Are Independent Parameter Draws Necessary for Multiple Imputation?

Jingchen Hu, Robin Mitra, Jerome Reiter*

Abstract

In typical implementations of multiple imputation for missing data, analysts create m completed data sets based on approximately independent draws of imputation model parameters. We use theoretical arguments and simulations to show that, provided m is large, the use of independent draws is not necessary. In fact, appropriate use of dependent draws can improve precision relative to the use of independent draws. It also eliminates the sometimes difficult task of obtaining independent draws; for example, in fully Bayesian imputation models based on MCMC, analysts can avoid the search for a subsampling interval that ensures approximately independent draws for all parameters. We illustrate the use of dependent draws in multiple imputation with a study of the effect of breast feeding on children's later cognitive abilities.

Key words: Bayesian, Missing, Nonresponse, Survey

1 Introduction

Multiple imputation (Rubin, 1987) is a widely used approach for handling missing data. The basic idea is to fill in missing values with $m > 1$ draws from predictive distributions, resulting in m completed data sets that can be analyzed or shared with others. When the imputation models meet certain conditions (Rubin, 1987, Chapter 4), analysts of the m completed data sets can make valid inferences using complete-data statistical methods and software. Specifically, the analyst computes point and variance estimates of interest with each data set and combines these estimates using simple formulas (Rubin, 1987). These formulas serve to propagate the uncertainty introduced by imputation through the analyst's inferences. See Rubin (1996), Barnard and Meng (1999), and Harel and Zhou (2007) for reviews of multiple imputation.

*Jingchen Hu is PhD candidate, Department of Statistical Science, Duke University, Durham, NC 27708 (e-mail: jh309@stat.duke.edu); Robin Mitra is Lecturer, Department of Mathematics, University of Southampton, Highfield Campus, Building 54, Southampton, United Kingdom S017 1B (e-mail: R.Mitra@soton.ac.uk); and, Jerome P. Reiter is Mrs. Alexander Hehmeyer Professor, Department of Statistical Science, Duke University, Durham, NC 27708 (e-mail: jerry@stat.duke.edu). This work was supported by the National Science Foundation grants CNS-10-12141 and SES-11-31897.

Typical approaches to multiple imputation presume either a joint model for all the data, such as a multivariate normal or log-linear model (Schafer, 1997), or use approaches based on chained equations (Van Buuren and Oudshoorn, 1999; Raghunathan *et al.*, 2001). With either approach, usually it is recommended that completed data sets be generated from approximately independent draws of the imputation model parameters (e.g., Schafer, 1997; Sinharay *et al.*, 2001; White *et al.*, 2011). Indeed, as we document later, this is the default procedure in most, if not all, popular multiple imputation software routines. This recommendation stems from Rubin (1987), who assumes independent draws when deriving the methods for multiple imputation inference. Independent draws allow for straightforward and valid inferences with modest m. This is essential for Rubin's (1987) motivating application for multiple imputation: a statistical agency sharing completed data files with the public. In those days and even today, agencies considering multiple imputation prefer to release only modest numbers of data sets ($m \leq 20$) to simplify secondary analysts' work and reduce storage needs.

More than 25 years later, the uses of multiple imputation have extended far beyond handling nonresponse in public use data (Reiter and Raghunathan, 2007), including purposes for which inconvenience does not present a real barrier to creating and using large m, such as when analysts do not intend to share data outside their research team. Indeed, recently researchers have recommended that analysts with flexibility generate larger numbers of completed data sets (say $m \approx 50$), so as to reduce variances and stabilize estimates (Graham *et al.*, 2007; White *et al.*, 2011). These recommendations continue to insist that imputations be based on approximately independent parameter draws.

In this article, we revisit the need for independent parameter draws when generating completed data sets in multiple imputation. Our examination is motivated by the following observations and questions. First, most (proper) multiple imputation procedures generate large numbers of completed data sets that are sparsely sampled to ensure independent draws. Might throwing out these already-generated, completed data sets to gain independence needlessly sacrifice inferential accuracy? Second, when imputation models involve many parameters, it can be time consuming and difficult to find a subsampling interval that ensures approximately independent draws for all parameters. Indeed, these decisions often are buried inside multiple imputation software routines so that analysts actually cannot check if the sampled parameter draws are approximately independent. Is this process and reliance on black-box routines for guessing at independence necessary, or can it be avoided by using all completed data sets?

