Bayesian Approaches to File Linking with Faulty Data

by

Nicole Dalzell

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Gale Boyd

_____
Katherine Heller

_____
Sayan Mukherjee

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
2017

Abstract

Bayesian Approaches to File Linking with Faulty Data

by

Nicole Dalzell

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Gale Boyd

_____
Katherine Heller

_____
Sayan Mukherjee

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
2017

Abstract

File linking allows analysts to combine information from two or more sources of information, creating linked data bases. From linking school records to track student progress across years (NCERDC 2013), to official statistics (Newcombe 1988) and linking patient health files (Newcombe et al. 1959), linked data bases allow analysts to use existing sources of information to perform rich statistical analysis. However, the quality of this inference is dependent upon the accuracy of the data linkage. In this dissertation, we present methods for file linking and performing inference on linked data when the variables used to estimate matches are believed to be inconsistent, incorrect or missing.

In Chapter 2, we present BLASE, a Bayesian file matching methodology designed to estimate regression models and match records simultaneously when categorical matching variables may not agree for some matched pairs. The method relies on a hierarchical model that includes (1) the regression of interest involving variables from the two files given a vector indicating the links, (2) a model for the linking vector given the true values of the matching variables, (3) a measurement error model for reported values of the matching variables given their true values, and (4) a model for the true values of the matching variables. We also describe algorithms for sampling from the posterior distribution of the model and illustrate the methodology using artificial data and data from education records in the state of North Carolina.

In Chapter 3, we present LFCMV, a Bayesian file linking methodology designed to

link records using continuous matching variables in situations where we do not expect values of these matching variables to agree exactly across matched pairs. The method involves a linking model for the distance between the matching variables of records in one file and the matching variables of their linked records in the second. This linking model is conditional on a vector indicating the links. We specify a mixture model for the distance component of the linking model, as this latent structure allows the distance between matching variables in linked pairs to vary across types of linked pairs. Finally, we specify a model for the linking vector. We describe the Gibbs sampling algorithm for sampling from the posterior distribution of this linkage model and use artificial data to illustrate model performance. We also introduce a linking application using public survey information and data from the U.S. Census Bureau's Census of Manufactures and use LFCMV to link the records.

The linkage techniques in Chapters 2 and 3 assume that all data involved in the linking is complete, i.e., contains no missing data. However, file linking applications can involve files prone to item non-response. The linking application in Chapter 3, for instance, involves a file that has been completed by imputing missing data. In Chapter 4, we use simulated data to examine the impact of imputations in a file linking scenario. Specifically, we frame linking multiply imputed data sets as a two-stage imputation scenario, and, within this context, conduct a simulation study in which we introduce missing data, impute the missing values, and compare the accuracy of estimating C on the imputed versus the fully observed data sets. We also consider the process of performing inference and discuss a Bayesian technique for posterior estimation on multiply-imputed linked data sets. We apply this technique to the simulated data and examine the effect of the imputations upon posterior estimates.

To my parents, my friends, and all who helped me on the way.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

The following symbols are used consistently across chapters.

F_1	File 1; one of the files to be linked.
F_2	File 2; the second file to be linked.
i	Generally denotes a record in F_1.
$i^{'}$	Generally denotes a record in F_2.
C	A permutation vector; this is used as the linkage structure throughout this dissertation.
$InvGamma$	Denotes an Inverse Gamma distribution.

Abbreviations

BLASE	Bayesian Linkage and Simultaneous Estimation; see Chapter 2.
BV	Blocking variable; see Chapter 2.
CMF	U.S. Census of Manufactures; see Chapters 3 - 4.
DPMPM	Dirichlet Process Mixture Model of Products of Multinomials; see Dunson and Xing (2009).
EM	Expectation-maximization; see Chapter 1.
FL	File linking.
FS	Record linkage technique of Fellegi and Sunter (1969); see Chapter 1.
GM	A record linkage technique developed by Gutman et al. (2013); see Chapter 2.

IAC	Industrial Assessment Center; see Chapters 3 - 4 and U.S. Department of Energy (2016)
LFCMV	Linking with faulty continuous matching variables; file linking methodology suggested in Chapter 3.
MAR	Missing at random; a missing data mechanism in which the missing data depends on observed data. See Little and Rubin (2002).
MCAR	Missing completely at random; see Little and Rubin (2002).
MCMC	Markov Chain Monte Carlo; a Bayesian technique for estimating a posterior distribution.
MH	Metropolis-Hasting
MI	Multiple imputation; see Chapter 1.
MIDL	Multiply imputed data linkage; see Chapter 4.
MNAR	Missing not at random; a missing data mechanism in which the missing data depends on observed and unobserved data. See Little and Rubin (2002).
MR	Match rate; an estimate of the percentage of records that are correctly matched after applying a file linking technique.
MV	Matching variable; see Chapter 3.
NCERDC	North Carolina Education Research Centers; see Chapter 2.
PB	Perfectly blocked data; see simulation studies in Chapter 2.
PMR	Posterior match rate; a posterior estimate of the percentage of records that are correctly matched after applying a Bayesian file linking technique.
RL	Record linkage; the process of combining information from two or more files of data.
UI	Unique identifiers; a specific variable that uniquely identifies the individual corresponding to a record. Example: Social Security number; see Chapter 1.

Acknowledgements

I am exceedingly grateful for all the opportunities I have had in the Department of Statistical Science. From faculty members always willing to stop to chat about life or research, to the staff who went out of their way to make sure I had all I needed to succeed, it has been a privilege to be a part of this department. I would especially like to thank my advisor, Jerry Reiter, for his time, guidance and support throughout this process. I would also like to thank Gale Boyd for all of his work during our collaborations. My research work was supported by NSF Grant SES 1131897, and by the Duke University Energy Initiative Energy Research Seed Fund, with co-funding from the Information Initiative at Duke. Data was provided by the North Carolina Education Research Data Center. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

In addition to my research work, this process has provided me opportunities to grow and develop as a statistics educator. I am especially grateful to Mine Çetinkaya-Rundel for her guidance, support and example. Through her efforts, I have had incredible opportunities to teach both online and in person, to develop materials and to discuss teaching strategies. Mine is a phenomenal educator, and I am very privileged to work with her. I am also grateful to the Duke Graduate School and the Certificate of College Teaching program for allowing me to explore new teaching

techniques and philosophies. I would also like to thank The Graduate School for the opportunities presented through the Bass Fellowship and the Preparing Future Faculty Fellowship. These programs allowed me to explore my future as an educator and to engage with students and faculty across the university who are excited about possibilities in teaching.

I could not have made it through this journey without the love and support of my family. My parents, Bill and Debbie, are my rocks in every storm. Thank you for working so hard to help me reach for my dreams, and for cheering me on every step on the way. Thank you to the best brother, Alex, for laughs and a lot of Mario Kart mental health runs. Thank you to my grandparents, and everyone in my wonderful family, for always being there, no matter what.

I have been incredibly blessed to have the support and love of the crazy, caring family that is the Duke Chapel Choir. One never has to look far at a choir event to find a familiar face ready to offer a supportive ear or share a laugh. From Messiah concerts to evening practice, I love every minute of the music we make together. Thank you to Jen for fabulous friendship, for hugs on the hardest days and for laughs that have made up some of the best. Thank you to Amie for being a supportive shoulder even when I didn't know I needed one, for crazy fun and early morning hikes. Thank you to Erin for constant support and encouragement throughout this process. You ladies are amazingly fabulous friends.

I am also grateful for the support and friendship offered by my colleagues. Thank you to Erika for being a wonderful, fun and caring friend. Thank you to Shih-Han, David, Allie, Mary Beth, Ted, Monika, Maria, Jacopo, Doug and Charisse, for sharing their journeys and continuing to be part of mine. Thank you to Jonathan, Lan, Chris and Gary, who walked with me from the first year straight through the end. Thank you to Natasha, Rossie, and Chien-Kuang for lovely weekly lunches as we moved together through the Ph.D. experience at Duke University.

And finally, thank you to every teacher who has helped me along the way, especially Dr. Steven Permuth. Dr. Brendan Nagle, Dr. Catherine Beneteau, Ms. Mirjam Darley and Mr. Michael Brooks. My teachers nurtured my love of learning, encouraged me to strive for my goals and showed me how much of a difference a teacher makes in the life of a student.

1

Introduction to file linking

File linking (FL), or record linkage (RL), is the process of combining information from two or more files of data. This combining, or linking, is executed by selecting a record i in one file and locating a record $i^{'}$ in the second file that corresponds to the same individual. A pair $(i, i^{'})$ corresponding to the same individual is called a match, while any pair $(i, i^{'})$ with records belonging to different individuals is called a non-match. The true match status of a record pair is generally unknown, and because of this, file linking typically involves estimating match status. Pairs $(i, i^{'})$ that are estimated to be matches are called links. Similarly, pairs that are estimated as non-matches are called non-links. The goal of file linking is to assign link status and to correctly identify as many matches as possible.

As large databases become increasingly prevalent, the potential of file linking as a tool to construct databases needed for research is rapidly increasing. From linking school records to track student progress across years (NCERDC, 2013), to official statistics (Newcombe, 1988) and linking patient health files (Newcombe et al., 1959), analyzing linked databases allows researchers to use existing sources of information

to perform rich statistical analysis. However, the quality of this analysis is dependent upon the accuracy of the linkage. Though advances in computing have increased the speed and ease of merging databases, data quality concerns continue to be problematic for linking applications. Winkler (2004) notes three specific quality concerns relevant to file linking: "duplicates, inconsistent information and missing information" (p. 533). These concerns impact both the accuracy with which the files can be linked as well as the quality of inference conducted on the linked data sets.

In this dissertation, we develop techniques for performing Bayesian file linking in the presence of faulty data. We focus on scenarios in which the variables used to estimate link status are believed to be inconsistent, incorrect or missing. In Chapter 2, we develop a Bayesian model for file linking and simultaneous regression parameter estimation when the categorical variables used for linking disagree for some matches. We use both simulated data and North Carolina education data to illustrate the performance of the method. In Chapter 3, we develop a model for linking records using continuous matching variables that may not exactly agree across matched pairs. We present simulation studies to demonstrate the method and apply the methodology to link data from the U.S. Census of Manufactures (CMF) to an external database. This linking is reflective of the reported CMF data but ignores the reality that, for many records, the values of the matching variables are imputations created by the U.S. Census Bureau. In Chapter 4, we revisit this application to consider the implications of linking when the matching variables are imputed. We use simulation studies to illustrate potential implications of linking files with imputed matching variables. We also discuss a Bayesian technique for performing analysis with linked, multiply imputed data sets. We conclude with a discussion of future research directions.

The remainder of this chapter serves as an overview of some core concepts of file linking. The foundation presented here is critical for the proceeding chapters. In

Section 1.1, we review fundamental concepts of probabilistic record linkage, including the work of Fellegi and Sunter (1969). In Section 1.2, we introduce the FL technique of blocking. In Section 1.3, we discuss techniques for estimating the quality of a linked data product. In Section 1.4, we discuss missing data in the context of FL and presents an overview of multiple imputation techniques (Rubin, 1987). In Section 1.5, we overview techniques for performing statistical analyses on linked data sets.

1.1 Foundations in file linking

When the match status of a record pair is unknown, file linking is conducted using a set of criteria to determine which pairs of records are possible matches. These criteria vary widely across applications and data types. In deterministic linkage, link status is assigned based on a set of fixed rules. In probabilistic linkage, match weights or other quantities are estimated and used to represent the probability that two records are a match. In this section, we review both deterministic and probabilistic approaches to file linking and explain the benefits and challenges involved with each linkage type.

1.1.1 Deterministic linkage

Records are composed of fields, such as first name, school, sex, or birth date. Deterministic linkage generally involves requiring exact agreement on some combination of these fields. In an ideal scenario, linkage can be conducted using a single field that corresponds to a unique identifier (UI). UIs are specific variables that uniquely identify the individual corresponding to each record; two examples are social security numbers and student identification numbers. In the presence of accurate UIs, file linking can be reduced to a deterministic task of merging together records with a common UI (Jaro, 1989).

Without UIs, deterministic linkage typically relies on examining a set of fields common to both files. These fields are referred to as key variables, comparison fields

or matching variables. Matching variables may contain a mixture of categorical, continuous, or string information. For instance, the North Carolina Education Research Data Center (NCERDC) utilizes matching variables including birth date, first name, and last name to link student records (NCERDC, 2013). A pair of records $(i, i^{'})$ is declared a link if there is agreement on the matching variables.

With advances in computing technology, deterministic linkage is often a straight forward option for file linking that can be executed relatively quickly. However, the choice to use deterministic linking relies heavily on the assumption that the UIs or matching variables used for linking are both accurate and agree across all matched pairs. In practice, matching variables may be faulty, i.e., the field values for true matched pairs may disagree (Newcombe et al., 1959). Examples of processes that can introduce faults include typographical errors (Jaro, 1989); actual changes in field value, such as a name change that occurs after a marriage and; discrepancies in the coding system used to record information (Winkler, 2004). For example, one corporation may use one set of codes to represent a product, while a second corporation may use another. Directly examining values of the matching variable for "product" will fail to identify some matches; matches that are classified as non-links are called missed matches. Without UIs or accurate matching variables, a high rate of missed matches is a common trait of deterministic linkage.

1.1.2 Probabilistic linkage

In the absence of UIs, probabilistic linkage is an alternative to deterministic approaches. Probablistic linkage uses matching variables or other non-unique identifiers to estimate match status for records (Winkler, 2015). The conceptual framework for probablistic linkage was laid by Newcombe et al. (1959) and Fellegi and Sunter (1969). Newcombe et al. (1959) propose the use of computer generated matching weights (or scores) as a means of determining how likely it is that two records are

a match when their matching variable values are not identical. These weights are an example of a similarity metric, a numerical quantity that determines the similarity between records. Defining such a metric is a key facet of a probabilistic linkage technique (Belin and Rubin, 1995).

The approach of Newcombe et al. (1959) assigns a match weight to each possible pair of records $(i, i^{'})$, where this weight represents the likelihood that the pair is a match. The match weight is composed of separate components, each representing the contribution of one matching variable. Agreement on a matching variable for a pair results in an increase in the match weight while disagreement results in a decrease in the weight. The magnitude of this increase or decrease is specific to the relevance of each field for determining match status. Agreement on surname, for instance, may be given higher importance than agreement on state of residence.

The work of Newcombe et al. (1959) provides the conceptual foundation for probabilistic record linkage, and these concepts are formalized into a theoretical framework by Fellegi and Sunter (1969). This file linking procedure, which we call FS, consists of two main components: (1) a measure of similarity between two records, and (2) a decision rule that determines how to classify pairs as links or non-links based on this measure (Belin and Rubin, 1995). More formally, consider two files F_1 and F_2. The goal of file linking is to partition the pairs in the product space $F_1 \times F_2$ into two disjoint subsets, M and U. M represents the set of true matches and U contains the true non-matches. The belief that file linking is feasible implies that some individuals in F_1 also have information in F_2, and hence M is non-empty.

Newcombe et al. (1959) use relative frequencies as a means of computing match weights for patterns of agreement in record pairs. Fellegi and Sunter (1969) consider a more general agreement pattern $\boldsymbol{\gamma}$ with possible values contained in a space Γ. For a set of matching variables, $\boldsymbol{\gamma}$ might represent a binary vector of agreement/disagreement patterns, the relative frequencies of field values, or some mea-

sure of distance between one field value and another. These vectors $\boldsymbol{\gamma}$ of agreement patterns are used in the FS approach to compute a ratio

$$R = \frac{P(\boldsymbol{\gamma} \in \Gamma | M)}{P(\boldsymbol{\gamma} \in \Gamma | U)} = \frac{m}{u}. \tag{1.1}$$

Fellegi and Sunter (1969) note that R or any monotonically increasing function of R may be used in file linking as the matching weight. Here, m is the probability of observing a particular value of $\boldsymbol{\gamma}$ given that a pair $(i, i^{'})$ is a match, while u is the probability of $\boldsymbol{\gamma}$ given that the pair is non-match.

Once a matching weight has been computed, the second task in a RL procedure is to make a decision about match status for each record pair conditional on its matching weight. In the FS approach, this decision rule incorporates two matching weight thresholds, an upper threshold and a lower threshold; Belin and Rubin (1995) note that FS is sensitive to the choice of thresholds. The matching weights for each record pair in $F_1 \times F_2$ are compared against these two threshold values in order to assign match status. Pairs associated with matching weights above the upper threshold are declared links, while pairs with matching weights below the lower threshold are declared non-links. If the weight for a given pair falls between the two thresholds, the pair is declared a possible link and clerical review is used to determine final match status. As noted in Newcombe et al. (1959), manual clerical review is time consuming and expensive, and hence the choice of optimal threshold for FS is designed to minimize the number of records assigned to this "possible link" status (Fellegi and Sunter, 1969).

Computing R requires the probabilities m and u from (1.1). Lahiri and Larsen (2005) note that a practical issue associated with FS is that these matching probabilities are conditional on match status, which is unknown. The probabilities m and u must therefore be estimated in order to estimate R. Fellegi and Sunter (1969) provide a means of estimating R under a conditional independence assumption. This

assumption states that, conditional on match status, agreement patterns on each matching variable are independent of agreements patterns on the other matching variables. For example, in the case of six matching variables, this means

$$\begin{aligned} P\big[\boldsymbol{\gamma} = (1,1,1,1,1,0)|x\big] &= P(\gamma_1 = 1|x) \times P(\gamma_2 = 1|x) \times P(\gamma_3 = 1|x) \\ &\quad \times P(\gamma_4 = 1|x) \times P(\gamma_5 = 1|x) \times P(\gamma_6 = 0|x), \end{aligned} \tag{1.2}$$

where x denotes match status, i.e., $x \in \{M, U\}$. However, this independence assumption can fail in many RL applications (Winkler, 1993b; Thibaudeau, 1993). Winkler (1993b) presents a general expectation-maximization (EM) algorithm that can be used to estimate m and u without the conditional independence assumption. This work allows the FS framework to be applied to Census and other practical applications (e.g., Jaro, 1989). Other work on using general EM methodology for estimating match weights include Winkler (1988) and Meng and Rubin (1993).

Fellegi and Sunter (1969) underline that the theoretical goal of file linking is to divide records into two disjoint subsets M and U such that M contains all true matches and U contains all true non-matches. The idea that a record must belong to one of two sub-populations readily translates into a mixture model framework for file linking in which records are divided into two mixture components, M and U (Winkler, 1993b; Larsen, 2004; Lahiri and Larsen, 2005). From such a perspective, the probability mass function for the comparison vector $\boldsymbol{\gamma}$ can be expressed as

$$P(\boldsymbol{\gamma}) = \sum_{h \in \{M,U\}} \pi_h P(\boldsymbol{\gamma}|h), \tag{1.3}$$

where π_h represents the probability that a record pair belongs to mixture component h for $h \in \{M, U\}$. $P(\boldsymbol{\gamma}|h)$ is the probability of observing comparison vector $\boldsymbol{\gamma}$ for a record pair in class h. In the FS approach, $P(\boldsymbol{\gamma}|h = M) = m$, and similarly, $P(\boldsymbol{\gamma}|h = U) = u$. Jaro (1989), Winkler (1993a), Winkler (1993b), and Thibaudeau (1993) use mixture models to apply the FS approach to U.S. Census data.

Also using a finite mixture model framework, Larsen and Rubin (2001) present an iterative procedure for record linkage. This procedure seeks to lessen the burden of performing clerical review on all records assigned "possible link" status by the FS approach. In addition to the three decision classes specified by FS, Larsen and Rubin (2001) create a fourth decision category by splitting the category of "possible links". A subset of the possible links is chosen to be sent to clerical review, while the remaining "possible link" records are not. For the selected subset, clerical review is performed to assign each record to status M or status U. Using this newly assigned status, the records are then shifted from the "possible link" category to category M or U. Based on these updated M and U subsets, matching probabilities are re-estimated, and the model to assign match status is refit. This iterative process of identifying links, sending a subset to clerical review, and then fitting the model is repeated until clerical review identifies very few new links.

In the FS file linking approach, record pairs are assigned a match status based on the decision thresholds. However, this rule does not stipulate that each record i in F_1 may be declared a match with at most record in $i^{'}$ in F_2. If, for distinct $i_1^{'}, i_2^{'}$ in F_2, record pairs $(i, i_1^{'})$ and $(i, i_2^{'})$ both have matching weights above the threshold, both pairs can be declared links. Many FL techniques handle this by imposing a one-to-one constraint that forces at most one link per record. Jaro (1989), for instance, imposes such a constraint within a linear-sum decision threshold that maximizes the sum of all composite weights associated with linked pairs. Choosing a decision threshold in such a manner is synonymous with a linear optimization problem with constraints that enforce one-to-one matching (Winkler, 1999). Winkler (1993a) introduces a variant of this algorithm that is more computationally efficient in practice.

The framework presented by Fellegi and Sunter (1969) continues to be used in many modern file linking applications, but some challenges inherent in this methodology have led to the development of multiple techniques for file linking. Many of these

techniques are inspired by applications in the U.S. Census Bureau (e.g., Jaro, 1989; Winkler and Thibaudeau, 1987; Scheuren and Winkler, 1997). These applications involve imperfect data in which missing data, typographical errors, or other factors complicate the linking process. For a detailed overview of historical and modern probablistic techniques, see Winkler (1999) and Winkler (2015).

1.2 Blocking

Techniques like FS theoretically require comparing every possible pair of records in $F_1 \times F_2$. Fellegi and Sunter (1969) and Lahiri and Larsen (2005) note that the size of this product space is often too large for such exhaustive comparison to be practical. Additionally, it is possible that certain pairs of records can be eliminated as possible links in advance of performing FS. For instance, if we are certain of the age of a given individual, it is prudent to limit our search for a match to records with that age.

Limiting the space of possible matches before linking is an application of a standard FL technique known as blocking. Blocking generally involves grouping records together based on some common characteristics, such as state of residence or race. Clustering techniques or other similarity metrics also may be used to define blocks. Once a record i is placed within a block k, every record in any block $k' \neq k$ is declared a non-match for record i. This reduces the number of comparisons from all pairs in $F_1 \times F_2$ to the number of possible pairs present within each of the K blocks. An overview of blocking techniques and their application to RL is presented in Steorts et al. (2014a); we will briefly summarize some of these here.

Perhaps the most basic type of blocking (which Steorts et al. (2014a) refer to as traditional blocking) is partitioning records into blocks based on agreement on certain shared variables. We refer to these variables as blocking variables, or BVs. In Jaro (1989), for instance, blocks are defined by Census geographic block and SOUNDEX codes of individual surnames. Traditional blocking with a set of categorical BVs

is the starting point for LFCMV in Chapter 3. Traditional blocking relies upon BVs that are reliable, meaning that BVs for true matches tend to agree. However, Fortini et al. (2001) note that, "in large data sets it is very likely to have errors in the answers, typing errors and inconsistencies between answers to the same question in different times" (p. 188). The presence of such errors in the BVs results in an increase in false non-match rates, as records in different blocks are declared non-links a priori. BLASE accounts for such errors through an edit-imputation scheme, which estimates the location of faults in the BV fields and proposes new values for these fields. This model will be described in detail in Chapter 2.

In addition to requiring high accuracy, traditional blocking requires BVs that divide records into fairly small blocks. Blocks that are too large are computationally inefficient and present a greater chance of linking records that are not true matches. As noted in Fortini et al. (2001), however, very small blocks increase the number of records which are a priori declared non-matches and hence, in the presence of errors in the BVS, increase false non-match rates.

Alternative blocking techniques include sorted neighborhood clustering for privacy-preserving record linkage (Vatsalan and Christen (2013)); locally sensitive hashing (Kim and Lee, 2010; Steorts et al., 2014b); and nearest-neighbor clustering (Karakasidis and Verykios, 2012).

1.3 Quantifying linkage quality

The techniques discussed in Section 1.1 estimate match status for pairs of records. In this section, we discuss a few techniques for evaluating the quality of estimated matches.

One metric used to evaluate the quality of a linked data product is the classification error rate, or the probability that records are placed incorrectly into M or U. In the Larsen and Rubin (2001) approach, these probabilities are estimated as part of

the model. For each record i assigned to latent class U, sum up the probability that each of these records belongs to class M. Dividing by the total number of records in class M yields an approximate classification error rate.

A second metric to evaluate a linked data product is the false-match rate (FMR), or the probability of linking two records when they are a non match. Procedures for estimating false-match rates are presented in Fellegi and Sunter (1969) and Newcombe (1988), but these techniques rely heavily on assumptions of independent fields that are not realistic for some data (Belin and Rubin, 1995). In order to avoid the unrealistic conditional independence assumption, Belin and Rubin (1995) utilize calibration data as a way to estimate false-match rates of a file linking procedure. Calibration data refers to data that have been carefully reviewed to assure proper match status. In this dissertation, we refer to such data as seed data. The quantity and quality of seed data are critical to the success the BLASE methodology in Chapter 2. An examination of Belin and Rubin (1995)'s calibration technique by Winkler (1999) notes that the FMR rate tends to be well estimated by the calibration technique when "there is substantial separation of the curve of the log frequency versus matching weight for matches and non matches" (Winkler 1999, p. 4).

Bayesian techniques for FL, including the work of Larsen and Rubin (2001), Fortini et al. (2001), Larsen (2004), Tancredi and Liseo (2011), Steorts et al. (2014b) and Gutman et al. (2013), utilize posterior distributions to quantity some of the uncertainty accrued during the linkage process. In particular, a measure of certainty in links can be quantified via the posterior probabilities that any given record pair is linked. Fortini et al. (2001) note that many of the computational complexities involved in frequentist FL techniques (e.g., enforcing one-to-one matching, parameter estimation in finite-mixture model settings) can be more readily handled in Bayesian settings through careful prior choices and posterior computations. For example, Tancredi and Liseo (2011) present a Bayesian approach to file linking that uses a

hierarchical approach by considering noisy continuous data in a capture-recapture scenario. In addition to estimating links, this approach simultaneously estimates the number of unique individuals present in the combined data sets. Bayesian techniques also allow match probabilities to be defined conditionally upon other matched pairs through the process of posterior updating (Fortini et al., 2001).

1.4 Missing data

In this section, we discuss a common data quality issue that affects file linking: missing data. Missing data in a file linking scenario can take a variety of forms. Individual records in one or both of the files may be missing information on specific fields. This unit missingness is also common in scenarios outside of file linking, and Little and Rubin (2002) provide a thorough overview of handling missing data in statistical analyses. File linking itself can also be formulated as missing data problem (e.g., Wu, 1995; McGlincy, 2004; Gutman et al., 2013). The desired quantity in FL is a linked data set. Suppose we are linking F_2 to F_1. Prior to linking, the information from file F_1 is present, but the information from F_2 is missing. This missingness is described by a linkage structure, a construct that declares how records in F_1 and F_2 will be linked. The goal of FL is to estimate this structure, i.e., to identify matched pairs, in order to address the missingness. A commonly used linkage structure is a binary matrix (Fortini et al., 2001; Tancredi and Liseo, 2011), in which a value of 1 in cell $(i, i^{'})$ indicates that records $i \in F_1$ and $i^{'} \in F_2$ correspond to the same individual.

In the context of unit-missingness or file linkage as a missing data problem, one approach to handling missing data is to impute the missing values. In this section, we review some concepts fundamental to missing data imputation. In Section 1.4.1, we review missing data mechanisms. In Section 1.4.2, we present an overview of multiple imputation and in Section 1.4.3, we review the techniques of Rubin (1987) for obtaining estimates of parameters using multiply imputed data sets. For an in-

depth exploration of missing data and imputation techniques, see Little and Rubin (2002) and Rubin (1976).

1.4.1 Missing data mechanisms

Imputing missing data relies on assumptions about the mechanisms that cause the data to be missing (Rubin, 1976). Three primary classes of missing data mechanisms are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little and Rubin, 2002). Let $Y = (y_1, \ldots, y_n)$ denote a data vector of length n, and let R be a binary vector of length n such that $R_i = 1$ if and only if y_i is missing. Here, R is a missing-data indicator. Characterizing the missing data mechanism involves the conditional distribution $f(R|Y, \phi)$, where ϕ denotes a set of unknown parameters related to R. We say data is MCAR if for all Y, ϕ,

$$f(R|Y, \phi) = f(R|\phi). \tag{1.4}$$

Here, (1.4) indicates that the reason that any data point y_i is missing does not depend on the data Y (Little and Rubin, 2002).

Now consider splitting Y into two components, the vector Y_{obs} of observed values and Y_{mis} of missing values; here $Y = (Y_{obs}, Y_{mis})$. We say data is MAR if for all Y_{mis}, ϕ,

$$f(R|Y, \phi) = f(R|Y_{obs}, \phi). \tag{1.5}$$

Here, (1.5) implies that the reason Y_{mis} is missing depends on the observed data, but not on the quantities of the missing values. If the reason the data is missing depends on the missing values, we say data is MNAR. These terms will be applied in Chapter 2 to describe faulty values of categorical comparison fields, as well as to describe simulation studies in Chapter 3.

1.4.2 Multiple Imputation

One option for handling missing data is imputation, or sampling values for record fields that are missing. The U.S. Census Bureau, for instance, uses a variety of techniques to fill in missing data in the Census of Manufactures (White, 2014). However, imputing a single value for a missing data point may under-represent the uncertainty in that value (Little and Rubin, 2002). An alternative is to create $m > 1$ imputations for each missing data point, resulting in m completed data sets reflecting potentially different values for the missing fields (Rubin, 1987). For unit missingness in file linking, this process of multiple imputation (MI) can be applied to correct for missing data within a given file. In Chapter 4, we create multiple imputations for the 2007 CMF using classification and regression trees and explore the potential impact of multiply imputed data in a file linking scenario.

If we consider FL itself from a missing data perspective, we can also create multiple imputations of the unknown linkage structure. Bayesian methodology and specifically Bayesian multiple imputation naturally facilitates the creation of multiple data sets through posterior sampling. Each posterior sample reflects a potentially different linking structure. Rather than relying on a single linked data set that ignores the matching uncertainty, multiple imputations facilitate comparisons of likely links for particular records, and allow examination of the frequency of similar links across the imputations. The concept of file linking as a multiple imputation scenario appears throughout in this dissertation.

1.4.3 Estimation with Multiple Imputation

Rubin (1987) provides steps for obtaining single parameter estimates from multiply imputed data sets. Suppose Q represents some scalar quantity of interest, like a regression coefficient or a mean. The intended statistic is $Q = Q(Y)$, with variance $U = U(Y)$. With complete data, estimates for Q and U would be $\hat{Q}$ and $\hat{U}$,

respectively. Assume that

$$(Q - \hat{Q})/U \sim N(0, 1). \tag{1.6}$$

For each of the m completed data sets, we estimate Q with $\hat{Q}^{(m)}$ and U with $\hat{U}^{(m)}$. Rubin (1987) describes how to combine these m estimates to obtain inferences for Q. Specifically, we need the following quantities:

$$\bar{Q}_m = \sum_{\ell=1}^{m} Q^{(\ell)}/m \tag{1.7}$$

$$B_m = \sum_{\ell=1}^{m} (Q^{(\ell)} - \bar{Q}_m)^2/(m-1) \tag{1.8}$$

$$\bar{U}_m = \sum_{\ell=1}^{m} U^{(\ell)}/m. \tag{1.9}$$

$\bar{Q}_m$ is used by the analyst as a point estimate for Q. Because this quantity uses point estimates from each of the m multiply imputed data sets, estimating the variance of Q requires considering variation within each data set (1.9) as well as between the data sets (1.8). More specifically, $\bar{U}_m$ estimates the variance of Q if the data were complete, and B_m captures the increase in the variance of Q that can be attributed to the missing data (Reiter and Raghunathan, 2007). The estimated total variance for Q can be expressed as

$$T_m = \left(1 + \frac{1}{m}\right) B_m + \bar{U}_m. \tag{1.10}$$

Test statistics and MI confidence intervals are based off of a student-T distribution,

$$(\bar{Q}_m - Q)/\sqrt{T_m} \sim t_\nu, \tag{1.11}$$

with degrees of freedom $\nu = (m-1)\big[1 + \bar{U}/((1+m^{-1})B)\big]$. For further information on estimation in multiple imputation scenarios, see Rubin (1987) and Reiter and Raghunathan (2007).

1.5 Post-linking analysis

In this section, we discuss techniques for conducting analysis on linked data sets. As an estimation procedure, file linking is subject to error, and these errors can result in biased estimated for analyses conducted using linked data (Scheuren and Winkler, 1993; Lahiri and Larsen, 2005; Chambers, 2009; Bohensky, 2015). For an extensive discussion on secondary analysis on linked data, see Chambers and Kim (2015).

If we approach file linking as a missing data scenario, the general MI framework presented in Section 1.4.2 provides a means of reflecting uncertainty in the estimation of the linkage structure into post-linking analyses. Let Y contain information (Y_1, Y_2), where $Y_1 \in F_1$ and $Y_2 \in F_2$. Without loss of generality, suppose we match to F_1 and let the missing information be Y_2. The data are missing because the linkage structure to connect Y_1 and Y_2 is unknown. In the MI framework, Y_2 are "imputed" at each iteration by estimating a linkage structure. Conditional on this structure, we link F_1 and F_2, effectively filling in the missing data. Each of the m completed data sets reflects a potentially different linkage structure. Statistical analyses can be run on these m data sets with estimates combined using (1.7)-(1.9).

In the absence of multiply imputed linked data sets, statistical analysis is typically run upon one data set reflecting a single estimate of the linkage structure. Neter et al. (1965) study the impact of such post-linking analysis in finite population sampling. They demonstrate that even small errors in linking can lead to bias in "estimating the relationship between response errors and the true values" (Lahiri and Larsen 2005, p. 222). Neter et al. (1965) present a potential correction to adapt for this bias, but as noted in Scheuren and Winkler (1993), the assumptions necessary to apply this correction make the technique impractical for real world applications.

Building upon the findings of Neter et al. (1965), Scheuren and Winkler (1993) present a framework for correcting ordinary least squares (OLS) estimates for linear

regression models run on linked data which contain some mismatches. The model of interest is a multivariate regression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. In a linkage context, the true pairs (x_i, y_i) are not necessarily observed. Instead, linear regression run on linked data may be expressed as fitting the model

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.12}$$

where

$$z_i = \begin{cases} y_i & : \text{with probability } q_{ii} \\ y_j & : \text{with probability } q_{ij}, i \neq j, j = 1, ..., n, \end{cases} \tag{1.13}$$

where $\sum_{j=1}^{n} q_{ij} = 1$ to enforce one-to-one matching. Scheuren and Winkler (1993) explore the bias associated with naively estimating $\boldsymbol{\beta}$ from (1.12), i.e., with $\hat{\boldsymbol{\beta}}_N = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$. This is the estimator an analyst would obtain if he or she used standard statistical methods to obtain an OLS fit on the linked data.

The expected value of z_i is $E(z_i) = \mathbf{w}_i^T\boldsymbol{\beta}$, where the column vector $\mathbf{w}_i = \sum_{j=1}^{n} q_{ij}\mathbf{X}_j^T$ is a weighted average of the predictors. Scheuren and Winkler (1993) compute the bias of the naive estimator $\hat{\boldsymbol{\beta}}_N$ as

$$\begin{aligned} Bias(\hat{\boldsymbol{\beta}}_N) &= E[\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}] \\ &= \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} - I\right)\boldsymbol{\beta} \\ &= \left(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}\mathbf{X} - I\right)\boldsymbol{\beta}. \end{aligned} \tag{1.14}$$

This shows that, in general, $\hat{\boldsymbol{\beta}}_N$ is not an unbiased estimator of $\boldsymbol{\beta}$, meaning that standard OLS techniques on linked data sets may lead to biased estimates of the regression coefficients. This underscores the need to account for the uncertainty accrued throughout the linkage process in post-linking statistical analysis.

Scheuren and Winkler (1993) provide an alternative to the naive estimator $\hat{\boldsymbol{\beta}}_N$ that has lower bias. Following the notation of Lahiri and Larsen (2005), we denote

this estimator $\hat{\boldsymbol{\beta}}_{SW}$. Suppose the true value of $\mathbf{Y}$ is known. Then

$$Bias(\hat{\boldsymbol{\beta}}_N|\mathbf{Y}) = E[\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right)|\mathbf{Y}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{B}, \tag{1.15}$$

where $\boldsymbol{B} = (B_1, ..., B_n)^T$, with $B_i = q_i^T Y - y_i$. Here, B_i is a difference between an average of $\mathbf{Y}$, weighted with the probability that each y_i was chosen as z_i for record i , and the true y_i. Conditional on an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$, the estimator of Scheuren and Winkler (1993) is

$$\hat{\boldsymbol{\beta}}_{SW} = \hat{\boldsymbol{\beta}}_N - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{B}}. \tag{1.16}$$

The authors suggest estimating $\mathbf{B}$ such that each B_i is estimated by

$$\hat{B}_i = (q_{ij_1} - 1)z_{j_1} + q_{ij_2}z_{j_2}. \tag{1.17}$$

Here, q_{ij_1} and q_{ij_2} denote the two highest elements of the vector $\mathbf{q}_i$, and z_{j_1} and z_{j_2} represent the corresponding elements of $\mathbf{z}$. In other words, the estimate uses only the two most likely values for y_i to estimate B_i. Lahiri and Larsen (2005) suggest using $s > 2$ candidate links, enough so that $\sum_{r=1}^{s} q_{ij_r} < p$ for some chosen $0 < p \leqslant 1$.

Lahiri and Larsen (2005) suggest an unbiased estimator,

$$\hat{\boldsymbol{\beta}}_U = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{z}, \tag{1.18}$$

where $w_i = \mathbf{q}_i^T\boldsymbol{X} = \sum_{j=1}^{p} q_{ij}X_j^T$. They explore the performance of both $\hat{\boldsymbol{\beta}}_U$ and $\hat{\boldsymbol{\beta}}_{SW}$, citing that both provide "useful" corrections in a variety of situations. Scheuren and Winkler (1997) further research linear regression coefficient estimates with linked data and present an iterative procedure that adjusts coefficient estimates for outliers.

