# Sequential Imputations and Bayesian Missing Data Problems

Augustine KONG, Jun S. LIU, and Wing Hung WONG*

For missing data problems, Tanner and Wong have described a data augmentation procedure that approximates the actual posterior distribution of the parameter vector by a mixture of complete data posteriors. Their method of constructing the complete data sets is closely related to the Gibbs sampler. Both required iterations, and, similar to the EM algorithm, convergence can be slow. We introduce in this article an alternative procedure that involves imputing the missing data sequentially and computing appropriate importance sampling weights. In many applications this new procedure works very well without the need for iterations. Sensitivity analysis, influence analysis, and updating with new data can be performed cheaply. Bayesian prediction and model selection can also be incorporated. Examples taken from a wide range of applications are used for illustration.

KEY WORDS: Bayesian inference; Importance sampling; Missing data; Predictive distribution; Sequential imputation.

## 1. SUMMARY

A standard method to handle Bayesian missing data problems is to approximate the actual incomplete data posterior distribution of the parameter vector by a mixture of complete data posterior distributions. The multiple complete data sets used in the mixture are ideally created by draws from the posterior distribution of the missing data conditioned on the observed data. Two existing and related methods for doing this are the Gibbs sampler (Geman and Geman 1984) and the data augmentation procedure proposed by Tanner and Wong (1987). Both of these procedures are iterative and are basically equivalent when the iteration is only between the parameter vector and the missing data (Gelfand and Smith 1990). In this setting the methods are analogous to the EM algorithm for finding maximum likelihood estimates (Dempster, Laird, and Rubin 1977). Similar to the EM algorithm, the convergence of these methods can be slow. Furthermore, in situations where the posterior distribution must be constantly updated with the arrival of new data with missing parts (Spiegelhalter and Lauritzen 1990), these methods can be highly inefficient, because the whole iteration process must be redone with additional data. To resolve some of these difficulties, we propose sequential imputation as an alternative and sometimes complementary method for creating multiple complete data sets. As a variation of importance sampling, sequential imputation does not require iterations. It is related to a method proposed by Rubin (1987a, 1987b) but tends to produce more stable importance weights. Moreover, with sequential imputation sensitivity analysis and updating with new data can be done cheaply. Bayesian prediction is automatically incorporated.

* Augustine Kong is Assistant Professor and Wing Hung Wong is Professor, Department of Statistics, University of Chicago, IL 60637. Jun S. Liu is Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This research was supported by National Science Foundation (NSF) Grant DMS-8902267 and National Institutes of Health Grant R01-GM46800-01A1. Computations for this document were performed using computing facilities supported in part by NSF grants DMS-8905292, DMS-8703942, DMS-8601732, and DMS-8404941 and by the University of Chicago Block Fund. The authors thank Professor Bruce Spencer for bringing our attention to the rule of thumb (14); Professor David Wallace for giving us many valuable suggestions and providing the missing data pattern displayed in Table 3, which he encountered in a real problem; and Professor Persi Diaconis for introducing us to the application in Section 3.2. The authors also thank an associate editor and two referees for their valuable comments that helped improve this article greatly.

Section 2 presents details related to the implementation of sequential imputation. Using a number of examples, two of them with data, Section 3 illustrates a range of problems to which the method can be applied. Section 4 investigates the relative efficiency of the method through the study of the variance of the importance sampling weights and also discusses the related problem of how many imputations need to be performed. Section 5 illustrates how sensitivity analysis, with respect to the prior distribution, and the study of case influence can be performed. Section 6 gives some final remarks, discussing in particular connections between sequential imputation and Gibbs sampling.

## 2. THE METHOD

Let $\theta$ denote the parameter vector of interest and let $\mathbf{x}$ denote complete data, so that the posterior distribution $p(\theta|\mathbf{x})$ is assumed to be simple. Suppose, however, that $\mathbf{x}$ is only partially observed and can be partitioned as $(\mathbf{y}, \mathbf{z})$, where $\mathbf{y}$ denotes the observed part and $\mathbf{z}$ represents the missing part. Now suppose that $\mathbf{y}$ and $\mathbf{z}$ can each be further decomposed into $n$ corresponding components so that

$$\mathbf{x} = (x_1, \ldots, x_n) = (y_1, z_1, \ldots, y_n, z_n) \overset{\text{def}}{=} (\mathbf{y}, \mathbf{z}),$$

where $x_t = (y_t, z_t)$ for $t = 1, \ldots, n$. In many applications $x_t$ are independent and identically distributed given $\theta$, but this is not a necessary assumption. Also, the missing pattern generally will be different for different $t$'s. Indeed, an observation $t$ can be complete so that $y_t = x_t$. We note that

$$p(\theta|\mathbf{y}) = \int p(\theta|\mathbf{y}, \mathbf{z}) p(\mathbf{z}|\mathbf{y})\, d\mathbf{z}.$$

Hence if we can draw $m$ independent copies of $\mathbf{z}$'s from the conditional distribution $p(\mathbf{z}|\mathbf{y})$ and denote them by $\mathbf{z}(1)$, $\mathbf{z}(2), \ldots, \mathbf{z}(m)$, then we can approximate $p(\theta|\mathbf{y})$ by

$$\frac{1}{m} \sum_{j=1}^{m} p(\theta|\mathbf{x}(j)),$$

where $\mathbf{x}(j)$ denotes the augmented complete data set $(\mathbf{y}, \mathbf{z}(j))$ for $j = 1, \ldots, m$. [Note that each $\mathbf{z}(j)$ has $n$ compo-


Journal of the American Statistical Association
March 1994, Vol. 89, No. 425, Theory and Methods

nents: $z_1(j), \ldots, z_n(j)$.] But drawing from $p(\mathbf{z}|\mathbf{y})$ directly is usually difficult. The Gibbs sampler or the data augmentation procedure mentioned earlier do this approximately by iterations. Sequential imputation achieves something similar by imputing the $z_i$'s sequentially and using importance sampling weights to avoid iterations. Generally, sequential imputation starts by drawing $z_1^*$ from $p(z_1|y_1)$ and computing $w_1 = p(y_1)$. Then, for $t = 2, \ldots, n$, the following steps are done sequentially:

a. Draw $z_t^*$ from the conditional distribution

$$p(z_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*, y_t).$$

Notice that the $z_t^*$'s had to be drawn sequentially, because each $z_t^*$ is drawn conditioned on the previously imputed missing parts $z_1^*, \ldots, z_{t-1}^*$.

b. Compute the predictive probabilities $p(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$ and

$$w_t = w_{t-1}p(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*). \quad (1)$$

Let $w = w_n$, so that

$$w = p(y_1)\prod_{t=2}^{n} p(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*).$$

Note that steps a and b are usually done simultaneously. Both steps are required to be computationally simple, which is often the case if the predictive distributions $p(x_1)$ and $p(x_t|x_1, \ldots, x_{t-1})$, $t = 2, \ldots, n$ are simple. This is the key to the feasibility of sequential imputation. Steps a and b are done repeatedly and independently $m$ times. Let the results be denoted by $\mathbf{z}^*(1), \mathbf{z}^*(2), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$, where $\mathbf{z}^*(j) = (z_1^*(j), \ldots, z_n^*(j))$ for $j = 1, \ldots, m$. We now estimate the posterior distribution $p(\theta|\mathbf{y})$ by the weighted mixture

$$\frac{1}{W}\sum_{j=1}^{m} w(j)p(\theta|\mathbf{x}^*(j)), \quad (2)$$

where $W = \sum w(j)$ and $\mathbf{x}^*(j)$ denotes the augmented data set $(\mathbf{y}, \mathbf{z}^*(j))$ for $j = 1, \ldots, m$. To understand why (2) is the appropriate approximation, we note that each independent imputation $\mathbf{z}^*(j)$ is not drawn from the actual conditional distribution $p(\mathbf{z}|\mathbf{y})$. Instead, from (A), $\mathbf{z}^*(j)$ is drawn from the "trial density"

$$p^*(\mathbf{z}^*(j)|\mathbf{y}) = p(z_1^*(j)|y_1)\prod_{t=2}^{n} p(z_t^*(j)|y_1, z_1^*(j), y_2, z_2^*(j), \ldots, y_{t-1}, z_{t-1}^*(j), y_t). \quad (3)$$

Using standard results from importance sampling, after the draws the different imputations are weighted by

$$w^*(j) = \frac{p(\mathbf{z}^*(j)|\mathbf{y})}{p^*(\mathbf{z}^*(j)|\mathbf{y})} \quad (4)$$

$$= \frac{p(\mathbf{y}, \mathbf{z}^*(j))}{p(\mathbf{y})}\frac{p(y_1)}{p(y_1, z_1^*(j))} \times \prod_{t=2}^{n}\frac{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_t^*(j))}$$

$$= \frac{p(\mathbf{y}, \mathbf{z}^*(j))}{p(\mathbf{y})}\frac{p(y_1)}{p(y_1, \ldots, y_n, z_1^*(j), \ldots, z_n^*(j))} \times \prod_{t=2}^{n}\frac{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j))}$$

$$= \frac{p(y_1)}{p(\mathbf{y})}\prod_{t=2}^{n} p(y_t|y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j))$$

$$= \frac{w(j)}{p(\mathbf{y})}. \quad (5)$$

Because $1/p(\mathbf{y})$ is the same across imputations and the weights have to be normalized, (2) gives the correct approximation. We now provide comments useful in the implementation of the two steps.