To offer insight on these questions, we investigate theoretical implications and perform several simulation studies of using dependent parameter draws in multiple imputation. The results suggest that, when m is large and parameter values are sampled from their posterior distributions, valid multiple imputation inferences can be ob-

tained from dependent draws using the formulas of Rubin (1987). In fact, one even can gain efficiency, particularly when the effective sample size of the dependent parameter draws exceeds the number of independent draws that would be otherwise used for multiple imputation. Importantly, these findings do not hold when m is too small; here, dependent draws can lead to underestimation of multiple imputation variances and below nominal confidence interval coverage rates (for the typical case of positively correlated parameter draws).

The remainder of the article is organized as follows. In Section 2, we review the theory of multiple imputation and discuss the role of independent draws. In Section 3, we present the results of simulation studies and theoretical arguments suggesting when independence is necessary and when it is not. In Section 4, we show how these issues can matter in practice with a multiple imputation analysis of data on the effects of breast feeding on children's cognitive development. In Section 5, we offer some final remarks about the implications of our findings.

2 Review of Multiple Imputation Inferences

To describe multiple imputation, we use notation that closely follows the presentation in Si and Reiter (2011). Let $\mathbf{Y}_{inc} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ be the $n \times p$ matrix of data for the n units included in some sample; $\mathbf{Y}_{obs}$ is the portion of $\mathbf{Y}_{inc}$ that is observed, and $\mathbf{Y}_{mis}$ is the portion of $\mathbf{Y}_{inc}$ that is missing. We assume arbitrary patterns of missing data, e.g., the same variables can be present in both $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$. Here, for simplicity, we ignore variables related to the sampling design, although these should be accounted for in imputation models (Reiter *et al.*, 2006). The analyst fills in values for $\mathbf{Y}_{mis}$ with draws from the posterior predictive distribution of $(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ or from approximations to that distribution such as the sequential regression approach of Raghunathan *et al.* (2001). These draws are repeated m times to obtain m completed data sets, $\mathbf{D}^{(l)} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(l)})$ where $l = 1, ..., m$. Let $\mathbf{S}^{(m)} = (\mathbf{D}^{(1)}, ..., \mathbf{D}^{(m)})$.

In standard practice, each $\mathbf{D}^{(l)}$ is generated from approximately independent draws of imputation model parameters. For imputations based on data augmentation for fully Bayesian models, these draws can be obtained from converged MCMC chains by (i) subsampling m values spaced so that autocorrelations among all parameters in successive draws are near zero or (ii) taking the final completed data set from each of m converged chains started at independently generated values. For example, PROC MI in SAS, which uses a multivariate normal model for imputation, offers analysts both options, with a default of sampling from one long chain with a subsampling interval of 100. This interval can be modified by the user. A version of the second strategy is typically used to sample completed data sets with (not fully Bayesian) chained equations approaches. For example, the software MICE in R and Stata saves the last completed data set in each of m independently initiated rounds of sequential imputation, where each round has $c > 1$ iterations through the chained equations. The software IVEWARE for SAS uses a

similar strategy. We note that all of these methods throw away potentially many completed data sets in the process of obtaining samples from approximately independent parameter draws.

From these m completed data sets, the analyst seeks inferences about some estimand Q, for example a population mean or regression coefficient. In each $\mathbf{D}^{(l)}$, the analyst estimates Q with some estimator $\hat{q}$ and the variance of $\hat{q}$ with some estimator $\hat{u}$. For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of $\hat{q}$ and $\hat{u}$ in $\mathbf{D}^{(l)}$. The following quantities are needed for inferences:

$$\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m \tag{1}$$

$$\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m \tag{2}$$

$$b_m = \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1). \tag{3}$$

The analyst uses $\bar{q}_m$ to estimate Q and $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate $\text{Var}(Q|\mathbf{S}^{(m)})$. Inferences are based on the t-distribution, $(Q - \bar{q}_m) \sim t_{\nu_m}(0, T_m)$, with $\nu_m = (m-1)(1 + \bar{u}_m/((1+1/m)b_m))^2$ degrees of freedom, mean zero, and squared scale parameter T_m.