The performance of (1.18) depends upon the assumption of homoskedastic errors (Chambers, 2009). Chambers (2009) explores this reality and presents several alternative frameworks for linear regression on linked data. Their work extends the notion of regression analysis on linked data to the more general setting of estimating functions on linked data and provides a detailed framework for adjusting standard

statistical analyses for applications involving linked data. Further work in Samart and Chambers (2014) develops techniques for performing linear regression in the presence of nested errors, or linear mixed effects models. Samart and Chambers (2014) provides modifications for Analysis of Variance (ANOVA), maximum likelihood, and restricted maximum likelihood methods for linked data with this framework.

1.6 Summary

In this chapter, we have provided an overview of several key concepts in file linking. We have also provided information in multiple imputation and other concepts that will be applied in the remaining chapters of this dissertation. We assume familiarity with these concepts when presenting material in the proceeding chapters.

2

BLASE: File linking and estimation with faulty categorical matching variables

In this chapter, we present a Bayesian file matching methodology designed to estimate regression models and match records simultaneously when categorical variables used for matching may be faulty. The method relies on a hierarchical model that includes (1) the regression of interest involving variables from the two files given a vector indicating the links, (2) a model for the linking vector given the true values of the matching variables, (3) a measurement error model for reported values of the matching variables given their true values, and (4) a model for the true values of the matching variables. We describe algorithms for sampling from the posterior distribution of the model and illustrate the methodology using artificial data and data from education records in the state of North Carolina. A version of this chapter appears on arXiv as the paper "Regression Matching and File Linking Using Possibly Erroneous Matching Variables" (Dalzell and Reiter, 2016).

2.1 Introduction

Increasingly, analysts seek to integrate data from two files comprising overlapping sets of individuals, for example, by linking subjects in planned studies to administrative databases like electronic health records or tax records. Often analysts are unable to base linkages on unique identifiers, such as Medicare identification numbers or social security numbers, because they are unavailable on at least one of the files, e.g., due to privacy requirements. Analysts instead may have to base linkages on combinations of categorical variables available in both files, such as demographic variables (e.g., age, race, sex) or group membership variables (e.g., county of residence, hospital, school attended). Depending on the richness and quality of these categorical matching variables, this file matching may not result in unique links for every record across the files.

In this chapter, we consider file matching settings where the ultimate goal is to regress some outcome Y_1 available in file F_1 on a set of predictors Y_2 available in the other file F_2 (and possibly other predictors available in F_1). In this setting, the most obvious approach is to base regression inferences on only the linked data from the unique matches. However, this discards information from the other records, resulting in inefficient and possibly even biased estimates. An alternative is the approach of Gutman et al. (2013), which simultaneously estimates the regression and performs linkages using the information from all records in the files; we describe this model, which we call the GM, in detail in Section 2.2. The GM not only takes advantage of information from non-unique matches, it also leverages relationships among Y_1 and Y_2 to improve the linking process.

The GM makes an assumption that may not hold in practice, namely that an individual present in both files should have the same values of the matching variables in those files. This may not be the case because of, for example, typographical

errors (Jaro, 1989), discrepancies in the coding systems used to record information (Winkler, 2004), and legitimate changes in field values, such as a student who moves to a new school. Such faulty data values can generate incorrect matches, which are well known to adversely impact the results of regression based on linked files (e.g., Neter et al., 1965; Scheuren and Winkler, 1993, 1997; Lahiri and Larsen, 2005; Chambers, 2009; Samart and Chambers, 2014).

We extend the GM to handle scenarios where the categorical matching variables, which we call BVs, may be faulty. The basic idea is to use a Bayesian hierarchical model that includes four layers. First, we specify a model for the regression of interest conditional on a vector C that indicates which records are linked across files. Second, within each block formed by the BVs, we specify a model for the corresponding elements of C. Third, we specify a measurement error model for the observed BV values. The measurement error model comprises a model for the probability that an observed BV field is faulty coupled with a reporting error model that describes the relationship between the observed and latent, true BV values. Finally, we specify a model for the true BV values using a mixture of products of multinomial distributions (Dunson and Xing, 2009), also known as a latent class model. This model is discussed in detail in Appendix A. We call the full hierarchical model BLASE, which stands for Bayesian linkage and simultaneous estimation. BLASE can be used to obtain posterior inferences for the regression model parameters, as well as for C.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce BLASE and indicate how it differs from GM. In Section 2.3, we describe algorithms for sampling from the posterior distribution of BLASE. In Section 2.4, we derive the acceptance ratio for the Metropolis-Hastings step of the posterior sampling algorithm. In Section 2.5, we use simulation studies to investigate the performance of BLASE, comparing it to the performance of GM across a range of illustrative scenarios. In Section 2.6, we apply BLASE to data on students in public schools in North Carolina.

In Section 2.7, we conclude with a summary of findings.

2.2 The BLASE model

After defining notation and key concepts in Section 2.2.1, we describe the sub-models in BLASE in Sections 2.2.2 to 2.2.5, followed by some intuition about the model in Section 2.2.6.

2.2.1 Notation and key concepts

Let the two files to be integrated, F_1 and F_2, comprise n_1 and n_2 records, respectively. We assume each record corresponds to a single true individual; see Steorts et al. (2014b) for linking data with replicates. For simplicity of notation, we assume that the matching is based on all BVs common to both files. However, this is not necessary for BLASE or GM.

Let B_{fij} denote the true BV value for record i in file f on field j, where $j = 1, \ldots, J$, $i = 1, \ldots, n_f$, and $f = 1, 2$. Let $\hat{B}_{fij}$ be the reported value of B_{fij}; these values could differ from B_{fij}. For all (f, i), let $\hat{B}_{fi} = (\hat{B}_{fi1}, \ldots, \hat{B}_{fiJ})$, and let $\hat{\boldsymbol{B}}_f = \{\hat{B}_{fi} : i = 1, \ldots, n_f\}$. We define B_{fi} and $\boldsymbol{B}_f$ analogously. Finally, let $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{B}}_1, \hat{\boldsymbol{B}}_2)$ and $\boldsymbol{B} = (\boldsymbol{B}_1, \boldsymbol{B}_2)$.

Through validation, data cleaning, or other techniques, it is typically possible to determine with high certainty that a subset of the records are not faulty on some or all of the J fields. A record is a *seed* on field j if it is known *a priori* that $\hat{B}_{fij} = B_{fij}$. If the relationship between $\hat{B}_{fij}$ and B_{fij} is unknown *a priori*, then i is a *non-seed* on field j. When i is a seed for all J fields, we call i a Type 2 (T_2) seed. The GM assumes that all records are T_2 seeds, whereas BLASE does not. We also define Type 1 (T_1) seeds to be records where it is known that records $i \in F_1$ and $i' \in F_2$ belong to the same individual, and that $\hat{B}_{fi} = B_{fi}$ for $f = 1, 2$. When it is known that an (i, i') pair belong to the same individual yet $\hat{B}_{1ij} \neq \hat{B}_{2i'j}$ for some j, the analyst can

select one of $(\hat{B}_{1ij}, \hat{B}_{2i'j})$ to be correct. As we shall discuss, the information in the T_1 and T_2 seeds is crucial for helping BLASE (and GM) estimate model parameters reliably.

Each unique combination of BVs, whether through $\hat{\boldsymbol{B}}$ or $\boldsymbol{B}$, can be used to define a block. For example, for BVs of sex, race and school, all records identified as belonging to female Caucasian students who attend the same school are placed in the same block k. The GM forms blocks from $\hat{\boldsymbol{B}}$, whereas BLASE forms blocks from the latent $\boldsymbol{B}$.

For a given $\boldsymbol{B}$, let $n_{1(k)}$ and $n_{2(k)}$ denote the number of records in block k from F_1 and F_2, respectively, that are not T_1 seeds. In general, $n_{1(k)}$ need not equal $n_{2(k)}$. Following Gutman et al. (2013), in each block k we match all records from the smaller sized file for that block to records from the larger sized file for that block; we treat the unmatched records as if they were missing at random from the smaller file. Thus, for the given $\boldsymbol{B}$, the total number of unique individuals in any block k is $n_{(k)} = c_{(k)} + u_{(k)}$, where $c_{(k)} = \max(n_{1(k)}, n_{2(k)})$ and $u_{(k)}$ is the number of T_1 seeds in block k. The total number of unique individuals in $F_1 \cup F_2$ is $N_B = \sum_{k=1}^{K} n_{(k)}$, where K is the number of unique blocks formed by $\boldsymbol{B}$. In GM, each $n_{(k)}$ and N_B are fixed since we assume $\boldsymbol{B} = \hat{\boldsymbol{B}}$; in BLASE, these can change with different plausible values of $\boldsymbol{B}$.

Let F_{1k} and F_{2k} represent the records in block k from F_1 and F_2, respectively. Without loss of generality, within each block k fix the ordering of the records from F_{1k}. For each block k, we define a $n_{(k)}$ element permutation vector C_k that specifies how records from F_{2k} must be re-organized to match their corresponding records in F_{1k}. To illustrate, suppose that $n_{1(k)} = n_{2(k)} = 2$ and we link the first record in F_{1k} to the second record in F_{2k}. Then, $C_k = (2, 1)$.

When $n_{1(k)} > n_{2(k)}$, we create $n_{1(k)} - n_{2(k)}$ dummy records and add them to

F_{2k}. We then match $n_{2(k)}$ records from F_{1k} to actual records from F_{2k}, and match the remaining $n_{1(k)} - n_{2(k)}$ records from F_{1k} to the dummy records. Likewise, when $n_{1(k)} < n_{2(k)}$, we add $n_{2(k)} - n_{1(k)}$ dummy records to F_{1k}. When $n_{1(k)} = n_{2(k)}$, no dummy records are needed. Thus, for a given $\boldsymbol{B}$, the linkages in both BLASE and GM are one-to-one and complete, in the sense that every record from F_1 is matched to an actual or dummy record from F_2, and vice versa. We refer to blocks with complete linkage as balanced.

For any given $\boldsymbol{B}$, we define $C = \{C_1, \ldots, C_K\}$ as the collection of all K permutation vectors. We emphasize that in BLASE C derives from $\boldsymbol{B}$; hence, it can change with different plausible values of $\boldsymbol{B}$. Similar linkage structures are used in other Bayesian record linkage approaches (e.g., Fortini et al., 2001; Larsen and Rubin, 2001; Larsen, 2004; Tancredi and Liseo, 2011; Steorts et al., 2014b; Sadinle, 2016).

Finally, for $i \in F_1$, let $\mathbf{Y}_{1i} = (Y_{1i1}, \ldots, Y_{1ip})$ comprise the p variables of interest for the analysis that are exclusive to F_1. For $i \in F_2$, let $\mathbf{Y}_{2i} = (Y_{2i1}, \ldots, Y_{2iq})$ comprise the q variables of interest for the analysis that are exclusive to F_2. Let $\mathbf{Y}_1 = \{\mathbf{Y}_{1i} : i \in F_1\}$ and $\mathbf{Y}_2 = \{\mathbf{Y}_{2i} : i \in F_2\}$.

2.2.2 Analysis model

In BLASE, we specify a joint distribution, $f(\mathbf{Y}_1, \mathbf{Y}_2 \mid \Theta, C, \boldsymbol{B})$, where Θ represents the parameters of the distribution. The GM does likewise, setting $\hat{\boldsymbol{B}} = \boldsymbol{B}$. For any record $i' \in F_2$ linked to a dummy record, we treat its value of $\mathbf{Y}_{1i'}$ as missing at random (Rubin, 1976); similarly, we treat $\mathbf{Y}_{2i}$ as missing at random for records $i \in F_1$ linked to dummy records in F_2.

In the examples in Sections 2.5 and 2.6, we make Y_1 univariate with a normal linear regression on $(Y_2, \boldsymbol{B})$, with parameters $(\boldsymbol{\beta}, \sigma_1)$. We make Y_2 univariate with a normal linear regression on $\boldsymbol{B}$, with parameters $(\boldsymbol{\eta}, \sigma_2)$. As a prior distribution for $\Theta = (\boldsymbol{\beta}, \sigma_1, \boldsymbol{\eta}, \sigma_2)$, we use $p(\Theta) \propto (1/\sigma_1^2)(1/\sigma_2^2)$.

2.2.3 Model for C

In any block k, by definition the T_1 seeds never change links, whereas the links for the non-T_1 seeds are unknown. Hence, there are $c_{(k)}!$ possible values of C_k, which we denote as C_{kl}, where $l \in \{1, \ldots, c_{(k)}!\}$. We assume each permutation is *a priori* equally likely given a realization of $\boldsymbol{B}$, so that

$$p(C_k = C_{kl} \mid \boldsymbol{B}) = \frac{1}{c_{(k)}!}. \tag{2.1}$$

We model each C_k independently, given $\boldsymbol{B}$. The GM model uses the same distribution, replacing $\boldsymbol{B}$ with $\hat{\boldsymbol{B}}$.

2.2.4 Measurement error model

The GM ends with the models in Section 2.2.3; however, in BLASE we specify a measurement error model for the observed BVs. For each (f, i) we define $E_{fi} = \{E_{fij} : j \in (1, \ldots, J)\}$, where each

$$E_{fij} = \begin{cases} 1 & : \text{ if } \hat{B}_{fij} \neq B_{fij} \\ 0 & : \text{ otherwise.} \end{cases} \tag{2.2}$$

For T_1 and T_2 seeds, $E_{fij} = 0$ for all j. For non-seeds, each E_{fij} is a random variable. We let $\mathbf{E}_f = \{E_{fi} : i \in F_f\}$ where $f = 1, 2$. We use independent beta-Bernoulli models for each E_{fij}. We have

$$E_{fij} | \gamma_{fj} \sim Bernoulli(\gamma_{fj}) \tag{2.3}$$

$$\gamma_{fj} \sim Beta(a_{\gamma_{fj}}, b_{\gamma_{fj}}), \tag{2.4}$$

where $E(\gamma_{fj}) = a_{\gamma_{fj}}/(a_{\gamma_{fj}} + b_{\gamma_{fj}})$ is the expected prior probability of observing a faulty value on non-seed field j in file f. We discuss selection of $(a_{\gamma_{fj}}, b_{\gamma_{fj}})$ in Section 2.5. This effectively assumes that BVs are faulty at random (Kim et al., 2015) at potentially different rates, which is a reasonable default model in the absence of specific information about the mechanisms that generate faulty values. If available,

such information could be encoded in the model for E_{fij}; for example, use a logistic regression on some function of $\boldsymbol{B}$.

For $\hat{\boldsymbol{B}}$, we use a model similar to that used in Tancredi and Liseo (2011), Steorts et al. (2014b), and Manrique-Vallier and Reiter (2016). We have

$$f\left(\hat{B}_{fij}|E_{fij}, B_{fij}\right) = \begin{cases} B_{fij} & : \text{if } E_{fij} = 0 \\ \text{Uniform discrete on all but } B_{fij} & : \text{if } E_{fij} = 1. \end{cases} \tag{2.5}$$

This model implies that reported values for faulty items are generated completely randomly. This also is a default model. If information were available on the mechanisms that introduce faults in the BVs, one could use other reporting error models as appropriate.

The models in (2.3) – (2.5) allow for records in both F_1 and F_2 to be faulty. When sensible, one convenient modification is to allow only one file, say F_2, to have faulty records. To do so, we set $E_{1ij} = 0$ and $\hat{B}_{1ij} = B_{1ij}$ for all i, j. This speeds up the posterior sampling in BLASE, as only F_2 is examined for faulty data. It also helps pin down the estimated distribution of $\boldsymbol{B}$, and enables a simplification in the block moves described in Section 2.3.

2.2.5 Model for $\boldsymbol{B}$

BLASE can use any model that adequately describes the distribution of $\boldsymbol{B}$. Here, we use the highly flexible and computationally efficient latent class model, which has been shown to be a useful tool for modeling and imputation of categorical data (e.g., Vermunt et al., 2008; Si and Reiter, 2013). We assume that B_{fi} follows the same distribution for all unique individuals i in the population defined by $F_1 \cup F_2$. When there is reason to believe that the distribution varies by file, one can specify separate models for $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$.

For each BV variable j, where $j = 1, \ldots, J$, without loss of generality assume its values take on elements in $\{1, \ldots, d_j\}$. As is usual in latent class models, assume each

individual belongs to one of H latent classes. Let $z_i \in \{1, \ldots, H\}$ represent the latent class assignment for record i. We assume that $Pr(z_i = h) = \pi_h$ for all i and h. Let $\pi = (\pi_1, \ldots, \pi_H)$. Within each latent class h, all BVs follow independent multinomial distributions. For $h = 1, \ldots, H$ and $j = 1, \ldots, J$, let $\phi_{hjb} = Pr(B_{ij} = b \mid z_i = h)$, where $b \in \{1, \ldots, d_j\}$. Let $\phi_{hj} = \left(\phi_{hj1}, \ldots, \phi_{hjd_j}\right)$. Mathematically, we have

$$B_{ij}|\phi_{hj}, z_i \sim Multinomial\left(1; \{1, \ldots, d_j\}, \phi_{z_ij}\right) \tag{2.6}$$

$$z_i|\pi \sim Multinomial(\pi_1, \ldots, \pi_H). \tag{2.7}$$

As prior distributions on π and each ϕ_{hj}, we use a truncated Dirichlet process mixture model representation (Dunson and Xing, 2009). We have

$$\pi_h = V_h \prod_{g<h}(1 - V_g), h = 1, \ldots, H \tag{2.8}$$

$$V_h \sim Beta(1, \alpha), \text{for } h = 1, \ldots, H-1;\ V_H = 1 \tag{2.9}$$

$$\alpha \sim Gamma(a_\alpha, b_\alpha) \tag{2.10}$$

$$\left(\phi_{hj1}, \ldots, \phi_{hjd_j}\right) \sim Dirichlet(a_{j1}, \ldots, a_{jd_j}). \tag{2.11}$$

Following Dunson and Xing (2009) and Si and Reiter (2013), we set $a_\alpha = b_\alpha = 0.25$ to represent low prior sample size, and $a_j = (a_{j1}, \ldots, a_{jd_j}) = (1, \ldots, 1)$ for all j. For H, we recommend starting with a large value, such as $H = 30$. When posterior runs indicate that all H classes are consistently occupied, we increase H and restart the sampling. We represent the collection of parameters as $\Psi = (\phi, \pi, \alpha, z)$. For more details about this model, see Appendix A.

2.2.6 Intuition about block moves and roles of seeds

The distributions of $(\hat{\boldsymbol{B}}_f, \mathbf{E}_f, \boldsymbol{B}_f)$ define models for the BVs that incorporate the possibility of faulty blocks. At each iteration s of the MCMC for BLASE, a posterior sample for $\boldsymbol{B}^{(s)}$ is drawn, and records are shifted into the blocks defined by $\boldsymbol{B}^{(s)}$. Thus, unlike the GM, BLASE allows records initially in a separate block from their true match potentially to shift into the block in which their true match lies.

Block moves are influenced by multiple levels of the hierarchical model. BLASE favors shifting records from a current block into one that represents a higher value of the likelihood function for the regression models of interest. Thus, the accuracy of block moves is reliant upon the strength of the relationships between $\boldsymbol{B}$ and Y_1, and between $\boldsymbol{B}$ and Y_2. When $\boldsymbol{B}$ is not important in the regression models, posterior samples for $\boldsymbol{B}^{(s)}$ tend not to move from $\hat{\boldsymbol{B}}$, resulting in few records shifting blocks. Block moves also are reliant on the measurement error model, particularly when (2.4) is highly informative. Finally, values of the BVs with high posterior probability under the latent class model are more likely to be sampled for $\boldsymbol{B}^{(s)}$, whereas unlikely combinations are sampled less often.

The T_1 and T_2 seeds inform multiple parts of the model. Because their true BV values are known, T_1 and T_2 seeds provide information on the joint distribution of the BVs. The T_1 seeds additionally provide essential information about Θ; effectively, they identify regions of support with non-trivial mass. At each posterior update, T_1 seeds are used along with estimated links to update C and the parameters of the linking analysis. In this way, the influence of the T_1 seed information persists throughout posterior estimation. Thus, it is crucial that the T_1 seeds are representative of the underlying relationships between Y_1 and Y_2. When this is not the case, strong T_1 seed information could encourage BLASE to favor choices of C that reflect the relationships among the seeds rather than the true Θ.

2.3 Posterior Sampling Algorithm

To estimate the posterior distribution of the parameters in BLASE, we leverage the MCMC algorithms described in Gutman et al. (2013). The GM uses a two-step Gibbs sampler for $(\Theta \mid C, \hat{\boldsymbol{B}})$ and $(C \mid \Theta, \hat{\boldsymbol{B}})$. BLASE, however, considers $\hat{\boldsymbol{B}}$ to be subject to measurement error and allows the blocking structure to change; hence, additional steps are needed. We present the sampler for the case where only F_2 is

subject to faulty values.

We initialize the MCMC sampler using information from the T_1 and T_2 seeds. Let $\mathbf{T}_1$ be the subset of $(\mathbf{Y}_1, \mathbf{Y}_2, \boldsymbol{B})$ corresponding to the T_1 seeds. We initialize $\Theta^{(0)}$ by taking a random draw from $f(\Theta \mid \mathbf{T}_1)$. We initialize $\boldsymbol{B}^{(0)} = \hat{\boldsymbol{B}}$, and put all records in $F_1 \cup F_2$ into the blocks defined by $\boldsymbol{B}^{(0)}$. We create dummy records as needed to balance all initial blocks. We initialize $C_k^{(0)}$ for each k using the step described in S2.3 below. Let $\boldsymbol{B}_{12}$ contain the rows of $\boldsymbol{B}$ corresponding to the T_1 and T_2 seeds. Conditioning on $\boldsymbol{B}_{12}$, we initialize $\Psi^{(0)}$ using one draw from the block Gibbs sampler (Ishwaran and James, 2001) as follows:

1. For $i \in 1, \ldots, N_B$, update $z_i \in \{1, \ldots, H\}$ from a multinomial full conditional distribution with sample size 1 and

$$Pr(z_i^{(s+1)} = h | \boldsymbol{B}^{(s+1)}, \phi^{(s)}, \pi^{(s)}, C^{(s+1)}) = \frac{\pi_h^{(s)} \prod_{j=1}^{J} \phi^{(s)}_{hjB^{(s+1)}_{i'j}}}{\sum_{h'=1}^{H} \pi_{h'}^{(s)} \prod_{j=1}^{J} \phi^{(s)}_{h'jB^{(s+1)}_{i'j}}}. \quad (2.12)$$

2. For $h \in \{1, \ldots, H-1\}$, update V_h from the full conditional,

$$V_h^{(s+1)} | z^{(s+1)}, \alpha^{(s)} \sim Beta\left(1 + m_h^{(s+1)}, \alpha^{(s)} + \sum_{h'=h+1}^{H} m_h^{(s+1)}\right), \quad (2.13)$$

   where $m_h^{(s+1)} = \sum_i I(z_i^{(s+1)} = h)$ denotes the number of individuals in latent class h. Set $V_H = 1$.

3. Compute $\pi_h^{(s+1)} = V_h^{(s+1)} \prod_{g<h}(1 - V_g^{(s+1)})$ for all $h \in \{1, \ldots, H\}$.

4. For $j \in \{1, \ldots, J\}$ and $h \in \{1, \ldots, H\}$, update ϕ_{hj} from the full conditional,

$$\phi_{hj}^{(s+1)} | z^{(s+1)}, \boldsymbol{B}^{(s+1)} \sim$$

$$Dirichlet\left(a_{j1} + \sum_{i:z_i=h} I(B_{ij}^{(s+1)} = 1), \ldots, a_{jd_j} + \sum_{i:z_i=h} I(B_{ij}^{(s+1)} = d_j)\right). \quad (2.14)$$

5. Update α from the full conditional,

$$\alpha^{(s+1)}|\pi^{(s+1)} \sim Gamma\left(a_\alpha + H - 1, b_\alpha - log(\pi_H^{(s+1)})\right). \tag{2.15}$$

The posterior sampling for BLASE includes the following seven steps to update from iteration (s) to iteration $(s+1)$. We outline the steps here, and additional details are presented in Sections 2.3.1 through 2.4.

S1. For each non-seed record in F_2, propose $(\mathbf{E}_2^*, \boldsymbol{B}_2^*)$ using a Metropolis Hastings (MH) step, resulting in the updated draw $(\mathbf{E}_2^{(s+1)}, \boldsymbol{B}_2^{(s+1)})$. See Section 2.3.2 for details.

S2. Shift records into blocks defined by $\boldsymbol{B}^{(s+1)}$. Add or delete dummy records as needed to ensure all blocks are balanced.

S3. Sample a new $C_k^{(s+1)}$ for each block k; see Section 2.3.1 for details.

S4. Impute missing values of Y_1 or Y_2 for records linked to dummy records, using the appropriate conditional distribution derived from $f(Y_1, Y_2 \mid C^{(s+1)}, \boldsymbol{B}^{(s+1)}, \Theta^{(s)})$. Let $\mathbf{Y}_1^{(s+1)}$ comprise $\mathbf{Y}_1$ and the imputed values of Y_1, and let $\mathbf{Y}_2^{(s+1)}$ comprise $\mathbf{Y}_2$ and the imputed values of Y_2.

S5. Sample $(\Theta^{(s+1)} \mid C^{(s+1)}, \boldsymbol{B}^{(s+1)}, \mathbf{Y}_1^{(s+1)}, \mathbf{Y}_2^{(s+1)})$ from its full conditional.

S6. Sample $\gamma^{(s+1)}$ from

$$\gamma_{2j} \mid \mathbf{E}_2^{(s+1)} \sim Beta\left(a_{\gamma_{2j}} + \sum_i E_{2ij}^{(s+1)}, b_{\gamma_{2j}} + \sum_i (1 - E_{2ij}^{(s+1)})\right). \tag{2.16}$$

S7. Sample $\Psi^{(s+1)}$ using the block Gibbs sampler described above.

2.3.1 Further details on sampling $C^{(s+1)}$

Given $\boldsymbol{B}^{(s+1)}$, to sample $C^{(s+1)}$ we use the strategies in Gutman et al. (2013), which we summarize here due to their importance in the algorithm. Let $\mathbf{Y}_{1(k)}^{(s)}$ and $\mathbf{Y}_{2(k)}^{(s)}$ be the values of $\mathbf{Y}_1^{(s)}$ and $\mathbf{Y}_2^{(s)}$ in block k defined by $\boldsymbol{B}^{(s+1)}$. We sample each C_k independently from

$$p(C_k = C_{k\ell} \mid \mathbf{Y}_1^{(s)}, \mathbf{Y}_2^{(s)}, \boldsymbol{B}^{(s+1)}, \Theta^{(s)}) = \frac{f(\mathbf{Y}_{1(k)}^{(s)}, \mathbf{Y}_{2(k)}^{(s)} \mid \Theta^{(s)}, C_{k\ell}, \boldsymbol{B}^{(s+1)})}{\sum_{l'=1}^{c_{(k)}!} f(\mathbf{Y}_{1(k)}^{(s)}, \mathbf{Y}_{2(k)}^{(s)} \mid \Theta^{(s)}, C_{k\ell'}, \boldsymbol{B}^{(s+1)})}. \tag{2.17}$$

We refer to sampling C_k from (2.17) as the Exact Step. It requires enumerating all possible permutations within block k, and computing the likelihood for the analysis model within the block for each possible permutation.

Evaluating all possible permutations becomes computationally expensive when the number of records within a block is even modestly large. To mitigate this for large blocks, we adapt the MH step in the algorithm for the GM, which itself is adapted from Wu (1995), as an alternative sampling technique for C_k. We refer to this alternative as a Switch Step. For any k in which $c_{(k)}$ is larger than a user specified threshold, we do the following.

1. Propose a new permutation $C_{k\ell^*}$ by selecting with uniform probability two within-block records from F_2, and swapping their positions.

2. Compute the ratio of the likelihoods under the current and proposed $C_{k\ell}$ and $C_{k\ell^*}$,

$$r = \frac{f(\mathbf{Y}_{1(k)}^{(s)}, \mathbf{Y}_{2(k)}^{(s)} \mid \Theta^{(s)}, C_{k\ell}{}^*, \boldsymbol{B}^{(s+1)})}{f(\mathbf{Y}_{1(k)}^{(s)}, \mathbf{Y}_{2(k)}^{(s)} | \Theta^{(s)}, C_{k\ell}, \boldsymbol{B}^{(s+1)})}. \tag{2.18}$$

3. Set $C_k = C_{k\ell}{}^*$ when $log(u) < log(r), u \sim Unif(0,1)$. Otherwise, set $C_k = C_{k\ell}$.

4. Repeat these three steps $R > 1$ times to encourage sampling high probability permutations. The constant R is user specified ($R = 30$ in our applications). The value of C_k in the Rth iteration is taken for $C_k^{(s+1)}$.

This avoids computing the likelihood for every possible C_{kl}, as only two likelihoods need to be computed. Following GM, we sample C_k for all blocks k with $c_{(k)} \geqslant 5$ using the Switch Step and for all smaller blocks using the Exact Step.

The collection $\{C_k^{(s+1)} : k = 1, \ldots, K\}$ defines the sampled value for $C^{(s+1)}$. We link F_1 and F_2 according to this permutation. We note that for BLASE, since $\boldsymbol{B}^{(s+1)}$ potentially can change at each MCMC iteration, we must introduce dummy variables to make balanced blocks prior to updating $C^{(s)}$.

2.3.2 Further details on sampling $(\mathbf{E}_2^{(s+1)}, \boldsymbol{B}_2^{(s+1)})$

The most complicated step of BLASE involves proposing values for $\mathbf{E}_2$ and $\boldsymbol{B}_2$. The basic strategy is as follows. For each record $i \in F_2$ that is not a T_1 or T_2 seed, we propose a new value of $((E_{2i})^*, (B_{2i})^*)$. We accept or reject that proposal using a MH step, with an acceptance ratio defined in Section 2.4. We repeat these two steps independently for all non-seed records in F_2, resulting in $(\mathbf{E}_2^{(s+1)}, \boldsymbol{B}_2^{(s+1)})$. Using independent, record by record updates allows for the updates to be done in parallel using multiple processors (although we do not utilize this feature in the examples).

We now describe the updating process for any record $i \in F_2$ that is not a T_1 or T_2 seed. For any eligible record i, we randomly select one non-seed field j^* uniformly from all non-seed fields for that record. Thus, at any iteration in the MCMC, $E_{2i}^{(s)}$ and $E_{2i}^{(s+1)}$ differ on at most one field, and record i may move only to some block k^* that disagrees with its current block k on exactly one field. This simplifies the computation of the MH acceptance ratio.

We propose a value for E_{2ij^*} from

$$(E_{2ij^*})^* | \gamma_2^{(s)}, j^* \sim Bernoulli(\gamma_{2j^*}^{(s)}). \tag{2.19}$$

When $(E_{2ij^*})^* = E_{2ij^*}^{(s)} = 0$, we set $E_{2i}^{(s+1)} = E_{2i}^{(s)}$ and $B_{2i}^{(s+1)} = B_{2i}^{(s)}$. Otherwise, we set $(E_{2ij})^* = E_{2ij}^{(s)}$ for all $j \neq j^*$, and propose a new $(B_{2ij^*})^*$ via one of three general patterns.

$$\text{If } (E_{2ij^*})^* = 0 \text{ and } E_{2ij^*}^{(s)} = 1 \Rightarrow \text{Propose } (B_{2ij^*})^* = \hat{B}_{2ij^*}. \tag{2.20a}$$

$$\text{If } (E_{2ij^*})^* = 1 \text{ and } E_{2ij^*}^{(s)} = 0 \Rightarrow \text{Propose } (B_{2ij^*})^* \neq \hat{B}_{2ij^*}. \tag{2.20b}$$

$$\text{If } (E_{2ij^*})^* = 1 \text{ and } E_{2ij^*}^{(s)} = 1 \Rightarrow \text{Propose } (B_{2ij^*})^* \notin \{\hat{B}_{2ij^*} \cup B_{2ij^*}^{(s)}\}. \tag{2.20c}$$

To aid understanding, call the block defined by $\hat{B}_{2i}$ the reported block for record i, and any block that is different from the reported block an alternate block. Hence, (2.20a) shifts i from an alternate block to a reported block, (2.20b) shifts from a reported block to an alternate block, and (2.20c) shifts i from one alternate block to another.

In the cases of (2.20b) and (2.20c), we propose $(B_{2ij^*})^*$ by drawing from the mixture model described in Section 2.2.5. However, we restrict $(B_{2ij^*})^*$ to ensure the conditions in (2.20b) and (2.20c) are not violated. We also disallow $(B_{2ij^*})^*$ that result in BV combinations not observed in $\hat{\boldsymbol{B}}_1$. In our experience, if we allow BLASE to create a block with only one observed record from F_2, that record typically stays in this newly created block over the MCMC runs, regardless of whether it should be there or not. This is because BLASE always has to base the MH ratio on an imputed value of Y_1 for that record, which by design is generated to be concordant with the analysis model, i.e., it has relatively high likelihood. As a result, the MH step tends not to move the record out of its (randomly sampled) unique block, which runs counter to our objective of exploring the space of possible block moves.

We define the set of *legal blocks* for each record i to contain blocks defined by $\hat{\boldsymbol{B}}_{\mathbf{1}}$ and blocks that fulfill the restrictions in (2.20a)-(2.20c). The move restriction requires adapting the proposal step for $\boldsymbol{B}^*$ to only allow proposals which lead records to move to legal blocks. In the proposal, we normalize each ϕ_{hj*} over the set of legal blocks. Specifically, define

$$D_{(b)} = \{x \in (1 : d_{j*}) | x \neq \hat{B}_{2ij*} \text{ and } B_{2ij*} = x \text{ implies } B_{2i} \in \hat{\boldsymbol{B}}_1\}. \tag{2.21}$$

$D_{(b)}$ contains the set of possible proposals for B_{2ij*} under (2.20b). Likewise, define

$$D_{(c)} = \{x \in (1 : d_{j*}) | x \neq \hat{B}_{2ij*}, x \neq {B_{2ij*}}^{(s)} \text{ and } B_{2ij*} = x \text{ implies } B_{2i} \in \hat{\boldsymbol{B}}_1\}. \tag{2.22}$$

$D_{(c)}$ then contains the set of possible proposals for B_{2ij*} under (2.20c). Define

$$\Phi_{ij*m}{}^{(s)} = \begin{cases} \phi_{hj*m}{}^{(s)} / \sum_{x \in D_{(b)}} \phi_{hj*x}{}^{(s)} & : m \in D_{(b)}, (E_{2ij*})^* = 1 \text{ and } E^{(s)}_{2ij*} = 0 \\ \phi_{hj*m}{}^{(s)} / \sum_{x \in D_{(c)}} \phi_{hj*x}{}^{(s)} & : m \in D_{(c)}, (E_{2ij*})^* = 1 \text{ and } E^{(s)}_{2ij*} = 1 \\ 0 & : \text{otherwise.} \end{cases} \tag{2.23}$$

Let $\Phi^{(s)}_{ij*} = \{\Phi^{(s)}_{ij*m} | m \in 1, \ldots, d^*_j\}$. We then obtain the proposal for B_{2ij*} as follows. Under (2.20a), propose $(B_{2ij*})^* = \hat{B}_{2ij*}$. Under (2.20b) or (2.20c), propose $(B_{2ij*})^*$ by sampling from $Multinomial(1; \{1, \ldots, d^*_j\}, \Phi^{(s)}_{ij*})$.

Based on the sampled value of $(B_{2ij*})^*$, we shift record i from the block k defined by ${B_{2i}}^{(s)}$ into the block k^* defined by $(B_{2i})^*$. Block k loses a record from F_2 while block k^* gains one. We then add or remove dummy records to rebalance both affected blocks. When block k^* contains at least one record from F_{1k^*} matched to a dummy record, we temporarily replace one of the dummy records with the record i shifted into the block. When block k^* does not contain a dummy record in F_{2k}, we impute the missing Y_1 for record i from $f(Y_1 | Y_2^{(s)}, C^{(s)}, \boldsymbol{B}^*, \Theta^{(s)})$. For block k, if record i was matched to a dummy record when in F_{2k}, we temporarily remove the dummy record from block k. Otherwise, we impute the missing Y_2 value for record i from

$f(Y_2|Y_1^{(s)}, C^{(s)}, \boldsymbol{B}^*, \Theta^{(s)})$. Let Y_1^* comprise Y_1 and the imputed values of Y_1 after the block move, and let Y_2^* comprise Y_2 and the imputed values of Y_2 after the block move. Similarly, let $(Y_{1(k)})^*$ and $(Y_{2(k)})^*$ comprise the values of Y_1^* or Y_2^* in block k after moving record i. We define $(Y_{1(k^*)})^*$ and $(Y_{2(k^*)})^*$ analogously.

Shifting records in and out of blocks requires two new permutations, $(C_k)^*$ and $(C_{k*})^*$. For small blocks, we propose these permutations using the Exact Step. For large blocks, we use a variation of the Switch Step, which we describe in Section 2.3.2.1 through 2.3.2.4. We accept or reject $(E_{2ij^*}^*, B_{2ij^*}^*)$ based on the MH acceptance ratio described in Section 2.4, the result of which we use for $(E_{2ij^*}^{(s+1)}, B_{2ij^*}^{(s+1)})$. When updating each non-seed's BVs in F_2 in serial, we return all values to those in $(E^{(s)}, B^{(s)}, C^{(s)}, \mathbf{Y}_1^{(s)}, \mathbf{Y}_2^{(s)})$ before moving to the proposal for the next record.

We now describe in detail the process of proposing $(C_k)^*$ and $(C_{k*})^*$. These permutations define the linkage structure within the proposed blocking structure defined by $(\boldsymbol{B}_2)^*$.

2.3.2.1 Block k: Case 1

Conditional on $C_k^{(s)}$, suppose record i is matched to a dummy record $*$ in block k. If i is shifted from the block, the dummy record is no longer necessary, and may be removed from k. Because the permutation $C_k^{(s)}$ defines the linkage structure for all remaining records in the block, we propose $(C_k)^* = C_k^{(s)}$, with the record i removed.

