters, rather than by its conditional mean. Because units are independent given the parameters, each missing y_{i2} is drawn independently as

$$y_{i2}^{(t+1)} \sim_{\text{ind}} N(\beta_{20.1}^{(t)} + \beta_{21.1}^{(t)} y_{i1}, \sigma_{22.1}^{(t)}),$$

where $\beta_{20.1}^{(t)}$, $\beta_{21.1}^{(t)}$, and $\sigma_{22.1}^{(t)}$ are the tth iterates of the regression parameters of Y_2 on Y_1. Analogously, each missing y_{i1} is drawn independently as:

$$y_{i1}^{(t+1)} \sim_{\text{ind}} N(\beta_{10.2}^{(t)} + \beta_{12.2}^{(t)} y_{i2}, \sigma_{11.2}^{(t)}),$$

where $\beta_{10.2}^{(t)}$, $\beta_{12.2}^{(t)}$, and $\sigma_{11.2}^{(t)}$ are the tth iterates of the regression parameters of Y_1 on Y_2.

In the P step of DA, these drawn values of the missing data are treated as if they were the actual observed values of the data, and one draw of the bivariate normal parameters is made from the complete-data posterior distribution, given in Example

[1] The definition of data augmentation used here differs slightly from the original version, which involves a multiple imputation step at each iteration, followed by multiple draws of the parameters from the current estimate of the posterior distribution.

6.21. In the limit, the draws are from the joint posterior distribution of the missing data and the parameters. Thus one run of data augmentation generates both a draw from the posterior predictive distribution of Y_{mis} and a draw from the posterior distribution of θ. Data augmentation can be run independently D times to generate D iid draws from the approximate joint posterior distribution of θ and Y_{mis}. The values of Y_{mis} are multiple imputations of the missing values, drawn from their posterior predictive distribution.

Note that unlike EM, estimates of the covariance matrix from the filled-in data can be computed without adding corrections to the variances. The reason is that draws from the predictive distribution are imputed in the I step of DA, rather than conditional means in the E step of EM. The loss of efficiency from imputing draws is limited when the posterior mean from DA is computed by averaging over many draws from the posterior distribution, and hence over many imputed data sets.

EXAMPLE 10.2. *Bayesian Computations for One-Parameter Multinomial Model (Example 9.1 continued).* Example 9.1 applied EM and SEM to the one-parameter multinomial model of Example 8.2; slightly different asymptotic approximations underlie the calculations in the raw and logit scales. With DA, this distinction is avoided, although different prior distributions yield different posterior distributions. The I step of DA imputes y_3 and $y_4 = 125 - y_3$ assuming the drawn value of θ, $\theta^{(t)}$, is true. Specifically, the I step for iteration $(t+1)$ of DA draws

$$y_3^{(t+1)} \sim \text{Bin}[125, \theta^{(t)}/(\theta^{(t)} + 2)],$$

which is analogous to the E step of EM given by Eq. (8.10). The complete-data likelihood is proportional to

$$(1/2 - \theta/2)^{y_1}(\theta/4)^{y_2}(\theta/4)^{y_3}(1/2)^{y_4}.$$

Hence with a Beta (Dirichlet) prior distribution proportional to $\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$, the complete-data posterior distribution of θ is proportional to

$$\theta^{y_2 + y_3 + \alpha_1 - 1}(1-\theta)^{y_1 + \alpha_2 - 1},$$

which is Beta. The P step of DA draws from this Beta distribution, with y_1, y_2, and y_3 fixed at their values from the previous I step, that is,

$$\theta^{(t+1)} \sim \text{Beta}(y_2 + y_3^{(t+1)} + \alpha_1, y_1 + \alpha_2),$$

using gamma or chi-squared deviates, as described in Example 6.20. This P step is analogous to the M step of EM, Eq. (8.11).

Histograms of 90,000 draws from the posterior distribution of θ and logit(θ), for the Jeffreys' prior distribution with $\alpha_1 = \alpha_2 = 0.5$, are displayed in Figure 10.1. Note that the posterior distribution on the logit scale looks more normal, although even on the raw scale the normal approximation is not far off. Table 10.1

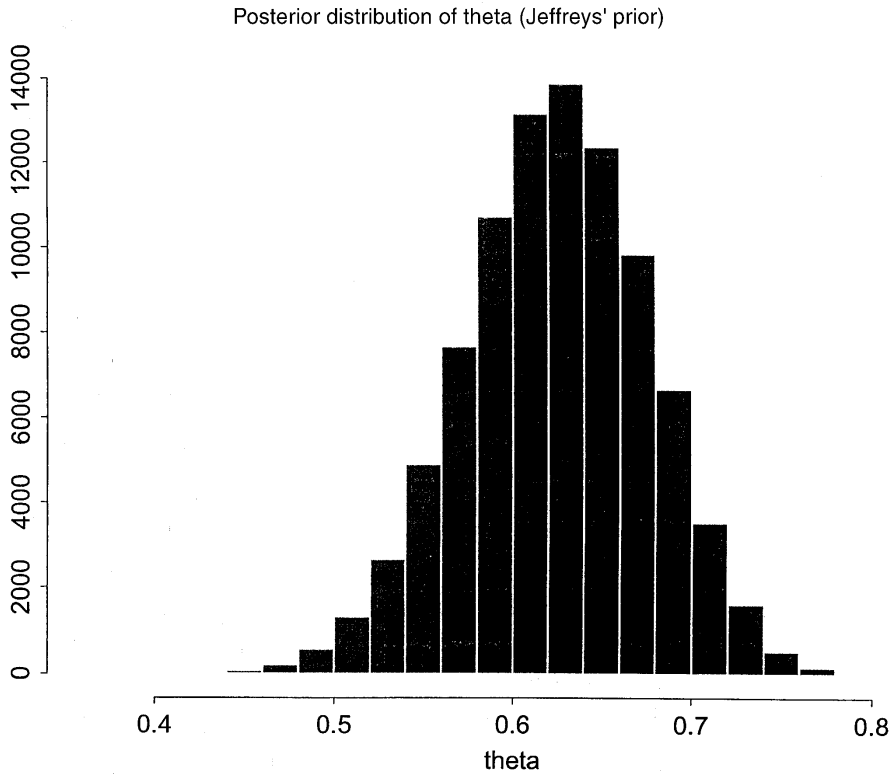

Figure 10.1. Posterior distribution of logit(θ) (Jeffreys' prior).

summarizes the posterior means and variances of θ and logit(θ) from this analysis, and the analysis based on the uniform prior for θ, $\alpha_1 = \alpha_2 = 1$. These are close to the ML estimate and asymptotic standard error from EM/SEM, displayed in the last column of the table.

