Journal of Privacy and Confidentiality (2010) **2**, Number 2, pp. 3–19

Model Selection when Multiple Imputation Is Used to Protect Confidentiality in Public Use Data

Satkartar K. Kinney*, Jerome P. Reiter†, and James O. Berger‡

Abstract. Several statistical agencies use, or are considering the use of, multiple imputation to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. For example, agencies can release partially synthetic datasets, comprising the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. We describe how secondary analysts of such multiply-imputed datasets can implement Bayesian model selection procedures that appropriately condition on the multiple datasets and the information released by the agency about the imputation models. We illustrate by deriving Bayes factor approximations and a data augmentation step for stochastic search variable selection algorithms.

Keywords: Confidentiality, Disclosure, Model selection, Multiple imputation, Synthetic data.

1 Introduction

Statistical agencies and other organizations that disseminate data to the public are ethically, practically, and often legally required to protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, Rubin (1993) and Little (1993) proposed that agencies utilize multiple imputation approaches. For example, agencies can release the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called partially synthetic datasets (Reiter, 2003a).

In recent years, statistical agencies have begun to use partially synthetic approaches to create public use data for major surveys. In 2007, the U.S. Census Bureau released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of Social Security benefits information and dozens of other highly sensitive variables.[1] The Census Bureau also released partially synthesized origin-destination matrices, i.e., where people live and work, available to the public as maps via the web.[2] The Census Bureau plans to protect the identities of people in group quarters (e.g., prisons, shelters) in the next release of public use files from the American Community Survey by replacing demographic data for people at high disclosure risk with imputations. Partially synthetic, public use datasets are in

*National Institute of Statistical Sciences, mailto:saki@niss.org
†Duke University, mailto:jerry@stat.duke.edu
‡Duke University, mailto:berger@stat.duke.edu

the development stage for the U.S. Census Bureau's Longitudinal Business Database, Longitudinal Employer-Household Dynamics survey, and American Community Survey veterans and full sample data. Statistical agencies in Canada (Mantel and Hidiroglou, 2008, Bocci and Beaumont, 2009), Germany (Drechsler et al., 2008), and New Zealand (Graham and Penny, 2005) also are investigating the approach. Other applications of partially synthetic data are described by Kennickell (1997), Abowd and Woodcock (2001; 2004), Abowd and Lane (2004), Little et al. (2004), Reiter (2005b), Mitra and Reiter (2006), Reiter and Mitra (2009), An and Little (2007), and Reiter and Raghunathan (2007).

Under certain conditions, analysts can obtain valid inferences for finite population quantities or model parameters from the multiple datasets by combining point and variance estimates computed with each dataset. The combining rules differ from those developed by Rubin (1987) for multiple imputation for missing data (see Reiter and Raghunathan, 2007). These methods are predicated on the analyst having specified a model of interest. What if this is not the case, and the analyst wants to compare many models and select ones that are best? How should the analyst utilize the multiple datasets for model selection, particularly when different imputed datasets may yield different model comparisons? Currently, analysts do not have any principled procedures for model selection with partially synthetic data.

This article provides a framework for constructing such procedures. The approach is to derive Bayesian posterior model probabilities by paying careful attention to the appropriate conditioning information. This includes recognizing and accounting for any differences between the models used by the agency to impute synthetic data and the model of interest to the secondary data analyst. This framework provides a starting point from which to implement different Bayesian model selection procedures, for example Bayes factors and stochastic search variable selection algorithms (George and McCulloch, 1997, Geweke, 1996).

The remainder of this article is organized as follows: Section 2 reviews model selection approaches with multiple imputation for missing data and motivates the appropriate marginal likelihood for partially synthetic data. This is used to develop Bayes factor approximations, described in Section 3, and an illustrative stochastic search variable selection procedure, described in Section 4. Section 5 concludes with some additional remarks and directions for future work.

2 The appropriate marginal likelihood

To set the stage for our discussion of model selection, we first describe the process of generating synthetic data. Let Y_{obs} be the $n \times p$ matrix of data for the n sampled units. We presume that Y_{obs} is fully observed; see Reiter (2004) for partial synthesis with missing data. Let $Z_l = 1$ if unit l is selected to have any of its data replaced with synthetic values, and let $Z_l = 0$ for those units with all data left unchanged. Let

[1] www.sipp.census.gov/sipp/synth/_data.html

[2] On The Map, http://lehdmap.did.census.gov/

$Z = (Z_1, \ldots, Z_n)$. Let Y_{rep} be the values of Y_{obs} which are to be replaced with m imputations; let $Y_{rep}^{(*)} = \{Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}\}$, where $Y_{rep}^{(i)}$ are the imputed (replaced) values in the ith synthetic dataset; and, let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}. The $Y_{rep}^{(i)}$ are generated from the conditional distribution of $(Y_{rep}^{(i)} \mid Y_{obs}, Z)$, or a close approximation of it. Each synthetic data set, $D_{syn}^{(i)}$, then comprises $(Y_{rep}^{(i)}, Y_{nrep}, Z)$. The entire collection of m data sets, $D_{syn} = \{D_{syn}^{(i)}, i = 1, \ldots, m\}$, is released to the public. Releasing multiple datasets enables secondary analysts to account for the uncertainty that results from replacing observed values with draws from statistical models; see Reiter and Mitra (2009) for further discussion. Typically, the agency also releases information about the models used to generate the synthetic data so that analysts can get a sense of what analyses are supported by the synthetic data. For example, agencies might include the code for synthetic data generation with public releases of data. Or, they might include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education."