The rationale for using independent parameter draws is evident in Rubin's (1987) derivations of these inferential methods. To see this, let Q_{inc} and U_{inc} be the approximately unbiased point estimate (posterior mean) of Q and its (posterior) variance if $\mathbf{Y}_{inc}$ was available. Assuming noninformative prior distributions for all parameters as in Rubin (1987), we have

$$E(Q|\mathbf{Y}_{obs}) = E(E(Q|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})|\mathbf{Y}_{obs}) = E(Q_{inc}|\mathbf{Y}_{obs}) \tag{4}$$

$$\begin{aligned} Var(Q|\mathbf{Y}_{obs}) &= Var(E(Q|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})|\mathbf{Y}_{obs}) + E(Var(Q|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})|\mathbf{Y}_{obs}) \\ &= Var(Q_{inc}|\mathbf{Y}_{obs}) + E(U_{inc}|\mathbf{Y}_{obs}). \end{aligned} \tag{5}$$

As suggested by Rubin (1987, p. 85), analysts can simulate (4) and (5) by sampling values of Q_{inc} and U_{inc} from their posterior distribution, $f(Q_{inc}, U_{inc}|\mathbf{Y}_{obs})$. In particular, suppose that each $(q^{(l)}, u^{(l)})$ is a draw from the posterior distribution of (Q_{inc}, U_{inc}). Then, following Rubin (1987), we have $\bar{q}_\infty = \lim \bar{q}_m = E(Q_{inc}|\mathbf{Y}_{obs})$ as $m \to \infty$; $\bar{u}_\infty = \lim \bar{u}_m = E(U_{inc}|\mathbf{Y}_{obs})$ as $m \to \infty$; and, $b_\infty = \lim b_m = Var(Q_{inc}|\mathbf{Y}_{obs})$ as $m \to \infty$. Thus, for large n, we can use a normal approximation for inferences about Q,

$$(Q|\bar{q}_\infty, b_\infty, \bar{u}_\infty) \sim N(\bar{q}_\infty, b_\infty + \bar{u}_\infty). \tag{6}$$

Rubin (1987) further presumes that each $(q^{(l)}, u^{(l)})$ are independently distributed according to

$$(q^{(l)}|\bar{q}_\infty, b_\infty) \sim N(\bar{q}_\infty, b_\infty) \tag{7}$$

$$(u^{(l)}|\bar{u}_\infty) \sim (\bar{u}_\infty, << b_\infty) \tag{8}$$

where the notation $x \sim (y, << z)$ in (8) means that x has mean y and variance much less than z.

Assuming noninformative prior distributions for all parameters, this independence implies that

$$(\bar{q}_\infty|\bar{q}_m, b_\infty) \sim N(\bar{q}_m, b_\infty/m) \tag{9}$$

$$((m-1)b_m/b_\infty|b_m) \sim \chi^2_{m-1} \tag{10}$$

$$(\bar{u}_\infty \mid \bar{u}_m) \sim (\bar{u}_m, << b_\infty/m). \tag{11}$$

Thus, from (6) and (9) we have

$$(Q|\mathbf{S}^{(m)}, b_\infty, \bar{u}_\infty) \sim N(\bar{q}_m, (1+1/m)b_\infty + \bar{u}_\infty). \tag{12}$$

From (11), one can replace $\bar{u}_\infty$ with $\bar{u}_m$ so that

$$(Q|\mathbf{S}^{(m)}, b_\infty) \sim N(\bar{q}_m, (1+1/m)b_\infty + \bar{u}_m). \tag{13}$$

The t-approximation to $f(Q|\mathbf{S}^{(m)})$ follows from (10) and (13), with degrees of freedom obtained by matching the first two moments of the posterior distribution of $(\nu_m T_m/((1+1/m)b_\infty + \bar{u}_m)|\mathbf{S}^{(m)})$ to those of a χ^2 distribution with ν_m degrees of freedom.

3 Theoretical Considerations and Simulations With Dependent Draws

For modest m, assuming independence in (7) is necessary to ensure that b_m is an unbiased estimate of b_∞, which in turn is necessary to substitute b_m for b_∞ in the variance in (13). However, and crucially for our argument, for large m Rubin's (1987) simulation approach does not require independent draws of Q_{inc}. Rather, it requires that analysts use simulation to construct consistent estimates of the expectation in (4) and variance in (5), which can be done with dependent draws. This is akin to summarizing a posterior distribution from a full (i.e., not thinned) scan of parameter draws generated from an MCMC algorithm: for long chains, the sample mean and sample variance

of the dependent parameter draws are consistent estimates of the posterior mean and posterior variance (Tierney, 1994). In fact, the analogy is precise for fully Bayesian imputation models estimated via MCMC, since Q_{inc} is an unknown parameter with a posterior distribution.