2.3.2.2 Block k: Case 2

Conditional on $C^{(s)}$, suppose record i is not matched to a dummy record in block k. We impute the missing Y_2 value for record i from $f(Y_2|Y_1^{(s)}, C^{(s)}, \boldsymbol{B}^*, \Theta^{(s)})$. $(Y_{2(k)})^*$ then comprises this imputed value and the values of $Y_{2(k)}^{(s)}$ which remain in block k after record i is removed. Y_1 is unaffected by this block move, so $(Y_{1(k)})^* = Y_{1(k)}^{(s)}$. If

block k is small, we propose $(C_k)^*$ using the Exact Step. In this case,

$$Pr((C_k)^* = C_{kl} \mid C_k^{(s)}, Y_1^*, Y_2^*, \boldsymbol{B}^*, \Theta^{(s)}) = \frac{f((Y_{1(k)})^*, (Y_{2(k)})^* \mid \Theta^{(s)}, C_{k\ell}, \boldsymbol{B}^*)}{\sum_{l'=1}^{c_k^*!} f((Y_{1(k)})^*, (Y_{2(k)})^* \mid \Theta^{(s)}, C_{k\ell'}, \boldsymbol{B}^*)} \tag{2.24}$$

If k is large, we propose $(C_k)^*$ by selecting two non-seed records and switching their positions. Let $\mathcal{C}_{(k)}$ denote the set of all permutations which can be obtained by such a process. Let c_k^* denote the value of c_k after the record move. Then for a large block, we have

$$Pr((C_k)^* = C_{kl} \mid C_k^{(s)}, Y_1^*, Y_2^*, \boldsymbol{B}^*, \Theta^{(s)}) = \begin{cases} \frac{(c_k^*-2)!2!}{c_k^*!} & : \text{ if } C_{k\ell} \in \mathcal{C}_k \\ 0 & : \text{ otherwise .} \end{cases} \tag{2.25}$$

This avoids computing the likelihood for every possible C_{kl}. To allow exploration of $\mathcal{C}_k$, for large blocks the process of accepting or rejecting (E_{2i}^*, B_{2i}^*) is as follows:

1. Propose a new permutation from (2.25).
2. Compute the acceptance ratio A as defined in (2.43)
3. If $log(u) < log(A), u \sim Unif(0,1)$, we accept (E_{2i}^*, B_{2i}^*). Otherwise, repeat the first two steps $R > 1$ times. The constant R is user specified ($R = 30$ in our applications). If after R iterations we fail to accept, we reject the block move and set $(E_{2i}^{(s+1)}, B_{2i}^{(s+1)}) = (E_{2i}^{(s)}, B_{2i}^{(s)})$.

2.3.2.3 Block k^*: Case 1

Conditional on $C_{k^*}^{(s)}$, suppose at least one record in $F_{1(k^*)}^{(s)}$ in block k^* is matched to a dummy record, denoted $*$. Conditional on $(B_{2i})^*$, we move record i into block k^* and replace the dummy record with i. If there is more than one F_2 dummy record in the block, we randomly select one to replace with i. Let $r^{'}$ denote this selected dummy

record. Then $(Y_{2(k^*)r'})^* = Y_{2i}$ and $(Y_{2(k^*)r})^* = Y^{(s)}_{2(k)r}$ for all $r = \{1, \ldots, c^{(s)}_{(k^*)} | r \neq r'\}$. $Y^{(s)}_1$ is unaffected by this block move, so $(Y_{1(k^*)})^* = Y^{(s)}_{1(k^*)}$. There are no dummy records created for this case. We propose $(C_{k^*})^*$ analogously to Section 2.3.2.2.

2.3.2.4 Block k: Case 2*

Conditional on $C^{(s)}_{k^*}$, suppose no records in $F^{(s)}_{1(k^*)}$ in block k^* are matched to a dummy record. Conditional on $(B_{2i})^*$, we move record i into block k^* and impute the missing Y_1 for record i from $f(Y_1|Y_2^*, C^{(s)}, \boldsymbol{B}^*, \Theta^{(s)})$. Here $(Y_{2(k^*)})^* = ((Y_{2(k^*)})^{(s)}, Y_{2i})$ and $(Y_{1(k^*)})^*$ comprises $Y^{(s)}_{1(k^*)}$ and the imputed Y_1 value. We propose $(C_{k^*})^*$ analogously to Section 2.3.2.2.

2.4 Sampling $(\mathbf{E}_2^*, \boldsymbol{B}_2^*)$: Deriving the Acceptance Ratio

Once values for $(E_{2i})^*$ and $(B_{2i})^*$ have been proposed, we simultaneously accept or reject both proposals. In this section, we derive the acceptance ratio A used in this MH step.

2.4.1 Notation

Let $\boldsymbol{B}^* = (\boldsymbol{B}_1, \boldsymbol{B}_2^*)$, where the i^{th} row of $\boldsymbol{B}_2^*$ is defined by the proposal $(B_{2i})^*$, and $(B_{2i'})^* = (B_{2i'})^{(s)}$ for all $i' \neq i$. Similarly, let $\mathbf{E}^* = (\mathbf{E}_1, \mathbf{E}_2^*)$, where the i^{th} row of $\mathbf{E}_2^*$ is defined by the proposal $(E_{2i})^*$, and $(E_{2i'})^* = (E_{2i'})^{(s)}$ for all $i' \neq i$.

For each proposed block move, the acceptance ratio A takes a standard MH form, namely $A = p(\cdot) \times J(\cdot)$. $J(\cdot)$, the ratio of transition probabilities from the proposal distribution, is

$$J(\cdot) = \frac{p(E^{(s)}_{2ij^*}, B^{(s)}_{fij^*}, C^{(s)} | \mathbf{E}^*, (C)^*, \boldsymbol{B}^*, Y_1^*, Y_2^*, \Psi^{(s)}, \hat{\boldsymbol{B}}, \Theta^{(s)}, \gamma^{(s)})}{p((E_{2ij^*})^*, (B_{2ij^*})^*, (C)^* | \mathbf{E}^{(s)}, C^{(s)}, \boldsymbol{B}^{(s)}, Y_1^{(s)}, Y_2^{(s)}, \Psi^{(s)}, \hat{\boldsymbol{B}}, \Theta^{(s)}, \gamma^{(s)})}. \tag{2.26}$$

The ratio of the probability of the new state to the current state is denoted $p(\cdot)$, with

$$p(\cdot) = A_1 \times A_2 \times A_3, \tag{2.27}$$

where

$$\begin{aligned} A_1 =& \frac{f(\hat{B}_{2ij*}|\mathbf{E}_2^*, \boldsymbol{B}_2^*)}{f(\hat{B}_{2ij*}|\mathbf{E}_2^{(s)}, \boldsymbol{B}_2^{(s)})} \times \frac{p(B_{2ij*} = (B_{2ij*})^*|\Psi^{(s)})}{p(B_{2ij*} = B_{2ij*}^{(s)}|\Psi^{(s)})} \times \\ & \frac{p(E_{2ij*} = (E_{2ij*})^*|\gamma_{2j*}^{(s)})}{p(E_{2ij*} = E_{2ij*}^{(s)}|\gamma_{2j*}^{(s)})} \times \frac{p(C = C^*|\boldsymbol{B}_2^*)}{p(C = C^{(s)}|\boldsymbol{B}_2^{(s)})}, \end{aligned} \tag{2.28}$$

$$A_2 = \frac{f(Y_1, Y_2|\Theta^{(s)}, C^*, \boldsymbol{B}^*)}{f(Y_1, Y_2|\Theta^{(s)}, C^{(s)}, \boldsymbol{B}^{(s)})}, \tag{2.29}$$

$$A_3 = \frac{p(\Psi^{(s)})p(\gamma^{(s)})p(\Theta^{(s)})}{p(\Psi^{(s)})p(\gamma^{(s)})p(\Theta^{(s)})}. \tag{2.30}$$

2.4.2 Component $p(\cdot)$

In (2.30), the priors for γ and Ψ are independent of $\mathbf{E}_2, \boldsymbol{B}_2$ or C, and therefore do not contribute to the acceptance ratio. The prior $p(\Theta)$ is user-specified. For example, if the linking analysis were a regression model, a g prior could be used for the coefficients. This choice for $p(\Theta)$ is conditional on the permutation C through reliance on the design matrix. In the simulation studies in Section 2.5, $\Theta = (\boldsymbol{\beta}, \sigma_1, \boldsymbol{\eta}, \sigma_2)$ and the prior $p(\Theta) \propto (1/\sigma_1^2)(1/\sigma_2^2)$ is independent of $\mathbf{E}_2, \boldsymbol{B}_2$ or C. Thus, (2.30) equals 1.

The ratio $p(\hat{B}_{2ij*}|\mathbf{E}_2^*, \boldsymbol{B}_2^*)/p(\hat{B}_{2ij*}|\mathbf{E}_2^{(s)}, \boldsymbol{B}_2^{(s)})$ in (2.28) is defined by the measurement error model in (2.5), i.e.,

$$f\left(\hat{B}_{2ij}|E_{2ij}, B_{2ij}\right) = \begin{cases} B_{2ij} & : \text{if } E_{2ij} = 0 \\ \text{Uniform discrete on all but } B_{2ij} & : \text{if } E_{2ij} = 1. \end{cases}$$

If $E_{2ij*} = 0$, $\hat{B}_{2ij*} = B_{2ij*}$ with probability 1. If $E_{2ij*} = 1$, (2.5) assigns equal mass

for all d_{j^*} choices of $\hat{B}_{fij}$ excluding B_{fij}. The ratio is then

$$\frac{f(\hat{B}_{2ij^*}|\mathbf{E}_2^*, \boldsymbol{B}_2^*)}{f(\hat{B}_{2ij^*}|\mathbf{E}_2^{(s)}, \boldsymbol{B}_2^{(s)})} = \begin{cases} 1 & : (E_{2ij^*})^* = E_{2ij^*}^{(s)} = 1 \\ 1/(d_{j^*}-1) & : E_{2ij^*}^{(s)} = 0, (E_{2ij^*})^* = 1 \\ d_{j^*}-1 & : E_{2ij^*}^{(s)} = 1, (E_{2ij^*})^* = 0. \end{cases} \tag{2.31}$$

The ratio A is not computed when $(E_{2ij^*})^* = E_{2ij^*}^{(s)} = 0$. With such a proposal, we set $B_{2ij^*}^{(s+1)} = \hat{B}_{2ij^*}$ and the MH step is complete.

The contribution of the prior on B to (2.28) is the ratio of the probability of $p(B_{2ij^*} = (B_{2ij^*})^*|\Psi^{(s)})$ to $p(B_{2ij^*} = B_{2ij^*}^{(s)}|\Psi^{(s)})$. Let the latent class assignment $z_i^{(s)} = h$. Then we have

$$\frac{p(B_{2ij^*} = (B_{2ij^*})^*|\Psi^{(s)})}{p(B_{2ij^*} = B_{2ij^*}^{(s)}|\Psi^{(s)})} = \frac{\phi^{(s)}_{hj^*(B_{2ij^*})^*}}{\phi^{(s)}_{hj^*B_{2ij^*}^{(s)}}}. \tag{2.32}$$

This ratio increases the MH acceptance ratio A if we propose shifting i to a value $(B_{2ij^*})^*$ which is more likely under $\Psi^{(s)}$ than its current $B_{2ij^*}^{(s)}$. The contribution of the model for $\mathbf{E}_2$ to (2.28) is similarly concise. For all $j \neq j^*$, $(E_{2ij})^* = E_{2ij}^{(s)}$. Hence, we have:

$$\frac{p(E_{2ij^*} = (E_{2ij^*})^*|\gamma_{2j^*}^{(s)})}{p(E_{2ij^*} = E_{2ij^*}^{(s)}|\gamma_{2j^*}^{(s)})} = \begin{cases} 1 & : (E_{2fij^*})^* = E_{2ij^*}^{(s)} = 1 \\ \frac{\gamma_{2j^*}}{1-\gamma_{2j^*}} & : E_{2ij^*}^{(s)} = 0, (E_{2ij^*})^* = 1 \\ \frac{1-\gamma_{2j^*}}{\gamma_{2j^*}} & : E_{2ij^*}^{(s)} = 1, (E_{2ij^*})^* = 0. \end{cases} \tag{2.33}$$

The final component of (2.28) is the ratio $p(C^*|\boldsymbol{B}^*)/p(C^{(s)}|\boldsymbol{B}^{(s)})$. Within each MH step, only record i moves blocks. Specially, record i moves from a block k to a block k^*. This means that the only difference between C^* and $C^{(s)}$ are the permutations C_k and C_{k^*}. Following our convention, let $c_{(k)}^{(s)}$ denote the length of permutation C_k before the record i move, and let $(c_{(k)})^*$ denote the length of permutation C_k after the proposed record i move. We then have

$$\frac{p(C = C^*|\boldsymbol{B}^*)}{p(C = C^{(s)}|\boldsymbol{B}^{(s)})} = \frac{\frac{1}{(c_{(k)})^*!}\frac{1}{(c_{(k^*)})^*!}}{\frac{1}{c_{(k)}^{(s)}!}\frac{1}{c_{(k^*)}^{(s)}!}} = \frac{c_{(k)}^{(s)}!c_{(k^*)}^{(s)}!}{(c_{(k)})^*!(c_{(k^*)})^*!}, \tag{2.34}$$

where (2.34) is a ratio permutation lengths. The difference $c_{(k)}^{(s)} - (c_{(k)})^* \in \{0, 1\}$, so

$$\frac{c_{(k)}^{(s)}!}{(c_{(k)})^*!} \in \{1, c_{(k)}^{(s)}\}. \tag{2.35}$$

Similarly $(c_{(k^*)})^* - c_{(k^*)}^{(s)} \in \{0, 1\}$, so

$$\frac{c_{(k^*)}^{(s)}!}{(c_{(k^*)})^*!} \in \{1, \frac{1}{(c_{(k^*)})^*}\}. \tag{2.36}$$

Combining (2.31) - (2.34), $p(\cdot)$ can be written as:

$$\begin{aligned} p(\cdot) = \frac{f(Y_1, Y_2|\Theta^{(s)}, C^*, \boldsymbol{B}^*)}{f(Y_1, Y_2|\Theta^{(s)}, C^{(s)}, \boldsymbol{B}^{(s)})} \times \frac{c_{(k)}^{(s)}! c_{(k^*)}^{(s)}!}{(c_{(k)})^*!((c_{(k^*)})^*!} \times \frac{\phi_{hj^*(B_{2ij^*})^*}^{(s)}}{\phi_{hj^*(B_{2ij^*})^{(s)}}^{(s)}} \times \\ \begin{cases} 1 & : E_{2ij^*}^{(s)} = (E_{2ij^*})^* = 1 \\ \gamma_{2j^*} / \left((d_{j^*} - 1)(1 - \gamma_{2j^*})\right) & : E_{2ij^*}^{(s)} = 0, (E_{2ij^*})^* = 1 \\ (d_{j^*} - 1)(1 - \gamma_{2j^*})/\gamma_{2j^*} & : E_{2ij^*}^{(s)} = 1, (E_{2ij^*})^* = 0. \end{cases} \end{aligned} \tag{2.37}$$

2.4.3 Component $J(\cdot)$

The transition probability $J(\cdot)$ in (2.26) may be expressed as

$$\begin{aligned} J(\cdot) &= \frac{Pr(E_{2ij^*} = E_{2ij^*}^{(s)}|\dots) Pr(B_{2ij^*} = B_{2ij^*}^{(s)}|E_{2ij^*}^{(s)}, \dots) Pr(C = C^{(s)}|\boldsymbol{B}^{(s)}, \dots)}{Pr(E_{2ij^*} = (E_{2ij^*})^*|\dots) Pr(B_{2ij^*} = (B_{2ij^*})^*|(E_{2ij^*})^*, \dots) Pr(C = C^*|\boldsymbol{B}^*, \dots)} \\ &= \frac{Pr(E_{2ij^*} = E_{2ij^*}^{(s)}|\gamma_{2j^*}^{(s)})}{Pr(E_{2ij^*} = (E_{2ij^*})^*|\gamma_{2j^*}^{(s)})} \times \frac{Pr(B_{2ij^*} = B_{2ij^*}^{(s)}|\mathbf{E}^{(s)}, \mathbf{E}^*, \Psi^{(s)}, \hat{\boldsymbol{B}}, \boldsymbol{B}^*)}{Pr(B_{2ij^*} = (B_{2ij^*})^*|\mathbf{E}^*, \mathbf{E}^{(s)}, \Psi^{(s)}, \hat{\boldsymbol{B}}, \boldsymbol{B}^{(s)})} \\ &\times \frac{Pr(C = C^{(s)}|C^*, \boldsymbol{B}^{(s)}, Y_1^{(s)}, Y_2^{(s)}, \Theta^{(s)})}{Pr(C = C^*|C^{(s)}, \boldsymbol{B}^*, Y_1^*, Y_2^*, \Theta^{(s)})} \end{aligned} \tag{2.38}$$

2.4.3.1 $(E_{2ij*})^*$

We propose $(E_{2ij*})^*|\gamma^{(s)}$ by sampling $(E_{2ij*})^*|\gamma_2^{(s)}, j^* \sim Bernoulli(\gamma_{2j^*}{}^{(s)})$. Hence, we have

$$\frac{Pr(E_{2ij*} = E_{2ij*}^{(s)}|\gamma_{2j*}^{(s)})}{Pr(E_{2ij*} = (E_{2ij*})^*|\gamma_{2j*}^{(s)})} = \begin{cases} 1 & : (E_{2ij*})^* = 1 = E_{2ij*}^{(s)} \\ (1-\gamma_{2j*})/\gamma_{2j*} & : E_{2ij*}^{(s)} = 0, (E_{2ij*})^* = 1 \\ \gamma_{2j*}/(1-\gamma_{2j*}) & : E_{2ij*}^{(s)} = 1, (E_{2ij*})^* = 0. \end{cases} \tag{2.39}$$

When multiplying (2.39) by (2.37), all terms involving γ cancel out of A.

2.4.3.2 $(B_{2ij*})^*$

The proposal distributions for $(B_{2ij*})^*$ are described in Section 2.3.2. We denote the probability of transitioning from B_{2ij*}^* to $(B_{2ij*})^{(s)}$ as $\Phi^{(*\to s)}_{ij*B_{fij}^{(s)}}$. This has the same form as (2.23) but with the roles of $(B_{2ij*})^*$ and $B_{2ij*}^{(s)}$ reversed. The ratio of proposal distributions for B_{2ij*} is then

$$\frac{Pr(B_{2ij*} = B_{2ij*}^{(s)}|\mathbf{E}^{(s)}, (\mathbf{E})^*, \Psi^{(s)}, \hat{\boldsymbol{B}}, \boldsymbol{B}^*)}{Pr(B_{2ij*} = (B_{2ij*})^*|(\mathbf{E})^*, \mathbf{E}^{(s)}, \Psi^{(s)}, \hat{\boldsymbol{B}}, \boldsymbol{B}^{(s)})} =$$

$$\begin{cases} \Phi^{(*\to s)}_{ij*B_{fij}^{(s)}}/\Phi^{(s)}_{ij*(B_{fij*})^*} & : (E_{fij*})^* = 1 = E_{fij*}^{(s)} \\ 1/\Phi^{(*\to s)}_{ij*B_{fij}^{(s)}} & : (E_{fij*})^* = 1, E_{fij*}^{(s)} = 0 \\ \Phi^{(s)}_{ij*(B_{fij*})^*} & : (E_{fij*})^* = 0, E_{fij*}^{(s)} = 1. \end{cases} \tag{2.40}$$

2.4.3.3 C^*

The final contribution to $J(\cdot)$ in (2.38) is the ratio

$$J_C = \frac{Pr(C = C^{(s)}|C^*, \boldsymbol{B}^{(s)}, Y_1^{(s)}, Y_2^{(s)}, \Theta^{(s)})}{Pr(C = C^*|C^{(s)}, \boldsymbol{B}^*, Y_1^*, Y_2^*, \Theta^{(s)})}. \tag{2.41}$$

Within each MH step, only one record moves from a block k to a block k^*. This means that the only differences between $C^{(s)}$ and C^* are the permutations in blocks k and k^*. Following our convention, let $C_k^{(s)}$ denote the length $c_{(k)}^{(s)}$ permutation C_k

before the record i move, and let $(C_k)^*$ denote the length $(c_k)^*$ permutation C_k after the proposed record i move. Then (2.41) is written as

$$\begin{aligned} J_C = & \frac{Pr(C_k = C_k^{(s)} | (C_k)^*, \boldsymbol{B}^{(s)}, Y_1^{(s)}, Y_2^{(s)}, \Theta^{(s)})}{Pr(C_k = (C_k)^* | C_k^{(s)}, \boldsymbol{B}^*, Y_1^*, Y_2^*, \Theta^{(s)})} \times \\ & \frac{Pr(C_{k^*} = C_{k^*}^{(s)} | (C_{k^*})^*, \boldsymbol{B}^{(s)}, Y_1^{(s)}, Y_2^{(s)}, \Theta^{(s)})}{Pr(C_{k^*} = (C_{k^*})^* | C_{k^*}^{(s)}, \boldsymbol{B}^*, Y_1^*, Y_2^*, \Theta^{(s)})}. \end{aligned} \tag{2.42}$$

Each of the four pieces of (2.42) represents the probability of proposing a given permutation conditional upon either the current or proposed state of the chain. These proposal distributions are described in Section 2.3.2. J_C will be a product of the probabilities described in Sections 2.3.2.1 - 2.3.2.4.

2.4.4 Final ratio

The final acceptance ratio A has one of three forms, dependent on E^* and $E^{(s)}$.

$$\begin{aligned} A &= p(\cdot) \times J(\cdot) \\ &= \frac{f(Y_1, Y_2 | \Theta^{(s)}, C^*, \boldsymbol{B}^*)}{f(Y_1, Y_2 | \Theta^{(s)}, C^{(s)}, \boldsymbol{B}^{(s)})} \times \frac{c_{(k)}^{(s)}! c_{(k^*)}^{(s)}!}{(c_{(k)})^*! ((c_{(k^*)})^*!} \times \frac{\phi_{hj^*(B_{2ij^*})^*}^{(s)}}{\phi_{hj^*B_{2ij^*}^{(s)}}^{(s)}} \times \\ &\times J_C \times \begin{cases} \Phi_{ij^*B_{fij}^{(s)}}^{(*\to s)} / \Phi_{ij^*(B_{fij^*})^*}^{(s)} & : (E_{fij^*})^* = 1 = E_{fij^*}^{(s)} \\ 1/\Phi_{ij^*B_{fij}^{(s)}}^{(*\to s)} & : (E_{fij^*})^* = 1, E_{fij^*}^{(s)} = 0 \\ \Phi_{ij^*(B_{fij^*})^*}^{(s)} & : (E_{fij^*})^* = 0, E_{fij^*}^{(s)} = 1. \end{cases} \end{aligned} \tag{2.43}$$

Table 2.1: Simulation study specifications and prior settings.

Sim Group	Prior Setting	Fault Level	T_1 Seed Count	$a_{\gamma_{2\text{prog}}}$	$b_{\gamma_{2\text{prog}}}$	Prior Mean	Non-Seed Fault %
	CA	40%	60%	90000	10000	90%	100%
HSHF	D	40%	60%	2	10	16.7%	
	CP	40%	60%	12500	87500	12.5%	
	CA	40%	20%	50000	50000	50%	50%
LSHF	D	40%	20%	2	10	16.7%	
	CP	40%	20%	6250	93750	6.25%	
	CA	5%	60%	12500	87500	12.5%	12.5%
HSLF	D	5%	60%	2	10	16.7%	
	CP	5%	60%	90000	10000	90%	
	CA	5%	20%	6250	93750	6.25%	6.25%
LSLF	D	5%	20%	2	10	16.7%	
	CP	5%	20%	50000	50000	50%	

Notes: (1) The diffuse informative (W) prior setting is consistent across all seed levels; (2) Fault Level: the overall percentage of records from F_2 initialized in the incorrect block; (3) T_1 Seed Count: the percentage of the 5000 pairs that are T_1 seeds; (4) $a_{\gamma_{2\text{prog}}}$, $b_{\gamma_{2\text{prog}}}$: the hyper-parameter values for $\gamma_{2\text{prog}}$; (5) Prior Mean: prior mean of $\gamma_{2\text{prog}}$ with the specified hyper-parameters; (6) Non-Seed Fault %: percentage of non-seeds from F_2 initialized in incorrect blocks.

2.5 Simulation studies

In this section, we present results of simulation studies designed to illustrate the performance of BLASE with varying amounts of faultiness, numbers of T_1 seeds, and hyper-parameter settings for the prior distribution of γ_2. We use two faultiness levels, high (40%) and low (5%), and two T_1 seed counts, high (60%) and low (20%), to define four primary simulation groups. For each simulation, we examine three prior distributions for γ_2, representing diffuse (D) prior beliefs, concentrated and appropriate (CA) prior beliefs, and concentrated but poorly specified (CP) beliefs. Table 2.1 presents the simulation specifications in detail.

Table 2.2: Variable descriptions for the hsbdemo data set

Name	Type	Description	Levels	In File
read	Cont	Reading test score		F_1 only
math	Cont	Math test score		F_2 only
female	Binary		0 = Male	Both
			1 = Female	
schtyp	Binary	Type of school	0 = Public	Both
			1 = Private	
ses	Cat	Level of school	1 = Lower	Both
			2 = Middle	
			3 = High	
prog	Cat	Program of study	1 = General	Both
			2 = Academic	
			3 = Vocational	
honors	Binary	Honors status	0 = Not in Honors	Both
			1 = In Honors	
cid	Cat	Unspecified	1-30	Both

Note: All six categorical variables are used as BVs.

2.5.1 Simulation set-up

We use variables from the hsbdemo dataset (UCLA: Stat Consulting Group, 2016) to create simulated files F_1 and F_2. This dataset focuses on school testing, as does the NC education data we use in Section 2.6. Table 2.2 describes the variables. Because hsbdemo has only 200 records, we simulate 5000 record pairs by sampling from predictive distributions estimated on the observed data. In each replication, we refer to generated data with no BV faults as the perfectly blocked (PB) data, meaning $\hat{\boldsymbol{B}} = \boldsymbol{B}$. To ensure small block sizes, we generate each replicated data set ensuring that $c_{(k)} \leqslant 10$ for all k. Specifically, the PB data sets are each generated using the following steps.

1. cid: Sample n values for cid from a Multinomial(1,π), where the probabilities π give equal probability to every integer in $\{1, \ldots, 30\}$.

2. female: Independent of cid, sample n values for female from a Binomial(n,0.545).

The success probability is the proportion of females in the hsbdemo data set.

$$\text{female} \sim Binomial(n, 0.545).$$

3. schtyp|cid: Using the hsbdemo data, fit a logistic model regressing schtyp on cid. Based on this model, estimate the probability for public and private school status for each of the cid values generated in Step 1, and sample n values for schtyp based on these probabilities.

4. prog|schtyp : Using the hsbdemo data, fit a multinomial regression model for prog on schtyp. Based on this model, estimate the probability for program status for each of the n schtyp values generated in Step 2, and sample n values for prog based on these probabilities.

5. ses|prog, schtyp: Using the hsbdemo data, fit a multinomial regression model for ses on schtyp and prog. Based on this model, estimate the probability for program status, and sample n values for ses based on these probabilities.

6. honors|ses, prog: Using the hsbdemo data, fit a logistic regression model for honors on ses and prog. Based on this model, estimate the probability of being enrolled in honors, and sample n values for honors based on these probabilities.

7. Let $\boldsymbol{B}$ be the $n \times 6$ matrix with columns containing the generated values for female, schtyp, ses, prog, honors , and cid.

8. math|prog, ses, female: Using the hsbdemo data, fit a normal regression model for math conditional on prog,ses,female. We obtain the model below.

$$\begin{aligned}\hat{\text{math}} &= 47.9 + 5.88I(\text{prog} = academic) - 3.84I(\text{prog} = vocational) \\ &\quad + 2.93I(\text{ses} = middle) + 4.57I(\text{ses} = high) - 0.20I(\text{female} = 1) + \epsilon_2,\end{aligned} \tag{2.44}$$

where $\epsilon_2 \sim N(0, 6.37^2)$. From (2.44), sample sample n values for $\mathsf{math}|\,\boldsymbol{B}$. Let F_2 be the $n \times 7$ matrix with columns containing the generated values for math, female, schtyp, ses, prog, honors , and cid.

9. read|math, prog: Using the hsbdemo data, fit a normal regression model for read conditional on math and prog. We obtain the following:

$$\begin{aligned}\widehat{\mathsf{read}} = 17.1 + 0.65\mathsf{math} + 2.02I(\mathsf{prog} = academic) \\ -1.20I(\mathsf{prog} = vocational) + \epsilon_1,\end{aligned} \tag{2.45}$$

where $\epsilon_1 \sim N(0, 6.25^2)$. From (2.45), sample n values for $\mathsf{read}|\,\mathsf{math}, \boldsymbol{B}$. Let F_1 be the $n \times 7$ matrix with columns containing the generated values for read, female, schtyp, ses, prog, honors , and cid.

To satisfy the block size requirement, for any k such that $c_{(k)} > 10$, we randomly sample $c_{(k)} - 10$ in-block records. Selected records are declared T_1 seeds. Let $t^* = \sum_k max(0, c_{(k)} - 10)$, or the total number of T_1 seeds needed to fulfill the block size restriction. Each simulation setting specifies a certain number N_{T_1} of T_1 seeds. We select the remaining $N_{T_1} - t^*$ required T_1 seeds at random from the non-seed pairs.

We introduce faulty values for the BV, prog. To do so, we select the record pairs with faults at random, and sample the faulty reported values of prog uniformly from the available incorrect values. We repeat this process independently 100 times at each simulation setting, generating new sets of 5000 records each time.

In each data set, we apply both BLASE and GM, with the latter assuming (incorrectly) that all records are T_1 or T_2 seeds. We also apply GM to the PB data as a baseline, as this is the best BLASE or GM could hope to do with the faulty data. For each replicate and each estimation method, we run a MCMC chain for 10000 iterations, thinned every two iterations, with 500 iterations used for burnin. When fitting BLASE, we only allow prog to change; that is, all F_1 records are T_2 seeds, and

all F_2 records are seeds for all $j \neq \mathsf{prog}$. This corresponds to the prior belief that $\gamma_{2j} = 0$ for all $j \neq \mathsf{prog}$. The linking analysis includes the linear regression of $Y_1 = \mathsf{read}$ on math and prog,

$$\begin{aligned} Y_1 =& \beta_0 + \beta_1 \mathsf{math} + \beta_2 I(\mathsf{prog} = academic) + \beta_3 I(\mathsf{prog} = vocational) + \epsilon_1 \\ =& \mathbf{X_1}\beta + \epsilon_1, \end{aligned} \tag{2.46}$$

where $\epsilon_1 \sim N(0, \sigma_1^2$, and a linear regression for $Y_2 = \mathsf{math}$ on female, prog, and ses,

$$\begin{aligned} Y_2 =& \eta_0 + \eta_1 I(\mathsf{female} = yes) + \eta_2 I(\mathsf{prog} = academic) + \\ & \eta_3 I(\mathsf{prog} = vocational) + \eta_4 I(\mathsf{ses} = Middle) + \eta_5 I(\mathsf{ses} = High) + \epsilon_2 \\ =& \mathbf{X_2}\eta + \epsilon_2, \end{aligned} \tag{2.47}$$

where $\epsilon_2 \sim N(0, \sigma_2^2)$. These models are the same as those used for data generation. Let $\mathbf{X_1}$ denote the design matrix for (2.46), and let $\mathbf{X_2}$ denote the design matrix for (2.47).

Let $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^T$, $\boldsymbol{\beta}_1 = [\beta_0, \beta_2, \beta_3]^T$, $\boldsymbol{\eta} = [\eta_0, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5]^T$. As a prior distribution for $\Theta = (\boldsymbol{\beta}, \sigma_1, \boldsymbol{\eta}, \sigma_2)$, we use $p(\Theta) \propto (1/\sigma_1^2)(1/\sigma_2^2)$. Conditional on $C^{(s+1)}$,$Y_1^{(s+1)}$, and $Y_2^{(s+1)}$, let $\mathbf{X_1}^{(s+1)}$ and $\mathbf{X_2}^{(s+1)}$ represent the design matrices of the regression of Y_1 and Y_2, respectively. Let $N_B^{(s)} = \sum_{k=1}^{K} n_{(k)}^{(s)}$ be the number of individuals as defined by $C^{(s+1)}$. The Gibbs sampling steps for the simulation then proceed as follows.

1. Impute missing values of Y_2 for records linked to dummy records. Let $\mathbf{V}^{(s+1)}$ denote $\mathbf{X_1}^{(s+1)}$ without the column corresponding to Y_2. Dropping the superscripts (s) on Θ, we impute using the following conditional distribution:

$$Y_2 | Y_1^{(s)}, \Theta^{(s)}, C^{(s+1)}, \boldsymbol{B}^{(s+1)} \sim N(\mu, \Sigma I), \tag{2.48}$$

$$\mu = \Sigma \times (\sigma_2^{-2} \mathbf{X_2}^{(s+1)} \boldsymbol{\eta} + \sigma_1^{-2} \beta_2 (Y_1^{(s)} - \mathbf{V}^{(s+1)} \boldsymbol{\beta}_1)), \tag{2.49}$$

$$\Sigma = (\sigma_2^{-2} + \sigma_1^{-2} \beta_2^2)^{-1}. \tag{2.50}$$

Here I is the identity matrix. Let $Y_2^{(s+1)}$ comprise Y_2 and the imputed values of Y_2.

2. Impute missing values of Y_1 for records linked to dummy records using the following conditional distribution:

$$Y_1|\mathbf{X_1}^{(s+1)}, C^{(s+1)}, \Theta^{(s)} \sim N(\mathbf{X_1}^{(s+1)}\boldsymbol{\beta}^{(s)}, (\sigma_1^2)^{(s)}I).$$

Here I is the identity matrix. Let $Y_1^{(s+1)}$ comprise Y_1 and the imputed values of Y_1.

3. Sample $\sigma_1^{(s+1)}$ from the full conditional

$$\sigma_1^2|Y_1^{(s+1)}, \mathbf{X_1}^{(s+1)}, \boldsymbol{\beta}^{(s)} \sim InvGamma\left(\frac{N_B^{(s+1)} - p}{2}, SSE_1/2\right) \tag{2.51}$$

$$SSE_1 = (Y_1^{(s+1)} - \mathbf{X_1}^{(s+1)}\boldsymbol{\beta}^{(s)})^T(Y_1^{(s+1)} - \mathbf{X_1}^{(s+1)}\boldsymbol{\beta}^{(s)}). \tag{2.52}$$

4. Sample $\boldsymbol{\beta}^{(s+1)}$ from the full conditional

$$\boldsymbol{\beta}|\sigma_1^{(s+1)}, Y_1^{(s+1)}, \mathbf{X_1}^{(s+1)} \sim N\left(\hat{\boldsymbol{\beta}}, (\mathbf{X_1}^{(s+1)^T}\mathbf{X_1}^{(s+1)})^{-1}(\sigma_1^2)^{(s+1)}\right), \tag{2.53}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X_1}^{(s+1)^T}\mathbf{X_1}^{(s+1)})^{-1}\mathbf{X_1}^{(s+1)^T}Y_1^{(s+1)}. \tag{2.54}$$

5. Sample $\sigma_2^{(s+1)}$ from the full conditional

$$\sigma_2^2|\mathbf{X_2}^{(s+1)}, \boldsymbol{\eta}^{(s)} \sim InvGamma\left(\frac{N_B^{(s+1)} - q}{2}, SSE_2/2\right), \tag{2.55}$$

$$SSE_2 = (Y_2^{(s+1)} - \mathbf{X_2}^{(s+1)}\boldsymbol{\eta}^{(s)})^T(Y_2^{(s+1)} - \mathbf{X_2}^{(s+1)}\boldsymbol{\eta}^{(s)}). \tag{2.56}$$

6. Sample $\boldsymbol{\eta}^{(s+1)}$ from the full conditional

$$\boldsymbol{\eta}|\sigma_2^{(s+1)}, Y_2^{(s+1)}, C^{(s+1)}, \boldsymbol{B}^{(s+1)} \sim N\left(\hat{\boldsymbol{\eta}}, (\mathbf{X_2}^{(s+1)^T}\mathbf{X_2}^{(s+1)})^{-1}(\sigma_2^2)^{(s+1)}\right), \tag{2.57}$$

$$\hat{\boldsymbol{\eta}} = (\mathbf{X_2}^{(s+1)^T}\mathbf{X_2}^{(s+1)})^{-1}\mathbf{X_2}^{(s+1)^T}Y_2^{(s+1)}. \tag{2.58}$$

2.5.2 Simulation results

We compare BLASE and GM across the simulation scenarios using several metrics. First, we compute the posterior match rate (PMR), which is the number of correctly matched pairs divided by the total number of matched pairs. Since for BLASE (but not GM) the number of matched pairs can vary at each MCMC iteration, we compute the PMR for each post-burnin posterior sample of C, and average this across posterior samples. We also evaluate posterior inferences for Θ, the coefficients in both regressions in the linking analysis model. Finally, we compute an average root mean square error (RMSE) for prediction on out-of-sample cases. Here, we create a single test data set of 500 records simulated from the models used to create the 5000 records. In each replication, we then compute

$$RMSE = \sqrt{\frac{1}{500}\sum_{i=1}^{500}\left(\hat{Y}_{1i} - Y_{1i}\right)^2}. \tag{2.59}$$

Each $\hat{Y}_{1i}$ is sampled from $f(Y_1|\hat{\Theta}, Y_{2i}, B_{2i})$, where $\hat{\Theta}$ equals the posterior mean of Θ per BLASE or GM for that replication. In each simulation setting and for each method, we average the 100 values of (2.59) to obtain the average RMSE.