## 2.1 Predictive Distributions

Let $\pi(\theta) = p(\theta)$ denote the prior distribution of the parameter vector $\theta$, with other notations as before. For simplicity we assume in this section that $x_t$ are iid given $\theta$. As mentioned earlier, the implementation of sequential imputation requires that the predictive distribution $p(x_t|x_1, \ldots, x_{t-1})$ is simple for all $t$ and any realization of the $x$'s. Interestingly, if $\pi(\theta)$ is conjugate to $p(x_t|\theta)$, which is necessary for the complete data posterior distribution to be simple, then the predictive distribution can usually be obtained in closed form. This is because

$$p(x_t|x_1, \ldots, x_{t-1}) = \frac{p(x_1, \ldots, x_t)}{p(x_1, \ldots, x_{t-1})}$$

and generally

$$p(x_1, \ldots, x_t) = \int_\Theta \prod_{i=1}^{t} p(x_i|\theta)\pi(\theta)\, d\theta$$

is the inverse of the normalizing constant of the complete data posterior distribution $p(\theta|x_1, \ldots, x_t)$; that is,

$$p(x_1, \ldots, x_t) = \frac{p(x_1, \ldots, x_t|\theta)\pi(\theta)}{p(\theta|x_1, \ldots, x_t)}$$

for any $\theta$. An essentially identical result was noted by Besag (1989). For example, in the standard conjugate setup for multivariate Gaussian data, the predictive distribution is multidimensional noncentral $t$. In problems where the parameters have Dirichlet prior distributions, it is often the case that

$$p(x_t|x_1, x_2, \ldots, x_{t-1}) = p(x_t|\hat{\theta}), \quad (6)$$

where $\hat{\theta} = E(\theta|x_1, x_2, \ldots, x_{t-1})$. Note that (6) means that the predictive distribution of $x_t$ given $(x_1, x_2, \ldots, x_{t-1})$ is the same as if $\theta$ is known to be $\hat{\theta}$. More details can be found in the examples presented in Section 3.

## 2.2 Weights and Prediction

Defining the importance weight recursively as in (1) frees us from having to think of a fixed sample size, which is con-

venient when the data really do arrive sequentially. Definition (1) has the natural interpretation that the importance weight is sequentially adjusted by how well a certain augmented data set predicts the next observation. Keeping track of the $w_t$'s has another potential advantage. Suppose that the data are processed sequentially for all $m$ independent augmented data sets simultaneously. At time $t$, $p(\theta \mid y_1, \ldots, y_t)$ is approximated by

$$\frac{1}{W_t} \sum_{j=1}^{m} w_t(j) p(\theta \mid \mathbf{x}_t^*(j)), \tag{7}$$

where $W_t = \sum_j w_t(j)$ and $\mathbf{x}_t^*(j) = (y_1, z_1^*(j), \ldots, y_t, z_t^*(j))$. Now some of the augmented data sets $\mathbf{x}_t^*(j)$ may have weights $w_t(j)$ that are substantially smaller than the others and are practically negligible. When the new observation $y_{t+1}$ arrives, instead of continuing to impute using the current augmented data sets $\mathbf{x}_t^*(j), j = 1, \ldots, m$, there is the following alternative. Draw $m$ copies of $\mathbf{x}_t$ independently from the collection $\mathbf{x}_t^*(j), j = 1, \ldots, m$, with probabilities proportional to $w_t(j)$. These drawn $\mathbf{x}_t$'s, denoted by $\mathbf{x}_t^{**}(j), j = 1, \ldots, m$, will now have *equal* weights. To process the $(t+1)$th observation, we draw $z_{t+1}^{**}(j), j = 1, \ldots, m$ from

$$p(z_{t+1} \mid \mathbf{x}_t^{**}(j), y_{t+1}).$$

The data sets $\mathbf{x}_{t+1}^{**}(j) = (\mathbf{x}_t^{**}(j), y_{t+1}, z_{t+1}^{**}(j)), j = 1, \ldots, m$ will then have weights $p(y_{t+1} \mid \mathbf{x}_t^{**}(j))$.

To end this section, we note that the problem of Bayesian prediction is automatically incorporated in sequential imputation. Obviously $p(x_{t+1} \mid y_1, \ldots, y_t)$ can be approximated by

$$\frac{1}{W_t} \sum_{j=1}^{m} w_t(j) p(x_{t+1} \mid \mathbf{x}_t^*(j)),$$

and $p(z_{t+1} \mid y_1, \ldots, y_{t+1})$ can be approximated by

$$\frac{1}{W_{t+1}} \sum_{j=1}^{m} w_{t+1}(j) p(z_{t+1} \mid \mathbf{x}_t^*(j), y_{t+1}).$$

### 2.3 Order of Imputation

When the data do not actually arrive sequentially, we are free to process the cases in any order. But some orders are better than others in the sense that for the same number of imputations, they tend to produce better approximations of the actual posterior distribution and hence are more efficient. The difference in efficiency can be substantial. Generally, it is desirable to have the trial distribution $p^*(\mathbf{z}^* \mid \mathbf{y})$ as close to the true distribution $p(\mathbf{z} \mid \mathbf{y})$ as possible. This usually means that the complete cases should be processed first, and the other cases should then be processed in the order of increasing missingness. This is because the missing $z_t$'s should be imputed conditioned on as much of $\mathbf{y}$ as possible. More details concerning the efficiency of sequential imputation and how it can be measured are given in Section 4.

### 2.4 Likelihood of Models

Our imputation and inference are based on a specific model. In situations where we want to compare models, it will be important to get likelihoods of different models. For a particular model $M$ the likelihood of $M$ given incomplete data $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$p_M(\mathbf{y}) = \int p_M(\mathbf{y} \mid \theta) \pi_M(\theta)\, d\theta, \tag{8}$$

where $\theta$ may be model-dependent. Suppose that we have applied sequential imputation based on model $M$. Then for all $j$ it is easy to see that

$$E_{p^*}(w^*(j)) = 1, \tag{9}$$

where $p^*$ and $w^*$ are as defined in (3) and (4). It then follows from (5) that $E_{p^*}(w(j)) = p_M(\mathbf{y})$. Thus

$$\hat{p}_M(\mathbf{y}) = \frac{1}{m} \sum_{j=1}^{m} w(j)$$

is an unbiased estimate of $p_M(\mathbf{y})$.

## 3. EXAMPLES

### 3.1 Functionals of the Multivariate Normal Covariance Matrix

This example was constructed by Murray (1977) in the discussion of Dempster, Laird, and Rubin (1977) and was again used by Tanner and Wong (1987). Table 1 contains 12 observations assumed to be drawn from a bivariate normal distribution with known means $\mu_1 = \mu_2 = 0$ and unknown covariance matrix. For notation, $\rho$ denotes the correlation coefficient and $\sigma_1$ and $\sigma_2$ denote the marginal variances. We are interested in the posterior distribution of $\rho$ given the incomplete data. Note that the information on $\sigma_1$ and $\sigma_2$ provided by the eight incomplete observations cannot be ignored in drawing likelihood-based inference about $\rho$. Following Tanner and Wong, the covariance matrix of the bivariate normal is assigned the Jeffreys's, noninformative prior distribution (Box and Tiao 1973)

$$\pi(\Sigma) \propto |\Sigma|^{-(k+1)/2} \tag{10}$$

where $k$ is the dimensionality and is equal to 2 here. The posterior distribution of $\Sigma$ given complete data is

$$p(Z \mid \text{complete data}) \propto Z^{(3/2)-1} \exp\left\{-\frac{1}{2} \operatorname{tr}[Z \cdot S]\right\},$$

where $S = (s_{ij})_{2\times 2}$ is the sample uncorrected covariance matrix and $Z = \Sigma^{-1}$.

