MODEL SELECTION AND MULTIVARIATE INFERENCE USING DATA MULTIPLY IMPUTED FOR DISCLOSURE LIMITATION AND NONRESPONSE

by

Satkartar K. Kinney

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Jerome P. Reiter, Supervisor

_____
Dr. James O. Berger

_____
Dr. Alan F. Karr

_____
Dr. David L. Banks

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2007

ABSTRACT

MODEL SELECTION AND MULTIVARIATE INFERENCE USING DATA MULTIPLY IMPUTED FOR DISCLOSURE LIMITATION AND NONRESPONSE

by

Satkartar K. Kinney

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Jerome P. Reiter, Supervisor

_____
Dr. James O. Berger

_____
Dr. Alan F. Karr

_____
Dr. David L. Banks

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2007

Abstract

This thesis proposes some inferential methods for use with multiple imputation for missing data and statistical disclosure limitation, and describes an application of multiple imputation to protect data confidentiality. A third component concerns model selection in random effects models.

The use of multiple imputation to generate partially synthetic public release files for confidential datasets has the potential to limit unauthorized disclosure while allowing valid inferences to be made. When confidential datasets contain missing values, it is natural to use multiple imputation to handle the missing data simultaneously with the generation of synthetic data. This is done in a two-stage process so that the variability may be estimated properly. The combining rules for data multiply imputed in this fashion differ from those developed for multiple imputation in a single stage. Combining rules for scalar estimands have been derived previously; here hypothesis tests for multivariate components are derived.

Longitudinal business data are widely desired by researchers, but difficult to make available to the public because of confidentiality constraints. An application of partially synthetic data to the U. S. Census Longitudinal Business Database is described. This is a large complex economic census for which nearly the entire database must be imputed in order for it to be considered for public release. The methods used and analytical results for synthetic data generated for a subgroup are described. Modifications to the multiple imputation combining rules for population data are also developed.

Model selection is an area in which few methods have been developed for use with multiply-imputed data. Careful consideration is given to how Bayesian model selection can be conducted with multiply-imputed data. The usual assumption of cor-

respondence between the imputation and analyst models is not amenable to model selection procedures. Hence, the model selection procedure developed incorporates the imputation model and assumes that the imputation model is known to the analyst.

Lastly, a model selection problem outside the multiple imputation context is addressed. A fully Bayesian approach for selecting fixed and random effects in linear and logistic models is developed utilizing a parameter expanded stochastic search Gibbs sampling algorithm to estimate the exact model-averaged posterior distribution. This approach automatically identifies subsets of predictors having nonzero fixed coefficients or nonzero random effects variance, while allowing uncertainty in the model selection process.

Acknowledgements

Firstly, I would like to thank my advisor, Jerry Reiter, without whom this could not be done, for his enthusiastic support and guidance throughout my graduate career. I would also like to thank my prelim advisor David Dunson for his direction on the random effects paper, included in Chapter 4, Professor Jim Berger for his help with Chapter 5, and all my committee members for helpful comments. Many of the NSF-ITR project participants were helpful with Chapter 3, including Javier Miranda, Ron Jarmin, Arnold Reznek, and Simon Woodcock. I am also grateful to many other faculty, staff, and students for being generally helpful, friendly, and providing a stimulating graduate school environment, including Michael Lavine, Merlise Clyde, Dalene Stangl, David Banks, Kris Moyle, Abel Rodriguez, Simon Lunagomez, Gavino Puggioni, and many others. My family deserves a special note of gratitude for their ongoing unconditional support of my endeavors, as does the Kinney family, and most especially, my husband Kris.

Some of the research in this thesis was conducted while the author was a Special Sworn Status researcher of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the author and do not necessarily reflect the views of the Census Bureau. This work has been screened to insure that no confidential data are revealed. This work was supported by NSF grant ITR-0427889.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multiple imputation was first proposed for handling nonresponse in large complex surveys. The goal was to facilitate valid inferences when the data producer and the ultimately many end users of the data were distinct entities, potentially having different knowledge, capabilities, and ideas about the cause of missingness, and likely conducting different analyses. In this scenario, the burden of modeling the missing data mechanism lies on the data producer, who may have skills and information unavailable to the users, while the users are able to focus on their analyses without learning new or complex missing data methods (Rubin, 1996).

Many agencies find multiple imputation an appealing approach for handling missing data in public-use files as they would like all users be able to obtain the same inferences. As the validity of inferences depends to an extent on the models used to impute the missing values, agencies releasing multiply-imputed data should include in their imputation models as many relevant variables from the dataset as possible, including design variables, and completely observed variables. This will increase the scope of valid analyses that may be of interest to future analysts. Releasing information about the imputation model will also aid analysts in assessing the analytic validity of analyses performed (Schafer, 1997; Meng, 1994; Reiter *et al.*, 2006).

Multiple imputation is now widely used to handle missing data by agencies as well as individual users. Several software packages, including R, SAS, and SPlus, have routines that simplify the process for both filling in missing values with multiple imputations and drawing inferences from completed datasets. In addition to missing data, multiple imputation is now used in other applications, including statistical

disclosure limitation (Rubin, 1993; Little, 1993) and measurement error (Clogg *et al.*, 1991; Cole *et al.*, 2006). These are reviewed in Reiter and Raghunathan (2007).

This thesis will focus on the applications to missing data and statistical disclosure limitation for large samples and populations. The main contributions of this thesis are to expand the inferential methods available to users of certain applications of multiply imputed data, and to undertake the generation of multiply-imputed public-use files for the U. S. Census Bureau Longitudinal Business Database. The remainder of this chapter provides background material on the different applications of multiple imputation for which inferential methods are proposed. Section 1.1 describes multiple imputation for missing data, Section 1.2 describes multiple imputation for disclosure limitation, and Section 1.3 describes two-stage multiple imputation. Section 1.4 describes the structure of the remaining chapters.

1.1 Multiple imputation for missing data

There is much literature on multiple imputation for missing data. Rubin (1987), the standard reference for multiple imputation for nonresponse, derives combining rules for obtaining inferences from multiply-imputed data for scalar and multivariate estimands. Additional testing procedures for multivariate estimands were developed by Li *et al.* (1991a) and Meng and Rubin (1992). Barnard and Rubin (1999) and Reiter (2007b) adapt the combining rules for small samples. Rubin (1996) discusses several issues surrounding multiple imputation for missing data and provides an extended bibliography. Schafer (1997) and Little and Rubin (2002) describe several procedures for generating imputations for missing data.

Multiple imputations for missing data are generated by replacing missing values with m draws from their posterior predictive distribution, resulting in m completed datasets, $D_{com}^{(i)}, i = 1, \ldots, m$. The use of multiple draws enables analysts to estimate

the additional uncertainty due to the imputation. Typically only a few imputations are needed and often $m = 5$ is sufficient (Rubin, 1987). Under regularity conditions described in Rubin (1987) and reviewed in Section 1.1.3, valid inferences can be obtained from completed datasets by using standard complete data methods on each completed dataset and applying the combining rules described in Section 1.1.1.

1.1.1 Inferences for scalar estimands

Let $q^{(i)}, i = 1, \ldots, m$, be estimates of a scalar parameter q obtained from each completed dataset $D_{com}^{(i)}$, and $u^{(i)}, i = 1, \ldots, m$, the estimates of the variance of q obtained from each $D_{com}^{(i)}$. Valid inferences for q may be obtained from $D_{com} = \{D_{com}^{(i)}, i = 1, \ldots, m\}$, using the following quantities from the completed data:

$$\bar{q} = \frac{1}{m} \sum_{i=1}^{m} q^{(i)} \tag{1.1}$$

$$\bar{u} = \frac{1}{m} \sum_{i=1}^{m} u^{(i)} \tag{1.2}$$

$$b = \frac{1}{m-1} \sum_{i=1}^{m} (q^{(i)} - \bar{q})^2 \tag{1.3}$$

where $\bar{q}$ is the average of the point estimates $q^{(i)}$, $\bar{u}$ is the mean of the variance estimates $u^{(i)}$, and b is the sample variance of $q^{(i)}$. The posterior distribution $(q|D_{com})$ used to make inferences about q is $t_{\nu_m}(\bar{q}, T_m)$ where $T_m = \bar{u} + (1 + 1/m)b$, $\nu_m = (m-1)(1 + 1/r_m)^2$, and $r_m = (1 + 1/m)b/\bar{u}$. When the sample size s is small, ν_m should be replaced with the degrees of freedom derived in Barnard and Rubin (1999).

1.1.2 Inferences for multivariate estimands

The results for scalar estimands can be generalized to multivariate estimands; however, estimation of the covariance B is poor unless m is large. A procedure for testing

the hypothesis $H_0 : Q = Q_0$, for a k-dimensional estimand Q, was derived in Rubin (1987) using the following quantities from the completed data:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} Q^{(i)} \tag{1.4}$$

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U^{(i)} \tag{1.5}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(Q^{(i)} - \bar{Q})(Q^{(i)} - \bar{Q})', \tag{1.6}$$

where terms are defined similarly to (1.1)-(1.3). A key assumption in Rubin's (1987) multicomponent test is that $B_\infty \propto \bar{U}$, or equivalently, the fractions of information missing for the k components of Q are the same. While not necessarily true in practice, it proves to be a reasonable assumption, and tests based on this assumption have been shown to be more robust than tests not based on it (Li *et al.*, 1991a).

Applying the proportionality assumption reduces the number of covariance parameters and allows for a closed-form approximation to a Bayesian p-value. When s is sufficiently large, the approximate Bayesian p-value is determined by $P(F_{k,w^*} > S_m)$, where

$$\begin{aligned} S_m &= (Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})/k(1 + r_m) \\ w_m^* &= k(m-1)(1 + 1/r_m)^2 \\ r_m &= (1 + 1/m)\mathrm{tr}(B\bar{U}^{-1})/k. \end{aligned}$$

An alternate degrees of freedom, which has been shown to have improved analytic validity for higher values of k, was developed by Li *et al.* (1991a). The degrees of freedom for this test are

$$w_m = 4 + (t-4)\{1 + (1 - t/2)/r_m\}^2$$

where $t = k(m-1)$. When $t \geq 4$ is undefined, w_m is undefined so w_m^* is used instead. The degrees of freedom derived in Reiter (2007b) should be used when the sample size s is small.

Another version of the test, not requiring the covariance matrices $U^{(i)}$, was derived by Meng and Rubin (1992). They use the asymptotic equivalence between the Wald test statistic and the log-likelihood ratio test statistic to derive a test statistic asymptotically equivalent to S_m, and a term asymptotically equivalent to r_m, so that S_m and the degrees of freedom w_m or w_m^* may be computed without access to the covariance matrices $U^{(i)}$.

1.1.3 Conditions for valid inferences

The inferential methods for multiply-imputed data are justified with Bayesian arguments; however, their analytic validity is usually considered from a frequentist viewpoint. As described in Raghunathan (2003) and Rubin (1987, Ch. 4), users of multiply-imputed data should be able to construct unbiased estimates of population quantities without access to the imputation models. Furthermore, nominal $100(1-\alpha)\%$ confidence intervals for population quantities should contain the quantity at least $100(1-\alpha)\%$ of the time. The inferential methods in Sections 1.1.1 and 1.1.2 result in randomization-valid inferences, as described in Rubin (1987) and Schafer (1997), when the following conditions hold:

- If no missing values were present, inferences about Q would be based on the posterior distribution $(Q - Q_{inc}) \sim N(0, U)$, where Q_{inc} is an estimate of a k-dimensional estimand Q obtained from the complete data D_{inc} and U is the covariance of $(Q - Q_{inc})$.

- The imputation procedure is proper, in the sense of Rubin (1987, p. 118):

– Under the posited response mechanism, with (X, Y, I) fixed and R random, as m becomes large, $(\bar{Q} - Q_{inc})$ is approximately $N(0, B)$ over the distribution of R, where I is a N-vector such that $I_l = 1, l = 1, \ldots, N$ indicates that unit l in the population of size N is selected in the sample and R is a $N \times p$ matrix such that $R_{lk} = 1, l = 1, \ldots, N; k = 1, \ldots, p$ indicates that unit l is missing its value for item k.

– Under the posited response mechanism, with (X, Y, I) fixed and R random, as m becomes large, $\bar{U}$ is a consistent estimate of U.

– With infinitely many imputations, the true between-imputation variance is stable over repeated random samples of the complete data, with variability of lower order than that of Q_{inc}.

Little and Rubin (2002, Ch. 10) describe several proper and improper procedures for generating multiple imputations. Drawing from $(Y_{mis}|Y_{obs})$ will yield proper imputations; however, doing so is often difficult. Methods that are easier to implement can still yield approximately valid inferences (Schafer, 1997; Little and Rubin, 2002; Raghunathan *et al.*, 2001). For example, sequential regression multivariate imputation (SRMI) is a popular approach for imputing missing values for complex data structures when the data are missing at random (Raghunathan *et al.*, 2001).

The best-case scenario for analytic validity occurs when the analysis procedure is *congenial* to the imputation model. Meng (1994) defines a congenial analysis procedure as one in which the procedure corresponds to the imputation model. When the analyst has different information and assumptions about the responses and nonresponses, an *uncongenial* procedure may arise. As long as the imputation model is reasonable and imputations are proper, using combined-data inferences from multiply-imputed data should limit any serious nonresponse bias (Meng, 1994). The

major danger of inconsistency, noted by Meng (1994) and Schafer (1997, Ch. 4), occurs when the imputer makes poorly grounded assumptions but the analyst does not. Hence it is recommended that the imputer refrain from imposing restrictions on unknown parameters likely to be the subject of further inquiry.

From a Bayesian perspective, analytic validity is achieved when the posterior distributions obtained from the multiply-imputed data match approximately what would be obtained using complete observed data. Schafer (1997) defines *Bayesianly proper* imputations as independent realizations of the posterior predictive distribution $p(Y_{mis}|Y_{obs})$ under some complete-data model and prior, and assumption of ignorability.

1.2 Multiple imputation for statistical disclosure control

Many national statistical agencies, survey organizations, and researchers—henceforth all called agencies—disseminate masked confidential microdata, i.e., data on individuals or establishments, in public use files. Agencies typically have the legal and ethical obligation to protect the confidentiality of the survey respondents, so the microdata must be modified to prevent confidential information from being disclosed. Simply removing all obvious identifying information is insufficient to protect the respondent identities and attributes due to the increasing sophistication of record-linking technologies and the profileration of databases available to ill-intentioned users intent on re-identifying survey respondents. Thus agencies additionally alter the data values to limit the disclosure risks using statistical disclosure control methods. The enormous amount of confidential data collected by agencies has a great deal of value to furthering scientific research and better informing public policy (Panel on Data Access for Research Purposes, 2005). Hence, good statistical disclosure control methods are

crucial to ensuring that agencies are able to share the information they hold with researchers and others agencies.

A common masking method used by agencies is data swapping, where a percentage of units have values of certain variables randomly swapped with other units (Dalenius and Reiss, 1982). This is appealing because of its simplicity; however, it distorts relationships between swapped and unswapped variables and disclosure risks may remain high when actual values are released. Several variations on this approach have been developed, including "data shuffling" (Muralidhar and Sarathy, 2006), where values are swapped in such as way that linear relationships can be preserved; however, uncertainty introduced by the shuffling is not accounted for. Examples of other methods employed by agencies to mask microdata before release include global recoding of variables, such as releasing ages in five year intervals or top-coding incomes above 100,000 as "100,000 or more" (Willenborg and de Waal, 2001), or adding random noise to continuous data values (Fuller, 1993). Another approach is microaggregation, where small groups of units with similar identifying characteristics have their confidential values replaced with the average over the group (Defays and Nanopoulos, 1992). Synthetic data methods, first proposed in Rubin (1993), use multiple imputation to limit disclosure risks, which allows for valid inferences to be made. Agencies may sometimes apply more than one of these methods to the same dataset.

Ideally the statistical disclosure control methods applied to confidential data to create public use files should allow analysts to obtain the same inferences using the altered data as they would with the unaltered data, using standard statistical methods. Many of the methods used to mask the data, however, have the unfortunate consequence of reducing the utility of the data by complicating analyses, making some analyses impossible, and severely distorting others. Some masked data may be

analyzed properly using the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs. Balancing the utility of the masked data with the risk of disclosure is a challenge in developing and applying statistical disclosure methods.

The utility of the datasets may be broadly thought of as the benefit to society of the released information, though a more feasible approach is to quantify what can be learned from the masked data relative to the observed data (Karr *et al.*, 2006). One approach to comparing the masked and observed data is to conduct several key analyses on both the observed and synthetic data and compare the results using metrics such as confidence interval overlap. It is difficult to anticipate and evaluate every analysis; hence, global measures of utility are desireable. These measures can be used by imputers to adjust the intensity of masking employed and compare different disclosure control methods. They can also be combined with risk measures to assess the tradeoff between risk and utility (Karr *et al.*, 2006). When released to users, utility measures can provide a sense of confidence in inferences obtained from the masked data. Karr *et al.* (2006) construct a framework for defining and comparing measures of utility and describe two measures for inference-specific comparisons and a global measure based on Kullback-Leibler divergences. Dobra *et al.* (2002) and Gomatam *et al.* (2005) develop inference-based utility measures for masked tabular data. Woo *et al.* (2007) propose several global measures of utility, finding the use of propensity score matching to compare the joint empirical distributions to be the most promising for general use. The other methods evaluated use cluster analysis and Kolmogorov-Smirnov-type statistics to compare the original and masked data.

Agencies typically assess the risk associated with releasing masked datasets prior

to making them available for public use. This may involve attempts to re-identify records using record-matching software, estimating the number of population uniques in the released data, or computing other measures of risk. Duncan and Lambert (1986) develop a framework for obtaining probabilities of identification for each sampled unit. Fienberg *et al.* (1997) describe a Bayesian framework for modeling identifications leading to disclosure. Reiter (2005a) describes shortcomings of using population uniques, further develops the approach of Duncan and Lambert (1986), and applies it to real data using the data swapping and noise addition masking methods. Reiter and Mitra (2007) propose a similar framework for assessing the risks associated with partially synthetic data, accounting for all the information in the released datasets and what information is released about the synthetic data generation models. Also of interest is understanding how the number of datasets released relates to disclosure risk.

Synthetic data methods are gaining increasing popularity among agencies that release large complex datasets in public-use files for their ability to provide a reasonable balance between risk and utility. Although *fully synthetic* data were originally proposed in Rubin (1993), *partially synthetic* data (Little, 1993) methods are more commonly implemented. Fully synthetic datasets are promising in their ability to protect confidentially as they contain no actual units or values. In this approach, new units are randomly and independently sampled from the sampling frame, and then the original survey data are used to impute data values for all variables on all of the sampled units. This is repeated multiple times and the set of imputed datasets is then released for public use. With appropriate synthetic data generation and the inferential methods developed by Raghunathan *et al.* (2003) and Reiter (2005c), users can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data

are described by Rubin (1993), Little (1993), Fienberg *et al.* (1998), Raghunathan *et al.* (2003), Abowd and Lane (2004), and Reiter (2002, 2005b). A drawback of the approach is the dependence of inferences on the models used, which can be difficult to specify well for complex datasets.

The partially synthetic data approach differs from fully synthetic data in that the original survey units and some actual values are released, and some values are replaced by multiple imputations. An advantage over fully synthetic data is reduced dependency on the model specification, and potentially a reduction in the complexity of the imputation models. There is, however, a tradeoff with increased disclosure risk when actual units and some of their true values are released. Partially synthetic datasets still are appealing because they maintain the primary benefits of fully synthetic data, specifically, they can protect confidentiality while allowing users to make valid inferences without learning complicated statistical methods or software.

While no agencies have released fully synthetic datasets as of this writing, several agencies have released partially synthetic public use data. Several examples are given in Reiter (2005c), including partially synthetic datasets released by the U.S. Federal Reserve Board and the U. S. Bureau of the Census. Current projects to release partially synthetic datasets at the Census Bureau include the Survey of Income and Program Participation (SIPP), described in Abowd *et al.* (2006), the "OnTheMap" public-use microdata released by the Longitudinal Employer-Household Dynamics (LEHD)program, and the Longitudinal Business Database (LBD), described in Chapter 3 of this thesis. A beta version of the SIPP synthetic data has recently been released while the LBD is still in the production stage, though a beta version is expected to completed within the next few months. In the SIPP synthetic beta release, a process for submitting discloseable analyses conducted on the synthetic data to be run on the confidential data is in place. Other current projects include the genera-

tion of partially synthetic public release files for the American Communities Survey group quarters data, underway at the U. S. Census Bureau, and the generation of fully synthetic public release datasets for establishment panel data, underway at the German Institute for Employment Research (Dreschler *et al.*, 2007).

Some methods for generating partially synthetic data include SMIKe, a general algorithm presented by Liu and Little (2002) and Little *et al.* (2004) for simulating multiple values of key identifiers for selected units. Reiter (2005d) describes the use of CART modeling to generate partially synthetic data. Other illustrations of partially synthetic data methods include Abowd and Woodcock (2004), Raghunathan (2003), and Mitra and Reiter (2006). As with multiple imputation for missing data, users can apply standard statistical methods to each imputed dataset with simple combining rules to obtain valid inferences. Combining rules for univariate estimands were developed in Reiter (2003), whose rules for combining point and variance estimates differ from those of Rubin (1987) and also from those of Raghunathan *et al.* (2003). Tests for multivariate components were derived in Reiter (2005c). These inferences are valid under regularity conditions similar to those of Rubin (1987) for multiple imputation for missing data, reviewed in Section 1.1.3, namely, that the observed-data inferences must be valid and the imputation procedure must be proper.

1.2.1 Inferences for scalar estimands

Reiter (2003) derived the combining rules used to make inferences about scalar estimands from partially synthetic data. Under regularity conditions, valid inferences about a scalar estimand q may be obtained from D_{syn}, the set of m partially synthetic datasets using the quantities in (1.1)-(1.3). The posterior distribution $(q|D_{syn})$ used to make inferences about q is $t_{\nu_p}(\bar{q}, T_p)$ where $T_p = \bar{u} + b/m$, $\nu_p = (m-1)(1+1/r_p)^2$, and $r_p = (1/m)b/\bar{u}$.

1.2.2 Inferences for multivariate estimands

Reiter (2005c) extended the methodologies of Li *et al.* (1991a) and Meng and Rubin (1992) to derive tests for multivariate estimands using partially synthetic datasets. These tests are similarly based on the assumption that $B_\infty \propto \bar{U}$. For partially synthetic datasets, an approximate Bayesian p-value for a Wald-type test of H_0 : $Q = Q_0$ for a k-variate parameter Q is determined by $P(S_p > F_{k,w_p})$ where

$$\begin{aligned} S_p &= (Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})/k(1 + r_p) \\ w_p &= 4 + (t-4)(1 + (1 - 2/t)/r_p)^2 \\ r_p &= (1/m)\mathrm{tr}(B\bar{U}^{-1})/k \end{aligned}$$

where $t = k(m-1)$, $t \geq 4$. An alternate degrees of freedom, in the form of w^* as in Section 1.1.2, for use when $t < 4$ has not been formally derived, but from similar work for two-stage imputation (Section 2.6), it can be seen to be $w_p^* = t(1 + 1/r_p)^2$. This is used in a simulation example with $m = 2$ in Section 3.6.4.

1.3 Two-stage multiple imputation

Nested or two-stage imputation refers to multiple imputation which is conducted in a nested fashion. In the first stage, m imputations are generated. In the second stage, n imputations are generated for each multiply-imputed data set generated in the first stage, resulting in a total of $M = mn$ multiply-imputed data sets.

Two-stage multiple imputation was originally developed to address computational efficiency for a missing data problem (Shen, 2000; Rubin, 2003); however, several potential uses have been noted, both within the missing data context and for other multiple imputation applications where an advantage is gained by partitioning the data to be imputed into two parts. Typically this occurs when the imputations for

each partition are generated from different posterior predictive distributions, generating two sources of variability in addition to the sampling variability. The two-stage approach to generating imputations allows the analyst to estimate properly the variability due to both types of imputations, and allows for differing numbers of imputations of each partition. Inferences for data multiply imputed in two stages use different combining rules than single-stage imputation as the imputations are not exchangeable.

Other applications where generating imputations in two stages has been found useful include synthetic data and measurement error. Reiter and Drechsler (2007) found that two-stage imputation could be used to reduce computational burden in the generation of fully or partially synthetic data. They also note that the approach can be applied to release fewer imputations for variables at high risk of disclosure, and more imputations for variables at lower risk of disclosure to improve analytic efficiency. Reiter (2007c) uses two-stage imputation to allow valid inferences to be made from data multiply-imputed for measurement error when the validation data are not made available to the analyst. Another way in which the two-stage imputation approach has been found useful is in using multiple imputation to address two issues at once, such as disclosure limitation and missing data (Reiter, 2004). Additional uses of two-stage multiple imputation for missing data as well as additional extensions are suggested in Harel and Schafer (2003) and Reiter and Raghunathan (2007). Two applications of two-stage imputation are addressed in this thesis: imputation of missing data in two stages and imputation for nonresponse and disclosure limitation.

1.3.1 Imputation of missing data in two stages

It is often the case that missing data are of different types, such as planned and unplanned nonresponse, which contribute qualitatively different types of variability. While one-stage imputation may still be used in these cases, the use of two-stage imputation can result in inferences with higher efficiency. It may also be useful in cases where imputation of one partition would be substantially easier if the other were known, and in other cases where different numbers of imputations are desired for two partitions of missing data. Nested multiple imputation may also be used to to isolate the effects of different types of missingness, evaluate different sources of variability, and measure the expected increase in information if one part were known, factors which could be useful for informing future studies (Harel and Schafer, 2003).

Nested multiple imputation was first proposed in Shen (2000), motivated in part by the multiple imputation of missing data in the National Medical Expenditure Survey. In this project, a large number of imputations were generated, with reduced computational burden, by splitting the missing data into two parts, where one part was computationally intensive and the other computationally inexpensive. First, a small number of imputations were generated for the computationally intensive portion, which included all the data except medical expenditures with missing disease codes, that initially took ten days per imputation to generate. Then, conditional on these imputated datasets, several imputations were generated for the inexpensive portion, comprising the missing disease codes and the associated expenditures, thus increasing the overall efficiency of analyses (Rubin, 2003; Shen, 2000).

Harel (2003) extends Rubin's (1976) concepts of ignorability and missing at random to nested imputation, examines the asymptotic behavior of rates of missing information, and applies the approach to longitudinal data with death and dropout. Shen (2000) develops a combining rule for univariate estimands and derives a test

for multicomponent estimands, noting that the analytic validity does not hold when the dimension of the estimands is high relative to the number of imputations. The methods of Shen (2000) are reviewed in Chapter 2 and an improved multivariate test for high-dimensional estimands is presented.

1.3.2 Two-stage imputation for nonresponse and disclosure limitation

When confidential datasets contain missing values, it is natural to use multiple imputation to address both missing data and disclosure limitation. Reiter (2004) describes a two-stage approach to handle them simultaneously. First, the agency uses multiple imputation to fill in the missing data, generating m multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with n multiple imputations, ultimately releasing mn multiply-imputed datasets. This approach is being used to create synthetic public use files for the U.S. Survey of Income and Program Participation (Abowd *et al.*, 2006).

Generating the imputations in two stages enables users to estimate properly all sources of uncertainty – the sampling variability, the variability due to imputing missing data, and the variability due to replacing sensitive values. The rules of Rubin (1987) and Reiter (2003) do not apply in this two-stage imputation scheme. Appropriate rules for scalar estimands, similar in nature to those for nested multiple imputation for missing data (Shen, 2000; Harel and Schafer, 2003; Rubin, 2003), are presented in Reiter (2004) and reviewed in Chapter 2. Also in Chapter 2, tests for multicomponent estimands are derived.

1.4 Overview

The remainder of this thesis expands upon the issues described above and presents several new inferential methods for use with multiply-imputed data.

Chapter 2 derives a Bayesian p-value for multivariate estimands using two-stage imputation where missing data and disclosure limitation are handled simultaneously. The analytic validity is illustrated by demonstrating the frequentist operating characteristics of the test. An improvement over an existing multivariate test for two-stage multiple imputation for missing data is presented. Estimates of the rates of imputed information are also given.

In Chapter 3, the generation of a partially synthetic public release file for economic census data from the U.S. Census Longitudinal Business Database is discussed. This is a large database containing longitudinal payroll and employment data over a 30-year period. When analyzing census data, parameters computed from the data are assumed to be the true values, possibly with measurement error, but without sampling error, as the data represent a population. This affects the generation and analysis of multiply-imputed census data as the existing rules were derived for survey samples. Modifications to the existing inferential methods for scalar and multivariate estimands are proposed for partially synthetic population data and extended to the missing data case.

Stepping away from multiple imputation, Chapter 4 describes the problem of selecting which predictors should be included in the fixed and random components of normal and logistic mixed effects models, accounting for model uncertainty. A fully Bayesian approach is implemented, utilizing a parameter-expanded stochastic search Gibbs sampling algorithm to estimate the exact model-averaged posterior distribution.

Chapter 5 develops a fully Bayesian model selection approach for partially syn-

thetic data. The assumption of agreement between the imputation and analyst models, described in Section 1.1.3, is not amenable to inferences about models; hence, the procedure developed incorporates the imputation model and assumes that the imputation model is known to the analyst. The development of a Bayes factor approximation is also discussed.

Chapter 2

Hypothesis testing when using multiple imputation for disclosure limitation and nonresponse

In this chapter, two-stage multiple imputation for handling nonresponse and disclosure limitation simultaneously, described in Section 1.3.2, is revisited. Users of data multiply imputed in this manner may seek to test multi-component null hypotheses, for example if several regression coefficients equal zero. Methods for performing such hypothesis tests exist when multiple imputation is used for missing data only (Rubin, 1987; Li *et al.*, 1991a,b; Meng and Rubin, 1992; Shen, 2000; Reiter, 2007b) and for synthetic data only (Reiter, 2005c). Here these tests are extended to the case when multiple imputation is used to handle missing data and disclosure limitation simultaneously. First, the two-stage imputation procedure and univariate combining rules of Reiter (2004) are reviewed, and then a Wald-type test and a log-likelihood ratio test for multicomponent estimands are proposed. Lastly, a modification to the test procedure of Shen (2000) for two-stage imputation for nonresponse only is proposed. Estimation of rates of missing information is addressed for both applications.

2.1 Review and notation

For a finite population of size N, let $I_l = 1$ if unit l is included in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, e.g. stratum or cluster indicators or size measures. The design variables X are assumed to be known approximately for the entire population, for example from census records or

the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ sub-matrix of Y for all units with $I_l = 1$, where Y_{obs} is the portion of Y_{inc} that is observed and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let R be an $N \times p$ matrix of indicators such that $R_{lk} = 1$ if the response for unit l to item k is recorded, and $R_{lk} = 0$ otherwise. The observed data is thus $D_{obs} = (X, Y_{obs}, I, R)$.

To generate the synthetic data, the agency first fills in values for Y_{mis} with draws from the conditional distribution of $(Y_{mis} \mid D_{obs})$, or approximations of that distribution such as those of Raghunathan *et al.* (2001). These draws are repeated independently $i = 1, \ldots, m$ times to obtain m completed datasets, $D_{com} = \{D_{com}^{(i)} = (D_{obs}, Y_{mis}^{(i)}), i = 1, \ldots, m\}$. Having dealt with the missing data, the agency limits disclosure risks by replacing selected values in each $D_{com}^{(i)}$ with multiple imputations. For each $D_{com}^{(i)}$, imputations are made independently $j = 1, \ldots, n$ times to yield n different partially synthetic data sets. 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 $Z = (Z_1, \ldots, Z_s)$. Let $Y_{rep}^{(i,j)}$ be all the imputed (replaced) values in the jth synthetic data set associated with $D_{com}^{(i)}$, and let $Y_{nrep}^{(i)}$ be all unchanged (unreplaced) values of $D_{com}^{(i)}$. The $Y_{rep}^{(i,j)}$ are generated from the conditional distribution of $(Y_{rep}^{(i,j)} \mid D_{com}^{(i)}, Z)$, or a close approximation of it. Each synthetic data set, $D_{syn}^{(i,j)}$, then comprises $(X, Y_{rep}^{(i,j)}, Y_{nrep}^{(i)}, I, R, Z)$. The entire collection of $M = mn$ datasets, $D_{syn} = \{D_{syn}^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, n\}$, with labels indicating the nests, is released to the public.

2.1.1 Inferences for scalar estimands

Combining rules for scalar estimands were developed by Reiter (2004). For moderate M, the following quantities are needed to make inferences for scalar parameter q:

$$\bar{q} = \sum_{i=1}^{m}\sum_{j=1}^{n} q^{(i,j)}/mn = \sum_{i=1}^{m} \bar{q}^{(i)}/m \quad (2.1)$$

$$\bar{u} = \sum_{i=1}^{m}\sum_{j=1}^{n} u^{(i,j)}/mn \quad (2.2)$$

$$\bar{w} = \sum_{i=1}^{m}\sum_{j=1}^{n} (q^{(i,j)} - \bar{q}^{(i)})^2/m(n-1) = \sum_{i=1}^{m} w^{(i)}/m \quad (2.3)$$

$$b = \sum_{i=1}^{m} (\bar{q}^{(i)} - \bar{q})^2/(m-1) \quad (2.4)$$

where $\bar{q}^{(i)}$ is the average of the point estimates in the nest of datasets indexed by i, and $\bar{q}$ is the average of the $\bar{q}^{(i)}$ across nests. The $w^{(i)}$ are the within-group variances of the point estimates in the nest of datasets indexed by i, and $\bar{w}$ is the average of the $w^{(i)}$, while b is the between-group variance of the $\bar{q}^{(i)}$ across nests and $\bar{u}$ is the average of the estimated variances $u^{(i,j)}$ of $q^{(i,j)}$ across all imputed datasets.

Using these quantities, an estimate of the variance of $\bar{q}$ is given by $T_s = (1 + 1/m)b - (1/n)\bar{w} + \bar{u}$. Note that in the case that $n = \infty$, T_s reduces to T_m, the standard combining rule for missing data of Rubin (1987) given in Section 1.1. When the sample size s is sufficiently large, inferences for q can be based on t-distributions with mean $\bar{q}$, variance T_s and degrees of freedom $\nu_s = \left\{ \frac{((1+1/m)b)^2}{(m-1)T_s^2} + \frac{(\bar{w}/n)^2}{m(n-1)T_s^2} \right\}^{-1}$.

2.1.2 Inferences for multicomponent estimands

The theory for scalar estimands can be extended directly to multivariate estimands; however, in practical applications the covariance estimation is poor. To extend the results of Reiter (2004) to multivariate estimands, let Q be a multivariate estimand, such as a vector of population means or regression coefficients. Let $Q^{(i,j)}$ be the

estimate of Q in data set $D_{syn}^{(i,j)}$, and let $U^{(i,j)}$ be the estimate of the covariance matrix associated with $Q^{(i,j)}$. The following quantities are needed for inferences.

$$\bar{Q} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} Q^{(i,j)} = \frac{1}{m}\sum_{i=1}^{m}\bar{Q}^{(i)} \tag{2.5}$$

$$\bar{U} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{(i,j)} \tag{2.6}$$

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n-1}\sum_{j=1}^{n}(Q^{(i,j)}-\bar{Q}^{(i)})(Q^{(i,j)}-\bar{Q}^{(i)})' = \frac{1}{m}\sum_{i=1}^{m} W^{(i)} \tag{2.7}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{Q}^{(i)}-\bar{Q})(\bar{Q}^{(i)}-\bar{Q})'. \tag{2.8}$$

Derivation of $(Q|D_{syn})$

To derive the posterior distribution $(Q|D_{syn})$ as in Reiter (2004) let $B_\infty = \lim B$ as $m \to \infty$ and $n \to \infty$, and let $W_\infty^{(i)} = \lim W^{(i)}$ as $n \to \infty$. Let $\bar{W}_\infty = \sum_{i=1}^{m} W_\infty^{(i)}/m$. Assuming flat prior distributions, the posterior distribution $p(Q|D_{syn})$ can be written as

$$p(Q|D_{syn}) = \iiint p(Q|D_{syn}, D_{com}, B_\infty, \bar{W}_\infty)p(D_{com}|D_{syn}, B_\infty, \bar{W}_\infty) \times p(B_\infty|D_{syn}, \bar{W}_\infty)p(\bar{W}_\infty|D_{syn})dD_{com}dB_\infty d\bar{W}_\infty. \tag{2.9}$$

The observed values are fixed and the integration is over the distributions of the missing values in D_{inc} and the values that are replaced with imputations in each $D_{com}^{(i)}$. We proceed to evaluate the integral in (2.9) by first determining the form of each distribution inside the integral.