10.1.2. The Gibbs' Sampler

The Gibbs' sampler is an iterative simulation method that eventually yields a draw from the joint distribution in the case of a general pattern of missing data, and

Table 10.1 Estimates from Bayesian and ML Analyses of Multinomial Example (Example 10.2)

Summary Quantity	Bayes, Jeffreys' Prior	Bayes, Uniform Prior	ML
Post. Mean/MLE of θ	0.624	0.623	0.626
Post. Var/Asympt SE of θ	0.00265	0.00258	0.00265
Post. Mean/MLE of logit θ	0.513	0.508	0.519
Post. Var/Asympt SE of logit θ	0.492	0.478	0.484

provides a Bayesian method analogous to the ECM algorithm for ML estimation. In some ways the Gibbs' sampler is simpler to understand than ECM because all of its steps involve draws of random variables.

The Gibbs' sampler eventually generates a draw from the distribution $P(x_1, \ldots, x_p)$ of a set of p random variables $X_1, \ldots, X_p$, in settings where draws from the joint distribution are hard to compute, but draws from conditional distributions $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_p), j = 1, \ldots, p$, are relatively easy to compute. Initial values $x_1^{(0)}, \ldots, x_p^{(0)}$ are chosen in some way. Then given values $x_1^{(t)}, \ldots, x_p^{(t)}$ at iteration t, new values are found by drawing from the following sequence of p conditional distributions:

$$
\begin{aligned}
x_1^{(t+1)} &\sim p(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\
x_2^{(t+1)} &\sim p(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\
x_3^{(t+1)} &\sim p(x_3|x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \ldots, x_p^{(t)}) \\
&\vdots \\
x_p^{(t+1)} &\sim p(x_p|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}).
\end{aligned}
$$

It can be shown that, under quite general conditions, the sequence of iterates $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$ converges to a draw from the joint distribution of $X_1, \ldots, X_p$. In this method, the individual components X_j can be sets of variables, not just scalar variables.

When $p = 2$, the Gibbs' sampler is essentially the same as data augmentation if $X_1 = Y_{\text{mis}}, X_2 = \theta$, and distributions condition on Y_{obs}. Then we can, in the limit, obtain a draw from the joint distribution of $(Y_{\text{mis}}, \theta|Y_{\text{obs}})$ by applying the Gibbs' sampler, where at iteration t for the dth imputed data set:

$$
Y_{\text{mis}}^{(d,t+1)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(d,t)}); \qquad \theta^{(d,t+1)} \sim p(\theta|Y_{\text{mis}}^{(d,t)}, Y_{\text{obs}}).
$$

As with DA, one run of Gibbs' iterates to a draw from the posterior predictive distribution of Y_{mis} and a draw from the posterior distribution of θ. The Gibbs' sampler can be run independently D times to generate D iid draws from the approximate joint posterior distribution of θ and Y_{mis}. The values of Y_{mis} are multiple imputations of the missing values, drawn from their posterior predictive distribution. The Gibbs' sampler can be used in more complex problems where DA is difficult to compute, but partitioning the missing data or the parameters into more than one piece helps computation. These ideas are illustrated by the following important example.

EXAMPLE 10.3. *A Multivariate Normal Regression Model with Incomplete Data (Example 8.6 continued).* Suppose we have n independent observations from the following K-variate normal model:

$$
y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma), \qquad i = 1, \ldots, n, \tag{10.2}
$$

where X_i is a known $(K \times p)$ design matrix for the ith observation, β is a $(p \times 1)$ vector of unknown regression coefficients, and Σ is a $(K \times K)$ unknown unstructured variance–covariance matrix. Example 8.6 discussed ML estimation for this problem. We assume the following Jeffreys' prior for the parameters $\theta = (\beta, \Sigma)$:

$$p(\beta, \Sigma) \propto |\Sigma|^{-(K+1)/2}.$$

Draws from the posterior distribution of θ can be obtained from the Gibbs' sampler, applied in three steps consisting of an imputation step (I) for Y_{mis} and two conditional posterior steps (CP1 and CP2) for drawing the values of β and Σ. Let $(Y_{\text{mis}}^{(d,t)}, \beta^{(d,t)}, \Sigma^{(d,t)})$ denote draws of the missing data and parameters after iteration t for creating multiple imputation d. The $(t+1)$th iteration then consists of the following three steps:

I Step: the conditional distribution of Y_{mis} given Y_{obs}, $\mu^{(d,t)}$ and $\Sigma^{(d,t)}$ is multivariate normal. Let $y_{\text{mis},i}$ denote the set of missing values in observation i. Then $y_{\text{mis},i}$ given Y_{obs}, $\beta^{(d,t)}$ and $\Sigma^{(d,t)}$ is independent over i, and multivariate normal with mean and residual covariance matrix based on the linear regression of $y_{\text{mis},i}$ on $y_{\text{obs},i}$ and X_i. Draws from this distribution are readily accomplished using the SWEEP operator, as discussed in detail in Section 11.2.

CP1 Step: The conditional distribution of β given Y_{obs}, $Y_{\text{mis}}^{(d,t)}$, and $\Sigma^{(d,t)}$ is normal with mean

$$\hat{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} Y_i^{(d,t)} \right\}, \tag{10.3}$$

where $Y_i^{(d,t)} = (Y_{\text{obs},i}, Y_{\text{mis},i}^{(d,t)})$, and covariance matrix

$$\Sigma_{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1}.$$

Hence $\beta^{(d,t+1)}$ is a random draw from this multivariate normal distribution.

CP2 Step: The conditional distribution of Σ given Y_{obs}, $Y_{\text{mis}}^{(d,t)}$, and $\beta^{(d,t+1)}$ is inverse Wishart with scale matrix given by the sum of squares and cross-products matrix of the residuals:

$$\Sigma^{(t+1)} = n^{-1} \sum_{i=1}^{n} (Y_i^{(d,t)} - X_i \beta^{(d,t+1)})(Y_i^{(d,t)} - X_i \beta^{(d,t+1)})^{\mathrm{T}} \tag{10.4}$$

and degrees of freedom n.