Although there are no existing investigations of model selection with partially synthetic data, there is some literature on model selection with multiple imputation for missing data. For example, Ball (2001) provides an ad-hoc approach for combining the BIC, and Wood et al. (2007) stack the completed datasets and use weighted regression to apply variable selection procedures. Yang et al. (2005) propose two Bayesian model selection approaches, one of which can be used when the analyst and imputer are distinct. In this approach, the analyst seeks $f(M|Y_{obs})$, where M is the analyst's model under consideration. Here, Y_{obs} represents the data for the n observations that are not missing. The analyst determines $f(M|Y_{obs})$ by computing the posterior probability for M in each of m completed datasets using MCMC techniques, and then applies the combining rules of Rubin (1987). This implicitly assumes agreement between the analyst and the imputer models. Further, this approach cannot be used with synthetic data since Y_{obs} is not available to the analyst and using $f(M|Y_{nrep})$ is generally not sensible. Thus, model inferences for synthetic data must be based on $f(M|D_{syn})$.

When generating partially synthetic data, the agency specifies a collection of imputation models, M^*, for imputing all variables in Y_{rep}. Clearly, different specifications of M^* can result in different realizations of D_{syn}. The analyst of D_{syn}, however, may posit some other model M, distinct from M^*, to explain the relationship between a particular response variable and potential predictors. When this occurs, the analysis model is said to be uncongenial (Meng, 1994) to the agency's imputation model. Several authors have discussed conditions under which, for fixed M and M^*, inferences for parameters can be valid with mismatched imputation and analysis models (Rubin, 1987, Meng, 1994, Reiter, 2003b, Raghunathan et al., 2001). Model selection on partially synthetic data, in which M is not fixed but uncertain, has not been previously discussed in the literature.

When the analyst knows the form of M^* from agency released meta-data, other potential specifications of M^* become irrelevant. The analyst can and indeed should consider M^* as data that provide additional information about M. Thus, Bayesian model posterior probabilities should explicitly account for M^* in the conditioning, i.e.,

$$f(M|D_{syn}, M^*) \propto f(D_{syn}|M, M^*)p(M|M^*). \tag{1}$$

This posterior probability can be used to construct Bayes factors and model search algorithms, as we illustrate in the sections that follow.

Explicitly conditioning on M^* highlights an important feature of model selection with synthetic data. The analyst's model selection decisions are sensitive to the agency's choice of synthesis models. For example, if M includes a dependence between an outcome and some predictor, but M^* sets the two variables to be independent, then model comparisons are likely to favor models excluding that predictor. This lack of "model congeniality" presents serious problems when significant relationships truly exist but are omitted from M^*, i.e., the imputer made poorly grounded assumptions.

One might sensibly ask why analysts given M^* would perform model selection at all; that is, why not simply use M^*? There are several settings where using different M makes sense. First, an analyst may seek high probability models for a regression of some response that has not been synthesized on many potential explanatory variables, some of which have been synthesized. For example, suppose that the data comprise synthetic values only for age, race, sex, and marital status, and the analyst seeks a model for income that might include those variables (and others not synthesized). In this case, M^* is for the covariates, and M is for the response. Second, even when the analyst's response is subject to synthesis, M^* may be tailored to particular subsets of individuals as opposed to the entire dataset. For example, An and Little (2007) synthesize monetary values only when they exceed threshholds, so that the released monetary values are a mix of observed and simulated data. Models tailored specifically to these records may not describe relationships across the entire distribution of monetary values. An analyst looking to describe the entire distribution may want to compare several models that differ from M^*. Third, as with multiple imputation for missing data (Rubin, 1987), agencies are advised to err on the side of being inclusive when building synthesis models, in the sense that it is safer to include irrelevant predictors in M^* than to exclude relevant ones. Thus, analysts may want to identify more parsimonious models than M^* for some responses.

3 Bayes factors

This section uses (1) to derive approximate Bayes factors useful for model selection. Although (1) is not readily available, we show that one can integrate over Y_{rep} and obtain a Monte Carlo approximation. We also present an alternate approximation using the synthetic data implicates $Y_{rep}^{(i)}, i = 1, \ldots, m$ instead of Monte Carlo draws. We illustrate the performance of the approximations using simulation studies.

Suppose that the analyst seeks to compute a Bayes factor to compare two models, M_1 and M_0, using partially synthetic data. Using the appropriate posterior model probabilities from (1), we have

$$\frac{f(M_1|Y_{nrep}, Y_{rep}^{(*)}, M^*)}{f(M_0|Y_{nrep}, Y_{rep}^{(*)}, M^*)} = \frac{f(Y_{nrep}, Y_{rep}^{(*)}|M_1, M^*)}{f(Y_{nrep}, Y_{rep}^{(*)}|M_0, M^*)} \times \frac{p(M_1|M^*)}{p(M_0|M^*)}.$$

As frequently done for Bayes factors, we let the prior odds ratio equal one, so that the Bayes factor is just the likelihood ratio. The likelihood $f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*)$ is not readily available but can be obtained by integrating over the replaced values Y_{rep}. We have

$$f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) = \int f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M, M^*) f(Y_{nrep}, Y_{rep}|M, M^*) dY_{rep}. \quad (2)$$