The distribution $p(Q|D_{syn}, D_{com}, B_\infty, \bar{W}_\infty) = P(Q|D_{com}, B_\infty)$ since given D_{com}, D_{syn} provide no additional information about Q. For the first stage of imputation,

we assume the conditions are met for valid inferences under multiple imputation for missing data of (Rubin, 1987), reviewed in Section 1.1. Imputations are assumed to be drawn such that

$$(Q_{com}^{(i)}|\bar{Q}_\infty, B_\infty) \sim N(\bar{Q}_\infty, B_\infty). \tag{2.10}$$

Applying the combining rules for missing data (Rubin, 1987), we have

$$(Q|D_{com}, B_\infty) \quad \sim \quad N\{\bar{Q}_{com}, \bar{U}_{com} + (1 + 1/m)B_\infty\} \tag{2.11}$$

where $\bar{Q}_{com} = \sum Q_{com}^{(i)}/m$ and $\bar{U}_{com} = \sum U_{com}^{(i)}$, and $Q_{com}^{(i)}$ and $U_{com}^{(i)}$ are the estimates of Q and their variances that would be obtained from their corresponding $D_{com}^{(i)}$ prior to replacement of confidential values. An implicit assumption here is that the $U_{com}^{(i)}$ have sufficiently low variability so that, approximately, $\bar{U}_{com} = U_{obs}$. This is the usual assumption in multiple imputation, motivated by the fact that posterior variances generally have lower order variability than posterior means (Rubin, 1987, p. 89).

Determining $P(\bar{Q}_{com}, \bar{U}_{com}|D_{syn}, B_\infty, \bar{W}_\infty)$ is sufficient for $P(D_{com}|D_{syn}, B_\infty, \bar{W}_\infty)$, since under the assumptions list in Section 1.1.3, the posterior of Q follows a normal distribution, determined by independent mean and variance parameters. For each $D_{com}^{(i)}$, n partially synthetic datasets are imputed, for which we assume the conditions are met for valid inferences under multiple imputation for partially synthetic datasets of Reiter (2003), described in Section 1.2. The imputations are assumed to be drawn such that:

$$(Q^{(i,j)}|D_{com}^{(i)}, W_\infty^{(i)}) \quad \sim \quad N(Q_{com}^{(i)}, W_\infty^{(i)}). \tag{2.12}$$

It follows directly from (2.12) that

$$(\bar{Q}^{(i)}|D_{com}^{(i)}, W_\infty^{(i)}) \sim N(Q_{com}^{(i)}, W_\infty^{(i)}/n) \tag{2.13}$$

and thus

$$\begin{aligned}(Q_{com}^{(i)}|D_{syn}, B_\infty, W_\infty^{(i)}) &\sim N(\bar{Q}^{(i)}, W_\infty^{(i)}/n) \\ (\bar{Q}_{com}|D_{syn}, B_\infty, \bar{W}_\infty) &\sim N(\bar{Q}, \bar{W}_\infty/mn). \end{aligned} \tag{2.14}$$

Under the low-variability assumption of the posterior variance, the $U^{(i,j)}$ have sufficiently low variance so that approximately, $\bar{U} = U^{(i,j)} = \bar{U}_{com} = U_{obs}$, and $(\bar{U}_{com}|D_{syn}, B_\infty, \bar{W}_\infty)$ is taken to be a random variable with an expected value of $\bar{U}$ and variance substantially less than $\bar{W}_\infty/mn$.

To obtain the distributions $(B_\infty|D_{syn}, \bar{W}_\infty)$ and $(\bar{W}_\infty|D_{syn})$, first combine (2.10) and (2.13) to obtain the sampling distribution

$$(\bar{Q}^{(i)}|\bar{Q}_\infty, B_\infty, W_\infty^{(i)}) \sim N(\bar{Q}_\infty, B_\infty + W_\infty^{(i)}/n). \tag{2.15}$$

Then observe that B and $W^{(i)}$ are the sample covariance matrices for (2.15) and (2.12), and assume that $W_\infty^{(i)} = \bar{W}_\infty, i = 1, \ldots, m$. This assumption is reasonable given that the variability of posterior variances tends to be of smaller order than the variability of posterior means (Reiter, 2004). With diffuse prior distributions and standard multivariate normal theory we have

$$\{B(B_\infty + \bar{W}_\infty/n)^{-1}|D_{syn}, \bar{W}_\infty\} \sim Wi(m-1, I) \tag{2.16}$$

$$\{\bar{W}(\bar{W}_\infty)^{-1}|D_{syn}\} \sim Wi(n-1, I). \tag{2.17}$$

Having determined the form of each distribution in the integral in (2.9), the evaluation is completed using iterated expectations and variances. From (2.11) and

(2.14) we obtain

$$\begin{aligned}
E(Q|D_{syn}, B_\infty, \bar{W}_\infty) &= E\{E(Q|D_{com}, D_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}, \bar{W}_\infty, B_\infty\} \\
&= E(\bar{Q}_{com}|D_{syn}, B_\infty, \bar{W}_\infty) \\
&= \bar{Q} \\
V(Q|D_{syn}, B_\infty, \bar{W}_\infty) &= E\{V(Q|D_{com}, D_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}, B_\infty, \bar{W}_\infty\} \\
&\quad +V\{E(Q|D_{com}, D_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}, B_\infty, \bar{W}_\infty\} \\
&= E\{(1+1/m)B_\infty + \bar{U}|D_{syn}, B_\infty, \bar{W}_\infty\} \\
&\quad +V(\bar{Q}_{com}|D_{syn}, B_\infty, \bar{W}_\infty) \\
&= \bar{U} + (1+1/m)B_\infty + \bar{W}_\infty/mn.
\end{aligned}$$

Thus, since the distributions involved are normal,

$$(Q|D_{syn}, B_\infty, \bar{W}_\infty) \sim N(\bar{Q}, T_\infty) \tag{2.18}$$

where $T_\infty = \bar{U} + (1+1/m)B_\infty + \bar{W}_\infty/mn$. For sufficiently large s, m, and n, we can replace B_∞ and $\bar{W}_\infty$ in the expression for T_∞ with their approximate expected values in (2.16) and (2.17), $B - \bar{W}/n$ and $\bar{W}$, respectively, resulting in the variance estimate $T_s = (1+1/m)B - (1/n)\bar{W} + \bar{U}$; thus, for sufficiently large s, m, and n, inferences for Q can be based on $(Q - \bar{Q}) \sim N(0, T_s)$.

2.1.3 Hypothesis testing

Using the M released datasets, an analyst seeks to test the null hypothesis $Q = Q_0$ for some k-component estimand Q, for example to test if k regression coefficients equal zero. Given the normal approximation for inferences about Q, it may appear reasonable to use a Wald test with test statistic $(Q_0 - \bar{Q})' T_s^{-1} (Q_0 - \bar{Q})$; however, this test is unreliable when k is large and m and n are moderate, as is frequently the case in practice, because B and $\bar{W}$ can have large variability. Estimating B or $\bar{W}$ in such

cases is akin to estimating a covariance matrix using few observations compared to the number of dimensions. This is a problem for small m even when no synthetic data are generated (Rubin, 1987; Li *et al.*, 1991a,b). The instability in T_s can be avoided by making m and n large; however, that may be impractical.

In the next section, we propose two approaches to significance testing for multivariate Q. The first is a test based on the Wald test, and thus requires access to all elements of the $U^{(i,j)}$ matrices. The second is a test based on likelihood ratio statistics, which is most useful when the elements of the $U^{(i,j)}$ are not available, or when the dimension of $U^{(i,j)}$ makes working with Wald statistics too cumbersome. For both tests, the test statistic and its reference distribution are presented, followed by the derivation.

2.2 Wald test

The test statistic for the Wald test is

$$S = (Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})/k(1 + r^{(b)} - r^{(w)})$$

where

$$r^{(b)} = (1 + 1/m)\mathrm{tr}(B\bar{U}^{-1})/k \tag{2.19}$$

$$r^{(w)} = (1/n)\mathrm{tr}(\bar{W}\bar{U}^{-1})/k. \tag{2.20}$$

The reference distribution is approximated by an F-distribution with k degrees of freedom in the numerator and w_s degrees of freedom in the denominator, where

$$w_s = 4 + \left\{1 + \frac{r^{(b)}\nu_b}{\nu_b - 2} - \frac{r^{(w)}\nu_w}{\nu_w - 2}\right\}^2 / \left\{\frac{(r^{(b)}\nu_b)^2}{(\nu_b - 2)^2(\nu_b - 4)} + \frac{(r^{(w)}\nu_w)^2}{(\nu_w - 2)^2(\nu_w - 4)}\right\} \tag{2.21}$$

for $\nu_b > 4$ and $\nu_w > 4$, and $\nu_b = k(m-1)$ and $\nu_w = km(n-1)$. The approximate Bayesian p-value for testing $Q = Q_0$ is given by $P(F_{k,w_s} > S)$. When n is large,

or when $\bar{W}$ is small, S and w_s approximately equal the test statistic and degrees of freedom w_m in Section 1.1.2 from the test developed by Li *et al.* (1991b) for multiple imputation for missing data only.

When $\nu_b \leq 4$ or $\nu_w \leq 4$, w_s is not defined. This can occur for small k when $m = 2$, a choice for m that is not recommended due to the high probability of $T_s < 0$ (Reiter, 2007a). In such cases, we suggest an alternate denominator degrees of freedom,

$$w_s^* = \left\{ \frac{(r^{(b)})^2}{\nu_b(1 + r^{(b)} - r^{(w)})^2} + \frac{(r^{(w)})^2}{\nu_w(1 + r^{(b)} - r^{(w)})^2} \right\}^{-1}. \tag{2.22}$$

This expression generalizes the degrees of freedom used in the t-distribution of Reiter (2004) for inferences for scalar Q, given in Section 1.3.2.

2.2.1 Derivation

The key idea in the derivation of the test statistic S and degrees of freedom w_s is to reduce the number of unknown parameters in B_∞ and the $W_\infty^{(i)}$ by assuming (i) equal fractions of missing information on each component of Q, and (ii) equal fractions of replaced information on each component of Q. Equivalently, B_∞ and $\bar{W}_\infty$ are proportional to $\bar{U}$. Similar proportionality assumptions are used in multiple imputation for missing data only (Rubin, 1987; Li *et al.*, 1991a,b; Shen, 2000) and for synthetic data only (Reiter, 2005c). The reference F-distribution for the test statistic is derived following the moment-matching approach proposed by Li *et al.* (1991b).

Conditional on T_∞ and using (2.18), the p-value for testing $Q = Q_0$ is $P\{\chi_k^2 > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q})\}$, where χ_k^2 is a chi-squared random variable on k degrees of freedom. Since T_∞ is generally not known, we obtain the p-value by averaging over

the distributions of $(B_\infty|D_{syn}, \bar{W}_\infty)$ and $(\bar{W}_\infty|D_{syn})$ in (2.16) and (2.17):

$$\iint P\{\chi^2_k > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) | D_{syn}, B_\infty, \bar{W}_\infty\} \times$$
$$P(B_\infty|D_{syn}, \bar{W}_\infty) P(\bar{W}_\infty|D_{syn}) dB_\infty d\bar{W}_\infty.$$

This integral can be evaluated using a numerical approach such as a Monte Carlo Markov Chain method, but it is desirable to have a closed-form approximation. In order to obtain a closed-form approximation, and to reduce the number of variance parameters to be estimated, we assume that $B_\infty = r_\infty^{(b)} \bar{U}_\infty$ and $W_\infty^{(i)} = r_\infty^{(w)} \bar{U}_\infty$ for all i, where $r_\infty^{(w)}$ and $r_\infty^{(b)}$ are scalar quantities, not assumed to be equal. Assuming $\bar{U}_\infty = \bar{U}$, and averaging the $W_\infty^{(i)}$ across nests, the proportionality assumption becomes $B_\infty = r_\infty^{(b)} \bar{U}$ and $\bar{W}_\infty = r_\infty^{(w)} \bar{U}$. The covariance matrix to be estimated is now $T_\infty = \bar{U}\{1 + (1 + 1/m) r_\infty^{(b)} + r_\infty^{(w)}/mn\}$; hence, in place of $k(k+1)/2$ covariance parameters to be estimated for each of B_∞ and $\bar{W}_\infty$, only one parameter needs to be estimated for each. The corresponding p-value is

$$\iint P\left\{\chi^2_k > \frac{(Q_0 - \bar{Q})' \bar{U}^{-1} (Q_0 - \bar{Q})}{1 + (1 + 1/m) r_\infty^{(b)} + (1/mn) r_\infty^{(w)}} | D_{syn}, r_\infty^{(b)}, r_\infty^{(w)}\right\} \times$$
$$P(r_\infty^{(b)}|D_{syn}, r_\infty^{(w)}) P(r_\infty^{(w)}|D_{syn}) dr_\infty^{(b)} dr_\infty^{(w)} \quad (2.23)$$
$$= \iint P\left\{(\chi^2_k/k) \frac{1 + (1 + 1/m) r_\infty^{(b)} + (1/mn) r_\infty^{(w)}}{(1 + r^{(b)} - r^{(w)})} > S | D_{syn}, r_\infty^{(b)}, r_\infty^{(w)}\right\} \times$$
$$P(r_\infty^{(b)}|D_{syn}, r_\infty^{(w)}) P(r_\infty^{(w)}|D_{syn}) dr_\infty^{(b)} dr_\infty^{(w)}.$$

The posterior distributions $(r_\infty^{(b)}|D_{syn}, r_\infty^{(w)})$ and $(r_\infty^{(w)}|D_{syn})$ can be obtained from (2.15) and (2.12). First the proportionality assumption and a linear transformation defined by $\bar{U}^{-1/2}$ are applied so that

$$\begin{aligned}
(\bar{Q}^{(i)} \bar{U}^{-1/2} | \bar{Q}_\infty, r_\infty^{(b)}, r_\infty^{(w)}) &\sim N(\bar{Q}_\infty \bar{U}^{-1/2}, (r_\infty^{(b)} + r_\infty^{(w)}/n) I) \\
(Q^{(i,j)} \bar{U}^{-1/2} | D_{com}^{(i)}, r_\infty^{(w)}) &\sim N(Q_{com}^{(i)} \bar{U}^{-1/2}, r_\infty^{(w)} I)
\end{aligned}$$

where I is a $k \times k$ identity matrix. Applying diffuse priors and standard multivariate normal theory for sample covariance matrices yields:

$$\begin{aligned}
(m-1)\frac{\sum_{i=1}^{m}((\bar{Q}^{(i)}-\bar{Q})\bar{U}^{-1/2})((\bar{Q}^{(i)}-\bar{Q})\bar{U}^{-1/2})'}{(m-1)(r_{\infty}^{(b)}+r_{\infty}^{(w)}/n)}|D_{syn}, r_{\infty}^{(w)} &\sim Wi(m-1, I) \\
(n-1)\frac{\sum_{j=1}^{n}((Q^{(i,j)}-\bar{Q}^{(i)})\bar{U}^{-1/2})((Q^{(i,j)}-\bar{Q}^{(i)})\bar{U}^{-1/2})'}{(n-1)r_{\infty}^{(w)}}|D_{syn} &\sim Wi(n-1, I),
\end{aligned}$$

for $i = 1, \ldots, m$. Taking the trace of the left-hand side of each of the above expressions and rearranging matrices inside the trace results in:

$$\begin{aligned}
\frac{k(m-1)}{r_{\infty}^{(b)}+r_{\infty}^{(w)}/n}\mathrm{tr}(B\bar{U}^{-1})/k|D_{syn}, r_{\infty}^{(w)} &\sim \chi^2_{k(m-1)} \\
\frac{k(n-1)}{r_{\infty}^{(w)}}\mathrm{tr}(W^{(i)}\bar{U}^{-1})/k|D_{syn} &\sim \chi^2_{k(n-1)}.
\end{aligned} \tag{2.24}$$

Averaging across nests, (2.24) becomes

$$\frac{km(n-1)}{r_{\infty}^{(w)}}\mathrm{tr}(\bar{W}\bar{U}^{-1})/k|D_{syn} \sim \chi^2_{km(n-1)}.$$

Rearranging terms gives the desired posterior distributions

$$(r_{\infty}^{(b)}|r_{\infty}^{(w)}, D_{syn}) \sim \chi^{-2}_{k(m-1)}k(m-1)\mathrm{tr}(B\bar{U}^{-1})/k - r_{\infty}^{(w)}/n \tag{2.25}$$

$$(r_{\infty}^{(w)}|D_{syn}) \sim \chi^{-2}_{km(n-1)}km(n-1)\mathrm{tr}(\bar{W}\bar{U}^{-1})/k. \tag{2.26}$$

Substituting (2.25) and (2.26) into (2.23) and rearranging terms, we obtain

$$\begin{aligned}
&\iint P\left\{\chi^2_k \frac{1+(1+\frac{1}{m})r^{(b)}_\infty + \frac{1}{mn}r^{(w)}_\infty}{k(1+r^{(b)}-r^{(w)})} > S|D_{syn}, r^{(b)}_\infty, r^{(w)}_\infty\right\} \times \\
&\qquad P(r^{(b)}_\infty|D_{syn}, r^{(w)}_\infty)P(r^{(w)}_\infty|D_{syn})dr^{(b)}_\infty dr^{(w)}_\infty \\
= &\int P\left\{\chi^2_k \frac{1+(1+\frac{1}{m})\left[\chi^{-2}_{\nu_b}\nu_b \mathrm{tr}(B\bar{U}^{-1})/k - \frac{1}{n}r^{(w)}_\infty\right] + \frac{1}{mn}r^{(w)}_\infty}{k(1+r^{(b)}-r^{(w)})} > S|D_{syn}, r^{(w)}_\infty\right\} \times \\
&\qquad P(r^{(w)}_\infty|D_{syn})dr^{(w)}_\infty \\
= &\; P\left\{\chi^2_k \frac{1+(1+\frac{1}{m})\chi^{-2}_{\nu_b}\nu_b \mathrm{tr}(B\bar{U}^{-1})/k - \frac{1}{n}\chi^{-2}_{\nu_w}\nu_w \mathrm{tr}(W\bar{U}^{-1})/k}{k(1+r^{(b)}-r^{(w)})} > S|D_{syn}\right\} \\
= &\; P\left((\chi^2_k/k)\frac{1+\nu_b r^{(b)}/\chi^2_{\nu_b} - \nu_w r^{(w)}/\chi^2_{\nu_w}}{1+r^{(b)}-r^{(w)}} > S|D_{syn}\right). \qquad (2.27)
\end{aligned}$$

The random variable in (2.27) is approximated as proportional to a F-distributed random variable, F_{k,w_s}. The approximation is obtained by matching the first two moments of $\delta F_{k,w_s}$ to those of the left-hand side of the inequality in the final expression of (2.27), for a proportionality constant δ. Equivalently, we approximate the quantity $(1+\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)})$ as $\eta\chi^{-2}_{w_s}$, for some proportionality constant η.

Matching the first moment, $E(\eta\chi^{-2}_{w_s}) = \eta/(w_s-2)$ and $E(1+\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)})$ $= E(E(1+\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)})|\chi^{-2}_{\nu_w}) = 1+\nu_b r^{(b)}/(\nu_b-2) - \nu_w r^{(w)}/(\nu_w-2)$. Hence,

$$\eta\chi^{-2}_{w_s} \;\approx\; 1 + \frac{\nu_b r^{(b)}}{\nu_b - 2} - \frac{\nu_w r^{(w)}}{\nu_w - 2}. \qquad (2.28)$$

Matching the second moment, $E((\eta\chi^{-2}_{w_s})^2) = \eta^2/(w_s-2)(w_s-4)$ and $\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)})^2) = V(1+\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)}) + (E(1+\chi^{-2}_{\nu_b}\nu_b r^{(b)} - \chi^{-2}_{\nu_w}\nu_w r^{(w)}))^2$.

Applying iterated expectations and variances, $E((\eta\chi_{w_s}^{-2})^2) =$

$$E\left(\frac{2(\nu_w r^{(w)})^2}{(\nu_w-2)^2(\nu_w-4)}\right) + V\left(1+\frac{\nu_b r^{(b)}}{\nu_b-2}-\frac{\nu_w r^{(w)}}{\nu_w-2}\right) + \left(1+\frac{\nu_b r^{(b)}}{\nu_b-2}-\frac{\nu_w r^{(w)}}{\nu_w-2}\right)^2$$

$$= \frac{2(\nu_w r^{(w)})^2}{(\nu_w-2)^2(\nu_w-4)} + \frac{2(\nu_b r^{(b)})^2}{(\nu_b-2)^2(\nu_b-4)} + \left(1+\frac{\nu_b r^{(b)}}{\nu_b-2}-\frac{\nu_w r^{(w)}}{\nu_w-2}\right)^2$$

and thus, matching moments to $\eta\chi_{w_s}^{-2}$,

$$\begin{aligned}\frac{\eta^2}{(w_s-2)(w_s-4)} &\approx \frac{2(\nu_w r^{(w)})^2}{(\nu_w-2)^2(\nu_w-4)} + \frac{2(\nu_b r^{(b)})^2}{(\nu_b-2)^2(\nu_b-4)} \\ &\quad + \left(1+\frac{\nu_b r^{(b)}}{\nu_b-2}-\frac{\nu_w r^{(w)}}{\nu_w-2}\right)^2. \qquad (2.29)\end{aligned}$$

Solving (2.28) and (2.29) yields the expression in (2.21) for w_s and $\eta = (w_s - 2)(1+\nu_b r^{(b)}/(\nu_b-2)-\nu_w r^{(w)}/(\nu_w-2))$. Substituting back into (2.27), $\delta = (\eta/w_s)/(1+ r^{(b)} - r^{(w)})$. When ν_b and ν_w are sufficiently large, $\delta \approx 1$, so that the approximate p-value is $P(F_{k,w_s} > S)$.

The alternate denominator degrees of freedom w_s^* is obtained by approximating $(1+\chi_{\nu_b}^{-2}\nu_b r^{(b)} - \chi_{\nu_w}^{-2}\nu_w r^{(w)})^{-1}$ as proportional to a chi-square random variable similar to Rubin (1987). Equivalently, the quantity $A = (1+r^{(b)}-r^{(w)})/(1+\chi_{k(m-1)}^{-2} r^{(b)}\nu_b - \chi_{km(n-1)}^{-2} r^{(w)}\nu_w$ is approximated as proportional to a mean-square random variable with degrees of freedom w_s^* by matching the first two moments of A to an $MS_{w_s^*}$ random variable. First, let $\alpha^{-1} = \chi_{\nu_b}^2/\nu_b$ and $\beta^{-1} = \chi_{\nu_w}^2/\nu_w$, so $\alpha^{-1} \sim MS_{\nu_b}$ and $\beta^{-1} \sim MS_{\nu_w}$. Then expand A in α^{-1} about $E(\alpha^{-1}) = 1$:

$$A = \frac{1+r^{(b)}-r^{(w)}}{1+\alpha r^{(b)}-\beta r^{(w)}} = \frac{1+r^{(b)}-r^{(w)}}{1+r^{(b)}-\beta r^{(w)}} + \frac{1+r^{(b)}-r^{(w)}}{(1+r^{(b)}-\beta r^{(w)})^2} r^{(b)}(\alpha^{-1}-1).$$

Then $E(A|\beta) = (1 + r^{(b)} - r^{(w)})/(1 + r^{(b)} - \beta r^{(w)})$ and $V(A|\beta) = (2(1 + r^{(b)} -$

$r^{(w)})^2(r^{(b)})^2)/(\nu_b(1+r^{(b)}-\beta r^{(w)})^4)$. Note that $E(A) = E(E(A|\beta)) = 1$, and

$$V(A) = E\left\{\frac{2(1+r^{(b)}-r^{(w)})^2(r^{(b)})^2}{k(m-1)(1+r^{(b)}-\beta r^{(w)})^4}\right\} + V\left\{\frac{1+r^{(b)}-r^{(w)}}{1+r^{(b)}-\beta r^{(w)}}\right\}.$$

Expanding $(1+r^{(b)}-r^{(w)})/(1+r^{(b)}-\beta r^{(w)})$ in β^{-1} about $E(\beta^{-1}) = 1$,

$$\frac{1+r^{(b)}-r^{(w)}}{1+r^{(b)}-\beta r^{(w)}} = 1 + \frac{r^{(w)}(\beta^{-1}-1)}{1+r^{(b)}+r^{(w)}}.$$

$$V(A) = 2\left\{\frac{(r^{(b)})^2}{(1+r^{(b)}-r^{(w)})^2\nu_b} + \frac{(r^{(w)})^2}{(1+r^{(b)}-r^{(w)})^2\nu_w}\right\}.$$

Then by setting $V(A) = 2w_s^*$, where $2w_s^*$ is the variance of a $MS_{w_s^*}$ random variable, we obtain w_s^* as given in (2.22). With the approximation $A \sim MS_{w_s^*}$, (2.27) becomes $P\{(\chi_k^2/k)/(\chi_{w_s^*}^2/w_s^*) > S\} = P(F_{k,w_s^*} > S)$.

2.3 Log-likelihood ratio test

Meng and Rubin (1992) developed an alternative test for conventional multiple imputation for missing data, based on the set of log-likelihood ratio test statistics from the completed datasets. This was extended to nested multiple imputation for missing data only by Shen (2000) and to synthetic data only by Reiter (2005c). In this section, we extend this test to the case of missing and synthetic data handled simultaneously.

Following the notation in Schafer (1997), let ψ be the vector of parameters in the analyst's model. Let $\hat{\psi}_0^{(i,j)}$ and $\hat{\psi}^{(i,j)}$ be the maximum likelihood estimates of Q computed with $D_{syn}^{(i,j)}$ under the null and alternative hypotheses, respectively. Let $\bar{\psi}^{(i)} = \sum_{j=1}^n \hat{\psi}^{(i,j)}/n$; $\bar{\psi}_0^{(i)} = \sum_{j=1}^n \hat{\psi}_0^{(i,j)}/n$; $\bar{\psi} = \sum_{i=1}^m \bar{\psi}^{(i)}/m$; and, $\bar{\psi}_0 = \sum_{i=1}^m \bar{\psi}_0^{(i)}/m$. We write the log-likelihood ratio statistic evaluated at any two values a and b for any data set $D_{syn}^{(i,j)}$ as $d'(a,b|D_{syn}^{(i,j)}) = 2\log f(D_{syn}^{(i,j)}|a) - 2\log f(D_{syn}^{(i,j)}|b)$.

The test statistic is

$$\tilde{S} = \bar{L}/k(1+\tilde{r}^{(b)}-\tilde{r}^{(w)}) \tag{2.30}$$

where

$$\tilde{r}^{(b)} = \frac{m+1}{k(m-1)}(\bar{L}_m - \bar{L})$$

$$\tilde{r}^{(w)} = (\bar{l} - \bar{L}_m)/k(n-1)$$

and

$$\bar{L} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d'(\bar{\psi}_0, \bar{\psi}|D_{syn}^{(i,j)})$$

$$\bar{L}_m = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)}|D_{syn}^{(i,j)})$$

$$\bar{l} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d'(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)}|D_{syn}^{(i,j)}).$$

The reference distribution for $\tilde{S}$ is an F-distribution with k degrees of freedom in the numerator and $\tilde{w}_s$ degrees of freedom in the denominator, where $\tilde{w}_s$ is the expression in (2.21) with the terms $r^{(b)}$ and $r^{(w)}$ replaced by $\tilde{r}^{(b)}$ and $\tilde{r}^{(w)}$. When $\nu_b \leq 4$ or $\nu_w \leq 4$, the alternate denominator degrees of freedom given in (2.22) is used, substituting in $\tilde{r}^{(b)}$ and $\tilde{r}^{(w)}$ for $r^{(b)}$ and $r^{(w)}$ as above.

2.3.1 Derivation

The derivation parallels the strategy of Meng and Rubin (1992), namely (i) find a test statistic asymptotically equivalent to S based only on the Wald statistics from each synthetic data set; (ii) use the asymptotic equivalence of the Wald and log-likelihood ratio test statistics for individual datasets to define the log-likelihood ratio test statistic $\tilde{S}$; and, (iii) find a reference F distribution as in the Wald tests.

To begin, let $d(Q^{(i,j)}, U^{(i,j)}) = (Q^{(i,j)} - Q_0)' U^{(i,j)-1} (Q^{(i,j)} - Q_0)$ for all (i, j). It follows from the asymptotic equivalence of the Wald and log-likelihood ratio test statistics that each $d(Q^{(i,j)}, U^{(i,j)})$ is asymptotically equivalent to its corresponding $d'(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)} | D_{syn}^{(i,j)})$. Furthermore, because of the low-order variability in the $U^{(i,j)}$, $U^{(i,j)}$ can be interchanged with $\bar{U}$ in any of $d(Q^{(i,j)}, U^{(i,j)})$, $d(\bar{Q}^{(i)}, U^{(i,j)})$, or $d(\bar{Q}, U^{(i,j)})$.

Let $\bar{d} = \sum_{i=1}^{m} \sum_{j=1}^{n} d(Q^{(i,j)}, U^{(i,j)})/mn$; let $\bar{d}^{(i)} = \sum_{j=1}^{n} d(\bar{Q}^{(i)}, U^{(i,j)})/n$; and, let $\hat{d} = \sum_{i=1}^{m} \sum_{j=1}^{n} d(\bar{Q}, U^{(i,j)})/mn$. Then S is equivalent to:

$$S^* = \frac{\frac{\bar{d}}{k} - (n-1) r_*^{(w)} - (m-1) r_*^{(b)}/(m+1)}{1 + r_*^{(b)} - r_*^{(w)}} \tag{2.31}$$

where $r_*^{(b)} = \frac{(m+1)}{k(m-1)} \left(\sum \bar{d}^{(i)}/m - \hat{d} \right)$ and $r_*^{(w)} = \frac{1}{k(n-1)} (\bar{d} - \sum \bar{d}^{(i)}/m)$.

This is shown by assuming without loss of generality that $Q_0 = 0$ and $\bar{U}$ is a $k \times k$ identity matrix, as in Rubin (1987, p. 100). Then, $S = \bar{Q}' \bar{Q} / k(1 + r^{(b)} - r^{(w)})$ and, using a sums-of-squares decomposition,

$$\begin{aligned} \bar{d} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (Q^{(i,j)} - \bar{Q}^{(i)})' (Q^{(i,j)} - \bar{Q}^{(i)}) \\ &\quad + \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (\bar{Q}^{(i)} - \bar{Q})' (\bar{Q}^{(i)} - \bar{Q}) + \bar{Q}' \bar{Q} \qquad (2.32) \\ &= k(n-1) r^{(w)} + \frac{k(m-1)}{m+1} r^{(b)} + \bar{Q}' \bar{Q}. \qquad (2.33) \end{aligned}$$

Computing $r^{(b)}$ and $r^{(w)}$ requires access to $\bar{U}$, which we do not want these tests to depend on. Obtaining the expressions $r_*^{(b)}$ and $r_*^{(w)}$ that are equivalent to $r^{(b)}$ and $r^{(w)}$ but rely only on Wald statistics is accomplished by using sums-of-squares decompositions. Under the canonical conditions, and without loss of generality, for

$r^{(b)}$ we have

$$\begin{aligned} r^{(b)} &= \frac{(m+1)}{km(m-1)} \sum_{i=1}^{m} (\bar{Q}^{(i)} - \bar{Q})'(\bar{Q}^{(i)} - \bar{Q}) \\ &= \frac{(m+1)}{km(m-1)} \left\{ \sum_{i=1}^{m} (\bar{Q}^{(i)'} \bar{Q}^{(i)}) - m\bar{Q}'\bar{Q} \right\} \\ &\approx \frac{(m+1)}{k(m-1)} \left(\sum \bar{d}^{(i)}/m - \hat{d} \right) = r_*^{(b)} \end{aligned}$$

since $\sum \bar{d}^{(i)}/m$ is asymptotically equivalent to $\sum_{i=1}^{m} (\bar{Q}^{(i)'} \bar{Q}^{(i)})$, and $\hat{d}$ is asymptotically equivalent to $\bar{Q}'\bar{Q}$. For $r^{(w)}$, we have

$$\begin{aligned} r^{(w)} &= \frac{1}{kmn(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (Q^{(i,j)} - \bar{Q}^{(i)})'(Q^{(i,j)} - \bar{Q}^{(i)}) \\ &= \frac{1}{kmn(n-1)} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} (Q^{(i,j)'} Q^{(i,j)}) - n \sum_{i=1}^{m} (\bar{Q}^{(i)'} \bar{Q}^{(i)}) \right\} \\ &\approx \frac{1}{k(n-1)} (\bar{d} - \sum \bar{d}^{(i)}/m) = r_*^{(w)}. \end{aligned}$$

Replacing $r^{(b)}$ and $r^{(w)}$ with $r_*^{(b)}$ and $r_*^{(w)}$ in (2.33) and substituting the the expression for $\bar{d}$ in (2.33) into (2.31) yields S.

The asymptotic equivalence between $\tilde{S}$ and S^* is obtained utilizing the asymptotic equivalence between the Wald statistic and the log-likelihood ratio statistic to show that $\bar{l} \approx \bar{d}$, $\bar{L} \approx \hat{d}$, and $\bar{L}_m \approx \bar{d}_m$. The equivalence of $\bar{l}$ and $\bar{d}$ follows directly from the asymptotic equivalence of the $d(Q^{(i,j)}, U^{(i,j)})$ and their corresponding $d'(\hat{\psi}^{(i,j)}, \hat{\psi}_0^{(i,j)} | D_{syn}^{(i,j)})$. The equivalence of $\bar{L}$ and $\hat{d}$, and of $\bar{L}_m$ and $\bar{d}_m = \sum \bar{d}^{(i)}/m$ is more subtle. Although $d(Q^{(i,j)}, U^{(i,j)}) \approx d'(\hat{\psi}^{(i,j)}, \hat{\psi}_0^{(i,j)} | D_{syn}^{(i,j)})$, the equivalence does

not hold if we replace the $Q^{(i,j)}$ and $\hat{\psi}^{(i,j)}$ with averages, i.e.,

$$\begin{aligned} d(\bar{Q}^{(i)}, U^{(i,j)}) &\not\approx d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)} | D_{syn}^{(i,j)}) \\ d(\bar{Q}, U^{(i,j)}) &\not\approx d'(\bar{\psi}_0, \bar{\psi} | D_{syn}^{(i,j)}). \end{aligned}$$

Using arguments similar to those of Meng and Rubin (1992) and Shen (2000),

$$d'(\bar{\psi}_0, \bar{\psi} | D_{syn}^{(i,j)}) \approx d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}, U^{(i,j)}) \tag{2.34}$$

$$d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)} | D_{syn}^{(i,j)}) \approx d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}^{(i)}, U^{(i,j)}). \tag{2.35}$$

The first equation is obtained by viewing the log-likelihood ratio $d'(\bar{\psi}_0, \bar{\psi} | D_{syn}^{(i,j)})$ as the log ratio of two multivariate normal densities with mean $\hat{\psi}^{(i,j)}$, where the numerator is evaluated at $\bar{\psi}_0$ and the denominator at $\bar{\psi}$. Then, decomposing ψ into two orthogonal components such that $\hat{\psi}^{(i,j)} - \hat{\psi}_0^{(i,j)} = (Q^{(i,j)}, 0)'$, and using a quadratic likelihood function, (2.34) follows (Meng and Rubin, 1992). Similarly, (2.35) follows by viewing $d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)} | D_{syn}^{(i,j)})$ as the log ratio of two multivariate normal densities where the numerator is evaluated at $\bar{\psi}_0^{(i)}$ and the denominator at $\bar{\psi}^{(i)}$.