EXAMPLE 10.4. *Univariate* t *Sample with Known Degrees of Freedom (Example 8.10 continued).* In Example 8.10 we applied the PX-EM algorithm to compute ML estimates for the univariate t model (8.22) with known degrees of freedom v, by

embedding the observed data X in a larger data set (X, W) from the expanded complete-data model:

$$(x_i|\mu_*, \sigma_*, \alpha, w_i) \sim_{\text{ind}} N(\mu_*, \sigma_*^2/w_i), \qquad (w_i|\mu_*, \sigma_*, \alpha) \sim_{\text{ind}} \alpha\chi_v^2/v, \tag{10.5}$$

with parameters $\phi = (\mu_*, \sigma_*, \alpha)$. This model reduces to the original model (8.23) when $\alpha = 1$. Applying DA to this expanded model yields the Bayesian analog to PX-EM, which is called parameter-expanded data augmentation (PX-DA). The steps of PX-DA in this example are as follows:.

The PX-I step is analogous to the PX-E step (8.42) of PX-EM: at iteration $(t + 1)$, draw the missing data w_i conditionally given x_i and the current draw of parameters $\phi^{(t)}$. From the E step in Example 8.10, this distribution is:

$$w_i^{(t)} \sim_{\text{ind}} \chi_{v+1}^2/(v + d_i^{(t)2}), \tag{10.6}$$

where

$$d_i^{(t)} = \sqrt{\alpha^{(t)}}(x_i - \mu_*^{(t)})/\sigma_*^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)},$$

as in Eq. (8.42).

The PX-M step maximizes the expected complete-data loglikelihood of the expanded model with respect to ϕ. The PX-P step of PX-DA draws ϕ from its complete-data posterior distribution, which is normal-inverse chi-squared as described in Example 6.16.

In Chapter 12 we generalize this PX-DA algorithm to provide a form of robust Bayes inference for multivariate data with missing values.

10.1.3. Assessing Convergence of Iterative Simulations

If the DA or Gibbs' sampler iterations have not proceeded long enough, the simulations may be seriously unrepresentative of the target distribution. Assessing convergence of the sequence of draws to the target distribution is more difficult than assessing convergence of an EM-type algorithm to the ML estimate, since there is no single target quantity to monitor like the maximum value of the likelihood. Methods have been proposed for assessing convergence of a single sequence (see for example Geyer, 1992, and discussion). However, these methods are only recommended for well-understood models and straightforward data sets. A more reliable approach is to simulate $D > 1$ sequences with starting values dispersed throughout the parameter space. The convergence of all quantities of interest can then be monitored by comparing variation between and within simulated sequences, until within variation roughly equals between variation. Only when the distribution of each simulated sequence is close to the distribution of all the sequences mixed together can they all be approximating the target distribution.

Gelman and Rubin (1992) develop an explicit monitoring statistic based on this idea. For each scalar estimand ψ, label the draws from D parallel sequences as $\psi_{d,t}$ $(d = 1, \ldots, D, t = 1, \ldots, T)$, and compute B and $\bar{V}$, the between and within sequence variances:

$$B = \frac{T}{D-1} \sum_{d=1}^{D} (\bar{\psi}_{d\cdot} - \bar{\psi}_{\cdot\cdot})^2,$$

where

$$\bar{\psi}_{d\cdot} = \frac{1}{T} \sum_{t=1}^{T} \psi_{d,t}, \qquad \bar{\psi}_{\cdot\cdot} = \frac{1}{D} \sum_{d=1}^{D} \bar{\psi}_{d\cdot}.$$

$$\bar{V} = \frac{1}{D} \sum_{d=1}^{D} s_d^2,$$

where

$$s_d^2 = \frac{1}{T-1} \sum_{t=1}^{T} (\psi_{d,t} - \bar{\psi}_{d\cdot})^2.$$

We can estimate $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$, the marginal posterior variance of the estimand, by a weighted average of $\bar{V}$ and B, namely

$$\widehat{\mathrm{Var}}^{+}(\psi|Y_{\mathrm{obs}}) = \frac{T-1}{T}\bar{V} + \frac{1}{T}B,$$

which *overestimates* the marginal posterior variance assuming the starting distribution is appropriately overdispersed, but is *unbiased* under stationarity (that is, if the starting distribution equals the target distribution). This is analogous to the classical variance estimate for cluster sampling. For any finite T, the within variance $\bar{V}$ should be an *underestimate* of $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$ because individual sequences have not had time to range over all the target distribution, and, as a result, should have smaller variance than B; in the limit as $T \to \infty$, the expectation of $\bar{V}$ approaches $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$. These facts suggest monitoring convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for ψ might be reduced if the simulations were continued in the limit as $T \to \infty$. This potential scale reduction is estimated by

$$\sqrt{\hat{R}} = \sqrt{\widehat{\mathrm{Var}}^{+}(\psi|Y_{\mathrm{obs}})/\bar{V}},$$

which declines to 1 as $T \to \infty$. If the potential scale reduction is high, then there is evidence that proceeding with further simulations should improve our inference about the target distribution. Thus, if $\sqrt{\hat{R}}$ is not near one for all the estimands of

interest, the simulation runs should be continued, or perhaps the simulation algorithm itself should be altered to make the simulations more efficient. Once $\sqrt{\hat{R}}$ is near 1 for all scalar estimands of interest, subsequent draws from all the multiple sequences should be collected and treated as draws from the target distribution. The way the condition that $\sqrt{\hat{R}}$ is "near" 1 is implemented depends on the problem at hand; for most examples, values below 1.2 are acceptable, but for an important analysis or data set, a higher level of precision may be required.

It is useful to monitor convergence by computing $\sqrt{\hat{R}}$ for the logarithm of the posterior density, as well as for particular quantities of interest. When monitoring scalar quantities of interest, it is best to transform them to be approximately normal (for example, take logarithms of all-positive quantities and logits of quantities that lie between 0 and 1). Note that simulation inference from correlated draws is generally less precise than from the same number of independent draws, because of serial correlation within the run. If the simulation efficacy is unacceptably low (in the sense of requiring too much real time on any computer to obtain approximate convergence of posterior inference for quantities of interest), seek ways to alter the algorithm to speed convergence (Gelman et al., 1995, p. 330; Liu and Rubin, 1996, 2002).