Figures 2.1 and 2.2 displays key results from the high seed, high fault (HSHF) scenario, which is where we expect the largest payoffs from using BLASE over GM. As evident in Figure 2.1, applying the GM to the PB data yields reasonable estimates of the coefficient for math in the regression model for Y_1; we focus on this coefficient because it is the most important predictor by far, as evident in the final row of Table 2.3. However, applying GM on the data with faulty BVs results in significant under-estimation of the math coefficient. BLASE estimates the coefficient more accurately than GM for all three prior settings, with the best (and comparable) results for the diffuse and appropriately concentrated priors. BLASE provides more reliable estimates of most of the remaining coefficients, as well. As a result, BLASE has

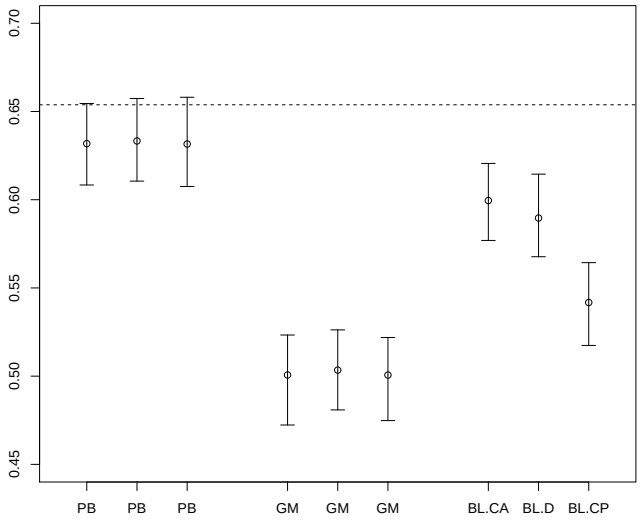


FIGURE 2.1: HSHF: Posterior means for math coefficient for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these posterior means. Dashed horizontal lines indicate true values of coefficients.

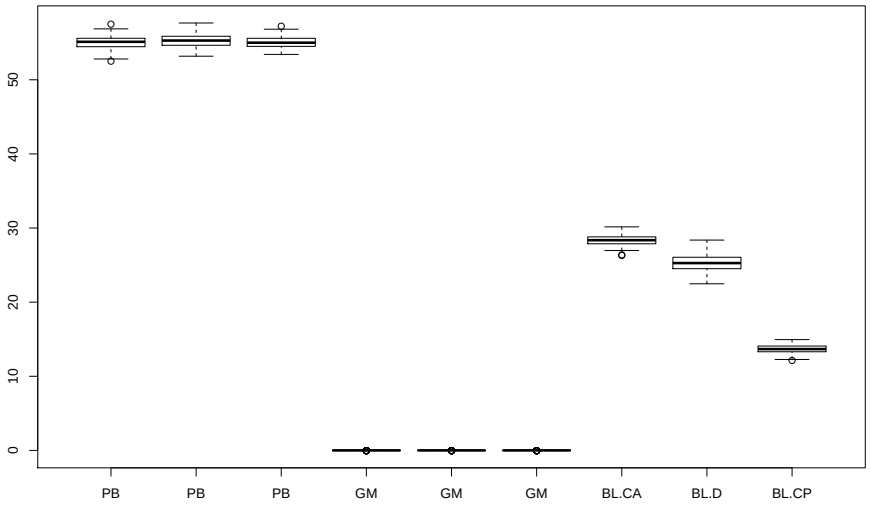


FIGURE 2.2: HSHF: Comparison of nonseed PMRs. Average posterior match rates from 100 simulations for GM and BLASE at each prior distribution.

smaller average RMSE than GM does, as shown in Table 2.3. We also note that BLASE has a larger average PMR than GM does (Figure 2.2).

Table 2.3: Summary of results of simulation studies. Positive values indicate BLASE is more accurate than GM, and negative values indicate the opposite. Bold values indicate statistically significant results (paired t-test p-value $< .05$). Entries under dPMR equal $100(PMR_{BLASE} - PMR_{GM})$. Entries under dRMSE are averages across 100 replications of $(RMSE_{GM} - RMSE_{BLASE})/RMSE_{PB}$. Entries under each predictor are averages across 100 replications of $100\left(\left|\hat{\theta}_{GM} - \theta\right| - \left|\hat{\theta}_{BLASE} - \theta\right|\right)/\theta$. The final row is the average t-values and R^2 values of the linking analysis across the 100 PB data sets.

		Y_1 Regression Linking Analysis Parameters					
		Intercept	Math	prog = Acad	prog = Voc	PMR	RMSE
	CA	**29.5**	**15.2**	**-14.7**	**42.8**	**11.4**	**3.3**
HSHF	D	**25.6**	**13.3**	**-14.0**	**41.0**	**10.1**	**2.5**
	CP	**11.9**	**6.3**	-3.1	**55.0**	**5.5**	**.50**
	CA	.3	.06	**-21.2**	1.2	**.95**	**.003**
LSHF	D	**5.9**	**2.8**	**-9.9**	**20.3**	**3.3**	**.008**
	CP	**9.9**	**4.8**	**-4.4**	**27.0**	**4.3**	**.011**
	CA	**.9**	**.5**	**2.1**	**4.4**	**-1.2**	.006
HSLF	D	.16	**.15**	**1.7**	**3.2**	**-1.6**	**-.02**
	CP	**-21.4**	**-11.0**	**-37.6**	**-69.1**	**-10.6**	**-.96**
	CA	**-12.6**	**-6.3**	.6	**-9.3**	**-1.9**	**-1.3**
LSLF	D	**-9.9**	**-5.0**	**-6.2**	**-13.0**	**-1.3**	**-.9**
	CP	**-40.0**	**-20.2**	**-45.7**	**-82.3**	**-9.9**	**-3.4**
t-values		27.7	56.3	8.9	-5.3	R^2	0.46

As is evident in Figures 2.1 and 2.2, using strong but inappropriate prior beliefs to specify $(a_{\gamma_{2j}}, b_{\gamma_{2j}})$ can degrade the performance of BLASE. This is true in most other simulation scenarios as well, as seen in Table 2.3. The most severe degradation occurs in the low seed, low fault (LSLF) scenario, where the poorly specified prior distribution is informative enough to cause many unnecessary changes in blocks. In contrast, across all scenarios, the diffuse prior distribution performs comparably to the better of the CA and CP prior distributions. Thus, in the absence of reliable prior information about the faultiness rates, we recommend using diffuse priors in

which the choice of $(a_{\gamma_{2j}}, b_{\gamma_{2j}})$ corresponds to a small prior sample size.

Focusing on results in Table 2.3 for the diffuse prior distribution, we can discern general patterns over the four simulation scenarios. First, in both HF scenarios, BLASE tends to result in significantly higher average PMR, lower average RMSE, and more accurate estimates of the Y_1 model coefficients than GM does. These improvements are smaller in the low fault scenario, in which the violation of the GM perfect blocking assumption is less problematic. Second, in the HSLF scenario, BLASE offers slight improvements over GM (using the CA or D priors). Apparently, even when only a modest number of records need to be moved, BLASE can leverage the information in the T_1 seeds to identify plausible moves. Third, in the LSLF scenario, BLASE is slightly outperformed by GM. Without adequate numbers of T_1 seeds, BLASE does not have enough information to determine accurate moves. We note that, although not apparent from Table 2.3, in scenarios with low numbers of T_1 seeds neither BLASE nor GM provide accurate estimates of the coefficients; see Appendix B for supporting results.

2.6 Illustrations with NC education data

In this section, we illustrate the performance of BLASE using data from the North Carolina Education Research Data Center (NCERDC, 2013). The data consist of 83,957 record pairs containing end-of-grade (EOG) math test scores for 7^{th} graders from 2011 (F_1) and their accompanying 8^{th} grade scores from 2012 (F_2). We work specifically with a subset of 77,998 record pairs that correspond to the three largest ethnic groups in the data set. These records are known from extensive clerical review to be perfectly matched, so we can treat the records as a truth for purposes of evaluating the performance of BLASE with genuine data.

We use seven categorical variables as the BVs: birth day, birth month, birth year, ethnicity (White, Black, Hispanic), sex, honors (1 if Honors enrolled, 0 if not), and a

school identifier (704 schools). Let Y_1 be the centered math EOG score for 2011, and let Y_2 be the centered math EOG score for 2012. As the linking analysis for BLASE and GM, as well as for model evaluation, we use

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 I(\text{Black}) + \beta_3 I(\text{Hispanic}) + \beta_4 I(\text{Male}) + \epsilon_1, \quad (2.60)$$

$$Y_2 = \eta_0 + \eta_1 I(\text{Black}) + \eta_2 I(\text{Hispanic}) + \eta_3 I(\text{Male}) + \epsilon_2, \quad (2.61)$$

where $\epsilon_1 \sim N(0, \sigma_1^2)$ and $\epsilon_2 \sim N(0, \sigma_2^2)$. In both models, we use a baseline of a female with ethnicity equal to white. In the perfectly matched data, the Y_1 regression has $R^2 = .69$, and the Y_2 regression has $R^2 = .1$.

The data contain a high number of seeds, with 62,276 pairs (roughly 80%) of the true matches agreeing uniquely on all seven BVs. Only about 9.2% of the true matches disagree on at least one BV field, with school being the most commonly different. This arises mainly because some students move schools over the year. We fit BLASE (using the diffuse prior) and GM to these data. Not surprisingly given the low error rates and high fraction of T_1 seeds, there is little difference in their performances. Additionally, in these data the school attended is not an important predictor of the other variables. As a result, BLASE has little information to encourage moves to the correct blocks.

As more stringent tests of BLASE, we generate more numerous and impactful errors by perturbing ethnicity. For all true matches $(i, i^{'})$ which are faulty on school, we set $\hat{B}_{2i'j} = \hat{B}_{1ij}$. We then make 15% of the values of ethnicity faulty; this corresponds to 75% of the values for the non-seeds. This variable is an important predictor in (2.60), so that errors in the variable could adversely affect the estimated regression. We sample 75% of the non-seeds completely at random, and for selected records sample a new value for ethnicity uniformly from incorrect values. We call this the block faults at random (FAR) scenario.

We also create a scenario in which the faulty value of ethnicity is not chosen

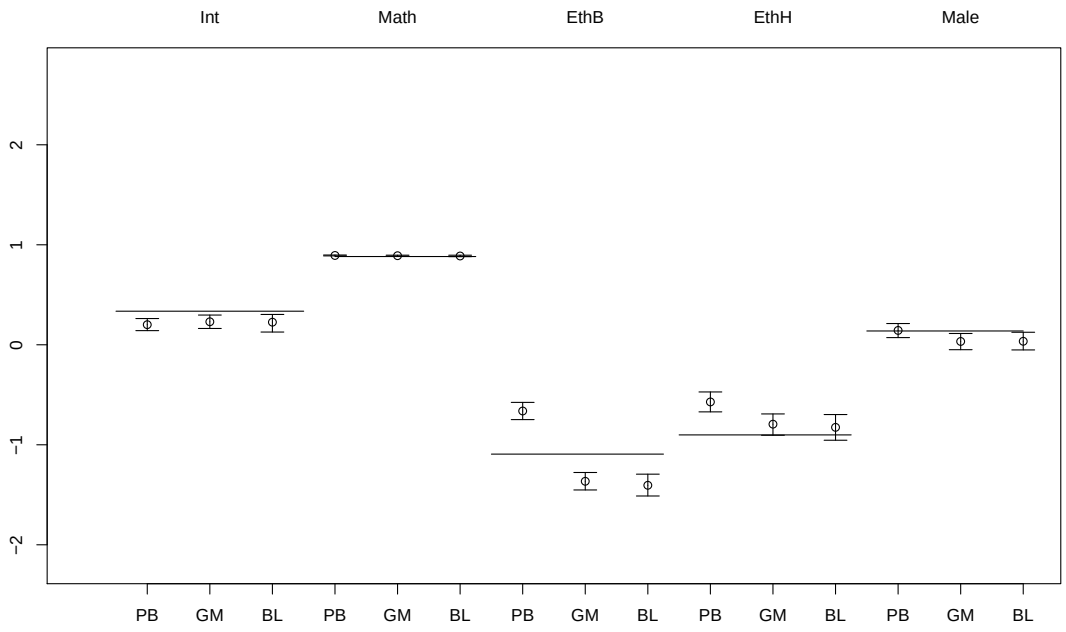


FIGURE 2.3: FAR: 95% posterior CIs for the parameters of the linking analysis. Results are from MCMC chain run for 10000 iterations, thinned every two draws, for each of the PB, GM and BLASE methods. The horizontal bars represent the OLS parameter estimates from the perfectly linked data. "EthB" stands for "Black", where "EthH" stands for "Hispanic". Note: For the math coefficient, the perfectly linked estimate is .883, with PB mean .893 (.889,.898), GM mean .891 (.886,.896), and BLASE mean .888 (.882,.896).

completely at random. Specifically, for each record chosen to be faulty, we make Black and Hispanic students have a reported value of white, and we make white students have a reported value of Black or Hispanic, chosen with equal probability. We call this the faulty not at random (FNAR) scenario. The FNAR scenario assesses the performance of BLASE when its model assumptions do not match the error generating process for the data.

Figure 2.3 displays the results of the FAR simulation. This simulation matches the characteristics of the high seed, low fault (HSLF) simulation from Section 2.5. As in that simulation, here BLASE and GM result in similar values of PMR, RMSE and point estimates of Θ for the Y_1 model. Apparently, the T_1 seeds are representative of

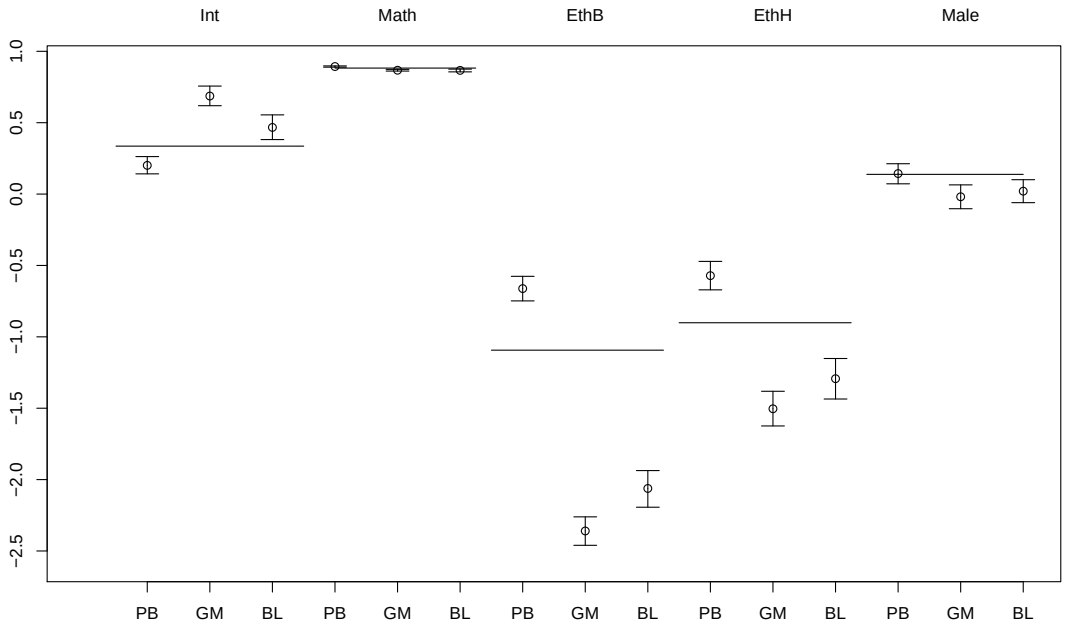


FIGURE 2.4: FNAR: 95% posterior CIs for the parameters of the linking analysis. Results are from one MCMC chain run for 10000 iterations, thinned every two draws, for each of the PB, GM and BLASE methods. The horizontal bars represent the OLS parameter estimates from the perfectly linked data. "EthB" stands for "Black", where "EthH" stands for "Hispanic". Note: For the math coefficient, the perfectly linked estimate is .883, with PB mean .893 (.889,.898), GM mean .867 (.862,.873), and BLASE mean .867 (.856,.875).

the relationships in the data, and the signal from these complete cases dominates the aberrant records with block faults. Interestingly, BLASE indicates larger posterior uncertainty about Θ than GM does. This actually is sensible, as the block faults generate increased posterior uncertainty that should be reflected in inferences.

Unlike the FAR simulation, in the FNAR simulation GM is unable to locate the true mode for the parameters of the linking analysis. As evident in Figure 2.4, GM overestimates the intercept resulting in underestimation for the ethnicity coefficients. BLASE estimates the model more accurately, albeit still with bias, as would be expected since the measurement error model in BLASE does not perfectly capture the FNAR error generation process. The increase in accuracy is accompanied by an

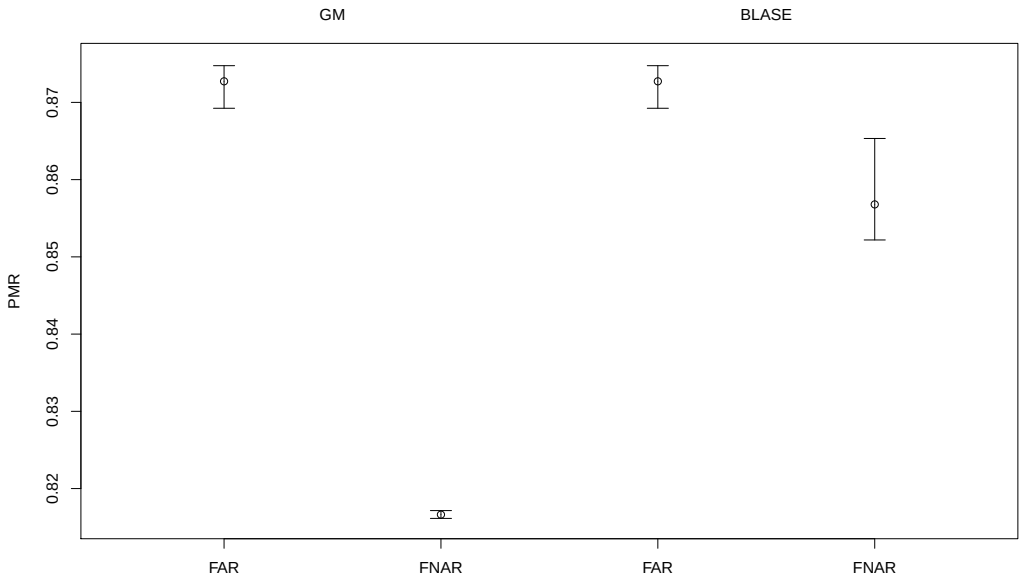


FIGURE 2.5: 95% central intervals for posterior match rates in the FAR and FNAR scenarios.

increase in PMR from 82% in GM to 86% for BLASE, as seen in Figure 2.5. Thus, by allowing blocks to move, BLASE uses information from the seeds to improve the file matching compared to GM, even though it uses the wrong measurement error model.

2.7 Conclusion

BLASE is a Bayesian approach to file matching designed to estimate regression models and match records simultaneously when categorical matching variables are subject to reporting error. Simulation results suggest that BLASE performs well relative to GM in scenarios with high numbers of T_1 seeds and a high number of BV faults, resulting in more accurate estimates of the linking analysis parameters and increased posterior match rates. Results also suggest that prior settings for γ can impact posterior results. When using BLASE, we recommend diffuse prior values especially

absent reliable information to think otherwise, as strongly concentrated prior distributions reflecting inaccurate prior beliefs can degrade the performance of BLASE.

3

File linking with faulty continuous matching variables

In this chapter, we consider linking scenarios with faulty continuous matching variables. We develop a hierarchical model that links files in the presence of such faults, and implement the technique on both simulated data and data from the U.S. Census Bureau's Census of Manufactures (CMF).

Like most surveys, the CMF is prone to item non-response, meaning that the collected data are often incomplete. For missing data, as well as data identified as faulty, the U.S. Census Bureau uses a variety of techniques to replace reported data or fill-in missing data with imputed values (White, 2014; White et al., 2015). Each missing or incorrect value is replaced with a single imputed value. The resultant complete data sets are used for research in a wide variety of disciplines, with analysts generally treating the imputations as true values. In this chapter, we also treat these Census Bureau imputations as being true, and any analysis is reflective of results that would be obtained by researchers working with the CMF. However, as discussed in Chapter 1, replacing missing data with a single imputation can underestimate variability (Rubin, 1987; Little and Rubin, 2002). In Chapter 4, we implement an

alternative imputation strategy and replace the single imputations with multiple imputations.

3.1 Introduction

The CMF is a census containing information on every U.S. manufacturing plant. In addition to information on sales, production, and product information, the CMF contains information related to energy usage. In the energy economics literature, these data are analyzed to better understand trends in energy usage and efficiency. Specifically, there is interest in examining the "energy-efficiency gap", or failure to implement cost-effective technologies to improve energy efficiency (Allcott and Greenstone, 2012; Gerarden et al., 2015).

For small to medium-sized manufacturers, information about corporate energy efficiency is provided publicly by the U.S. Department of Energy. This information is obtained through no-cost energy assessments provided by the U.S. Department of Energy Industrial Assessment Centers (IAC). Assessments are conducted by teams of university faculty and students who visit manufacturing plants to assess productivity, energy use and waste. Visits result in a report detailing all cost-saving opportunities identified during the assessment (U.S. Department of Energy, 2016).

Manufacturing plants that receive IAC energy assessments represent a subset of the plants in the United States, and hence a subset of the plants in the CMF. Linking a record in the IAC to a record in the CMF is synonymous with declaring that the manufacturing plant associated with the Census record received an IAC energy assessment. Such a link would provide information on the energy efficiency recommendations provided to a plant by IAC, enabling researchers to examine CMF energy data within the context of the IAC recommendations.

Records from the IAC assessments are publicly available. However, due to privacy restrictions, identifying information such as plant name have been removed from the

released IAC records. Linking IAC records to CMF records therefore relies on a set of categorical and continuous variables common to both files. However, linking in this scenario is complicated by a few characteristics of the data. First, the number of IAC assessments conducted in a given year is much smaller than the number of CMF records available for the same year. This means that linking on common categorical variables rarely results in unique matches in the CMF. Second, the continuous variables common to both files do not necessarily agree for true matches. For instance, values in the IAC relating to sales information are often rounded, and information for each completed form is provided at potentially different times of the year. This means that even for true matches, we do not expect values of common continuous variables to agree exactly.

In this chapter, we develop a file linking methodology suitable for linking records from a small file to records from a larger file in the presence of faulty, continuous matching variables. The method utilizes common categorical variables, which we call blocking variables (BVs), to define blocks of records in the larger file that contain possible matches for each record in the smaller file. In contrast to the BLASE model in Chapter 2, we assume that these BVs are fault-free. Once blocks of records are identified, we utilize common continuous variables, which we call matching variables (MVs), to estimate which in-block record in the larger file is a match for each record in the smaller. We assume that these MVs may not agree exactly across matched pairs. We refer to the entire model as LFCMV, which stands for linking with faulty continuous matching variables.

LFCMV linking proceeds in four main steps. First, we group records into blocks based on patterns of agreement in the BVs. Second, we utilize a linking model that models the distance between the MVs of records in the smaller file and the MVs of their matching records in the larger file. This linking model is conditional on a vector C that indicates which record in the larger file is matched to each record in

the smaller. Third, we specify a mixture model for the distance component of the linking model. This latent structure allows the distance between MVs in matched pairs to vary across types of matched pairs. Finally, within each block defined by the BVs, we specify a model for the corresponding elements of C.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce the LFCMV methodology. In Section 3.3, we describe the Gibbs sampling steps used to obtain posterior estimates of the linkage structure. In Section 3.4, we use simulation studies to illustrate the performance of LFCMV. In Section 3.5, we present results of additional studies that investigate the performance of LFCMV under a variety of fault scenarios. In Section 3.6, we apply the model to link records from the 2007/2008 IAC assessments to the 2007 CMF and conclude with a discussion of future research directions.

3.2 Model

In this section, we describe the LFCMV methodology. Notation and key concepts are presented Section 3.2.1. We then present the model for C in Section 3.2.2, and the linking model in Section 3.2.3.

3.2.1 Notation

Let the two files to be linked be denoted F_1 and F_2, containing n_1 and n_2 records, respectively. Without loss of generality, we assume that $n_1 < n_2$. We further assume that each record $i \in F_1$ has a match $i^{'} \in F_2$, that is, the set of individuals in F_1 is a subset of the set of individuals in F_2. This assumption is motivated by the application in which the IAC assessment records (F_1) represent a subset of the records in the CMF (F_2). We assume each record corresponds to a single true individual.

Our goal is to link each record $i \in F_1$ to a record $i^{'} \in F_2$. To reduce the number of possible matches for each i, we place each of the n_1 records in F_1 in its own block.

Let (i) denote the block defined by $i \in F_1$. For each record i, we limit the possible matches to a set of records $i^{'} \in F_2$ which are assigned to block (i). Possible matches $i^{'} \in F_2$ are assigned to (i) as follows.

Let J denote the number of BVs, i.e., the number of categorical variables common to F_1 and F_2 that are used for blocking. It is not necessary to use all common categorical variables for blocking; analysts can select variables appropriate for blocking for each application. Let B_{fik} denote the BV value for record i in file f on field k, where $k = 1, \ldots, J$, $i = 1, \ldots, n_f$, and $f = 1, 2$. For all (f, i), let $B_{fi} = (B_{fi1}, \ldots, B_{fiJ})$. For each $i \in F_1$, define $F_{2(i)}$ as the set of all $i^{'} \in F_2$ such that $B_{1ik} = B_{2i^{'}k}$ for all $k = 1, \ldots, J$. Assign all records in $F_{2(i)}$ to block (i). In other words, for each i, we restrict the set of possible matches to records $i^{'}$ in F_2 with the same BV values as record i.

Under this blocking structure, it is possible that a record $i^{'}$ in F_2 can be considered a possible match for more than one record i in F_1. Each record i in F_1 is placed in its own block, and the set of possible matches $F_{2(i)}$ are defined only by the BV combination B_{fi}. If two records a and b in F_1 have the same BV combination, their sets of possible matches, $F_{2(a)}$ and $F_{2(b)}$, contain the same records from F_2. This yields the possibility that both a and b can be linked to the same record in F_2. For our application, this is not a concern. Our goal is to identify records in the CMF that have received an IAC assessment, not necessarily to generate a set of matched pairs. If unique matches are required for a given application, one could imagine adapting this blocking structure to suit this need. For instance, blocks may be defined such that more than one F_1 record may be assigned to a block. Matching would then assign each record in $F_{2(i)}$ to at most one record i in F_1.

With n_1 records in F_1, we have n_1 blocks of records from F_2. Each block $F_{2(i)}$ represents potential matches for the accompanying record $i \in F_1$. Let $n_{(i)}$ be the number of records in $F_{2(i)}$. For each $i \in F_1$, the goal of file linking is to determine

which of the $n_{(i)}$ records is a match for i. We define $C_i = i^{'}$ such that $i^{'} \in F_{2(i)}$ and $(i, i^{'})$ is a match. The n_1 element vector $C = (C_1, \ldots, C_{n_1})$ then specifies which records from F_2 match the records in F_1. This is similar to the linkage structure used for BLASE in Chapter 2.

Finally, let p be the number of common continuous variables (MVs) used for matching. For all $j = 1, \ldots, p$, let Y_{ij} denote the standardized value for MV j for record i in F_1. We provide details on this standardization in Section 3.4.1. For all $i = 1, \ldots, n_1$, let $Y_i = (Y_{i1}, \ldots, Y_{ip})$. Similarly, let $\mathbf{Y}_j = (Y_{ij}, \ldots, Y_{n_1 j})$ for all j. We define $\mathbf{Y}$ as the $n_1 \times p$ matrix of MV values for F_1. Similarly, for all j, let $X_{i^{'}j}$ denote the standardized value for MV j for record $i^{'} \in F_2$ where $i^{'} = 1, \ldots, n_2$. Let $X_i = (X_{i1}, \ldots, X_{ip})$ and $\mathbf{X}_j = (X_{ij}, \ldots, X_{n_1 j})$. We define $\mathbf{X}$ as the $n_2 \times p$ matrix of MV values for F_2.

3.2.2 Model for C

Within a block $F_{2(i)}$, there are $n_{(i)}$ possible values of C_i. We denote these values as ℓ, where $\ell \in \{1, \ldots, n_{(i)}\}$. We assume that *a priori*, each $i^{'}$ in $F_{2(i)}$ is equally likely to be linked with i, so that for all ℓ,

$$p(C_i = \ell \mid B_{1i}) = \frac{1}{n_{(i)}}. \tag{3.1}$$

Conditional on B_{1i}, which defines the block for record i, each C_i is modeled independently.

3.2.3 Linking model

Conditional on C, we assume that for each matching pair $(i, i^{'} = C_i)$, the values of Y_i are related to X_{C_i} through a linking model. This model reflects the belief that for true matches $(i, i^{'} = C_i)$, we may have $Y_{ij} \neq X_{i^{'}j}$ for some $j = 1, \ldots, p$. Such a discrepancy could occur due to incorrect recording, time differences in data collection, different accuracy thresholds, etc. The notion of linking based on examining

distances between variables common to both files is an example of distance-based record linkage. Distance-based record linkage is commonly used for re-identification purposes in disclosure and privacy research (e.g., Pagliuca and Seri, 1999; Domingo-Ferrer and Torra, 2002a, 2003; Torra et al., 2006). In these applications, distances between each record i and $i^{'}$ are computed using some distance metric, and the record pair that has the smallest distance according to this metric is considered a link (Domingo-Ferrer and Torra, 2002b).

We model Y_i with a multivariate normal distribution centering Y_i at its corresponding X_{C_i}. This choice of linking model favors matched pairs with similar values of Y and X. The variance component of the linking model represents the distance between Y_i and X_{C_i} that is considered plausible for matched pairs. In our application, each IAC record corresponds to a manufacturing plant. It is reasonable to assume that the distance in the MVs across matched pairs may vary with state, plant type, or other characteristics of the records. In more general applications, the distance may vary across certain blocks, or with other features in the data. To allow the linking model to capture these distributional features, we utilize a mixture of multivariate normals. Mixtures of multivariate normals are highly flexible and, with enough components, can represent any distribution. For this application, the mixture framework allows the distance across types of matched pairs to vary with latent class.

Following standard conventions for a mixture model, assume each pair $(i, i^{'} = C_i)$ belongs to one of H latent classes. For convenience, we associate the latent class with record i. Let $z_i \in \{1, \ldots, H\}$ represent the latent class assignment for $i \in F_1$, with $z = (z_1, \ldots, z_{n_1})$. For $h = 1, \ldots, H$, we assume that $Pr(z_i = h) = \pi_h$ for all i. Let $\pi = (\pi_1, \ldots, \pi_H)$.

Conditional on z and C, we model Y_i with a multivariate normal distribution centered at X_{C_i} and class-specific variance Σ_h. This mixture model can be written

as

$$Y_i|\mathbf{X}, C, \mathbf{\Sigma}, z_i \sim N_p\left(X_{C_i}, \Sigma_{z_i}\right) \tag{3.2}$$

$$z_i|\boldsymbol{\pi} \sim Multinomial\left(1; \pi_1, ..., \pi_H\right). \tag{3.3}$$

This is an adaption of the standard mixture of multivariate normals in which each Y_i is modeled as $Y_i \mid \boldsymbol{\mu}, \mathbf{\Sigma}, z \sim N_p\left(\mu_{z_i}, \Sigma_{z_i}\right)$. In the standard formulation, each Y_i is assumed to have a component specific mean μ_{z_i}. In our framework, we center each Y_i at its corresponding X_{C_i}. This reflects a belief that for matched pairs, Y_i and X_{C_i} should be close together. In our application, $p = 2$, so (3.2) is bivariate normal.

Following Kim et al. (2014), we select a conjugate prior structure for each Σ_h as follows:

$$\Sigma_h \sim InvWishart\left(f_0, G_0\right), \tag{3.4}$$

where $G_0 = Diag(\phi_1, ..., \phi_p)$ and

$$\phi_j \sim Gamma(a_\phi, b_\phi), \tag{3.5}$$

with $E(\phi_j) = a_\phi / b_\phi$. This conjugate prior structure facilitates posterior updates, as discussed in Section 3.3. In order to ensure a proper posterior distribution, we set $f_0 = p + 1$. We set the hyper-parameters $a_\phi = b_\phi = 0.25$, a choice which allows substantial prior mass at modest sized variances (Kim et al., 2014).

For the latent class weights π, we use a stick-breaking representation of the truncated Dirichlet process (Sethuraman, 1994; Ishwaran and James, 2001). In this framework, the mixture probabilities are

$$\pi_h = V_h \prod_{g<h} (1 - V_g), h = 1, ..., H \tag{3.6}$$

$$V_h \sim Beta(1, \alpha), V_H = 1 \tag{3.7}$$

$$\alpha \sim Gamma(a_\alpha, b_\alpha). \tag{3.8}$$

We set $a_\alpha = b_\alpha = 0.25$, reflecting a low prior sample size and hence a vague prior for α. As noted in Escobar and West (1995), such a choice allows the data to dominate in the posterior.

The stick-breaking representation allows the data to inform the number of latent classes that are occupied at any iteration of the posterior sampler. As H represents the maximum number of occupied classes, we recommend starting with a large value, such as $H = 30$. When posterior runs indicate that all H classes are consistently occupied, we increase H and restart the sampling.

There are a few considerations to take into account when using LFCMV for file linking. First, as is true for BLASE and indeed file linking methodologies in general, block size is important to the performance of the model. Smaller blocks tend to lead to a higher match rate, while for extremely large blocks, the model often selects a match essentially at random. It is therefore important to select a blocking scheme that results in small blocks. Second, the linking model described in Section 3.2.3 is based on distance between Y_i and its corresponding X_{C_i}. As we illustrate in Section 3.4, in the posterior, the model tends to assign high probability to matches with small distances across matched pairs. The determination of what is a reasonable distance across a matched pair is influenced by the class-specific variance component of the linking model.

3.3 Posterior sampling

In this section, we describe the Gibbs sampler used to estimate the posterior distribution of the LFCMV model.

3.3.1 Initialization

We initialize $C^{(0)}$ such that

$$C_i^{(0)} = \left(i^{'} \mid d(i, i^{'}) = min \left(d(i, \ell), \ell \in F_{2(i)} \right) \right), \tag{3.9}$$

where

$$d(i,i^{'}) = \sqrt{\sum_{j=1}^{p}\left(\frac{Y_{ij} - X_{i^{'}j}}{Y_{ij}}\right)^2}. \tag{3.10}$$

In other words, within a block, we initialize C_i such that $d(i,i^{'})$ is minimized. The two closest records according to this metric will be initialized as links. Because Y_i and X_{C_i} are standardized variables, $d(i,i^{'})$ assigns equal weight to the distances in each of the p MVs. The importance of such standardization in distance-based linkage is discussed in Pagliuca and Seri (1999).

For the linking model, we set $\alpha^{(0)} = 1$, and initialize V, π, z, and $\mathbf{\Sigma}$ by drawing from the appropriate distributions from Section 3.2.3.

3.3.2 Posterior sampling algorithm

With the truncated representation of the Dirichlet process, posterior sampling for the entire model is facilitated with a Gibbs Sampler as follows.

1. For $h = 1, ..., H$, update Σ_h from the full conditional

$$\Sigma_h^{(s+1)} | \mathbf{Y}, \mathbf{X}, C^{(s)}, z^{(s)} \sim InvWishart\left(f_0 + n_h^{(s)}, G_0 + S_h^{(s)}\right), \tag{3.11}$$

   where $n_h^{(s)} = \sum_i I(z_i^{(s)} = h)$ denotes the number of individuals in latent class h at iteration (s) and $S_h^{(s)} = \sum_{i:[z_i^{(s)}=h]} (Y_i - X_{C_i^{(s)}})(Y_i - X_{C_i^{(s)}})^T$.

2. For $h \in \{1, ..., H-1\}$, update V_h from the full conditional,

$$V_h^{(s+1)} | z^{(s)}, \alpha^{(s)} \sim Beta\left(1 + n_h^{(s)}, \alpha^{(s)} + \sum_{g=h+1}^{H} n_g^{(s)}\right). \tag{3.12}$$

   Set $V_H^{(s+1)} = 1$.

3. Set $\pi^{(s+1)} = V_h^{(s+1)} \prod_{g<h} (1 - V_g^{(s+1)})$ for all $h \in \{1, ..., H\}$ per (3.6).

4. For $j = 1, \ldots, p$, update ϕ_j from the full conditional

$$\phi_j^{(s+1)} | \mathbf{\Sigma}^{(s+1)} \sim Gamma \left(a_\phi + \frac{1}{2} H f_0, b_\phi + \frac{1}{2} \sum_{h=1}^{H} \left(\Sigma_h^{-1}[j,j]^{(s+1)} \right) \right), \quad (3.13)$$

where $\left(\Sigma_h^{-1}[j,j] \right)^{(s+1)}$ represents the j^{th} diagonal entry of $(\Sigma_h^{-1})^{(s+1)}$.

5. Update α from the full conditional,

$$\alpha^{(s+1)} | \pi^{(s+1)} \sim Gamma \left(a_\alpha + H - 1, b_\alpha - log(\pi_H^{(s+1)}) \right). \quad (3.14)$$

6. For $i \in 1, ..., n$, sample the latent class indicator $z_i \in \{1, ..., H\}$ from a multinomial full conditional,

$$z_i^{(s+1)} | \phi^{(s+1)}, C^{(s)} \sim Multinomial(\pi_1{}^*, \pi_2{}^*, ..., \pi_H{}^*), \quad (3.15)$$

where

$$\pi_h{}^* = \frac{\pi_h^{(s+1)} f(Y_i | X_{C_i^{(s)}}, \Sigma_h^{(s+1)})}{\sum_{g=1}^{H} \pi_g^{(s+1)} f(Y_i | X_{C_i^{(s)}}, \Sigma_g^{(s+1)})}. \quad (3.16)$$

7. For $i = 1, ..., n_1$, we update each C_i. For each $i' \in F_{2(i)}$,

$$Pr \left(C_i = i' | Y_i, \mathbf{X}, z_i^{(s+1)}, \mathbf{\Sigma}^{(s+1)} \right) = \frac{f(Y_i | \mathbf{X}, C_i = i', \Sigma_h^{(s+1)})}{\sum_{\ell \in F_{2(i)}} f(Y_i | \mathbf{X}, C_i = \ell, \Sigma_h^{(s+1)})}. \quad (3.17)$$

The update in (3.17) represents draws from the multinomial Bernoulli posterior distribution of C. Because each block contains only one Y_i, computing (3.17) requires enumerating $n_1 \times \left(\sum_i n_{(i)} \right)$ multivariate normal likelihoods at each MCMC iteration. Unlike many file linking applications in which there are multiple records from each file in each block, for LFCMV this direct updating is fairly efficient in terms of computation time.