Since $Y_{rep}^{(*)}$ are generated independently of M, $f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*, M) = f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*)$. Furthermore, since (Y_{rep}, Y_{nrep}) are original data values, they do not depend on M^*, so that $f(Y_{nrep}, Y_{rep}|M^*, M) = f(Y_{nrep}, Y_{rep}|M)$. Thus, (2) simplifies to

$$f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) = \int f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*) f(Y_{nrep}, Y_{rep}|M) dY_{rep}. \quad (3)$$

We can interpret this integral as the average of the original data marginal likelihood over the distribution of Y_{rep} implied by $Y_{rep}^{(*)}$ and M^*. Values of Y_{rep} that could not feasibly generate $Y_{rep}^{(*)}$ have relatively low density, so that the first part of the integral serves to discount implausible values of Y_{rep} (under M^*) for the averaging.

We now re-express this integral in a way conducive to Monte Carlo simulation. First, we note that

$$f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*) = \frac{f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) f(Y_{rep}^{(*)}, Y_{nrep}|M^*)}{f(Y_{rep}, Y_{nrep}|M^*)}. \quad (4)$$

We can substitute the right hand side of (4) for the first term inside the integral in (3), so that

$$f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) \propto \int f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) \frac{f(Y_{nrep}, Y_{rep}|M)}{f(Y_{nrep}, Y_{rep}|M^*)} dY_{rep} \quad (5)$$

We drop $f(Y_{rep}^{(*)}, Y_{nrep}|M^*)$ because it does not depend on M or Y_{rep} and is constant for all model comparisons. The integral in (5) can be approximated with a Monte Carlo estimate as

$$f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) \quad \propto \quad \frac{1}{K} \sum_{k=1}^{K} \frac{f(Y_{nrep}, Y_{mrep}^{(k)}|M)}{f(Y_{nrep}, Y_{mrep}^{(k)}|M^*)} \quad (6)$$

where $Y_{mrep}^{(k)}$ is a draw from $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ and K is the number of Monte Carlo draws. The term $f(Y_{nrep}, Y_{mrep}^{(k)}|M^*)$ represents a density rather than a likelihood since

M^* is assumed to be known and fixed. This density is evaluated only for data that are part of Y_{rep}; for example, if the agency replaces all values above a threshold t by simulating from a truncated normal distribution, the analyst computes $f(Y_{nrep}, Y^{(k)}_{mrep}|M^*)$ only for those replaced values. The expression in (6) can be computed by the analyst without access to the confidential data, provided that M^* is made available by the imputer.

The distribution $f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*)$ can be complicated because $Y^{(*)}_{rep}$ is simulated from models with parameters that are functions of Y_{rep}. To illustrate with a simple example, suppose that the data comprise one variable $Y \sim N(\mu, \sigma^2)$. To create synethetic data the agency replaces all n values of Y_{obs} by drawing from the posterior predictive distribution for new Y, which for large n is approximately $f(Y|Y_{obs}, M^*) = N(\bar{y}, (1 + 1/n)s^2)$. Given an infinite number of imputed datasets, i.e., $m = \infty$, the analyst can learn $\bar{y}$ and s^2 by averaging the sample means and variances of the datasets. Thus, $f(Y_{rep}|Y^{(*)}_{rep}, M^*) = f(Y_{rep}|\bar{Y}_{rep} = \bar{y}, Var(Y_{rep}) = s^2, M^*)$, so that the analyst needs to draw values of Y_{rep} having sample mean of $\bar{y}$ and sample variance of s^2. With more complicated data settings and finite numbers of imputed datasets (so that the parameters of M^* are not known exactly), this distribution can be quite complex and computationally expensive to simulate from.

To simplify matters, we obtain draws of $f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*)$ by using M^* as the model that generates Y_{rep} (see Section 5 for further discussion of this). Let γ be the parameters in M^* if it were the model that generated Y_{rep}. We use

$$f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*) = \int f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*, \gamma) f(\gamma|Y_{nrep}, Y^{(*)}_{rep}, M^*) d\gamma. \qquad (7)$$

We estimate $f(\gamma|Y_{nrep}, Y^{(*)}_{rep}, M^*)$ using standard techniques for partially synthetic data. That is, in each $D^{(i)}_{syn}$, $i = 1 \ldots, m$, we estimate the posterior mean and variance of γ. We combine these means and variances using the methods of Reiter (2003a), which ultimately result in a normal approximation for the posterior distribution of γ. Thus, each $Y^{(k)}_{mrep}, k = 1, \ldots, K$ is obtained by taking a draw of γ, say γ_k, followed by a draw from $f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*, \gamma_k)$.