Formally, and assuming a fully Bayesian imputation procedure, suppose that we have a set of m dependent draws of $(q^{(l)}, u^{(l)})$ derived from a MCMC algorithm that has converged to the limiting distribution, $f(Q_{inc}, U_{inc} \mid \mathbf{Y}_{obs})$. We note that such convergence also is assumed when using independent draws. By the ergodic theorem, $\bar{q}_m$ is consistent for $E(Q_{inc} \mid \mathbf{S}^{(m)})$, $\bar{u}_m$ is consistent for $E(U_{inc} \mid \mathbf{S}^{(m)})$, and b_m is consistent for $Var(Q_{inc}|\mathbf{Y}_{obs})$. Thus, for large n and infinite m, we can continue to base inferences on (6), even with dependent draws.

Because all $(q^{(l)}, u^{(l)})$ are not jointly independent, we cannot assume (9) through (11), hence nor (13). However, for sufficiently large m and a converged MCMC sampler, it is reasonable to assume that the Monte Carlo errors in the sampled moments $(\bar{q}_m, \bar{u}_m, b_m)$ are inconsequential as proportions of $b_\infty + \bar{u}_\infty$; that is, we assume $\bar{q}_m \approx \bar{q}_\infty$, $\bar{u}_m \approx \bar{u}_\infty$, and $b_m \approx b_\infty$. With this assumption, we can replace (6) with

$$(Q|\mathbf{S}^{(m)}) \sim N(\bar{q}_m, b_m + \bar{u}_m), \tag{14}$$

which can be used directly for inferences. We note that when m is large, the usual multiple imputation reference distribution, $(Q - \bar{q}_m) \sim t_{\nu_m}(0, T_m)$, is essentially equivalent to (14), since b_m is generally modest for typical amounts of missing information. This approximate equivalence offers analysts the convenience of using existing software routines for multiple imputation inferences, even with dependent draws.

This argument suggests that it is sensible to use all the completed data sets generated during the data augmentation steps in fully Bayesian imputation models. Can it be advantageous? To approach this question, we again turn to the literature on MCMC. In particular, Geyer (1992) and MacEachern and Berliner (1994) show that using the full set of parameter draws sampled in a converged MCMC (after tossing out the burn-in) generally results in more precise summaries of posterior distributions, including posterior means and variances, than using only draws from subsamples of the full chain. Further, using all samples gets around the difficulties of choosing a subsampling interval. Hence, in addition to being feasible, using all completed data sets offers potential benefits.

To illustrate the validity and potential benefits of multiple imputation with many dependent draws, we turn to a simple simulation scenario. We generate 10000 data sets, each comprising $n = 1000$ observations and two

variables distributed as

$$y_{1i} \sim N(0,1) \tag{15}$$

$$y_{2i} \sim Bernoulli(\pi_i), \quad \pi_i = \frac{e^{1+2y_{1i}}}{1+e^{1+2y_{1i}}}. \tag{16}$$

Let $r_i = 1$ if y_{2i} is missing, and let $r_i = 0$ otherwise. In each complete data set, we randomly generate missing values for y_2 by independently sampling from Bernoulli distributions with

$$p(r_i = 1) = \frac{e^{-.5+.5y_{1i}}}{1+e^{-.5+.5y_{1i}}}. \tag{17}$$

In any data set, this generates about 40% missing values in y_2 under a missing at random mechanism (Rubin, 1976).

We implement multiple imputation of missing y_2 using a Bayesian logistic regression,

$$y_{2i} \sim Bern(p_i), \quad log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 y_{1i} \tag{18}$$

with independent, diffuse priors for β_0 and β_1. We perform the data augmentation by (i) sampling values of (β_0, β_1) conditional on a current version of $\mathbf{Y}_{inc}$ and (ii) sampling values of $\mathbf{Y}_{mis}$ conditional on a current version of (β_0, β_1). To sample from the full conditional distribution of (β_0, β_1), we use a standard Metropolis step with two independent $N(0, .025)$ as a proposal distribution. Across the 10000 data sets, this results in roughly a 39% acceptance rate. Trace plots of parameters suggest convergence generally with 2500 consecutive draws. Autocorrelations among all parameters die down after lag 30. We implement three multiple imputation scenarios: $m = 10$ independent draws, $m = 50$ independent draws, and all $m = 2500$ dependent draws.