Let complete data be $x_1, \ldots, x_{12}$, where $x_t = (u_t, v_t)$ for $t = 1, \ldots, 12$. Thus $v_t$ are missing for observations 5–8, whereas $u_t$ are missing for observations 9–12. The predictive distribution of $x_{t+1}$ given complete data $x_1, \ldots, x_t$ is

$$x_{t+1} \mid x_1, \ldots, x_t \sim t_2\left(0, \frac{S_t}{t-1}, t-1\right),$$

where $S_t = (s_{ij})$, $s_{ij} = \sum_{s=1}^{t} x_s(i) x_s(j)$, and $t_2$ is the bivariate

*Table 1. Twelve Bivariate Normal Observations*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | −1 | −1 | 2 | 2 | −2 | −2 | * | * | * | * |
| 1 | −1 | 1 | −1 | * | * | * | * | 2 | 2 | −2 | −2 |

* Indicates that the value is missing.

$t$ distribution. It follows that, conditioning on $x_1, \ldots, x_t$, marginal and conditional distributions are

$$u_{t+1}|\mathbf{x}_t \sim t_1\left(0, \frac{s_{11}}{t-1}, t-1\right)$$

and

$$v_{t+1}|\mathbf{x}_t, u_{t+1} \sim t_1\left[\frac{s_{12}}{s_{11}}u_{t+1}, \frac{|S_t|}{t \cdot s_{11}}\left(1+\frac{u_{t+1}^2}{s_{11}}\right), t\right].$$

Similar results can be obtained for $v_{t+1}|\mathbf{x}_t$ and $u_{t+1}|\mathbf{x}_t v_{t+1}$. Based on these distributional results, steps a and b both can be easily implemented.

The complete-data posterior distribution of $\rho$ is still hard to compute, because it has the form

$$p(\rho|\mathbf{x}) \propto (1-\rho^2)^{(\nu-2)/2}\int_0^\infty \omega^{-1}\left(\omega+\frac{1}{\omega}-2\rho r\right)^{-(\nu+1)} d\omega,$$

where $r$ is the sample correlation coefficient, $\nu = n-2 = 10$. To avoid numerical integration of this distribution, the algorithm proposed by Odell and Feiveson (1966) can be used to generate observations from inverse Wishart distribution. The observed-data posterior distribution of $\rho$ can be obtained analytically up to a renormalizing constant:

$$p(\rho|\text{data in Table 1}) \propto \frac{[(1-\rho^2)^{4.5}]}{[(1.25-\rho^2)^8]}.$$

Figure 1 gives the exact posterior distribution of $\rho$ and the approximated posterior distribution based on sequential imputation. The approximation is a weighted mixture of $m = 1{,}000$ complete data posterior distributions. Our approximation is comparable to that of Tanner and Wong (1987), which is based on 6,400 imputations and 15 iterations. Furthermore, we do not have to draw from an inverse Wishart distribution, which is an unavoidable step for Tanner and Wong. The sequential imputation part of this example took 10 seconds on a Sparc Station II.

### 3.2 Nonparametric Bayesian Analysis of Binomial Data

Here we consider the binomial model

$$y_t \sim \text{Binomial}(l_t, \zeta_t) \qquad 1 \le t \le n,$$

where the $\zeta_t$'s are assumed to be random and independent and to have a common distribution $F$. Our interests are in drawing inference about both $F$ and the $\zeta_t$'s based on the observed data $y_t$, $t = 1, \ldots, n$. For example, imagine randomly drawing $n$ baseball players from a population of professional players. Then $\zeta_t$ can be interpreted as the inherent unobserved long-run batting average of player $t$. The observed data for player $t$ is $y_t$, the number of hits out of a total of $l_t$ official at bats. Note that the number of at bats need not be the same for all the players. Efron and Morris (1975) used normal theory-based empirical Bayes method to analyze data of the first 45 at bats in a season for $n = 18$ major league players. Taking a nonparametric Bayesian approach, we assign $F$, which is an infinite dimensional parameter, a Dirichlet prior distribution $\mathcal{D}(\alpha)$ where $\alpha$ is some measure on the interval $[0, 1]$. Note that $\mathcal{D}(\alpha)$ is a probability measure on $\mathcal{P}$, where $\mathcal{P}$ is the set of all probabilities on the Borel sets of $[0, 1]$. Readers not familiar with this special class of Dirichlet prior distributions and nonparametric Bayesian inference are referred to the work of Ferguson (1974). In this problem $\zeta_t$ plays the role of the missing data $z_t$ and $F$ corresponds to $\theta$. What is special is that $F$ has infinite dimension. Also, the posterior distribution of the missing data $\zeta = (\zeta_1, \ldots, \zeta_n)$ may itself be of interest. In the case of batting averages, we may want to estimate $\zeta_t$ for a certain player $t$ given his performance in the first 45 at bats to predict his performance in the remainder of the season. Given $\zeta$, the posterior distribution of $F$ is simply $\mathcal{D}(\alpha')$ with $\alpha' = \alpha + \sum_{t=1}^n \delta_{\zeta_t}$, where $\delta_{\zeta_t}$ is a delta measure with mass 1 located at $\zeta_t$. For simplicity, suppose that $\alpha$ is uniform on $[0, 1]$ with

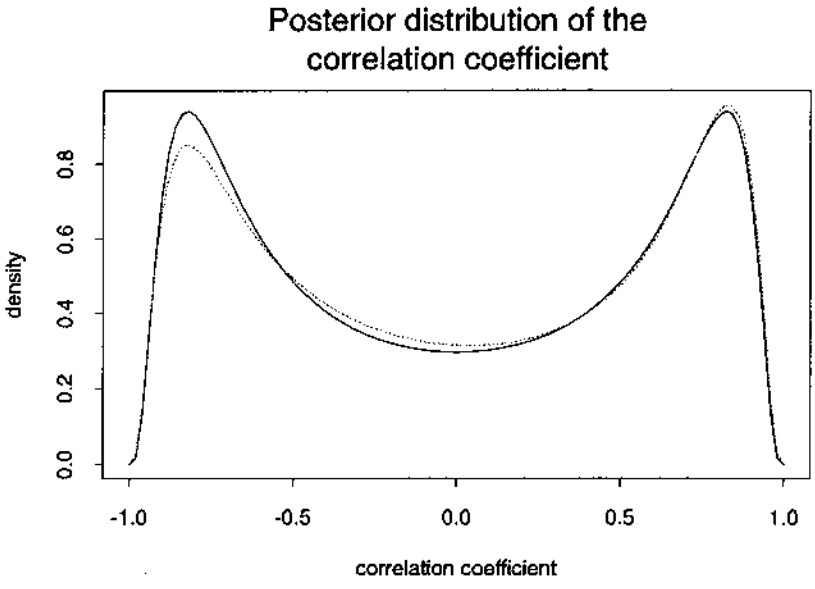


*Figure 1. An Approximation to the Posterior Distribution of the Sample Correlation Coefficient $\rho$ for Murray's Data. Solid line: true density; dashed line: approximation.*

Table 2. Batting Averages and Their Estimates

| t | Batting average for first 45 at bats (1) | Batting average for remainder of season (2) | Stein's estimator (3) | Efron–Morris's estimator (4) | Dirichlet prior with $\alpha = 2 \times \text{Beta}(2, 6)$: Posterior mean (5) | Dirichlet prior with $\alpha = 2 \times \text{Beta}(2, 6)$: Posterior standard deviation (6) |
|---|---|---|---|---|---|---|
| 1 | .400 | .346 | .290 | .334 | .315 | .063 |
| 2 | .378 | .298 | .286 | .313 | .304 | .056 |
| 3 | .356 | .276 | .281 | .292 | .294 | .050 |
| 4 | .333 | .222 | .277 | .277 | .286 | .045 |
| 5 | .311 | .273 | .273 | .273 | .279 | .041 |
| 6 | .311 | .270 | .273 | .273 | .279 | .041 |
| 7 | .289 | .263 | .268 | .268 | .272 | .039 |
| 8 | .267 | .210 | .264 | .264 | .267 | .037 |
| 9 | .244 | .269 | .259 | .259 | .261 | .036 |
| 10 | .244 | .230 | .259 | .259 | .261 | .036 |
| 11 | .222 | .264 | .254 | .254 | .255 | .037 |
| 12 | .222 | .256 | .254 | .254 | .255 | .037 |
| 13 | .222 | .303 | .254 | .254 | .255 | .037 |
| 14 | .222 | .264 | .254 | .254 | .255 | .037 |
| 15 | .222 | .226 | .254 | .254 | .255 | .037 |
| 16 | .200 | .285 | .249 | .249 | .249 | .039 |
| 17 | .178 | .316 | .244 | .233 | .241 | .042 |
| 18 | .156 | .200 | .239 | .208 | .231 | .048 |