10.1.4. Some Other Simulation Methods

When draws from the sequence of conditional distributions that form a Gibbs' algorithm are not easily computed, other simulation approaches are needed. Drawing from complicated multivariate distributions is a very rapidly developing field of statistics (Liu, 2001), with many applications outside what might be considered missing-data problems. However, a variety of the methods have their roots in the missing-data formulation, such as sequential imputation in computational biology (Kong, Liu and Wong, 1994; Liu and Chen, 1998). Here we give a brief overview of some of the main ideas, with references.

Suppose that draws of θ are sought from a target distribution $f(\theta)$, but are hard to compute. However, draws are easily obtained from an approximation to the target distribution, say $g(\theta)$, with the same support as $f(\theta)$, and both $f(\theta)$ and $g(\theta)$ can be evaluated up to some proportionality constant. For example, in the context of Bayesian inference, $f(\theta)$ may be the posterior distribution of logistic regression coefficients, and $g(\theta)$ could be its large sample normal approximation. A helpful idea involves the use of importance weights to improve the draws from $g(\theta)$, so they can be used as approximate draws from $f(\theta)$. Suppose D^* draws $\theta^*_1, \ldots, \theta^*_{D^*}$ are made from $g(\theta)$, where $D^* \gg D =$ the desired number of draws from $f(\theta)$, and let $R_d \propto f(\theta_d)/g(\theta_d)$. If D values of θ are drawn from the D^* draws $\theta^*_1, \ldots, \theta^*_{D^*}$ with probability proportional to the "importance" ratios or weights, R_d, then in the limit as $D/D^* \to 0$, the resulting D draws will be from $f(\theta)$.

This simple use of importance weights is known as Sampling Importance Resampling (SIR, see Rubin, 1987b; Gelfand and Smith, 1990; Smith and Gelfand, 1992). More sophisticated uses of these weights involve sequentially accepting or rejecting

the draws depending on whether R_d is greater than or less than some constant (rejection sampling, attributed to Von Neumann, 1951), or embedding rejection sampling within a Gibbs' sampler (the Metropolis–Hastings algorithm, see Metropolis et al., 1953; Hastings, 1970). The Gibbs' sampler and more complex extensions such as the Metropolis–Hastings algorithm are often referred to generically as "Markov Chain Monte Carlo" (MCMC) algorithms, because the sequence of iterates $\theta_{d,1}, \theta_{d,2}, \ldots$ forms a Markov chain. Gelman et al. (1995, Chapter 11) provide details.

The idea of using the draws from an incorrect distribution to build a bridge to the target distribution is the central idea behind bridge sampling, discussed in Meng and Wong (1996). An extension, path sampling, builds a path of distributions between the drawing distribution and the target distribution (Gelman and Meng, 1998).

Another approach for obtaining approximate draws from a target distribution is to create a set of initial independent parallel draws from a MCMC sequence and analyze them well before they have had a chance to converge to the target distribution. Assuming approximate normality of the target distribution, this estimation is straightforward (Liu and Rubin, 1996, 2002) and can be used to create a dramatically improved starting distribution. This "Markov normal" analysis may also reveal subspaces in which the proposed MCMC method is hopelessly slow to converge, and where alternative methods must be used.

10.2. MULTIPLE IMPUTATION

10.2.1. Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

The iterative simulation methods we have discussed eventually create draws from the posterior distribution of θ. If inferences for θ are based on the empirical distribution of the draws (for example, a 95% posterior interval of a parameter based on the 2.5 and 97.5 percentiles of the empirical distribution of that parameter), then a large number of independent draws is required, say in the thousands. If, on the other hand, we can assume approximate normality of the observed-data posterior distribution, we need only enough draws to reliably estimate the mean and variance of the posterior distribution, say a few hundred. Intermediate numbers of draws might suffice to estimate the posterior distribution by smoothing the empirical distribution, for example by fitting a parametric model such as the t family, or by semiparametric methods.

In those cases where inference from the complete-data posterior distribution is based on multivariate normality (or the multivariate t), posterior moments of θ can be reliably estimated from a surprisingly small number, D, of draws of the missing data Y_{mis} (e.g., $D = 2$–10), if the fraction of missing information is not too large. This approach creates D draws of (θ, Y_{mis}) and applies combining rules for multiple imputation introduced in Section 5.4.

The idea, first proposed in Rubin (1978b), is to relate the observed-data posterior distribution (10.1) to the complete-data posterior distribution that would have been obtained if we had observed the missing data Y_{mis}, namely:

$$p(\theta|Y_{\text{obs}}, Y_{\text{mis}}) \propto p(\theta)L(\theta|Y_{\text{obs}}, Y_{\text{mis}}). \tag{10.7}$$

Equations (10.1) and (10.7) can be related by standard probability theory as:

$$\begin{aligned} p(\theta|Y_{\text{obs}}) &= \int p(\theta, Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \\ &= \int p(\theta|Y_{\text{mis}}, Y_{\text{obs}})p(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}}. \end{aligned} \tag{10.8}$$

Equation (10.8) implies that the posterior distribution of θ, $p(\theta|Y_{\text{obs}})$, can be simulated by first drawing the missing values, $Y_{\text{mis}}^{(d)}$, from their joint posterior distribution, $p(Y_{\text{mis}}|Y_{\text{obs}})$, imputing the drawn values to complete the data set, and then drawing θ from its completed-data posterior distribution, $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$. When the posterior mean and variance are adequate summaries of the posterior distribution, Eq. (10.8) can be effectively replaced by

$$E(\theta|Y_{\text{obs}}) = E[E(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}], \tag{10.9}$$

and

$$\text{Var}(\theta|Y_{\text{obs}}) = E[\text{Var}(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}] + \text{Var}[E(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}]. \tag{10.10}$$

Multiple imputation effectively approximates the integral (10.8) over the missing values as the average:

$$p(\theta|Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} p(\theta|Y_{\text{mis}}^{(d)}, Y_{\text{obs}}), \tag{10.11}$$

where $Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}})$ are draws of Y_{mis} from the posterior predictive distribution of the missing values.