In each data set, we estimate $E(y_2)$, β_0, and β_1 using maximum likelihood estimation, which we then feed into Rubin's (1987) multiple imputation inferences. Table 1 summarizes the simulated coverage rates and average lengths of 95% confidence intervals across the 10000 replications. For all parameters, the coverage rates when using $m = 2500$ dependent draws are well-calibrated, as is also the case for the independent draws. Generally, the intervals based on $m = 2500$ dependent draws have similar properties as those based on $m = 50$ independent draws, with a suggestion of very slight gains in precision due to the larger m. Compared to $m = 10$ independent draws, however, using the larger m with dependent draws offers roughly 3% reduction in interval length. This reduction comes essentially for free, since for any replication we already generate the 2500 data sets when running the MCMC.

Estimand	Dependent $m = 2500$	Independent $m = 50$	Independent $m = 10$
$E(y_2)$	94.80 (.068)	94.72 (.069)	94.73 (.070)
β_0	95.21 (.477)	94.39 (.481)	94.86 (.497)
β_1	94.79 (.659)	94.76 (.663)	94.47 (.683)

Table 1: Results of simulation study of multiple imputation based on large number of dependent draws. Entries in the table are the percentage of the 95% confidence intervals that cover the true parameter and, in parentheses, the average length of the 95% confidence interval. All numbers based on 10000 replications.

Estimand	Consecutive $m = 10$	Consecutive $m = 25$	Consecutive $m = 50$	Independent $m = 50$
$E(y_2)$	93.55	93.90	94.56	94.72
	.00023, .00006	.00023, .00007	.00023, .00007	.00023, .00008
β_0	91.53	92.88	93.82	95.39
	.00888, .0042	.00888, .0049	.00888, .0053	.00887, .0060
β_1	91.75	92.76	93.58	94.76
	.01809, .0071	.01808, .0082	.01808, .0090	.01807, .0103

Table 2: Results of simulation study of multiple imputation based on small numbers of dependent draws. Entries in the table are the percentage of the 95% confidence intervals that cover the true parameter followed on the next line by, in order, the average values of $\bar{u}_m$ and b_m. All numbers based on 10000 replications.

While the results in Table 1 suggest that dependent draws can be used for multiple imputation, we emphasize that m must be large for this to be the case. In particular, dependent draws from MCMC may not offer reliable estimates for small numbers of dependent samples, since the consistency results depend on large m. To illustrate this empirically, we repeat the simulation from Table 1 but now use consecutive draws taken from the chain after convergence. We consider three cases with consecutive samples, namely $m \in \{10, 25, 50\}$. Table 2 summarizes the simulated coverage rates along with the average values of $(\bar{u}_m, b_m)$ across 10000 replications. The results based on $m = 50$ independent draws from Table 1 are shown as a baseline. Using a too small number of consecutive dependent draws results in coverage rates below the nominal 95%. This is not due to underestimation of $\bar{u}_m$: for any parameter, its expectation is nearly identical across all cases. Rather, as evident in Table 2, b_m is the culprit. It tends to underestimate b_∞ with too small a number of dependent draws, with increasing bias as m gets smaller. This is not surprising, since the positive autocorrelation among consecutive draws generally reduces the variance of consecutive sets of $q^{(l)}$.

The empirical results confirm that analysts planning to use dependent draws from an MCMC must ensure a sufficiently large number of them, so that the completed data quantities in (1) – (3) closely estimate $(\bar{q}_\infty, \bar{u}_\infty, b_\infty)$. To assess this convergence and, hence, if (14) is plausible for a particular set of m dependent samples, one useful and convenient indicator is the effective sample size (ESS) of $\bar{q}_\infty$. This can be computed with the m values of $q^{(l)}$ using standard routines, for example with the *coda* package in R. Intuitively, if the m values of $q^{(l)}$ have a small

	Consecutive		Subsampled	
Estimand	$m = 2500$	$m = 500$	$m = 500$	$m = 100$
$E(y_2)$	94.80 (.068)	94.91 (.068)	94.96 (.068)	94.96 (.069)
Avg. ESS	482	116	324	79
β_0	95.21 (.477)	95.19 (.476)	95.30 (.478)	95.27 (.478)
Avg. ESS	353	89	264	67
β_1	94.79 (.659)	94.68 (.658)	94.79 (.659)	94.64 (.660)
Avg. ESS	312	81	241	63

Table 3: Results of simulation study of multiple imputation based on dependent draws with varying effective sample sizes. Entries in the table are the percentage of the 95% confidence intervals that cover the true parameter; in parentheses, the average length of the 95% confidence interval; and, on the next line the average effective sample size (ESS) of the point estimates. All numbers based on 10000 replications.