norm $\|\alpha\| = 1$. Then, conditioned on $\zeta_1, \ldots, \zeta_{t-1}$, $F$ is distributed as $\mathcal{D}(\alpha_{t-1})$, with $\alpha_{t-1} = \alpha + \sum_{i=1}^{t-1} \delta_{\zeta_i}$. This implies that

$$\zeta_t | \zeta_1, \ldots, \zeta_{t-1} \sim \frac{1}{t}\left(\alpha + \sum_{i=1}^{t-1} \delta_{\zeta_i}\right), \tag{11}$$

which should be interpreted as a probabilistic mixture of $\alpha$ and delta measures concentrated at the $\zeta_i$'s. It follows that

$$\zeta_t | \zeta_1, \ldots, \zeta_{t-1}, y_t \sim c[B(y_t + 1, l_t - y_t + 1)\text{Beta}(y_t + 1, l_t - y_t + 1) + \sum_{i=1}^{t-1} \zeta_i^{y_t}(1 - \zeta_i)^{l_t - y_t}\delta_{\zeta_i}], \tag{12}$$

where

$$B(y_t + 1, l_t - y_t + 1) = \int_0^1 \zeta^{y_t}(1 - \zeta)^{l_t - y_t}\, d\zeta$$

is the Beta function, $\text{Beta}(\cdot, \cdot)$ is the standard Beta distribution, and

$$c = \frac{1}{B(y_t + 1, l_t - y_t + 1) + \sum_{i=1}^{t-1} \zeta_i^{y_t}(1 - \zeta_i)^{l_t - y_t}}$$

is the normalizing constant. Note that (12) is a mixture of a Beta distribution and discrete point masses. From (11) we also get

$$p(y_t | \zeta_1, \ldots, \zeta_{t-1}) = \frac{1}{t} B(y_t + 1, l_t - y_t + 1) + \frac{1}{t}\sum_{i=1}^{t-1} \zeta_i^{y_t}(1 - \zeta_i)^{l_t - y_t},$$

which is the term needed for updating the importance sampling weights. Hence both steps a and b of sequential imputation can be easily implemented. Note that a direct application of Gibbs sampling is not feasible here because of the difficulties in drawing samples of $F$, which is infinite dimensional. Escobar (1991) described a way of implementing Gibbs sampling without drawing the $F$'s directly, which also takes advantage of the simplicity of the predictive distributions (11) and (12). For a related problem where the $\zeta_i$'s are ordered, Gelfand and Kuo (1991) also used a similar idea to avoid sampling the infinite-dimensional $F$.

Sequential imputation is applied to the same set of data analyzed by Efron and Morris (1975). The batting averages based on the first 45 at bats, $y_t/45$, of 18 major league players in the 1970 season are displayed in Table 2. Also displayed are the batting averages for the remainder season. We performed $m = 1{,}000$ imputations. Because $F$ is infinite dimensional, there is no easy way to display its approximated posterior distribution. Figure 2 shows part of our results, which is the approximate of the mean curve of the posterior distribution of $F$. More precisely, the posterior distribution of $F$ is approximated by $(1/W)\sum_{j=1}^{m} w(j)\mathcal{D}(\alpha(j))$, where $\alpha(j) = \alpha + \sum_{i=1}^{n} \delta_{\zeta_i(j)}$. The curves in Figure 2 are $1/[(n + \|\alpha\|)W]\sum_{j=1}^{m} w(j)\alpha(j)$, which are approximations to $E(F|\mathbf{y})$. Note that $E(F|\mathbf{y})$ is also the predictive distribution $p(\zeta_{n+1}|\mathbf{y})$ of the long-term batting average for the next randomly chosen player. Two different priors were used. The first has $\alpha$ as a uniform measure on [0, 1] with $\|\alpha\|$

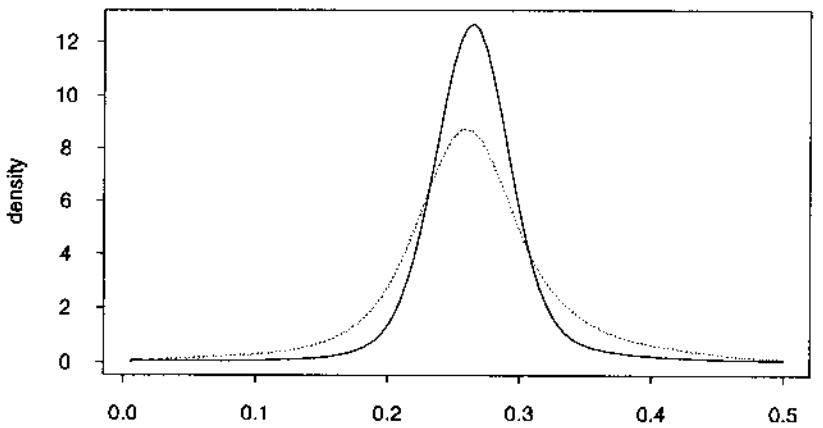


Figure 2. Approximation of $E(F|\mathbf{y})$, the Predictive Distribution for the Batting Average of Another Randomly Chosen Player. Solid line: $\alpha$ uniform; dotted line: $\alpha = 2 \times \text{Beta}(2, 6)$.

= 1. The second $\alpha$, which better reflects our prior knowledge, has a density proportional to a Beta(2, 6) distribution, but $\|\alpha\| = 2$. (This will be referred to as $2 \times$ Beta(2, 6).) A complete analysis studying the sensitivity of the results to both the shape and the norm of $\alpha$ was provided by Liu (1993).

For a given $t$ the posterior distribution $p(\zeta_t|\mathbf{y})$ can be estimated by the weighted mixture

$$\frac{1}{W}\sum_{j=1}^{m} w(j)p(\zeta_t|\mathbf{y}, \zeta^*_{[-q]}(j)),$$

where $\zeta^*_{[-q]}(j) = (\zeta^*_1(j), \ldots, \zeta^*_{t-1}(j), \zeta^*_{t+1}(j), \ldots, \zeta^*_n(j))$. From this mixture approximates of both the posterior mean $E(\zeta_t|\mathbf{y})$ and the posterior standard deviation $\sqrt{\text{var}(\zeta_t|\mathbf{y})}$ can be obtained. These estimates for all 18 players, computed using the results from having $\alpha = 2 \times$ Beta(2, 6), are displayed in Table 2. For comparison the results from using Stein's estimator and one of the estimators constructed by Efron and Morris (1975) are also displayed. These are point estimates of the $\zeta_t$'s. Relative to the maximum likelihood estimates, the observed batting averages, Stein's estimator has the well-known effect of shrinking the estimates toward the center of the group. Efron and Morris put a threshold on the amount of shrinkage and hence deviated from Stein's estimator for the extreme observations. Our posterior means are somewhat between the two. Note that, as is reasonable, the posterior standard deviations are bigger for the more extreme $y_t$'s. Also note that for all 18 players, the batting averages for the remainder of the season, which can be treated as surrogates to the actual $\zeta_t$'s, fall within two posterior standard deviations of the posterior means.

The computations performed for this example took about 15 seconds on a Sparc station SLC.

### 3.3 Graphical Models and Genetics

We end this section by mentioning two large classes of problems where sequential imputation can be and has been applied.

Graphical models, directed and undirected, were introduced by Kiiveri, Speed, and Carlin (1984) and Darroch, Lauritzen, and Speed (1980). Directed and undirected graphs are used to illustrate causal relations and conditional independence relations among variables. Undirected graphical models with discrete data form a special class of log-linear models. For undirected decomposable graphical models with complete data, Dawid and Lauritzen (1993) demonstrated how conjugate prior distributions can be set up so that both the posterior distributions of parameters and the predictive distribution of the next observation are simple. In particular, they showed that (6) is satisfied for models with discrete data and Dirichlet prior distributions. Similar theoretical results for directed graphical models were presented by Spiegelhalter and Lauritzen (1990), who also discussed the problem of missing data and the possibility that the observations may actually arrive sequentially under a medical diagnostic setting. Using the results derived by the two aforementioned references, sequential imputation can be easily implemented where there are missing data. Because these models have some other special characteristics that require elaborate illustration, a separate report on them with real examples is under preparation.