Similarly, the mean and variance equations (10.9) and (10.10) can be approximated, for large D, using the simulated values of Y_{mis} as follows:

$$E(\theta|Y_{\text{obs}}) \approx \int \theta \frac{1}{D}\sum_{d=1}^{D} p(\theta|Y_{\text{mis}}^{(d)}, Y_{\text{obs}})d\theta = \bar{\theta}, \tag{10.12}$$

where $\bar{\theta} = \sum_{d=1}^{D} \hat{\theta}_d / D$, and $\hat{\theta}_d = E(\theta | Y_{\text{mis}}^{(d)}, Y_{\text{obs}})$ is the estimate of θ from the dth completed data set, and for scalar θ:

$$\text{Var}(\theta | Y_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^{D} V_d + \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2 = \bar{V} + B, \tag{10.13}$$

say, where V_d is the complete-data posterior variance of θ calculated for the dth data set $(Y_{\text{mis}}^{(d)}, Y_{\text{obs}})$, $\bar{V} = \sum_{d=1}^{D} V_d / D$ is the average of V_d over the MI data sets, and $B = \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2 / (D-1)$ is the between-imputation variance. When D is small, the posterior mean is still approximated by Eq. (10.12), but an improved approximation for the posterior variance (10.13) is obtained by multiplying the between-imputation component by $(1 + D^{-1})$, that is:

$$\text{Var}(\theta | Y_{\text{obs}}) \approx \bar{V} + (1 + D^{-1}) B. \tag{10.14}$$

The ratio of estimated between-imputation to total variance, $\hat{\gamma}_D = (1 + D^{-1})B / (\bar{V} + (1 + D^{-1})B)$, estimates the fraction of missing information. For vector θ, the variance V_d is replaced by a covariance matrix, and $(\hat{\theta}_d - \bar{\theta})^2$ is replaced by $(\hat{\theta}_d - \bar{\theta})(\hat{\theta}_d - \bar{\theta})^{\text{T}}$.

A further refinement for small D is to replace the normal reference distribution by a t distribution with degrees of freedom given by

$$v = (D-1) \left(1 + \frac{D}{D+1} \frac{\bar{V}}{B} \right)^2. \tag{10.15}$$

When the completed data sets are based on limited degrees of freedom, say v_{com}, an additional refinement replaces v with:

$$v^* = (v^{-1} + \hat{v}_{\text{obs}}^{-1})^{-1},$$

where

$$\hat{v}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{v_{\text{com}} + 1}{v_{\text{com}} + 3} \right) v_{\text{com}}. \tag{10.16}$$

The theoretical basis for Eq. (10.15) is given in Rubin and Schenker (1986), and for Eq. (10.16) is given in Barnard and Rubin (1999).

EXAMPLE 10.5. *Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missing Data (Example 10.1 continued).* Suppose that the algorithm of Example 10.1 is run independently five times to create five joint draws of θ and Y_{mis}. Five draws are far too few to generate a reliable empirical distribution for estimating the actual posterior distribution of θ. However, the five draws of Y_{mis} can be quite adequate for generating MI inferences based on the

methods of this section, provided the fraction of missing information is modest, as when the fractions of cases with Y_1 or Y_2 missing are limited. In that case, the draws of Y_{mis} yield five completed data sets, the dth with sample means, variances, and covariance that we denote $\{(\bar{y}_1^{(d)}, \bar{y}_2^{(d)}, s_{11}^{(d)}, s_{22}^{(d)}, s_{12}^{(d)}), d = 1, \ldots, 5\}$. The resulting estimate of μ_1 from Eq. (10.12) is

$$\tilde{\mu}_1 = \sum_{d=1}^{5} \bar{y}_1^{(d)}/5,$$

With associated standard error from Eq. (10.14)

$$\text{Var}(\mu_1) = (1/5) \sum_{d=1}^{5} (s_{11}^{(d)}/n) + (6/5)(1/4) \sum_{d=1}^{5} (\bar{y}_1^{(d)} - \tilde{\mu}_1)^2.$$

If the original sample size n is large, a 95% interval estimate of μ_1 is given by

$$\tilde{\mu}_1 \pm t_{v,0.975} \sqrt{\text{Var}(\mu_1)},$$

where v is given by Eq. (10.15) with $D = 5$. For small n, the more refined approximation (10.16) should be used.

10.2.2. Approximations Using Test Statistics

In additional to interval estimation, it is often of interest to summarize the posterior distribution for a multi-component estimand by calculating a test statistic with an associated P value. Some multivariate analogs of the expressions given for scalar quantities are listed in Rubin (1987a, Section 3.4). Meng and Rubin (1992) developed methods for likelihood ratio testing when the available information consists of point estimates and the evaluation of the complete-data loglikelihood ratio statistic as a function of these estimates and the completed data. With large data sets and large models, such as in the common situation of a multiway contingency table, the complete-data analysis may produce only a test statistic or P value, and no parameter estimates. With such limited information, Rubin (1987a, Section 3.5) provided initial methods and Li et al. (1991) developed improved methods that require only the D completed-data chi-squared statistics (or equivalently, the D completed-data P values) that result from testing a null hypothesis using each of the D completed data sets. These methods, however, are less accurate than methods that use the completed-data statistics $\hat{\theta}_d, V_d$. Hence we start with a summary of the more accurate methods.

For θ with $k > 1$ components, significance levels for null values of θ can be obtained from D completed-data estimates, $\hat{\theta}_d, d = 1, \ldots, D$, and variance–covariance matrices, $V_d, d = 1, \ldots, D$, using multivariate analogs of the previous expressions. First, let θ_0 be the null value of θ, and let

$$W(\theta_0, \bar{\theta}) = \frac{(\theta_0 - \bar{\theta})^{\text{T}} \bar{V}^{-1} (\theta_0 - \bar{\theta})}{(1 + r)k}, \tag{10.17}$$

where $r = (1 + D^{-1})\ \text{trace}(B\bar{V}^{-1})/k$, and $\text{trace}(B\bar{V}^{-1})/k$ is the average diagonal element of $B\bar{V}^{-1}$. Equation (10.17) is an estimated Wald statistic, as defined in Section 6.1.3. The P value is then

$$\Pr[F_{k,\ell} > W(\theta_0, \bar{\theta})], \tag{10.18}$$

where $F_{k,\ell}$ is an F random variable with k and ℓ degrees of freedom with

$$\ell = 4 + (k(D-1) - 4)\left(1 + \frac{a}{r}\right)^2, \qquad a = \left\{1 - \frac{2}{k(D-1)}\right\}; \tag{10.19}$$

if $k(D-1) \leqslant 4$, let $\ell = (k+1)v/2$. Rubin (1987a) and Li, Raghunathan, and Rubin (1991) provide motivation for this test statistic and its reference distribution.