The expression in (6) can be numerically unstable and difficult to compute. We use Laplace approximations to simplify computations. Let θ be the d-dimensional parameter vector for M. Going back to (2), we have

$$f(Y_{nrep}, Y^{(*)}_{rep}|M, M^*) = \iint f(Y_{nrep}, Y^{(*)}_{rep}|M, M^*, \theta, \gamma) p(\theta|M) p(\gamma|M^*) d\theta d\gamma$$

so that

$$\log f(Y_{nrep}, Y^{(*)}_{rep}|M, M^*) \quad \approx \quad \log \sum_{k=1}^{K} \frac{f(Y_{nrep}, Y^{(k)}_{mrep}|M, \bar{\theta})}{f(Y_{nrep}, Y^{(k)}_{mrep}|M^*, \bar{\gamma})} - \frac{d}{2} \log n. \qquad (8)$$

where $\bar{\theta}$ and $\bar{\gamma}$ are the maximum likelihood estimates of θ and γ obtained from $(Y_{nrep}, Y^{(*)}_{rep})$

under M and M^*, respectively. Terms that are $O_n(1)$ or less in (8) are dropped, as is usually done with the BIC.

For two models M_0 and M_1 with dimensions k_0 and k_1, respectively, we can approximate -2 times the logarithm of the Bayes factor using (8). We call this Approximation 1, which is given by

$$-2\log\frac{f(Y_{nrep},Y_{rep}^{(*)}|M_1,M^*)}{f(Y_{nrep},Y_{rep}^{(*)}|M_0,M^*)} \approx -2\log\left(\sum_{k=1}^{K}\frac{f(Y_{nrep},Y_{mrep}^{(k)}|M_1,\bar{\theta}_1)}{f(Y_{nrep},Y_{mrep}^{(k)}|M^*,\bar{\gamma})}\right)+ 2\log\left(\sum_{k=1}^{K}\frac{f(Y_{nrep},Y_{mrep}^{(k)}|M_0,\bar{\theta}_0)}{f(Y_{nrep},Y_{mrep}^{(k)}|M^*,\bar{\gamma})}\right)+(k_1-k_0)\log n. \tag{9}$$

As the imputations $Y_{rep}^{(i)}, i=1,\ldots,m$ are readily available, an approximation using these in place of $Y_{mrep}^{(k)}, k=1,\ldots,K$ in (9) is simpler to compute. We call this Approximation 2, which is

$$-2\log\frac{f(Y_{nrep},Y_{rep}^{(*)}|M_1,M^*)}{f(Y_{nrep},Y_{rep}^{(*)}|M_0,M^*)} \approx -2\log\left(\sum_{i=1}^{m}\frac{f(Y_{nrep},Y_{rep}^{(i)}|M_1,\bar{\theta}_1)}{f(Y_{nrep},Y_{rep}^{(i)}|M^*,\bar{\gamma})}\right) +2\log\left(\sum_{i=1}^{m}\frac{f(Y_{nrep},Y_{rep}^{(i)}|M_0,\bar{\theta}_0)}{f(Y_{nrep},Y_{rep}^{(i)}|M^*,\bar{\gamma})}\right)+(k_1-k_0)\log n. \tag{10}$$

This may be similar to Approximation 1 since $Y_{rep}^{(*)}$ and $Y_{mrep}^{(k)}$ are drawn from similar distributions.

We now illustrate the Bayes factor approximations in (9) and (10) using simple simulation scenarios. The basic set-up is to repeatedly generate observed datasets, generate partially synthetic datasets for each observed dataset, and then compute Approximations 1 and 2 using each set of $m = 5$ partially synthetic datasets.

Let the observed data have $n = 10,000$ records and seven variables. For each unit j we generate the first six variables, X_j, from standard normal distributions and the seventh variable, y_j, from $N(X_j\beta, 1)$. Here, β is drawn from a mixture distribution such that, for $l = 1,\ldots,6$, $\beta_l = 0$ with probability π_l, and $\beta_l \sim N(0,1)$ with probability $1-\pi_l$. Each π_l is drawn from independent $Beta(2,2)$ distributions. This allows for a range of models and coefficients to be selected for the true model in each simulation. We presume that the agency replaces all n values of Y, i.e., $Y_{rep} = (y_1,\ldots,y_n)$, but leaves all $n \times 6$ values of X unchanged, i.e., $Y_{nrep} = (X_1,\ldots,X_n)$. The agency draws $m = 5$ copies of $Y_{rep}^{(*)} = Y_{rep}^{(1)},\ldots,Y_{rep}^{(m)}$ from the posterior predictive distribution, $f(Y_{rep}|X,Y_{obs},M^*)$, using the saturated model, $Y \sim N(X\gamma,\tau)$ as M^*.

For each simulation, we compute the approximate Bayes factors under Approximation 1 and 2 (using $K = 100$) for all 64 possible models for the regression of Y on X. Because we know the true model in any simulation run, we can determine the estimated rank of the true model, which ideally should be first. Table 1 displays the frequencies of the estimated rankings of the true model for both approximations for 1000 true models. As a baseline, we also include the frequencies when using the BIC on Y_{obs} (which would not be possible for analysts to do). Model selection based on Approximations 1 and 2 performs well, resulting in rankings close to those based on the observed data. The two approximations have similar properties in this scenario. We note that this evaluation does not take into account cases where there are multiple top models with similar Bayes factors.

Table 1: Comparison of Bayes factor approximations

Est. ranks:	1	2	3	4	5	6+
Obs. BIC	916	46	18	8	5	7
Approx 1	879	52	37	16	9	7
Approx 2	874	80	22	10	7	7

4 Stochastic search algorithm

Stochastic search variable selection (SSVS), which we review briefly below, is a popular Bayesian model selection approach. This section illustrates how to implement SSVS algorithms for partially synthetic data. We present SSVS algorithms for linear models, though the framework can be extended to other models. We illustrate the algorithms using simulation studies.