Thus, we can show $\bar{L}$ is asymptotically equivalent to $\hat{d}$:

$$\begin{aligned} \bar{L} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d'(\bar{\psi}_0, \bar{\psi} | D_{syn}^{(i,j)}) \\ &\approx \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \{ d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}, U^{(i,j)}) \} \qquad (2.36) \\ &\approx \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \{ d(Q^{(i,j)}, \bar{U}) - d(Q^{(i,j)} - \bar{Q}, \bar{U}) \} \qquad (2.37) \\ &\approx d(\bar{Q}, \bar{U}) \approx \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d(\bar{Q}, U^{(i,j)}) = \hat{d}. \qquad (2.38) \end{aligned}$$

We can get from (2.36) to (2.37) because we can replace the covariance matrix in the Wald statistic with any consistent estimator, and under the assumption of low

variability of U, $\bar{U} \approx U^{(i,j)}$, for all i, j. To get from (2.37) to (2.38), note that the quantity $\sum_{i=1}^{m}\sum_{j=1}^{n} d(Q^{(i,j)}, \bar{U})$ can be considered as a total sum of squares, to be decomposed into a sum of squares about the mean, $\bar{Q}$, and a sum of squares for the mean, hence $\sum_{i=1}^{m}\sum_{j=1}^{n} d(Q^{(i,j)}, \bar{U}) = \sum_{i=1}^{m}\sum_{j=1}^{n} d(Q^{(i,j)} - \bar{Q}) + d(\bar{Q}, \bar{U})$.

Similar reasoning shows that $\sum \bar{d}^{(i)}/m$ is asymptotically equivalent to $\bar{L}_m$:

$$\begin{aligned}
\bar{L}_m &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)} | D_{syn}^{(i,j)}) \\
&\approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \{d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}^{(i)}, U^{(i,j)})\} \\
&\approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \{d(Q^{(i,j)}, \bar{U}) - d(Q^{(i,j)} - \bar{Q}^{(i)}, \bar{U})\} \\
&\approx d(\bar{Q}^{(i)}, \bar{U}) \approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d(\bar{Q}^{(i)}, U^{(i,j)}) = \bar{d}_m.
\end{aligned}$$

Thus, we can replace $\bar{l}$ with $\bar{d}$, $\bar{L}$ with $\hat{d}$, and $\bar{L}_m$ with $\bar{d}_m$ to obtain the test statistic $\tilde{S}$ and reference F distribution.

2.4 Rates of information replaced

Estimates of the fraction of missing information about Q are useful diagnostic tools for assessing how missing data contribute to inferential uncertainty about Q (Schafer, 1997, p. 110). Rubin (1987) addressed estimation of rates of missing information for a scalar estimand q in conventional single-stage multiple imputation for missing data. Harel (2003) addressed asymptotic rates of information in two-stage imputation for missing data for scalar estimands. In this section, estimates for single-stage imputation from Rubin (1987) for a finite or infinite number of imputations are reviewed and extended to multivariate estimands, and then further extended to two-

stage imputation for nonresponse and disclosure limitation. Estimates of the rate of missing information, the rate of information replaced due to disclosure controls, and the overall rate of information imputed are given.

Let the subscript *com* denote quantities derived from D_{com} as in Rubin (1987), prior to the replacement of any data for disclosure limitation purposes. The Fisher information observed for q is defined to be $(\bar{u}_\infty + b_\infty)^{-1}$, and the total information that would be present if Y_{mis} were also observed is $\bar{u}_\infty^{-1}$; hence the rate of missing information, when $m = \infty$, is

$$\gamma_{mis} = \{\bar{u}_\infty^{-1} - (\bar{u}_\infty + b_\infty)^{-1}\}/\bar{u}_\infty^{-1} = b_\infty(\bar{u}_\infty + b_\infty)^{-1} \tag{2.39}$$

which can be estimated as $\hat{\gamma}_{mis} = b_{com}/(\bar{u}_{com} + b_{com})$. Using the posterior distribution $(q|D_{com}) \sim t_\nu(\bar{q}_{com}, T_{com} = \bar{u}_{com} + (1 + 1/m)b_{com})$, the total information about q is $(\nu+1)(\nu+3)^{-1}T_{com}^{-1}$, hence an estimate accounting for a finite number of imputations is given by

$$\hat{\gamma}_{mis} \;=\; \{\bar{u}_{com}^{-1} - (\nu_m + 1)(\nu_m + 3)^{-1}T_{com}^{-1}\}/\bar{u}_{com}^{-1} \tag{2.40}$$

where $\nu_m = (m-1)(1+1/r_m)^2$ as in Section 1.1 and $r_m = (1+1/m)b_{com}/\bar{u}_{com}$. The expression in (2.40) can also be written as

$$\hat{\gamma}_{mis} = \frac{r_m + 2/(\nu_m + 3)}{1 + r_m}. \tag{2.41}$$

For multivariate estimands, the posterior of Q generalizes to a multivariate t-distribution, where component q_l of Q has posterior $t_\nu(\bar{q}_l, T_l)$, $\bar{q}_l$ is the lth component of $\bar{Q}_{com}$ and T_l is the lth diagonal element of T_{com}. The degrees of freedom $\nu_m^{(l)}$ for the lth component are $(m-1)(1+1/r_m^{(l)})^2$, where $r_m^{(l)} = (1+1/m)b_{com}^{(l)}/\bar{u}_{com}^{(l)}$. As the degrees of freedom $\nu_m^{(l)}$ are the same for each component, we can obtain an improved

estimate of ν_m by averaging the $r_m^{(l)}$ across components, yielding

$$r_m = (1 + 1/m)/k \sum_{l=1}^{k} b_{com}^{(l)}/\bar{u}_{com}^{(l)} = (1 + 1/m)tr(B_{com}\bar{U}_{com}^{-1})/k. \tag{2.42}$$

Similarly, under the proportionality assumption of Section 2.2.1, γ_{mis} is the same across components, and hence, to estimate γ_{mis} for multivariate Q, we average the information in Q across components and use (2.41) to estimate γ_{mis}, with r_m as defined in (2.42). In the case that Q is a scalar quantity, this reduces to the expression in (2.40).

If we would like to estimate the fraction of missing information, prior to the replacement of values for disclosure limitation, we cannot compute r_m if we have D_{syn} and not D_{com}. To estimate γ_{mis} for Q, we note that when using D_{com}, we have B_{com}, an unbiased estimate of B_∞, while when using D_{syn}, we have B, an unbiased estimate of $B_\infty + \bar{W}_\infty/n$, and $\bar{W}$, an unbiased estimate of $\bar{W}_\infty$. Thus B_∞ is estimated by $B - \bar{W}/n$ and $\hat{\gamma}_{mis} = (B - \bar{W}/n)/(\bar{U} + B - \bar{W}/n)$. To estimate γ_{mis} taking into account the finite number of imputations from D_{syn}, we use (2.41), replacing r_m with $(1 + 1/m)tr((B - \bar{W}/n)\bar{U}^{-1})$. For scalar q we also use (2.41), and r_m reduces to $(1 + 1/m)(b - \bar{w}/n)/\bar{u}$.

With an infinite number of imputations, the fraction of information replaced in the second stage goes to zero; thus, the total fraction of information for Q replaced in both stages is the same as for missing data, i.e., $\gamma_{tot} = \gamma_{mis}$. Hence, $\hat{\gamma}_{tot} = \hat{\gamma}_{mis}$. To obtain an estimate that takes into account the finite number of imputations, we can directly extend the missing data case in (2.40), replacing T_{com} with T_s and substituting the appropriate degrees of freedom, as the posterior distribution $(Q|D_{syn})$ follows a similar t-distribution to that of $(Q|D_{com})$. This yields an estimated total

fraction of missing information for Q, in the form of (2.41):

$$\hat{\gamma}_{tot} = \frac{r^{(b)} - r^{(w)} + 2/(\nu_s + 3)}{1 + r^{(b)} - r^{(w)}} \tag{2.43}$$

where $r^{(b)}$ and $r^{(w)}$ are as defined in (2.19) and (2.20) and

$$\nu_s = \left\{ \frac{(r^{(b)})^2}{(m-1)(1 + r^{(b)} - r^{(w)})^2} + \frac{(r^{(w)})^2}{m(n-1)(1 + r^{(b)} - r^{(w)})^2} \right\}^{-1},$$

which reduces to the degrees of freedom for the posterior of scalar estimand q from Reiter (2004) when Q is a scalar quantity.

As the denominators of all the rates of missing information considered here are the same, and equal to the total information about Q in the posterior distribution had all the data been observed, $\bar{U}_\infty^{-1}$ or $\bar{U}^{-1}$, estimation of the fraction of information replaced in the second stage, γ_{syn}, can be accomplished by subtraction: $\gamma_{syn} = \gamma_{tot} - \gamma_{mis}$.

Estimation of fractions of missing information tends to be unstable, except for large values of m and n. Typically several imputations will be required to obtain useful estimates. This is because the estimates depend on between-imputation variance components which are estimated with $m - 1$ and $m(n - 1)$ degrees of freedom. When m is small, estimates of fractions of missing information should be used as a rough guide only (Schafer, 1997, p. 200).

2.5 Simulation Studies

In this section, the performance of the Wald test for multicomponent estimands is evaluated using simulations. Since the likelihood ratio test is asymptotically equivalent to the Wald test, for large samples it should have similar performance.

For sample size $s = 1000$, the complete data $\{Y_0, Y_1, \ldots, Y_{20}\}$ are simulated from independent normal distributions with $E(Y_i) = 0$ for all i, $Var(Y_0) = 1$, and

$Var(Y_i) = 2$ for $i > 0$. To simulate missing data, for computational simplicity 30% of the observations have their values of $\{Y_1, \ldots, Y_{20}\}$ missing completely at random and Y_0 is always fully observed. The set of completed datasets, D_{com} is obtained by drawing values of the missing data from $f(Y_1, \ldots, Y_{20}|D_{obs})$, using a multivariate normal distribution with an unrestricted covariance matrix. To simulate partial synthesis in the second stage, all values of Y_0 are replaced. The replacement imputations for each $D_{syn}^{(i,j)}$ are drawn independently from $f(Y_0|D_{com}^{(i)})$. The number of imputations is varied with $m \in (4, 8)$ and $n \in (2, 4, 8)$. By design, this simulation satisfies both proportionality assumptions.

The hypothesis tested is $H_0 : Q = 0$, where Q is the vector of coefficients for the regression of Y_0 on $Y_1, \ldots, Y_k$, excluding the intercept, for $k \in (5, 10, 20)$. As this null hypothesis is true, the expected nominal rejection of H_0 is $100\alpha\%$, for a given significance level α. Table 2.1 summarizes the simulated nominal significance levels of the Wald test using 10000 runs of the simulation for each combination of m, n, and k, for $\alpha \in (.01, .05, .10)$. The simulated significance levels are close the desired significance levels. The rates are low when $n = 2$, suggesting the tests may be conservative in these cases. The conventional Wald test, conducted by refering the test statistic $(\bar{Q} - Q_0)'T_s(\bar{Q} - Q_0)$ to a chi-square distribution on k degrees of freedom, requires a much larger number of imputations to yield correct levels. As shown in Table 2.4, this test has dramatically high rejection rates for the realistic values of m and n used in the simulation.

The test results shown in Table 2.1 were obtained using the denominator degrees of freedom w_s. Simulations run using the alternate degrees of freedom w_s^* produced similar results. The observed rejection rates are also given for reference purposes in Table 2.3. Using w_s over the alternate degrees of freedom w_s^* is recommended except when w_s is undefined. While this simulation produced satisfactory results

Table 2.1: Nominal rejection rates for given significance level α using Wald-type test with denominator degrees of freedom w_s

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	0.1	0.2	0.7	1.9	2.5	3.8	5.5	6.7	9.3
	$n = 4$	1.0	1.0	1.3	5.0	5.5	5.6	10.1	10.2	10.7
	$n = 8$	1.4	1.1	1.3	5.6	5.4	5.6	11.1	10.9	11.1
$m = 8$	$n = 2$	0.3	0.5	0.9	3.2	4.2	5.1	7.7	8.9	10.7
	$n = 4$	1.4	1.2	1.2	5.2	5.7	5.8	10.5	10.7	11.4

for both, previous results from Li *et al.* (1991a) and Shen (2000) indicate that tests using degrees of freedom analogous to w_s^*, for conventional missing data imputation, perform poorly in general when $k > 5$. In the context of two-stage imputation for missing data and disclosure limitation, further research is needed to better assess the relative performance of w_s and w_s^*.

Li *et al.* (1991a) show for multiple imputation for missing data only, that Wald tests based on the proportionality assumption are robust in cases of practical interest even when the proportionality assumption fails. To evaluate the robustness of the test to violations of the proportionality assumptions in the context of imputation for missing data and disclosure control, a simulation in which the proportionality assumption is not met for the synthetic replacement data is performed. In this second simulation, $Y_0, \ldots, Y_{10}$ are replaced in entirety and $Y_{11}, \ldots, Y_{20}$ are left intact. The imputations are generated from D_{com} by taking draws from $(Y_{10}|Y_{11}, \ldots, Y_{20})$, $(Y_9|Y_{10}, \ldots, Y_{20}), \ldots, (Y_0|Y_1, \ldots, Y_{20})$. The test $H_0 : Q = 0$ is carried out as above, with $k = 20$. Table 2.5 gives the nominal rejection rates for this scenario, which are seen to be similar to those in Table 2.1.

2.6 Two-stage imputation for nonresponse

As described in Section 1.3.1, two-stage imputation was originally developed to address computational efficiency for a missing data problem in an unpublished thesis

Table 2.2: Nominal rejection rates for given significance level α using Wald-type test with denominator degrees of freedom w_s^*

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	0.1	0.2	0.6	1.5	2.8	4.3	4.7	7.3	9.3
	$n = 4$	1.0	1.2	1.3	4.3	5.5	5.6	9.1	10.6	11.1
	$n = 8$	1.2	1.3	1.0	5.1	5.5	5.7	10.0	10.6	10.6
$m = 8$	$n = 2$	0.4	0.7	1.1	3.3	4.4	5.6	7.2	9.1	10.5
	$n = 4$	1.0	1.3	1.4	5.2	5.5	6.1	10.5	10.8	11.4

Table 2.3: Nominal rejection rates for given significance level α using standard Wald test on observed data

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	1.1	0.9	1.2	5.2	4.9	5.5	10.3	10.0	10.4
	$n = 4$	1.1	0.9	1.2	5.2	4.9	5.5	10.3	10.0	10.4
	$n = 8$	1.0	1.1	1.1	5.0	5.4	5.0	9.7	10.3	9.8
$m = 8$	$n = 2$	1.1	0.9	1.2	5.2	4.9	5.5	10.3	10.0	10.4
	$n = 4$	1.1	0.9	1.2	5.2	4.9	5.5	10.3	10.0	10.4

Table 2.4: Nominal rejection rates for given significance level α using standard Wald test on imputed data using covariance matrix T

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	9.8	11.6	22.4	15.6	21.5	39.7	20.1	28.5	50.0
	$n = 4$	26.0	32.6	33.5	36.4	41.2	40.4	43.0	45.9	43.8
	$n = 8$	7.9	21.5	54.1	19.2	38.1	71.7	27.9	48.3	79.4
$m = 8$	$n = 2$	12.4	11.0	10.2	17.9	15.5	17.0	22.0	19.2	23.5
	$n = 4$	11.7	36.1	40.2	22.3	48.3	46.9	30.2	55.7	50.6

Table 2.5: Nominal rejection rates for $k = 20$ and given significance level α using Wald-type test, where proportionality assumption not met

		$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
$m = 4$	$n = 2$	0.6	4.2	9.2
	$n = 4$	1.2	5.3	10.7
	$n = 8$	1.0	5.4	10.5
$m = 8$	$n = 2$	0.9	5.1	10.3
	$n = 4$	1.2	5.6	10.9

(Shen, 2000). In this section the existing methodology for two-stage multiple imputation for nonresponse is reviewed, and a modified Wald-type test for multivariate estimands is proposed that demonstrates improved analytic validity for high-dimensional estimands. The work of Harel and Schafer (2003) on rates of missing information is also reviewed and extended. Thus this section serves to extend the existing methodology for this application to that developed earlier in this chapter for the application of nonresponse and disclosure limitation. The structure and derivations of the proposed tests are similar to those presented in previous sections, and thus are presented in less detail.

2.6.1 Imputing missing values in two stages

We use the same notation as in Section 2.1, but we now let $R = (R^{(A)}, R^{(B)})$, where $R^{(A)}$ is an $N \times p$ matrix of indicators such that $R^{(A)}_{lk} = 1$ if the response for unit l to item k is missing and to be imputed in the first stage and $R^{(A)}_{lk} = 0$ otherwise, and $R^{(B)}$ be the corresponding $N \times p$ matrix of indicators for the second stage of imputation and partition Y_{mis} into $Y^{(A)}_{mis}$ and $Y^{(B)}_{mis}$.

To generate the imputations, first $Y^{(A)}_{mis}$ is replaced with m draws from the posterior distribution $(Y^{(A)}_{mis}|D_{obs})$, resulting in m partially completed datasets, $D_{pcom} = \{D^{(i)}_{pcom}, i = 1, \ldots, m\}$, where $D^{(i)}_{pcom}$ is comprised of D_{obs}, $Y^{(B)}_{mis}$, and the ith imputation of $Y^{(A)}_{mis}$. Then for each $D^{(i)}_{pcom}$, $Y^{(B)}_{mis}$ is replaced with n draws from the posterior predictive distribution $(Y^{(B)}_{mis}|D^{(i)}_{pcom}, R^{(B)})$, resulting in a total of $M = mn$ imputed datasets $D_{com} = \{D^{(i,j)}_{com}, i = 1, \ldots, m; j = 1, \ldots, n\}$, where $D^{(i,j)}_{com}$ is comprised of D_{obs}, the ith imputation of $Y^{(A)}_{mis}$ and the jth imputation of $Y^{(B)}_{mis}$.

Shen (2000) describes a second, equivalent method of generating imputations in two stages. In this procedure, m imputations of $Y^{(A)}_{mis}$ and $Y^{(B)}_{mis}$ are drawn in the first stage from the joint posterior distribution $(Y^{(A)}_{mis}, Y^{(B)}_{mis}|D_{obs})$. In the second

stage, an additional $n-1$ conditionally independent imputations are drawn from $(Y_{mis}^{(B)}|D_{pcom}^{(i)}, R^{(B)})$. This approach is advantagous when it is easier to specify or draw from $(Y_{mis}^{(A)}, Y_{mis}^{(B)}|D_{obs})$ than from $(Y_{mis}^{(A)}|D_{obs})$.

When nested imputation is used for the purpose of improving computational efficiency, the computationally intensive portion is naturally chosen to be imputed first so as to minimize the number of imputations, with $m < n$. Absent computational concerns, for randomization validity it makes sense to impute the portion with a greater proportion of missing values first, with $m > n$. Harel (2003) suggests setting $n = 2$ and then choosing m to obtain the desired precision, unless the rate of missing information in the first stage is thought to be much smaller than in the second. In the similar setting of two-stage imputation for missing data and disclosure limitation, Reiter (2007a) found improved efficiency in inferences when $m > n$, particularly for large fractions of missing data in the first stage.

2.6.2 Existing inferential methods

The combining rules for multiply imputed data from Rubin (1987) do not apply to data imputed in two stages as the imputations are not exchangeable. Appropriate combining rules were first derived in Shen (2000).

Inferences for scalar parameters

Inferences for a scalar parameter q are made using the quantities given in (2.1) through (2.4). The estimate of q is given by $\bar{q}$ and the variance of $\bar{q}$ by $T_n = (1 + 1/m)b + (1 - 1/n)\bar{w} + \bar{u}$. Note that as $\bar{w}$, the between-imputation variance, goes to zero, or n becomes large, T_n reduces to T_m, the standard combining rule for missing data (Rubin, 1987) given in Section 1.1. When the sample size s is sufficiently large, inferences for q can be based on t-distributions with mean $\bar{q}$, variance T_n and

degrees of freedom $\nu_n = \left\{ \frac{((1+1/m)b)^2}{(m-1)T_n^2} + \frac{((1-1/n)\bar{w})^2}{m(n-1)T_n^2} \right\}^{-1}$.

Inferences for multivariate parameters

Generalizing from the univariate case, let Q be a multicomponent estimand, such as a vector of regression coefficients. The quantities given in (2.5) through (2.8) are used to make inferences about Q, with the expected value given by $\bar{Q}$. An unbiased estimate of the variance of $\bar{Q}$ is given by $T_n = (1 + 1/m)B + (1 - 1/n)W + \bar{U}$.

When testing $H_0 : Q = Q_0$, for multivariate parameter Q, it may seem reasonable to use a Wald test with test statistic $(Q_0 - \bar{Q})T_n^{-1}(Q_0 - \bar{Q})$ when the sample size s is sufficiently large; however, T_n is a poor estimate of the variance except when m and n are excessively large. Hence Wald tests based on this covariance estimate perform poorly in cases of practical interest and a modification is needed. When the covariances matrices $U^{(i,j)}$ are available, Wald-type tests test statistics may be used to test $Q = Q_0$. The test statistic is

$$S_n = (Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})/k(1 + r_n^{(b)} + r_n^{(w)}) \tag{2.44}$$

where

$$r_n^{(b)} = (1 + 1/m)\mathrm{tr}(B\bar{U}^{-1})/k \tag{2.45}$$

$$r_n^{(w)} = (1 - 1/n)\mathrm{tr}(\bar{W}\bar{U}^{-1})/k. \tag{2.46}$$

Shen (2000) proposes an approximate Bayesian p-value extending the approach of Rubin (1987). This is obtained by referring S_n to an F_{k,w_n^*} distribution, where

$$w_n^* = \left\{ \frac{(r_n^{(b)})^2}{\nu_b(1 + r_n^{(b)} + r_n^{(w)})^2} + \frac{(r_n^{(w)})^2}{\nu_w(1 + r_n^{(b)} + r_n^{(w)})^2} \right\}^{-1} \tag{2.47}$$

and $\nu_b = k(m - 1)$ and $\nu_w = km(n - 1)$.

Shen (2000) also derived a log-likelihood ratio test similar to that of Section 2.3. In this test an approximate Bayesian p-value is found by referring $\tilde{S}_n$ to an $F_{k,\tilde{w}_n^*}$-distribution, where

$$\begin{aligned}
\tilde{S}_n &= \bar{L}/(k(1+\tilde{r}^{(b)}+\tilde{r}^{(w)})) \\
\tilde{r}_n^{(b)} &= \frac{m+1}{k(m-1)}(\bar{L}_m - \bar{L}) \\
\tilde{r}_n^{(w)} &= (\bar{l} - \bar{L}_m)/k(n-1)
\end{aligned}$$

and

$$\begin{aligned}
\bar{L} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(2\log f(D_{com}^{(i,j)}|\bar{\psi}) - 2\log f(D_{com}^{(i,j)}|\bar{\psi}_0)) \\
\bar{L}_m &= \frac{1}{m}\sum_{j=1}^{n}(2\log f(D_{com}^{(i,j)}|\bar{\psi}^{(i)}) - 2\log f(D_{com}^{(i,j)}|\bar{\psi}_0^{(i)})) \\
\bar{l} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 2\log f(D_{com}^{(i,j)}|\hat{\psi}^{(i,j)}) - 2\log f(D_{com}^{(i,j)}|\hat{\psi}_0^{(i,j)})
\end{aligned}$$

where $\hat{\psi}_0^{(i,j)}$ and $\hat{\psi}^{(i,j)}$ are the maximum likelihood estimates of Q under the null and alternative hypotheses, respectively; $\bar{\psi}^{(i)} = \sum_{j=1}^{n}\hat{\psi}^{(i,j)}/n$, $\bar{\psi}_0^{(i)} = \sum_{j=1}^{n}\hat{\psi}_0^{(i,j)}/n$, $\bar{\psi} = \sum_{i=1}^{m}\bar{\psi}^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^{m}\bar{\psi}_0^{(i)}/m$. The denominator degrees of freedom $\tilde{w}_n^*$ is defined as w_n^* in (2.47) with the terms $r_n^{(b)}$ and $r_n^{(w)}$ replaced by $\tilde{r}_n^{(b)}$ and $\tilde{r}_n^{(w)}$.

2.6.3 Proposed method

Shen (2000) found that the Wald-type test above exhibited poor frequentist properties when k was large relative to m, and the corresponding test for single-stage multiple imputation is known to have the same problem. Li *et al.* (1991a) proposed

an alternate denominator degrees of freedom to that of Rubin (1987) for conventional missing data which has been shown to provide better analytic validity, in the sense of Chapter 1, and has entered into wide use for testing multicomponent estimands. We extend this approach to two-stage multiple imputation and use a new denominator degrees of freedom given by:

$$w_n = 4 + \left\{1 + \frac{r_n^{(b)}\nu_b}{\nu_b - 2} + \frac{r_n^{(w)}\nu_w}{\nu_w - 2}\right\}^2 \Big/ \left\{\frac{(r_n^{(b)}\nu_b)^2}{(\nu_b-2)^2(\nu_b-4)} + \frac{(r_n^{(w)}\nu_w)^2}{(\nu_w-2)^2(\nu_w-4)}\right\} \quad (2.48)$$

Note that when n is large or $\bar{W}$ is small then S_n and w_n approximately equal the test statistic S_m and degrees of freedom w_m for missing data imputed in one stage (Li *et al.*, 1991a) that are given in Section 1.1.2.

When $\nu_b \leq 4$ or $\nu_w \leq 4$ then w_n is not defined; however, this only occurs for a few cases with $m = 2$ and k small. If a user is faced with a situation where w_n is undefined, then w_n^* can still be used.

Similarly, when ν_b and ν_w are defined, the likelihood ratio test of Shen (2000), described in Section 2.6.2, can be modified by replacing $\tilde{w}_n^*$ with $\tilde{w}_n$, defined as w_n in (2.48) with the terms $r_n^{(b)}$ and $r_n^{(w)}$ replaced by $\tilde{r}^{(b)}$ and $\tilde{r}^{(w)}$.

Derivation

The derivation given here for the test statistic S_n is similar to that presented in Shen (2000); however, the derivation of the reference distribution is substantially different. Most notably we do not ignore the lack of independence between the variance parameters corresponding the between- and within-nest variances. Details that are the same as those in Section 2.2.1 are omitted.

Let $B_\infty = \lim B$ as $m \to \infty$ and $n \to \infty$ and let $\bar{W}_\infty = \sum W_\infty^{(i)}/m$ where $W_\infty^{(i)} = \lim W^{(i)}$ as $n \to \infty$. Assuming the conditions for valid inferences un-

der multiple imputation (Rubin, 1987; Harel, 2003), the posterior distribution of $(Q|D_{com}, B_\infty, \bar{W}_\infty)$ is $N(\bar{Q}, T_\infty)$, where $T_\infty = \bar{U}_\infty + (1 + 1/m)B_\infty + (1 + 1/mn)\bar{W}_\infty$.

If T_∞ were known, then the Bayesian p-value$(Q_0|D_{com}, T_\infty)$ for testing $H_0 : Q = Q_0$ would be $P(\chi^2_k > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q}))$. Since T_∞ is generally not known, the p-value is obtained by integrating over the conditional distributions of the variance parameters $(B_\infty|D_{syn}, \bar{W}_\infty)$ and $(\bar{W}_\infty|D_{syn})$:

$$\iint P\{\chi^2_k > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q})|D_{com}, B_\infty, \bar{W}_\infty\} \times$$
$$P(B_\infty|D_{com}, \bar{W}_\infty)P(\bar{W}_\infty|D_{com})dB_\infty d\bar{W}_\infty. \tag{2.49}$$

In order to obtain a closed-form approximation, and reduce the number of variance parameters to be estimated, we assume equal fractions of missing information *in each stage* contribute to each component of Q, i.e., we assume the between-nest variance B_∞ and within-nest variance $\bar{W}_\infty$ are both proportional to the total variance and hence to $\bar{U}_\infty$:

$$B_\infty = r_\infty^{(b)}\bar{U}_\infty, \bar{W}_\infty = r_\infty^{(w)}\bar{U}_\infty \tag{2.50}$$

for scalar quantities $r_\infty^{(w)}$ and $r_\infty^{(b)}$, not assumed to be equal. Under (2.50), (2.49) reduces to

$$\iint P\left\{\chi^2_k > \frac{(Q_0 - \bar{Q})'U_\infty^{-1}(Q_0 - \bar{Q})}{1 + (1 + \frac{1}{m})r_\infty^{(b)} + (1 + \frac{1}{mn})r_\infty^{(w)}}|D_{com}\right\}$$
$$P(r_\infty^{(b)}|D_{com}, r_\infty^{(w)})P(r_\infty^{(w)}|D_{com})dr_\infty^{(b)}dr_\infty^{(w)}. \tag{2.51}$$

Under asymptotic theory for the sampling distribution of the posterior variance, which tends to have lower-order posterior variance than the mean, $\bar{U}_\infty$ can be replaced with $\bar{U}$ (Rubin, 1987, p.89). Generalizing from the theory for univariate estimands, assuming (2.50), and standard multivariate normal theory, and averaging across nests,

the conditional distributions of $r_\infty^{(b)}$ and $r_\infty^{(w)}$ are:

$$\begin{aligned}(r_\infty^{(b)}|r_\infty^{(w)}, D_{com}) &\sim \chi_{\nu_b}^{-2}\nu_b(\mathrm{tr}(B\bar{U}^{-1})/k) - r_\infty^{(w)}/n \\ (r_\infty^{(w)}|D_{com}) &\sim \chi_{\nu_w}^{-2}\nu_w(\mathrm{tr}(\bar{W}\bar{U}^{-1})/k)\end{aligned}$$

Using the above and (2.51), and substituting in (2.44), (2.45) and (2.46) gives

$$P\left\{(\chi_k^2/k)\frac{(1+\chi_{\nu_b}^{-2}\nu_b r_n^{(b)} + \chi_{\nu_w}^{-2}\nu_w r_n^{(w)})}{(1+r_n^{(b)}+r_n^{(w)})} > S_n\right\}. \tag{2.52}$$

The left-hand side of the inequality in (2.52) can be approximated as proportional to an F_{k,w_n} distribution by matching the first two moments of each, so that the approximate p-value is $P(\delta F_{k,w_n} > S_n)$, for a proportionality constant δ. Equivalently, the quantity $(1+\chi_{\nu_b}^{-2}\nu_b r_n^{(b)} + \chi_{\nu_w}^{-2}\nu_w r_n^{(w)})$ is approximately proportional to an inverse chi-square distributed random variable with degrees of freedom w_n by matching the first two moments of $\eta\chi_w^{-2}$, for proportionality constant η:

$$\begin{aligned}E(\eta\chi_w^{-2}) &= \eta/(w_n-2) \\ &\approx 1+\nu_b r_n^{(b)}/(\nu_b-2) + \nu_w r_n^{(w)}/(\nu_w-2) \\ E\{(\eta\chi_w^{-2})^2\} &= \eta^2/(w_n-2)(w_n-4) \\ &\approx \frac{2(\nu_w r_n^{(w)})^2}{(\nu_b-2)^2(\nu_w-4)} + \frac{2(\nu_b r_n^{(b)})^2}{(\nu_b-2)^2(\nu_w-4)} + \left(1+\frac{\nu_b r_n^{(b)}}{\nu_b-2}+\frac{\nu_w r_n^{(w)}}{\nu_w-2}\right)^2\end{aligned}$$

Solving these expressions gives the expression for w_n in (2.48) and $\eta = (w_n-2)(1+\nu_b r_n^{(b)}/(\nu_b-2)+\nu_w r_n^{(w)}/(\nu_w-2))$. Substituting into (2.52), $\delta = (\eta/w_n)/(1+r_n^{(b)}+r_n^{(w)})$. For sufficiently large ν_b and ν_w, $\delta \approx 1$ S_n is referred to the F_{k,w_n} distribution.

When $\nu_b \le 4$ or $\nu_w \le 4$ and w_n is undefined, the denominator degrees of freedom w_n^* of (2.47) can be used. The derivation of w_n^* in Shen (2000) uses Satterthwaite's approximation, assuming independence between $r_\infty^{(b)}$ and $r_\infty^{(w)}$, or rather, ignoring the

lack of independence. The degrees of freedom w_n^* can also be obtained following a procedure similar to the derivation of w_s^* in Section 2.2.1, without making any assumptions about independence. Following that derivation procedure, w_n^* is obtained by approximating $(1+\chi_{\nu_b}^{-2}\nu_b r_n^{(b)}+\chi_{\nu_w}^{-2}\nu_w r_n^{(w)})^{-1}$ as proportional to a chi-square random variable as in Rubin (1987).

2.6.4 Rates of missing information

Harel (2003) considers the population rates of missing information for scalar estimands, but not multivariate estimands or estimates of the rates accounting for the finite number of imputations. These rates for both scalar and multivariate estimands can be easily derived in a manner similar to that of Section 2.4. In this section, the rates of missing information for multivariate estimands are derived, which give the rates for scalar estimands when the dimension $k = 1$.

The estimate of the fraction of missing information due to $Y_{mis}^{(A)}$ is same as γ_{mis} in Section 2.4 since in both applications missing data is imputed in the first stage, and in both cases B_∞ is estimated by $B - \bar{W}/n$. Thus, $\gamma_{mis}^{(A)} = B_\infty(\bar{U}_\infty + B_\infty)^{-1}$ and $\hat{\gamma}_{mis}^{(A)} = (B - \bar{W}/n)(\bar{U} + B - \bar{W}/n)^{-1}$. The finite-imputation estimate is given by 2.41 with $r_m = (1 + 1/m)tr((B - \bar{W}/n)\bar{U}^{-1})$.

The total fraction of missing information for Q due to both $Y_{mis}^{(A)}$ and $Y_{mis}^{(B)}$ is determined similar to (2.40) as $\gamma_{tot} = (B_\infty + \bar{W}_\infty)(\bar{U}_\infty + B_\infty + \bar{W}_\infty)^{-1}$. Since B_∞ is estimated by $B - \bar{W}/n$, an estimate of this fraction is given by $\hat{\gamma}_{tot} = (B + (1 - 1/n)\bar{W})/(\bar{U} + B + (1 - 1/n)\bar{W})$.

The assumption of equal fractions of missing information across components in each stage of imputation, also leads to equal fractions of total missing information. To estimate γ_{tot} accounting for the finite number of imputations, T_{com} in (2.40) is replaced with $T_n = \bar{U} + (1 + 1/m)B + (1 - 1/n)\bar{W}$ and the appropriate degrees of

freedom,

$$\nu_n = \left\{ \frac{(r_n^{(b)})^2}{(1+r_n^{(b)}+r_n^{(w)})^2} + \frac{(r_n^{(w)})^2}{(1+r_n^{(b)}+r_n^{(w)})^2} \right\}.$$

yielding the estimate

$$\hat{\gamma}_{tot} = \frac{2/(\nu_n+3) + r_n^{(b)} + r_n^{(w)}}{1+r_n^{(b)}+r_n^{(w)}}$$

where $r_n^{(b)}$ and $r_n^{(w)}$ are as defined in (2.45) and (2.46), and ν_n reduces to the degrees of freedom for univariate q as in Shen (2000) when Q is a scalar quantity.