ESS, we cannot count on the corresponding $\bar{q}_m$ being close to $\bar{q}_\infty$, nor b_m being close to b_∞. In the simulations in Table (2), the average ESS of all parameters are between 27 and 30 when $m = 50$, and between 21 and 23 when $m = 25$; these are quite small values. In contrast, the average ESS in the simulations in Table 1 when $m = 2500$ all exceed 300.

As a rule of thumb for basing multiple imputation inferences on dependent draws, we suggest that analysts require the minimum $ESS \geq 100$ for all $\bar{q}_\infty$ of interest. Using (7) and (10) with $m = ESS$ as (very) rough approximations to the sampling distributions of $\bar{q}_m$ and b_m, this would imply a standard error of $1/\sqrt{100} \approx 10\%$ of b_∞ when approximating $\bar{q}_\infty$ with $\bar{q}_m$; this is typically a small number. Similarly, with $ESS = 100$ we expect the ratio b_m/b_∞ to have a standard error of $\sqrt{2/99} \approx 14\%$. When b_∞ is modest compared to $\bar{u}_\infty$, which is usually the case in missing data settings, the approximation error in b_m typically should be small compared to $b_\infty + \bar{u}_\infty$.

To explore this further, we consider three additional scenarios that use the simulation runs from Table 1 with 2500 draws. First, we select samples of $m = 500$ consecutive draws to represent a case with smaller ESS than the $m = 2500$ scans and larger ESS than the $m = 50$ scans. Second and third, we thin the resulting 500 and original 2500 length scans by keeping every fifth draw, resulting in samples of $m = 100$ and (thinned) $m = 500$ draws. The thinned scans have reduced autocorrelations, thus representing additional ESS for comparison. Table 3 summarizes key results over the 10000 simulation runs. Even with a minimum ESS in the neighborhood of 60 or 80, the 95% confidence intervals are well-calibrated. Taken together with the undercoverage in the $(m \leq 50, \max(ESS) \leq 30)$ scans from Table 2, the results are in reasonable accord with the proposed rule of thumb.

The simulation studies involve fairly simple models with small p for computational convenience; running MCMC samplers until convergence in repeated sampling studies can be computationally expensive. Since the theoretical arguments in support of using dependent draws do not depend on p or particular distributional assumptions—as long as posterior distributions of the quantities of interest are approximately Gaussian—we expect the overall

trends in the simulations to hold for other settings. However, with more complex data, analysts are likely to require a larger m to ensure convergence of the MCMC sampler and sufficient effective sample size.

We also repeat the simulation from Table 1 with $n = 100$. Results exhibit the same pattern: using a large number of dependent draws offers efficiency gains. For too small n, however, (6) may not hold due to failure of the normal approximation. In this case, Barnard and Rubin (1999) show that multiple imputation inference with independent draws should be based on $(Q - \bar{q}_m) \sim t_{\hat{v}_m}(0, T_m)$, where $\hat{v}_m = (1/v_m + 1/v_{obs})^{-1}$ and v_{obs} is an estimate of the observed-data degrees of freedom. Following similar logic as the large-n case, we can substitute consistent estimates of $(\bar{u}_\infty, b_\infty)$ from dependent draws in the expressions for (v_m, v_{obs}). For large m, we conjecture that doing so can generate efficiency gains over using independent draws.

4 Multiple Imputation in Breastfeeding Study

We now apply multiple imputation with dependent draws to handle missing data in a study of the effect of breastfeeding on children's later cognitive development. These data were previously used by Mitra and Reiter (2011, 2012) to develop methods for propensity score matching with multiply-imputed data. Our description of the data closely follows their presentation, although we do not employ matching techniques here.

The data comprise a subset of the National Longitudinal Survey of Youth (NLSY). This survey began in 1979 with a nationally representative sample of 12686 young men and women in the U.S. aged 14 to 22 years at that time. This cohort was interviewed annually until 1994 and biannually afterwards. After 1986, the NLSY collected detailed information on children born to women in the study. These children represent the unit of analysis for our application. We include only first born children to avoid complications due to birth order and family nesting. In addition, we discard 307 children with missing breastfeeding duration and children born before 1979. The resulting data set comprises 3748 children, of whom 1306 have completely observed data.