3.4 Illustrative simulation

In this section we illustrate the performance of the LFCMV model. For this simulation, we use IAC data to create both F_1 and F_2. F_1 and the corresponding matches are constructed using records from 2007 and 2008, while the remainder of F_2 is constructed using records from the remaining IAC years. In Section 3.4.1, we describe the process of creating F_1 and F_2 and introduce the BVs and MVs used in linking. Simulation results are presented in Section 3.4.2.

3.4.1 Data

The complete downloaded IAC database (up through the year 2016) contains 17583 records with 56 variables. In this simulation, we use two categorical variables, NAICS code and state, as the BVs. The state variable refers to the state in which a manufacturing plant is located. NAICS code refers to a 6-digit numerical code corresponding to the specific products made by a given plant (U.S. Census Bureau, 2016). NAICS codes are nested up to the 4-digit level, meaning that all plants that produce a certain type of product, say dairy products, must agree on at least the first 4 digits of their NAICS codes. In this simulation we define blocks using state and the first 4-digits of the 6-digit NAICS code.

Two continuous variables, sales and number of employees, serve as our MVs. Sales is a continuous variable for the sales of each plant in U.S. dollars, and employees refers to the total number of employees for the plant. We let $j = 1$ refer to sales and $j = 2$ to number of employees.

3.4.1.1 Create F_1

To create F_1, we use the IAC records from 2007 and 2008, comprising 797 records. We reduce this set to the final set of 512 F_1 records using the steps outlined in the second column of Table 3.1. For this simulation, we use use 16786 IAC records from

Table 3.1: Summary of the process of creating F_1 and F_{2U}. The "Action" column describes the data cleaning step. The center column describes the number of records in F_1 after each cleaning step. The far right column describes the number of records in F_{2U} after each cleaning step. The final row gives the number of records in the F_1 and F_{2U} used for linking.

Action	F_1 Records	F_{2U} Records
Initialize	All 2007/2008: 797	All other years: 16786
Remove records from Puerto Rico	786	16779
Remove States Unique to F_1 or F_2	778	16440
Remove Missing States	778	13622
Remove Missing Sales	764	13476
Remove Missing Employees	764	13464
Remove Missing Electricity Usage	763	13429
Remove Missing Cost of Energy	763	13422
Require Matching State/4-Digit NAICS	512	1170
Final Total:	512	1170

years other than 2007 or 2008 to create possible matches for F_1. If a record in F_1 has a state and 4-digit NAICS combination which is not found in these 16786 records, we remove these records from consideration for F_1. For simulation purposes, we also exclude from F_1 any records containing missing data. The process of incorporating missing data imputation into the file linking process adds an additional level of uncertainty; this is described in detail in Chapter 4. The selection process for F_1 results in the final count of $n_1 = 512$ records.

3.4.1.2 Create F_2

We create F_2 in two stages. First, for each record i in F_1, we create a matching record by adding noise to the MVs in record i. The collection of these n_1 matches is denoted F_{2M}. Second, we use the 16786 IAC records from years other than 2007 or 2008 to create sets of possible matches for each record i in F_1. These records represent the false matches, or possible options for i based on state and NAICS agreement which are in fact not a match for i. The collection of these records is denoted F_{2U}. We let $F_2 = (F_{2U}, F_{2M})$.

We create F_{2M} as follows. For each record i in F_1, we generate a matching record using the following steps. Assign each record i in F_1 to one of $H = 30$ latent classes z_i with

$$z_i \sim Multinomial(1; 1/H, \ldots, 1/H). \tag{3.18}$$

To generate the variance matrix, we sample from the following model:

$$\Sigma_h \sim InvWishart\left(50, G_{0h}\right), \tag{3.19}$$

where $G_{0h} = Diag(\phi_{1h}, \phi_{2h})$ and

$$\phi_{1h} \sim Gamma(1,1), \phi_{2h} \sim Gamma(25,3), \tag{3.20}$$

with $E(\phi_j) = a_\phi / b_\phi$. Let $\mathbf{Y}^{(0)}$ be the $n_1 \times p$ matrix containing all non-standardized MV values in F_1. For each i with latent class $z_i = h$, we define

$$F_{2M,i} = Y_i^{(0)} + \epsilon_i, \epsilon_i \sim N_2((0,0), \Sigma_{z_i}). \tag{3.21}$$

Here $F_{2M,i}$, $Y_i^{(0)}$ and ϵ_i are vectors of length $p = 2$. We repeat this match generating process 100 times for each i, creating 100 replicates $F_{2M,i}^{(m)}$, $m = 1, \ldots, 100$. Let $F_{2M}^{(m)} = \{F_{2M,i}^{(m)} \mid i = 1, \ldots n_1\}$.

We create F_{2U} using the 16786 IAC records from years other than 2007 or 2008. After removing missing data and other records as outlined in Table 3.1, we select 1170 records which have a state/4-digit NAICS combination that is observed in the F_1 data. Denote these 1170 records as F_{2U}. We let $F_2^{(m)} = (F_{2U}, F_{2M}^{(m)})$, where $m = 1, \ldots, 100$. For each replicate m, we link F_1 and $F_2^{(m)}$.

Each of the n_1 records from F_1 is assigned to a block (i). Based on the BVs, we assign each of the 1170 records in F_{2U} to these blocks. For each record $i \in F_1$, let $F_{2U(i)}$ denote the subset of F_{2U} assigned to block (i). As seen in Table 3.2, this results in blocks which range in size from 2 to 26. Blocks sizes, F_{2U}, and F_1 are consistent across replicates.

Table 3.2: Block sizes for the illustrative simulation. Block Size: the number of records from F_{2U} assigned to each block (i). Count: the number of blocks of each block size.

Block Size	2	3	4	5	6	7	8	9	10	11	12	13	14
Count	129	106	76	39	20	17	30	14	11	23	4	4	10
Block Size	16	17	18	19	24	25	26						
Count	2	3	4	4	8	3	5						

3.4.1.3 Standardization of Y and X

Before linking F_1 and $F_2^{(m)}$, we log transform the MVs. Let $\tilde{\mathbf{Y}}_j$ refer to the log-transformed MV values for field j in F_1, and $\tilde{\mathbf{X}}_j^{(m)}$ refer to the log-transformed MV values for field j in $F_2^{(m)}$. For ease of posterior computation in the linking model, we standardize $\tilde{\mathbf{Y}}_j$ and $\tilde{\mathbf{X}}_j^{(m)}$ as follows. Denote $S^{(m)} = (\tilde{\mathbf{Y}}_1, \tilde{\mathbf{X}}_1^{(m)})$ and $E^{(m)} = (\tilde{\mathbf{Y}}_2, \tilde{\mathbf{X}}_2^{(m)})$. Let $\bar{E}^{(m)}$ be the mean of all $n_1 + n_2$ elements of $E^{(m)}$ and let $\bar{S}^{(m)}$ be the mean of all $n_1 + n_2$ elements of $S^{(m)}$, with $sd(E^{(m)})$ and $sd(S^{(m)})$ representing the standard deviations of $E^{(m)}$ and $S^{(m)}$, respectively. We then standardize each transformed value as

$$\mathbf{Y}_1^{(m)} = (\tilde{\mathbf{Y}}_1 - \bar{S}^{(m)})/sd(S^{(m)}),\ \mathbf{X}_1^{(m)} = (\tilde{\mathbf{X}}_1^{(m)} - \bar{S}^{(m)})/sd(S^{(m)}), \tag{3.22}$$

$$\mathbf{Y}_2^{(m)} = (\tilde{\mathbf{Y}}_2 - \bar{E^{(m)}})/sd(E^{(m)}),\ \mathbf{X}_2^{(m)} = (\tilde{\mathbf{X}}_2^{(m)} - \bar{E}^{(m)})/sd(E^{(m)}). \tag{3.23}$$

3.4.2 Results

For each of 100 replicates, we apply the LFCMV model to link F_1 and the replicate specific $F_2^{(m)}$. We run the Gibbs sampler described in Section 3.3.2 for 5000 iterations with a burn in length of 200 iterations. For purposes of comparison, we also link each pair of data sets using a "naive" matching method. Specifically, for the naive method, we select a match $i' = C_i$ such that

$$C_i = \left(i' \mid d(i, i') = min\left(d(i, \ell), \ell \in F_{2(i)} \right) \right), \tag{3.24}$$

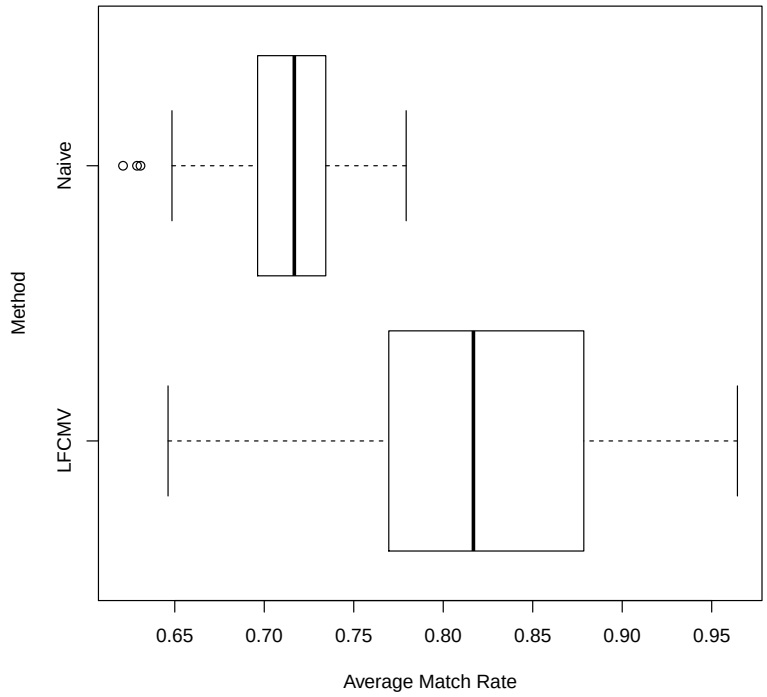


FIGURE 3.1: Match rates. Upper Boxplot: Match rate for each replicate obtained by linking records by the naive method. Lower Boxplot: Composed of average MR for each of 100 replicates of LFCMV.

where

$$d(i, i^{'}) = \sqrt{\sum_{j=1}^{p} \left(\frac{Y_{ij} - X_{i^{'}j}}{Y_{ij}} \right)^2}. \tag{3.25}$$

The primary metric used in this simulation is the match rate (MR), or the percentage of the n_1 records from F_1 that are correctly matched. For each replicate, the naive matching method yields one estimate of C. Conditional on this C, we compute a single estimate of the match rate. As a Bayesian procedure, LFCMV yields a set of posterior draws for C. For each replicate, we compute the MR for each posterior draw of C. We then average these MR values to obtain one estimate of the posterior MR for each replicate.

Results are displayed in Figure 3.1. For all but 2 of the 100 replicates, the MR

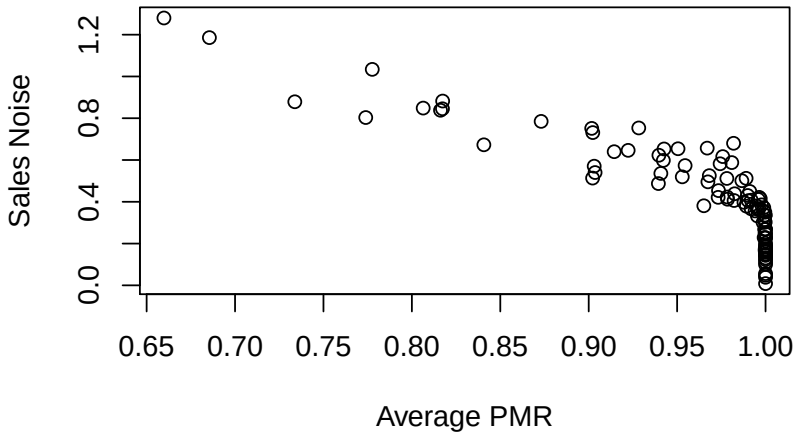


FIGURE 3.2: Average posterior match rates by average distance across matches pairs for sales from the 100 replicates. Lower PMR values are associated with greater distance.

obtained using LFCMV is higher than the MR obtained using the naive approach. A 95% posterior interval for the improvement in MR using LFCMV over the naive method is (.2%, 22%) with an average increase in match rate of 9%. This translates to an average of 47 additional records that are correctly matched using LFCMV.

As seen in Figure 3.1, the range of MR values is fairly wide. This range is due to the process of generating the X_i values. The average distance across matched pairs, i.e., the average value of $Y_{ij} - X_{C_i j}$, is different across each replicate. Consider Figure 3.2. Replicates with smaller average PMR have, on average, more distance between the MV values for sales across matched pairs. By design, LFCMV is searching for a match with small distance between Y_i and X_{C_i}. If the data generation process creates a matching X_{C_i} which is far from Y_i, the model sometimes selects records $X_{i' \neq C_i}$ with a smaller distance between Y and X. This behavior is consistent with the intuition discussed in Section 3.2.3.

3.5 Sensitivity to distance across matched pairs

In this section, we illustrate the performance of LFCMV linkage when the mechanism that introduces distance across matched pairs of MVs is not normally distributed. The creation of F_1 and F_{2U} is the same as the simulation study in Section 3.4. However, we vary the process of creating F_{2M} by using different distributions to create the distance across matched pairs. In Section 3.5.1, we apply uniform distance at three levels: 10%, 20% and 30%. For this simulation, we are interested in examining the performance of LFCMV with increasing distances between matches pairs. In Section 3.5.2, we introduce distance across matched pairs dependent upon the number of employees in a plant. This simulation leverages the ability of LFCMV to allow the distance between matched pairs to vary across types of pairs.

3.5.1 Uniform distance

In Section 3.4.2, the MR obtained by LFCMV tends to vary with the distance across matched pairs. To further explore this concept, we conduct a simulation in which distance between matched Y and X is generated uniformly at three different levels. Specifically, for each replicate, we generate $F_{2M}^{(m)}$ using

$$F_{2M,i1}^{(m)} = Y_{i1}^{(0)}(1+\kappa_{ij}), \kappa_{ij} \sim Uniform(0,u) \tag{3.26}$$

$$F_{2M,i2}^{(m)} = Y_{i2}^{(0)}(1+\kappa_{ij}), \kappa_{ij} \sim Uniform(0,u), \tag{3.27}$$

where $u = .1, .2$, and $.3$. We run 100 replicates at each setting of u.

Figure 3.3 displays the results across all three settings of u. As is true in Section 3.4, LFCMV generally yields higher match rates than the naive approach, and this result holds across all three settings of u. As u increases and true matched pairs have a greater distance between the MVs of matched pairs, both the naive and LFCMV linking methods have reduced performance. As previously mentioned, block size also has an impact on MR. Table 3.3 shows the breakdown of posterior MR by block size.

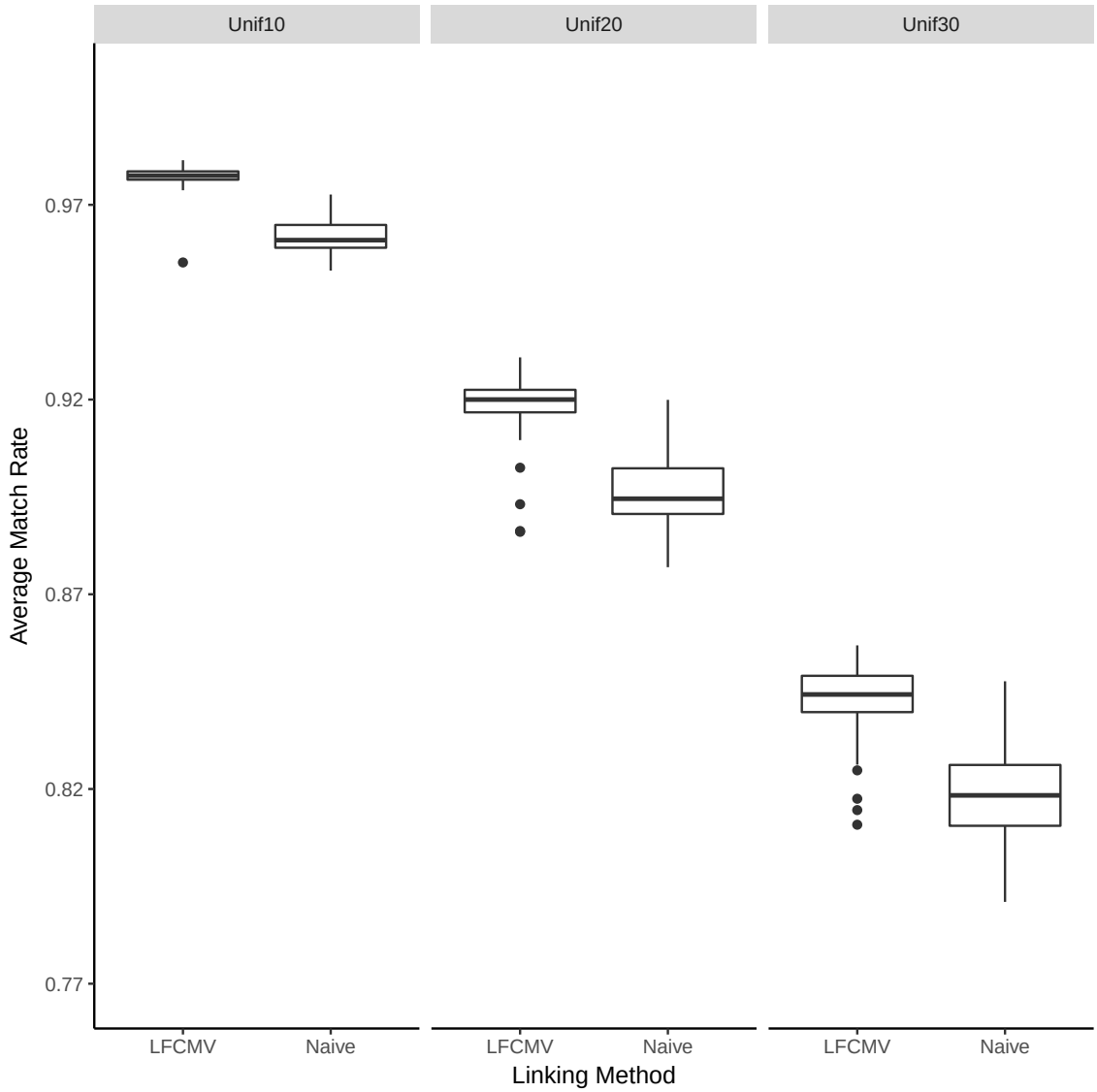


FIGURE 3.3: Match rates under uniform distance. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.

The average PMRs for smaller blocks tend to be higher than average PMR for larger blocks, with some variation attributable to the number of blocks of a given size.

Table 3.3: Average posterior match rate for each block size. The first column indicates the size of the block, with the number of blocks of that size in parentheses. Block sizes are consistent for each simulation. Remaining columns denotes the average posterior MR, as computed by averaging the block-specific MR across each of the 100 replicates.

Block Size (Count)	10%	20%	30%
2 (*129*)	.99	.98	.94
3 (*106*)	.99	.97	.93
4 (*76*)	.99	.94	.88
5 (*39*)	.98	.92	.80
6 (*20*)	.97	.90	.82
7 (*17*)	.98	.90	.78
8 (*30*)	.95	.89	.80
9 (*14*)	.91	.78	.66
10 (*11*)	.82	.56	.41
11 (*23*)	.93	.79	.64
12 (*4*)	1	1	1
13 (*4*)	.97	.85	.67
14 (*10*)	.99	.89	.72
16 (*2*)	1	.80	.57
17 (*3*)	1	.98	.82
18 (*4*)	1	.89	.61
19 (*4*)	.91	.50	.30
24 (*8*)	.99	.69	.40
25 (*3*)	1	.83	.57
26 (*5*)	1	.95	.80
Overall PMR	.98 (.003)	.92 (.007)	.84 (.008)

3.5.2 Size dependent distance

One of the features of LFCMV is the ability to model distances between the MVs that vary with types of matched pairs. To illustrate this capability, we conduct a simulation in which we introduce correlation between the number of employees in a plant and the distance across matched pairs of MVs. We classify the 261 plants in F_1 with more than 130 employees as "large" and classify the remaining 251 plants in F_1 with 130 or fewer employees as "small". We run a simulation, denoted LD, in which MVs for matched pairs corresponding to large plants tend to be farther apart than the MVs for matched pairs corresponding to small plants. We then run a second simulation, denoted SD, in which the opposite is true. For simulation LD, we generate $F_{2M}^{(m)}$ from

$$F_{2M,i,j}^{(m)} = Y_{ij}(1+\kappa_{ij}), \kappa_{ij} \overset{iid}{\sim} \begin{cases} Uniform(0,.4) & : \text{ if } Y_{i2}^{(0)} > 130 \\ Uniform(0,.15) & : \text{ otherwise.} \end{cases} \tag{3.28}$$

Similarly, for simulation SD, we generate $F_{2M}^{(m)}$ from,

$$F_{2M,i,j}^{(m)} = Y_{ij}(1+\kappa_{ij}), \kappa_{ij} \overset{iid}{\sim} \begin{cases} Uniform(0,.4) & : \text{ if } Y_{i2}^{(0)} \leqslant 130 \\ Uniform(0,.15) & : \text{ otherwise.} \end{cases} \tag{3.29}$$

We run 100 replicates of the LD and SD simulation settings.

With these simulations, we are interested in assessing two aspects of LFCMV. First, we examine the match rate for different types of matched pairs for both the SD and LD simulations. The matching results are summarized in Table 3.4. As with the results from Section 3.4, larger distance across matched pairs tends to be associated with lower MRs than smaller distance across matched pairs. However, an examination of the match rate by plant size reveals that the naive approach and LFCMV perform differently in the different size groups. Consider Figure 3.4, displaying the match rates for the LD replicates. For large plants, which have greater

Table 3.4: Average posterior match rate for the SD and LD simulations. Bold values indicate the PMR associated with the plant size with smaller distance between matched pairs in each simulation. Large plants: the PMR for large plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Small plants: the PMR for small plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Overall PMR: the PMR for all plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Values in parentheses represent the standard error.

	SD	LD
Large plants	**.90**(.008)	.78(.01)
Small plants	.76(.009)	**.88**(.003)
Overall PMR	.85(.006)	.84(.006)

distance across matched pairs in this simulation, LFCMV matches more accurately than the naive approach. However, the opposite trend is evident for smaller plants. This suggests that LFCMV overestimates the distance reasonable for matched pairs in the smaller plants, resulting in a lower match rate. Figure 3.5 shows similar tendencies for the SD simulation; LFCMV's match rate is better on average for the smaller plants, with the naive approach matching more accurately for the larger plants with less distance across matched pairs. Despite this behavior, LFCMV results in an overall improvement in match rate over the naive approach in both the SD and LD replicates.

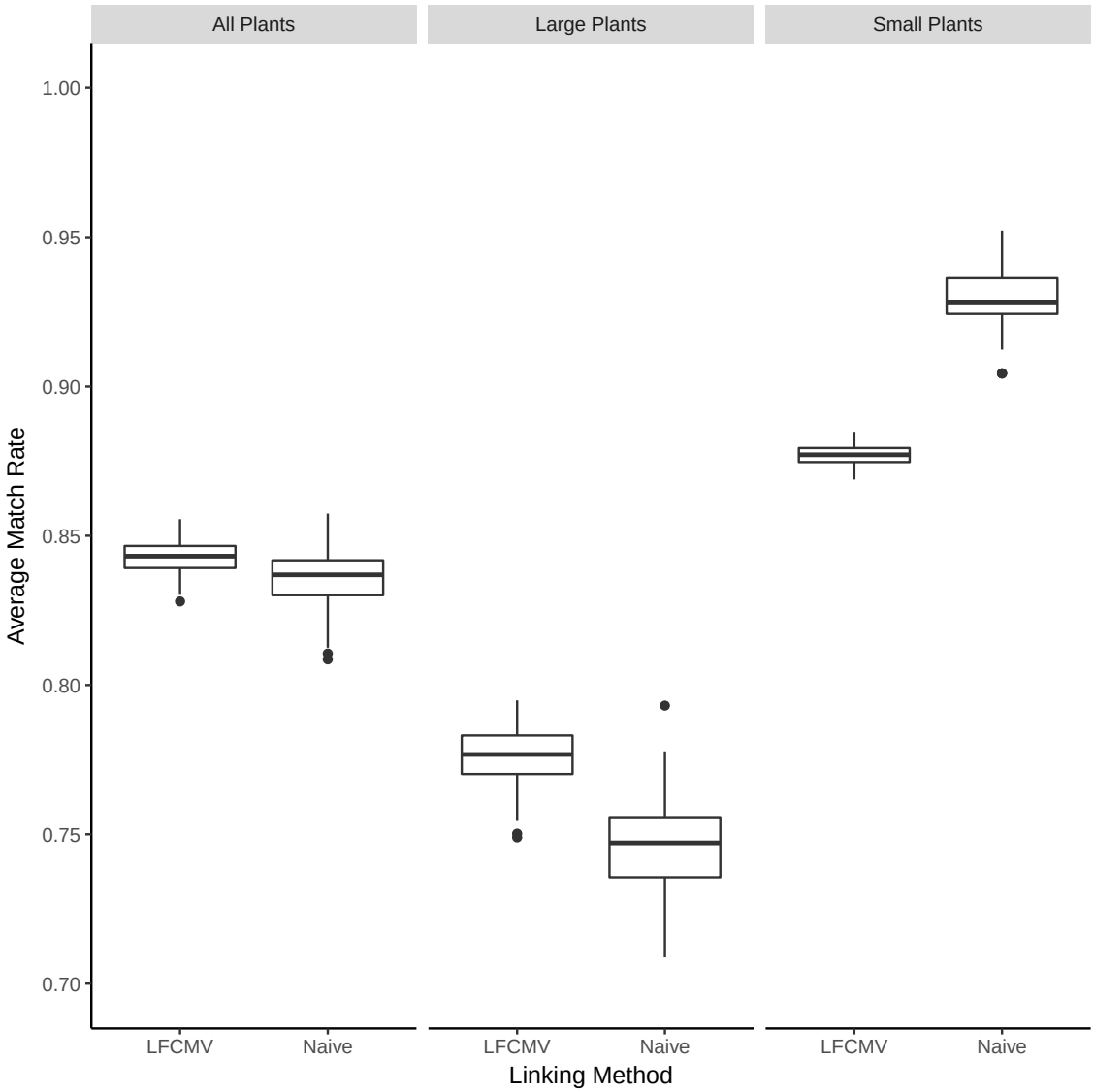


FIGURE 3.4: LD: Match rates under when large plants have greater distance across matched pairs than small plants. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.

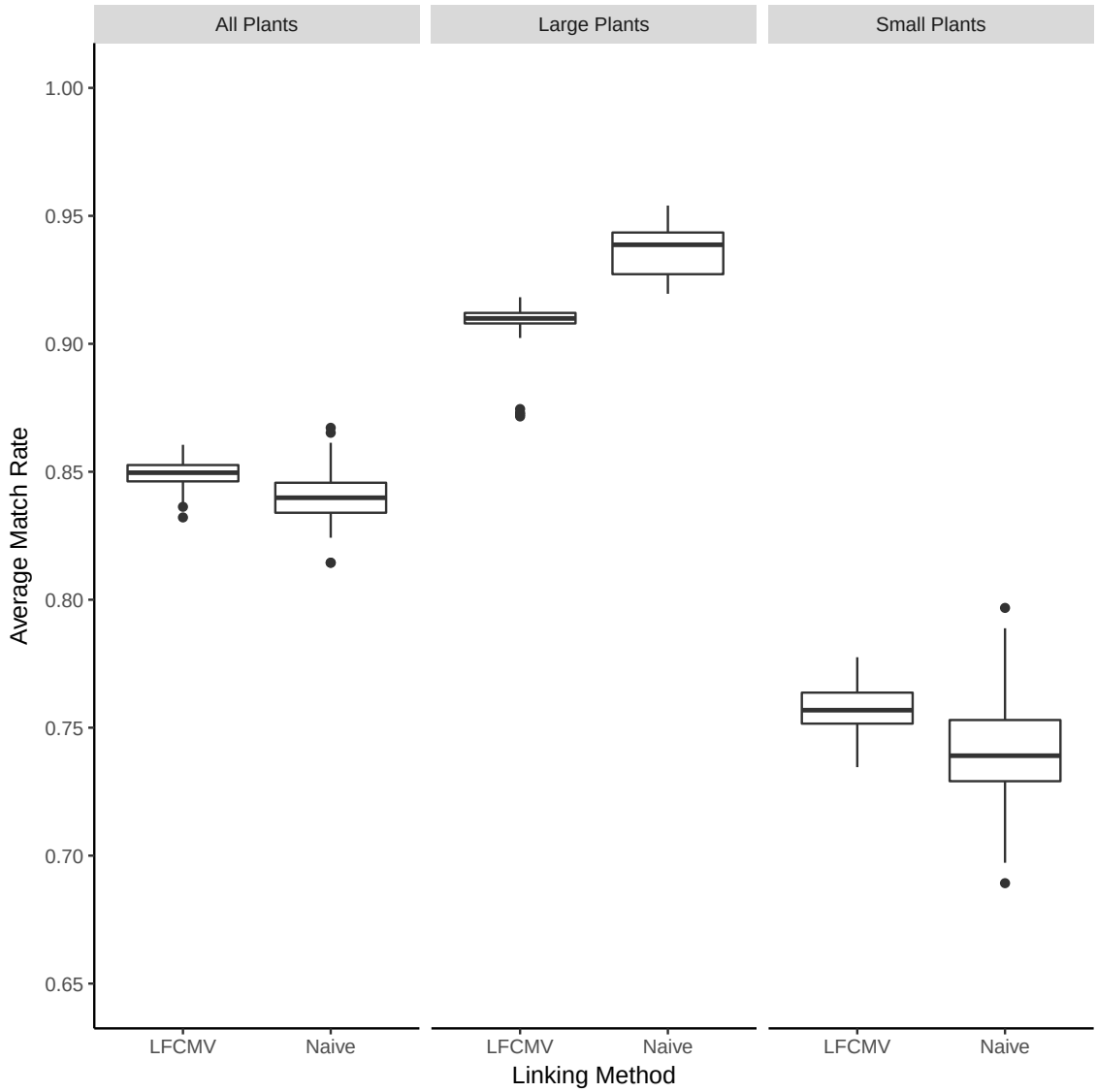
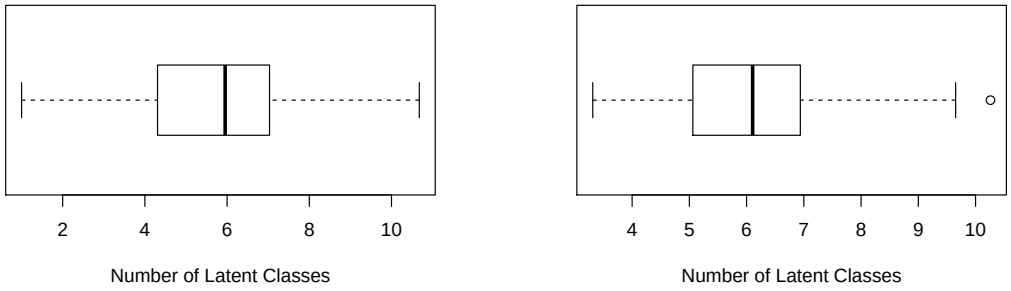


FIGURE 3.5: SD: Match rates under when small plants have larger distance across matched pairs than larger plants. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.



(a) SD (b) LD

FIGURE 3.6: Average number of latent classes for the SD and LD simulations. For each of the 100 replicates of simulation SD (Figure 3.6a) and LD (Figure 3.6b), we compute the number of latent classes at each posterior iteration with more than 5 records. We then average these counts to obtain a single estimate of the average number of occupied latent classes for each replicate. The boxplots are composed of the 100 replicate-specific counts.

The second aspect of LFCMV we illustrate with the SD and LD simulations is the ability of LFCMV to capture the relationships inherent in the different types of matched pairs using the mixture model formulation of the linking model. In the simulation studies described in Section 3.5.1, the distance across matched pairs is uniform, so latent classes are not needed to describe the distribution of distances across matched pairs. Accordingly, the number of occupied mixture components collapses to a single component during the posterior sampling. However, for the SD and LD simulations, there are two clearly defined types of matched pairs, as for each simulation, small and large plants are associated with a certain degree of distance across matches. Accordingly, Figure 3.6 indicates that multiple mixture components are occupied in the SD and LD simulations. In Figure 3.7, we examine these components for a single iteration of a replicate of the SD simulation. The figure indicates that latent class 1 is associated with smaller plants than latent class 2, while latent class 3 contains a blend of plant sizes. As distance across matched pairs is

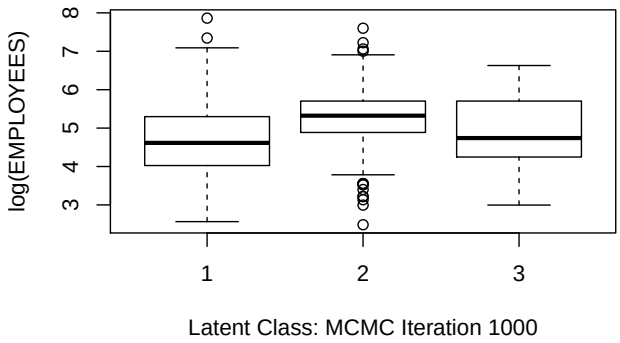


FIGURE 3.7: SD: The box plots display the log(employees) value for the three latent classes at MCMC iteration 1000. Each boxplot represents a latent class, and the y-axis represents the value of $\tilde{\mathbf{X}}_2$, denoted log(employees), for each record in that latent class. A paired t-test comparing the log(employees) value of latent classes 1 and 2 leads statistically significant results ($p < 0.05$). Specifically, latent class 1 tends to contain smaller plants than latent class 2.

directly related to plant size, these results suggest that the latent class structure of the linking model may be useful for capturing features in the data associated with differing amounts of distance across matched pairs.

3.5.3 Summary of findings

The simulations conducted in Sections 3.5.1 and 3.5.2 highlight a few features of LFCMV linking. First, as is evident in Section 3.4.2, large distances among the MVs for matched pairs can lead to a reduction in PMR. This underscores the importance of the considerations discussed in Section 3.2.3. Second, the simulations in Section 3.5.2 provide an example the utility of the latent class structure of the linking model. The latent class structure provides the potential for distribution in distance to vary across matched pairs. If such a structure is unnecessary in an application, the sampler tends to collapse to using single class to estimate $\mathbf{\Sigma}$.

3.6 Data application

In this section, we describe an application of the LFCMV methodology to link records from the 2007/2008 IAC and 2007 CMF. For simplicity of notation, unless otherwise noted we let CMF denote the 2007 CMF records and IAC denote the 2007/2008 IAC records. In Section 3.6.1, we describe the data processing performed prior to linking the data. In Section 3.6.2, we describe the links resulting from applying LFCMV. In Section 3.6.3, we perform a post-linking analysis exploring energy efficiency. We conclude with a discussion of future work in Section 3.6.4.

3.6.1 Data cleaning, blocking and transformation

The purpose of linking in this application is to identify which records in the CMF received IAC energy assessments. However, the IAC and CMF records were collected by two different organizations. Some variables common to both files are recorded using an organization specific scale, variable name, and accuracy level. For instance, in the IAC, NAICS code is recorded as a 6-digit identifier, while in the CMF, up to 9-digits are recorded for each plant. Because of these discrepancies, pre-processing is required to convert the IAC and CMF records into the subsets F_1 and F_2, respectively, used for linking. We use F_1 to denote the subset of IAC records used for linking, while F_2 denotes the subset of CMF records considered as possible matches for F_1.

3.6.1.1 BVs and MVs

To create F_1 and F_2, we begin by examining the BVs used to block the records. As in the simulations in Section 3.4, we use the BV fields of state and NAICS code. In the CMF, the state where an establishment is located is denoted using the standard notation of two upper case letters. A small number of records are missing the variable for state, and we exclude these records from consideration for F_2. In the IAC, however, the BV "state" is recorded with a variety of upper-lower case

combinations. For instance, "AL","Al" and "al" all refer to the state Alabama. We convert the IAC labels to match the double capital format in the CMF, i.e., for plants referring to Alabama, state is recorded as "AL". Once the state variable is standardized, we create a list of all states associated with plants in the IAC. We then exclude from F_2 all CMF records from states not on this list. In addition to records pertaining to the 50 states, the IAC contains 15 records from plants in Puerto Rico. As there are no records corresponding to Puerto Rican establishments in the CMF, these records are also excluded from F_1.

After blocking on state, we proceed to the BV of NAICS code. For each record in the IAC, we create a list of CMF records pertaining to the same 6-digit NAICS code and state as the IAC record. However, such direct blocking leads to a number of IAC records with no possible matches in the CMF. The IAC records are a subset of the CMF, and as such all records in the IAC should have a match in the CMF. The lack of matches for some records after direct blocking suggests that some of the NAICS codes in IAC or CMF may have some uncertainty. To account for this, we considering blocking on three different levels of NAICS agreement. Specifically, we create subsets with agreement on 4-digit, 5-digit and 6-digit NAICS codes. We refer to these as levels of NAICS blocking.

Let $F_1^{(x)}$ denote the subset of IAC records that have a non-empty block $F_{2(i)}^{(x)}$ when blocking is performed on state and x-digit NAICS code. Here, $x = 4, 5, 6$. Similarly, let $F_2^{(x)}$ denote the set of possible matches for $F_1^{(x)}$ when blocking on state and x-digit NAICS. Recall that NAICS codes are nested up to the 4-digit level, meaning that every record in $F_2^{(6)}$ will also be in $F_2^{(5)}$ and $F_2^{(4)}$. By the same logic, $F_2^{(4)}$ is the largest subset of CMF records to be linked, 5-digit blocking restricts this linking to a smaller subset $F_2^{(5)}$, and 6-digit blocking yields the smallest subset $F_2^{(6)}$.

Once records are moved into blocks defined by state and x-digit NAICS, MVs are used to determine which records in $F_{2(i)}^{(x)}$ are a match for each i in $F_1^{(x)}$. The MVs

Table 3.5: Description of the matching variables used to link the data in Section 3.6. The first column gives the name of the variable in IAC database. The second columns lists the name of the associated variable in the CMF, with the abbreviation for that variable in the third column.