In this application, missing data is imputed in the second stage, so unlike the partially synthetic data case, the fraction of missing information does not go to zero as n goes to infinity. An estimate of the fraction of missing information due to $Y_{mis}^{(B)}$ if $Y_{mis}^{(A)}$ were known, assuming an infinite number of imputations, is given by $\gamma_{mis}^{(B)} = \bar{W}_\infty/(\bar{U}_\infty + \bar{W}_\infty)$. Since $\bar{W}_\infty$ is estimated by $\bar{W}$, an estimate of this fraction is given by $\hat{\gamma}_{mis}^{(B)} = \bar{W}/(\bar{U}+\bar{W})$. As in Section 2.4, a finite-imputation estimate of the fraction of missing information due to $Y_{mis}^{(B)}$ if $Y_{mis}^{(A)}$ were to be observed can be obtained by subtraction: $\hat{\gamma}_{mis}^{(B)} = \hat{\gamma}_{tot} - \hat{\gamma}_{mis}^{(A)}$.

2.6.5 Simulation Studies

In this section the improvement in the frequentist performance of the Wald test when using w_n instead of w_n^* is demonstrated with a few simulations. Extensive simulations in Shen (2000) demonstrate the performance of Wald tests using w_n^* for $k \leq 5$, for varying values of m, n, and rates of missing information. Additionally, the ability of two-stage imputation to improve upon single-stage imputation is demonstrated and the power of the test is evaluated. The robustness of tests to violations of the proportionality assumptions used here has been demonstrated for single-stage

multiple imputation for missing data in one stage by Li *et al.* (1991a) and for two-stage multiple imputation by Shen (2000). Similar robustness is expected for the alternate reference distribution proposed.

For a sample size $s = 1000$, the complete data $\{Y_0, \ldots, Y_{20}\}$ are simulated from independent normal distributions with $E(Y_i) = 0$ for all i, $V(Y_0) = 1$ and $V(Y_i) = 2$ for $i > 0$. For computational simplicity missingness is simulated by letting $Y_{mis}^{(A)}$ be the first 20% of Y_0 and $Y_{mis}^{(B)}$ be the last 30% of $Y_1, \ldots, Y_{20}$. The partially completed datasets $D_{pcom}^{(i)}, i = 1, \ldots, m$, are generated by drawing values from $f(Y_0|D_{obs})$ using a multivariate normal distribution with an unrestricted covariance matrix. The completed datasets $D_{com}^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, n$ are then generated by drawing from $f(Y_1, \ldots, Y_{20}|D_{pcom}^{(i)})$, again from a multivariate normal distribution. The number of imputations is varied, with $m \in (2, 5, 10, 20)$ and $n \in (2, 5, 10, 20)$.

The hypothesis tested is $H_0 : Q = 0$, where Q is the vector of coefficients for the regression of Y_0 on $Y_1, \ldots, Y_k$, excluding the intercept, for $k \in (5, 10, 20)$. As this null hypothesis is true in the simulated data, the nominal rejection rate is expected to be close to $100\alpha\%$, for a given significance level α. Table 2.6 compares the simulated nominal significance levels for 1000 iterations using each combination of m, n, and k, for $\alpha \in (.01, .05, .10)$ using denominator degrees of freedom w_n and w_n^*. The simulated significance levels using w_n are seen to be generally closer to the expected significance levels than when w_n^* is used.

2.7 Concluding Remarks

Popular software packages contain routines for obtaining confidence intervals for scalar quantities and p-values for multi-component tests from conventional multiply-imputed datasets. These routines can be easily modified to perform the tests proposed here. Analysts should use the Wald-type test when possible, because the likelihood

Table 2.6: Comparison of rejection rates for tests using w_n and w_n^* for 2-stage multiple imputation for missing data only

		w_n			w_n^*		
		$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
$\alpha = .01$							
	$n=2$	0.4	1.0	0.9	0.4	0.0	0.0
$m=2$	$n=5$	0.2	0.9	0.6	0.3	0.3	0.0
	$n=10$	0.2	1.2	1.0	0.4	0.3	0.0
	$n=20$	0.3	1.1	1.0	0.6	0.3	0.0
	$n=2$	1.0	1.2	1.5	0.4	0.1	0.0
$m=5$	$n=5$	1.0	1.1	1.4	0.3	0.1	0.0
	$n=10$	0.9	1.2	1.3	0.2	0.1	0.0
	$n=20$	0.9	1.1	1.5	0.5	0.1	0.0
	$n=2$	0.7	1.0	1.4	0.5	0.4	0.1
$m=10$	$n=5$	0.7	1.0	1.4	0.5	0.6	0.1
	$n=10$	0.6	1.1	1.3	0.3	0.6	0.1
$\alpha = .05$							
	$n=2$	1.4	4.1	4.9	1.9	0.5	0.0
$m=2$	$n=5$	1.9	5.1	5.7	2.4	1.0	0.1
	$n=10$	1.9	4.6	6.5	2.8	1.3	0.1
	$n=20$	2.0	4.6	6.2	2.5	1.4	0.1
	$n=2$	5.3	6.0	6.3	2.3	1.4	0.4
$m=5$	$n=5$	4.9	5.6	7.2	2.6	2.0	1.0
	$n=10$	5.0	5.9	7.0	2.9	2.1	1.3
	$n=20$	5.2	5.6	6.4	3.0	2.2	1.3
	$n=2$	6.0	4.9	6.9	4.2	2.8	2.3
$m=10$	$n=5$	5.7	4.9	7.4	4.0	3.5	3.0
	$n=10$	5.6	5.3	7.3	4.3	3.7	3.2
$\alpha = .10$							
	$n=2$	5.7	9.1	11.1	5.1	1.4	0.1
$m=2$	$n=5$	5.5	10.3	11.6	5.1	2.6	0.2
	$n=10$	5.1	11.0	11.4	4.6	2.6	0.4
	$n=20$	4.8	11.0	11.4	4.6	2.7	0.3
	$n=2$	10.6	12.7	13.1	7.0	5.3	2.7
$m=5$	$n=5$	10.7	12.3	13.8	7.6	5.9	3.6
	$n=10$	10.9	12.7	14.3	7.7	5.7	3.7
	$n=20$	10.6	12.8	13.6	7.9	6.2	3.9
	$n=2$	9.9	12.3	12.6	9.1	7.7	6.5
$m=10$	$n=5$	11.0	12.4	13.0	9.2	8.3	8.0
	$n=10$	10.7	11.8	12.8	9.5	8.8	8.1

ratio test involves further approximations. Software distributed with partially synthetic datasets can make the Wald-type test the default option.

The simulations suggest that the Wald-type tests have appropriate rejection rates when the null hypothesis is true. To get a sense of the power properties of these tests, we can turn to the results of Li *et al.* (1991b) and Shen (2000). These tests are derived from similar assumptions and approximations as the Wald-type tests proposed here. Based on extensive simulation studies, Li *et al.* (1991b) report that power curves for their tests are similar to the power curves for Wald-type tests based on the observed data. The greatest losses in power occur when the data deviate substantially from the proportionality assumption. The losses are largest when m is small, and mostly disappear for large m. Shen (2000) reported similar findings for nested imputation, with greatest power loss for small m and n and for large deviations from proportionality. The tests proposed here are expected to have similar properties.

Chapter 3

Generating partially synthetic public release files for the Longitudinal Business Database

This chapter describes the generation of partially synthetic public release files for the U. S. Census Longitudinal Business Database (LBD). The synthesis approach is described and a preliminary assessment given of the utility and risk for a synthetic dataset generated for two industry subgroups, and plans for extending the imputation strategies to the entire LBD are discussed. The combining rules for partially synthetic data of Reiter (2003) and Reiter (2005b) were developed for random samples and thus are modified here for population data.

The LBD contains longitudinal payroll and employment data for U. S. establishments with employees, covering nearly all non-farm private economy and some public sector activities. It was generated by creating longitudinal linkages for all employer establishments contained in the Census Bureau's Standard Statistical Establishment List (Business Register), which serves as the data frame, and it draws information from various administrative records, economic censuses and surveys. The LBD is a unique and valuable source of information on the structure of the U.S. economy, in particular for examining gross job flows and entry and exit of establishments by industry (Jarmin and Miranda, 2002). Currently, controlled access to this data is granted to researchers by special agreement with the U.S. Census Bureau at one of nine Census Research Data Centers. No public use files are available.

The goal in generating public release files for the LBD is to allow researchers to gain access to the information it contains without having to spend time and money

gaining access to the confidential data. It is not expected that synthetic data inferences for individual establishments or highly refined subgroups will be valid, but that broad analyses on employment and payroll levels and trends will be valid for moderately sized industrial and geographic groups. Users desiring detailed analyses will need to apply for access to the data at a Census Research Data Center; however, the public data will still be of benefit to them. As the structure of the public-use files will be the same as the observed data, users in the process of obtaining access, considering applying for access, or that have access but are not located near a Research Data Center, will be able to write and refine their analyses using the public data. The analyses can then be repeated on the confidential data prior to publishing research results.

The remainder of this chapter is organized as follows. In Section 3.1, some of the methods for generating synthetic data that were used or explored in the generation of the synthetic LBD are described. The details of the LBD synthesis are given in Section 3.2. The utility of the synthetic data is illustrated with some economic analyses using confidential and synthetic data for a portion of the LBD in Section 3.3 and a preliminary risk assessment is discussed in Section 3.4. Section 3.5 describes future plans for completing the synthesis of the LBD. Lastly, Section 3.6 describes modifications to the combining rules of Rubin (1987) for missing data and of Reiter (2003) for partially synthetic data for use with population data.

3.1 Data generating methods

Partially synthetic datasets are constructed by replacing selected values in the observed data with m independent draws from their posterior predictive distribution. Let $Z_j = 1, j = 1, \ldots, N$ indicate that unit j has been selected to have any observed values replaced with imputations. Imputations should only be made from the pos-

terior predictive distribution of those units with $Z_j = 1$. For the LBD, all units have their values of confidential variables replaced, i.e., $Z_j = 1$ for all units. Let $Y = (y_1, \ldots, y_d)$ be the matrix of confidential variables that will be replaced with imputations and X the matrix of variables that will not be replaced, and assume that all the units are fully observed, i.e., no missing values are present. Let $Y_{rep}^{(i)}$ be the imputed values of Y in the ith synthetic dataset, $i = 1, \ldots, m$, assumed to be drawn from the posterior predictive distribution of Y. Each of m synthetic datasets, $D_{syn}^{(i)}$, is comprised of $(X, Y_{rep}^{(i)})$. The set $D_{syn} = \{D_{syn}^{(i)}, i = 1, \ldots, m\}$ is what will be released to the public.

As several variables are considered confidential, i.e., Y has dimension $N \times d$, specification of the joint posterior density $Y|X$ is difficult. Writing the joint distribution as a product of conditional densities simplifies the specification. For $Y = (y_1, \ldots, y_d)$, sampling from $Y|X$ is thus achieved by sampling from a series of conditional distributions, $f(y_1|X)$, $f(y_2|y_1, X), \ldots, f(y_d|y_1, \ldots, y_{d-1}, X)$. This allows complex relationships to be modeled in a computationally feasible fashion.

Ideally, imputation models should be as general and saturated as possible in order for analysts to be able to make a broad range of valid inferences (Meng, 1994). Some simplifications may be needed to make modelling the data tractable; however, they should be applied sparingly in order to maximize the utility of the partially synthetic data. In the LBD synthesis, some predictors are omitted from the imputation model where an independence relationship is reasonable. For example, for the longitudinal variables, values in year t are assumed to be dependent on values in year $t-1$, but not on values in previous years. Similar practical considerations in choosing imputation models, encountered in the imputation of missing data for the National Health and Nutritional Examination Survey (NHANES), are described in Schafer *et al.* (1993).

The remainder of this section describes several approaches for modeling different

types of data that are used in the imputation of the LBD. The choice of methods used depends on the form of the data, computational concerns, and the order in which the variables are imputed.

3.1.1 Normal Method

A common approach for generating imputations for continuous variables is to model the posterior distribution using a normal linear regression model, possibly on transformed data. Given the highly skewed nature of payroll and employment data in the LBD, the kernel density estimation procedure of Abowd and Woodcock (2004) is used to transform the response variables so that marginally they approximately follow standard normal distributions, and a normal model can be used.

Using the normal approach with a KDE transform, a synthetic variable $\tilde{y}_k^{(i)}$ is generated from $(X, y_1, \ldots, y_{k-1}, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$ by drawing from the posterior predictive distribution of y_k as follows:

1. Apply the KDE transform to the response variable and any needed tranformation functions to the predictors to satisfy approximately linear regression assumptions. For simplicity, the tranformations performed on the predictors are not notated here, though the models used are given in Section 3.2. When the KDE transforms were applied to the linear predictors, the observed-data correlations were not preserved in the synthetic data.

   For each observed value $y_{k,l}, l = 1, \ldots, n$, of response variable y_k, the transformed values $y_{k,l}^*$ are computed as $\Phi^{-1}(\hat{K}(y_{k,l}))$, where Φ denotes the standard normal cumulative distribution function and $\hat{K}(y_{k,l})$ is a kernel density estimate of $y_{k,l}$.

2. Fit a linear model, $f(y_k^*|Z, \beta, \sigma^2) = N(Z\beta, \sigma^2)$ to the observed data, where $Z = (X, y_1, \ldots, y_{k-1})$, and obtain estimates of β and σ^2.

3. For each imputation, draw new values $\tilde{\sigma}^{2(i)}$ and $\tilde{\beta}^{(i)}$ from the posterior distributions $f(\sigma^2|X, Y)$ and $f(\beta|\sigma^2, Y, X)$.

4. Draw $\tilde{y}_k^{(i)*}$ from $N(\tilde{Z}^{(i)}\tilde{\beta}^{(i)}, \tilde{\sigma}^{2(i)})$, where $\tilde{Z}^{(i)} = (X, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$.

5. Apply the inverse KDE transform, $\tilde{y}_{k,l}^{(i)} = \hat{K}^{-1}(\Phi(\tilde{y}_{k,l}^{(i)*}))$, to return to the original scale of y_k.

Step 3 can be considered optional for census data as the parameters are considered to be known and can be computed from the data. In Section 3.6, modifications to the partially synthetic data combining rules for population data are given and a simple simulation study suggests that these can still yield valid inferences when Step 3 is skipped. Including this step can potentially reduce disclosure risks by increasing between-imputation variance. For speed and simplicity, this step is omitted in the imputation of the LBD. Similarly, the transformation function applied depends on the data, and thus contains uncertainty when imputing random samples; hence, Abowd and Woodcock (2004) draw a Bayesian bootstrap sample to estimate the transformation in each implicate to account for this additional uncertainty. This step is also skipped in the synthesis of the LBD.

More flexible approaches utilizing a Generalized Additive Model (GAM) were considered for the LBD but ultimately the preservation of the correlation structure as well as the computational intensity proved unsatisfactory, while the normal method with the KDE transformation proved to be fast and effective. Computational speed is an important consideration in the LBD synthesis as there are over 21 million records.

3.1.2 Nonnormal Models

The normal approach can be modified for nonlinear models by replacing the normal model with a nonlinear one. For binary and categorical responses without very many categories, one can sample from binomial and multinomial distributions, using appropriate generalized linear models to obtain the sampling probabilities.

The synthetic variable $\tilde{y}_k$ for binary response y_k is generated by approximating draws from $f(y_k|X, y_1, \ldots, y_{k-1}, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$ as follows:

1. Use the observed data to fit a logistic model, $\text{logit}(p(y_k = 1)) = Z\beta$, where $Z = (X, y_1, \ldots, y_{k-1})$, to obtain $\hat{p}_l(Z_l), l = 1, \ldots, N$.

2. Update model parameters by taking draws from their posterior distributions. As before, this step is considered optional and omitted in the LBD synthesis.

3. Use the observed data model to obtain $\hat{p}_l(Z_l^{(i)})$, where $Z_l^{(i)} = (X_l, \tilde{y}_{1,l}^{(i)}, \ldots, \tilde{y}_{k-1,l}^{(i)})$.

4. Obtain $\tilde{y}_k^{(i)}$ by sampling from $Bin(1, \hat{p}_l(\tilde{Z}_l^{(i)})), l = 1, \ldots, N$.

For categorical responses, the same approach can be used, with a generalized logit model used in place of the logistic model to obtain the posterior probabilities, $\hat{p}_{lj}(x_l, y_{1,l}, \ldots, y_{k-1,l}), l = 1, \ldots, n; j = 1, \ldots, c$, where c is the number of categories in the response. A multinomial distribution is used in place of the binomial.

3.1.3 Dirichlet-multinomial method

When there are many categories in the response, and many categorical predictors, the generalized logit model can become computationally infeasible. The simpler and faster Dirichlet-multinomial approach provides a convenient framework for sampling

from the posterior predictive distribution for a categorical y_k when the predictors in $X, y_1, \ldots, y_{k-1}$ are categorical.

Let c be the number of categories in the response y_k. Let l be the number of unique categories determined by the predictors in $X, y_1, \ldots, y_{k-1}$. Assuming a flat prior on the cell probabilities, $\tilde{y_k}$ is generated as follows:

1. Use the observed data to determine the cell counts $n_j = (n_j^1, \ldots, n_j^c)$, $j = 1, \ldots, l$.

2. Draw new values of the cell probabilities $p_j = (p_j^1, \ldots, p_j^c)$ from a Dirichlet(n_j).

3. For each unit in the synthetic data, look up the appropriate cell probabilities p_j based on the values of X and $\tilde{y}_1, \ldots, \tilde{y}_k$.

4. Sample from a multinomial distribution with cell probabilities p_j.

As in the previous methods, we skip Step 2. In Step 3, if an exact cell match is not found in the observed data, a possibility depending on the disclosure control applied to $y_1, \ldots, y_{k-1}$, then the cell is collapsed untill a match is found. Hence, in Step 1, cell counts must be determined for one or more sets of broader categories as well.

This approach is very fast computationally and appears to yield good predictions with sufficient disclosure control when used in the LBD synthesis. With sufficient variability in the observed data, disclosure control is provided by sampling from the multinomial distribution and by the disclosure control methods applied to any predictors. In some cases, this method fails to provide sufficient disclosure protection. When there are a large number of categories and categorical variables, numerous

units are uniquely determined by their values of the categorical predictors, yielding predictions that are "too good." For example, let C be a unique category determined by categorical predictors in X and let y_C be the observed values of a categorical response variable corresponding to the n_C units in C. If $n_C = 1$, or $y_{Ci}, i = 1, \ldots, n_C$ all have the same value, then a categorical model will impute synthetic values $\tilde{y}_C$ for y_C such that $\tilde{y}_C = y_C$ in each implicate. This creates a high risk of re-identification of y_C.

Disclosure control in this case is improved by using an informative prior distribution to add a positive probability that, for a given category C, the $\tilde{y}_C$ generated may contain values not present in y_C. The prior is estimated by replacing one of the categorical predictors with a coarsened version and using this to determine the prior cell counts. For example, if County is a predictor, the prior could be obtained from state-level cell counts. The prior counts are normalized to represent a small number of units to reduce the sensitivity to the prior. This serves to add noise in a controlled fashion, meeting the goal of reducing disclosure risks with minimal loss of utility.

Let c be the number of categories in the response y_1. Let l be the number of unique categories determined by the predictors in X and let p be the number of unique categories in a coarsened version of X, X_p, i.e., with one or more of the predictors dropped or coarsened, so that $l > p$ and X_p has fewer categories than X and larger cell counts. Generate draws from $f(y_1|X)$ as follows:

1. Using the observed data, determine the cell counts $n_k = (n_k^1, \ldots, n_k^c)$, $k = 1, \ldots, p$. Normalize these cell counts so that $\sum_{i=1}^{p} n_k^i = a$, where a is a small number. Larger values will give more weight to the prior.

2. Using the observed data, determine the cell counts $n_j = (n_j^1, \ldots, n_j^c)$, $j = 1, \ldots, l$. Add each n_j to its corresponding normalized n_k to obtain the posterior

counts $m_j, j = 1, \ldots, l$.

3. Draw new values of the cell probabilities $p_j = (p_j^1, \ldots, p_j^c)$ from a Dirichlet$(m_j = (m_j^1, \ldots, m_j^c))$.

4. For each unit in the synthetic data, look up the appropriate cell probabilities p_j based on the values of X.

5. Sample from a multinomial distribution with cell probabilities p_j.

As before, for the LBD, we omit drawing parameters in Step 3, and sample from a multinomial distribution with cell probabilities given by $m_j / \sum_{i=1}^{c} m_j^i$.

3.2 Imputation of the LBD

This section describes in detail how the methods of Section 3.1 are used to generate public use files for longitudinal establishment data from the LBD. The version of the LBD being synthesized is based on the most recent release, made available to authorized users by the Center for Economic Studies (CES) in May 2007. The variables intended for inclusion in an eventual public-release file are described in Table 3.1. The existence of industries by type and county is already public information, so county and industry codes are not synthesized but all other variables must be synthesized for the data to be considered for public release. While a number of establishments exist that can be uniquely determined by their county and industry type, all of their attributes will be synthesized. Special attention may need to paid to these establishments when a thorough risk assessment is conducted. The final form of the public release files will be determined during the risk assessment and disclosure review stages.

Table 3.1: LBD Variable Descriptions

Variable	Name	Type	Description
$x1$	County	categorical	Geographic Location
$x2$	SIC	categorical	Industry Code
$y1$	Firstyear	categorical	First Year Establishment is Observed
$y2$	Lastyear	categorical	Last Year Establishment is Observed
$y3$	Multiunit	categorical	Multiunit Status
$y4$	Employment	continuous	March 12 Employment (26 years)
$y5$	Payroll	continuous	Annual Payroll (26 years)

Establishments started after 2001 will not be included in the initial synthetic public release files due to changes in the industry coding systems. The system for coding industry types in the LBD changed from the Standard Industrial Classification (SIC) system to the North American Industry Classification System (NAICS) in 2001. The mapping between NAICS and SIC is neither one-to-one or onto and thus the initial synthetic public release file includes only the establishments started in 2001 or prior, for which SIC codes are available. Future releases may be generated based on NAICS once a method for linking NAICS and SIC has been established.

When imputing the LBD, we assume that there are no missing values, and thus impute the synthetic data in a single stage. In actuality, a small percentage of missing values are present. If a large proportion of values were missing, it would be appropriate to use two-stage multiple imputation, as in Reiter (2004) and Chapter 2, for handling both missing values and disclosure limitation. As the percentage of missing values is very small, only a few percent compared to replacement of nearly all the data for disclosure control, the improvement to be gained from imputing in two stages is neglible, while the effort would be substantial. The missing values are thus imputed during the synthetic data imputation. Units that are missing values for industry code or geography are discarded, as those values are required for a unit to be included in the data frame.

Some additional data cleaning is also performed on the observed data prior to gen-

erating the synthetic data using code provided by CES (Jarmin and Miranda, 2007). This includes applying a smoothing filter so that values for large establishments with unrealistic growth are smoothed to an average of the previous year's value and the following year's value. In a given year, establishments with employment greater than 500 and growth rate higher than 1.67 have their employment smoothed to the average of the previous and following year values. The growth rate is defined as the magnitude of the change in number of jobs from the previous year over the average of the total jobs over the current and previous year (Davis *et al.*, 1996). A similar step is used to smooth values of payroll for units that exhibit high growth when the payroll value is greater than 2,500. Remaining values that are still unrealistically high are deleted. This includes payroll values greater than 5,000,000, payroll values greater then 500,000 corresponding to a payroll-employment ratio greater than 4,500, and payroll values greater than 10,000 with a payroll-employment ratio greater than 80,000.

The SIC codes, and the NAICS codes, are six-digit codes with a nested structure. The first two digits designate the highest-level division, the third digit a sub-division, the fourth digit a smaller industry group. As the LBD is very large, the variable SIC is used to define subgroups of the LBD that are imputed separately. At present, synthesis of the entire LBD has not been completed, so as the imputation of each variable is described below, results comparing marginal distributions are presented for a subgroup of approximately 130,000 retail establishments in one 3-digit SIC group (Group 1). This subgroup was used for most of the model development. Results for a subgroup of approximately 25,000 manufacturers in one 3-digit group (Group 2) are given for comparison and to evaluate how well the models perform on a group with completely different characteristics. Additional evaluation of risk and utility is conducted in Sections 3.3 and 3.4. For the purpose of providing a better comparison

of sums, the observed data have had their missing values imputed using the synthetic data model prior to the synthetic data generation. This step will not be conducted in the actual synthesis.

The imputation strategy is to build up the joint distribution for each by-group as follows:

1. Impute Firstyear using the Dirichlet-multinomial approach to approximate draws from $f(y_1|x_1, x_2)$.

2. Impute Lastyear using the simple multinomial approach to approximate draws from $f(y_2|y_1, x_1, x_2)$.

3. Impute a categorical Multiunit status using the simple multinomial approach to draw from $f(y_3|y_2, y_1, x_1, x_2)$.

4. Impute Employment and Payroll using normal linear regression masking, with a kernel density estimator transformation applied to the response to draw from $f(y_4^{(t)}|y_4^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$ and $f(y_5^{(t)}|y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$, where t indicates a year between 1976 and 2001.

A constraining factor in the generation of synthetic data for the LBD is the population size. With over 21 million records in the dataset to be synthesized, there are limited software programs and algorithms that can be used to generate the data in a timely manner. Modeling approaches such as Monte Carlo Markov Chain algorithms which can take hours or days on much smaller datasets are infeasible for imputing the entire LBD, even with the supercomputer available to this project. With the bulk of the system dedicated to the effort, it is estimated that with the current methods implemented, the entire LBD synthesis will still take days or weeks to run. Due to the Census Bureau preference for software development, the methods used were programmed using SAS.

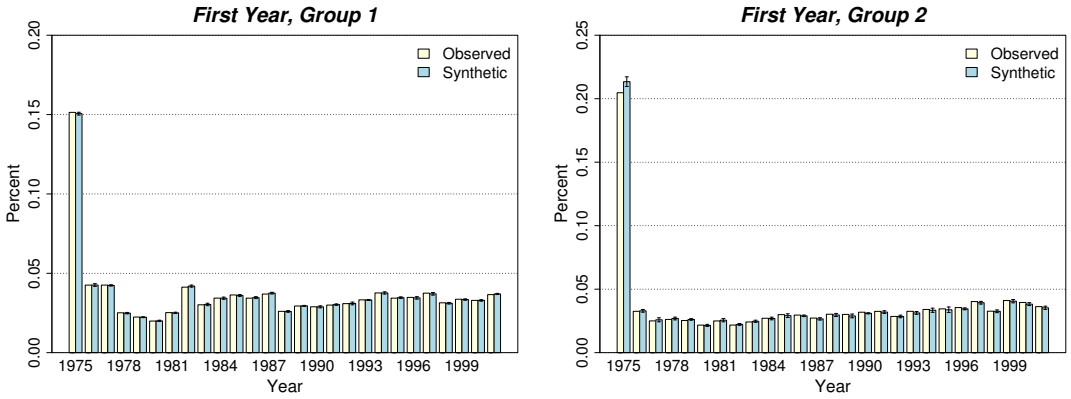

Figure 3.1: Observed and Synthetic Distributions of First Year, Groups 1 and 2

3.2.1 Firstyear

The variable Firstyear contains 27 categories, namely the years 1975 through 2001, representing the first year an establishment is observed. This is predicted conditional on 4-digit SIC and County. There are over 3000 counties in the United States and numerous SIC groups. This results a large number of unique county-SIC groups, that with 27 categories in the response, it is not always possible to use a generalized logit or similar model to predict the response. Furthermore, there are many county-SIC groups for which all the observed units have the same observed value of Firstyear. Using standard categorical models could result in imputed values that were all idential to the observed value, a serious disclosure risk. Hence the Dirichlet-multinomial approach with an informative prior, described in Section 3.1.3, is used to handle this case.

The marginal distribution is well-preserved for both Group 1 and Group 2 using this approach as seen in Figure 3.1. The synthetic data distributions shown are based on the mean percent of units born in each year, across five implicates, with 95-percent standard error bars indicating a small between-implicate variability. The large spikes in 1975 are due to censoring.

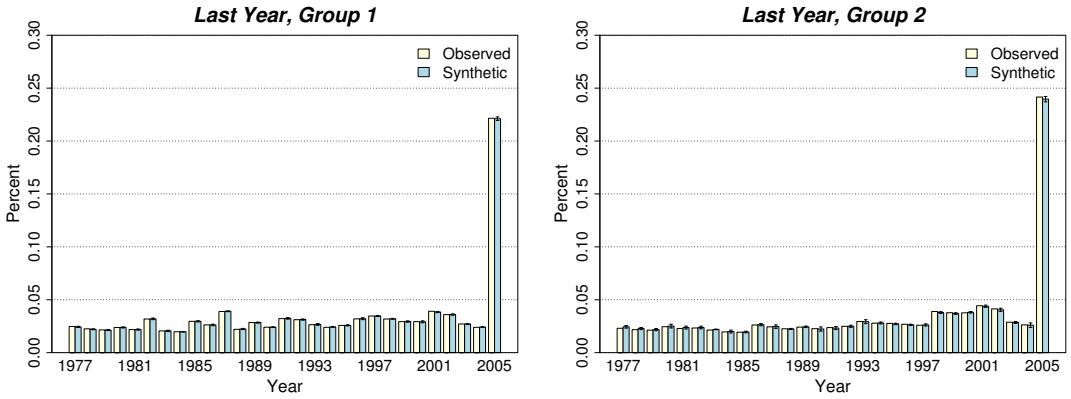

Figure 3.2: Observed and Synthetic Distributions of Last Year, Groups 1 and 2

3.2.2 Lastyear

The variable Lastyear contains 30 categories, the years 1976 through 2005, representing the last year an establishment is observed. Imputations are generated for Lastyear using the Dirichlet-multinomial approach with a flat prior. Frequencies of Lastyear values are determined for each category determined by a combination of Firstyear and 4-digit SIC using the observed data. Dependencies on geographic variables are not accounted for. For logical consistency, the probability that the value of Lastyear for a given unit can be less than the imputed value of Firstyear is set to zero and the remaining cell probabilities are normalized. Figure 3.2 shows the close correspondence between the observed frequencies of the variable Lastyear and the synthetic frequencies. Additionally, the marginal distribution of Lifetime, where Lifetime = Lastyear − Firstyear, is preserved, as shown in Figure 3.3. Figure 3.4 gives a sample comparison of a conditional distribution. As before, the synthetic data distributions are averaged over five implicates, with 95 percent standard error bars indicating a small between-implicate variability. The spikes at 2005 in Figure 3.2, 30 in Figure 3.3 and 15 in Figure 3.4 are due to censoring.

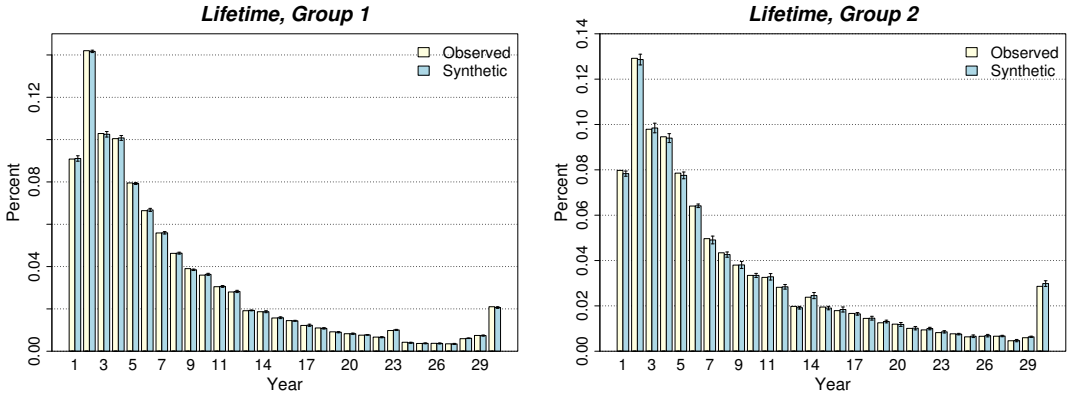

Figure 3.3: Observed and Synthetic Distributions of Lifetime, Groups 1 and 2

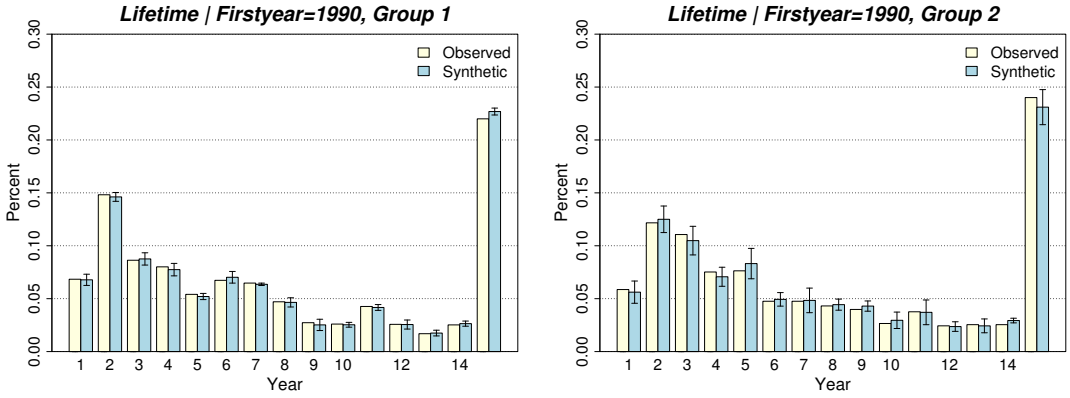

Figure 3.4: Observed and Synthetic Distributions of Lifetime given Firstyear=1990, Groups 1 and 2

3.2.3 Multiunit

The variable Multiunit indicates whether or not an establishment was ever part of a multi-unit firm, i.e., whether an establishment was ever part of a parent enterprise conducting business at multiple locations. In the observed data, multiunit status is given by a series of longitudinal binary indicators of multiunit status for each year. To facilitate synthesis, a categorical variable was defined such that a value of 1 indicates an establishment was never part of a multi-unit firm; values of 2-4 indicate a change in multi-unit status at some point in the lifetime of the establishment; and a value of 5 indicates the establishment was always part of a multi-unit firm.

The synthesis of this categorical variable using the Dirichlet-multinomial approach with a flat prior was straightforward and proved to be much faster and just as effective as the nonnormal method using a generalized logit model. The predictors used include Firstyear, Life (Lastyear − Firstyear), 4-digit SIC, and State. Some combinations of these variables using the synthetic Firstyear and Life variables did not correspond to categories in the observed data. For these cases, predictors were dropped until a match could be found.

For units that are predicted to change their multiunit status over the course of their lifetime, the year when the change occurs is also of interest and is planned for synthesis. Firm structure and linkages between establishments in the same firm are not planned for synthesis. Table 3.2 shows the close correspondence between the observed frequency of Multiunit compared to 95-percent confidence intervals obtained from five synthetic implicates.