Stochastic search variable selection algorithms search for models having high posterior probability by (i) starting with the full model containing all p candidate predictors; (ii) choosing mixture priors that allow predictors to drop out by zeroing their coefficients; and (iii) running a Gibbs sampler (Gelfand and Smith, 1990) relying on conditional conjugacy to sample from the posterior distribution. The resulting draws will differ in the subset of predictors having non-zero coefficients and, after discarding initial burn-in draws, one can estimate the posterior model probabilities using the proportion of MCMC draws spent in each model. In general, all 2^p models will not be visited; hence, many or most of the candidate models will be estimated to have zero posterior probability. Although there is no guarantee that the model with highest posterior probability will be visited when p is large, SSVS tends to quickly locate good models. Model-averaged estimates may also be obtained for model coefficients by averaging the parameter estimates over all MCMC draws, and marginal inclusion probabilities for each predictor estimated by the proportion of draws spent in models containing that predictor.

In order to use the SSVS approach for synthetic data, we augment the parameter space with Y_{rep} so that after drawing plausible values of Y_{rep} conditional on the synthetic data and M^*, the Gibbs sampler can proceed as in the observed data case.

Suppose an analyst is interested in finding a parsimonious subset of predictors that adequately predicts some response variable. Let θ_M be the parameters of candidate model M, and as before, let M^* be the agency's imputation model used to generate the synthetic data $Y_{rep}^{(*)}$. The posterior distribution of interest is $f(\theta_M, M|Y_{nrep}, Y_{rep}^{(*)}, M^*)$. Augmenting this distribution with Y_{rep}, we have

$$\begin{aligned} & f(\theta_M, Y_{rep}, M|Y_{nrep}, Y_{rep}^{(*)}, M^*) \\ &\quad = f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) f(\theta_M, M|Y_{nrep}, Y_{rep}^{(*)}, Y_{rep}, M^*) \qquad (11) \\ &\quad = f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) f(\theta_M, M|Y_{nrep}, Y_{rep}) \qquad (12) \end{aligned}$$

The simplification from (11) to (12) follows because if the observed data Y_{rep} are known, there is no use for the synthetic data $Y_{rep}^{(*)}$ and imputation model M^*. We approximate the distribution of $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ as in (7). The Gibbs sampler proceeds by drawing Y_{rep} from $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$, and then drawing from $f(\theta_M, M|Y_{nrep}, Y_{rep})$ as in the observed data case.

To obtain the full conditional posterior distributions for the Gibbs sampler, the analyst needs to specify a prior distribution, $p(\theta_M, M)$. Proper distributions are desired for Bayes factors to be well-defined (Pauler et al., 1999), but otherwise any reasonable prior specification for an observed-data model selection problem may be used.

4.1 Simulation 1: Only dependent variable synthesized

To illustrate the approach, we implement the SSVS algorithm on simulated datasets using the design of Section 3.2. As in that simulation, we randomly select true models and tabulate the number of times the true model is assigned the highest posterior probability. We also examine model probabilities for several simulation runs in detail. For comparison, we run similar SSVS algorithms on both the observed and partially synthetic data. We use a convenient prior specification that has performed well in other observed-data stochastic search algorithm problems; details are in the Appendix. We use the same prior structure for both the observed and partially synthetic data algorithms.

We generated 1000 true models and corresponding observed datasets. All 64 possible models were drawn multiple times, with different coefficients in each draw. The SSVS algorithm with the synthetic data assigned the highest probability to the true model in 884 out of these 1000 runs. As a baseline, the true model was ranked highest in 902 cases when using the observed data. The true model was ranked below fifth 19 times when using the synthetic data and 15 times when using the observed data. The lowest ranking of the true model was 11th in the synthetic data and 8th in the observed data.

Capturing the true model is not the only measure of success for a model search algorithm. Often, several models have approximately the same posterior probability, so that the differences among them are minimal. We examined the runs in which the observed and/or synthetic data model searches failed to find the true model, and we found nearly all of these runs involved true null models or models with very small coefficients, so that there was little difference between them and other models with zero or small coefficients.

Table 2: Simulation 1 SSVS posterior model probabilities, null model true

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_2	0.146	X_2	0.129
X_3	0.135	X_6	0.122
null	0.116	X_5	0.114
X_1	0.115	X_1	0.085
X_6	0.115	X_3	0.084
X_5	0.105	null	0.070
X_4	0.099	X_4	0.062
X_2, X_6	0.018	X_2, X_5	0.045
X_3, X_5	0.017	X_4, X_5	0.030
X_1, X_2	0.016	X_1, X_6	0.027

Table 3: Simulation 1 SSVS marginal inclusion probabilities, null model true

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	0.167	0.215	0.185	0.159	0.173	0.165
Synthetic	0.187	0.288	0.180	0.175	0.322	0.202

To illustrate the performance on a single dataset, we generate Y independently of X so that $\beta = (0, 0, 0, 0, 0, 0)$. We then generate partially synthetic data as before from the saturated model, and we run the observed and synthetic data Gibbs sampling algorithms. In one run, the observed data algorithm visited 33 models in 1000 iterations, and the synthetic data algorithm visited 39. The top ten models are given in Table 2 and the marginal inclusion probabilities in Table 3. The top ten models and the inclusion probabilities differ only slightly in the observed and synthetic data SSVS algorithms. Neither the observed nor synthetic data algorithms selected the true null model as the highest posterior probability model; however, in both cases the results provide little evidence to support large models.