IAC Variable	**CM Variable**	**Abbr.**
Sales (in $)	Total Value of Shipments (in thousands of $)	TVS
Number of Employees	Total Employment	TE

are listed in Table 3.5, and we use the abbreviations from the CMF to denote the MVs. The scale of the TVS variable is different in the IAC than it is in the CMF. Specifically, the CMF variable TVS is recorded in thousands of dollars while the IAC variable is recorded in dollars. We divide the IAC value by 1000 to match the scale of the CMF variable.

3.6.1.2 Agreement filters

Blocking with respect to 4-, 5- and 6-digit NAICS all lead to some large blocks. To reduce the block size, we apply what we call an agreement filter to narrow the possible options for each IAC record. Let TVS_{fi} denote the recorded value for TVS in file f for record i, and similarly let TE_{fi} denote the recorded value for employment. For each i in $F_1^{(x)}$ and each $i^{'}$ in $F_{2(i)}^{(x)}$, we compute

$$d_1(i, i^{'}) = \frac{TVS_{1i} - TVS_{2i'}}{TVS_{1i}}, \tag{3.30}$$

and

$$d_2(i, i^{'}) = \frac{TE_{1i} - TE_{2i'}}{TE_{1i}}. \tag{3.31}$$

Here, $d_1(i, i^{'})$ and $d_2(i, i^{'})$ are the percent difference between IAC record i and the observed Census record $i^{'}$ on TVS and TE, respectively. If either (3.30) or (3.31) is greater than .3, we declare i and $i^{'}$ a non-link and remove $i^{'}$ from $F_{2(i)}^{(x)}$. Otherwise, we allow i and $i^{'}$ to be possible links. While this agreement filter does reduce the

Table 3.6: The number of IAC records to be matched at each level of NAICS blocking, i.e., the cardinality of $F_1^{(x)}$ for $x = 4, 5, 6$. There are 786 original IAC records from 2007/2008, excluding records for plants in Puerto Rico.

NAICS Level	4-digit	5-digit	6-digit
IAC Records	550	480	330

block sizes, it also results in some IAC records which have no possible matches in the CMF. We drop these IAC records from F_1. Table 3.6 shows the resulting size of F_1 at each level of NAICS blocking.

The agreement filter chosen here reflects a belief that for true matches, the values of TVS and TE in the IAC and CMF should be similar. The threshold of 30% allows some room for noise, recording error and time variation. In general applications, the choice to apply an agreement filter, as well as the choice of appropriate threshold, is at the discretion of the analyst.

3.6.1.3 Transformations and standardization

After applying the agreement filter described in Section 3.6.1.2, we have our final subsets $F_1^{(x)}$ and $F_2^{(x)}$ for each x. Before linking, we transform the MVs by taking the log. Taking the log puts the data on a similar scale, meaning that the mixture model should require fewer components to describe the distribution of the data. In a final pre-processing step, we standardize the MVs, as necessitated by the linking model in Section 3.2.3. The standardization is performed as described in Section 3.4.1. Let $j = 1$ refer to TVS and $j = 2$ refer to TE. Let $Y_{ij}^{(x)}$ denote the standardized value for MV j for record i in $F_1^{(x)}$. Similarly, let $X_{i'j}^{(x)}$ denote the standardized value for MV j for record i' in $F_2^{(x)}$.

3.6.2 Evaluating the links

We run LFCMV to link each pair of data sets $(F_1^{(x)}, F_2^{(x)})$ for $x = 4, 5, 6$. As described in Section 3.4, for $x = 4, 5, 6$, we initialize C by selecting, for each record i in $F_1^{(x)}$, the record $i^{'}$ in $F_{2(i)}^{(x)}$ such that

$$d(i, i^{'}|x) = \sqrt{\sum_{j=1}^{2} \left(\frac{Y_{ij}^{(x)} - X_{i^{'}j}^{(x)}}{Y_{ij}^{(x)}} \right)}, \tag{3.32}$$

is minimized. We then fit the LFCMV model using the posterior sampling steps described in Section 3.3.2.

Applying LFCMV results in 5000 post burn-in posterior estimates of C at each level of NAICS blocking. We use these estimates to compute posterior probabilities that each record $i^{'} \in F_{2(i)}$ is a match to record i in $F_1^{(x)}$ by computing the proportion of posterior draws such that each $i^{'}$ in $F_{2(i)}^{(x)}$ is linked to its corresponding record i. For each i, we look at the posterior probabilities associated with possible matches $i^{'}$ in $F_{2(i)}^{(x)}$. If at least one of the potential matches in $F_{2(i)}^{(x)}$ has a posterior probability $> .5$, we say i has a high probability match. Otherwise, we say i has low probability matches. Table 3.7 shows a summary of the counts of IAC records having high and low probability matches at each NAICS level. Counts have been rounded to satisfy disclosure protocols.

As we shift from one level of NAICS blocking to another, the subset of records $F_1^{(x)}$ and $F_2^{(x)}$ result in different possible links. For instance, a record in $F_1^{(4)}$ that has a possible match in $F_2^{(4)}$ may not be in $F_1^{(5)}$, i.e., may have any possible matches in the CMF when blocking on 5-digit NAICS. This can be seen in the difference in Tthe total column in Table 3.7. For instance, when shifting from 4-digit to 5-digit blocking, there are 70 records that, according to our choice of blocking and agreement filter, no longer have any possible matches in the CMF. Table 3.8 displays

Table 3.7: Displays the types of links obtained after linking at each level of NAICS blocking. We define a link as high probability if the posterior probability of linking $(i, i^{'})$ is greater than .5.

	Link Type		
NAICS Level	High Prob	Low Prob	Total
4-digit	320	230	550
5-digit	330	150	480
6-digit	270	60	330

Table 3.8: Displays the types of links from IAC records that cannot be matched as we move from one level of NAICS blocking to a more restrictive blocking level. We define a link as high probability if the posterior probability of linking $(i, i^{'})$ is greater than .5.

	Link Type		
NAICS Change	High	Low	Total
4 to 5	60	10	70
5 to 6	110	40	150
4 to 6	150	70	220

the number of IAC records which are lost, i.e., no longer have possible matches in the CMF, as the blocking level changes. Table 3.8 also displays rounded counts for the number of these lost records that correspond to high probability matches.

While some record links are lost as we compare results across different levels of NAICS blocking, some record links remain consistent. We use the posterior probabilities associated with each CMF record to determine which record in $F_2^{(x)}$ is the most likely match to each i in $F_1^{(x)}$. In other words, we determine which record in $F_2^{(x)}$ is the *highest* probability match. We note that these matches are not necessarily *high* probability matches. For some IAC records, the highest probability match remains consistent across different levels of NAICS blocking. Table 3.9 indicates the counts of such records at 4-digit to 5-digit NAICS blocking, as well as 5-digit to 6-digit blocking.

Table 3.9: At each NAICS blocking level, for each i in $F_1^{(x)}$ there is a highest-probability match in $F_2^{(x)}$. The counts above indicate how many of these highest-probability matches stay the same across x-digit NAICS levels.

NAICS Level	4 to 5	5 to 6
Count	350	250

At each level of NAICS blocking, there are also some singleton IAC records, or records in $F_1^{(x)}$ that have exactly one option in $F_{2(i)}^{(x)}$. Similar to the seeds discussed in Section 2.2.6 in the context of the BLASE model, singleton pairs are linked independently of the process of estimating C; the pairs are declared links at every iteration. These singletons therefore inform the process of estimating $\mathbf{\Sigma}$, providing iteration-consistent links from which to estimate distance across matched pairs. Singletons are dependent both on x and the choice of agreement filter threshold. Because we have applied the agreement filter described in Section 3.6.1.2, all singletons are to be within 30% of the F_1 value for TVS and TE. Ideally, the relationship in the linked singleton pairs will be reflective of relationships present in the rest of the data. However, the latent class structure of LFCMV means that if the relationship in the singletons is different than the relationship across pairs in other blocks, the model can assign these records to their own latent class (or classes) with a reflective covariance matrix. The number of singletons varies across x-digits NAICS blocking. To satisfy disclosure protocols, we do not provide counts of singletons at any level of NAICS blocking.

3.6.3 Post-linking analysis

The links estimated in Section 3.6.2 can be used to construct identifiers for the records in the CMF, specifying which records received an IAC energy assessment. In this section, we examine energy efficiency in plants that received IAC energy

assessments versus those that did not. In addition to examining the data to see if there are year-specific trends, we are also interested in examining that trend across time. As the CMF is conducted every 5 years, we examine energy efficiency trends in 2007 and 2012. We do not make a causal argument about the relationship between receiving an assessment and energy efficiency, as many factors, including plant size, may impact the relationships observed in the data.

As a basis for our analysis, we begin with all CMF records from 2007 and 2012 that have a 4-digit NAICS code observed in the IAC. This limits our analysis to industries that were assessed in the time period of interest. We further limit our analysis sample to plants meeting IAC standards for assessment. IAC assessments are typically performed for plants with no more than 500 employees and with a gross annual sales of no more than $100 million (Muller, 2001, pg. 13). Let A_{2007} denote the set of CMF records from 2007 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. Similarly, let A_{2012} denote the set of CMF records from 2012 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. We refer to A_{2007} and A_{2012} as our analysis samples.

3.6.3.1 Energy metric

For each record $i^{'}$ in A_{2007} and A_{2012}, we compute an energy efficiency metric as follows:

$$EF_{i'} = log\left(\frac{TVS_{i'}}{CF_{i'} + EE_{i'}}\right). \tag{3.33}$$

Here, (3.33) represents the log ratio of the total value of shipments divided by the total cost of energy. The denominator represents total energy cost and is computed by adding the cost of fuels (CF) to the electricity expenditure (EE).

Energy efficiency is known to vary across different industries, or, equivalently,

across NAICS codes. Producing aluminum foil, for instance, has different energy requirements from producing potato chips. This suggests that rather than directly comparing values of (3.33) across plants, an analysis might be conducted at the NAICS specific level. However, there are not enough IAC-assessed plants within most NAICS codes to make such an analysis feasible. Instead, we adjust the metric in (3.33) so that it can be compared across all NAICS codes. To do this, we convert the efficiency metric in (3.33) into a standardized energy score as follows.

For each record $i^{'}$ in A_{2007}, we sort the records into groups defined by 6-digit NAICS codes. Within each group, we compute the mean and standard deviation of values of (3.33) for CMF records in the group. Let $\overline{EF}_g$ denote the group g specific mean, and similarly let $SD(EF)_g$ be the group specific standard deviation. For $EF_{i'}$ in group g let

$$\widetilde{EF}_{i'} = \frac{EF_{i'} - \overline{EF}_g}{SD(EF)_g}, \tag{3.34}$$

be the standardized value of $EF_{i'}$. We refer to the value of (3.34) as the energy score. The energy scores are on a common scale, describing how many standard deviation each plant's value of $EF_{i'}$ is from the group specific mean. This allows each corporation to receive an energy score relative to their peers as defined by 6-digit NAICS. For records in A_{2012}, we create energy scores for 2012 analogously.

3.6.3.2 Results

The records in A_{2007} and A_{2012} are divided into two categories. The first category, labeled "With Assessments", refers to plants in a given year that received an IAC assessment. The remaining plants are categorized as "Without Assessments", meaning that these plants did not receive an IAC assessment in 2007 or 2008. The process of assigning records in A_{2007} and A_{2012} to these categories is dependent upon the linking structures estimated in Section 3.6.2. For each x-digit NAICS blocking level,

and for each posterior sample $s = 1, \ldots, M = 5000$ of C, we create a binary vector $I^{(s)}$ of length a_{2007}, where a_{2007} denotes the number of records in A_{2007}. Here, a 1 in position $i^{'}$ means that record $i^{'}$ in A_{2007} is linked to an IAC record at iteration s; such plants are categorized as "With Assessment" records for iteration s. All other plants are categorized as "Without Assessment" records.

In order to compare the energy scores in 2007 and 2012, we also consider records in A_{2012}. Linking is performed on the 2007 records, meaning that the indicators for IAC assessments are associated with a 2007 CMF record. However, the Census Bureau uses a longitudinal identifier called SURVU_ ID that can be used to identify records for a given plant across time. For each iteration, we identify the set of SURVU_ ID numbers that, conditional on each $I^{(s)}$, received an energy assessment in 2007. The 2012 records corresponding to these unique identifiers compose the 2012 "With Assessment" samples.

Data cleaning is necessary before using SURVU_ ID as a longitudinal identifier. Exploration of the data revealed that SURVU_ ID is not unique within either the 2007 nor the 2012 CMF. This means multiple records with the same SURVU_ ID identifier exist within each year-specific data set. However, these records appear to be true duplicates, meaning every field value is the same. We delete all but one record with each SURVU_ ID to create a unique identifier.

For each iteration s and each level of NAICS blocking, we compute the density of energy scores for "With Assessment" plants, and separately for "Without Assessment" plants, for 2007 and 2012. We use fixed bins across the iterations. Within each year, and each of the two categories, we then average the density in each bin at each iteration and plot the results. The resultant average density curves for 2007 are displayed in Figures 3.8, 3.10 and 3.12, and the resultant averaged density curves for 2012 are displayed in Figures 3.9, 3.11 and 3.13.

In Figures 3.8 through 3.13, solid lines are used for the "Without Assessment"

Table 3.10: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment" group at each level of NAICS blocking. Similar trends are evident across all three levels of NAICS blocking.

NAICS4		Mean	95% Interval
2007	With Assessments	-0.18	(-.27,-.09)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.03	(-0.04,0.11)
	Without Assessments	0.008	(0.006,0.011)
NAICS5		Mean	95% Interval
2007	With Assessments	-0.17	(-0.26,-0.08)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.02	(-0.06,0.10)
	Without Assessments	0.009	(0.006,0.011)
NAICS6		Mean	95% Interval
2007	With Assessments	-0.18	(-0.29,-0.08)
	Without Assessments	-0.007	(-0.009,0.005)
2012	With Assessments	0.04	(-0.05,.12)
	Without Assessments	0.009	(0.06,.011)

group and dashed lines are used for the "With Assessment" group. Each curve has a plotted 95% interval, representing the 95% interval for the density within each bin. The trends in the density curves are very similar across the three levels of NAICS blocking. In 2007, the density curve for the "With Assessment" group appears to be shifted further left than the "Without Assessment" curve. This suggests that the energy scores for assessed plants seem smaller than plants without assessments. In other words, relative to their peer group, plants receiving IAC audits appear less energy efficient. This trend appears to vanish in the plots corresponding to 2012. In Figures 3.9, 3.11 and 3.13, there is essentially no difference in the curves between the "Without" and "With Assessment" groups. We stress that we do not make a causal argument about the relationship between plants with and without assessments.

To supplement the graphical exploration of the data, we leverage the combining rules of Rubin (1987) to perform inference across the M estimated linkage structures

at each level of NAICS blocking. Using the "With" and "Without Assessment" groups described above, we compute the average energy score for each of the two groups for each of the M iterations. Using the combining rules of Rubin (1987), we compute 95% confidence intervals for the mean of each group. Results are displayed in Table 3.10 by NAICS blocking level. The intervals in the table suggest the same general trends observed in the density curves. In 2007, the average energy score tends to lower for the "With Assessment" groups, with non-overlapping intervals for the "With" and "Without" groups. This difference is much less pronounced in 2012, and the intervals overlap, suggesting no statistically significant difference in the means between the groups.

3.6.4 Future work

The data application in Section 3.6 is not causal. We compare the distribution of energy scores of plants receiving assessments versus those that do not. However, there are many fewer plants in the "With Assessment" group than the "Without Assessment" group. In order to better explore the relationship between IAC assessments and energy efficiency, it might be useful to reduce the size of the "Without Assessment" comparison group to approximate a sort of matched pairs analysis. Another direction is to fit a model assessing energy score on assessment group (with or without) as well as other features in the CMF data. These are directions for future work, all of which are enabled by the creation of the linked data product.

Additionally, as noted early in this chapter, the CMF contains imputations created by the U.S. Census Bureau to replace missing data and correct for faulty values. In Chapter 4, we implement an alternative imputation strategy that replaces the single Census Bureau imputation with multiple imputations. We then perform the analysis in Section 3.6 and compare the results.

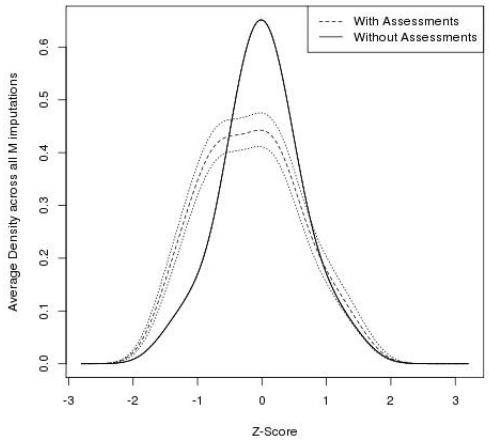


FIGURE 3.8: NAICS 4 Blocking, 2007: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

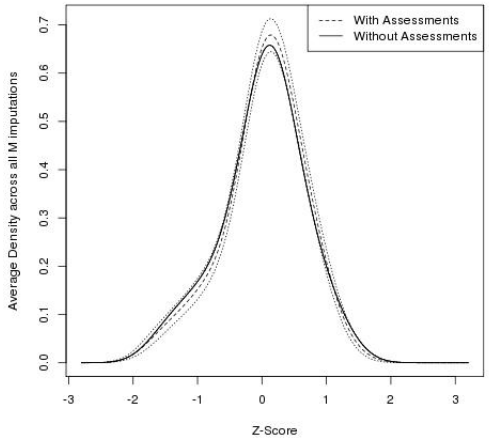


FIGURE 3.9: NAICS 4 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

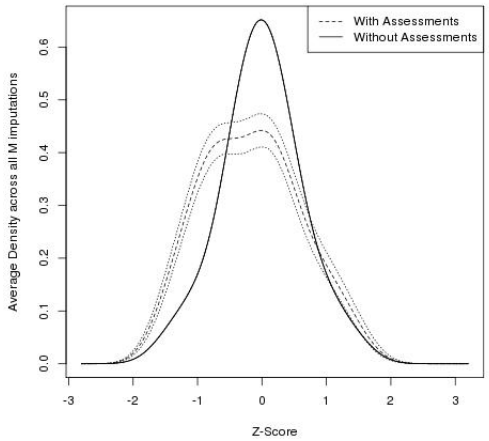


FIGURE 3.10: NAICS 5 Blocking, 2007 Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

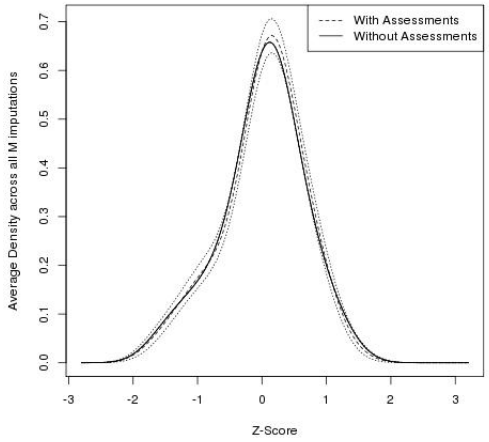


FIGURE 3.11: NAICS 5 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

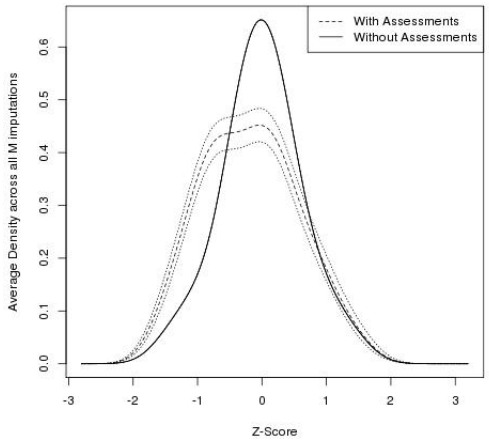


FIGURE 3.12: NAICS 6 Blocking, 2007: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

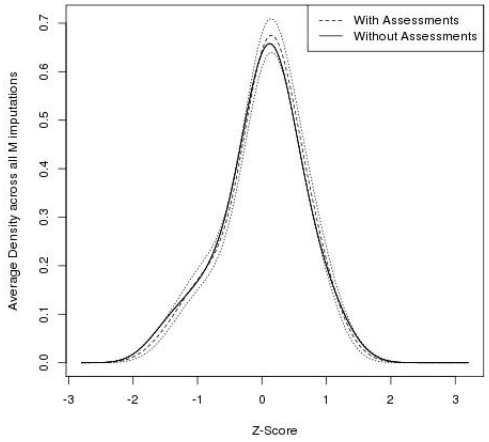


FIGURE 3.13: NAICS 6 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

4

Multiple imputation and file linking

The application discussed in Chapter 3 involves the U.S. Census of Manufactures (CMF). The results of both the linking and the analysis for these data are conditional upon the values contained in the CMF. However, some of these values are imputed. As neither the true relationship between the variables in the IAC and CMF, nor the distribution of the actual values in the CMF, is known, it is difficult to assess the impact of the CMF imputations on the resultant analysis. However, understanding this impact is critical to assessing the reliability of results obtained from the linked data sets. In this chapter, we use simulated data to examine the impact of imputations in a file linking scenario. We introduce missing data, impute the missing values, and compare the accuracy of estimating C on the imputed versus the fully observed data sets. We also consider the process of performing inference and discuss a Bayesian technique for posterior estimation on multiply-imputed linked data sets. We apply this technique to the simulated data and examine the effect of the imputations upon posterior estimates. The results suggest that while the performance of both linking and estimation can be negatively impacted by the presence of imputations, general trends in the data tend to be preserved. Within the context of

these findings, we revisit the analysis from Section 3.6. We replace the CMF imputations with multiple imputations and compare the results of the resultant analysis with those from Section 3.6.

4.1 Introduction

The Census Bureau examines collected CMF data for missing values and applies a series of checks, called edit rules, to identify reported values that may be faulty (White, 2014; Kim et al., 2015). For values that fail these rules, as well as for missing values, the Census Bureau uses a variety of techniques to replace or fill-in the data with imputed values (White et al., 2015). Some records are imputed using administrative records or information for a plant from a previous year. However, as discussed in White et al. (2015), some records are imputed using either "industry-average or univariate regressions using only current-year reported data" (pg. 2). These techniques can be problematic for analysts seeking to use the completed data sets for regression and other typcs of analysis. As discussed in both Schafer and Graham (2002) and Little and Rubin (2002), such techniques can result in distorted covariance and correlation estimates and underestimated standard errors. For the CMF specifically, White et al. (2015) illustrate that "functions of key variables in the completed CM[F] data show evidence of attenuation and under-estimation of variability" (pg. 2).

In addition to the potential impact on analysis, in a file linking scenario there are other potential implications of treating the CMF imputations as true. The linking technique in Chapter 3 relies on values of two MVs, total value of shipments (TVS) and total employment (TE). Values of these variables are used to determine which CMF record is matched to each IAC record. If the CMF imputed values for TE or TVS are not reflective of the underlying distributions in the data, the linking may result in inaccurate matches. For the CMF data, we examine this possibility by

blocking on state and 6-digit NAICS code. Such blocking produces some IAC plants that have exactly one possible match in the CMF. Looking at these matches, we see that, for some pairs, the values reported for variables relating to employment and sales are very different across the files. As noted in Chapter 3, we do not expect the IAC and CMF records to exactly agree on the field values, but there should be a similarity evident across matched pairs.

Because the exact relationship between the MVs in the IAC and CMF is unknown, it is difficult to assess the impact of the CMF imputations upon the estimates of the linking structure obtained in Chapter 3. However, the extreme differences displayed across several pairs, along with the work in White et al. (2015), suggests replacing the imputations in the CMF. One option is to replace each of the CMF imputations with multiple imputations to better reflect the variability. In this chapter, we illustrate the impact of multiple imputations upon the performance of LFCMV using simulated data. For these data, the relationship between F_1 and F_2 is known, allowing us to compare the quality of links estimated using imputed values versus those estimated on the fully observed data. The imputations in the simulation study are created by drawing from appropriate conditional distributions, meaning that within F_2, marginal and conditional relationships among related variables are preserved. However, the results of the study suggest that the relationship between F_1 and F_2 is not necessarily preserved by the imputations, and this leads to a reduction in accuracy in estimating C.

In addition to examining estimates of C, we are also interested in the effect of imputations on posterior analysis conducted on linked data sets. The concept of combining estimates obtained from multiply imputed data sets is well explored (e.g., Rubin, 1976, 1987; Little and Rubin, 2002; Reiter and Raghunathan, 2007; Zhou and Reiter, 2010). However, the process becomes more complex in a file linking scenario. Suppose F_1 contains no missing data, but F_2 has been completed using multiple

imputations. In such an instance, the file linking process involves linking each of the M completed versions of F_2 to F_1. We refer to this process as multiply imputed data linkage (MIDL). For a Bayesian linking technique like LFCMV, MIDL results in S estimates of C for each completed data set $m = 1, \ldots, M$. An analyst looking to perform inference after linking will therefore have $M \times S$ linked data sets to use. Furthermore, as each of the posterior samples for C is conditional on the completed data set used for linking, some of the $M \times S$ completed sets will be correlated.

In this chapter, we review a Monte Carlo technique of Gelman et al. (2004) and adapt this framework to enable posterior estimation in an MIDL setting. Specifically, we frame MIDL as a two-stage imputation scenario and discuss the inference technique within this framework. The first stage of imputation in MIDL is the completion of F_2 through the creation of M completed data sets. The second stage of imputation is estimating C conditional on the completed data sets. Using this structure and our simulated data, we examine the impact of imputations upon posterior estimates obtained on the imputed versus fully observed data.

The remainder of this chapter proceeds as follows. In Section 4.2, we introduce notation and frame MIDL as a two-stage imputation scenario. In Section 4.3, we review the inference technique of Gelman et al. (2004) and discuss how the technique can be applied to MIDL scenarios. In Section 4.4, we use a simulation study to illustrate the impact of imputations upon both the estimation of the linkage structure in LFCMV as well as on posterior estimates obtained on the linked data. Finally, in Section 4.5, we return to the application described in Chapter 3. We create multiple imputations for the 2007 CMF data and describe the process of linking these imputations to the IAC. We then conduct an examination of energy efficiency and compare the results with these imputed data sets to the results obtained in Chapter 3. In Section 4.5.4, we conclude with a discussion of future work.

4.2 Two-stage imputation

Using our standard notation, let F_1 and F_2 denote the two data sets to be linked. Without loss of generality, assume that F_2 contains missing information in some of the MVs, but F_1 is fully observed. This is motivated by the application in which the CMF (F_2) contains missing and imputed values but the IAC (F_1) is fully observed. Let $F_2 = (F_{2obs}, F_{2mis})$, where F_{2obs} contains all fully observed variables in F_2 and F_{2mis} contains all variables with some missing information. Assume F_{2mis} has been completed via multiple imputation, yielding M completed data sets. Let $\mathcal{D}^{(m)}$ represent each of these completed data sets, where $m = 1, \ldots, M$. An analyst then uses some application specific file linking technique to combine each pair of data sets $(F_1, \mathcal{D}^{(m)})$. For Bayesian file linking techniques like LFCMV, each linking process results in S posterior draws for C. This means that, in total, MIDL results in $M \times S$ linked, completed data sets.

The process of creating such linked, completed data sets is an example of two-stage, or nested, imputation. Nested multiple imputation (NMI) refers to the process of generating completed data sets by partitioning the missing data into two subsets (Shen, 2000; Rubin, 2003; Harel and Schafer, 2003). Imputation is then performed in two stages. First, the first subset of missing data is imputed. Each of the resultant first-stage imputed data sets is complete with respect to the first subset of variables. Every first-subset complete data set is called a nest. Within each nest, a second round of imputations is conducted to fill in missing information for the second subset of missing values. This results in fully completed data sets.

In the work of Shen (2000), these two imputation rounds are composed of "expensive" and "inexpensive" imputations, respectively. "Expensive" variables require several days to impute, while "inexpensive" variables can be generated much more efficiently. The first stage of imputation is then to create a few, say M_1, imputations

for each of the expensive variables. Then, conditional on each of the M_1 imputations, M_2 imputations are generated for each of the "inexpensive" variables. In another perspective on two-stage imputation, Harel and Schafer (2003) discuss NMI in the context of imputing different types of data at each of the two stages. The authors also consider applications in which each stage of imputation can reflect potentially different missingness mechanisms. Reiter (2004) and Kohnen and Reiter (2009) use nested multiple imputation in the context of synthetic data sharing.

MIDL can also be framed as an NMI scenario. There are two components that require imputation. First, we impute the missing information in F_{2mis} to create M completed data sets $\mathcal{D}^{(m)}$, $m = 1, \ldots, M$, using some multiple imputation technique; this technique is application specific. In our application with the CMF, creating the completed data sets requires several days, making this the "expensive" component of the process. The first imputation stage results in M pairs of data sets $(F_1, \mathcal{D}^{(m)})$. Because we are interested in linking F_1 and $\mathcal{D}^{(m)}$, the information that will be gained by linking the files is considered missing. Conditional on each $\mathcal{D}^{(m)}$, i.e., within each nest, we impute this missing information by generating imputations of the linkage structure, i.e., estimating C. In essence, we impute the missing information for the linked variables conditional on the estimated C. Let $C^{(m,s)}$ denote the s^{th} estimate of C for $(F_1, \mathcal{D}^{(m)})$. Here, $s = 1, \ldots, S$, where the number S of estimates for C is assumed to be the same for each nest. Let $\mathcal{C}^{(m)} = (C^{(m,1)}, \ldots, C^{(m,S)})$ denote the set of estimates of C obtained for $(F_1, \mathcal{D}^{(m)})$.

4.3 MIDL and MCMC

In this section, we use the context of two-stage imputation to discuss a technique for performing inference after MIDL. This technique is based on an approach of Gelman et al. (2004) for performing Bayesian inference after multiple imputation. As illustrated in simulation studies by Zhou and Reiter (2010), with an appropriate

choice of M the technique of Gelman et al. (2004) performs well in MI settings even when the normality assumption required for the combining rules of Rubin (1987) is not satisfied. In Section 4.3.1, we review the approach suggested in Gelman et al. (2004). In Section 4.3.2, we re-frame the approach to suit the MIDL setting.

4.3.1 Bayesian inference after MI

Let $D = (D_{obs}, D_{mis})$ be a data set of interest. Here, D_{obs} contains all fully observed variables in D and D_{mis} contains all variables with some missing information. Let Q be some parameter of interest whose estimation relies on D. Let α be a desired quantile of the posterior distribution for Q, where $0 < \alpha < 1$. We can express α as

$$\alpha = \int_{-\infty}^{Q_\alpha} f(Q|D)dQ = \int_{-\infty}^{Q_\alpha} \int_{D_{mis}} f(Q|D_{obs}, D_{mis}) f(D_{mis}|D_{obs}) dD_{mis} dQ, \qquad (4.1)$$

for some value Q_α of Q. Though simple in form, depending on the distributions involved, it can be difficult to compute the integral in (4.1). Because of this, analysts often use Monte Carlo methods to approximate the integral. Such an approximation involves two steps. First, draw a value of D_{mis} from its posterior distribution $f(D_{mis}|D_{obs})$. Second, conditional on the draw of $D^{(mis)}$, draw a value of Q from its posterior distribution $f(Q|D_{obs}, D_{mis})$. Repeating these two steps for a large number of iterations yields an approximation to (4.1).

This Monte Carlo approach also facilitates the estimation of Q_α. Let K be the number iterations of the two approximation steps, where K is large. The approximation process then yields K estimates of Q. To estimate Q_α, sort the K draws of Q and select the αK^{th} element. This provides an estimate of Q_α.

In addition to a large number of draws for Q, the approximation also produces K draws of $D^{(mis)}$. In an MI application, sampling $D^{(mis)}$ corresponds to the process of creating a completed data set. This imputation process can be computationally expensive. Indeed, in a typical MI application, the number M of completed data sets

that are created is generally small, such as $M = 10$. To avoid the need to generate K completed data sets, Gelman et al. (2004) suggest sampling more than one value from $f(Q|D_{obs}, D_{mis})$ for each draw of $f(D_{mis}|D_{obs})$. Suppose we have M completed data sets $D^{(m)} = (D_{obs}, D_{mis}^{(m)})$, $m = 1, \ldots, M$. Following Zhou and Reiter (2010), the integral in (4.1) can be expressed as

$$\alpha = \lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} \int_{-\infty}^{Q_\alpha} f(Q|D_{obs}, D_{mis}^{(m)}) dQ. \tag{4.2}$$

For any Q_α, the integral can be approximated as follows.

1. Conditional on each $D^{(m)}$, sample J values of Q, where J is large.
2. For each $D^{(m)}$, calculate the percentage of these J draws that are less than Q_α.
3. Average these percentages across all M completed data sets, where $M \to \infty$.

The approximation to (4.2) relies on an infinte M; in practical applications, M will be finite. Accordingly, Gelman et al. (2004) suggest the following process to create an approximation $\widetilde{Q_\alpha}$ of Q_α.

1. Conditional on each $D^{(m)}$, $m = 1, \ldots, M$, sample J values of Q from $f(Q|D^{(m)})$ where J is large. Let $\hat{f}(Q^{(m)})$ denote the set of these J draws.
2. Mix $\hat{f}(Q^{(m)})$ across all m. Let $\hat{f}(Q^{(all)})$ be the resultant set of these $M \times J$ draws.
3. Sort $\hat{f}(Q^{(all)})$ and select the $\alpha(MJ)^{th}$ element as $\widetilde{Q_\alpha}$.

The properties of this estimator are explored in Zhou and Reiter (2010). The authors use simulation studies to illustrate that the estimation procedure works well when M is large, but caution that for small choices of M, estimates of the posterior may be unreliable. Accordingly, in our simulations and application we produce at least $M = 20$ completed data sets.

4.3.2 Bayesian inference after MIDL

The technique discussed in Section 4.3.1 can be applied to an MIDL scenario. Let D be the data set resulting from linking F_1 and F_2. This means that D depends on F_1, F_2 and C. The integral in (4.1) can then be written as

$$\alpha = \int_{-\infty}^{Q_\alpha} f(Q|D)dQ = \int_{-\infty}^{Q_\alpha} \int_C f(Q|F_1, F_2, C)f(C)dCdQ. \tag{4.3}$$

The above expression is appropriate for a typical file linking scenario with fully observed F_1 and F_2. To extend this to a MIDL scenario, we assume that F_2 contains missing data. In this case, we have,

$$\begin{aligned}\alpha = \int_{-\infty}^{Q_\alpha} \int_C \int_{F_{2mis}} & f(Q|F_1, F_{2obs}, F_{2mis}, C)f(C|F_1, F_{2obs}, F_{2mis}) \\ & f(F_{2mis}|F_{2obs})dF_{2mis}dCdQ.\end{aligned} \tag{4.4}$$

The extension from (4.3) to (4.4) reflects the two-stage imputation; there is missing data in the form of both F_{2mis} and C. Using the notation defined in Section 4.2, we can equivalently write

$$\alpha = \lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} \left(\lim_{m\to\infty} \frac{1}{m} \sum_{m=1}^{M} \int_{-\infty}^{Q_\alpha} f(Q|F_1, C^{(m,s)}, \mathcal{D}^{(m)})dQ \right). \tag{4.5}$$

Here, (4.5) suggests integrating over all missing data F_{2mis} and all possible permutations C. In practice, this is often extremely difficult to perform analytically. Instead, the framework of Gelman et al. (2004) can be extended to approximate (4.5).

Suppose we have M completed data sets $\mathcal{D}^{(m)}$, $m = 1, \ldots, M$, for F_2. For each pair $(F_1, \mathcal{D}^{(m)})$, suppose we also have S draws $C^{(m,s)}$, $s = 1, \ldots, S$. Following the framework of Gelman et al. (2004), we approximate (4.5) as follows:

1. Conditional on each $\mathcal{D}^{(m)}$, $m = 1, \ldots, M$, and each $C^{(m,s)}$, $s = 1, \ldots, S$, sample a value of Q from $f(Q|F_1, C^{(m,s)}, \mathcal{D}^{(m)})$. Let $\hat{f}(Q^{(m)})$ denote the set of the S draws for each m.

2. Mix $\hat{f}(Q^{(m)})$ across all M completed data sets. Let $\hat{f}(Q^{(all)})$ be the resultant set of these $M \times S$ estimates.

3. Sort $\hat{f}(Q^{(all)})$ and select the $\alpha(MS)^{th}$ element as the estimate of Q_α.

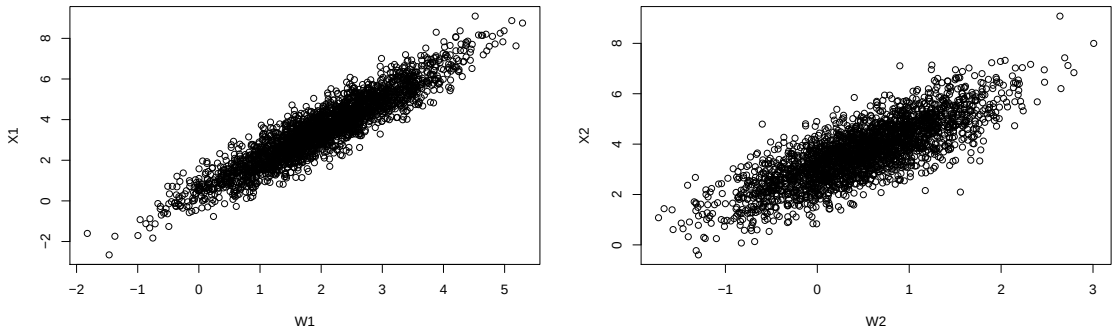
With appropriate choices of prior distributions, this approximation has the advantage of being computationally simple to implement. For analysts familiar with Bayesian approaches, this is a straight forward Monte Carlo approximation that can be implemented after MIDL.

4.4 Simulation: MIDL and MCMC

In the application with the CMF, we do not know the true model for the relationship between the MVs in F_1 and F_2, nor do we know the mechanism that introduced the missing data in F_2. In this section, we conduct a simulation study to examine the performance of MIDL in a more controlled scenario. Specifically, we impute the first-stage imputations from the correct conditional distributions, and generate the variables such that there is a small distance between the MVs in F_1 and F_2. The distance component of the LFCMV linking model penalizes large distances between the MVs, so generating the MVs with a small distance facilitates the accurate estimation of C. In Section 4.4.1 we describe the process of generating the data. In Section 4.4.2 we discuss the introduction and imputation of missing data in F_2. Results are discussed in Section 4.4.3.