The examples presented so far concentrate on situations where $x_i$, $i = 1, \ldots, n$ are iid given the unknown parameter $\theta$. In an application in genetics linkage involving the analysis of pedigree data on multiple loci, the missing data $z_i$'s are high-dimensional binary vectors and form an inhomogeneous Markov chain given $\theta$. In one special case $\theta$ is univariate and denotes the location of a decrease-causing gene relative to a collection of genetic markers. For this problem the exact evaluation of a single likelihood value is prohibitive, because it involves summing over a very high-dimensional space. Kong, Irwin, Cox, and Frigge (1992) applied sequential imputation conditioned on a fixed parameter value of interest denoted by $\theta^*$. Similar to the results presented in Section 2.4, it can be shown that $E_{p^*}(w(j)) = p(\mathbf{y}|\theta^*)$, which implies that the average of the $w(j)$'s is an unbiased estimate of the likelihood $L(\theta^*) = p(\mathbf{y}|\theta^*)$. In this way accurate estimates of likelihood values can be obtained even with data involving more than four highly polymorphic genetic markers, something that could not be done before.

## 4. EFFICIENCY OF SEQUENTIAL IMPUTATION AND THE CHOICE OF $M$

### 4.1 Effective Sample Size

Because sequential imputation is a form of importance sampling, one way to measure its efficiency as a method to perform imputations is to compare it with direct sampling from $p(\mathbf{z}|\mathbf{y})$. To be specific, let $h(\theta)$ be some function of $\theta$ and suppose that the posterior mean $E(h(\theta)|\mathbf{y})$, denoted by $\mu_h$, is of interest. If sequential imputation is applied, then

$$\tilde{\mu}_h = \frac{1}{W}\sum_{j=1}^{m} w(j)E(h(\theta)|\mathbf{x}^*(j))$$

is a natural estimate of $\mu_h$. For comparison suppose $\mathbf{z}(j)$, $j = 1, \ldots, m$ are independent draws from $p(\mathbf{z}|\mathbf{y})$; then

$$\hat{\mu}_h = \frac{1}{m}\sum_{j=1}^{m} E(h(\theta)|\mathbf{x}(j)),$$

where $\mathbf{x}(j) = (\mathbf{y}, \mathbf{z}(j))$, is the natural unbiased estimate of $\mu_h$. Hence the ratio $\text{var}_{p^*}(\tilde{\mu}_h)/\text{var}_p(\hat{\mu}_h)$, where $p^*$ is as defined in (3) and $p$ denotes $p(\cdot|\mathbf{y})$, measures the relative efficiency of sequential imputation. Although this ratio generally depends on $h$, by applying the delta method and using only the first two moments of $w$ and $h$ we get the approximation

$$\frac{\text{var}_{p^*}(\tilde{\mu}_h)}{\text{var}_p(\hat{\mu}_h)} \approx 1 + \text{var}_{p^*}(w^*(j)), \tag{13}$$

where $w^*(j)$ is as defined in (4) (see Kong 1992 for details). Although there is no guarantee that the remainder term in this approximation is ignorable, what is nice about (13) is that it does not involve $h$. This makes it particularly useful as a measure of relative efficiency when many different $h$'s are of potential interest. Indeed, following a rule of thumb in sampling, we define

$$\text{ESS} = \frac{m}{1 + \text{var}_{p^*}(w^*(j))} \tag{14}$$

as the effective sample size of sequential imputation. In general, $\text{var}_{p^*}(w^*(j))$ is impossible to obtain. But because of (9) and the fact that $w(j)$ is proportional to $w^*(j)$, $\text{var}_{p^*}(w^*(j))$ is equal to the square of the coefficient of variation of $w(j)$. Define the standardized weights as

$$w^{st}(j) = \frac{w(j) \cdot m}{W}.$$

The standardized weights have average equal to 1. Hence their sample variance

$$s^2(w^{st}(j)) = \left(\frac{m}{W}\right)^2 s^2(w(j)) \tag{15}$$

can be used to approximate $\text{var}_{p^*}(w^*(j))$. Consider the examples in Section 3. In the bivariate normal example the variance of the standardized weights is .08, which translates into an effective sample size of $m/1.08 = .93m$, a very high efficiency. In the baseball example the variances of the standardized weights for the two priors are 2.95 and 3.45, which correspond to effective sample sizes of $m/3.95 = .25m$ and $m/4.45 = .22m$.

Here we discuss the practical problem of how to choose $m$ assuming that the posterior distribution of $h(\theta)$ for some function $h$ is of interest. One key aspect of the posterior distribution is the posterior mean $\mu_h = E(h(\theta)|\mathbf{y})$. The variance $\text{var}_{p^*}(\tilde{\mu}_h)$, is a measure of the noise created by Monte Carlo approximation. In most applications the Monte Carlo noise will have a negligible effect on the overall inference about $h(\theta)$ if $\text{var}_{p^*}(\tilde{\mu}_h)$ is small relative to the posterior variance $\text{var}(h(\theta)|\mathbf{y})$. More specifically, the ratio

$$\begin{aligned}\frac{\text{var}_{p^*}((\tilde{\mu}_h)}{\text{var}(h(\theta)|\mathbf{y})} &= \frac{\text{var}_p(\hat{\mu}_h)}{\text{var}(h(\theta)|\mathbf{y})}\,\frac{\text{var}_{p^*}(\tilde{\mu}_h)}{\text{var}_p(\hat{\mu}_h)} \\ &= \frac{\text{var}(E(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y})}{E(\text{var}(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y}) + \text{var}(E(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y})}\,\frac{1}{m}\,\frac{\text{var}_{p^*}(\tilde{\mu}_h)}{\text{var}_p(\hat{\mu}_h)}\end{aligned} \tag{16}$$

should not be greater than .05, preferably smaller than .01. [Note that $E$ and var are taken with respect to $p(\cdot)$, and $\hat{\mu}_h$, as before, is the estimate of $\mu_h$ assuming that one can sample directly from $p(\mathbf{z}|\mathbf{y})$.] The term

$$\frac{\text{var}(E(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y})}{E(\text{var}(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y}) + \text{var}(E(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y})}$$

is what Rubin (1987b) referred to as the fraction of missing information with respect to $h$. From (13) and (14) the term $(1/m)(\text{var}_{p^*}(\tilde{\mu}_h)/\text{var}_p(\hat{\mu}_h))$ is approximately equal to $1/\text{ESS}$. Thus (16) can be interpreted as the fraction of missing information divided by the effective sample size. Note that the fraction of missing information is specific to the data, and the effective sample size is specific to the method used to perform the imputations. Because the fraction of missing information is bounded above by 1 for all $h$, it implies that (16) will be smaller than .01 if the effective sample size is larger than 100. If needed, the fraction of missing information can be approximated from the samples. The conditional expectation $E(\text{var}(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y})$ can be approximated by $(1/W)\sum_{j=1}^{m} w(j)\text{var}(h(\theta)|\mathbf{x}^*(j))$. Also,

$$\begin{aligned}&\text{var}(E(h(\theta)|\mathbf{y}, \mathbf{z})|\mathbf{y}) \\ &\quad = E[(E(h(\theta)|\mathbf{y}, \mathbf{z}) - \mu_h)^2|\mathbf{y}] \\ &\quad = E_{p^*}[w^* \times (E(h(\theta)|\mathbf{y}, \mathbf{z}^*) - \mu_h)^2|\mathbf{y}],\end{aligned}$$

which can be approximated by $(1/W)\sum_{j=1}^{m} w(j) \times (E(h(\theta)|\mathbf{x}^*(j)) - \tilde{\mu}_h)^2$. For the bivariate normal example presented in Section 3.1, taking $h(\Sigma)$ to be $\sigma_1^2$, the fraction of missing information is estimated to be approximately .42.