Finally, we illustrate the algorithms' performance on a dataset with only one important predictor, now setting $\beta = (1, 0, 0, 0, 0, 0)$. In one run, the observed data algorithm visited 16 models while the synthetic data algorithm visited 25. Table 4 shows the top ten models and their posterior probabilities, and Table 5 gives the marginal inclusion probabilities. The results in both cases convincingly identify the true model as the highest posterior probability model.

4.2 Simulation 2: Dependent and independent variables synthesized

Using the same data generation methods as in Simulation 1, we now let $Y_{rep} = (Y, X_1)$ and $Y_{nrep} = (X_2, \ldots, X_6)$. The imputation procedure generates $f(Y, X_1 | X_2, \ldots, X_6)$

Table 4: Simulation 1 SSVS posterior model probabilities, one important predictor

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_1	0.601	X_1	0.413
X_1, X_3	0.210	X_1, X_3	0.248
X_1, X_2	0.052	X_1, X_2	0.071
X_1, X_4	0.033	X_1, X_2, X_3	0.047
X_1, X_2, X_3	0.023	X_1, X_4	0.038
X_1, X_6	0.023	X_1, X_3, X_4	0.032
X_1, X_5	0.019	X_1, X_3, X_5	0.029
X_1, X_3, X_4	0.014	X_1, X_6	0.026
X_1, X_3, X_5	0.004	X_1, X_5	0.024
X_1, X_2, X_4	0.004	X_1, X_3, X_6	0.017

Table 5: Simulation 1 SSVS marginal inclusion probabilities, one important predictor

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	1.000	0.081	0.254	0.062	0.034	0.034
Synthetic	1.000	0.165	0.380	0.088	0.085	0.068

using a bivariate normal regression. We use the same analysis model as in Simulation 1, so that the stochastic search algorithm is also the same. However, since M^* and Y_{rep} are different, the specification of $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is different. The simulation of Y_{rep} is described in the appendix.

We proceed as in Simulation 1, with model searches run on observed data and synthetic data for 1000 draws of data and models. Each of 64 possible true models was drawn between 6 and 23 times. The observed data algorithm ranked the true model highest 791 times while the synthetic data model search picked the true model 735 times. As before, we examine the Gibbs sampler output for one run of each algorithm when the true model is the null model. Table 6 and Table 7 display the results, which tell a similar story as those from Simulation 1.

5 Discussion

In these derivations, we presumed that the form of the agency's model M^* is known to the user. It is possible that the agency may not release M^*, perhaps out of fear that this would increase confidentiality risks too much, or out of convenience if the models are complicated to describe. Without M^*, the analyst has two possible approaches. First, the analyst can make a reasonable guess about M^*. The general advice to agencies on specifying imputation models is to include as many variables as possible (Meng, 1994, Schafer, 1997); hence, the analyst could assume this advice has been followed and use a

Table 6: Simulation 2 posterior model probabilities, null model true

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_6	0.141	X_2	0.128
X_5	0.140	X_4	0.118
X_2	0.131	X_5	0.112
X_4	0.125	X_3	0.091
X_3	0.124	X_1	0.088
X_1	0.118	X_6	0.071
null	0.104	null	0.065
X_2, X_4	0.017	X_2, X_4	0.039
X_1, X_2	0.012	X_1, X_4	0.029
X_1, X_4	0.010	X_3, X_4	0.025

Table 7: Simulation 2 marginal inclusion probabilities, null model true

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	0.157	0.185	0.162	0.177	0.176	0.170
Synthetic	0.216	0.271	0.196	0.289	0.188	0.182

series of saturated, chained regression models (Raghunathan et al., 2001), perhaps with reasonable transformations and interaction terms as a reasonable approximation to M^*. Second, the analyst can take ad hoc approaches similar to those described in Section 2. Even with M^*, some analysts may prefer simpler approaches than those presented in Section 3. Evaluating the efficacy of ad hoc approaches, both mathematically and via simulation, is an important topic for future study.

As with all analyses based on synthetic data, particularly when a substantial portion of a dataset has been replaced with simulated values, the validity of the model selection procedures depends on the validity of assumptions embedded in the synthesis models. In particular for the model selection procedures, the approximation in (7) implicitly assumes that M^* is in some sense close to the “correct” model for Y_{rep}. If this is not the case, the simplification in (7) may result in inaccurate estimates of the posterior probabilities of the models. For example, if an important predictor is omitted from M^*, model selection procedures will tend to favor models without the predictor. In our simulations, (7) was reasonable; further study is needed to characterize the sensitivity of model selection to degrees of violations of (7).

In some cases agencies may intentionally alter the posterior predictive distributions used in M^* for the purposes of increasing disclosure protection; for example, Machanavajjhala et al. (2008) and Abowd et al. (2009) alter the parameters of informative prior distributions to ensure probabilistic differential privacy. In these cases M^* may differ substantially from the “correct” model and as above, (7) may yield inaccurate posterior probabilities unless the perturbation parameters are made available and included in M^*. In extreme cases or if the details of the perturbation are concealed, then biases may still result.