4.4.1 Data generation

For this simulation, we generate a pair of data sets F_1 and F_2, where F_2 contains four continuous variables (W_1, W_2, X_1, X_2) and F_1 contains a pair of continuous variables $\mathbf{Y} = (Y_1, Y_2)$. Let Y_1 and Y_2 be the MVs in F_1, and X_1 and X_2 be the corresponding MVs in F_2. Let $\mathbf{X} = (X_1, X_2)$.



(a) X_1 versus W_1 (b) X_2 versus W_2

FIGURE 4.1: Visualizing the relationship between X_1 and W_1 (Figure 4.1a) and X_2 and W_2 (Figure 4.1b), respectively.

To create the $n_2 = 2500$ records for F_2, we begin by generating values for two continuous variables W_1 and W_2 where,

$$W_{1i'} \overset{iid}{\sim} N(2, 1^2), W_{2i'} \overset{iid}{\sim} N(.5, (.7)^2). \tag{4.6}$$

Conditional on W_1 and W_2, we generate $X_1|W_1$ and $X_2|W_2$ from

$$X_{1i'}|W_1 \sim Normal(.5 + 1.5 \times W_{1i'}, \sigma_1^2 = (.6)^2) \tag{4.7}$$

$$X_{2i}|W_2 \sim Normal(3 + 1.4 \times W_{2i'}, \sigma_2^2 = (.7)^2), \tag{4.8}$$

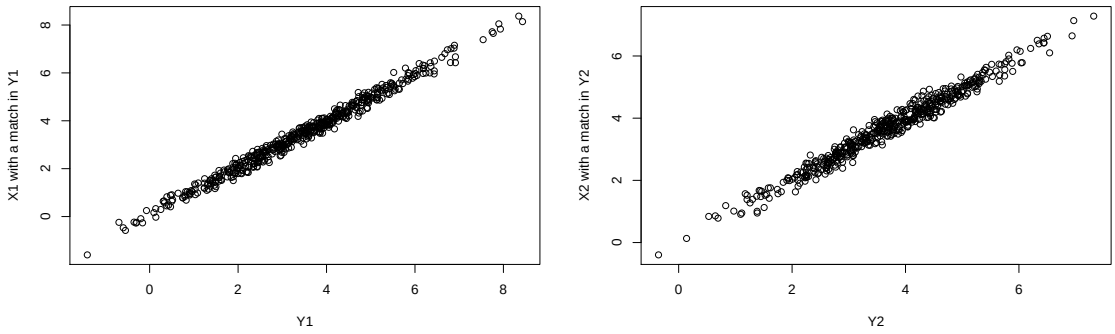
where $X_1 = (X_{11}, \ldots, X_{1n_2})$ and $X_2 = (X_{21}, \ldots, X_{2n_2})$ are vectors of length n_2. Figure 4.1 displays the relationship between X_1 and W_1, and X_2 and W_2.

Let $\boldsymbol{V_1} = (1, W_1)$ and $\boldsymbol{V_2} = (1, W_2)$ be $n_2 \times 2$ matrices, and let $\boldsymbol{\eta_1} = (.5, 1.5)^T$ and $\boldsymbol{\eta_2} = (3, 1.4)^T$. Then we have

$$X_1|W_1, \boldsymbol{\eta_1}, \sigma_1^2 \sim Normal(\boldsymbol{V_1}\boldsymbol{\eta_1}, \sigma_1^2) \tag{4.9}$$

$$X_2|W_2, \boldsymbol{\eta_2}, \sigma_2^2 \sim Normal(\boldsymbol{V_2}\boldsymbol{\eta_2}, \sigma_2^2). \tag{4.10}$$

The choice of generation model for X_1 and X_2 yields conditional distributions that are easily sampled from for the purposes of filling in missing data. This allows us to create first-stage imputations from the generating model for X_1 and X_2.



(a) X_1 with corresponding matches in Y_1. (b) X_2 with corresponding matches in Y_2.

FIGURE 4.2: Visualizing the relationship between X and Y. Figure 4.2a: X_1 and Y_1; Figure 4.2b: X_2 and Y_2.

After generating F_2, we generate F_1 as follows. Randomly partition the n_2 records $(X_{1i'}, X_{2i'}, W_{1i'}, W_{2i'})$ into $K = 500$ blocks of size 5. Within each of the K blocks, we randomly choose one record $i^{'}$ to be assigned a match i in F_1. For all such records $i^{'}$, we create a matching (Y_{1i}, Y_{2i}) where $Y_{ij}|X_{i'j} \overset{iid}{\sim} N(X_{i'j}, (.2)^2)$. Let $Y_1 = (Y_{11}, \ldots, Y_{1n_1})$ and $Y_2 = (Y_{21}, \ldots, Y_{2n_1})$ be vectors of length $n_1 = K$. Based on the generation of X_1 and X_2, the choice of variance represents a generation process for Y_{ij} in which each value is roughly 5% away from its corresponding value of $X_{i'j}$, as illustrated in Figure 4.2. The generating model for Y_1 and Y_2 also agrees with the model assumed by LFCMV, meaning that linking model for LFCMV is "correct" for these data.

We generate F_1 and F_2 100 times, creating 100 replicates. For each replicate, let $I = (I_1, \ldots, I_{n_2})$ be a vector of binary variables, where

$$I_{i'} = \begin{cases} 1 & : \text{ if } i^{'} \text{ has a match in } F_1 \\ 0 & : \text{ otherwise.} \end{cases} \tag{4.11}$$

For each $i^{'}$ in F_2, we generate a random variable

$$R_i^{'} \sim Normal(\beta_0 + \beta_1 \times I_{i'} + \beta_2 \times X_{2i'}, \sigma_1^2 = (.5)^2), \tag{4.12}$$

where $(\beta_0, \beta_1, \beta_2) = (1, .7, .2)$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. In this MIDL scenario, we are interested in estimating $\boldsymbol{\beta}$.

Let $X_2^{(m)}$ be the completed-data vector for X_2 for complete data set $\mathcal{D}^{(m)}$. Let $I^{(m,s)}$ be a vector containing the values of I conditional on $C^{(m,s)}$. Let $\mathbf{V}^{(m,s)} = (1, I^{(m,s)}, X_2^{(m)})$ be the $n_2 \times 3$ design matrix for the length n_2 vector $R = (R_1, \ldots, R_{n_2})$. Using the prior $p(\boldsymbol{\beta}, \sigma_1^2) \propto 1/\sigma_1^2$, posterior estimates for $\boldsymbol{\beta}$ may be obtained by sampling a value of $\boldsymbol{\beta}$ and σ_1^2 from the appropriate conditional distributions as follows.

$$\boldsymbol{\beta} | C^{(m,s)}, X_2^{(m)}, \sigma_1^2 \sim N\left(\widehat{\mathbf{B}}^{(m,s)}, \sigma_1^2 ({\mathbf{V}^{(m,s)}}^T \mathbf{V}^{(m,s)})^{-1}\right), \tag{4.13}$$

$$\widehat{\mathbf{B}}^{(m,s)} = ({\mathbf{V}^{(m,s)}}^T \mathbf{V}^{(m,s)})^{-1} {\mathbf{V}^{(m,s)}}^T {X_2}^{(m)}, \tag{4.14}$$

$$\sigma_1^2 | X_2^{(m)} \sim InvGamma\left(\frac{n_2 - 3}{2}, \frac{\left({X_2}^{(m)} - \widehat{\mathbf{B}}^{(m,s)}\right)^T \left({X_2}^{(m)} - \widehat{\mathbf{B}}^{(m,s)}\right)}{2}\right). \tag{4.15}$$

We do not condition on $I^{(m,s)}$, as $I^{(m,s)}$ is completely defined by $C^{(m,s)}$.

4.4.2 Missing data and first stage imputation

We introduce missing data in F_2 using a MCAR missing data mechanism. Specifically, we randomly select 20% of the X_1 values and, independently, select 20% of the X_2 values as missing. We blank out the value of the pertinent variable for each of the selected records. To create each first-stage completed data set, we begin by filling in the missing values for X_1. Let $X_1 = (X_{1obs}, X_{1mis})$, where X_{1obs} contains all $n_{2(1obs)}$ observed values of X_1 in F_2 and X_{1mis} denotes the $n_{2(1mis)}$ missing values; similarly partition $W_1 = (W_{1obs}, W_{1mis})$. Let $\boldsymbol{V_{1obs}} = (1, W_{1obs})$ be the $n_{2(1obs)} \times 2$ matrix of predictors for X_{1obs}. Using the prior $p(\boldsymbol{\eta_1}, \sigma_1^2) \propto 1/\sigma_2^2$ and conditioning on $\boldsymbol{X_{1obs}}$ and $\boldsymbol{V_{1obs}}$, we sample a value of $\boldsymbol{\eta_1}$ and σ_1^2 from the appropriate conditional

distributions as follows.

$$\boldsymbol{\eta_1}|\boldsymbol{V_{1obs}}, X_{1obs}, \sigma_1^2 \sim N\left(\widehat{\boldsymbol{\eta_1}}, \sigma_1^2(\boldsymbol{V_{1obs}}^T\boldsymbol{V_{1obs}})^{-1}\right), \tag{4.16}$$

$$\widehat{\boldsymbol{\eta_1}} = (\boldsymbol{V_{1obs}}^T\boldsymbol{V_{1obs}})^{-1}\boldsymbol{V_{1obs}}^T X_{1obs}, \tag{4.17}$$

$$\sigma_1^2|X_{1obs} \sim InvGamma\left(\frac{n_{2(1obs)}-2}{2}, \frac{(X_{1obs}-\widehat{\boldsymbol{\eta_1}})^T(X_{1obs}-\widehat{\boldsymbol{\eta_1}})}{2}\right). \tag{4.18}$$

Conditional on the sampled $\boldsymbol{\eta_1}$ and σ_1^2, we then sample X_{1mis} from the normal distribution $X_{1mis}|W_{1mis}, \boldsymbol{\eta_1}, \sigma_1^2 \sim N(\boldsymbol{V_{1mis}}\boldsymbol{\eta_1}, \sigma_1^2)$ Here, $\boldsymbol{V_{1mis}} = (1, W_{1mis})$ is the $n_{2(1mis)} \times 2$ matrix of predictors for X_{1mis}.

We use the same process to fill in the missing X_2 values. Let $X_2 = (X_{2obs}, X_{2mis})$, where X_{2obs} contains all $n_{2(2obs)}$ observed values of X_2 in F_2 and X_{2mis} denotes the $n_{2(2mis)}$ missing values; similarly partition $W_2 = (W_{2obs}, W_{2mis})$. Let $\boldsymbol{V_{2obs}} = (1, W_{2obs})$ be the $n_{2(2obs)} \times 2$ matrix of predictors for X_{2obs}. Using the prior $p(\boldsymbol{\eta_2}, \sigma_2^2) \propto 1/\sigma_2^2$ and conditioning on $\boldsymbol{X_{2obs}}$ and $\boldsymbol{V_{2obs}}$, we sample a value of $\boldsymbol{\eta_2}$ and σ_2^2 from the appropriate conditional distributions as follows.

$$\boldsymbol{\eta_2}|\boldsymbol{V_{2obs}}, X_{2obs}, \sigma_2^2 \sim N\left(\widehat{\boldsymbol{\eta_2}}, \sigma_2^2(\boldsymbol{V_{2obs}}^T\boldsymbol{V_{2obs}})^{-1}\right), \tag{4.19}$$

$$\widehat{\boldsymbol{\eta_2}} = (\boldsymbol{V_{2obs}}^T\boldsymbol{V_{2obs}})^{-1}\boldsymbol{V_{2obs}}^T X_{2obs}, \tag{4.20}$$

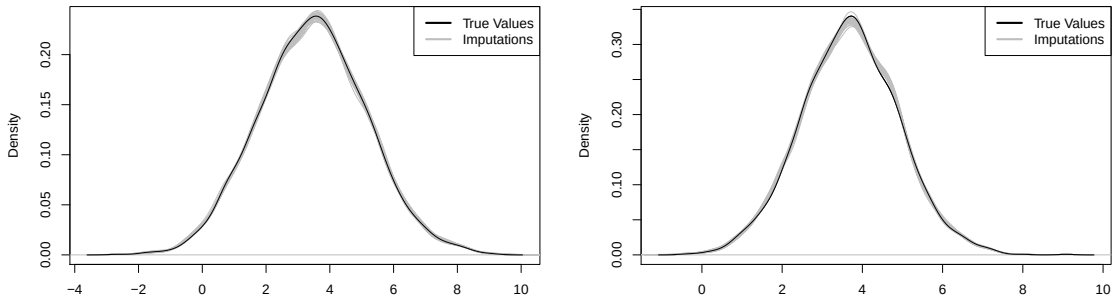
$$\sigma_2^2|X_{2obs} \sim InvGamma\left(\frac{n_{2(2obs)}-2}{2}, \frac{(X_{2obs}-\widehat{\boldsymbol{\eta_2}})^T(X_{2obs}-\widehat{\boldsymbol{\eta_2}})}{2}\right). \tag{4.21}$$

Conditional on the sampled $\boldsymbol{\eta_2}$ and σ_2^2, we sample X_{2mis} from the normal distribution $X_{2mis}|W_{2mis}, \boldsymbol{\eta_2}, \sigma_2^2 \sim N(\boldsymbol{V_{2mis}}\boldsymbol{\eta_2}, \sigma_2^2)$.

We repeat the steps of filling in the missing data multiple times to create $M = 30$ complete versions $\mathcal{D}^{(m)}$ of F_2.

4.4.2.1 Examining the imputations

For MIDL, the ability to impute from the correct distribution is an ideal scenario for first-stage imputation. In this case, we also assume knowledge not present in the

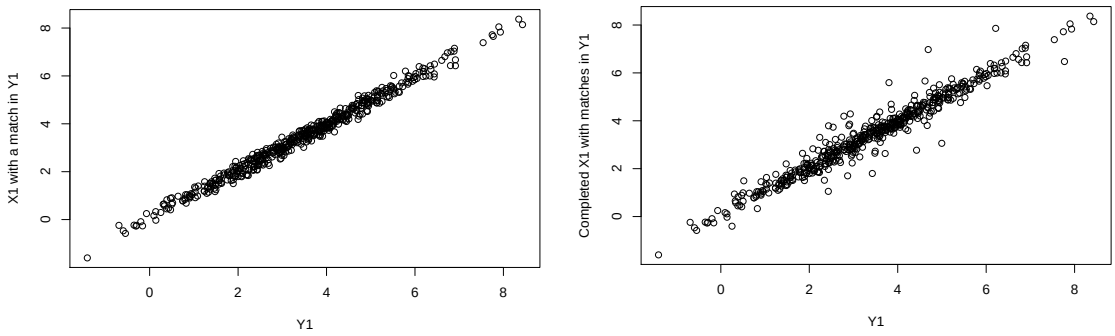


(a) X_1 (b) X_2

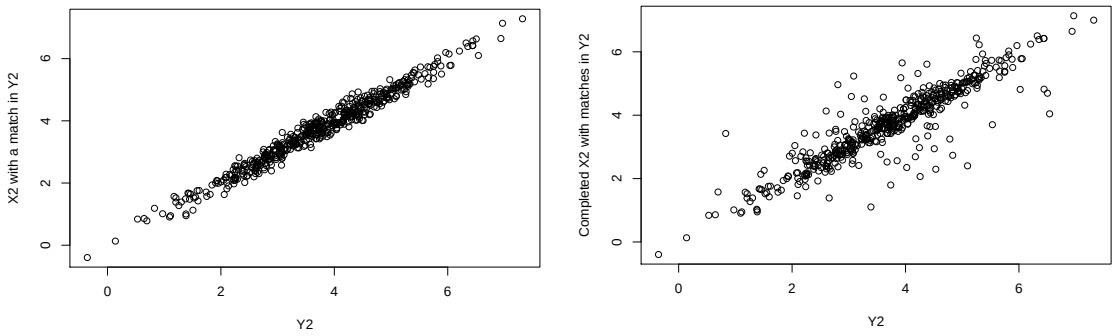
FIGURE 4.3: Comparison of the distribution of X_1 (Figure 4.3a) and X_2 (Figure 4.3b), respectively, to the imputed data sets. The black line is the empirical density of the true empirical density of X_1 or X_2, while each of the S grey lines is the empirical density of the appropriate variable in $\mathcal{D}^{(s)}$, $s = 1, \ldots, 30$.

data, namely that X_1 and X_2 are independent. For record pairs in F_2 with missing data on both $X_{1i'}$ and $X_{2i'}$, we impute $X_{1i'}$ and $X_{2i'}$ independently. Figures 4.3a and 4.3b illustrate the distribution of the true values of X_1 and X_2 versus their imputed values. In this scenario, the distribution of the imputations well approximates the marginal distribution of the true values.

In addition to the relationship between F_{2obs} and F_{2mis}, we are interested in examining whether or not the relationship between F_2 and F_1 are preserved by the imputations. Techniques for file linking that are not structured as MIDL settings, such as Gutman et al. (2013) and Dalzell and Reiter (2016), merge the process of imputing missing data with the process of estimating C, enabling the relationship between parameters in both files to be involved in the imputations. However, in MIDL, the second-stage missing data, i.e., the need to estimate C, means that the missing data in the first stage cannot be imputed conditional on the relationship between F_1 and F_2. In Figure 4.4, we see that the trend of the relationship between Y and X is preserved, but that the imputations have greater variability between Y



(a) X_1 with corresponding matches in Y_1. (b) $X_1^{(1)}$ with corresponding matches in Y_1.



(c) X_2 with corresponding matches in Y_2. (d) $X_2^{(1)}$ with corresponding matches in Y_2.

FIGURE 4.4: Visualizing the relationship between Y_1 and the imputed $X_1^{(s)}$. Figure 4.4a shows the original relationship between X_1 and Y_1; Figures 4.4c similarly displays the relationship between X_2 and Y_2. Figure 4.4b shows the relationship between Y_1 and $X_1^{(1)}$, i.e., the first completed data set. Figures 4.4d similarly display the relationship for Y_2 and $X_2^{(1)}$.

and X. This is not unexpected, as the values are imputations, but as we will see in Section 4.4.3, the imputations impact the linking results.

4.4.3 Results

There are a few facets to consider when analyzing results in an MIDL scenario. First, the data sets involved in the linking process have been completed via multiple imputation, meaning that the links estimated in the second-stage are conditional upon the quality of the imputations from the first round. It is therefore important for analysts to evaluate the quality of the first-round imputations before proceeding to linking.

For this simulation, we examine the first-stage imputations as follows. For each replicate, we estimate $\boldsymbol{\beta}$ using the data pairs $(F_1, \mathcal{D}^{(m)})$ for $m = 1, \ldots 30$ but conditioning on the true C. We call this the "Known C" run. Examining these results allows us to evaluate whether the relationships expressed by $\boldsymbol{\beta}$ are preserved in the first-round imputations. If this is not the case, any estimates obtained after MIDL will inaccurately estimate the parameters of interest. For each of the 100 replicates, we compute a replicate-specific estimate of $\widehat{\boldsymbol{\beta}}$ by drawing from the distributions in (4.13) and (4.15), but conditioning on the true C. For each completed data set $\mathcal{D}^{(m)}$, we draw 1000 estimates of $\boldsymbol{\beta}$. The empirical 95% intervals of these 100 replicate estimates for $\boldsymbol{\beta}$ are displayed in Table 4.1. The results suggest that the "Known C" runs estimate all three components of $\boldsymbol{\beta}$ decently well. The missing data has been imputed using the techniques described in Section 4.4.2, but second round imputations are not needed as we condition on the true C. The reasonable estimates of $\boldsymbol{\beta}$ suggest that the first-round imputations preserve the relationships expressed by these parameters.

Results in an MIDL scenario are also influenced by the accuracy of the links estimated in the second round of imputations. In other words, the linking technique selected for second-round imputations needs to be appropriate for the data. To allow us to examine the accuracy with which LFCMV is able to estimate C, for each

Table 4.1: 95% empirical intervals for $\widehat{\boldsymbol{\beta}}$. For each replicate, we average the estimates of $\boldsymbol{\beta}$ to produce a replicate-specific estimate $\widehat{\boldsymbol{\beta}}$. The intervals presented below are the 95% empirical intervals of these 100 estimates. Known C: the estimated intervals from when the missing data has been multiply imputed, but C is known. No Missing Data: the estimated intervals when there is no missing data, but C is estimated. MIDL: the estimated intervals from the entire MIDL scenario. True: the estimates intervals from fitting a standard linear regression model to the fully observed, correctly linked data.

	$\beta_0 = 1$	$\beta_1 = .7$	$\beta_2 = .2$
Known C	(1.0,1.08)	(.68,.72)	(.18,.20)
No Missing Data	(.98,1.04)	(.62,.68)	(.19,.21)
MIDL	(1.04,1.11)	(.54,.59)	(.18,.20)
True	(.97,1.02)	(.68,.72)	(.19,.21)

replicate we apply LFMCV to estimate C for the fully observed data (F_1, F_2). We call this the "No Missing Data" run. The results of this run also allow us to investigate how the estimates of C impact the estimation of $\boldsymbol{\beta}$. For each replicate, we run the LFCMV Gibbs sampler for 5500 iterations with a burn-in length of 500. For each of the 100 replicates, we compute a replicate-specific estimate of $\widehat{\boldsymbol{\beta}}$ by drawing from the distributions in (4.13) and (4.15), but conditioning on the true X_2 rather than $X_2^{(s)}$.

The results for the "No Missing Data" runs in the second row of Table 4.1 are comparable to the "Known C" estimates for β_0 and β_2. However, this is not the case for β_1. These runs condition on the fully observed data, so β_0 and β_2 are well estimated, but C, the parameter directly related to β_1, is unknown. As displayed in Table 4.1, 95% of the estimates for β_1 are between .62 and .68, slightly lower than the true value of .7. This bias is not unexpected. C must be estimated to determine the relationship between $\mathbf{Y}$ and $\mathbf{X}$. Because this relationship is, in actuality, very strong, we expect some attenuation as a result of the estimation process. Any missed matches will corrupt the estimates of C and likely result in smaller estimates of β_1. Overall, the results of the "Known C" runs suggest that LFCMV is a reasonable

linking technique for these data.

Once we have evaluated the first and second-stage imputations, we apply LFCMV to the full MIDL scenario. For each of the $M = 30$ nests generated in Section 4.4.2 for each of the 100 replicates, we apply the LFCMV model to link F_1 and the replicate specific $\mathcal{D}^{(m)}$. We call this the MIDL run. As with the "No Missing Data" run, we apply the LFCMV Gibbs sampler for 5500 iterations with a burn-in length of 500. For each of the 100 replicates, we draw one estimate from the distributions in (4.13) and (4.15) for each $M \times S$ completed data sets.

The third row of Table 4.1 contains results from the MIDL run. The estimates for β_2 are very similar across all three runs types, but there are differences in the estimation of β_0 and β_1. As in the "No Missing Data" runs, C is estimated in the MIDL runs, leading to an underestimation of β_1. Indeed, the estimates for β_1 in the MIDL scenario are smaller than the "No Missing Data" estimates. This in turn results in a slight overestimation of β_0.

The underestimation of β_1 in the MIDL scenario as compared to the "No Missing Data" scenario suggests that the presence of missing data hinders the process of estimating C. To see this, for each of the 100 replicates and each of the three run types, we compute a 95% credible interval for the PMR. For the "Known C" and "No Missing Data" runs, these intervals are computed using the approach of Gelman et al. (2004). For the MIDL runs, we create the intervals using the techniques discussed in Section 3.2. For ease of notation, we refer to the technique of Gelman et al. (2004) described in Section 4.3.1 as Estimation Technique 1 (ET1) and the technique discussed in Section 4.3.2 as Estimation Technique 2 (ET2).

Results are displayed in Table 4.2. The results in the table suggest that despite the same linking technique being used across the two run types, the missing data in the MIDL scenario results in a reduction in match rate. Part of the reason for this reduction is illustrated in Figure 4.5. In general, the imputations in X_2 result in a

Table 4.2: 95% intervals for the PMR. **No Missing Data**: estimated intervals on fully observed data but C is estimated. For each replicate, we compute the PMR obtained by linking data sets (F_1, F_2) by permutation $C^{(s)}$. We use estimation technique ET1 to obtain an estimate of the $(\alpha \times 100)^{th}$ quantile of the PMR, where $\alpha = .025, .25, .975$. The intervals presented are the 95% empirical intervals constructed from the 100 replicate specific estimate for each quantile. **MIDL**: the estimated intervals from the entire MIDL scenario. For each replicate, we compute the PMR obtained by linking data sets $(F_1, \mathcal{D}^{(m)})$ by permutation $C^{(m,s)}$. We then use estimation technique ET2 to obtain an estimate of the $(\alpha \times 100)^{th}$ quantile of the PMR, where $\alpha = .025, .25, .975$. The intervals presented are the 95% empirical intervals constructed from the 100 replicate specific estimate for each quantiles.

	$\alpha = .025$	$\alpha = .5$	$\alpha = ..975$
No Missing Data	(.92,.94)	(.93,.96)	(.95,.97)
MIDL	(.79,.82)	(.82,.86)	(.86,.89)

increase in the distance between Y_{2i} and $X_{2i'}$ for matched pairs $(i, i^{'})$ as compared to the true values. This in turn results in an increase in the estimation of the distance component of the linkage model and hence more variable estimates of C. This is not necessarily undesirable. In an MIDL scenario, it is important that estimates of C reflect uncertainty in the estimates of the first-stage imputations. For this simulation, this results in an attenuation in the estimation of β_1, but preserves the general trends. This suggests that in an MIDL scenario, the estimated magnitude of an effect may be inexact due to increased variability from the first-round imputations, but the general trends suggested by the results can be reflective of the analysis of interest.

To explore this behavior in more detail, we focus one replicate of the MIDL runs. As seen in Table 4.3 the imputations for X_1 and X_2 are sometimes quite different from the true values. For instance, the third record in $F_{2(i)}$ has a true X_1 value of 3.800, but an imputed value of 4.325. These differences impact the sampling of C. This can be seen in the final two columns of the table which display the proportion of the posterior draws for C_i matching i and each $i^{'}$ in $F_{2(i)}$ for the true data set F_2 versus the imputed $\mathcal{D}^{(1)}$. The third record in $F_{2(i)}$ is the correct match, but because X_1 and

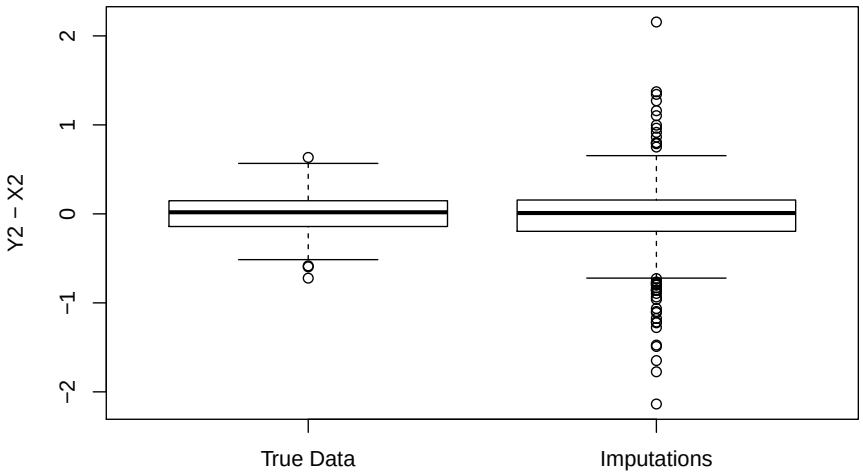


FIGURE 4.5: Illustrating the distance across matches pairs for the MV TE ($j = 2$).

X_2 are imputed, LFCMV only assigns 21% probability to that match. However, the match favored by LFCMV in the MIDL scenario favors the same relationship as the true match.

Finally, we examine the width of the credible intervals for each of the three run types. The width reflects the estimated uncertainty in the estimates for $\boldsymbol{\beta}$ for each run type. Figure 4.6 indicates similar widths for the "Known C" and missing data runs for β_0 and β_2, with slightly wider intervals for MIDL. Since β_0 and β_2 have similar estimates across the run types, similar width intervals are reasonable. Figure 4.6 shows an increase in the interval width for "No Missing Data" over "Known C", reflecting the increase in uncertainty from estimating C. The interval corresponding to the MIDL runs is the widest interval, reflecting the increased uncertainty from both the missing data an the estimation of C.

Table 4.3: Replicate 6: $i = 2$, $m = 6$. A comparison of the record value and estimates of C within this block. The true match is the 3rd record in $F_{2(i)}$. Row 1: Y_1 and Y_2 for a record i in F_1. Columns F_2 contain the true value of X_1 and X_2, respectively, in F_2. Columns labeled $\mathcal{D}^{(1)}$ contain the imputed values of X_1 and X_2 from the $\mathcal{D}^{(1)}$. If the values for X_1 or X_2 are observed, the column $\mathcal{D}^{(1)}$ is blank. Final two columns: proportion of the posterior draws for C_i matching i and each $i^{'}$ in $F_{2(i)}$ for the true data set (column F_2) the imputed (column $\mathcal{D}^{(1)}$).

	$Y_{1i} = 2.96$		$Y_{2i} = 4.16$		Proportion	
	X_1		X_2		$C_i = i^{'}$	
$F_{2(i)}$	F_2	$\mathcal{D}^{(1)}$	F_2	$\mathcal{D}^{(1)}$	F_2	$\mathcal{D}^{(1)}$
1	4.321		3.814	3.493	0	≈0
2	4.532	3.305	3.881		0	0.55
3	4.865		2.260		.98	0
4	2.564		3.393		.001	0.07
5	2.900	3.413	4.340		**0.99**	**.38**

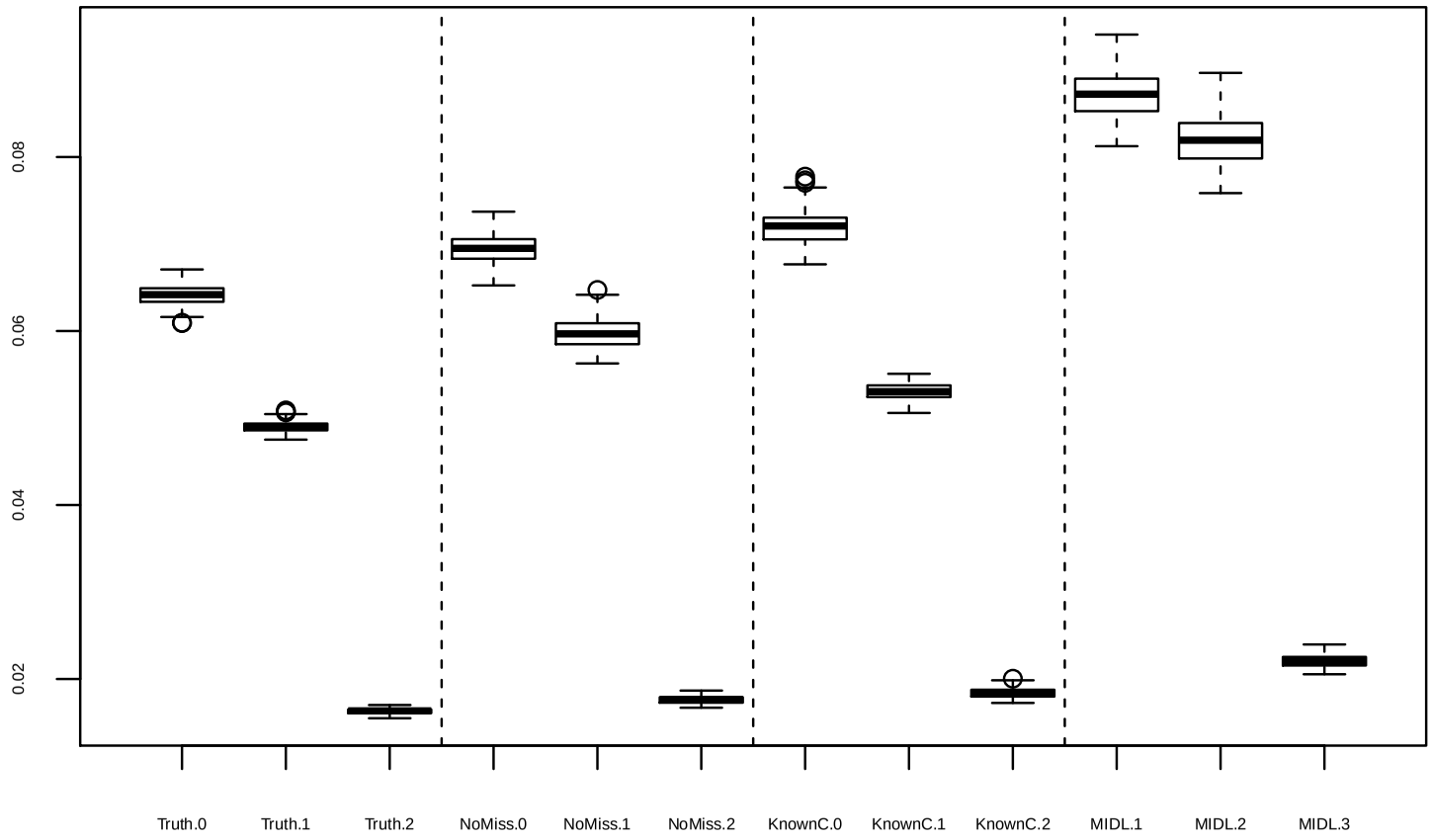


FIGURE 4.6: Width of 95% credible intervals for $\boldsymbol{\beta}$. Boxplots are composed of 100 points each, where each point corresponds to the width of the CI for one replicate. Numbers after the run name correspond to the parameters, i.e., β_0, β_1. or $beta_2$. Truth: width of the intervals estimated on the fully-observed data conditional on the true C. Known C: width of intervals estimated from multiply imputed data, but C is known. No Missing Data: width of intervals estimated on fully observed data but estimated C. MIDL: width of intervals estimated from the entire MIDL scenario. Known C and No Missing Data intervals are constructed using ET1. MIDL intervals are constructed using ET2.

4.4.4 Non-normal parameter of interest

One of the advantages of the Monte Carlo technique in Section 4.3.2 is that it does not rely on the assumption that the distribution of the parameter of interest is normal. Using the same data and completed data sets from Section 4.4.1, we choose a parameter Q that is not normally distributed. Recall that $I = (I_1, \ldots, I_{n_2})$ is a vector of binary variables, where

$$I_{i'} = \begin{cases} 1 & : \text{ if } i' \text{ has a match in } F_1 \\ 0 & : \text{ otherwise.} \end{cases} \tag{4.22}$$

We generate a Bernoulli random variable G, $G = (G_1, \ldots, G_{n_2})$, where the probability that $G_{i'} = 1$ depends on whether i' is a match to a record i in F_1. Specifically, $P(G_{i'} = 1) = \rho_M = .8$ if $I_{i'} = 1$, and $P(G_{i'} = 1) = \rho_U = .3$ otherwise. We take ρ_M as our parameter of interest Q.

Let $I_M^{(m,s)}$ be a set containing all records in F_2 that, according to $C^{(m,s)}$, have a match in F_1. For each of the $M \times S$ completed data sets for each replicate, and using a uniform prior for ρ_M, we draw one value of ρ_M from its posterior distribution conditional on $(F_1, \mathcal{D}^{(m)}, C^{(m,s)})$, where

$$\rho_M | C^{(m,s)} \sim Beta\left(\sum_{i' \in I_M^{(m,s)}} (G_{i'}) + 1, n_2 - \sum_{i' \in I_M^{(m,s)}} (G_{i'}) + 1 \right). \tag{4.23}$$

4.4.4.1 Results

In Section 4.4.3, the parameter of interest depends directly on both C and the imputations of X_2. For this simulation, estimating ρ_M depends on the imputations of X_2 only through the quality of estimates of C. For this reason, we do not perform the "Known C" run in this simulation. If we condition on a known C, ρ_M will be correctly estimated. We also do not discuss PMR in this section. The data and estimation of C are identical to the results discussed in Section 4.4.3.

Table 4.4: Empirical intervals. The intervals display the intervals containing 95% of the estimates of each quantile of ρ_M for each run type. No Missing Data: the estimated quantiles from the fully-observed data when C is estimated. Quantiles estimated using ET1. MIDL: the estimated quantiles from the entire MIDL scenario. Quantiles estimated using ET2.

	α =.025	α = .25	α = .5	α = .75	α = .975
No Missing Data	(.69,.76)	(.72,.79)	(.74,.80)	(.75,.81)	(.77,.83)
MIDL	(.64,.70)	(.67,.73)	(.69,.75)	(.70,.76)	(.73,.79)
Truth	(.72,.79)	(.75,.81)	(.76,.82)	(.77,.84)	(.80,.86)

In addition to the MIDL runs, we do perform "No Missing Data" runs to examine the influence of estimating C on the estimates of ρ_M. For the "No Missing Data" runs, there is only one source of missing data (C), so we apply ET1 to estimate ρ_M. Specifically, for each replicate, we mix the $M \times S$ draws of ρ_M and choose the $\alpha(MS)^{th}$ element to estimate the α^{th} quantile of ρ_M. To examine the estimation of the varaiblity of ρ_M, we use several choices of α. Here, we let $\alpha = (.025, .25, .5, .75, 975)$. For the MIDL runs, we apply ET2 and use the same set of values of α. Results are presented in Table 4.4.

The intervals presented are the 95% empirical intervals for the estimates of the $(\alpha \times 100)^{th}$ quantile of ρ_M. We see that, as discussed in Section 4.4.3 with respect to β_1, the uncertainty in the "No Missing Data" and MIDL scenarios results in an underestimation of the parameter of interest. However, this underestimation is fairly modest. As in Section 4.4.3, there is some bias in the estimation of ρ_M, as expected due to the relationship between $\mathbf{Y}$ and $\mathbf{X}$, but the general patterns in ρ_M are preserved in the estimations.

Figure 4.7 displays the width of 95% credible intervals for ρ_M for each of the 100 replicates. The MIDL intervals are wider than both the true and "No Missing Data" intervals, reflecting increased uncertainty in due to missing data in both F_2 and C.