With large data sets and large models, such as occur often with multiway contingency-table data, each complete-data analysis may not produce the complete-data variance–covariance matrix V_d, but a P value for $\theta = \theta_0$ may still be desired. Two general methods are available, one asymptotically as precise as $W(\theta_0, \bar{\theta})$, and one less precise but simpler to use. We describe the more accurate method first.

Typically in multiparameter problems, in addition to the parameter of interest θ, there will be nuisance parameters ϕ, which are estimated by different values when $\theta = \theta_0$ than when $\theta \neq \theta_0$. Let $\hat{\phi}$ be the complete-data estimate of ϕ when $\theta = \hat{\theta}$, and $\hat{\phi}_0$ be the complete-data estimate of ϕ when $\theta = \theta_0$. Assume the complete-data analysis produces the estimates $(\hat{\theta}, \hat{\phi})$, the null estimates $(\theta_0, \hat{\phi}_0)$ and the P value for $\theta = \theta_0$ based on the likelihood-ratio χ^2 statistic,

$$\text{P value} = \Pr(\chi^2_k > \text{LR}), \tag{10.20}$$

where $\text{LR} = \text{LR}[(\hat{\theta}, \hat{\phi}), (\theta_0, \hat{\phi}_0)]$, using the notation of Section 6.1.3, and χ^2_k is a χ^2 random variable on k degrees of freedom. Let the average values of $\hat{\theta}, \hat{\phi}, \hat{\phi}_0$, and LR across the D sets of multiple imputations be denoted by $\bar{\theta}, \bar{\phi}, \bar{\phi}_0$, and $\overline{\text{LR}}$. Assume that the function LR can be evaluated for each of the D completed data sets at $\bar{\theta}, \bar{\phi}, \theta_0, \bar{\phi}_0$ to obtain D values of $\text{LR}[(\bar{\theta}, \bar{\phi}), (\theta_0, \bar{\phi}_0)]$ whose average across the D imputations is $\overline{\text{LR}}_0$. Then

$$\overline{\text{LR}}_0 \Big/ \left[k + \frac{(D+1)(\overline{\text{LR}} - \overline{\text{LR}}_0)}{(D-1)}\right] \tag{10.21}$$

is identical in large samples to $W(\theta_0, \bar{\theta})$ and can be used exactly as if it were $W(\theta_0, \bar{\theta})$ (Meng and Rubin, 1992).

In some cases, the complete-data method of analysis may not produce estimates of the general function $\text{LR}(\cdot, \cdot)$, but only the value of the likelihood ratio statistic, so that the multiple imputations result in D values $\text{LR}_1, \ldots, \text{LR}_D$. If so, the following

procedure due to Li et al. (1991) can be used. Let the repeated-imputation P value be $\Pr(F_{k,b} > \widetilde{\mathrm{LR}})$, where

$$\widetilde{\mathrm{LR}} = \frac{(\overline{\mathrm{LR}}/k) - (1 - D^{-1})\mathrm{v}}{1 + (1 + D^{-1})\mathrm{v}}, \tag{10.22}$$

and v is the sample variance of $(\mathrm{LR}_1^{1/2}, \ldots, \mathrm{LR}_D^{1/2})$, and

$$b = k^{-3/D}(D-1)\left\{1 + [(1 + D^{-1})\mathrm{v}]^{-1}\right\}^2. \tag{10.23}$$

10.2.3. Other Methods for Creating Multiple Imputations

We now return to the problem of creating the multiple imputations. The theory of the previous section suggests that we draw the missing values as

$$Y_{\mathrm{mis}}^{(d)} \sim p(Y_{\mathrm{mis}} | Y_{\mathrm{obs}}), \tag{10.24}$$

that is, from their joint posterior predictive distribution. Unfortunately, it is often difficult to draw from this predictive distribution in complicated problems, because of the implicit requirement in Eq. (10.24) to integrate over the parameters θ. Data augmentation accomplishes this by iteratively drawing a sequence of values of the parameters and missing data until convergence. Although this approach is theoretically preferable when the underlying model is well justified, in situations with multivariate data involving nonlinear relationships, building one coherent model for the joint distribution of the variables, programming the draws, and assessing convergence may be difficult and time-consuming. Simpler methods that approximate draws from Eq. (10.24), although less formally rigorous, may be easier to implement and yield approximately valid inferences when used in conjunction with the combining rules in Sections 10.2.1 and 10.2.2. Such methods may even be more effective than rigorous MI inference under a full model, if the full model is not a good reflection of the data.

A trivial example of an approximate method is to run the simulation for a fixed number of iterations or fixed time, without formally assessing convergence. We now list some other alternatives:

1. **Improper MI.** An approximate method is to draw:

$$Y_{\mathrm{mis}}^{(d)} \sim p(Y_{\mathrm{mis}} | Y_{\mathrm{obs}}, \tilde{\theta}), \tag{10.25}$$

where $\tilde{\theta}$ is an estimate of θ, for example the ML estimate, or an easy-to-compute estimate such as that from the complete cases. This is a reasonable approximation with small fractions of missing information, but Rubin (1987a, Chapter 4) shows that it does not provide valid frequentist inferences in general, since uncertainty in estimating θ is not propagated. Rubin (1987a) calls methods that do not propagate this uncertainty *improper*.

2. **Use the posterior distribution from a subset of the data.** Often it is relatively simple to draw θ from its posterior distribution based on a subset of the data close to the full data. The method propagates uncertainty about θ, but does not use all the available information to draw θ. For example, we have seen in Chapter 7 that the posterior distribution of θ may have a simple form for a monotone missing data pattern. This suggests discarding values to create a data set $Y_{\text{obs-mp}}$ with a monotone pattern, and then drawing θ from its posterior distribution given $Y_{\text{obs-mp}}$. Then draw $Y_{\text{mis}}^{(d)}$ as follows:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}), \tag{10.26}$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{obs-mp}}).$$

An even simpler but less accurate example of this approach is to draw θ from its posterior distribution given the complete cases, that is,

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}), \tag{10.27}$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{obs-cc}}),$$

where $Y_{\text{obs}-\text{cc}}$ represents data from the complete cases. For the multivariate normal problem with missing values, Eq. (10.27) can be viewed as a stochastic version of Buck's method (see Example 4.3), and is related to a class of pattern–mixture models involving complete-case missing value restrictions, as discussed in Little (1993b).