We seek to estimate a linear regression of Peabody individual assessment test math scores (PIATM), which is administered to children ages 5 or 6, on fifteen covariates. These include five categorical variables: the child's race (Hispanic, black, or other), the mother's race (Hispanic, black, Asian, white, Hawaiian/Pacific Islander/American Indian, or other), child's sex, and two variables indicating the presence of a spouse/partner or grandparents at birth. We categorize three of the ten continuous variables: mother's weeks of work in the previous year (worked 0 weeks, worked less than 48 weeks, worked no less than 48 and less than 52 weeks, and worked 52 weeks), weeks preterm at birth (0 weeks preterm, less than 5 weeks preterm, 5 or more weeks preterm), and weeks of breastfeeding (less than 24 weeks of breastfeeding, at least 24 weeks of breastfeeding). The remaining seven continuous variables

include the number of years between 1979 and when the mother gave birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, the number of days that the child spent in hospital, the number of days that the mother spent in hospital, and family income. The full set of variables, along with fraction of missing values in each, is reported in Table 4.

We implement multiple imputation with the R software package "mix," which can be freely downloaded at `http://sites.stat.psu.edu/~jls/misoftwa.html`. This uses a general location model (Schafer, 1997) for imputation, which is a joint model in which the categorical variables follow a log-linear model and the continuous variables (after transformation) follow multivariate normal distributions with common variance and means given by linear functions of the categorical variables. Specifically, in the log-linear model we include all main effects plus two interactions (mother's and child's race, and presence of spouses and grandparents) suggested by exploratory data analyses. Following Mitra and Reiter (2011), we apply Box-Cox transformations to several of the continuous variables to improve normality assumptions. These transformations are used both for multiple imputation and for the linear regression model. In the multivariate normal models, we include main effects for all levels of the categorical variables in the regression for the mean. We generate 5000 completed data sets using the built-in MCMC routines in "mix," derived from the $m = 5000$ dependent draws. The minimum effective sample size of $\bar{q}_{5000}$ across all coefficients is 1274. We also obtain repeated realizations of (approximately) independent draws by subsampling 500 sets of $m = 10$ completed data sets via systematic sampling, leaving a gap of 500 draws between successive data sets.

Table 4 displays the point estimates and interval lengths for the case with $m = 5000$ dependent draws, as well as the corresponding average and standard deviation of interval lengths across the 500 sets of $m = 10$ independent draws. For any estimand, by design $\bar{q}_{5000} = \sum_{h=1}^{500} \bar{q}_{10}^{(h)}/500$ where h indexes a set of ten completed data sets. Although not shown here, each $\bar{u}_m$ hardly varies across imputation scenarios, as expected. Further, we see that b_{5000} approximately equals the average of the 500 sets of b_{10}, again as expected since MCMC theory suggests that b_{5000} converges to b_∞ (and b_{10} is an unbiased estimator of b_∞). However, using smaller m to ensure independent draws has a cost: the intervals based on the independent draws are longer than those based on dependent draws. Moreover, using $m = 10$ results in additional instability in inferences, as can be seen in the standard deviations of the interval lengths in Table 4. For example, the 95% confidence interval length for the coefficient of the logarithm of family income ($+.5$) plausibly could be less than 1.0 in one set of $m = 10$ completed data sets and more than 1.4 in another set. With $\bar{q}_m$ values expected to be centered on .67, this variability could result in intervals containing (or nearly containing) zero for some sets and not close to containing zero in other sets.

Finally, although not shown here, results based on systematic subsamples with $m = 100$ approximately inde-

Estimand	% missing	$\bar{q}_m$	Avg. CI Length (SD) $m = 5000$	$m = 10$
Intercept	—	79.32	13.24	14.04 (1.83)
Mother's race - black	.9	1.76	10.90	11.43 (1.41)
Mother's race - Asian	.9	4.90	13.80	14.69 (2.12)
Mother's race - white	.9	3.26	6.35	6.61 (.77)
Mother's race - Hawaiian/PI/American Indian	.9	2.79	7.43	7.73 (.82)
Mother's race - other	.9	1.50	7.07	7.35 (.82)
Child's race - black	0	-2.08	10.62	11.14 (1.36)
Child's race - other	0	.45	5.89	6.13 (.71)
Child's sex - female	0	.91	1.97	2.05 (.21)
Spouse/partner present at birth	4.2	.99	4.55	4.76 (.60)
Spouse/partner not know about child until after birth	4.2	.78	3.24	3.40 (.41)
Grandparents in house 1 yr. before birth - Yes	4.1	-.98	3.26	3.42 (.41)
Weeks mother worked in yr. before birth - 1-48 weeks	23.5	.70	3.24	3.45 (.54)
Weeks mother worked in yr. before birth - 49-51 weeks	23.5	.59	4.12	4.37 (.68)
Weeks mother worked in yr. before birth - 52 weeks	23.5	1.89	3.69	3.92 (.59)
Weeks preterm - 1-4 weeks	4.8	.98	2.70	2.81 (.31)
Weeks preterm - >5 weeks	4.8	.91	6.59	6.87 (.84)
Breastfeeding at least 24 weeks - Yes	0	1.09	2.84	2.95 (.31)
Sq. root(mother's age - mother's age in 1979)	0	-.40	1.33	1.41 (.19)
Sq. root(mother's AFQT score)	4.9	1.15	.65	.68 (.08)
Child's birth weight	1.4	.01	.06	.06 (.01)
Log(number of days child spent in hospital+.5)	6.6	-1.34	3.01	3.12 (.33)
Log(number of days mother spent in hospital+.5)	6.8	.08	3.29	3.40 (.35)
Mother's attained education	4.3	.49	.63	.65 (.08)
Log(family income+.5)	24.6	.67	1.08	1.18 (.21)