3.2.4 Payroll and Employment

Payroll and employment data are imputed for each active establishment in every year between 1976 and 2001. If the synthetic values of Firstyear and Lastyear indicate an

Table 3.2: Observed and Synthetic Distributions of Multiunit, Groups 1 and 2

	Group 1		Group 2	
Value	Obs Pct	Synth Pct	Obs Pct	Synth Pct
1	0.696	(0.696,0.698)	0.924	(0.924, 0.930)
2	0.011	(0.011,0.012)	0.016	(0.016, 0.017)
3	0.002	(0.002,0.002)	0.001	(0.001, 0.001)
4	0.001	(0.001,0.001)	0.003	(0.002, 0.003)
5	0.289	(0.287,0.290)	0.056	(0.050, 0.056)

establishment was inactive in a given year, then no payroll or employment value is generated. The employment variables are imputed first, in ascending order by year, followed by the payroll variables. Separate regressions are estimated for different subgroups of data. For example, employment for births and continuers in a given year are predicted separately as employment in the previous year is a predictor for continuers but obviously not for births. Single-unit and multi-unit establishments are treated separately since they have substantially different characteristics, and are typically analyzed separately in economic analyses. Initially, establishments that change their multiunit status over time are treated as multiunit establishments in the imputation of payroll and employment. When the year of status change is imputed then the longitudinal binary indicators can be reconstructed and the status in the current year used to predict payroll and employment for a given year.

While general, saturated models are desired for generating imputations yielding a broad range of valid inferences, models with good predictive properties are needed for any inferences to be valid. Different models provided better predictions in different years, so to keep the modelling procedure as flexible as possible, a variable selection procedure based on the Bayesian Information Criterion (BIC) is employed. Thus parsimonious models yielding good predictions are used to impute payroll and employment.

Births

For establishments in their first year, first-year employment and payroll are predicted from observed data corresponding to units in their first year. The predictors in the saturated model for employment births currently include 4-digit SIC, years till death, and indicators for whether or not the firstyear or lastyear are censored, and an indicator for the penultimate year. The predictors in the saturated model for payroll are the same, with the addition of the log of current year's employment.

As the employment variables are defined as the number of employees as of March 12th in a given year, a large percentage of establishments start their first year after March 12 and hence have zero employment recorded for the first year. In order to reflect this in the synthetic data and prevent the zeros from influencing the imputations for establishments with nonzero employment in their first year, the birth model is imputed in two stages. First, a logistic regression is used, following the Nonnormal method, to predict whether or not units in their first year had zero employment, conditional on 4-digit SIC and years to death. Then, the observed data for units with nonzero employment are used to impute employment for units predicted to have nonzero employment using the Normal method, with a KDE transform on the response variable.

Continuers

The predictors in the saturated employment model for continuers are currently 4-digit SIC, age, years to death, log of previous year's employment, indicator variables indicating the first year is censored or the last year is censored, as well as indicators for the second year, the penultimate year, and the last year, and interactions of these three indicators with the log of previous year's employment. The predictors in the saturated model for payroll include all 4-digit SIC, age, years to death, log of current

year's employment, log of previous year's payroll, indicator variables indicating the first year is censored, the last year is censored, if the current year is the second year, the penultimate year, or the last year, and interactions of payroll and employment with the last year and penultimate year indicators.

Imputations are generated using normal models with KDE transformations applied to the response variables to satisfy approximately the normal linear model assumptions. When establishments have zero employment in their first year, it is assumed that this is due to the recording of the number of employees on March 12, and not due to any other establishment characteristic. Hence, establishments in their second year that had zero employment in their first year are imputed separately using the employment model for births.

Imputing all of the continuers in a given year together adequately preserved the marginal distributions and correlations for most of the variables; however, for some variables the means were different enough to throw off analyses of job creation and destruction, which depend on sums of employment values for units in their first and last year, and sums of expansions and contractions. The root cause of this was the presence of extreme outliers, which can affect the mean and sum if either too many outliers are imputed or too few. Hence continuers were imputed in two stages, for both payroll and employment. This is improves the quality of predictions for many continuer subgroups, though outliers still present a problem. Imputation of continuers proceeds as follows:

1. The observed data are divided into high and low groups based on a 95th-percentile threshold of the response variable.

2. Using all of the observed continuers to build a model, the Nonnormal method with a logistic model, and the predictors from the employment and payroll

continuer models, was used to impute whether or not a synthetic continuer should be in the high group or the low group.

3. Using the observed high group, employment was imputed for units imputed to be in the high group using the normal method with a KDE transform.

4. Using the observed low group, employment was imputed for units imputed to be in the low group using the normal method with a KDE transform.

Small subgroups

As just described, the LBD is broken down in many subsets to facilitate the modelling process. To summarize, for each year of employment and payroll, establishments in operation that year are broken down by multi-unit status, which is then separated into births and continuers. For births, zero births and nonzero births are separately imputed. For continuers, second-year establishments that had zero employment in the first year are separated for the imputation of employment. When sufficient continuers remain in the observed data for the saturated model to be full rank, the remaining units are split into the highest 5% and the lowest 95% of employment level. While over 21 million units are being imputed, when broken down into so many subgroups, many are quite small, as small as a single unit.

Small subgroups are addressed by incorporating an informative prior for the vector of regression coefficients β into the Normal method. For a given 3-digit SIC group, a comparable subgroup is found in the corresponding 2-digit SIC group, which is used to estimate the prior, and 4-digit SIC is dropped from the imputation model. For example, if there are too few single-unit nonzero births in a given year in the 3-digit SIC group being imputed, the prior is estimated from all single-unit nonzero births in the same year from the corresponding 2-digit SIC group. This is similar to the

common practice of using information from previous experiments, external surveys, and censuses to determine prior values.

Using a unit information prior allows for all of the available data to be used to estimate a prior mean and variance for the regression coefficient without overwhelming the data. The unit information prior has the same amount of information about β as contained in a single observation. In this case the information is in the sample used to estimate the prior, which has the form

$$\begin{aligned} p(\beta|\sigma^2) &= N(\beta_0, \sigma^2\Sigma_0) \\ p(\sigma^2) &= \chi^{-2}(n_0 - k, s_0^2) \end{aligned}$$

where $\beta_0 = (X_0'X_0)^{-1}X_0'Y_0$, $\Sigma_0 = n_0(X_0'X_0)^{-1}$, X_0 and Y_0 are the prior data for X and Y, s_0^2 is the sample variance $(Y_0 - X_0\beta_0)'(Y_0 - X_0\beta_0)/(n_0 - k)$, and n_0 is the prior data sample size.

The resulting posterior $(\beta, \sigma^2|Y, X)$, used to draw from the posterior predictive distribution, is given by

$$\begin{aligned} p(\beta|\sigma^2, Y, X) &= N(\hat{\beta}, \hat{\Sigma}) \\ p(\sigma^2|Y, X) &= \chi^{-2}(n + n_0 - k, s^2) \end{aligned}$$

where $\hat{\beta} = \hat{\Sigma}(\Sigma_0^{-1}\beta_0 + X'Y)$, $\hat{\Sigma} = (\Sigma_0^{-1} + X'X)^{-1}$, and $s^2 = \{(n_0 - k)s_0^2 + (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - \beta_0)\Sigma_0^{-1}(\hat{\beta} - \beta_0)\}/(n + n_0 - k)$.

When $n_0 \geq k$, this gives a full-rank model for drawing from the posterior predictive distribution under an informative prior. In addition to providing a full-rank model for small subgroups where $n < k$, this provides a degree of disclosure protection by using information from external data to build the model. If additional noise is desired, more weight can be given to the prior by replacing n_0 in the prior specification with $n_p < n_0$. If $n_0 < k$, predictors may be dropped to obtain a full-rank model, or the group used to estimate the prior may be expanded.

Results

Figures 3.5 and 3.6 illustrate the preservation of the marginal distributions of payroll and Figures 3.7 and 3.8 compare the marginal distributions of employment for one year of of the LBD. The examples shown for Group 1 in Figures 3.5 and 3.7 show that the marginal distributions are well preserved. The employment distribution shown for Group 2 in Figure 3.8 also does a reasonable job of preserving the overall shape of the distribution; however, the payroll distributions shown in Figure 3.6 differ in the lower range of the data. While other synthetic payroll variables do better match the observed data, this gives an example of the type of problem that arises due to extreme outliers. The outliers are not visible in these graphs as the tails of the distributions are compressed for disclosure purposes. In general, analyses excluding the outliers are comparable to analyses on the observed data that also exclude outliers. In analyses involving sums, such as the economic analyses in Section 3.3, the outliers are not excluded. Table 3.3 compares a sample of correlations computed using the observed data and the synthetic data. The correlations in Group 1 match very closely, with some attenuation. Greater attenuation and between-implicate variability is seen in the synthetic data correlations for Group 2. This is due to the presence of extreme outliers in the observed data that are imputed in different frequencies in different implicates.

3.3 Economic analyses

This section describes some economic analyses that are viewed by economists at the Center for Economic Studies as key analyses that should be approximately preserved in the synthetic data (Jarmin and Miranda, 2007) and compares these analyses performed on the confidential and synthetic data. All of the imputed variables must be well-modeled in order for the synthetic data results to match the confidential data

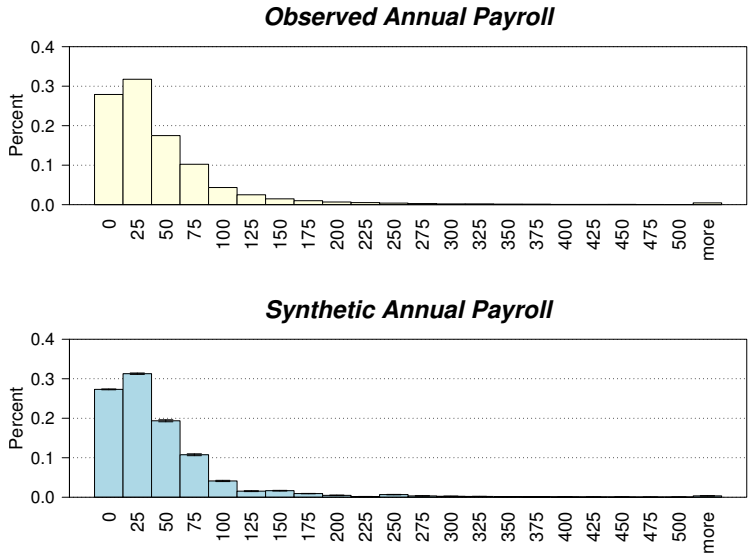

Figure 3.5: Observed and Synthetic Distributions of Annual Payroll (in $1000), One Year, Group 1

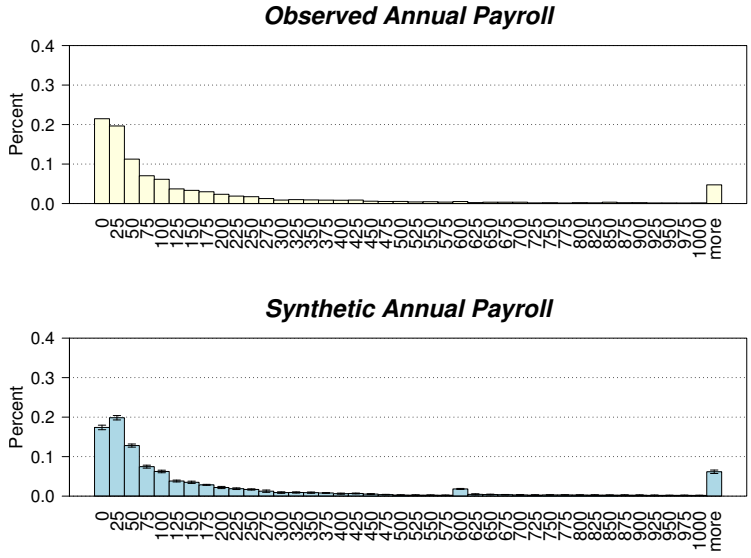

Figure 3.6: Observed and Synthetic Distributions of Annual Payroll (in $1000), One Year, Group 2

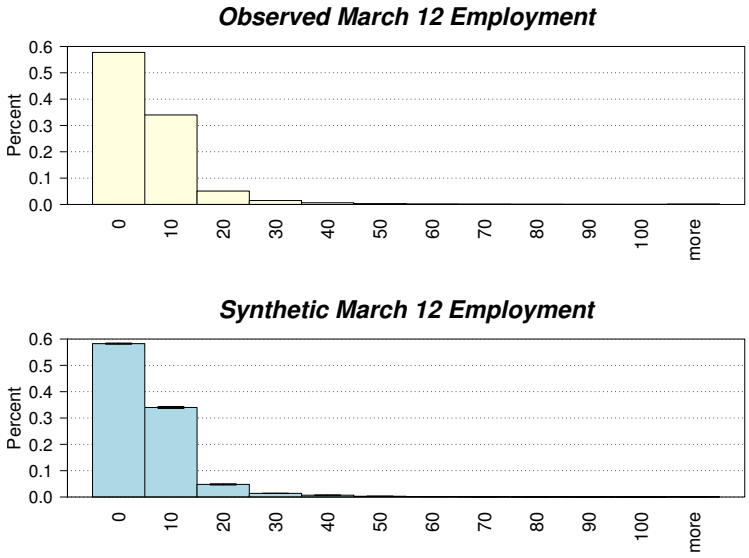

Figure 3.7: Observed and Synthetic Distributions of March 12 Employment, One Year, Group 1

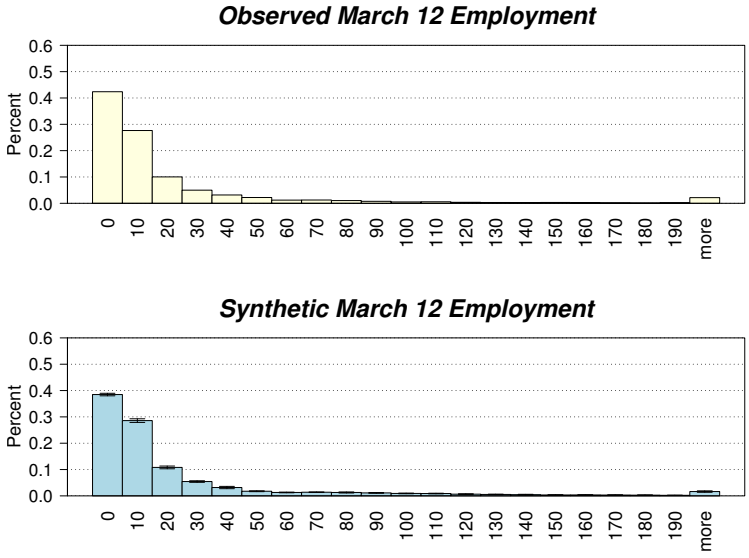

Figure 3.8: Observed and Synthetic Distributions of March 12 Employment, One Year, Group 2

Table 3.3: Sample Correlations on Observed and Synthetic Data, Groups 1 and 2

	Group 1		Group 2	
Variables	Obs.	Syn. Mean & CI	Obs.	Syn. Mean & CI
1978 Emp., 1979 Emp.	0.76	0.66 (0.59, 0.73)	0.96	0.93 (0.82,0.97)
1978 Emp., 1978 Payroll	-0.04	0.00 (-0.03,0.04)	0.54	0.38 (0.02,0.66)
1978 Emp., 1979 Payroll	0.61	0.57 (0.46,0.67)	0.91	0.71 (0.33,0.90)
1979 Emp., 1978 Payroll	-0.06	-0.02 (-0.04,0.01)	0.55	0.39 (0.02,0.67)
1979 Emp., 1979 Payroll	0.70	0.71 (0.58,0.81)	0.94	0.74 (0.42,0.91)
1978 Payroll, 1979 Payroll	-0.04	0.06 (0.01,0.12)	0.60	0.55 (0.17,0.80)
1997 Emp., 1999 Emp.	0.84	0.80 (0.77,0.82)	0.95	0.79 (0.67,0.87)
1997 Emp., 1997 Payroll	0.76	0.74 (0.66,0.81)	0.90	0.62 (-0.11,0.93)
1997 Emp., 1999 Payroll	0.63	0.74 (0.70,0.77)	0.87	0.67 (0.18,0.91)
1999 Emp., 1997 Payroll	0.76	0.63 (0.58,0.67)	0.90	0.55 (-0.15,0.90)
1999 Emp., 1999 Payroll	0.75	0.71 (0.69,0.73)	0.90	0.67 (0.19,0.90)
1997 Payroll, 1979 Payroll	0.76	0.70 (0.66,0.74)	0.98	0.70 (0.18,0.92)

results. The analyses are briefly described and then the results for each shown below.

3.3.1 Job creation and destruction

Job creation and destruction series are used to analyze job flows by economic sector. Using the growth rate definitions of Davis *et al.* (1996), job creation in year t for a given sector is the sum of employment over all establishments started in year t and employment gains for establishments that expanded between year $t-1$ and year t. The job creation rate is the job creation divided by the gross employment level, which is the sum of the average of the total jobs in year $t-1$ and year t. Similarly, job destruction in year t is the sum of employment over all establishments that are last observed in year t and employment losses for establishments that contracted between year $t-1$ and year t. The job destruction rate equals job destructions divided by the gross employment level. Net job flow is computed as the job creation rate minus the job destruction rate.

Figure 3.9 shows the job creation rate by year for Group 1 and Group 2. The rates

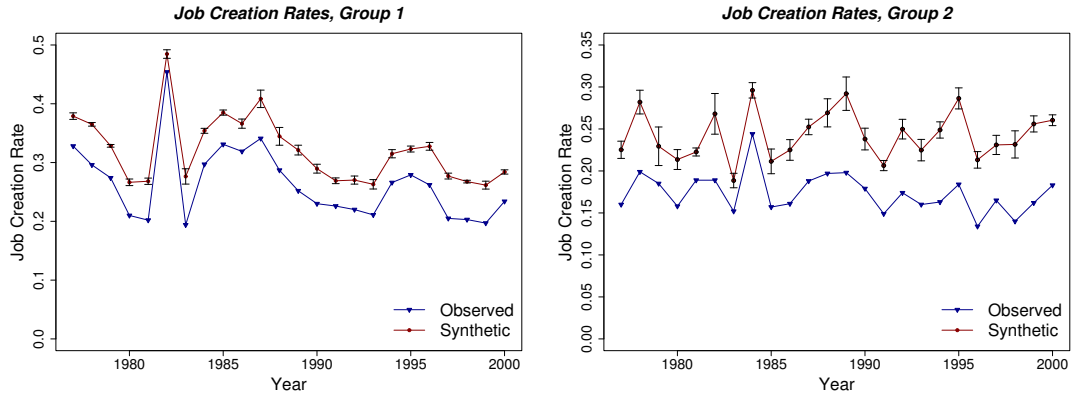

Figure 3.9: Job creation rate by year, Groups 1 and 2

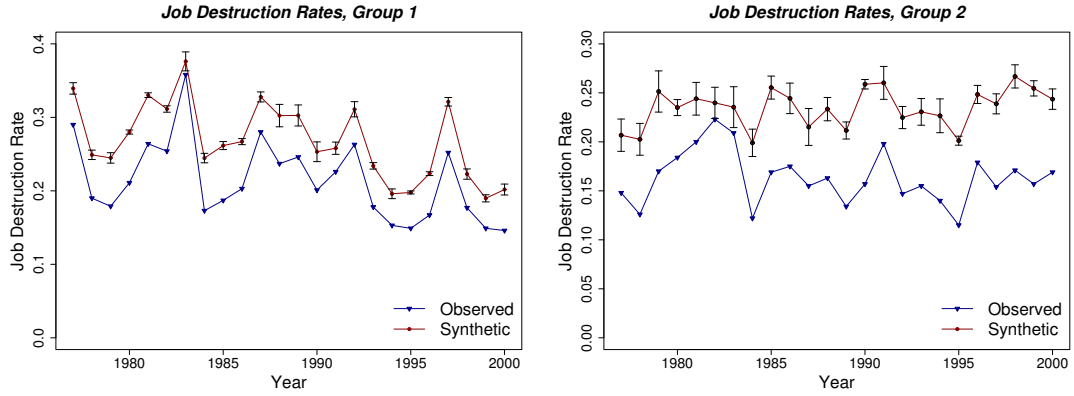

Figure 3.10: Job destruction rate by year, Groups 1 and 2

are seen to be consistently higher in the synthetic data, although the general trends are preserved. Similar results are seen for the job destruction series in Figure 3.10. Further investigation is needed to determine if this is occurs uniformly througout the LBD, and if it can be adjusted for. The discrepancy is largely due to establishment contractions and expansions; job creations and destructions due to births and deaths are generally well preserved. The synthetic net jobs flows, or the difference between the job creation and job destruction rates, are seen to match the observed data closely in Figure 3.11, as do the gross employment levels, shown in Figure 3.12.

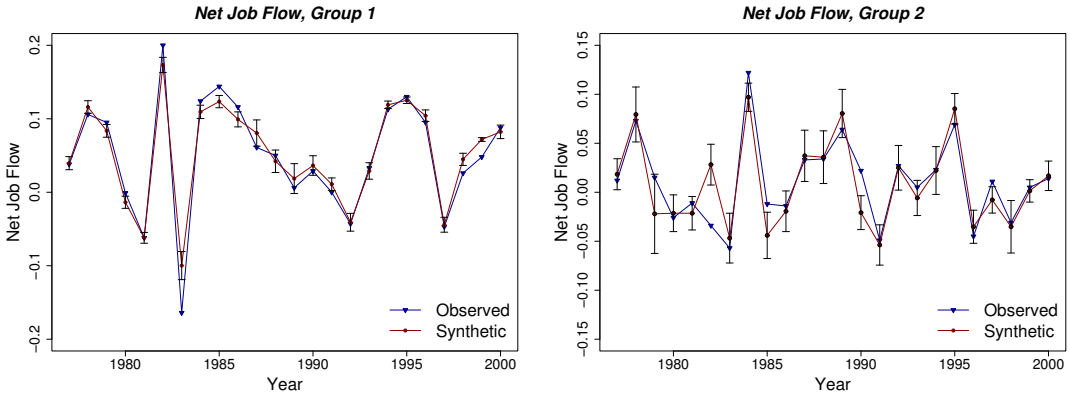

Figure 3.11: Net job flow by year, Groups 1 and 2

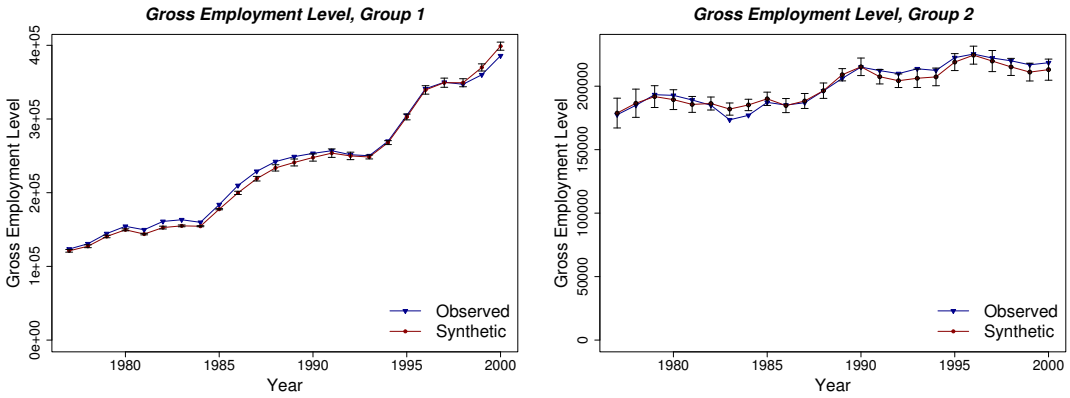

Figure 3.12: Gross employment level by year, Groups 1 and 2

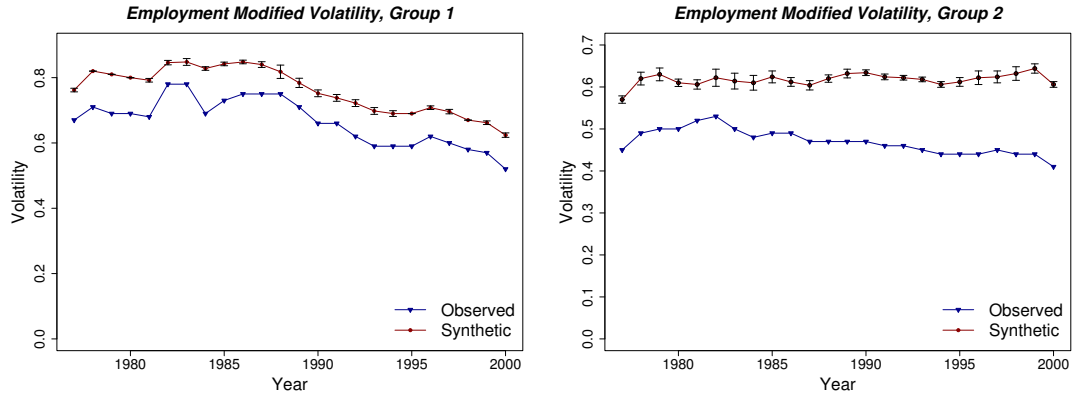

Figure 3.13: Employment volatility, Groups 1 and 2

3.3.2 Employment volatility

Employment volatility provides a measure of how much establishments expand and contract over time in terms of number of employees. The volatility measure used here is a measure of deviance from a ten-year moving average. Given the results seen for the job creation and job destruction series, it is not surprising that the volatility levels seen in Figure 3.13 also are consistently higher in the synthetic data than in the observed data. While the gross employment levels in the synthetic data tend to match the observed employment levels (Figure 3.12), the year-to-year changes at the establishment level tend to be greater in the synthetic data. The observed data contains many units with little or no change over time, a feature not captured in the synthetic data.

For estimates of means, the combining rules of Reiter (2003) give an estimate of the added variability due to the synthetic data imputations. Accounting for uncertainty in the estimation of variances is an area for future work. Information provided to users of the synthetic data should include information about what types of analyses are likely to be valid.

3.4 Risk assessment

While utility of the data is of primary interest to researchers, it is equally important that the disclosure risk be adequately addressed. This section discusses issues relevant to risk assessment but does not formally evaluate the risks associated with releasing synthetic data for the LBD. Quantitative measures such as re-identification probabilities are not provided. During the model development stage, the key concern has been that sufficient variability is generated and that the implicates do not look too much like the observed data, so that accurate re-identification will be difficult. A few simple illustrations of the difference between observed and synthetic that can be given without disclosing confidential values are shown. A thorough risk assessment will be conducted prior to releasing data for public use.

The imputation of entry and exit information is an important step in reducing attribute disclosure. A unit may have imputed lifetimes that are completely different across implicates and from the observed lifetime. For example, a unit with an actual lifetime of 1982 to 1987 could have imputed lifetimes of 1975 to 1990, 1983 to 1989, and 1996 to 2001. Similarly, in each implicate, the payroll and employment levels will vary, so that the probability of re-identification should be small. Figure 3.14 shows the distribution of one implicate of synthetic Firstyear for all units in the test sample having true Firstyear of 1995. It can be seen that a wide range of values are imputed, with only a small percent actually imputed to be 1995 in a given implicate.

With repeated imputations, a potential re-identification method for categorical variables is to look across imputations and take the mode. This is a consideration that should be taken into account when conducting a risk assessment and determining the number of implicates to release. Computing the percent of identifications using the mode that are successful is one way to obtain a measure of risk. An additional step that will confound re-identification using the mode is that unit identifiers will be

stripped from the data prior to release and replaced with random identifiers that are similar in structure to the actual identifiers, for use by persons planning to eventually run their analyses on the confidential data. The identifiers will be different across imputations and establishments listed in random order. It is not intended that the analytic validity extend to unit-level analyses, and this will deter users from attempting to average values for units across imputations. Averaging across imputations, or taking the mode for categorical variables, is a potential intrusion method which is likely to result in only a small percentage of true matches, though the number of perceived matches may be higher.

Figures 3.15 and 3.16 show the distribution of observed Firstyear for units that have synthetic Firstyear values of 1995 on one implicate and two implicates, respectively. These figures show that if a user is able to match units across implicates and attempts to re-identify values based on the mode, the repetition of values across implicates will have only a small probability of corresponding to a true value. Matching across three implicates yields a similarly dispersed distribution. These results do not necessarily extend to the rest of the LBD; however, they suggest that matching probabilities should be acceptably low.

3.5 Current status and future plans

At present, the first stage of model development is complete and efforts are underway to extend the imputation to the entire LBD. This task primarily involves identifying and addressing various conditions in the data which have not been accounted for in the program and thus will result in computer errors. In addition, I and other project members are working on parallelization of the code, and considering additional risk assessment that must be done prior to requesting the release of the synthetic data from the disclosure review boards of the U. S. Bureau of the Census and Internal

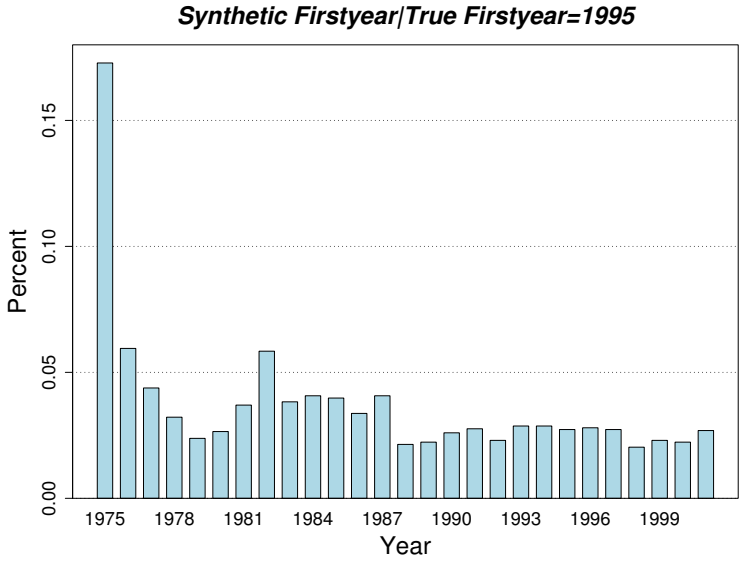

Figure 3.14: Distribution of synthetic Firstyear, one implicate, for units with observed Firstyear = 1995, Group 1

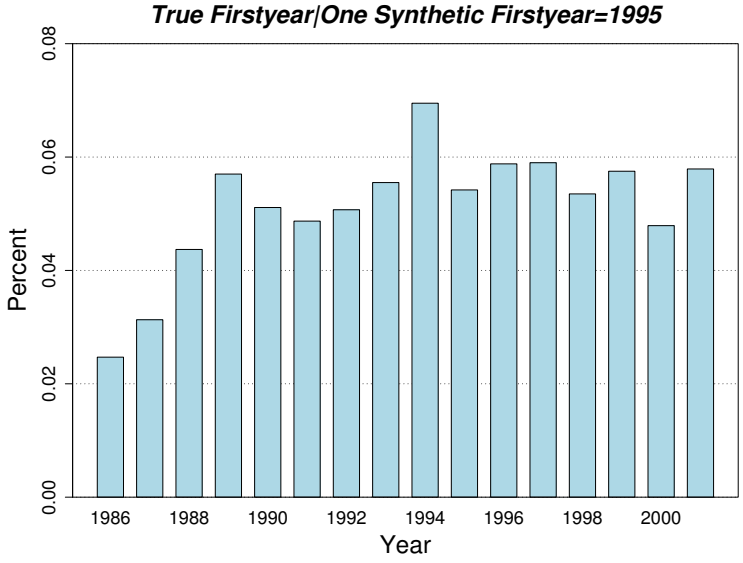

Figure 3.15: Distribution of observed Firstyear for units with synthetic Firstyear=1995 on one implicate, Group 1

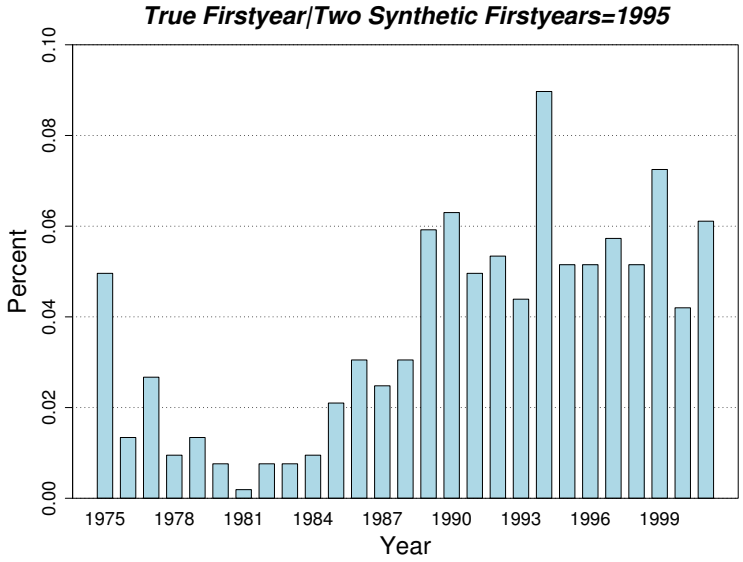

Figure 3.16: Distribution of observed Firstyear for units with synthetic Firstyear=1995 on two implicates, Group 1

Revenue Service.

The data utility will also be evaluated further, both before and after the initial data release. Past synthetic data projects have required several iterations to improve the utility of the data. A "beta" release stage, similar to the current beta release of synthetic data for the Survey of Income and Program Participation, would be helpful in flushing out where further modeling efforts should be focused. This would allow users to gain access to valuable, discloseable information without having to gain access to the confidential data, and allow the project team to gain feedback on the utility of the synthetic data.

3.6 Inferential methods for multiply-imputed population data

This chapter has described ongoing work to develop synthetic data for longitudinal establishment data from the U. S. Census Bureau's Longitudinal Business Database

for public release. Preliminary results illustrate the feasibility of using synthetic data for releasing microdata for public use while protecting confidentiality and allowing valid inferences to be made. The methodology is flexible and can be adapted for other datasets.

The remainder of the chapter discusses modifications to the combining rules for synthetic data of Reiter (2003) and Reiter (2005c), reviewed in Section 1.2. These were developed for random samples from populations. As the LBD is a census and is considered to be population data, there are no sampling errors, and so modified rules are needed. If a population is reasonably believed to be a random sample from a super-population, and the true parameter values are assumed to be determined by the super-population and not the observed population, then the standard combining rules are appropriate.

3.6.1 Inferences for scalar estimands

For a given scalar population parameter q, if D_{inc} is a population dataset then the value of q is given by q_{inc}, the value of the parameter computed from D_{inc}. For partially synthetic data we assume no missing values; hence, $D_{inc} = D_{obs}$ and $q_{inc} = q_{obs}$, where q_{obs} is computed from the observed data D_{obs}, prior to the replacement of any values for disclosure control. The variance u of $(q|D_{obs})$ is thus zero resulting in the posterior distribution $(q|D_{syn}) \sim t_{m-1}(\bar{q}, b/m)$, where $\bar{q}$ and b are defined in (1.1) and (1.3).

Derivation

To derive the combining rule for a scalar estimand q, we return to the derivation of the combining rules for partially synthetic survey data in Reiter (2003) to determine where the derivation differs for population data. The key difference is that

$V(q|D_{obs}, D_{syn}, b_\infty) = 0$, and not $\bar{u}$. Let $b_\infty = V(\hat{q}^{(i)})$ as $m \to \infty$ and observe that since $V(q|D_{obs}, D_{syn}, b_\infty) = 0$, and not $\bar{u}$, it follows that

$$\begin{aligned} V(q|D_{syn}, b_\infty) &= E(V(q|D_{syn}, D_{obs}, b_\infty)|D_{syn}, b_\infty) + V(E(q|D_{obs}, D_{syn}, b_\infty)|D_{syn}, b_\infty) \\ &= 0 + V(q_{obs}|D_{syn}, b_\infty) \\ &= b_\infty/m \end{aligned}$$

and

$$\begin{aligned} E(q|D_{syn}, b_\infty) &= E(E(q|D_{syn}, D_{obs}, b_\infty)|D_{syn}, b_\infty) \\ &= \bar{q}. \end{aligned}$$

From standard multivariate normal theory we have that $\{(m-1)bb_\infty^{-1}|D_{syn}\} \sim \chi^2_{m-1}$, and thus it follows that $(q|D_{syn}) \sim t_{m-1}(\bar{q}, b/m)$. The simplification resulting from setting $\bar{u} = 0$ means that the approximations needed to obtain the degrees of freedom for the posterior distribution in Reiter (2003) are not needed for population data and the degrees of freedom is exact.