3. **Filling in data to create a monotone pattern.** In some situations where a monotone missing-data pattern is destroyed by a small number of missing values, an attractive option is to impute these nonmonotone missing values using one of the single imputation methods of Chapter 4, preferably as draws from an approximation to their posterior predictive distribution:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{aug-mp}}),$$

where $Y_{\text{aug-mp}}$ is the observed data augmented to create a monotone pattern. This method could be combined with method 2 in various ways.

4. **Use the asymptotic distribution of the ML estimate.** Suppose the ML estimate $\hat{\theta}$ of θ is available, together with a consistent estimate of its large-sample covariance matrix $C(\hat{\theta})$, as discussed in Section 6.1.2. Then $\theta^{(d)}$ can be drawn from its asymptotic normal posterior distribution:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} \sim N[\hat{\theta}, C(\hat{\theta})].$$

Draw d has the form $\theta^{(d)} = \hat{\theta} + z^{(d)}$, where $z^{(d)}$ is multivariate normal with mean 0 and covariance matrix $C(\hat{\theta})$. In large samples, this method is clearly preferable to method 1 and often preferable to method 2, since it correctly propagates asymptotic uncertainty in the ML estimate of θ.

5. **Refining approximate draws using importance sampling.** Methods 2 to 4 draw pairs $(Y_{\text{mis}}^{(d)}, \tilde{\theta}^{(d)})$ from a joint distribution where the draw of $Y_{\text{mis}}^{(d)}$ given $\tilde{\theta}^{(d)}$ is correct but the draw of $\tilde{\theta}^{(d)}$ is from an approximating density, say $g(\theta)$. A refinement is obtained by drawing a substantial set (e.g., 100–1000) of draws $Y_{\text{mis}}^{(d)}$, and then subsampling a smaller number (for example 2–10) from this set, with probability of selection of draw d proportional to $w_d \propto p(\tilde{\theta}^{(d)})\text{L}(\tilde{\theta}^{(d)}|Y_{\text{obs}})/g(\tilde{\theta}^{(d)})$. This is a version of sampling importance resampling (see Section 10.1.4). As the ratio of the initial set to the final number of draws gets large, the final draws are correct under mild support conditions.

6. **Substituting ML estimates from bootstrapped samples.** If EM is used to estimate θ and the large-sample covariance matrix is not readily available, then an approximate draw from the posterior distribution can be obtained as the estimate from applying EM to a bootstrapped sample $Y_{\text{obs}}^{(\text{boot},d)}$ of the cases (complete and incomplete), that is, a random sample with replacement of the same size as the observed sample. In symbols,

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} = \hat{\theta}(Y_{\text{obs}}^{(\text{boot},d)}).$$

This procedure is proper in that the ML estimates from the bootstrap samples are asymptotically equivalent to a sample from the posterior distribution of θ. This method may provide some robustness to model misspecification, since the bootstrap provides estimates of uncertainty asymptotically equivalent to the sandwich estimator (6.17). However, if a substantial fraction of the bootstrap samples do not yield unique ML estimates and are discarded, the

standard errors based on the remaining samples can be severely underestimated.

7. **Drawing from pragmatic conditional distributions.** With real multivariate data, it is often possible to formulate a set of conditional distributions relating each variable to a set of the other variables, which are reasonable when taken one at a time, but incoherent in the sense that they cannot be derived from a single joint distribution. Such models, even when incoherent, may be useful for creating multiple imputations as if they were coherent. For each of the variables, a draw of parameters and then missing data is made, the missing data are imputed for that variable, and the procedure cycles through the variables, replacing variables that are being conditioned in any regression by the observed or currently imputed values. A number of practical implementations of this idea include Kennickell (1991); MICE (Van Buuren and Oudshoorn, 1999); and IVEWARE (Raghunathan et al., 2001). Rubin (2002) proposes limiting the possibly incoherent draws to the creation of a monotone pattern, that is, to the creation of Yang-mp.

10.2.4. Use of Different Models for Imputation and Analysis

If the entire rationale for doing multiple imputation were for the computation of Bayesian posterior distributions in large samples, it would be an important but relatively limited tool. As the examples in Section 10.2.3 suggest, often a method can be chosen for creating multiple imputations without consideration of the precise model to be used for the analysis of the multiply-imputed data. When the model chosen to impute the data and the model chosen for analysis are identical, the theory is as described in Section 10.2.1. A theoretically interesting and practically important setting occurs when the imputation method does *not* perfectly align with the complete-data analysis conducted by the ultimate user. That is, the ultimate user of the multiply-imputed data could apply a variety of potentially complicated complete-data analyses to the multiply-imputed data, and then use the combining rules and combined results even though the multiple imputations were created under a different model.

Somewhat surprisingly, this approach can be very successful, especially with relatively limited fractions of missing information, as suggested by theoretical results and as documented by empirical examples. A simple example can be used to illustrate this phenomenon.

EXAMPLE 10.6. *Inference Under the Approximate Bayesian Bootstrap (Example 5.8 continued).* Suppose that the approximate Bayesian bootstrap (ABB) method of Example 5.8 is used to create multiple imputations within adjustment cells, but that the complete-data analysis will be based on the large sample normality of the sample mean. Assuming MAR (i.e., MCAR within adjustment cells), it is simple to show that the combining rules give valid frequentist inferences. In fact, this result holds for a variety of other multiple imputation methods: fully normal, the Bayesian bootstrap, a mean and variance-adjusted hot-deck, etc. (see Examples 4.1 to 4.4 in Rubin, 1987a).

When the imputation method uses more information than the complete-data analysis and this information is correct, the complete-data analyses will tend to be more efficient than anticipated: for instance, confidence intervals will have greater than the nominal convergence. This phenomenon was noted in Rubin and Schenker (1987) and Fay (1992, 1996), and termed "superefficiency" in Rubin (1996).

The general situation is called "uncongeniality" of the imputer's and ultimate user's models by Meng (1995). Usually, uncongeniality leads to conservative inferences, although in special circumstances it can lead to invalid (i.e., anticonservative) inferences. The following example conveys some intuition; other examples are discussed by Meng (2002) and Robins and Wang (2002).