Arguably, it is essential for agencies releasing data to provide information about M^*. This would enable analysts to understand (at least somewhat) the limitations of the released data for their analyses. Indeed, as suggested by a reviewer, the importance of M^* for model selection—and inference in general—makes the case for sharing M^*. In general, the disclosure risks and data utility associated with releasing such information have not been quantified (Karr, 2009). Reiter and Mitra (2009) illustrated small increased risk of identification disclosure associated with releasing M^* with synthetic data; however, it is not clear if or when this additional risk is large enough to offset potential gains in utility.

Throughout the article, we focused on selecting the predictors to include in models. We did not discuss estimation of parameters associated with selected models. Given M, analysts could use the combining rules of Reiter (2003a) for parameters with approximately normal posterior distributions. Alternatively, analysts could use parameter estimates obtained during the Bayes factor or SSVS computations. We did not compare the accuracy of these two approaches.

6 Appendix

6.1 Prior specification and posterior computation

This section describes the prior specification used in the simulation examples of the stochastic search variable selection algorithm and describes the posterior computation. The full conditional posterior distributions used for both simulations are in the sections that follow. Let $J = (J_1, \ldots, J_6)$ be a vector of indicator variables such that $J_l = 1$ if variable l is included in the current model and $J_l = 0$ otherwise, where $l = 1, \ldots, 6$. Let β_J be the vector of nonzero elements of β in the model. We use a Zellner-type prior, given by $(\beta_J | J, \sigma^2) \sim N(0, \sigma^2(X'_J X_J)^{-1}/g)$, where the $n \times k_J$ matrix X_J is the matrix X with columns corresponding to $J_l = 0$ excluded, $g \sim G(\frac{1}{2}, \frac{N}{2}), (\sigma^2 | J) \propto \frac{1}{\sigma^2}$, and $J_l \sim Be(p_0), l = 1, \ldots, p$, with $Be(p_0)$ denoting a Bernoulli distribution with prior probability p_0 and $G(a, b)$ denoting the Gamma distribution with mean a/b and variance a/b^2.

By updating the full conditional posterior of J in the Gibbs sampler, the algorithm is able to move between models with different dimensions (Smith and Kohn, 1996).

The Gibbs sampler proceeds by iteratively sampling from the full conditional posterior distribution of Y_{rep}, followed by the full conditional posterior distributions of β and σ^2, as well as J and g. The details of these distributions are given below. After discarding draws from 'an initial burn-in period, the draws of J can be used to determine both the posterior model probabilities using the percent of times each model is visited and the marginal inclusion probabilities for a l-th predictor using the percent of the time that $J_l = 1$. Model-averaged estimates of the parameter coefficients and associated uncertainties may also be obtained from the draws of β.

6.2 Full conditional posterior distributions, Simulation 1

The full conditional posterior distribution of Y_{rep} is $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ which is,

$$\begin{aligned} & f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) = \\ & \qquad \int f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*, \gamma, \tau) f(\gamma, \tau|Y_{nrep}, Y_{rep}^{(*)}, M^*) d\gamma d\tau. \end{aligned} \tag{13}$$

The form of this distribution for the illustrative example is $N(X\gamma, \tau)$, where $p(\gamma|Y_{nrep}, Y_{rep}^{(*)}, M^*) = N(\bar{\gamma}, T_p)$, and $\bar{\gamma}$ and T_p are the posterior mean and variance of γ computed with the methods in Reiter (2003a). The distribution $p(\tau|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is taken to be $(n-p)\bar{s}^2\chi_{n-p}^{-2}$, where $\bar{s}^2 = \sum_{i=1}^m (Y_{rep}^{(i)} - X\hat{\gamma}^{(i)})'(Y_{rep}^{(i)} - X\hat{\gamma}^{(i)})/m(n-1)$, and $\hat{\gamma}^{(i)}$ is the estimate of γ obtained from $D_{com}^{(i)}$.

The remaining full conditional posteriors follow from the joint posterior distribution $f(\beta_J, \sigma^2, J, Y_{rep}|Y_{nrep}, Y_{rep}^{(*)})$ and prior specification through straightforward algebraic routes and are given by:

- $f(\beta_J|Y_{nrep}, Y_{rep}, \sigma^2, M, g) = N(\hat{\beta}_J, V_J)$, where $\hat{\beta}_J = (X_J'X_J)^{-1}X'Y_{rep}$ and $V_J = (X_J'X_J)^{-1}(1/\sigma^2 + g)^{-1}$.
- $p(J_l = 1|J_{-l}, Y_{nrep}, Y_{rep}, \beta, \sigma^2, g) = 1/(1+h_l)$, obtained by integrating out β_J and σ^2 as in Smith and Kohn (1996), where

$$\begin{aligned} h_l &= \frac{1-p_{0l}}{p_{0l}}\left(1+\frac{1}{g}\right)^{1/2}\frac{S(J_l=0)}{S(J_l=1)}, && (14) \\ S(J) &= (Y_{rep}'Y_{rep} - \hat{\beta_J}'V_J^{-1}\hat{\beta_J})^{-n/2}, && (15) \end{aligned}$$