Table 4: Regression coefficient estimates and 95% confidence interval lengths after multiple imputation with one set of $m = 5000$ dependent draws, and the corresponding averages and standard deviations for 500 disjoint sets of $m = 10$ independent draws. The response variable PIATM has 36% missing values.

pendent draws closely resemble the results from the $m = 5000$ dependent draws.

5 Concluding Remarks

For multiple imputation based on Bayesian joint models, the theoretical and simulation results indicate that analysts can obtain valid inferences using dependent draws, provided that m is large. Large m is often available in such settings, because analysts using MCMC typically run (multiple) long chains to ensure convergence. There can be advantages to using dependent draws, as analysts can avoid the task of identifying appropriate subsampling intervals and possibly increase accuracy by using larger m. Of course, analysts always can run the chain long enough to ensure a large number of independent draws after appropriate subsampling, in which case the inferences from dependent and independent draws likely will be very similar. There also can be disadvantages to using many

dependent draws when each completed data analysis is computationally expensive. The cost in terms of timeliness and computing resource usage from repeating the completed data analysis many times could outweigh the benefits from using the dependent draws.

As a practical guideline, we recommend that analysts estimate the amount of time and computing resources needed per completed data analysis with a trial run based on a modest-sized subsample of the completed data sets from the converged MCMC sampler. The analyst then can project a value of m for which computation costs are acceptable, and determine if the effective sample size is large enough at that m. Our simulations required a minimum effective sample size of at least 50 for valid inferences with dependent draws, but to be safe we recommend minimum effective samples sizes of at least 100 (along with careful checking that the MCMC sampler has converged). If using all completed data sets from the converged chain is too expensive, analysts can use subsamples of completed data sets. We generally expect m draws from a thinned chain to yield higher effective sample sizes than m consecutive draws.

The theoretical arguments of Section 3, based on ergodic theorems for MCMC, do not automatically apply for approximations to full Bayesian models like chained equations. Although these approaches mimic Gibbs samplers, the collection of conditional models may not actually correspond to a proper joint distribution (Liu *et al.*, 2012). Nonetheless, since chained equations approaches have been shown empirically to perform comparably to proper Bayesian imputation models, at least for relatively straightforward modeling tasks (Van Buuren *et al.*, 2006), we conjecture that using dependent draws, i.e., using more than the final completed data set in each cycle of iterations, should offer similar advantages.

References

Barnard, J. and Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research* **8**, 17–36.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–955.

Geyer, C. (1992). Practical Markoc chain Monte Carlo. *Statistical Science* **7**, 473–483.

Graham, J., Olchowski, A., and Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* **8**, 206–213.

Harel, O. and Zhou, X. H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* **26**, 3057–3077.

Liu, J., Gelman, A., Hill, J., and Su, Y. S. (2012). On the stationary distribution of iterative imputations (arxiv.org/abs/1012.2902v2). Tech. rep., Department of Statistics, Columbia University.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician* **488**, 188–190.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine* **30**, 6, 627–641.

Mitra, R. and Reiter, J. P. (2012). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research* (online early).

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Si, Y. and Reiter, J. P. (2011). A comparison of posterior simulation and inference by combining rules for multiple imputation. *Journal of Statistical Theory and Practice* **5**, 335–347.

Sinharay, S., Stern, H., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods* **6**, 317–329.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

Van Buuren, S. and Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Tech. rep., Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 4, 377–399.