### 4.2 A Simulation Study

As demonstrated, the key to the efficiency of sequential imputation is the variance of the importance sampling weights. Considering that both the bivariate normal example and the baseball example have rather small sample sizes $n$, to better understand the behavior of the weights we simulated 269 vectors from a six-dimensional multivariate distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} 2 & 0 & \frac{1}{2} & \frac{3}{2} & \frac{3}{4} & -1 \\ 0 & 8 & -\frac{1}{3} & 2 & 0 & -4 \\ \frac{1}{2} & -\frac{1}{3} & \frac{5}{9} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{3}{2} & 2 & \frac{1}{3} & 4 & 1 & -\frac{1}{2} \\ \frac{3}{4} & 0 & 0 & 1 & \frac{3}{4} & -\frac{1}{2} \\ -1 & -4 & \frac{1}{3} & -\frac{1}{2} & -\frac{1}{2} & 6 \end{pmatrix}.$$

It is also assumed that the data have the missing pattern displayed in Table 3. For example, 88 vectors are completely observed, 40 vectors are missing only the sixth variable, and 26 vectors are missing both the first and the second variable. (This missing data pattern came from a real application in social science, but the multivariate normal assumption is not quite appropriate for the actual data set.) Both the mean and the covariance matrix are assumed to be unknown and are assigned the Jeffreys's noninformative prior distribution

$$\pi(\mu, \Sigma) \propto |\Sigma|^{-(k+1)/2},$$

where $k = 6$. Similar to the bivariate normal example in Section 3.1, the predictive distributions $p(x_{t+1}|x_1, \ldots, x_t)$ are all multivariate noncentral $t$ (see Box and Tiao 1973). Because of that, both steps a and b can be easily implemented.

Our focus here is the variance of the standardized weights instead of the actual posterior distribution of the parameters. We started by applying sequential imputation to the full data set, which has 88 complete observations and 181 incomplete observations. We chose $m$ to be 1,000 and processed the data in order from left to right in Table 3. In other words the 88 complete cases were processed first, which did not require

*Table 3. Missing Pattern for the Six-Dimensional Multivariate Normal Data.*

| Dim/No. | 88 | 40 | 22 | 22 | 4 | 3 | 23 | 26 | 7 | 3 | 3 | 2 | 10 | 5 | 2 | 2 | 1 | 2 | 2 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | ? | | ? | | | ? | | ? | | | ? | ? | ? | ? | ? | ? |
| 2 | | | | | | | | ? | | | | | ? | | | ? | ? | ? | ? | ? | ? |
| 3 | | | | | ? | | | | | | | ? | | | ? | | | | | ? | ? |
| 4 | | | | ? | | | | | ? | ? | | | | ? | | ? | | | ? | | |
| 5 | | | ? | | | | ? | | | ? | | | | ? | ? | | ? | ? | ? | | ? |
| 6 | | ? | | | | | ? | | ? | | ? | ? | ? | ? | ? | | | ? | | ? | |

NOTE: A question mark represents missing data.

any imputations, followed by the 40 observations missing the sixth variable, and so on. The overall computing time was about 21 minutes on a Sparc Station II. The sample variance of the standardized weights came out to be about .12. This corresponds to an effective sample size of about $m/1.12 = .83m$. To investigate further, we redid the imputations, this time using only 20 of the original 88 complete observations but keeping all the incomplete cases. This reduced $n$ from 269 to 201. With this change, the sample variance of the standardized weights increased to about 4.0, implying that the effective sample size decreases to $m/5 = .2m$, which is still perfectly acceptable. We then further reduced the number of complete observations to 10. Here the variance of the standardized weights becomes as large as 140, corresponding to less than 1% efficiency.

These results are not surprising. The main concern with sequential imputation is how the early imputations ($z_t$ for small $t$) are done. Because the early imputations are done only conditional on the early part of the observations, the trial distribution they are drawn from can be very far from the actual conditional distribution $p(\cdot \mid \mathbf{y})$. This can lead to highly varied importance weights. When we have 88 complete cases to start with, the first time we have to actually impute is with $z_{89}$. The trial density we draw from, $p(z_{89} \mid x_1, \ldots, x_{88}, y_{89})$, is likely to be not too far from the actual conditional distribution $p(z_{89} \mid x_1, \ldots, x_{88}, y_{89}, \ldots, y_{269})$. The situation improves further for the latter imputations. Hence the variance of the importance weights tends to be small. In contrast, when the number of complete cases decreases to near 0, the situation deteriorates because the early imputations are basically drawn from the flat prior distribution, which can be very far from the actual conditional distribution, especially in the case of normal data. This is what we see when the number of complete cases is reduced to 10, which is quite extreme for this problem because it takes 6 observations to ensure the sample covariance matrix is nonsingular, a necessary condition for the predictive distribution to be proper.

The preceding discussion does not imply that having a certain percentage of complete observations is necessary for sequential imputation to work well. For example, in the problem with the baseball data there are no complete observations, but the weights are still well behaved. The reason is partly that the observed data are discrete and the missing data are bounded and partly that the latter observations do not provide a huge amount of information about the early missing data. Indeed the worst situation is when there are a significant portion of complete cases, but they are processed after the incomplete observations. As mentioned earlier, if the data do not actually arrive sequentially, then the data should be processed in the order of increasing missingness.

Another interesting aspect is how the variance of the weights behave as a function of $t$. Define

$$w_t^*(j) = \frac{p(\mathbf{z}_t^*(j) \mid \mathbf{y}_t)}{p^*(\mathbf{z}_t^*(j) \mid \mathbf{y}_t)},$$

where $\mathbf{z}_t^*(j) = (z_1(j), \ldots, z_t(j))$ and $\mathbf{y}_t = (y_1, \ldots, y_t)$. So $w^*(j)$, as defined in (4), is equal to $w_n^*(j)$. It can be easily checked that $w_t^*(j)$ satisfies the recursive relationship

$$w_{t+1}^*(j) = w_t^*(j) \frac{p(\mathbf{z}_t^*(j) \mid \mathbf{y}_t, y_{t+1})}{p(\mathbf{z}_t^*(j) \mid \mathbf{y}_t)} \tag{17}$$

and, as a generalization of (9),

$$E_{p^*}(w_t^*(j)) = 1. \tag{18}$$

Because of (18), the sample variance of the standardized $w_t(j)$'s, as defined in (1), can be used to approximate $\mathrm{var}_{p^*}(w_t^*(j))$. For the multivariate normal data set generated based on Table 3, the sample variance of the standardized $w_t(j)$'s is plotted as a function of $t$ in Figure 3. (Because the first 88 observations do not require any imputations, the Time in the Figure starts with the 89th observation.) The results of two separate runs, each with $m = 1{,}000$, are presented for comparison. One striking feature is that the two plots are very similar. In both simulations the sample variance of the standardized weights are 0 up to Time = 40. This is not a coincidence. In the case of multivariate normal data it can be shown that if the missing data pattern is monotone (Little and Rubin 1987), which means that one can find an ordering of the data so that the missing covariates are nested in the order of increasing missingness, and the data are processed in such order, then $p^*(\mathbf{z}^* \mid \mathbf{y})$ is actually the same as $p(\mathbf{z}^* \mid \mathbf{y})$, and so the importance weights have zero variance! In our example, for up to the first 40 incomplete cases the missing data pattern is monotone. Another feature we see in the plots is that the sample variance of the standardized weights tend to increase in time but not strictly so. Both phenomena can be explained by the following theorem.

*Theorem.* The importance sampling weight $w_t^*(j)$ resulting from the method of sequential imputation is a martingale sequence in $t$ with both $\mathbf{z}^*(j)$ and $\mathbf{y}$ treated as **random**. This implies that its variance is an increasing function of $t$.

*Proof.* For simplicity we suppress the argument $(j)$ here and let $\mathcal{F}_t = \sigma\{z_1^*, \ldots, z_t^*, y_1, \ldots, y_t\}$ be the $\sigma$-field gen-