3.6.2 Inferences for multivariate estimands

The significance test of Reiter (2005c) for multivariate estimands Q for partially synthetic data, reviewed in Section 1.2, can also be extended to partially synthetic population data. Simply setting setting $\bar{U} = 0$ yields a test statistic of zero, so an additional modification is needed. The proposed test for the hypothesis $H_0 : Q = Q_0$ is conducted by referring the test statistic

$$S_c = \frac{(Q_0 - \bar{Q})'(Q_0 - \bar{Q})}{kr_c}$$

to an $F_{k,k(m-1)}$ distribution, where $r_c = \frac{1}{m}tr(B)/k$, k is the dimension of Q and $\bar{Q}$ and B are defined in (1.4) and (1.6).

Derivation

A key assumption in the hypothesis test for multicomponent estimands proposed by Reiter (2005c), and in similar tests for other applications of multiple imputation, is that B_∞ is proportional to $\bar{U}_\infty$, i.e., the proportion of information replaced with imputations is the same across components of Q. When imputing population data, $\bar{U}_\infty = 0$, so an alternate assumption is needed. A comparable assumption is to assume that $B_\infty = r_\infty I$, for some scalar quantity r_∞ and k-dimensional identity matrix I. In other words, the between-imputation variance is constant across components of Q. This is slightly stronger than the assumption that $B_\infty = r_\infty \bar{U}$ as the covariance matrix is forced to be diagonal. Simulations in Section 3.6.4 suggest that valid inferences can still result when this is not true. Under this assumption the Bayesian p-value is given by

$$\int P(\chi^2_k > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) | D_{syn}, B_\infty) P(B_\infty | D_{syn}) dB_\infty \qquad (3.1)$$

$$= \int P\left(\chi^2_k > \frac{(Q_0 - \bar{Q})' I (Q_0 - \bar{Q})}{r_\infty / m} | D_{syn}, r_\infty\right) P(r_\infty | D_{syn}) dr_\infty$$

$$= \int P\left(\frac{\chi^2_k}{k} \cdot \frac{r_\infty}{m r_c} > S_c | D_{syn}, r_\infty\right) P(r_\infty | D_{syn}) dr_\infty. \qquad (3.2)$$

Thus the proportionality assumption reduces the number of variance parameters to be estimated from $k(k-1)/2$ to 1 and allows for the closed-form approximation of the integral in (3.1). To complete the integration, we need the distribution of $(r_\infty | D_{syn})$. Extending the scalar case in Reiter (2003), the sampling distribution of $\hat{Q}^{(i)}$, the estimate of Q obtained from $D^{(i)}_{syn}$, is given by $(\hat{Q}^{(i)} | Q_{obs}, B_\infty) \sim N(Q_{obs}, B_\infty)$. Under the proportionality assumption, this becomes $(\hat{Q}^{(i)} | Q_{obs}, r_\infty) \sim N(Q_{obs}, r_\infty I)$. With diffuse priors and standard multivariate normal theory for sam-

ple covariance matrices, we obtain

$$(m-1)\frac{\sum_{i=1}^{m}(\hat{Q}^{(i)}-\bar{Q})(\hat{Q}^{(i)}-\bar{Q})'}{(m-1)r_\infty}|D_{syn} \sim Wish(m-1, I).$$

Taking the trace of each side yields

$$\frac{k(m-1)}{r_\infty}\frac{tr(B)}{k}|D_{syn} \sim \chi^2_{k(m-1)}.$$

Hence

$$(r_\infty|D_{syn}) \sim \chi^{-2}_{k(m-1)}km(m-1)r_c$$

Integrating over r_∞ in (3.2) yields a Bayesian p-value of

$$P\left(\frac{\chi^2_k}{k}\frac{k(m-1)}{\chi^2_{k(m-1)}} > S_c|D_{syn}\right) = P(F_{k,k(m-1)} > S_c|D_{syn})$$

3.6.3 Extension to missing data

While imputation for missing values is not done for the LBD, it is worth noting that multiple imputation of missing values in population data requires similar modifications to the combining rules. As we only need to account for the uncertainty due to imputation, and not random sampling, the modification and justification for the scalar combining rules is the same, namely, set $\bar{u} = 0$ in the combining rules for missing data of Rubin (1987). To obtain the test for multicomponent estimands, we apply the same proportionality assumption and set $B_\infty = r_\infty I$.

The posterior for scalar estimand q is given by $(q|D_{com}) \sim N(\bar{q}, (1 + 1/m)b)$, where D_{com} is the set of completed datasets. The corresponding test for multivariate estimands Q is given by $P(F_{k,k(m-1)} > S_q|D_{com})$ where

$$S_q = \frac{(Q_0-\bar{Q})'(Q_0-\bar{Q})}{kr_q},$$

and $r_q = (1 + 1/m)tr(B)/k$. The next section describes results for simulated data in order to illustrate the validity of these formulas, for both missing value imputation and partially synthetic data.

3.6.4 Simulation study

In this section, simple simulation examples illustrate the analytic validity of the proposed combining rules. In addition, the effect on the analytic validity of drawing from the posterior predictive distribution $P(Y|X) = \int P(Y|X,\theta)p(\theta)d\theta$ versus $P(Y|X,\hat{\theta})$ when generating imputations is evaluated. Both cases are evaluated for both partially synthetic data and the missing data cases. Although the missing data case is not needed for the LBD, the results are of interest both on their own and in comparison with the partially synthetic data results.

For a population of size $N = 50,000$, five predictors $X = \{X_1, \ldots, X_5\}$ are generated from a standard normal distribution and a response variable Y is generated from $N(X\beta, 1)$, $\beta = (1, 1, 2, 2, -1)$. For the missing data case, the vector R, where $R_i = 1$ if unit i has missing values, $i = 1, \ldots, N$, is generated from $Bin(1, p)$, where $p = \exp(Xa)/(1 + \exp(Xa))$, $a = (-2, -2, 1, 1, 2)$, so that missingness occurs at random (MAR).

While inferential methods for multiple imputation are derived from a Bayesian perspective, their analytic validity is usually considered from a frequentist one. If we consider (X, Y) to be population data, then the frequentist properties of interest are the repeated sampling properties under repeated sampling of imputations. Using $m = 5$, in each of 5000 iterations, m imputations are drawn and the combining rules applied to the estimands of interest. For the missing data simulations, the response indicator R is also redrawn in each iteration.

The scalar estimands evaluated are $E(Y)$, $\beta_l, l = 1, \ldots, 5$, $P(Y > 1)$, and $E(Y|X >$

1). For each scalar estimand, the confidence intervals computed in each iteration are expected to contain the true value $100(1-\alpha)\%$ of the time, where the significance level $\alpha = .05$. Since (X, Y) represents a population, the true values are computed from (X, Y).

Random sampling scenarios are also simulated for comparison purposes. When (X, Y) are considered to be a random sample of size $n = 50,000$, at each iteration, (X, Y) are generated from the distributions described above, prior to drawing R and generating the m imputations. The true values used to assess the coverage rates are determined from the known population parameters used to generate (X, Y).

The hypothesis test for multiple components is evaluated in a similar simulation scenario. For a population of $N = 50000$, $X = (X_1, \ldots, X_k)$ and Y are drawn from standard normal distributions. Missingness is simulated to be completely at random, with $P(R_l = 1) = .3, l = 1, \ldots, s$. For each of 5000 iterations, m imputations are drawn for $m \in (2, 5, 10)$ and a hypothesis test conducted for $H_0 : Q = 0$, where Q is the vector of regression coefficients, excluding the intercept, of the regression of Y on X and has dimension k, $k \in (2, 5, 20)$. By design, H_0 is true, so H_0 should be rejected $100\alpha\%$ of the time, for significance level $\alpha = .05$.

Partially synthetic data

Let Y be a confidential response variable and X be unreplaced predictors. Then Y_{syn} is generated by drawing independently from the posterior predictive distribution $f(Y|X)$ assuming a normal linear model. Two cases are evaluated, one in which the model parameters β and σ^2 are drawn from their posterior distribution in each imputation and one in which they are fixed at their maximum likelihood estimates. For comparison, the simulations are repeated in a random sampling scenario, using the combining rules for random samples of Reiter (2003) and Reiter (2005c). For the

case $m = 2$, the degrees of freedom in Reiter (2005c) is undefined so the degrees of freedom w_p^* given in Section 1.2.2 is used for conducting the multicomponent test. Such a distinction is not necessary for population data.

Table 3.4 gives the nominal 95% coverage rates for confidence intervals computed for population data using the proposed combining rules, and compares them to the the coverage rates using the combining rules for random samples. Table 3.5 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to the significance level 0.05. From these results it appears that the proposed combining rules for population data perform well in terms of providing valid inferences. Not shown are the coverage rates when the rules from random samples are applied to populations, which were observed to be quite high, typically 1, in the simulations conducted.

From both sets of results, it is seen that using the MLE in place of drawing from the posterior of the regression parameters does not affect analytic validity. Thus it is expected that the decision not to draw parameters in the generation of the LBD will not affect analytic validity. It may seem surprising that drawing imputations from $p(Y|X, \hat{\theta})$ versus $p(Y|X)$ appears to provide valid inferences in the random sampling scenario since population parameters are not assumed to be known; however, Little and Rubin (2002) note for missing data imputation that using the distribution $p(Y_{mis}|Y_{obs}, \hat{\theta})$ can provide a reasonable approximation for $p(Y_{mis}|Y_{obs})$ when the fraction of missing information is small. Extending that principle to the synthetic data case with no missing values, drawing imputations from $p(Y|X, \hat{\theta})$ can still result in approximately proper imputations. For small samples, or other cases where the sampling error is large, improper imputations could result. Since drawing parameters does not appear to have a negative effect on the analytic validity and can improve disclosure control by increasing between-imputation variance, a suggested rule of thumb

is to draw them whenever feasible.

Table 3.4: Comparison of nominal 95% coverage rates for estimands computed from partially synthetic data for population data and random samples, impute with parameters drawn and without

	$E(Y)$	β_1	β_2	β_3	β_4	β_5	$P(Y>1)$	$E(Y\|X>1)$
Population data								
Draws	0.9488	0.9504	0.9394	0.9528	0.9508	0.9508	0.9480	0.9500
No draws	0.9532	0.9476	0.9502	0.9508	0.9474	0.9510	0.9492	0.9496
Random sampling								
Draws	0.9842	0.9526	0.9488	0.9486	0.9486	0.9426	0.9756	0.9842
No draws	0.9866	0.9522	0.9488	0.9488	0.9488	0.9472	0.9884	0.9866

Table 3.5: Comparison of nominal 5% rejection rates for tests using partially synthetic data for population data, imputed with parameters drawn and without

		$k=2$	$k=5$	$k=20$
Population data				
Draws	$m=2$	0.0468	0.0562	0.0524
	$m=5$	0.0494	0.0574	0.0544
	$m=10$	0.0456	0.0512	0.0470
No draws	$m=2$	0.0606	0.0522	0.0550
	$m=5$	0.0456	0.0502	0.0520
	$m=10$	0.0522	0.0534	0.0522
Random sampling				
Draws	$m=2$	0.0668	0.0616	0.0598
	$m=5$	0.0540	0.0522	0.0498
	$m=10$	0.0470	0.0494	0.0488
No draws	$m=2$	0.0548	0.0570	0.0528
	$m=5$	0.0534	0.0520	0.0524
	$m=10$	0.0468	0.0492	0.0518

Missing data

The missing values of Y are imputed from the posterior predictive distribution $f(Y_{obs}|X)$ assuming a normal linear model. As in the partially synthetic data case, two cases are evaluated, one in which the model parameters β and σ^2 are drawn

from their posterior distribution in each step and one in which they are fixed at their maximum likelihood estimates. For comparison, the simulations are repeated in a random sampling scenario, using the combining rules for random samples of Rubin (1987) and Li *et al.* (1991a). For the case $m = 2$, the degrees of freedom used for the multivariate test comes from Rubin (1987); otherwise, the degrees of freedom from Li *et al.* (1991a) is used. No distinction is necessary in the population data case.

Table 3.6 gives the nominal 95% coverage rates for confidence intervals computed for the population data using the proposed combining rules, and compares them to the the coverage rates using the combining rules of Rubin (1987) for random samples. Table 3.7 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to 0.05. From these results it appears that the proposed combining rules for population data provide valid inferences.

Unlike the partially synthetic simulation results, these results indicate that drawing imputations from $f(Y_{mis}|Y_{obs})$ instead of $f(Y_{mis}|Y_{obs}, \hat{\theta})$ is necessary for analytic validity when the fraction of information missing is substantial. When population data contain missing values, $\beta_{obs} \neq \beta_{inc}$ and the additional uncertainty in the estimation of the maximum likelihood estimate is unaccounted for when the parameters are not drawn, resulting in improper imputations. This result is well known for imputation of random samples (Rubin, 1987; Little and Rubin, 2002), and these results confirm the same holds true for population data.

Robustness

The combining rules proposed for use with multiply-imputed population data were derived based on a simple extension from the existing combining rules for random samples. Given the success of similar tests, it is expected that the analytic validity of

Table 3.6: Comparison of nominal 95% coverage rates for estimands computed from completed population data and random samples, impute with parameters drawn and without, for the missing data case

	$E(Y)$	β_1	β_2	β_3	β_4	β_5	$P(Y>1)$	$E(Y\|X>1)$
Population data								
Draws	0.9334	0.9584	0.9432	0.9736	0.9758	0.8416	0.9492	0.9240
No draws	0.6134	0.8410	0.7820	0.9056	0.9356	0.5082	0.8738	0.8778
Random sampling								
Draws	0.9480	0.9498	0.9510	0.9486	0.9498	0.9470	0.9554	0.9572
No draws	0.9458	1	1	1	1	1	0.8726	0.9554

the combining rules will hold when the imputations are proper in the sense of Rubin (1987) and that the multicomponent test will be robust to moderate violations of the assumption of proportionality used. Further work is warranted to make a definitive assessment.

The robustness of the multicomponent test to the assumption of independence between components of a multivariate estimand Q, not assumed in the corresponding test for random samples, is evaluated in a simulation test. Using the same simulation scenarios for Tables 3.5 and 3.7, X is drawn from $N(0, .5I)$, so that the assumption of independence between components of Q is no longer satisfied. The nominal 5% rejection rates for the test $H_0 : Q = Q_0$ where Q_0 is true are computed for both the missing data and synthetic data cases, where $m = 5$ and parameters have been drawn from their posterior distributions. The results, shown in Table 3.8 show the rejection rates to be quite close to .05 for $k \in \{2, 5, 20\}$.

The tests are proposed for the case where sampling error is not present; however, other sources of error may be present and should be addressed if necessary. Another area of future work is to extend the combining rules to imputation of population data in other applications, such as two-stage multiple imputation.

Table 3.7: Comparison of nominal 5% rejection rates for tests using completed population data, impute with parameters drawn and without

		$k = 2$	$k = 5$	$k = 20$
Population data				
Draws	$m = 2$	0.0490	0.0430	0.0538
	$m = 5$	0.0460	0.0486	0.0522
	$m = 10$	0.0524	0.0522	0.0488
No draws	$m = 2$	0.0642	0.0776	0.1288
	$m = 5$	0.0898	0.1214	0.2362
	$m = 10$	0.1022	0.1378	0.2874
Random sampling				
Draws	$m = 2$	0.0608	0.0560	0.0530
	$m = 5$	0.0556	0.0520	0.0524
	$m = 10$	0.0480	0.0496	0.0508
No draws	$m = 2$	0.0666	0.0782	0.0910
	$m = 5$	0.0652	0.0696	0.0992
	$m = 10$	0.0636	0.0716	0.0990

Table 3.8: Nominal 5% rejection rates for tests with correlated data

	$k = 2$	$k = 5$	$k = 20$
Missing data	0.0646	0.0494	0.0538
Synthetic data	0.0644	0.0508	0.0554

Chapter 4

Bayesian model uncertainty in mixed effects models

Often in linear and nonlinear mixed effects models, random effects are chosen to control for specific factors which are expected to cause random variation in the coefficients, such as batch effects and within-subject variation in repeated measurements. Choosing a subset of predictors from a larger set of potential predictors is often desired to succinctly express the relationship between the response and predictors and identify the important predictors (Mitchell and Beauchamp, 1988). It is a more difficult question how to decide which predictors have coefficients that vary among subjects. Standard model selection criteria and test procedures are not appropriate for comparing models with different numbers of random effects due to constraints on the parameter space of the variance components. For example, in the model selection context one typically is interested in testing the hypothesis that the variance component is zero, a boundary condition.

A challenge in using likelihood approaches for the estimation of mixed effect models, and hence in model selection, is that the likelihood cannot be computed analytically. Several approximation methods have been developed. Sinharay and Stern (2001) summarize the major approaches, including marginal maximum likelihood, restricted maximum likelihood, and quasilikelihood. The marginal maximum likelihood approach evaluates the likelihood using quadrature or a Laplace approximation and computes maximum likelihood estimates of model parameters using traditional numeric optimization approaches. This approach tends to underestimate variance parameters; hence, Stiratelli *et al.* (1984) suggest an approximate E-M algorithm for

computing the restricted maximum likelihood estimate of the variance matrix. Another alternative from Breslow and Clayton (1993) is the quasilikelihood method, in which a Laplace approximation is used to integrate out the random effects.

Using these approaches for approximating the likelihood, likelihood ratio test statistics can be computed; however, the problem with testing of variance parameters on the boundary of their support remains. Potentially one could get around this problem by using a parametric bootstrap to simulate values from the null distribution of the likelihood ratio statistics (Sinharay and Stern, 2001). Several other frequentist tests for homegeneity of variance components have been proposed. In the setting of linear mixed models with one variance component, Crainiceanu and Ruppert (2004) derived finite and asymptotic distributions of likelihood ratio test statistics. Such results are not yet available for logistic mixed models. Lin (1997) proposed a score test for whether all variance components in generalized linear mixed model (GLMM) are zero (See also Verbeke and Molenberghs (2003) and Hall and Praestgaard (2001)); however, such methods cannot be used for general comparisons of models with different numbers of random effects. Jiang *et al.* (2006) recently proposed an innovative "fence" method to select predictors with random effects in linear mixed models; however, this approach does not allow inferences on whether a given predictor has a random component, and uncertainty in the model selection process is not accounted for.

Given the practical difficulties that arise in implementing a frequentist approach to this problem, we focus on Bayesian methods. An advantage of Bayesian model selection is that one can account for the uncertainty in the selection process, hence the term, "model uncertainty." In the Bayesian approach one bases inferences on estimates of the exact posterior distribution obtained using an MCMC algorithm, with the estimation accuracy improving with the number of MCMC iterations. An-

other advantage is the greater flexibility in comparing models with differing numbers of variance components. Potential disadvantages include computational burden and sensitivity to the prior.

This chapter discusses the selection of fixed and random effects in mixed effects models. Bayesian model uncertainty is reviewed in Section 1, and in the context of mixed models in Section 2. Section 3 describes a Bayesian approach for linear mixed models and discusses prior specification. A modification for binary logistic models in Section 4. Section 5 provides a simulation example and Section 6 a data example. Additional extensions are discussed in Section 7 and concluding remarks are given in Section 8.

4.1 Bayesian Model Uncertainty

Let us first consider a normal linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\epsilon, \epsilon \sim N(0, \sigma^2)$, with no random effects. From the Bayesian perspective, the model parameters are considered random variables with probability distributions; hence, uncertainty is expressed in terms of probability. When fitting the model, prior distributions are assigned to each parameter and then the posterior distributions induced are used to make inferences about the parameters of interest. Often the posterior distributions are obtained numerically using MCMC methods and then inferences are made by computing summaries of these distributions, such as posterior means and probability intervals.

In the Bayesian framework model uncertainty can be addressed simultaneously with parameter uncertainty by placing priors $p(M_k)$ on each possible model $M_1, \ldots, M_K$ in addition to the model parameters $p(\boldsymbol{\beta}|M_k, \sigma^2)$ and $p(\sigma^2)$. The posterior model probabilities are determined by

$$p(M_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_k)p(M_k)}{\sum_k p(\boldsymbol{y}|M_k)p(M_k)},$$

where

$$p(\boldsymbol{y}|M) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, M)p(\boldsymbol{\beta}|\sigma^2, M)p(\sigma^2)d\boldsymbol{\beta}, d\sigma^2$$

is the marginal likelihood of M_k. This marginal likelihood is available analytically for normal linear regression models when conjugate normal inverse-gamma priors are chosen for $(\boldsymbol{\beta}, \sigma^2)$; however, in generalized linear models and in normal linear models with random effects, the marginal likelihood will not be available analytically. In such cases, it is common to rely on the Laplace approximation, or to use simulation-based approaches to approximate the marginal likelihood and/or posterior model probabilities. The model posteriors $p(M_1|\boldsymbol{y}), \ldots, p(M_K|\boldsymbol{y})$ can be used to select one or more models having high posterior probability and provide a measure of the model uncertainty.

A challenge in model selection is that the number of models 2^p increases rapidly with p. Efficient MCMC algorithms such as Gibbs sampling (Gelfand and Smith, 1990) are used to rapidly search for models with high posterior probability, so that every model need not be visited. George and McCulloch (1997) propose a promising Bayesian approach for subset selection called stochastic search variable selection (SSVS). This method uses a Gibbs sampler to 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 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.

4.1.1 Bayes factors

When proper priors are used, posterior model probabilities can be used to compare two models by computing the posterior odds, or Bayes factor

$$\frac{p(M_1|\boldsymbol{y})}{p(M_0|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_0)} \times \frac{p(M_1)}{p(M_0)}.$$

Often the prior odds ratio is taken to be 1, representing no preference for either model, and hence the Bayes factor is the ratio $p(\boldsymbol{y}|M_1)/p(\boldsymbol{y}|M_0)$. Bayes factors are commonly used to summarize the relative evidence provided by the data in support of one model against another. Their interpretation as posterior odds tends to be more intuitive than that of the p-value. They can also be used to quantify evidence for a null hypothesis, an important distinction from failing to find evidence against it. Bayes factors also have the advantage over p-values of allowing multiple hypotheses to be simultaneously compared and can be used to compare non-nested models with differing numbers of terms.

Approximations to the Bayes factor such as the Bayesian Information Criterion (BIC) are popular for model comparisons given their relative ease of computation. They are also appealing in that they provide a default approach for approximating posterior model probabilities that does not require the specification of the prior distribution. While the BIC provides a good approximation in a wide range of problems (Raftery, 1995), it is not appropriate for some common models, including hierarchical

models (Pauler *et al.*, 1999), and models where the number of parameters increases with sample size or other irregular asymptotics occur (Berger *et al.*, 2003).

Bayes factors are sometimes criticized for their sensitivity to the prior specification and inability to handle improper priors. It is important to either choose an informative prior based on subject matter knowledge or to choose a proper default prior, chosen to yield good Bayesian and/or frequentist properties. In subset selection for normal linear regression models, the Zellner-Siow prio (Zellner and Siow, 1980) is a commonly-used default, with recent work proposing alternative mixtures of g-priors (Liang *et al.*, 2005). Several alternative approaches for obtaining default approximations have been developed to address this including local Bayes factors (Smith and Spiegelhalter, 1980), expected posterior prior (Perez and Berger, 2000), intrinsic Bayes factors (Berger and Pericchi, 1996), and fractional Bayes factors (O'Hagan, 1995). Berger and Pericchi (2001) compare several of these approaches with the BIC for conventional linear models. These approaches have wide applicability, particularly in non-nested models or where conventional prior distributions are unavailable (Clyde and George, 2004).

4.2 Approaches for mixed effects models

While a great body of work exists on Bayesian model selection for fixed effects, there is very little work on selection of random effects. Pauler *et al.* (1999) compare variance component models using Bayes factors and Sinharay and Stern (2001) consider the problem of comparing two GLMMs using the Bayes factor. Motivated by sensitivity to the choice of prior, Chung and Dey (2002) develop an intrinsic Bayes factor approach for balanced variance component models. Chen and Dunson (2003) developed a more general stochastic search variable selection (SSVS) (George and McCulloch, 1993; Geweke, 1996) approach to the linear mixed effects model. Rely-

ing on Taylor series approximators to intractable integrals, Cai and Dunson (2006) recently extended this approach to all GLMMs.

4.2.1 Bayes factor approximations

The BIC is not appropriate for comparing models with differing numbers of random effects as the required regularity conditions are not met when the parameter lies on the boundary (Pauler *et al.*, 1999). Several Bayes factor approximations for testing variance components are reviewed in Sinharay and Stern (2001). Most of these involve estimation of $p(\boldsymbol{y}|M_1)$ and $p(\boldsymbol{y}|M_0)$ to obtain the Bayes factor. A modification to the Laplace approximation used to obtain the BIC (Raftery, 1995) which accomodates the boundary case is proposed by Pauler *et al.* (1999). As $p(\boldsymbol{y}|M)$ is an integral, numerical approaches for integral evaluation are available, such as quadrature and importance sampling.

A practical issue with importance sampling is the selection of the target distribution. Meng and Wong (1996) extend the importance sampler idea and suggest a bridge sampling approach for approximating $p(\boldsymbol{y}|M)$. An MCMC algorithm using Gibbs sampling was developed by Chib (1995). A harmonic estimator, consistent for simulations though otherwise unstable, is proposed by Newton and Raftery (1994). Lastly, an approach suggested by Green (1995) is described which computes the Bayes factors directly using a reversible-jump MCMC algorithm which can move between models with parameter spaces of differing dimension. This is likely to be computationally intensive, and in Sinharay and Stern (2001) indeed it was the slowest approach, whereas the Laplace approximation was the fastest.

4.2.2 Stochastic search variable selection

In extending Bayesian model selection procedures for linear models to linear mixed effects models the two primary considerations are the prior specification and posterior computation. The structure of the random effects covariance matrix needs to be considered, and the model parameterizations and prior structure carefully chosen so that the MCMC algorithm may move between models with both differing fixed effects and random effects. The efficiency of the posterior computation also needs to be considered; algorithms that explore the model space efficiently and quickly locate areas of high posterior probability are needed.

As described in Section 4.1, stochastic search variable selection (SSVS) is a promising approach for Bayesian model uncertainty using Gibbs sampling. The SSVS approach has been applied successfully in a wide variety of regression applications, including challenging gene selection problems. One challenge in developing SSVS approaches for random effects models is the constraint that the random effects covariance matrix $\mathbf{\Omega}$ be positive semi-definite. Chen and Dunson (2003) addressed this problem by using a modified Cholesky decomposition of $\mathbf{\Omega}$:

$$\mathbf{\Omega} = \mathbf{\Lambda\Gamma\Gamma'\Lambda}, \tag{4.1}$$

where $\mathbf{\Lambda}$ is a positive diagonal matrix with diagonal elements $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)'$ proportional to the random effects standard deviations, so that setting $\lambda_l = 0$ is equivalent to dropping the lth random effect from the model. $\mathbf{\Gamma}$ is a lower triangular matrix with diagonal elements equal to 1 and free elements that describe the random effects correlations. In the case of independent random effects, $\mathbf{\Gamma}$ is simply the identity matrix $\boldsymbol{I}$ and the diagonal elements $\lambda_l, l = 1, \ldots, q$ of $\mathbf{\Lambda}$ equal the random effects standard deviations.

In the next section, we revisit the SSVS approach of Chen and Dunson (2003) for linear mixed models, with additional consideration given to the prior structure and posterior computation. We will then discuss an extension to logistic models.

4.3 Linear mixed models

If we have n subjects under study, each with n_i observations, $i = 1, \ldots, n$, let y_{ij} denote the jth response for subject i, $\boldsymbol{X}_{ij}$ a $p \times 1$ vector of predictors, and $\boldsymbol{Z}_{ij}$ a $q \times 1$ vector of predictors. Then the linear mixed effects (LME) model is denoted as

$$y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{a}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{4.2}$$

where $\boldsymbol{a}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$. Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the fixed effects and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iq})'$ are the random effects. In practice $\boldsymbol{Z}_{ij}$ is typically chosen to be a subset of the predictors in $\boldsymbol{X}_{ij}$ believed to have random effects, often only the intercept for simplicity. If we let $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ include all candidate predictors, then the problem of interest is to locate a subset of these predictors to be included in the model.

Using the covariance decomposition in (4.1) so we can use SSVS, we write (4.2) as

$$y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{4.3}$$

where $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$. Chen and Dunson (2003) show that by rearranging terms, the diagonal elements, $\lambda_l, l = 1, \ldots, q$, of $\boldsymbol{\Lambda}$ can be expressed as linear regression coefficients, conditional on $\boldsymbol{\Gamma}$ and $\boldsymbol{b}_i$. Similarly, the free elements $\gamma_k, k = 1, \ldots, q(q-1)/2$, of $\boldsymbol{\Gamma}$ can be expressed as linear regression coefficients, conditional on $\boldsymbol{\Lambda}$ and $\boldsymbol{b}_i$. Hence the variance parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ have desirable conditional conjugacy

properties for constructing a Gibbs sampling algorithm for sampling the posterior distribution and we are able to use the SSVS approach.

4.3.1 Priors

Prior selection is a key step in any Bayesian analysis; however, in this context it is particularly important as problems can arise when default priors are applied without caution. In particular, flat or excessively diffuse priors are not recommended for hierarchical models given the potential for an improper posterior and the difficulty of verifying propriety due to the intractable nature of the density, even when the output from a Gibbs chain seems reasonable (Hobert and Casella, 1996). Proper distributions are also desired for Bayes factors to be well-defined (Pauler *et al.*, 1999). The arbitrary multiplicative constants from improper priors carry over to the marginal likelihood $p(\boldsymbol{y}|M)$ resulting in indeterminate model probabilities and Bayes factors (Berger and Pericchi, 2001).

A mixture of a point mass at zero and a normal or heavier-tailed distribution is a common choice of prior for fixed effects coefficients, $\beta_l, l = 1, \ldots, p$, in Bayesian model selection problems. Smith and Kohn (1996) introduce a vector $\boldsymbol{J}$ of indicator variables, where $J_l = 1$, indicates that the lth variable is in the model, $l = 1, \ldots, p$, and assign a Zellner g prior (Zellner and Siow, 1980) to $\boldsymbol{\beta_J}$, the vector of coefficients in the current model. As a notational convention, we let $\boldsymbol{\beta}$ denote the $p \times 1$ vector $(\{\beta_l : J_l = 1\} = \boldsymbol{\beta_J}, \{\beta_l : J_l = 0\} = \mathbf{0})$. Hence, conditional on the model index $\boldsymbol{J}$, the prior for $\boldsymbol{\beta}$ is induced through the prior for $\boldsymbol{\beta_J}$.

Consistency issues can arise when comparing models based on these priors; however, for linear models, placing a conjugate gamma prior on g induces a t prior on the coefficients. In the special case where the t distribution has degrees of freedom equal 1, the Cauchy distribution is induced, which has been recommended for Bayesian

robustness (Clyde and George, 2004). This can be considered a special case of mixtures of g-priors, proposed by Liang *et al.* (2005) as an attractive computational solution to the consistency and robustness issues with g-priors, and an alternative to the Cauchy prior, which does not yield a closed-form expression for the marginal likelihood. As choosing g can affect model selection, with large values concentrating the prior on small models with a few large coefficients and small values of g concentrating the prior on saturated models with small coefficients, several approaches for handling g have been proposed (Liang *et al.*, 2005). Recommendations include the unit information prior (Kass and Wasserman, 1995), which in the normal regression case corresponds to choosing $g = n$, leading to Bayes factors that behave like the BIC and the hyper-g prior of Liang *et al.* (2005). Foster and George (1994) recommend calibrating the prior based on the risk inflation criterion (RIC) and Fernandez *et al.* (2001) recommend a combination of the unit information prior and RIC approach. Another alternative is a local empirical Bayes approach, which can be viewed as estimating a separate g for each model, or global empirical Bayes, which assumes a common g but borrows strength from all models (Liang *et al.*, 2005).

For standard deviation parameters in hierarchical models, Gelman (2005) recommends a family of folded-t prior distributions over the commonly used inverse gamma family, due to their flexibility and behavior when random effects are very small. These priors are induced using a parameter-expansion approach which has the added benefit of improving computational efficiency by reducing dependence among the parameters (Liu *et al.*, 1998; Liu and Wu, 1999). This yields a Gibbs sampler less prone to slow mixing when the standard deviations are near zero. The Chen and Dunson (2003) approach had the disadvantages of (i) relying on subjective priors that are difficult to elicit, and (ii) computational inefficiency due to slow mixing of the Gibbs sampler; hence a parameter-expanded model is used to address these two

problems.

Extending the parameter expansion approach proposed by Gelman (2005) for simple variance component models to the LME model, (4.3) is replaced with:

$$y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{4.4}$$

where $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{D})$ and $\boldsymbol{A} = \text{diag}(\alpha_1, \ldots, \alpha_q)'$ and $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_q)'$ are diagonal matrices, $\alpha_l \sim N(0,1), l = 1, \ldots, q$, and $d_l \sim IG(\frac{1}{2}, \frac{N}{2}), l = 1, \ldots, q$, IG denoting the inverse gamma distribution. Note that the latent random effects have been multiplied by a redundant multiplicative parameter. In this case the implied covariance decomposition is $\boldsymbol{\Omega} = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\Gamma}'\boldsymbol{A}$.

The parameters $\alpha_l, l = 1, \ldots, q$, are proportional to λ_l and thus to the random effects standard deviations, so setting $\alpha_l = 0$ effectively drops out the random effects for the lth predictor. When random effects are assumed to be uncorrelated, i.e., $\Gamma = I$ and $\lambda_l, l = 1, \ldots, q$ equal the random effects standard deviations, a folded t prior on $\lambda_l = |\alpha_l|\sqrt{d_l}, l = 1, \ldots, q$ is induced, as described in Gelman (2005). Generalizing to the case of correlated random effects, a folded-t prior is not induced; however, improved computational efficiency is still achieved, as illustrated in Section 4.5.

In the proposed prior structure a Zellner-type prior is used for the fixed effects components. Specifically, $\boldsymbol{\beta_J} \sim N\left(\mathbf{0}, \sigma^2(\mathbf{X^{J}}'\mathbf{X^{J}})^{-1}/g\right)$, $g \sim G(\frac{1}{2}, \frac{N}{2})$, $\sigma^2 \propto \frac{1}{\sigma^2}$, and $J_l \sim Be(p_0), l = 1, \ldots, p$, with Be denoting the Bernoulli distribution and $G(a, b)$ denoting the Gamma distribution with mean a/b and variance a/b^2. Let $\alpha_l, l = 1, \ldots, q$, have a zero-inflated half-normal prior, $ZI - N^+(0, 1, p_{l0})$, where p_{l0} is the prior probability that $\alpha_l = 0$. Lastly, the free elements of $\boldsymbol{\Gamma}$ are treated as a $q(q-1)/2$-vector with prior $p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = N(\boldsymbol{\gamma}_0, \boldsymbol{V}_\gamma) \cdot 1(\boldsymbol{\gamma} \in \boldsymbol{R_\alpha})$ where $\boldsymbol{R_\alpha}$ constrains elements of $\boldsymbol{\gamma}$ to be zero when the corresponding random effects are zero. For simplicity, uncertainty in which random effects are correlated is not allowed.