  and $S(J_l = 0)$ is equivalent to $S(J)$ but with the element J_l of J set to 0, so $\hat{\beta_J}$ and V_J may need to be recomputed to correspond to $J_l = 0$. Similarly for $S(J_l = 1)$.
- The hyperparameter g has a Gamma posterior given by

$$G\left(\frac{k_J+1}{2}, \frac{\beta_J'X_J'X_J\beta_J/\sigma^2 + n}{2}\right),$$

  where $k_J = \sum_{l=1}^p I(J_l = 1)$.
- The posterior $f(\sigma^2|Y_{nrep}, Y_{rep}, \beta, J, M, g)$ is given by

$$G\left(\frac{k_J+n}{2}, \frac{(Y_{rep} - X_J\beta_J)'(Y_{rep} - X_J\beta_J) + g\beta_J'X_J'X_J\beta_J}{2}\right).$$

6.3 Full conditional posteriors: Simulation 2

We factor $f(Y_{rep}|Y_{nrep}, Y^{(*)}_{rep}, M^*)$ as $f(Y|X_1, \ldots, X_p, Y^{(*)}_{rep}, M^*)$ $f(X_1|X_2, \ldots, X_p, Y^{(*)}_{rep}, M^*)$, where $Y^{(*)}_{rep}$ are the synthetic values of (Y, X_1), the distribution $f(Y|X, Y^{(*)}_{rep}, M^*)$ is $N(X\gamma_1, \tau_1)$, and $f(X_1|X_2, \ldots, X_p, Y^{(*)}_{rep}, M^*) = N(X_{2:p}\gamma_2, \tau_2)$.

Draws of Y_{rep} are updated in the Gibbs sampler as follows:

1. Draw τ_2 from $(n-p-1)\bar{s}^2_2\chi^{-2}_{n-p-1}$, where $\bar{s}^2_2 = \sum_{i=1}^{m}(X_1^{(i)} - X_{2:p}\hat{\gamma}_2^{(i)})'(X_1^{(i)} - X_{2:p}\hat{\gamma}_2^{(i)})/m(n-1)$.
2. Draw γ_2 from $N(\bar{\gamma}_2, T_2)$, where $\bar{\gamma}_2$ and T_2 are the posterior mean and variance of γ_2, as defined in Reiter (2005a).
3. Draw X_1 from $N(X_{2:p}\gamma_2, \tau_2)$.
4. Draw τ_1 from $(n-p)\bar{s}^2_1\chi^{-2}_{n-p}$, where $\bar{s}^2_1 = \sum_{i=1}^{m}(Y^{(i)}_{rep} - X\hat{\gamma}_1^{(i)})'(Y^{(i)}_{rep} - X\hat{\gamma}_1^{(i)})/m(n-1)$.
5. Draw γ_1 from $N(\bar{\gamma}_1, T_1)$, where γ_1 and T_1 are the posterior mean and variance of γ_1, as defined in Reiter (2005a).
6. Draw $Y^{(k)}_{mrep}$ from $N(X\gamma_1, \tau_1)$.

The rest of the Gibbs sampler steps are the same as in Simulation 1.

References

Abowd, J. M., Gehrke, J., and Vilhuber, L. (2009). Parameter exploration for synthetic data with privacy guarantees for *OnTheMap*. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.*

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag. 282–289.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland. 251–277.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag. 290–297.

An, D. and Little, R. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170(4):923–940.

Ball, R. D. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: Approximate analysis using the BIC. *Genetics*, 159:1351–1364.

Bocci, C. and Beaumont, J.-F. (2009). Synthetic data creation for the Cross National Equivalent File. In *Symposium 2009, Longitudinal Surveys: From Design to Analysis*. Statistics Canada.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008). A new approach for disclosure control in the IAB Establishment Panel–Multiple imputation for a better data access. *Advances in Statistical Analysis*, 92:439–458.

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.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374.

Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5 – Proceedings of the Fifth Valencia International Meeting*. Oxford University Press. 609–620.

Graham, P. and Penny, R. (2007). Multiply imputed synthetic data files. *Official Statistics Research Series*, 1. Wellington, NZ: Statistics New Zealand. http://www.stats.govt.nz/sitecore/content/statisphere/Home/official-statistics-research/series/volume-1-2007.aspx

Karr, A. F. (2009). The role of transparency in statistical disclosure limitation. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*. Washington, D.C.: National Academy Press. 248–267.

Little, R., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: Wiley. 141–152.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice *OnTheMap*. In *IEEE 24th International Conference on Data Engineering*. 277–286. doi:10.1109/ICDE.2008.4497436.

Mantel, H. and Hidiroglou, M. (2008). Synthetic data for confidentiality. Technical report, Statistics Canada Methodology Branch Working Paper.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9:538–558.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrar, ed., *Privacy in Statistical Databases 2006*, vol. 4302 of *LNCS*. New York: Springer-Verlag. 177–188.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the Americal Statistical Association*, 94:448.

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. (2003a). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189.

Reiter, J. P. (2003b). Model diagnostics for remote access servers. *Statistics and Computing*, 13:371–380.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30:235–242.

Reiter, J. P. (2005a). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131:365–377.

Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21:441–462.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1:99–110.

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

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

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468.

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

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343.

Wood, A. M., White, I. R., and Royston, P. (2007). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27:3227–3246.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61:498–506.