EXAMPLE 10.7. *Effects of a Misspecified Imputation Model.* Suppose we have a sample of values of (X, Y) where X is fully observed but Y is half missing due to a MAR process. In truth, Y is a monotone but nonlinear function of X, $Y = \exp(X)$. Multiple imputations of missing Y values are created using a linear model relating Y to X. Clearly, the residual variability of Y on X will be overestimated due to lack of fit. The true residual variability is zero, and if an exponential model were fit, this would be found. The extra residual variability in the linear model has two consequences on multiple imputation. First, the between-imputation variability (e.g., of the slope of the linear model for Y on X) will be greater than if the true model were being fit, and second, for each set of imputations, the individual imputations (on and off the regression line) will be more variable than if the correct model were being used. Thus, both between- and within-variability are exaggerated relative to when the correct model is being applied to create the imputations. Because the linear fit often gives a decent approximation to the truth for global estimands, such as the grand mean or median, using an incorrect model for multiple imputation typically leads to overestimated variability, and thus, overcoverage of interval estimates. This result is seen in simulations with real data (e.g., Raghunathan and Rubin, 1998). With estimands in the tails of the distribution, such as extreme quartiles, this approximate validity may not hold.

In our experience with real and artificial data sets (e.g., Ezzati-Rice et al., 1995), the practical conclusion appears to be that multiple imputation, when carefully done, can be safely used with real problems even when the ultimate user may be applying models or analyses not contemplated by the imputer.

PROBLEMS

10.1. Reproduce the posterior distributions in Figure 10.1, and compare the posterior mean and variance with that given in Table 10.1. Recalculate the posterior distribution of θ using the improper prior distribution with $\alpha_1 = \alpha_2 = 0$. Is the resulting posterior distribution proper?

10.2. Consider a simple random sample of size n with r respondents and $m = n - r$ nonrespondents, and let $\bar{y}_R$ and s_R^2 be the sample mean and variance of the

respondents' data, and $\bar{y}_{NR}$ and s^2_{NR} the sample mean and variance of the imputed data. Show that the mean and variance $\bar{y}_*$ and s^2_* of all the data can be written as

$$\bar{y}_* = \frac{(r\bar{y}_R + m\bar{y}_{NR})}{n}$$

and

$$s^2_* = \frac{[(r-1)s^2_R + (m-1)s^2_{NR} + rm(\bar{y}_R - \bar{y}_{NR})^2/n]}{(n-1)}.$$

10.3. Suppose in Problem 10.2, imputations are randomly drawn with replacement from the r respondents' values.

(a) Show that $\bar{y}_*$ is unbiased for the population mean $\bar{Y}$.

(b) Show that conditional on the observed data, the variance of $\bar{y}_*$ is $ms^2_R(1-r^{-1})/n^2$, and that the expectation of s^2_* is $s^2_R(1-r^{-1})[1+rn^{-1}(n-1)^{-1}]$.

(c) Show that conditional on the sample sizes n and r (and the population Y values), the variance of $\bar{y}_*$ is the variance of $\bar{y}_R$ times $[1+(r-1)n^{-1}(1-r/n)(1-r/N)^{-1}]$, and show that this is greater than the expectation of $U_* = s^2_*(n^{-1}-N^{-1})$.

(d) Assume r and N/r are large, and show that interval estimates of $\bar{Y}$ based on U_* as the estimated variance of $\bar{y}_*$ are too short by a factor $(1+nr^{-1}-rn^{-1})^{1/2}$. Note that there are two reasons: $n > r$, and $\bar{y}_*$ is not as efficient as $\bar{y}_R$. Tabulate true coverages and true significance levels as functions of r/n and nominal level.

10.4. Suppose multiple imputations are created using the method of Problem 10.3 D times, and let $\bar{y}^{(d)}_*$ and $U^{(d)}_*$ be the values of $\bar{y}_*$ and U_* for the dth imputed data set. Let $\bar{\bar{y}}_* = \sum_{d=1}^{D} \bar{y}^{(d)}_*/D$, and T_* be the multiple imputation estimate of variance of $\bar{\bar{y}}_*$. That is,

$$T_* = \bar{U}_* + (1+D^{-1})B_*,$$

where

$$\bar{U}_* = \sum_{d=1}^{D} U^{(d)}_*/D, \qquad B_* = \sum_{d=1}^{D} (\bar{y}^{(d)}_* - \bar{\bar{y}}_*)^2.$$

(a) Show that, conditional on the data, the expected value of B_* equals the variance of $\bar{y}_*$.

(b) Show that the variance of $\bar{\bar{y}}_*$ (conditional on n, r, and the population Y values) is $D^{-1}\mathrm{Var}(\bar{Y}_*) + (1-D^{-1})\mathrm{Var}(\bar{y}_R)$, and conclude that $\bar{\bar{y}}_*$ is more efficient than the single-imputation estimate $\bar{y}_*$.

(c) Tabulate values of the relative efficiency of $\bar{\bar{y}}_*$ to $\bar{y}_R$ for different values of D, assuming large r and N/r.

(d) Show that the variance of $\bar{\bar{y}}_*$ (conditional on n, r, and the population Y values) is greater than the expectation of T_* by approximately $s_R^2(1 - r/n)^2/r$.

(e) Assume r and N/r are large, and tabulate true coverages and significance levels of the multiple imputation inference. Compare with the results in Problem 10.3, part (d).

10.5. Modify the multiple imputation approach of Problem 10.4 to give the correct answer for large r and N/r. (Hint: For example, add $s_R r^{-1/2} z_d$ to the imputed value for observation i, where the z_d are independent standard normal deviates.)

10.6. Consider the situation where the complete-data analysis is nonparametric, and produces no estimates but just a P value for a null hypothesis, for example, the P value for a Wilcoxon test in a randomized two-treatment experiment. Suppose that the missing data in this experiment have been multiply imputed with $D = 2$, and the two P values are p_1 and p_2. Let z_1 and z_2 be such that $\Pr(z < z_1) = p_1$, $\Pr(z < z_2) = p_2$, where z is standard normal, and $a = 3(z_1 - z_2)^2/4$. Show that a multiple-imputation combined P value can be found from treating

$$\sqrt{\frac{z_1 z_2}{1 + a}}$$

as a t random variable with $(1 + a^{-1})$ degrees of freedom. (Hint: consider Eq. (10.22) in this setting.)