4.3.2 Posterior computation

The joint posterior distribution for $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$ is given by:

$$p(\boldsymbol{\theta}|y) \propto \prod_{i=1}^{n} N_p(\boldsymbol{\xi}_i; \mathbf{0}, \boldsymbol{D}) \prod_{j=1}^{n_i} \left\{ N(y_{ij}; \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i, \sigma^2) \right\}$$
$$\times p(\sigma^2) p(\boldsymbol{\beta}, \boldsymbol{J}, g) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) p(\boldsymbol{D}) \tag{4.5}$$

This distribution has a complex form which we cannot sample from directly; instead we employ a parameter-expanded Gibbs sampler (Liu *et al.*, 1998; Liu and Wu, 1999). The Gibbs sampler proceeds by iteratively sampling from the full conditional distributions of all parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2$, hyperparameters g and $\boldsymbol{J}$, and the diagonal elements $d_l, l = 1, \ldots, q$ of $\boldsymbol{D}$.

The full conditional posterior distributions are given in Appendix A and follow from (4.5) using straightforward algebraic routes. After discarding draws from the burn-in period, posterior summaries of model parameters can be estimated the usual way from the Gibbs sampler output. Models with high posterior probability can be identified as those appearing most often in the output and considered for further evaluation. Marginal inclusion probabilities for a given coefficient may also be calculated using the proportion of draws in which the coefficient is nonzero.

4.4 Binary Logistic Mixed Models

Logistic mixed models are widely used, flexible models for unbalanced repeated measures data. The proposed approach for logistic mixed models is to formulate the model in such a way that its coefficients are conditionally linear and the SSVS approach can again be applied. This entails the use of a data augmentation strategy and approximation of the logistic density, with approximation error corrected for using

importance weights. The covariance decomposition in (4.1) and parameter expansion approach described in Section 4.3.1 are again used.

Defining terms as in (4.3), the logistic mixed model for a binary response variable y is written as:

$$\text{logit}(P(y_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{a}_i)) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{a}_i, \quad \boldsymbol{a}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{4.6}$$

We would like to be able to apply the SSVS approach as in the normal case. If we apply the covariance decomposition in (4.1) to the logistic mixed model, we have:

$$\text{logit}(P(y_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{b}_i)) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{I}) \tag{4.7}$$

In this case the model is nonlinear and we do not immediately have conditional linearity for the variance parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ as in the normal case. In order to obtain conditional linearity for the model coefficients, we take advantage of the fact that the logistic distribution can be closely approximated by the t distribution (Albert and Chib, 1993; Holmes and Knorr-Held, 2003; O'Brien and Dunson, 2004), and that the t distribution can be expressed as a scale mixture of normals (West, 1987).

First, note that (4.7) is equivalent to the specification:

$$y_{ij} \quad = \quad \begin{cases} 1 & w_{ij} > 0 \\ 0 & w_{ij} \leq 0 \end{cases},$$

where w_{ij} is a logistically distributed random variable with location parameter $\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i$ and density function

$$\mathcal{L}(w_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \frac{\exp\{-(w_{ij} - \boldsymbol{X}'_{ij}\boldsymbol{\beta} - \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i)\}}{\{1 + \exp[-(w_{ij} - \boldsymbol{X}'_{ij}\boldsymbol{\beta} - \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i)]\}^2}.$$

Then, as w_{ij} is approximately distributed as a non-central t_ν with location parameter $\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{b}_i$ and scale parameter $\tilde{\sigma}^2$, we can express it as a scale mixture

of normals and write:

$$w_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{\Lambda\Gamma b}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \tilde{\sigma}^2/\phi_{ij}) \tag{4.8}$$

where $\phi_{ij} \sim G(\frac{\nu}{2}, \frac{\nu}{2})$. Setting $\nu = 7.3$ and $\tilde{\sigma}^2 = \pi^2(\nu-2)/3\nu$ makes the approximation nearly exact. The approximation error, though negligible except in the extreme tails, may be corrected for by importance weighting when making inferences. Under this model formulation, we have a model in which all coefficients are conditionally normal, and we are able to apply SSVS to the problem. We also are able to take advantage of the improved computational efficiency of a parameter expanded model as in (4.4). Applying the parameter expansion to (4.8) we have:

$$w_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{A\Gamma\xi}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \tilde{\sigma}^2/\phi_{ij})$$

where terms are defined as in (4.4) and (4.8). We will use this model formulation to propose a prior structure and compute posterior distributions.

4.4.1 Priors and posterior computation

We use the same priors for the random effects parameters as in the normal case, and similar priors for the fixed effects parameters. We specify $\boldsymbol{\beta_J} \sim N\left(0, (\mathbf{X^{J'}X^J})^{-1}/g\right)$, $g \sim G(\frac{1}{2}, \frac{N}{2})$, and $J_l \sim Be(p_0), l = 1, \ldots, p$. Using the t-distribution to approximate the likelihood as previously described, the joint posterior distribution for $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi})$ is given by:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{J}, g)p(\boldsymbol{\gamma}, \boldsymbol{\alpha})p(\boldsymbol{D}) \left(\prod_{i=1}^{n} N_q(\boldsymbol{\xi}_i; \mathbf{0}, \boldsymbol{D}) \right.$$
$$\left. \times \prod_{j=1}^{n_i} \left[N\left(w_{ij}; \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{A\Gamma\xi}_i, \frac{\tilde{\sigma}^2}{\phi_{ij}} \right) \{1(w_{ij} > 0)y_{ij} + 1(w_{ij} \leq 0)(1 - y_{ij})\} p(\phi_{ij}) \right] \right). \tag{4.9}$$

Again we have a complex posterior from which we cannot directly sample and we employ a Gibbs sampler. In introducing a latent variable w_{ij} we have applied a data augmentation strategy related to Albert and Chib (1993) and used for multivariate logistic models by O'Brien and Dunson (2004). This auxiliary variable is updated in the Gibbs sampler and its full conditional posterior follows immediately from (4.9) as a normal distribution truncated above or below by 0 depending on y_{ij}:

$$
\begin{aligned}
&p(w_{ij}|\boldsymbol{\theta}, y_{ij}) \\
&= \frac{N\left(w_{ij}; \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i, \frac{\tilde{\sigma}^2}{\phi_{ij}}\right) \cdot 1\left((-1)^{y_{ij}} w_{ij} < 0\right)}{\Phi\left(0; \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i, \frac{\tilde{\sigma}^2}{\phi_{ij}}\right)^{1-y_{ij}} \left\{1 - \Phi\left(0; \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i, \frac{\tilde{\sigma}^2}{\phi_{ij}}\right)\right\}^{y_{ij}}}
\end{aligned} \tag{4.10}
$$

where $\Phi(\cdot)$ indicates the normal cumulative distribution function. The Gibbs sampler proceeds by iteratively sampling from the full conditional distributions of all parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\phi}$, hyperparameters g and $\boldsymbol{J}$, as well as the latent variable $\boldsymbol{\xi}_i, i = 1, \ldots, n$ and the diagonal elements $d_l, l = 1, \ldots, q$ of $\boldsymbol{D}$. The full conditional posterior distributions follow from (4.9). They are similar in form to the normal case and are given in Appendix A.

This Gibbs sampler generates samples from an approximate posterior as we have approximated the logistic likelihood in (4.8). To correct for this, importance weights (Hastings, 1970) may be applied when computing posterior summaries to obtain exact inferences. If we have M iterations of the Gibbs sampler, excluding the burn-in interval, then the importance weights $r^{(t)}, t = 1, \ldots, M$ can be computed as:

$$
r^{(t)} = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{\mathcal{L}(w_{ij}; \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i)}{\mathcal{T}_\nu(w_{ij}; \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i, \tilde{\sigma}^2)}
$$

where $\mathcal{L}(\cdot)$ is the logistic density function and $\mathcal{T}_\nu(\cdot)$ is the t density function with degrees of freedom ν.

Posterior means, probabilities, and other summaries of the model parameters can be estimated from the Gibbs sampler output using an importance-weighted sample average. For example, the posterior probability for a given model m is the sum of the weights corresponding to each occurence of model m in the posterior sample, divided by the sum of all M weights. The approximation is very close and hence the weights are close to one. In the simulation and data examples I found very little difference between weighted and unweighted results.

In lieu of approximating the logistic distribution with the t distribution, the slice sampler for sampling from the exact posterior distribution as applied by Gerlach *et al.* (2002) to variable selection for logistic models was also considered. In this approach, the model is considered linear with response variable $v_{ij} = \text{logit}(p(y_{ij} = 1))$, the vector of log odds, and $v_{ij} = \text{logit}(p(y_{ij} = 1)) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\Lambda\Gamma\boldsymbol{b}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$. The vector $v_{ij}$is updated in a data-augmented Gibbs sampler where an auxiliary variable $u_{ij} \sim U\left(0, \frac{1}{1+\exp(v_{ij})}\right)$ is introduced so that the full conditional posterior distribution for v_{ij} is simplified to a truncated normal distribution as follows:

$$\begin{aligned}
p(v_{ij}|y_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma^2) &\propto p(y_{ij}|v_{ij}) \cdot p(v_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma^2) \\
&\propto \left(\frac{e^{v_{ij}y_{ij}}}{1 + e^{v_{ij}}}\right) \cdot N(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\Gamma\boldsymbol{\xi}_i, \sigma^2) \\
p(v_{ij}|u_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma^2) &\propto p(u_{ij}|v_{ij})p(v_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma^2) \\
&\propto N(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{A}\Gamma\boldsymbol{\xi}_i + \sigma^2 y_{ij}, \sigma^2) \\
&\quad \times 1\left(v_{ij} < \log\left(\frac{1 - u_{ij}}{u_{ij}}\right)\right)
\end{aligned}$$

While slice sampling in general has been noted to have appealing theoretical properties (Neal, 2000; Mira and Tierney, 2002), it demonstrated unsatisfactory convergence properties due to asymmetries induced by the likelihood (Green, 1997). In simulations using the slice sampler approach, the correct models were quickly lo-

cated; however, the Gibbs chains for nonzero model coefficients were extremely slow to converge.

4.5 Simulation Examples

This section evaluates the proposed approach using a simulation example for a binary response logistic model. Three covariates are generated from $U(-2, 2)$ for 30 observations on each of 200 subjects, so $\boldsymbol{X}_{ij} = (1, X_{ij1}, X_{ij2}, X_{ij3})'$. Let $\boldsymbol{Z}_{ij} = \boldsymbol{X}_{ij}$, $\boldsymbol{\beta} = (1, 0, 1, 1)'$ and $\alpha_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$, with a range of realistic values chosen for the random effects variances:

$$\boldsymbol{\Omega} = \begin{pmatrix} .90 & .48 & .06 & 0 \\ .48 & .40 & .10 & 0 \\ .06 & .10 & .10 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The response $\text{logit}(P(y_{ij} = 1))$ is generated according to model (4.6) and then y_{ij} is drawn from $\text{Be}(p(y_{ij}))$. Following the prior specification outlined in Section 4.1, heavy-tailed priors are induced on the fixed effects coefficients and random effects variances. These default priors do not require subjective choice of hyperparameter values, with the exception of the prior inclusion probabilities, which can be chosen as $p = 0.5$ to give equal probability to inclusion and exclusion, and the prior mean and variance of $\boldsymbol{\gamma}$. This prior specification does include an informative normal prior for $\boldsymbol{\gamma}$; however, $\boldsymbol{\gamma}$ is scaled in the parameter-expanded model and hence an informative prior can reasonably be chosen. A prior that modestly shrinks the correlations towards zero is desirable for stable estimation while still allowing the data to inform the relationships between the random effects. As a reasonable choice, the prior mean and variance for $\boldsymbol{\gamma}$ are chosen to be $\mathbf{0}$ and $0.5\boldsymbol{I}$, which can be used as a default in other applications.

The Gibbs sampler is run for 20,000 iterations, after a burnin period of 5000 iterations. Three chains with dispersed starting values were run and found to converge after a few thousand iterations. The resulting MCMC chains for the random effects variances are shown in Figure 4.1 and the posterior means for the fixed effects coefficients and random effects variances are given in Table 4.1, along with the PQL estimates computed by glmmPQL in R.

The simulation results are compared to the penalized quasi-likelihood (PQL) approach (Breslow and Clayton, 1993), as this approach is widely used for estimating GLMMs. Although the focus of this chapter is on selection and inferences in the variance components allowing for model uncertainty, which is not addressed by current frequentist methods, model-averaged coefficient estimates are also obtained. Based on the limited number of simulations run, these estimates tend to be less biased, or closer to the true values than the PQL estimates, which are also known to be biased (Breslow, 2003; Jang and Lim, 2005). The MCMC algorithm is too computationally intense to run a large enough simulation to definitively assess the frequentist operating characteristics of the proposed approach.

Credible intervals for the random effects variances can also be computed. To my knowledge, methods for estimating valid frequentist confidence intervals for variance components remain to be developed. In addition, the proposed method allows simultaneous computation of the marginal posterior inclusion probabilities for both the fixed effects and random effects and correctly locates the true model as the one with highest posterior probability.

To evaluate sensitivity to the prior inclusion probability, the simulation is repeated with prior probabilities set to 0.2 and 0.8, with very little effect on the posterior means shown in Table 4.1. Posterior model probabilities were slightly different when the prior inclusion probabilities were changed; however there was no difference in

Table 4.1: Simulation results

Parameter	True	PQL	95% CI	Post. Mean	95% CI	Pr(inc.)
β_0	1.0	0.901	(0.753,1.048)	0.892	(0.759, 1.027)	1.000
β_1	0.0	0.031	(-0.062,0.125)	0.001	(0.000, 0.006)	0.044
β_2	1.0	0.900	(0.820,0.980)	0.929	(0.845, 1.016)	1.000
β_3	1.0	0.961	(0.896,1.025)	0.990	(0.920, 1.061)	1.000
ω_1	0.9	0.899		0.958	(0.721, 1.252)	1.000
ω_2	0.4	0.298		0.315	(0.221, 0.427)	1.000
ω_3	0.1	0.143		0.136	(0.072, 0.215)	1.000
ω_4	0.0	0.026		0.000	(0.000, 0.000)	0.008

parameter estimates, inferences or model ranking. In each case the true model had the highest posterior probability.

To evaluate the effect of using the priors induced by the parameter expanded model, simulation results are compared between two Gibbs samplers, one including and one excluding the redundant multiplicative parameter in the random effects component. As expected, there is no real difference in the point estimates; however, as seen in Figure 4.2, the parameter expansion approach resulted in improved computational efficiency and MCMC chains for the random effects variances. Table 4.2 shows the reduction in autocorrelation in the Gibbs chains. Note we have not directly drawn from the posterior distribution of the variances, rather we have computed them from the MCMC draws for $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$ and $\mathbf{d}$. The overparameterization causes the Gibbs chains for these parameters to mix poorly, but in combination they produce well-behaved chains for the random effects variances.

4.6 Epidemiology Application

As a motivating example, we consider data from the Collaborative Perinatal Project (CPP) conducted between 1959 and 1966. The desired inference is the effect of DDE, a metabolite of DDT, as measured in maternal serum, on pregnancy loss, a binary

Table 4.2: Autocorrelations in Gibbs chains, with and without parameter expansion

	Lag:	1	2	3	4	5	6	7	8	9
ω_1	w/o par exp	0.902	0.810	0.726	0.645	0.574	0.511	0.451	0.392	0.342
	w/par exp	0.422	0.350	0.288	0.252	0.208	0.177	0.154	0.142	0.132
ω_2	w/o par exp	0.783	0.653	0.558	0.484	0.422	0.369	0.324	0.286	0.251
	w/par exp	0.563	0.461	0.375	0.326	0.290	0.251	0.222	0.184	0.160
ω_3	w/o par exp	0.853	0.756	0.682	0.618	0.572	0.529	0.487	0.450	0.422
	w/par exp	0.811	0.711	0.639	0.574	0.520	0.477	0.441	0.417	0.388
ω_4	w/o par exp	0.808	0.629	0.439	0.335	0.228	0.162	0.087	0.038	0.008
	w/par exp	0.595	0.399	0.358	0.295	0.198	-0.001	-0.001	-0.001	-0.001

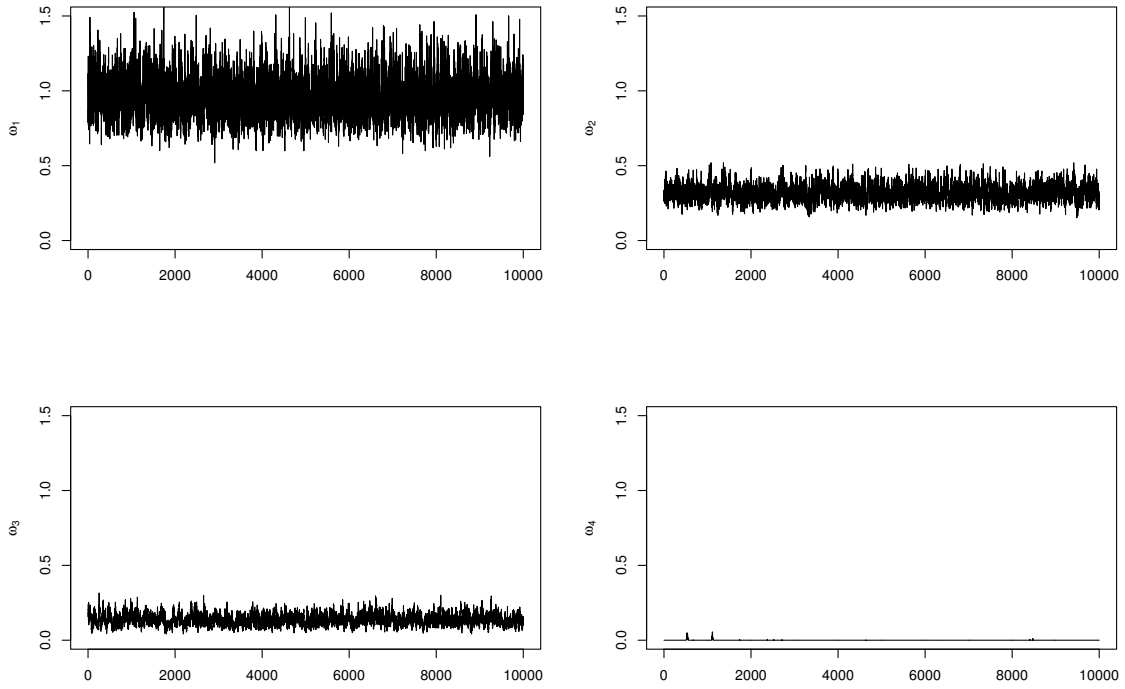

Figure 4.1: Gibbs chains for random effects variances

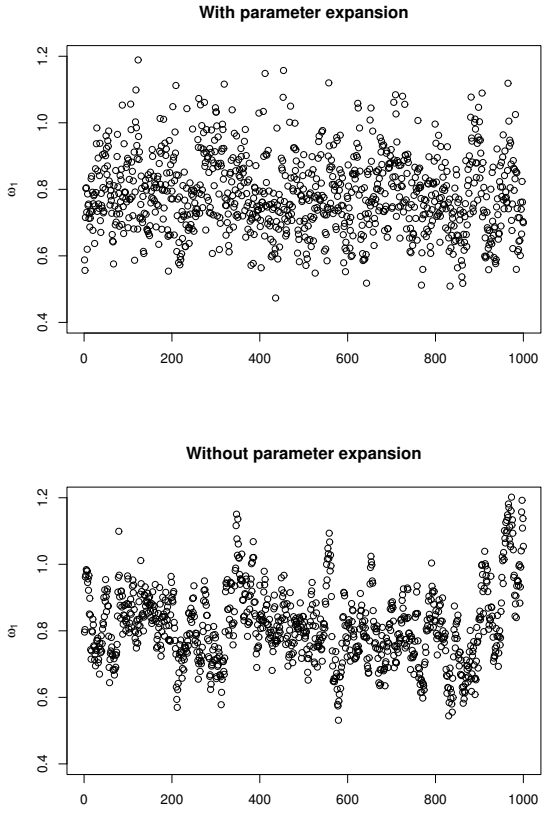

Figure 4.2: Illustration of parameter expansion effect on mixing of the Gibbs sampler

response variable. Potential confounding variables include mother's age, body mass index, smoking status, and serum levels of cholesterol and triglycerides. Data were collected across twelve different study centers and there is potential for heterogeneity across centers. The problem of interest is to select a logistic mixed effects model relating DDE levels and pregnancy loss, accounting for heterogeneity among study centers in those factors that vary in their effects across centers. In addition, inferences on whether predictors such as DDE vary in their effect is of substantial interest.

Let $y_{ij} = 1$ be a binary response variable indicate pregnancy loss for participant j in study center i, $i = 1, \ldots, 12; j = 1, \ldots, n_i$, for 5389 total participants. The covariate vector is $\boldsymbol{X}_{ij} = (1, X_{ij1}, \ldots, X_{ij5})'$ where X_{ij1} is the level of DDE, and $X_{ij2}, \ldots, X_{ij5}$ are the potential confounding variables. All covariates are continuous and centered at their means, and we let $\boldsymbol{Z}_{ij} = \boldsymbol{X}_{ij}$, thus considering all coefficients, including the intercept, for possible heterogeneity among centers.

Priors were chosen as in the simulation example and the Gibbs sampler run for 30,000 iterations after a burnin period of 5,000. The Gibbs sampling results indicate that there is no heterogeneity present among study centers and that a fixed effects model is appropriate. The preferred model, as shown in Table 4.3, includes only the intercept, body mass index, and age, as predictors. The posterior means for all variances are close to zero. A few models with nonzero posterior probability do contain a random effect. The posterior means for the fixed effect are similar to the PQL results returned by glmmPQL in R for the full model, shown in Table 4.4. These results also show that DDE did not have an appreciable effect on pregancy loss in the CPP study. The PQL results indicate that DDE had a very small but statistically significant effect; however, this may be due to bias in the PQL approach. Applying the BIC criteria to select the best fixed effects model yields the high posterior probability model shown in Table 4.3.

Table 4.3: Models with highest posterior probability

Prob	Model
.58	X_0, X_{bmi}, X_{age}
.16	X_0, X_{age}
.09	$X_0, X_{bmi}, X_{age}, X_{dde}$
.05	$X_0, X_{chol}, X_{bmi}, X_{age}$
.03	X_0, X_{age}, X_{dde}
.02	$X_0, X_{tg}, X_{bmi}, X_{age}$
.01	$X_0, X_{bmi}, X_{age}, Z_{chol}$
.01	X_0, X_{chol}, X_{age}
.01	$X_0, X_{chol}, X_{bmi}, X_{age}, X_{dde}$
.01	X_0, X_{age}, Z_{bmi}

Table 4.4: Posterior summary of fixed effects in CPP example

	PQL	95% CI	Mean	95% CI	$p(\beta_l = 0)$
β_0	-1.813	(-1.943, -1.700)	-1.793	(-1.871 ,-1.716)	0.000
β_{tg}	0.014	(-0.087, 0.101)	0.000	(0.000 , 0.000)	0.968
β_{chol}	-0.081	(-0.219, -0.001)	-0.002	(-0.034, 0.000)	0.932
β_{bmi}	-0.138	(-0.229, -0.055)	-0.096	(-0.210 , 0.000)	0.239
β_{age}	0.295	(0.211, 0.372)	0.279	(0.205 , 0.352)	0.000
β_{dde}	0.088	(0.009, 0.189)	0.005	(0.000 , 0.067)	0.876

4.7 Other models

4.7.1 Logistic models for ordinal data

This framework can also be adapted to accommodate logistic mixed models with ordinal response variables $y_{ij} \in \{1, \ldots, C\}$:

$$\text{logit}(P(y_{ij} \leq c | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{a}_i, \boldsymbol{\tau})) = \tau_c - \boldsymbol{X}'_{ij}\boldsymbol{\beta} - \boldsymbol{Z}'_{ij}\boldsymbol{a}_i, \quad c \in \{1, \ldots, C\} \quad (4.11)$$

where terms in the linear predictor are as defined in (4.3) and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{C-1})'$ where $\tau_1 = 0$ for identifiability and $-\infty = \tau_0 < \tau_1 < \cdots < \tau_C = \infty$ are threshold parameters for the ordered categories. The data augmentation stochastic search Gibbs sampler can be applied to (4.11) with modifications to truncate w_{ij} to $[\tau_{c-1}, \tau_c]$ for $y_{ij} = c$ and to update the threshold parameters $\boldsymbol{\tau}$. Although updating of $\boldsymbol{\tau}$ can potentially proceed after augmentation as described in Albert and Chib (1993), such an approach has a tendency to mix very slowly (Johnson and Albert, 1999). A modification in which the latent variables $\{w_{ij}\}$ are integrated out and a Metropolis-Hastings step is used yields better results. An alternative, which allows the baseline parameters $\boldsymbol{\tau}$ to be updated jointly from a multivariate normal posterior after augmentation, is to consider a continuation-ratio logit formulation of the form $\text{logit}(P(y_{ij} = c | y_{ij} \geq c, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}, \boldsymbol{a}_i)) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{a}_i$, instead of (4.11) (Agresti, 1990). Such formulations characterize the ordinal distribution in terms of the discrete hazard so are natural in time to event applications (Albert and Chib, 2001).

4.7.2 Probit models

Logistic models are often preferred over probit models due to the more intuitive interpretation of their regression coefficients in terms of odds ratios; however, it is worth noting that the approach for normal models is easily modified to accomplish model selection for probit mixed models by applying the well-known data augmentation Gibbs

sampler described in Albert and Chib (1993). For example, using a binary response probit model of the form $P(y_{ij} = 1) = \Phi(\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\boldsymbol{Z}'_{ij}\boldsymbol{a}_i)$, we introduce a latent variable v_{ij} such that $y_{ij} = 1(v_{ij} > 0)$ and $v_{ij} \sim N(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{a}_i, 1)$, yielding a conditional posterior distribution for v_{ij} of $N(\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\boldsymbol{Z}'_{ij}\boldsymbol{a}_i, 1)\cdot\{1(v_{ij} > 0)y_{ij}+1(v_{ij} < 0)(1-y_{ij})\}$. After updating v_{ij}, the MCMC algorithm proceeds as in the normal case, except that $\sigma^2 = 1$. In simulations this algorithm exhibited good mixing and convergence properties. This algorithm could also be adapted for ordinal probit models as described in the preceeding section.

4.8 Discussion

The Bayesian framework for model selection with mixed effects models discussed here is advantageous in that it allows for fixed and random effects to be selected simultaneously. Additionally it allows for marginal posterior inclusion probabilities to be computed for each predictor along with model-averaged coefficient estimates. Posterior model probabilities can be used to compare models; whereas frequentist testing for variance components is more limited.

In addition to model selection and averaging, the proposed prior structure and computational algorithm should be useful for efficient Gibbs sampling for fitting single mixed effects models. In particular, the prior and computational algorithm represent a useful alternative to approaches that rely on inverse-Wishart priors for variance components (e.g.Gilks *et al.* (1993)). There is an increasing realization that inverse-Wishart priors are a poor choice, particularly when limited prior information is available. Although this chapter focused on LMEs of the Laird and Ware (1982) type, it is straightforward to adapt the methods for a broader class of linear mixed models, accomodating varying coefficient models, spatially correlated data, and other applications (Zhao *et al.*, 2006).

Gibbs sampling chains from random effects model parameters tend to exhibit slow mixing and convergence. Gelfand *et al.* (1996) recommend hierarchical centering for improved convergence and posterior surface behavior. Vines *et al.* (1994) also propose a transformation of random effects to improve mixing. A challenge in implementing the hierarchically centered model is to efficiently update the correlation matrix in the context of random effects selection where we are interested in separating out the variances. One solution is proposed by Chib and Greenberg (1998); however, it is prohibitively slow for more than a couple random effects. Further work is needed to develop fast approaches that can be easily implemented and incorporated into software packages.

Chapter 5

Model selection with partially synthetic data

This chapter considers the problem of conducting model selection with partially synthetic data. As described in Section 1.2, partially synthetic datasets are constructed by replacing a portion of a confidential observed dataset with multiple imputations, for the purpose of reducing the risks of disclosing confidential information. While several inferential methods have been developed for various applications of multiple imputation (See Reiter and Raghunathan, 2007), few methods have been developed for conducting model selection with multiply-imputed data, and as far as I know, none for partially synthetic data. As different implicates of a multiply-imputed dataset may yield different model comparisons, a method for combining these inferences is desired.

The approach taken here is to give careful consideration to the correct fully Bayesian approach to obtaining posterior model probabilities from partially synthetic datasets. This provides a starting point from which to implement different Bayesian model selection procedures, such as a stochastic search variable selection algorithm (George and McCulloch, 1997; Geweke, 1996), which allows for both parameter and model uncertainty to be accounted for simultaneously, as discussed in Chapter 4. In addition to posterior model probabilities, model-averaged parameter estimates can be computed, along with marginal inclusion probabilities for each predictor. Also of interest is a method for computing Bayes factors, or posterior odds, which are commonly used to summarize the relative evidence provided by the data in support of one model against another. While these can be computed using posterior proba-

bilities estimated from a stochastic search algorithm, Bayes factor approximations, such as the BIC, are popular for their relative ease of computation and default prior specification. The availability of a simple Bayes factor approximation such as the BIC would be a valuable tool for users of partially synthetic data.

The remainder of this chapter is organized as follows. Section 5.1 gives notation and the general form for the appropriate likelihood and posterior model probability. These are used to develop a stochastic search variable selection procedure, described in Section 5.2, and Bayes factor approximations, described in Section 5.3. Section 5.4 provides concluding remarks and directions for future work.

5.1 Notation and motivation

For a finite population of size N, let $I_l = 1$ if unit l is included in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $n = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, e.g. stratum or cluster indicators or size measures. The design variables X are assumed to be known approximately for the entire population, for example from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population and Y_{inc} be the $n \times p$ matrix of survey data for the units sampled. Let $Y_{inc} = Y_{obs}$, i.e., assume that all all selected units are observed and no missing values are present. 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 $Z = (Z_1, \ldots, Z_n)$. The observed data is thus $D_{obs} = (X, Y_{obs}, I, Z)$.

Let Y_{rep} be the values of Y_{obs} which are to be replaced with multiple imputations, let $Y_{rep}^{(*)} = \{Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}\}$, where $Y_{rep}^{(i)}$ are all 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 $(Y_{rep}^{(i)} \mid D_{obs}, Z)$, or a close ap-

proximation of it. Each synthetic data set, $D_{syn}^{(i)}$, then comprises $(X, Y_{rep}^{(i)}, Y_{nrep}, I, Z)$. The entire collection of m data sets, $D_{syn} = \{D_{syn}^{(i)}, i = 1, \ldots, m\}$ is released to the public.

The available literature on model selection for multiply-imputed data is limited. Aside from a few ad-hoc attempts to obtain a combined BIC in the applied literature (e.g. Ball, 2001), there is one recent paper, Yang *et al.* (2005), where two model selection approaches are proposed for use with multiply-imputed data for nonresponse. In one approach, referred to as "simultaneously impute and select," or SIAS, the steps for imputation and Bayesian variable selection are embedded in a single Gibbs sampler. This approach is unsuitable for partially synthetic data as the imputer has access to the observed data, in which case the partially synthetic data are irrelevant, and the analyst has access only to the synthetic data. Another approach proposed by Yang *et al.* (2005), referred to as "impute then select," or ITS, is conducted with only the imputed data. The derivation, however, is based on the likelihood $p(D_{obs}|M)$, which does not represent a true likelihood in the partially synthetic data case since D_{obs} is not fully observed by the analyst, and sensible likelihoods cannot generally be constructed from the portion that is observed. Thus it is unclear that the method can be extended to partially synthetic data.

Nonetheless, the ITS method is illustrative of the challenges in carrying out model selection with multiply-imputed data. The method is conducted by computing posterior model probabilities with each completed dataset using a stochastic search variable selection algorithm, and then applying the standard combining rules of Rubin (1987) to obtain a combined posterior model probability. The justification for ITS relies on

the approximation

$$\begin{aligned} p(M|D_{obs}) &= \int p(M|D_{obs}, D_{mis})p(D_{mis}|D_{obs})dD_{mis} \\ &= \iint p(M|D_{obs}, D_{mis}, \theta)f(D_{mis}|D_{obs}, \theta)p(\theta|D_{obs})d\theta dD_{mis} \\ &\propto \sum_{i=1}^{m} \int p(M|D_{obs}, D_{com}^{(i)}, \theta)d\theta \end{aligned}$$

where D_{mis} is the portion of the data missing, D_{obs} is the portion observed, and $D_{com}^{(i)}, i = 1, \ldots, m$ are the imputed datasets. While this approximation may be reasonable in many cases, it is not theoretically justified. If $D_{com}^{(i)}$ is considered to be a draw from $f(D_{mis}|D_{obs}, \theta)$, then the approximation follows as a simple Monte Carlo estimate; however, as M is a random variable variable in the expression above, $p(D_{mis}|D_{obs})$ is correctly interpreted as $\sum_M p(D_{mis}|D_{obs}, M)$. Further, the approximation implicitly assumes that $D_{com}^{(i)}$ are drawn from $f(D_{mis}|D_{obs}, M)$. In actuality the imputations are drawn under a specific imputation model M^*.

In many imputation scenarios, the imputation model M^* is a collection of models $\{M_l^*, l = 1, \ldots, p\}$, used to model the joint distribution of the confidential data and generate multiple imputations. In contrast, M is one of 2^p possible models posited to explain the relationship between a specific response variable Y and p potential predictors. Existing multiple imputation combining rules for parameters (e.g. Rubin, 1987; Reiter, 2003) assume agreement between M and M^*; however, this assumption is not sensible for model inferences when there are several models under consideration.

As the assumption of agreement between the analyst and imputer models clearly is inadequate for inferences about models, the dependence of the imputations on M^* must be explicitly accounted for. The combining rules for inferences about population parameters are based on the posterior $f(\theta|Y_{nrep}, Y_{rep}^{(*)})$. Inferences for models $M_j, j = 1, \ldots, 2^p$ are thus based on the posterior model probabilities $f(M_j|Y_{nrep}, Y_{rep}^{(*)}, M^*) \propto$

$f(Y_{nrep}, Y_{rep}^{(*)}|M_j, M^*)f(M_j|M^*)$. This posterior probability is used to construct a model search algorithm and a Bayes factor in the sections that follow.

Although combining rules for parameters assume agreement between the analyst and imputer models, valid inferences are still obtained as long as imputations are *proper* (Rubin, 1987) and analyses *congenial* (Meng, 1994), as described in Section 1.1.3. A similar notion of congeniality is necessary between M^* and the models under consideration, even though the imputation model is explicitly factored into the analysis. For example, if M specifies a relationship between two variables that is assumed to be zero in M^*, then model comparisons are likely to favor models excluding those predictors. This only presents a serious problem if a significant relationship truly exists, i.e., the imputer has made poorly grounded assumptions.

5.2 Bayesian model uncertainty

This section describes the application of a Bayesian model uncertainty procedure to partially synthetic data. A stochastic search algorithm is computed for an illustrative example of a linear model in a particular imputation scenario. This differs from Yang *et al.* (2005) in that a single combined stochastic search algorithm is used rather than running a separate Gibbs sampler on each imputed dataset.

Let $Y_{rep} = Y$, for some response variable Y, and $Y_{nrep} = X$ for a $n \times p$ matrix of predictor variables X. Suppose an analyst is interested in locating a subset of predictors in X that parsimoniously describes the relationship between Y and X using the linear model $Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$. Assume that M^* is the saturated model, $Y = X\gamma + \delta$, $\delta \sim N(0, \tau)$, so that $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}$ are drawn from from $N(X\hat{\gamma}_{obs}, \hat{\tau}(I + X'(X'X)^{-1}X))$, I now indicating a p-dimensional identity matrix. The joint posterior distribution $f(\beta, \sigma^2, M|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is not easily determined, however, the posterior $f(\beta, \sigma^2, M|Y_{nrep}, Y_{rep})$ is well known, and a stochastic search

Gibbs sampler can easily be performed. By augmenting the posterior with Y_{rep}, we have $f(\beta, \sigma^2, M, Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$, which can be factored as

$$\begin{aligned} & f(\beta, \sigma^2, Y_{rep}, M|Y_{nrep}, Y_{rep}^{(*)}, M^*) \\ & \quad = f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)p(\beta, \sigma^2, M|Y_{nrep}, Y_{rep}^{(*)}, Y_{rep}, M^*) \qquad (5.1) \\ & \quad = f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)p(\beta, \sigma^2, M|Y_{nrep}, Y_{rep}) \qquad (5.2) \\ & \quad \propto f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)p(Y_{rep}, Y_{nrep}|M, \beta, \sigma^2)p(M, \beta, \sigma^2). \qquad (5.3) \end{aligned}$$

The simplification from (5.1) to (5.2) follows because if the observed data Y_{rep} is known, then there is no use for the synthetic data $Y_{rep}^{(*)}$ and imputation model M^*.