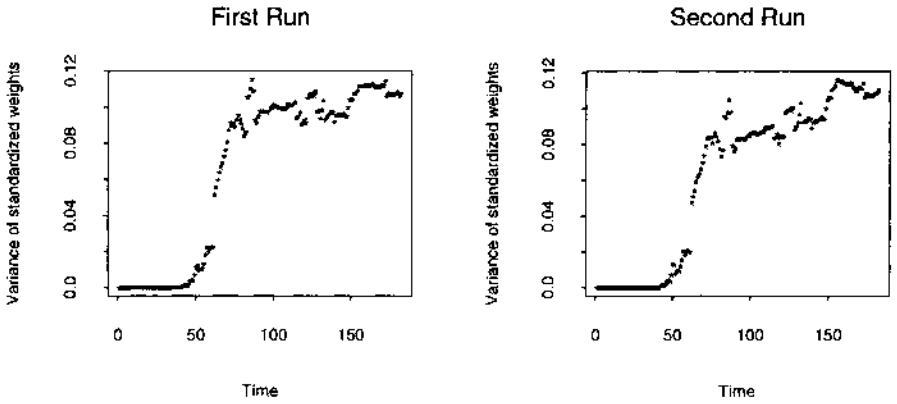


*Figure 3: The Increasing Trend of the Variance of the Standardized Weights.*

erated by all the observed and imputed random variables up to time $t$. Because $z_1^*, \ldots, z_t^*$ are imputed from $y_1, \ldots, y_t$, it is valid to think of $y_{t+1}$ as conditionally independent of $\mathbf{z}_t^*$ given $y_1, \ldots, y_t$. Hence from (17) we have

$$\begin{aligned} E(w_{t+1}^* \mid \mathcal{F}_t) &= w_t^* \int \frac{p(\mathbf{z}_t^* \mid y_t, y_{t+1})}{p(\mathbf{z}_t^* \mid \mathbf{y}_t)} p(y_{t+1} \mid \mathbf{y}_t)\, dy_{t+1} \\ &= w_t^* \int \frac{p(\mathbf{z}_t^*, y_{t+1} \mid \mathbf{y}_t)}{p(\mathbf{z}_t^* \mid \mathbf{y}_t)}\, dy_{t+1} \\ &= w_t^* \int p(y_{t+1} \mid \mathbf{z}_t^*, \mathbf{y}_t)\, dy_{t+1} = w_t^*, \end{aligned}$$

which shows that $w_t^*$ is a martingale in $t$. Therefore,

$$\operatorname{var}(w_t^*) = \operatorname{var}(E(w_{t+1}^* \mid \mathcal{F}_t)) \le \operatorname{var}(w_{t+1}^*),$$

which completes the proof.

The theorem shows that the variance of $w_t^*$, unconditional on $\mathbf{y}$, is an increasing function of $t$. But the variance in expression (14) and its estimate (15), although not explicitly indicated there, is conditional on $\mathbf{y}$, or $\mathbf{y}_t$ here. From the variance decomposition we have

$$\operatorname{var}(w_t^*) = E(\operatorname{var}(w_t^* \mid \mathbf{y}_t)) + \operatorname{var}(E(w_t^* \mid \mathbf{y}_t)).$$

But $E(w_t^* \mid \mathbf{y}_t) = 1$ for all $\mathbf{y}_t$. So $E(\operatorname{var}(w_t^* \mid \mathbf{y}_t)) = \operatorname{var}(w_t^*)$, which explains why the estimates of $\operatorname{var}(w_t^* \mid \mathbf{y}_t)$ plotted in Figure 3 have an increasing trend but are not strictly increasing.

## 5. SENSITIVITY AND INFLUENCE ANALYSES

In Bayesian inference we often will be interested in how sensitive the posterior distribution is to the prior distribution. In the missing data setting this distribution can be highly inefficient if we have to create separate augmented data sets, either by sequential imputation or other techniques, for each prior distribution we would like to consider. Next we demonstrate that the multiple augmented data sets that we constructed based on one prior distribution can actually be used for approximating the posterior distribution of $\theta$ under a different prior distribution.

Assume that the prior distribution used for the imputations is $\pi(\theta) = p(\theta)$ and that we are interested in the posterior distribution of $\theta$ if the prior distribution is $f(\theta)$ instead. In general, let $f$ denote the distribution of variables under the prior distribution $f(\theta)$. For example,

$$f(\mathbf{y}, \mathbf{z}) = \int_\Theta p(\mathbf{y}, \mathbf{z} \mid \theta) f(\theta)\, d\theta.$$

Again, based on standard importance sampling theory, the correct approach is to weight the augmented complete data sets by

$$\begin{aligned} \frac{f(\mathbf{z}^*(j) \mid \mathbf{y})}{p^*(\mathbf{z}^*(j) \mid \mathbf{y})} &= \frac{f(\mathbf{z}^*(j) \mid \mathbf{y})}{p(\mathbf{z}^*(j) \mid \mathbf{y})} \frac{p(\mathbf{z}^*(j) \mid \mathbf{y})}{p^*(\mathbf{z}^*(j) \mid \mathbf{y})} \\ &= \frac{p(\mathbf{y})}{f(\mathbf{y})} \frac{f(\mathbf{z}^*(j), \mathbf{y})}{p(\mathbf{z}^*(j), \mathbf{y})} \frac{w(j)}{p(\mathbf{y})} = \frac{f(\mathbf{x}^*(j))}{p(\mathbf{x}^*(j))} \frac{w(j)}{f(\mathbf{y})}, \end{aligned}$$

where $w(j)$, the original weight computed, is as defined in (1). Now, because $f(\mathbf{y})$ does not depend on $j$, we can simply weight the augmented data sets by

$$w^\diamond(j) = \frac{f(\mathbf{x}^*(j))}{p(\mathbf{x}^*(j))} w(j). \tag{19}$$

Computing the new weight $w^\diamond(j)$ requires evaluation of $f(\mathbf{x}^*(j))$ and $p(\mathbf{x}^*(j))$, which is easy if both $f(\theta)$ and $\pi(\theta)$ are conjugate prior distributions or mixtures of conjugate prior distributions. The posterior distribution $f(\theta \mid \mathbf{y})$ can then be approximated by

$$\frac{1}{W^\diamond} \sum_{j=1}^{m} w^\diamond(j) f(\theta \mid \mathbf{x}^*(j)), \tag{20}$$

where $W^\diamond = \sum_{j=1}^m w^\diamond(j)$. For example, consider the case of multivariate normal data studied earlier. Suppose that the missing data are imputed sequentially based on Jeffreys's prior distribution as given in (10) and that we are interested in the posterior distribution of the parameters under a new conjugate prior distribution of the form $f(\Sigma) \propto |\Sigma|^{-(k+1+b)/2} \exp(-\operatorname{tr}(\Sigma^{-1}A))$. Then

$$p(\mathbf{x}) \propto |S_t|^{-(t/2)} \quad \text{and} \quad f(\mathbf{x}) \propto |S_t + A|^{-(t+b)/2}.$$

So (19) can be computed easily. We apply this to the bivariate normal example in Section 3.1, where we choose $f$ to have $b = 1$ and

$$A = \begin{pmatrix} .5 & 1.0 \\ 1.0 & .5 \end{pmatrix}$$

Note that this alternative prior density is biased toward a positive $\rho$. By applying (20) to the augmented data sets in Section 3.1, the new posterior distribution of $\rho$ is approximated by the distribution displayed in Figure 4. The histograms of the standardized $w(j)$'s and $w^{\diamond}(j)$'s are also displayed in Figure 4. Recall that the standardized $w(j)$'s have sample variance equal to .08. In comparison, the sample variance of the standardized $w^{\diamond}(j)$'s is about .36, which still corresponds to an effective sample size of about $1{,}000/1.36 = 735$.

As pointed out by a referee, the technique of reweighting used for sensitivity analysis can also be applied to study case influence (Kass, Tierney, and Kadane 1989). For $1 \le t \le n$, $1 \le j \le m$, let $\mathbf{y}_{[-t]} = (y_1, \ldots, y_{t-1}, y_{t+1}, \ldots, y_n)$, $\mathbf{z}^*_{[-t]}(j) = (z^*_1(j), \ldots, z^*_{t-1}(j), z^*_{t+1}(j), \ldots, z^*_n(j))$, and $\mathbf{x}^*_{[-t]}(j) = (\mathbf{y}_{[-t]}, \mathbf{z}^*_{[-t]}(j))$. The posterior distribution with case $t$ deleted, $p(\theta|\mathbf{y}_{[-t]})$, can be approximated by

$$\frac{1}{W_{[-t]}} \sum_{j=1}^{m} w_{[-t]}(j) p(\theta|\mathbf{x}^*_{[-t]}(j)), \tag{21}$$

where

$$w_{[-t]}(j) = \frac{w(j)}{p(y_t|\mathbf{x}^*_{[-t]}(j))} \tag{22}$$

and $W_{[-t]} = \sum_{j=1}^{m} w_{[-t]}(j)$. The reasoning leading up to the adjusted weight (22) is as follows. To begin, $\mathbf{z}^*(j)$ can be interpreted as a sample taken from $p(\mathbf{z}|\mathbf{y})$ with associated weight $w(j)$. As a component of $\mathbf{z}^*(j)$, $\mathbf{z}^*_{[-t]}(j)$ can be considered as a sample drawn from $p(\mathbf{z}_{[-t]}|\mathbf{y})$ with associated weight $w(j)$. To delete case $t$, samples taken from $p(\mathbf{z}_{[-t]}|\mathbf{y}_{[-t]})$ are needed. This requires adjusting the original weight $w(j)$ by the factor

$$\frac{p(\mathbf{z}^*_{[-t]}(j)|\mathbf{y}_{[-t]})}{p(\mathbf{z}^*_{[-t]}(j)|\mathbf{y})} = \frac{p(\mathbf{x}^*_{[-t]}(j))}{p(y_t, \mathbf{x}^*_{[-t]}(j))} \times \frac{p(\mathbf{y})}{p(\mathbf{y}_{[-t]})}$$

$$= \frac{1}{p(y_t|\mathbf{x}^*_{[-t]}(j))} \times \frac{p(\mathbf{y})}{p(\mathbf{y}_{[-t]})}.$$

Ignoring the factor $p(\mathbf{y})/p(\mathbf{y}_{[-t]})$, which does not depend on $j$, leads to (22).