The distribution $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ can be computed by analysts without access to D_{obs}. The distribution $p(\beta, \sigma^2, M|Y_{rep}, Y_{nrep})$ is simply $p(\beta, \sigma^2, M|D_{obs})$, so by inserting draws of Y_{rep} from its conditional posterior distribution, we can then compute $p(\beta, \sigma^2, M|Y_{nrep}, Y_{mrep}^{(k)})$, where $Y_{mrep}^{(k)}$ is draw from $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$, as if $(Y_{nrep}, Y_{mrep}^{(k)})$ represents an observed dataset. Running this data augmentation Gibbs sampler in effect obtains the joint posterior $f(\beta, \sigma^2, M|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ by averaging over Y_{rep}.

In order to obtain the full conditional posterior distributions, the prior $p(\beta, \sigma^2, M)$ needs to be specified. Proper distributions are desired in order 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. A common choice of priors for regression coefficients in Bayesian model selection problems is a mixture of a point mass at zero and a normal or heavier-tailed distribution. A similar formulation, used here and in Chapter 4, is to place a Zellner-type prior on β_J, where β_J is the vector of coefficients corresponding to the current model M_J and J is a vector of indicator variables J, such that $J_l, l = 1, \ldots, p$ indicates that the l-th predictor is in the model. 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).

For the illustrative example, the prior specification is given by $(\beta|J,\sigma^2) \sim N(0,\sigma^2(X_J'X_J)^{-1}/g)$, 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.

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. After discarding an initial burn-in period, the draws of J can be used to determine 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.

The full conditional posterior distribution of Y_{rep} is $f(Y_{rep}|Y_{nrep},Y_{rep}^{(*)},M^*)$ which is determined as $f(Y_{rep}|Y_{nrep},Y_{rep}^{(*)},M^*) = \int f(Y_{rep}|Y_{nrep},Y_{rep}^{(*)},M^*,\gamma,\tau) \times p(\gamma,\tau|Y_{nrep},Y_{rep}^{(*)},M^*)d\gamma d\tau$. 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, as defined in (1.4) and (1.6). 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

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

  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

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

5.2.1 Simulation example 1

Using the illustrative example, an SSVS algorithm is implemented on a simulated dataset. For comparison purposes, similar algorithms are conducted both on the observed data and partially synthetic data. The procedure is run repeatedly on models drawn at random to compare the percent of time that the true model has the highest posterior probability, and then a few cases are examined in detail. Let

$n = 1000$ and $p = 6$. The $n \times p$ matrix of predictors $X = Y_{nrep}$ is generated from a standard normal distribution. A dependent variable $Y = Y_{rep}$ is generated from $N(X\beta, I)$, where β is generated from a standard normal distribution and $J_l, l = 1, \ldots, p$ is drawn from a $Beta(2, 2)$.

The imputations $Y_{rep}^{(*)} = Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}$, with $m = 5$, are generated from the posterior predictive distribution of $Y|X$, using the saturated model $Y = X\gamma + \delta, \delta \sim N(0, \tau)$, by drawing τ from $(n-p)s^2/\chi^2_{n-p}$, γ from $N(X\hat{\gamma}, (X'X)^{-1}\tau)$, and $Y_{rep}^{(i)}$ from $N(X\gamma, \tau)$. The data augmentation Gibbs sampling algorithm is used on the partially synthetic dataset $(X, Y_{rep}^{(*)})$ to compute the posterior model probabilities. A similar algorithm is run on the observed data (X, Y). The observed data algorithm used is the same as the synthetic data algorithm except that Y_{rep} is known and fixed. After a burnin period of 100 iterations, 1000 iterations are saved.

When 1000 models are drawn, all 64 possible models are visited between 7 and 29 times, with different coefficients each time. The coefficient vectors are saved so that unusual results can be re-examined. The observed data algorithm assigns the highest probability to the true model 758 times while the synthetic data algorithm assigns it 702 times. The true model is ranked below fifth 19 times in the observed data and 35 times in the synthetic data algorithm. The lowest ranking of the true model is 17th in the observed data and 26th in the synthetic data.

Capturing the "truth" is not the only measure of success for a model search algorithm, so by itself it is not very alarming if the true model does not have the highest posterior probability. Often there are several models with approximately the same posterior probability, so that the difference between them is nominal. Further examination of the cases in which the observed and/or synthetic data model searches fail to find the true model reveals that nearly all of these involve null models or models with very small coefficients so that there is little difference between them and

Table 5.1: Posterior model probabilities, null model true, Example 1

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_4	0.133	X_1	0.119
X_3	0.133	X_3	0.118
X_5	0.118	X_4	0.092
X_1	0.118	X_6	0.090
X_2	0.117	X_2	0.082
X_6	0.109	X_5	0.081
null	0.109	null	0.080
X_1, X_3	0.026	X_1, X_3	0.033
X_3, X_4	0.018	X_3, X_6	0.024
X_2, X_3	0.016	X_2, X_3	0.023

Table 5.2: Marginal inclusion probabilities, null model true, Example 1

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	0.361	0.417	0.361	0.278	0.389	0.389
Synthetic	0.426	0.362	0.447	0.404	0.426	0.426

other models with zero or small coefficients. This information is readily available from the Gibbs sampler output.

To illustrate, generate Y independently of X so that $\beta = (0, 0, 0, 0, 0, 0)$, generate partially synthetic data as before, and run the observed and synthetic data Gibbs sampling algorithms. In one run, the observed data algorithm visited 36 models in 1000 iterations while the synthetic data algorithm visited 47. The top 10 models are given in Table 5.1 and the marginal inclusion probabilities in Table 5.2. The results from both cases are seen to be similar, although the probabilities differ slightly. In neither case is the null model selected as the highest posterior probability model; however, in both cases it is clear that there is no evidence to support the inclusion of any predictors.

In contrast, let Y be generated from $N(X\beta, \sigma^2)$ with $\beta = (1, 0, 0, 0, 0, 0)$, generate partially synthetic data, and run the observed and synthetic data model Gibbs sam-

Table 5.3: Posterior model probabilities, one predictor in true model, Example 1

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_1	0.767	X_1	0.552
X_1, X_4	0.070	X_1, X_6	0.104
X_1, X_6	0.041	X_1, X_2	0.095
X_1, X_2	0.038	X_1, X_3	0.052
X_1, X_3	0.034	X_1, X_5	0.048
X_1, X_5	0.023	X_1, X_4	0.041
X_1, X_4, X_6	0.006	X_1, X_2, X_6	0.023
X_1, X_4, X_5	0.004	X_1, X_3, X_6	0.017
X_1, X_3, X_6	0.003	X_1, X_4, X_6	0.010
X_1, X_2, X_5	0.002	X_1, X_2, X_5	0.008

Table 5.4: Marginal inclusion probabilities, one predictor in true model, Example 1

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	1.000	0.051	0.039	0.085	0.033	0.053
Synthetic	1.000	0.015	0.097	0.074	0.082	0.176

pling algorithms. In one run, the observed data algorithm visited 16 models while the synthetic data algorithm visited 25. The results in both cases convincingly show that the true model is the high posterior probability model. Table 5.3 shows the top 10 models and their posterior probabilities and Table 5.4 gives the marginal inclusion probabilities.

5.2.2 Simulation example 2

This example considers a variation of the illustrative example where two variables are imputed in entirety. Let $Y_{rep} = (Y, X_1)$, and $Y_{nrep} = (X_2, \ldots, X_p)$ for some response variable Y and a $n \times p$ matrix of predictor variables X. The imputation procedure is to generate $f(Y, X_1|X_2, \ldots, X_p) = f(Y|X_1, \ldots, X_p)f(X_1|X_2, \ldots, X_p)$ using normal linear models as in Example 1. The analysis model is the same, so the augmented posterior distribution $p(\beta, \sigma^2, Y_{rep}, M|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is the same as given in (5.1) to

(5.3), and the stochastic search algorithm is the same, but since M^* and Y_{rep} are different, the specification of the distribution $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is different.

In this example, $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ is factored 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 thus updated in the Gibbs sampler as follows:

1. Draw τ_2 from $(n-p-1)\bar{s}_2^2\chi_{n-p-1}^{-2}$ 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 (1.4) and (1.6).

3. Draw X_1 from $N(X_{2:p}\gamma_2, \tau_2)$.

4. Draw τ_1 from $(n-p)\bar{s}_1^2\chi_{n-p}^{-2}$, where $\bar{s}_1^2 = \sum_{i=1}^m (Y_{rep}^{(i)} - X\hat{\gamma}_1^{(i)})'(Y_{rep}^{(i)} - X\hat{\gamma}_1^{(i)})/m(n-1)$.

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 (1.4) and (1.6).

6. Draw $Y_{mrep}^{(k)}$ from $N(X\gamma_1, \tau_1)$.

The rest of the Gibbs sampler steps are the same as in Example 1. This procedure can potentially be generalized for a variety of imputation scenarios and inference models. There is no requirement that the distributional assumptions of the imputation model and analysis model agree.

Table 5.5: Posterior model probabilities, null model true, Example 2

Observed		Synthetic	
Top 10 Models	$P(M\|Data)$	Top 10 Models	$P(M\|Data)$
X_1	0.146	X_6	0.099
X_2	0.129	X_5	0.083
X_6	0.128	X_2	0.073
X_3	0.126	X_1	0.070
null	0.115	null	0.064
X_5	0.108	X_3	0.059
X_4	0.106	X_4	0.057
X_2, X_6	0.014	X_1, X_5	0.032
X_3, X_5	0.012	X_2, X_3	0.030
X_1, X_5	0.011	X_2, X_5	0.030

Table 5.6: Marginal inclusion probabilities, null model true, Example 2

	X_1	X_2	X_3	X_4	X_5	X_6
Observed	0.192	0.174	0.177	0.154	0.163	0.185
Synthetic	0.260	0.269	0.264	0.241	0.315	0.295

The simulation is carried out as in Example 1, with model searches run on observed data and synthetic data over 1000 draws of data and models. In 1000 draws, 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 pick the true model 735 times. As in Example 1, we can examine the Gibbs sampler output for one run of each algorithm when the true model is the null model. The results are shown in Table 5.5 and Table 5.6 and the results are seen to be similar to those of Example 1.

5.3 Bayes factors

Using the appropriate posterior model probabilities to construct a Bayes factor to compare two models M_1 and M_0, 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, the prior odds ratio is taken to equal one, so that the Bayes factor is just the likelihood ratio

$$\frac{f(Y_{nrep}, Y_{rep}^{(*)}|M_1, M^*)}{f(Y_{nrep}, Y_{rep}^{(*)}|M_0, M^*)}. \tag{5.4}$$

The likelihood $f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*)$ is not readily available but can be obtained in an analytically tractable form by integrating over the replaced values Y_{rep}:

$$\begin{aligned} f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) &= \int f(Y_{rep}, Y_{nrep}, Y_{rep}^{(*)}|M^*, M) dY_{rep} && (5.5) \\ &\propto \int f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*) f(Y_{nrep}, Y_{rep}|M) dY_{rep} && (5.6) \\ &\propto \int f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*) \frac{f(Y_{nrep}, Y_{rep}|M)}{f(Y_{nrep}, Y_{rep}|M^*)} dY_{rep} && (5.7) \end{aligned}$$

To get from (5.5) to (5.6), note that $f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*, M) = f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*)$ since $Y_{rep}^{(*)}$ are generated from $f(Y_{rep}^{(*)}|Y_{nrep}, Y_{rep}, M^*)$ independently of M. Additionally, $f(Y_{nrep}, Y_{rep}|M^*, M) = f(Y_{nrep}, Y_{rep}|M)$, as Y_{rep} and Y_{nrep} do not depend on M^*, and $f(M^*, M)$ is assumed to be a constant. The integral in (5.7) follows directly from Bayes rule and can be approximated with a Monte Carlo estimate as

$$f(Y_{nrep}, Y_{rep}^{(*)}|M^*, M) \propto \frac{1}{K} \sum_{k=1}^{K} \frac{f(Y_{nrep}, Y_{mrep}^{(k)}|M)}{f(Y_{nrep}, Y_{mrep}^{(k)}|M^*)} \tag{5.8}$$

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. Note that 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 expression can be computed by an analyst without access to the confidential data, provided that M^* is made available by the imputer, though it is numerically difficult to compute for large n, and in test runs yielded unstable estimates with small n. Applying Laplace approximations simplifies the computation and provides a good approximation, up to a constant, of the observed-data likelihood.

Using the Laplace method yields the Bayes factor approximation of

$$-2\log\frac{f(Y_{nrep}, Y_{rep}^{(*)}|M_1, M^*)}{f(Y_{nrep}, Y_{rep}^{(*)}|M_0, M^*)} \approx -2\log\left(\sum_{k=1}^{K}\frac{f(Y_{nrep}, Y_{mrep}^{(k)}|M_1, \bar{\theta})}{f(Y_{nrep}, Y_{mrep}^{(k)}|M^*, \bar{\gamma})}\right) + 2\log\left(\sum_{k=1}^{K}\frac{f(Y_{nrep}, Y_{mrep}^{(k)}|M_0, \bar{\theta})}{f(Y_{nrep}, Y_{mrep}^{(k)}|M^*, \bar{\gamma})}\right) + (k_1 - k_0)\log n. \quad (5.9)$$

where $\bar{\theta}$ and $\bar{\gamma}$ are the maximum likelihood estimates of θ and γ obtained from $(Y_{nrep}, Y_{rep}^{(*)})$ under M and and M^*, respectively. These are used over the MLEs obtained from $(Y_{nrep}, Y_{mrep}^{(k)})$ as overall the likelihood of interest is $f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*)$.

To illustrate, using the Laplace method to integrate $f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*)$ over θ and γ yields

$$\begin{aligned} f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*) &= \iint f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*, \theta, \gamma)p(\theta|M)p(\gamma|M^*)d\theta d\gamma \\ &\approx C \cdot \int f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*, \bar{\theta}, \gamma) \cdot n^{-k/2} \cdot p(\gamma|M^*)d\gamma \\ &\approx C' \cdot f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*, \bar{\theta}, \bar{\gamma}) \cdot n^{-k/2}n^{-p/2} \end{aligned}$$

where C and C' represent the terms in the Laplace approximation that are $O_n(1)$ or less. Integrating $f(Y_{nrep}, Y_{rep}^{(*)}|M, M^*, \bar{\theta}, \bar{\gamma})$ over Y_{rep} as in (5.5) to (5.7) yields the approximation in (5.9).

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 (5.9) will yield an estimate that is simpler to compute. The distributions of $Y_{rep}^{(i)}$ and $Y_{mrep}^{(k)}$ in this example are equivalent under infinitely many imputations, suggesting this might be a reasonable approximation. Further assessment is needed to make a general assertion. An alternative approximation to (5.9) using the $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}$ in place of $Y_{mrep}^{(k)}$ is given by

$$-2\log\frac{f(Y_{nrep}, Y_{rep}^{(*)}|M_1, M^*)}{f(Y_{nrep}, Y_{rep}^{(*)}|M_0, M^*)} \approx -2\log\left(\sum_{i=1}^{m}\frac{f(Y_{nrep}, Y_{rep}^{(i)}|M_1, \bar{\theta})}{f(Y_{nrep}, Y_{rep}^{(i)}|M^*, \bar{\gamma})}\right)$$
$$+2\log\left(\sum_{i=1}^{m}\frac{f(Y_{nrep}, Y_{rep}^{(i)}|M_0, \bar{\theta})}{f(Y_{nrep}, Y_{rep}^{(i)}|M^*, \bar{\gamma})}\right) + (k_1 - k_0)\log n. \quad (5.10)$$

5.3.1 Simulation examples

This section evaluates the performance of the Bayes factor approximations in (5.9) and (5.10) using a few simple simulation examples. Let $n = 10,000$ and $p = 6$. The $n \times p$ matrix of predictors $X = Y_{nrep}$ is generated from a standard normal distribution. A dependent variable $Y = Y_{rep}$ is generated from $N(X\beta, I)$, where β is generated from a standard normal distribution and $J_l, l = 1, \ldots, p$ is drawn from a $Beta(2, 2)$.

The imputations $Y_{rep}^{(*)} = Y_{rep}^{(1)}, \ldots, Y_{rep}^{(m)}, m = 5$, are generated from the posterior predictive distribution of $Y|X$, using the saturated model $Y = X\gamma + \delta, \delta \sim N(0, \tau)$, by drawing τ from $(n-p)s^2/\chi^2_{n-p}$, γ from $N(\hat{\gamma}_{obs}, (X'X)^{-1}\tau)$, and $Y_{rep}^{(i)}$ from $N(X\gamma, \tau)$. Under repeated draws of models, the performance of the approximations, using $K = 5$, is evaluated relative to the observed-data BIC by comparing the estimated rank of the true model, which ideally should be ranked first. For computational simplicity, the approximations are only estimated for the ten highest ranked models as determined by the observed-data BIC.

Table 5.7 gives the frequencies of the estimated rankings of the "true" model for

Table 5.7: Comparison of Bayes factor approximations

Est. ranks:	1	2	3	4	5	6+
Obs. BIC	465	19	9	4	2	1
Approx 1	393	55	30	17	2	3
Approx 2	453	28	10	6	1	2

the observed-data BIC and both approximations for 500 true models. Interestingly, Approximation 2 in (5.10) appears to perform slightly better than Approximation 1 in (5.9) relative to the BIC. This evaluation does not take into account cases where there are multiple top models with similar values of BIC.

5.4 Discussion

This chapter has described some model selection procedures for use with partially synthetic data. The stochastic model search proposed is general and can be applied to different imputation scenarios; however, the specification of the distribution $f(Y_{rep}|Y_{nrep}, Y_{rep}^{(*)}, M^*)$ may be more difficult than in the illustrative examples. The algorithm described is based on an observed-data algorithm with an augmentation step. It is expected that the augmentation step could be applied to other Bayesian observed-data models. Certainly it could be applied to linear model estimation, and potentially extended to analyses currently unavailable to users of multiply-imputed data. Further work is needed to evaluate such procedures, and assess when they yield analytically valid inferences. Similar procedures are also desired for other applications of multiple imputation, including missing data and two-stage imputation.

The Bayes factor approximations presented provide reasonable results in the cases demonstrated; however, additional work is needed to determine the appropriate approximation for the term $f(Y_{nrep}, Y_{mrep}^{(k)}|M^*)$ under more complex imputation procedures M^*. Additionally, determining the proper numerical approach for computing

the Monte Carlo estimate of the Bayes factor in (5.8) will provide a close estimate of the Bayes factor and provide a better metric than the observed-data BIC, itself an approximation, for assessing the accuracy of the approximations in (5.9) and (5.10).

The development of model selection procedures when M^* is unknown is also left for future work. In the absence of a better approach, the best an analyst typically can do is to make a reasonable guess as to M^*. Imputation models are recommended to be as general and saturated as possible (Meng, 1994; Schafer, 1997), so a saturated model is a good place to start, perhaps with reasonable transformation and interaction terms appropriate for the scientific context. It is also plausible that agencies may release partial information about the imputation procedure, enough to yield a reasonable approximation. Further evaluation of the concepts of congeniality and proper imputations for model inferences is needed to determine explicitly when model inferences are valid and provide a deeper understanding of model selection with data imputed under a specific model.

Appendix A

Full conditional posterior distributions

Full conditional distributions are presented for the parameter-expanded Gibbs sampling algorithm of Chapter 4 for a binary logistic model. Except where noted, the corresponding distribution for the normal case is obtained by substituting y_{ij} for w_{ij}, and σ^2 for $\tilde{\sigma}^2/\phi_{ij}$. Additionally, σ^2 is sampled only in the normal case, and w_{ij} and ϕ_{ij} only in the logistic case. Let $\boldsymbol{X}_{ij}^{\mathbf{J}}$ denote the subvector of X_{ij}, $\{X_{ijl} : J_l = 1\}$ and $\boldsymbol{\psi}$ be the N-vector such that $\psi_{ij} = w_{ij} - \mathbf{X^{J'}}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}'_{ij}\boldsymbol{A\Gamma\xi}_i$.

- The full conditional posterior $p(\boldsymbol{\beta_J}|\boldsymbol{J}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ is $N(\hat{\boldsymbol{\beta}}_{\mathbf{J}}, \boldsymbol{V}_{\mathbf{J}})$ where

$$\hat{\boldsymbol{\beta}}_{\mathbf{J}} = \left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{\phi_{ij}}{\tilde{\sigma}^2} \psi_{ij} \mathbf{X^{J'}}_{ij} \right) \cdot \boldsymbol{V}_{\mathbf{J}} \text{ and } \boldsymbol{V}_{\mathbf{J}} = \left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \mathbf{X^{J}}_{ij} \mathbf{X^{J'}}_{ij} \left(\frac{\phi_{ij}}{\tilde{\sigma}^2} + g \right) \right)^{-1}$$

- To calculate the posterior for $\boldsymbol{J}$ we need to update each J_l individually. We calculate $p(J_l = 1|\boldsymbol{J}_{-l}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ for $l = 1, \ldots, p$, by integrating out $\boldsymbol{\beta}$ as in Smith and Kohn (1996) and obtaining $p(J_l = 1|\boldsymbol{J}_{-l}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \frac{1}{1+h_l}$, where $J_{-l} = \{J_i : i \neq l\}$ and $h_l = \frac{1-p_{l0}}{p_{l0}} \cdot c^{1/2} \cdot \frac{S(J_l=0)}{S(J_l=1)}$. For the logistic case we set $c = \frac{1}{g}$ and define

$$S(\boldsymbol{J}) = |\mathbf{X^{J'}X^{J}}|^{1/2} \cdot |\boldsymbol{V}_{\mathbf{J}}|^{1/2} \exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \phi_{ij} \psi_{ij}^2 - \hat{\boldsymbol{\beta}}'_{\mathbf{J}} \boldsymbol{V}_{\mathbf{J}}^{-1} \hat{\boldsymbol{\beta}}_{\mathbf{J}} \right) \right\}$$

and for the normal case we set $c = 1 + \frac{1}{g}$. We integrate out σ^2 and define:

$$S(\boldsymbol{J}) = \left(\boldsymbol{\psi}'\boldsymbol{\psi} - \hat{\boldsymbol{\beta}}_{\mathbf{J}} \boldsymbol{V}_{\mathbf{J}}^{-1} \hat{\boldsymbol{\beta}}_{\mathbf{J}} \right)^{-N/2}$$

$S(J_l = 0)$ is equivalent to $S(\boldsymbol{J})$ but with the element J_l of $\boldsymbol{J}$ set to 0, so $\boldsymbol{\psi}$, $\mathbf{X^J}$, $\hat{\boldsymbol{\beta}}_{\mathbf{J}}$ and $\boldsymbol{V}_{\mathbf{J}}$ may need to be recomputed to correspond to $J_l = 0$. Similarly for $S(J_l = 1)$.

- The gamma prior on g yields a conjugate gamma posterior:

$$\Gamma\left(\frac{p_{\mathbf{J}}+1}{2}, \frac{\boldsymbol{\beta_J}'\mathbf{X^{J}}'\mathbf{X^J}\boldsymbol{\beta_J}/\sigma^2 + N}{2}\right)$$

where $p_{\mathbf{J}} = \sum_{l=1}^{p} 1(J_l = 1)$. Set σ^2 =1 for the logistic case.

- Similary each ϕ_{ij} has a conjugate gamma posterior:

$$G\left(\frac{\nu+1}{2}, \frac{(w_{ij} - \boldsymbol{Z}_{ij}\boldsymbol{A\Gamma\xi}_i - \boldsymbol{X}'_{ij}\boldsymbol{\beta})^2/\tilde{\sigma}^2 + \nu}{2}\right)$$

- The posterior for $p(\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{J}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ is:

$$IG\left(\frac{N + p_{\mathbf{J}}}{2}, \frac{\boldsymbol{\psi}'\boldsymbol{\psi} + g\boldsymbol{\beta_J}'\mathbf{X^{J}}'\mathbf{X^J}\boldsymbol{\beta_J}}{2}\right)$$

- The full conditional posterior $p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ is given by $N(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{V}}_{\gamma}) \cdot 1(\boldsymbol{\gamma} \in \boldsymbol{R_\lambda})$ where

$$\hat{\boldsymbol{V}}_{\gamma} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{\phi_{ij}}{\tilde{\sigma}^2}\boldsymbol{u}_{ij}\boldsymbol{u}'_{ij} + \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1}\right)^{-1}$$

and

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{\phi_{ij}}{\tilde{\sigma}^2}(w_{ij} - \mathbf{X^J}_{ij}\boldsymbol{\beta_J})u'_{ij} + \boldsymbol{\gamma}_0\boldsymbol{V}_{\boldsymbol{\gamma}}^{-1}\right) \cdot \hat{\boldsymbol{V}}\boldsymbol{\gamma}$$

The $q(q-1)/2$ vector $\boldsymbol{u}_{ij}$ is defined as $(\xi_{il}\alpha_m Z_{ijm} : l = 1, \ldots, q, m = l+1, \ldots, q)'$ so that the random effects term $\boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i$ can be written as $\boldsymbol{u}'_{ij}\boldsymbol{\gamma}$.

- The latent variables $\boldsymbol{\xi}_i$ have posterior $p(\boldsymbol{\xi}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ given by $N(\hat{\boldsymbol{\xi}}_i, \boldsymbol{V}_\xi)$ where

$$\hat{\boldsymbol{\xi}}_i = \sum_{j=i}^{n_i} \phi_{ij}(w_{ij} - \mathbf{X}^{\mathbf{J}'}_{ij}\boldsymbol{\beta}_{\boldsymbol{J}})\boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{V}_\xi\tilde{\sigma}^{-2}$$

and

$$\boldsymbol{V}_\xi = \left(\sum_{j=1}^{n_i} \phi_{ij}\boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{Z}_{ij}\boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\tilde{\sigma}^{-2} + \boldsymbol{D}^{-1}\right)^{-1}$$

- Each α_l must be updated individually. The zero-inflated truncated normal prior for α_l yields a conjugate posterior $p(\alpha_l|\boldsymbol{\alpha}_{-l}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = ZI - N^+(\hat{\alpha}, V_{\alpha l}, \hat{p}_l)$ where

$$\hat{\alpha} = \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\phi_{ij}t_{ijl}T_{ij}}{\tilde{\sigma}^2}\right)V_{\alpha l}, \qquad V_{\alpha l} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\phi_{ij}t^2_{ijl}}{\tilde{\sigma}^2} + 1\right)^{-1}$$

$$\hat{p}_l = \frac{p_{al}}{p_{al} + (1-p_{al})\frac{N(0;0,1)}{N(0;\hat{\alpha},V_{\alpha l})} \cdot \frac{1-\Phi(0;\hat{\alpha},V_{\alpha l})}{1-\Phi(0;0,1)}}$$

where $T_{ij} = w_{ij} - \mathbf{X}^{\mathbf{J}'}_{ij}\boldsymbol{\beta}_{\boldsymbol{J}} - \sum_{k\neq l} t_{ijk}\alpha_k$ and $N(0; m, v)$ denotes the normal density with mean m and variance v evaluated at 0 and $\Phi(0; m, v)$ is the normal cumulative distribution function with mean m and variance v evaluated at 0. The q vector

$$\boldsymbol{t}_{ij} = \left(Z_{ijl}\left(\xi_{il} + \sum_{m=1}^{l-1}\xi_{im}\gamma_{ml}\right) : l = 1, \ldots, q\right)^T$$

is defined so that the random effects term $\boldsymbol{Z}'_{ij}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\xi}_i$ can be written as $\boldsymbol{t}'_{ij}\boldsymbol{\alpha}$.

- The diagonal elements of $\boldsymbol{D}$ have inverse gamma priors $IG(\frac{1}{2}, \frac{N}{2})$; hence the posterior is given by $p(d_l|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{y}) = IG\left(\frac{1}{2} + \frac{n}{2}, \frac{N}{2} + \frac{\sum_{i=1}^{n} \xi_{il}^2}{2}\right)$

Bibliography

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*, 282–289. New York: Springer-Verlag.

Abowd, J. M., Stinson, M. H., and Benedetto, G. L. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.

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.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 422, 669–679.

Albert, J. H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57**, 3, 829–836.

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

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

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference* **112**, 241–258.

Berger, J. O. and Pericchi, L. R. (1996). The instrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In P. Lahiri, ed., *Model Selection*, vol. 38 of *IMS Lecture Notes - Monograph Series*, 135–193. Institute of Mathematical Statistics.

Breslow, N. (2003). Whither PQL? *UW Biostatistics Working Paper Series* Working Paper 192.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62**.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. and Greenberg, E. (1998). Bayesian analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Chung, Y. and Dey, D. (2002). Model determination for the variance component model using reference priors. *Journal of Statistical Planning and Inference* **100**, 49–65.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* **86**, 68–78.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.

Cole, S., Chu, H., and Greenland, S. (2006). Multiple imputation for measurement error correction. *International Journal of Epidemiology* **35**, 1074–1081.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society B* **66**, 165–185.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Davis, S. J., Haltiwanger, J. C., and Schuh, S. (1996). *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Defays, D. and Nanopoulos, P. (1992). Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, 195–204.

Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002). Software systems for tabular data releases. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* **10**, 529–544.

Dreschler, J., Dundler, A., and Bender, S. (2007). A new approach for disclosure control in the IAB establishment panel: Multiple imputation for a better data access. Tech. rep., Institute for Employment Research (IAB).

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **81**, 10–28.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381–427.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parameterizations for generalized linear mixed models. *Bayesian Statistics* **5**.

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.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 2, 1–19.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

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

Gerlach, R., Bird, R., and Hall, A. (2002). Bayesian variable selection in logistic regression: Predicting company earnings direction. *Australian and New Zealand Journal of Statistics* **42**, 2, 155–168.

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

Gilks, W., Wang, C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**, 441–453.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005). Data swapping as a decision problem. *Journal of Official Statistics* **21**, 635–655.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P. J. (1997). Discussion of "The EM algorithm - an old folk song sung to a fast new tune," by Meng and van Dyk. *Journal of the Royal Statistical Society, Series B* **59**, 3, 554–555.

Hall, D. and Praestgaard, J. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika* **88**, 739–751.

Harel, O. (2003). *Strategies for Data Analysis with Two Types of Missing Values*. Ph.D. thesis, The Pennsylvania State University.

Harel, O. and Schafer, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 436.

Holmes, C. and Knorr-Held, L. (2003). Efficient simulation of Bayesian logistic regression models. Tech. rep., Ludwig Maximilians University Munich.

Jang, W. and Lim, J. (2005). Estimation bias in generalized linear mixed models. Tech. rep., Institute for Statistics and Decision Sciences, Duke University.

Jarmin, R. S. and Miranda, J. (2002). The Longitudinal Business Database. Tech. Rep. 02-17, U.S. Census Center for Economic Studies.

Jarmin, R. S. and Miranda, J. (2007). Personal communication.

Jiang, J., Rao, J., Gu, Z., and Nguyen, T. (2006). Fence methods for mixed model selection. Tech. rep., Department of Statistics, University of California, Davis.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer-Verlag.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**, 224–232.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Li, K. H., Raghunathan, T. E., Meng, X. L., and Rubin, D. B. (1991a). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* **1**, 65–92.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991b). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* **86**, 1065–1073.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2005). Mixtures of g-priors for Bayesian variable selection. Tech. Rep. 05-12, ISDS, Duke University.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–326.

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 Persepectives*, 141–152. New York: John Wiley & Sons.

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755–770.

Liu, F. and Little, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.

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

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

Mira, A. and Tierney, L. (2002). Efficiency and convergence properties of slice samplers. *Scandinavian Journal of Statistics* **29**, 1–12.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* **83**, 1023–1036.

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 (Lecture Notes in Computer Science)*, 177–188. New York: Springer-Verlag.

Muralidhar, K. and Sarathy, R. (2006). Data shuffling - A new masking approach for numerical data. *Management Science* **52**, 658–670.

Neal, R. M. (2000). Slice sampling. Tech. rep., Department of Statistics, University of Toronto.

Newton, M. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B* **56**, 3–48.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60**, 739–746.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B* **57**, 99–138.

Panel on Data Access for Research Purposes (2005). Expanding access to research data: Reconciling risks and opportunities.

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.

Perez, J. and Berger, J. O. (2000). Expected posterior prior distributions for model selection. Tech. Rep. 00-08, ISDS, Duke University.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology* **25**, 111–163.

Raghunathan, T. E. (2003). Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach. Tech. rep. Report for the National Academy of Sciences Panel on Access to Confidential Research Data.

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.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

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

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

Reiter, J. P. (2005a). Estimating identification risks for microdata. *Journal of the American Statistical Association* **100**.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

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

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

Reiter, J. P. (2007a). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters (forthcoming)* .

Reiter, J. P. (2007b). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94**, 502–508.

Reiter, J. P. (2007c). Two-stage multiple imputation for correcting measurement error with external validation data. Tech. rep., Department of Statistical Science, Duke Universtiy.

Reiter, J. P. and Drechsler, J. (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. Tech. rep., Institute for Employment Research (IAB).

Reiter, J. P. and Mitra, R. (2007). Estimating risks of identification disclosure in partially synthetic data. Tech. rep., Department of Statistical Science, Duke University.

Reiter, J. P. and Raghunathan, T. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association (forthcoming)* .

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

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

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

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

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

Schafer, J. L., Khare, M., and Ezzati-Rice, T. M. (1993). Multiple imputation of missing data in NHANES III. In *Proceedings of the Annual Research Conference*, 459–487. Bureau of the Census, Washington, DC.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.

Sinharay, S. and Stern, H. S. (2001). Bayes factors for variance component testing in generalized linear mixed models. *Bayesian Methods with Applications to Science, Policy, and Official Statistics* 507–516.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B* **42**, 213–220.

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

Stiratelli, R., Laird, N. M., and Ware, J. H. (1984). Random-effects model for several observations with binary response. *Biometrics* **40**, 961–971.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254–262.

Vines, S., Gilks, W., and Wild, P. (1994). Fitting Bayesian multiple random effects models. Tech. rep., Biostatistics Unit, Medical Research Council, Cambridge.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–648.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Woo, M., Reiter, J. P., Oganian, A., and Karr, A. F. (2007). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* (forthcoming).

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.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*.

Zhao, Y., Staudenmayer, J., Coull, B., and Wand, M. (2006). General design Bayesian generalized linear mixed models. *Statistical Science* **21**, 35–51.

Biography

Satkartar Kinney graduated from the University of California, Berkeley, in 1997 with an A.B. in Mathematics. From 1998 to 2003 she worked in the Environmental Energy Technologies Division of Lawrence Berkeley National Laboratory. In 2002 she completed a M.S. Statistics at California State University, East Bay. She has co-authored the following articles:

1. Kinney, S. K. and J. P. Reiter (2007). Making public use synthetic files of the Longitudinal Business Database. *JSM Proceedings, Government Statistics Section [CD-ROM]*. Alexandria, VA: American Statistical Association.

2. Kinney, S. K. and D. B. Dunson (2007). Fixed and random effects selection in linear and logistic models, *Biometrics* **63**, 690-698.

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