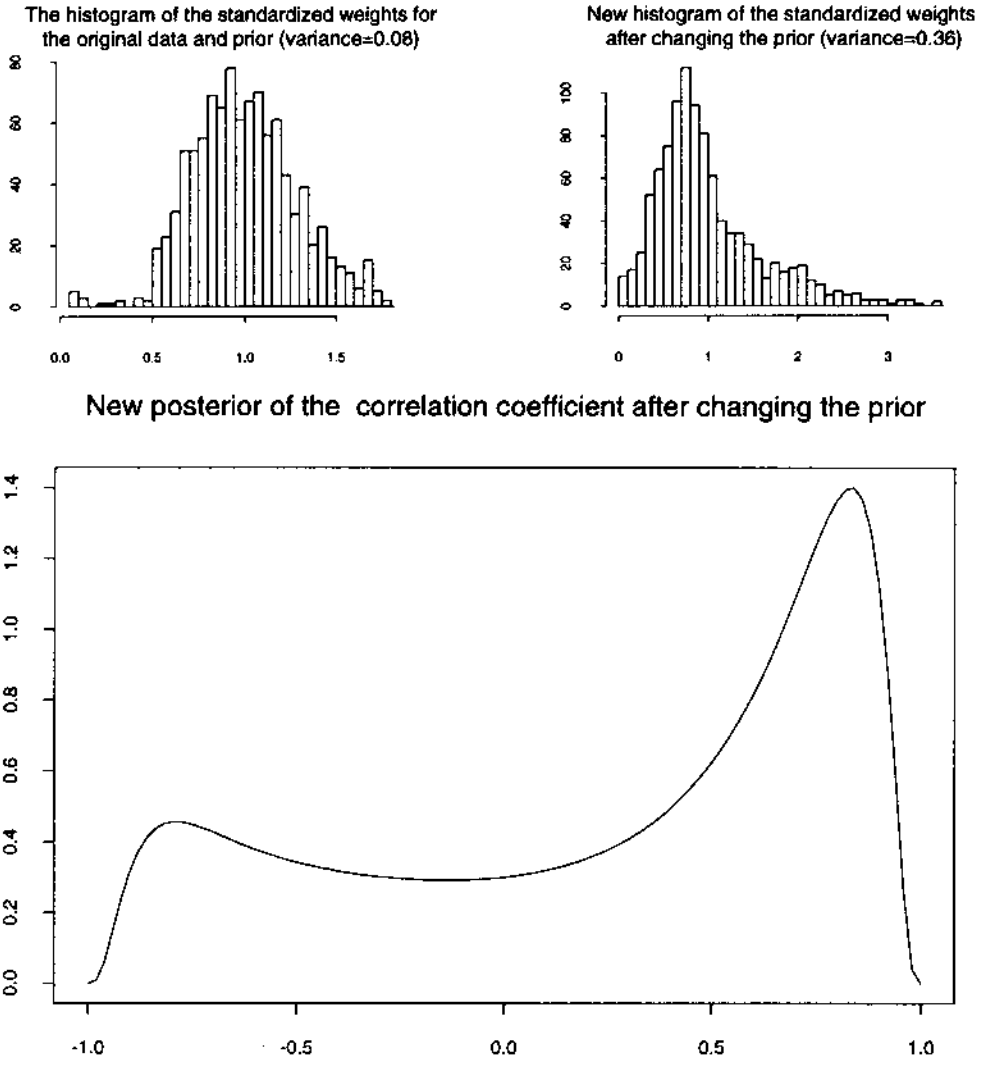


*Figure 4. Sensitivity Analysis on Murray's Data.*

Finally, it should be emphasized that the idea of reweighting also applies to augmented data sets created by Gibbs sampling. In that case, with $w(j)$ set to be 1 for all $j$, (19)–(22) are still appropriate.

## 6. DISCUSSION

For many problems where sequential imputation can be applied, Gibbs sampling can also be done. Although sequential imputation produces independent samples with different weights, Gibbs sampling generates dependent samples with equal weights. Which method is more efficient will in general depend on the problem at hand. Gibbs sampling can have problems if the serial correlations are too high (see Liu, Wong, and Kong 1994); in the case of sequential imputation the variance of the importance weights is the main concern.

Sequential imputation is most useful when the data actually arrive sequentially. In contrast, suppose that the Gibbs sampler is applied to a given set of data to generate multiple complete data sets. Suppose then that some new data are collected. If we insist on using the Gibbs sampler alone, then the previously imputed data sets will have to be abandoned and all the iterations redone by incorporating the new data. In this situation a much more efficient alternative is to simply use the augmented data sets generated from the first batch of data, which have equal weights to start with, to sequentially impute the new data. We will then have multiple complete data sets with different weights. This procedure is very similar to that described in the paragraph following (7). As long as the new data do not contain a lot more information than the first batch of data, the importance weights should be well behaved. This shows that the methods of sequential imputation and Gibbs sampling can sometimes be combined and actually complement each other. Indeed, as noted earlier, the reweighting schemes we gave for performing sensitivity and influence analyses apply equally well to complete data sets generated by the Gibbs sampler. Finally, we note that Gibbs sampling does not give direct estimates of model likelihoods as defined in (8), which is another strong point of sequential imputation.

*[Received October 1991. Revised May 1993.]*

## REFERENCES


Besag, J. (1989), "A Candidate's Formula: A Curious Result in Bayesian Prediction," *Biometrika*, 76, 183.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980), "Markov Fields and Log-Linear Interaction Models for Contingency Tables," *The Annals of Statistics*, 8, 522–539.

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, Vol. 21.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B., and Morris, C. (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Escobar, M. D. (1991), "Estimating Normal Means With a Dirichlet Process Prior," Technical Report No. 512, Carnegie Mellon University, Dept. of Statistics.

Ferguson, T. S. (1974), "Prior Distribution on Space of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Gelfand, A. E., and Kuo, L. (1991), "Nonparametric Bayesian Bioassay Including Ordered Polytomous Response," *Biometrika*, 78, 657–666.

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*, 6, 721–741.

Kass, R. E., Tierney, L., and Kadane, J. B., (1989), "Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis," *Biometrika*, 76, 663–674.

Kiiveri, H., Speed, T. P., and Carlin, J. B., (1984), "Recursive Causal Models," *Journal of the Australian Mathematical Society*, Ser. A, 36, 30–52.

Kong, A. (1992), "A Note on Importance Sampling Using Standardized Weights, Technical Report No. 348, University of Chicago, Dept. of Statistics.

Kong, A., Irwin, M., Cox, N., and Frigge, M. (1993), "Analysis of Multiple Loci Data Using Sequential Imputation," *Genetic Epidemiology*, Vol. 10.

Lauritzen, S. L., and Spiegelhalter, D. J. (1988), "Local Computation With Probabilities on Graphical Structures and Their Application to Expert Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 50, 157–224.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Liu, J. S. (1993), "Nonparametric hierarchical Bayes Via Sequential Imputations," Technical Report R-429, Harvard University, Dept. of Statistics.

Liu, J. S., Wong, W., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes," submitted to *Biometrika*.

Murray, G. D. (1977), Comment on "Maximum Likelihood from Incomplete Data Via the EM Algorithm" by A. P. Dempster, N. M. Laird, and D. B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, 39, 27–28.

Odell, P. L., and Feiveson, A. H. (1966), "A Numerical Procedure to Generate a Sample Covariance Matrix," *Journal of the American Statistical Association*, 61, 199–203.

Rubin, D. B. (1987a), "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm, Comment on Tanner and Wong (1987)," *Journal of the American Statistical Association*, 52, 543–546.

——— (1987b), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Spiegelhalter, D. J., and Lauritzen, S. L. (1990), "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks*, 20, 579–605.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.