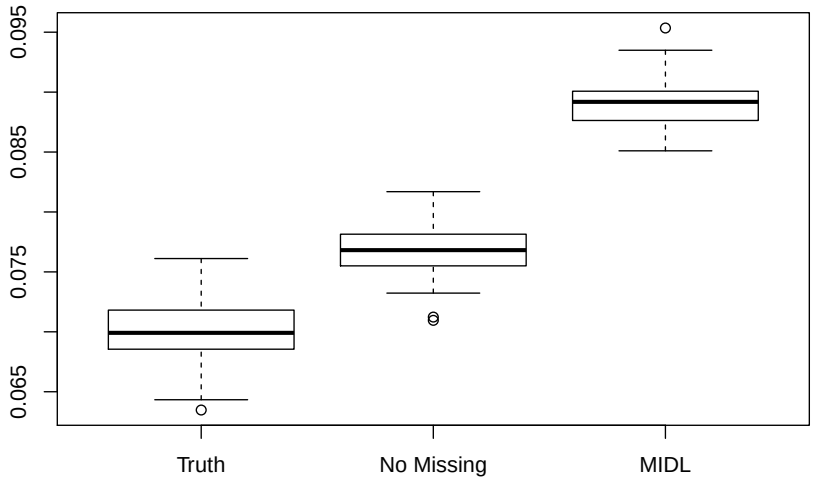


FIGURE 4.7: Width of 95% Credible intervals for ρ_M. Box plots are composed of 100 points, each corresponding to the width of the interval for each replicate. Truth: the width of the intervals for the parameters estimated on the fully-observed data conditional on the true C. No Missing Data: the estimated intervals on the fully-observed data when C is estimated. MIDL: the estimated intervals from the entire MIDL scenario. Known C and No Missing Data intervals are constructed using ET1. MIDL intervals are constructed using ET2.

4.4.5 Summary

The goal in MIDL settings is often to conduct inference involving data from both F_1 and F_2. However, because C is missing in an MIDL scenario, imputing values in one of the files can distort these cross-file relationships. From a linkage perspective, this can result in a loss of accuracy in estimating C, resulting in a lower PMR. From an analysis perspective, this loss of accuracy in C can lead to a reduction in accuracy when conducting posterior inference. However, in our simulations, MIDL did capture the main patterns in the relationship between $\mathbf{Y}$ and $\mathbf{X}$, albeit with some attenuation.

4.5 Data application

The simulation study in Section 4.4 illustrates that, with reasonable imputations, estimation after MIDL can preserve patterns present in the data. In the CMF application, however, the imputations themselves may not be reasonable. To help mitigate this concern, we use a non-parametric technique to re-impute some of the CMF imputations. In this section, we focus on examining the links obtained on the multiply-imputed data and comparing the results of an MIDL analysis to the results obtained in Chapter 3. However, as we saw in the simulations, even with the new imputations, it it important to be careful about interpreting the results of the MIDL analysis.

The remainder of this section proceeds as follows. In Section 4.5.1, we provide an overview of the Census Bureau imputations, including a list of the edit flags identifying the imputation techniques used in the CMF. We also identify the set of these edit flags for which we create new imputations. In Section 4.5.2, we detail the process of creating the imputations for the linking application. In Section 4.5.3, we apply the LFCMV linking methodology of Chapter 3 to link each of these M completed CMF data sets to the IAC. In Section 4.5.4, we perform the same energy-efficiency analysis as in Chapter 3 and compare the results. In Section 4.5.5 we conclude with a summary of future work.

4.5.1 Data cleaning and imputed values in the CMF

Prior to converting the IAC and CMF into the subsets F_1 and F_2 used for linking, we perform data cleaning as described in Section 3.6.1.1. This process standardizes the notation used for the BVs and removes from consideration for F_1 records in the IAC corresponding to plants in Puerto Rico.

In Chapter 3, we use three levels of NAICS blocking to account for uncertainty

in the reported NAICS codes. Specifically, we block based on 4-digit, 5-digit and 6-digit NAICS codes. However, as seen in Section 3.6.3, the results do not vary greatly across NAICS code. For this reason, in this section we block only on 4-digit NAICS code. Of the three choices of NAICS blocking used in the previous chapter, 4-digit is the least-restrictive blocking level, allowing the greatest number of possible links for each IAC record.

Once all BVs have been cleaned, we proceed to construct F_2 from the records in the 2007 CMF. We begin with the set of all plants in the CMF with a 4-digit NAICS code present in F_1, where F_1 contains all fully-observed records in the IAC from 2007 or 2008 that are not from Puerto Rico. Some of these CMF plants correspond to administrative records (AR), or plants with fewer than 5 employees that are not sent a survey for the CMF (White, 2014). A common practice in statistical analysis using the CMF data is to remove AR plants from analysis. We follow this practice and remove AR records from consideration for F_2.

Some records in consideration for F_2 records contain imputations constructed by the Census Bureau. The choice of the mechanism used to construct these imputations varies across records and variables. For this reason, records in the CMF contain edit/imputation flags associated with certain variables. These flags identify whether or not the value of the variable is an imputation, and, if so, which imputation technique was used to generate the value. A summary of these flags is presented in Table 4.5, with more detailed explanations presented in Tables 4.6 and 4.7. Some of imputation flags, like "analyst corrected" and "goldplated", correspond to an imputation created by an analyst directly correcting or imputing a value from expert information. For such flags, we choose to treat the CMF imputations as fixed. For other flags, such as "industry average", it is known that imputations constructed using these techniques can lead to poor analytical results on the completed data (Schafer and Graham, 2002; Little and Rubin, 2002; White et al., 2015). For such

Table 4.5: Description of edit imputation flags. The first three columns are based on Table 3 from White (2014), pg. 23.

Code	Name	Category	Re-impute Flag
(blank)	No flag set	Not imputed	
A	Administrative Records Data	Imputed	
B	Beta (Cold Deck Statistical)	Imputed	Y
C	Analyst Corrected	Non-imputed	
D	Donor Model Record	Imputed	Y
E	Endpoints of Limits (Upper/Lower)	Imputed	Y
G	Goldplated	Non-imputed	
H	Historic Values	Imputed	Y
J	Subject Matter Rule	Imputed	
K	Raked	Non-imputed	Y
L	Logical	Imputed	Y
M	Midpoints of Limits	Imputed	Y
N	Rounded	Non-imputed	
O	Override Edit with Reported Data	Non-imputed	
P	Prior Year Administrative Records Data	Imputed	Y
S	Direct Substitution	Imputed	Y
T	Trim and Adjust Algorithm	Imputed	Y
U	Unable to Impute	Non-imputed	Y
V	Industry Average	Imputed	Y
W	Warm Deck Statistical	Imputed	Y
X	Unusable	Non-imputed	Y
Z	Acceptable Zero	Non-imputed	

flags, which we call re-impute flags, we create multiple imputations for the values of the corresponding variable prior to performing linkage. For a list of flags that are designated re-impute flags, see the final column of Table 4.5.

Table 4.6: Detailed Description of edit imputation flags A - S.

Code	**Meaning**
A	the item is imputed by direct substitution of corresponding administrative data (for the same establishment/record)
B	the item is imputed from a statistical (regression/beta) model based on historic data.
C	the reported value fails an edit, and an analyst directly corrects the (reported or imputed) value.
D	the item is imputed using hot deck methods.
E	the item is imputed by direct substitution of value near (high or low) endpoints of imputation range.
G	the reported value for the item is "protected" from any changes by the edit. The value of a goldplated item is not changed by the editing system, even if the item fails one or more edits. In general, the goldplate flag is set by an analyst.
H	the item is imputed by ratio imputation using historic data for the same establishment (for example, prior year data imputation in Manufacturing)
J	the item is imputed using a subject matter defined rule (e.g. $y = 1/2x$).
K	the sum of a set of detail items do not balance to the total. The details are then changed proportionally to correct the imbalance. This preserves the basic distribution of the details.
L	the item's imputation value is defined by an additive mathematical relationship (e.g., obtaining a missing detail item by subtraction).
M	the item is imputed by direct substitution of midpoint of imputation range
N	the reported value is replaced by its original value divided by 1000
O	the reported value fails an edit. Either an analyst interactively restores the originally reported value of an edit (set by the interactive update system) or the ratio module later "imputes" originally reported data for an item which was imputed in the previous edit pass.
P	the item is imputed by ratio imputation using corresponding administrative data from prior year (for same establishment).
S	the item is imputed by direct substitution of another item's value (from within the same questionnaire.)

Table 4.7: Detailed Description of edit imputation flags T-Z.

Code	Meaning
T	the item was imputed using the Trim-and Adjust balancing algorithm (balance module default)
U	the reported item is blank or fails an edit, and the system cannot successfully substitute a statistically reasonable value for the original data.
V	the item is imputed by ratio imputation using an industry average.
W	the item is imputed from a statistical (regression/beta) model based on current data.
X	the sum of a set of detail items cannot be balanced to the total because none of the scripted solutions achieved a balance
Z	the reported value for an item is zero, and the item has passed a presence (zero/blank) test. This often occurs with part time reporters (e.g., births, deaths, idles). The zero value will not be changed, even if it fails one or more edits.

For our first attempt at imputation for the re-impute flags, we grouped all possible records for F_2 by 4-digit NAICS code. We then blanked out all CMF imputations corresponding to re-impute flags, and attempted to re-impute the values. However, for some 4-digit NAICS groups, all or nearly all of the values for some of the variables of interest correspond to re-impute flags. This means that there is little or no information to use to construct new imputations for this variable within the NAICS code groups. In order to create more stable imputations, we instead apply a filter requiring a certain percentage of the values for TE, TVS, CF and EE to be fixed, i.e., to correspond to a flag that is not a re-impute flag. Specifically, we sort all candidate F_2 records into groups defined by 6-digit NAICS code; 6-digit NAICS code defines the industry code for a plant. Within each 6-digit group, and for each of the four variables of interest, we compute the percentage of records corresponding to re-impute flags. If more than 40% of the values of TE or TVS have re-impute flags, or more than 50% of the values for CF or EE have re-impute flags, we remove all records corresponding to the 6-digit NAICS code from consideration for F_2. By extension, plants which such 6-digit NAICS codes are also excluded from considera-

Table 4.8: A description of the percentage of variables in our imputation sample have re-impute flags. Percents have been rounded to satisfy disclosure protocols.

	TVS	TE	CF	EE	At least one	None
Re-impute flags	30%	26%	34%	35%	40%	60%

tion for F_1. We consider these industries as having too much missing data to yield accurate re-imputations, as attempting to impute with such low levels of observed data will likely results in high variability imputations.

After removing records with too much missing data, we have a pool of candidates for F_2. We call this pool our imputation subset, as these records will be used to construct the M completed data sets in the first-stage of imputation. Table 4.8 displays the percentage of records in the imputation subset with re-impute flags on each of the variables of interest. Percentages have been rounded to satisfy disclosure protocols. From the table, we see that roughly 60% of the plants in the imputation subset have fixed values for all four variables of interest. The remaining 40% have at least one variable of interested correspond to a re-impute flag.

4.5.2 First-stage imputation with CART

For each record in the imputation subset described in Section 4.5.1, we group the records by 4-digit NAICS code. Within 4-digit NAICS code group, we impute all values of TE, TVS, CF and EE with re-impute flags using classification and regression trees (CART) (Burgette and Reiter, 2010). Imputations based on CART rely on estimating conditional distributions for imputed values. CART is also used to impute CMF data in White et al. (2015).

CART is an example of a chained approach to MI. However, unlike standard mulitple imputation with chained equations (MICE), CART is not restricted to parametric models. Instead, CART uses multiple predictors to approximate the condi-

tional distribution of each variable containing missing values. During the process of approximating these conditional distributions, CART automatically considers non-linear terms and well as interactions among the possible predictors. This presents a greater flexibility for approximating joint distributions while requiring little manual variable selection on the part of the analysts. In a complex data set like the CMF, it is often difficult to define a parametric model specifying the relationships among the variables. Using CART eliminates the need for such pre-imputation specification. However, CART does rely on the assumption that the data to be imputed is missing at random. See Section 1.4.1 to review missing data mechanisms.

CART classifies units, in this case manufacturing plants, into groups, called "leaves" which are relatively homogeneous. The process begins by deleting all values that need to be re-imputed. These values are then filled in with an initial guess; (Burgette and Reiter, 2010) provides suggestions for filling in these values. Variables are then ordered in terms of increasing percentage of missing data. This means that we begin by conditioning on all variables except the one least observed. Call this variable Y_1. Fit a regression tree for Y_1 by conditioning on all other variables. The tree grows by finding splits in the plants which minimize the variance of Y_1 in the leaves. The process of splitting leaves stops when (1) it is impossible to keep at least $k = 5$ plants in each leaf or (2) the variance of the fitted Y_1 is less than 0.00001 times the variance of the marginal distribution of Y_1. Once the tree is grown, values for Y_1 for any missing or faulty values are imputed from the farthest leaf containing the plant.

In the application, we apply CART separately to each 4-digit NAICS group of manufacturing plants. For each group, we use 100 iterations to grow each tree and create 20 completed data sets for each NAICS code group. To creation the final completed data sets $\mathcal{D}^{(m)}$, $m = 1, \ldots, 20$, of F_2, we combine the completed data set m for each of the NAICS code groups.

Table 4.9: The percentage of MVS in F_2 that have re-impute flags. Here, F_2 represents the set of all CMF plants with a possible state, 4-digit NAICS match in F_1. Percents have been rounded to satisfy disclosure protocols.

	TVS	TE	At least one	None
Re-impute flags	31%	27%	33%	67%

4.5.2.1 Agreement filters

Creating the completed data sets $\mathcal{D}^{(m)}$ produces M subsets of records that we will link to F_1. We use the BVs of state and 4-digit NAICS code to determine which records in $\mathcal{D}^{(m)}$ are considered possible matches for each i in F_1. Let $F_{2(i)}^{(m)}$ denote the set of possible matches in $\mathcal{D}^{(m)}$ for each i in F_1. Because the BVs contain no missing data, at this stage $F_{2(i)}^{(m)}$ contains the same plants for all m. We consider all plants in an $F_{2(i)}^{(m)}$ for at least one i as our final F_2. For this set F_2 of CMF plants, we compute the percentage of the MVs in these records that contain re-impute flags. This expresses the percentage of the records in F_2 that will vary across m. Results are displayed in Table 4.9, with the reported percentages rounded to satisfy disclosure protocols.. For roughly 67% of the records in F_2, all MVs are fixed across choices of M. The remaining 33% of the records have an imputed value for at least one MV.

As in Chapter 3, blocking on only state and 4-digit NAICS produces large blocks for some records i. Therefore, we apply the same agreement filter to narrow the possible options for each IAC record, with some new notation to reflect the MIDL setting. Specifically, let $TVS_{2i'}^{(m)}$ denote the value for TVS in completed data set $\mathcal{D}^{(m)}$ for record i', and similarly let $TE_{2i'}^{(m)}$ denote the value for employment. For each i in F_1 and each i' in $F_{2(i)}^{(m)}$, we compute

$$d_1(i,i')^{(m)} = \frac{TVS_{1i} - TVS_{2i'}^{(m)}}{TVS_{1i}}, \tag{4.24}$$

and

$$d_2(i, i^{'})^{(m)} = \frac{TE_{1i} - TE_{2i^{'}}^{(m)}}{TE_{1i}}. \tag{4.25}$$

Here, $d_1(i, i^{'})^{(m)}$ and $d_2(i, i^{'})^{(m)}$ are, respectively, the percent difference between IAC record i and the value of Census record $i^{'}$ on TVS and TE in completed data set $\mathcal{D}^{(m)}$. If either (4.24) or (4.25) is greater than .3, we declare i and $i^{'}$ a non-link with respect to $\mathcal{D}^{(m)}$ and remove $i^{'}$ from $F_{2(i)}^{(m)}$. Otherwise, we allow i and $i^{'}$ to be possible links. At this stage, $F_{2(i)}^{(m)}$ may contain different plants across different values of m. This underlines the correlation between the estimated values of $\mathcal{C}^{(m)}$ and the first-stage imputations.

4.5.2.2 Transformations and standardization

After applying the agreement filter described in Section 4.5.2.1, we have our final subset $F_2^{(m)}$ for each $m = 1, \ldots, M$. As in Chapter 3, before linking, we transform the MVs by taking the log and standardize the MVs. The standardization is performed as described in Section 3.4.1. Recall that $j = 1$ refers to TVS and $j = 2$ refers to TE. Let $Y_{ij}^{(m)}$ denote the standardized value for MV j for record i in conditional on $\mathcal{D}^{(m)}$. Similarly, let $X_{i^{'}j}^{(m)}$ denote the standardized value for MV j for record $i^{'}$ conditonal on $\mathcal{D}^{(m)}$.

4.5.3 Second-stage imputation: linkage

For each pair $(F_1, \mathcal{D}^{(m)})$, we run LFCMV to link the data sets using the posterior sampling steps described in Chapter 3. For each m, we initialize $C^{(m,0)}$ by selecting, for each record i in F_1, the record $i^{'}$ in $F_{2(i)}^{(m)}$ such that

$$d(i, i^{'}|m) = \sqrt{\sum_{j=1}^{2} \left(\frac{Y_{ij}^{(m)} - X_{i^{'}j}^{(m)}}{Y_{ij}^{(m)}} \right)}, \tag{4.26}$$

Table 4.10: Conditional on each completed data set, for each i in F_1 there is a highest-probability match in $F_2^{(x)}$. The counts above indicate how many of these highest-probability matches stay the same across all $M = 20$ completed data sets. Counts have been rounded to satisfy disclosure protocols.

	All Data Sets	At least 50% of the data sets
Count	190	370

is minimized. We then fit the LFCMV model using the posterior sampling steps described in Section 3.3.2. We run the sampler for 5000 iterations, with a burn-in of 1000, resulting in $S = 4000$ posterior samples of C for each m, where $\mathcal{C}^{(m)}$ denotes the set of these S samples.

Conditional on $\mathcal{C}^{(m)}$, we compute the posterior probability that each record i' in $F_2^{(m)}$ is linked to each i in F_1. This probability may differ across the $M = 20$ completed data sets based on the values of the imputations for TE and TVS. Indeed, as discussed in Section 4.5.2.1, a record i' that is in $F_{2(i)}^{(m)}$ for one choice of m may not even be an option for record i for another choice of m based on the application of the agreement filter. This is an example of the correlation structure involved in two-stage imputation. However, some links are consistent across m. Within each completed data set, we identify the highest probability link for each i in F_1. There are roughly $n_1 = 500$ records in F_1, where this count has been rounded to satisfy disclosure protocols. We then look across the data sets to determine if the highest probability link agrees across m. Results are displayed in Table 4.10. Roughly 190 of the 500 F_1 records are consistently linked with the same record in the CMF across choices of m.

The fact that there are similar links across the imputations is indicative of two features of the data. First, not all of the data involved in linking is imputed. Table 4.9 gives the percentage of records in F_2 that have imputed values for TE or TVS. If all possible matches for a record in the IAC are fully observed, linking results

will be consistent across data sets. However, there are 33% of the potential F_2 links which involve missing data. For these, we expect the probability of being linked to vary across data sets. The degree of this variation can be partially attributed to the variability in the imputations, and some to the data itself. If there are two very similar choices for F_2 for a given record i, we expect and indeed want the probability of a match to reflect this. The MIDL scenario allows us to conduct linking in such a way that we integrate not only the uncertainty in the links, but also uncertainty in the underlying F_2.

4.5.4 Post-linking analysis

As was done in Chapter 3, the links estimated in Section 4.5.3 can be used to construct identifiers, specifying, conditional on each completed data set, which records received an IAC energy assessment. In the context of two-stage linking, each completed data set represents a nest, and within that nest, each sampled C is a second-stage imputation. This means that we have $M \times S$ completed, linked data sets upon which to perform analysis.

As a basis for our analysis, we use the same analysis samples as Chapter 3, i.e., we begin with all CMF records from 2007 and 2012 that have a 4-digit NAICS code observed in the IAC for 2007 or 2008. This limits our analysis to industries that were assessed in the time period of interest. We further limit our analysis sample to plants meeting IAC standards for assessment. Recall that IAC assessments are typically performed for plants with no more than 500 employees with a gross annual sales of no more than \$100 million (Muller, 2001, pg. 13). Let A_{2007} denote the set of CMF records from 2007 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. Similarly, let A_{2012} denote the set of CMF records from 2012 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. We refer to A_{2007} and A_{2012}

as our analysis samples.

For each completed data set, we construct a set of energy scores as described in Chapter 3. To satisfy disclosure protocols, we are required to exclude from analysis plants corresponding to the top 5% and bottom 5% of energy scores in each year. For each iteration s, we compute the density of all other energy scores for "With Assessment" plants, and separately for "Without Assessment" plants, for 2007 and 2012. We use fixed bins across the iterations and across the completed data sets. Separately for each year, and separately for each of the two assessment groups, we average the density in each bin across iterations and across completed data sets. This choice mimics the logic in Zhou and Reiter (2010) of mixing together all posterior draws and estimating quantiles by sorting the draws. Within each bin in the density plot, we mix together all $M \times S$ values and plot the mean as well as the 95% interval for the scores within each bin.

The resultant average density curves for 2007 are displayed in Figure 4.8, and the resultant averaged density curves for 2012 are displayed in Figures 4.9. Solid lines are used for the "Without Assessment" group and dashed lines are used for the "With Assessment" group. The trends in the density curves are very similar to the results from Chapter 3. In 2007, the density curve for the "With Assessment" group appears to be shifted further left than the "Without Assessment" curve. This suggests that the energy scores for assessed plants seem smaller than plants without assessments. In other words, relative to their peer group, plants which receive IAC audits appear less energy efficient. This trend appears to vanish in the plots corresponding to 2012. In Figures 4.9, there is essentially no difference in the curves between the "Without" and "With Assessment" groups. We stress that we do not make a causal argument about the relationship between plants with and without assessments. Indeed, based on the simulations conducted in Section 4.4, we note that the imputations may be confounding some aspects of the analysis. However, the results in Section 4.4

Table 4.11: Reports with matches estimated with re-imputed data: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment" group.

Re-imputed Data		Mean	95% Interval
2007	With Assessments	-0.17	(-.27,-.07)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.03	(-0.052,0.113)
	Without Assessments	0.009	(0.006,0.011)

Table 4.12: Results with matches estimated with original CMF values: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment".

CMF values		Mean	95% Interval
2007	With Assessments	-0.18	(-.27,-.09)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.03	(-0.04,0.11)
	Without Assessments	0.008	(0.006,0.011)

systematically under-estimated the relationships of interest relating to the linkage structure. The fact that we see any differences here motivates further analysis into what facets of the data might lead to the distinctions between the two curves in 2007, and the erasure of this trend in 2012.

In addition to the graphical exploration of the data, we apply the combining rules of Shen (2000) to construct confidence intervals for the mean energy scores of the "Without Assessment" and "Without Assessment" groups. Results are displayed in Table 4.11. For comparison, we also present the results from Chapter 3 in Table 4.12. The intervals in the table suggest the same general trends observed in the density curves. In 2007, the average energy score tends to lower for the "With Assessment" groups, with non-overlapping intervals for the "With" and "Without" groups. This difference is much less pronounced in 2012, and the intervals overlap, suggesting no statistically significant difference in the means between the groups.

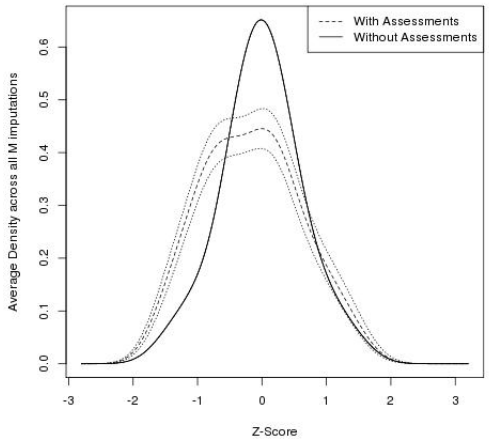


FIGURE 4.8: NAICS 4 Blocking, 2007: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

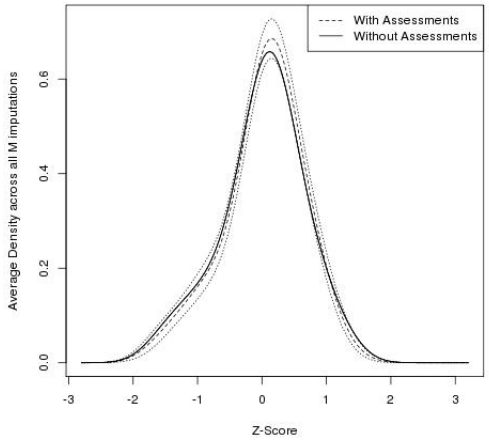


FIGURE 4.9: NAICS 4 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

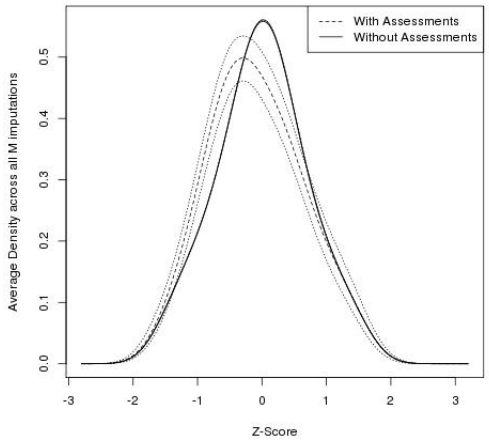


FIGURE 4.10: NAICS 4 Blocking, 2007: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

4.5.5 Future work

The data trends in Figure 4.8 are very similar to those results in Figure 3.8 in Chapter 3. The only difference in the construction of these two plots is that the linkage structure in Figure 4.8 is estimated using our re-imputed versions of the CMF. However, the energy scores used to construct the curves in both figures are the released CMF values.

To see the impact of the imputations on the energy scores, for any records in A_{2007} with a re-impute flag, we replace the corresponding CMF value with the M imputations constructed in Section 4.5.2. We then re-compute the energy scores conditional on these new values. The resultant plot is in Figure 4.10. We can see that this curve is quite different from the curves in Figure 4.8. While the distributions for the "With" and "Without" Assessments groups are slightly different, the clear different in distributions evident in Figure 4.8 is not as evident in Figure 4.10. The intervals corresponding to these two analyses are in Table 4.13. This suggests that

Table 4.13: Reports with matches estimated with re-imputed data: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment" group. Row 1-2: Using the CMF data and CMF imputations. Row 3-4: Using the CMF data and CART imputations.

2007		Mean	95% Interval
CMF Data	With Assessments	-0.17	(-.27,-.07)
	Without Assessments	-0.007	(-0.01,-0.005)
CART	With Assessments	-.09	(-0.19,0.004)
	Without Assessments	-0.007	(-0.007,0.010)

the imputations for CF and EE, two variables involved in the computation of the energy but not used in the linking, may have changed the distributions of these variables in A_{2007}. Understanding why this occurs is a topic for future work. We also want to explore the impact of imputations using the data in A_{2012}. This requires jointly imputing the data from 2007 and 2012 before performing and analysis of energy scores.

Appendix A

Dirichlet Process Mixture Model of Products of Multinomials

Traditional blocking conditions on a set of comparison fields. In Chapter 2, the comparison fields used for blocking are dependent categorical variables assumed to contain disagree among some true matches. This disagreement is corrected by re-imputing values for fields identified as being plausibly faulty using a Bayesian non-parametric model presented by Dunson and Xing (2009), which is designed to handle a potentially sparse table of dependent categorical variables. This model, which we call the Dirichlet Process Mixture Model of Products of Multinomials (DPMPM), is a latent class model in which categorical data are grouped into classes based on dependencies among the values. Once a record has been assigned to a latent class, field values may be drawn independently from a class-specific multinomial distribution.

Suppose record i has a vector $B_i = (B_{i1}, B_{i2}, ..., B_{iJ})$ of unordered categorical comparison fields. Let d_j be the number of options, or levels, for each field $j = 1, ..., J$. Suppose record i belongs to latent class z_i, where $z_i \in \{1, ..., H\}$. The DPMPM

models each record field B_{ij} as,

$$B_{ij}|\phi, z_i \sim Multinomial\left(\{1, ..., d_j\}, \phi_{z_i j 1}, ..., \phi_{z_i j d_j}\right) \tag{A.1}$$

$$z_i|\pi \overset{ind}{\sim} Multinomial(\{1, ..., H\}, \pi_1, ..., \pi_H). \tag{A.2}$$

Here, B_{ij} is a multinomial random variable with the sample size equal to 1. z_i is a multinomial random variable whose value specifies a latent class assignment for a given record i. Dunson and Xing (2009) allow z_i to reflect a potentially infinite number of latent classes. However, just as the contingency table for the BVs is sparse, in reality only a finite number of latent classes are necessary. The methodology of Ishwaran and James (2001) describes a finite mixture model approach to the DPMPM which populates a finite number H^* of latent classes (Dunson and Xing 2009, Si and Reiter 2013). The value H^* is fixed by the user, and its selection is based on a guess at the number of classes $\mathcal{H}$ that might be necessary to explain the dependence structure in the data. Ideally, $H^* >> \mathcal{H}$. At each iteration of the sampler, the DPMPM prior will populate, or assign records to, the H^* latent classes. H^* is well chosen if exploration of the runs indicates that the number of populated latent classes at each iteration of the Markov Chain Monte Carlo (MCMC) is fewer than H^*. If the sampler populates H^* latent classes, this suggests that $\mathcal{H} > H^*$. The model requires more latent classes to describe the dependence structure among the categorical variables than H^* allows. In such a case, H^* needs to be increased.

Describing the H^* latent classes in terms of the data relationships they represent is difficult. On a surface level, one latent class may seem to contain many individuals from a certain school who are female, while a second may contain primarily males. However, the relationships which define the latent class assignments are often complex and difficult to quantify in the posterior. Additionally, at each iteration latent class "1" may encode a different relationship. This introduces a label-switching concern, which complicates tracking the class relationships during the MCMC. Indeed,

the number of latent classes utilized at each iteration of the MCMC varies, signifying that different relationships may be used to define the class structure.

Understanding why a record has been assigned to a latent class is not of primary interest in our use of the DPMPM. The strength of the latent class structure for our application is that once a record has been placed within a class, values for the latent BVs can be drawn independently from a class specific multinomial distribution. For each latent class $h \in \{1, ..., H^*\}$, the multinomial probabilities ϕ from (A.1) are modeled as

$$\phi_{hj} \sim Dirichlet(a_{j1}, ..., a_{jd_j}). \tag{A.3}$$

Here, the vector $\phi_{hj} = (\phi_{hj1}, ..., \phi_{hjd_j})$ describes the probabilities that variable j for record i should take field value 1 through d_j respectively. The recommended default setting of Dunson and Xing (2009) for the hyper-parameters is $a_{j1} = ... = a_{jd_j} = 1$ for all h.

In (A.2), for every record i there is some probability π_h that i belongs to latent class h, where $\pi = (\pi_1, ..., \pi_{H^*})$. In the finite mixture model representation suggested in Ishwaran and James (2001), π is sampled using a stick-breaking technique, and the probability that $z_i \in \{1, ..., H^*\}$ must be 1. Probabilities are assigned to the H^* classes utilizing the following structure:

$$\pi_h = V_h \prod_{g<h} (1 - V_g), h = 1, ..., H^* \tag{A.4}$$

$$V_h \sim Beta(1, \alpha), V_{H^*} = 1. \tag{A.5}$$

To ensure that $\sum_{h=1}^{H^*} \pi_h = 1$, enforce $V_{H^*} = 1$ (Ishwaran and James, 2001).

While it allows all H^* classes to be populated, the finite mixture model representation of the DPMPM actually favors the allocation of practical probability mass to few classes. The speed with which the mass for new classes diminishes is controlled by the concentration parameter α in (A.5). α is a modeled as gamma distributed

random variable, $\alpha \sim Gamma(a_\alpha, b_\alpha)$. With an expected value of $\frac{1}{\alpha}$, the probability π_1 of being assigned to the first latent class increases as α decreases. This means a large α encourages more populated latent classes than a smaller α. A large α leaves more mass to be allocated to new classes while a small α leaves a smaller remaining probability mass for new latent classes. With a model on α, the number of occupied latent classes (classes to which records are actually assigned) can vary at each iteration of the MCMC. This allows for uncertainty in the effective number of latent classes. Dunson and Xing (2009) suggest hyper-parameter settings of $a_\alpha = b_\alpha = 0.25$.

The Gibbs sampling update steps for the DPMPM are as follows:

1. For $i \in 1, ..., N$, sample the latent class indicator $z_i^{(s+1)} \in \{1, ..., H^*\}$ from a multinomial distribution with sample size 1 and

$$Pr(z_i^{(s+1)} = h | B, \phi^{(s)}, \pi^{(s)}) = \frac{\pi_h^{(s)} \prod_{j=1}^{J} \phi^{(s)}_{hjB_{i'j}}}{\sum_{h'=1}^{H^*} \pi_{h'}^{(s)} \prod_{j=1}^{J} \phi^{(s)}_{h'jB_{i'j}}}. \tag{A.6}$$

2. For $h \in \{1, ..., H^* - 1\}$, sample

$$V_h^{(s+1)} | z^{(s+1)}, \alpha^{(s)} \sim Beta\left(1 + m_h^{(s+1)}, \alpha^{(s)} + \sum_{h'=h+1}^{H^*} m_h^{(s+1)}\right). \tag{A.7}$$

   where $m_h^{(s+1)} = \sum_i I(z_i^{(s+1)} = h)$ denotes the number of individuals in latent class h. Set $V_{H^*} = 1$.

3. Compute $\pi_h^{(s+1)} = V_h^{(s+1)} \prod_{g<h}(1 - V_g^{(s+1)})$ for all $h \in \{1, ..., H^*\}$.

4. For $j \in \{1, ..., J\}$ and $h \in \{1, ..., H^*\}$, sample

$$\phi_{hj}^{(s+1)} | z^{(s+1)}, B^{(s)} \sim Dirichlet\left(a_{j1} + \sum_{i:z_i=h} I(B_{ij}^{(s)} = 1), ..., a_{jd_j} + \sum_{i:z_i=h} I(B_{ij}^{(s)} = d_j)\right).$$

5. $\alpha^{(s+1)} | \pi^{(s+1)} \sim Gamma\left(a_\alpha + H^* - 1, b_\alpha - log(\pi_{H^*}^{(s+1)})\right)$.

Appendix B

Extended Simulation Results for BLASE

In this appendix we present further simulation results from the BLASE simulation discussed in Chapter 2. These results are summarized in Table 2.3 in the main paper. For each of the four simulation groups, we provide figures for (1) PMR, (2) RMSE and (3) parameter estimation for the Y_1 regression. Supplementary Table B.1 summarizes the parameter estimation from the Y_2 regression, but because the behavior mimics that observed in the Y_1 regression estimates, we do not include plots for the Y_2 estimates. Some of this material is repeated from Chapter 2 for completeness.

Table B.1: Summary of results: Y_2 regression

		Int	prog = Acad	prog = Voc	ses = middle	ses = high	female
	CA	**-.10**	**42.6**	**16.5**	**.68**	**16.11**	**5.9**
HSHF	D	**-.11**	**42.6**	**15.7**	.48	**-12.1**	**5.4**
	CP	**-.13**	**37.2**	**11.2**	.20	**-9.5**	**4.8**
	CA	**-1.3**	**-21.3**	**8.5**	**5.3**	**.003**	**.11**
LSHF	D	**-1.2**	**-29.2**	**15.9**	**8.0**	**.04**	**.02**
	CP	**-1.3**	**-13.5**	**17.0**	**8.9**	**.03**	**.02**
	CA	**-.053**	**3.3**	**-.80**	.20	-.71	-.35
HSLF	D	**-.072**	**4.6**	**-1.3**	0.02	-1.3	.35
	CP	**-2.2**	**-60.3**	**-15.2**	**-9.7**	**.03**	**-.006**
	CA	**-.1**	**6.3**	**2.4**	0.05	**23.1**	**-6.1**
LSLF	D	**-.78**	1.0	-.33	**-1.3**	**.02**	**.001**
	CP	**-.70**	**-72.3**	**-16.3**	**-3.6**	**.03**	**-.001**
t-values		218.9	7.9	-21.2	6.5	2.6	-6.4

NOTE: Positive values indicate BLASE is more accurate than GM, and negative values indicate the opposite. Bold values indicate statistically significant results (paired t-test p-value $< .05$). Entries under each predictor are averages across 100 replications of $100\left(|\hat{\eta}_{GM} - \eta| - |\hat{\eta}_{BLASE} - \eta|\right)/\eta$. The final row is the average t-values values of the Y_2 linking analysis across the 100 PB data sets. The average R^2 value is .14.

B.1 High fault

The high fault (HSHF) scenario is where we expect the largest payoffs from using BLASE over GM. BV faults necessarily correspond to a reduction in PMR for GM, as records in faulty blocks cannot be linked. While some increase in PMR in BLASE over GM is expected due to chance, Figure B.1 shows a sizeable increase in PMR at all three prior settings. The greatest improvement in PMR corresponds to the CA prior, corresponding to a large number of move proposals at each iteration of the MCMC. With a high fault level, many records need to shift blocks, so the large number of proposals is advantageous. The diffuse prior setting performs comparably to the CA setting. As discussed in the main paper, this trend is evident across simulations. The increase in PMR for BLASE in the HSHF scenario is accompanied by a reduction in RMSE (Figure B.2) and more accurate estimates of three of the four linking parameters (Figure B.3).

The role of T_1 seeds in the posterior performance of BLASE and GM is highlighted in the contrast of the results of the HSHF and LSHF simulation groups. With low seeds and high faults, both GM and BLASE perform poorly. While there is statistically significant improvement in BLASE over GM for PMR (Figure B.4), RMSE (Figure B.5) and parameter estimation (Figure B.6) for the diffuse prior, this improvement is markedly less than that observed in the HS simulation.

In contrast to the HF results, the LSHF scenario associates the best results with the poorly concentrated rather than appropriately concentrated prior setting. This is reasonable here. The CA setting corresponds with many move proposals at each iteration, yet the low seeds provide little signal with which to accept or reject these moves. The CP prior setting favors only a few move proposals, and therefore fewer records are moved into incorrect blocks. The weak prior setting performs comparably to the CP setting.

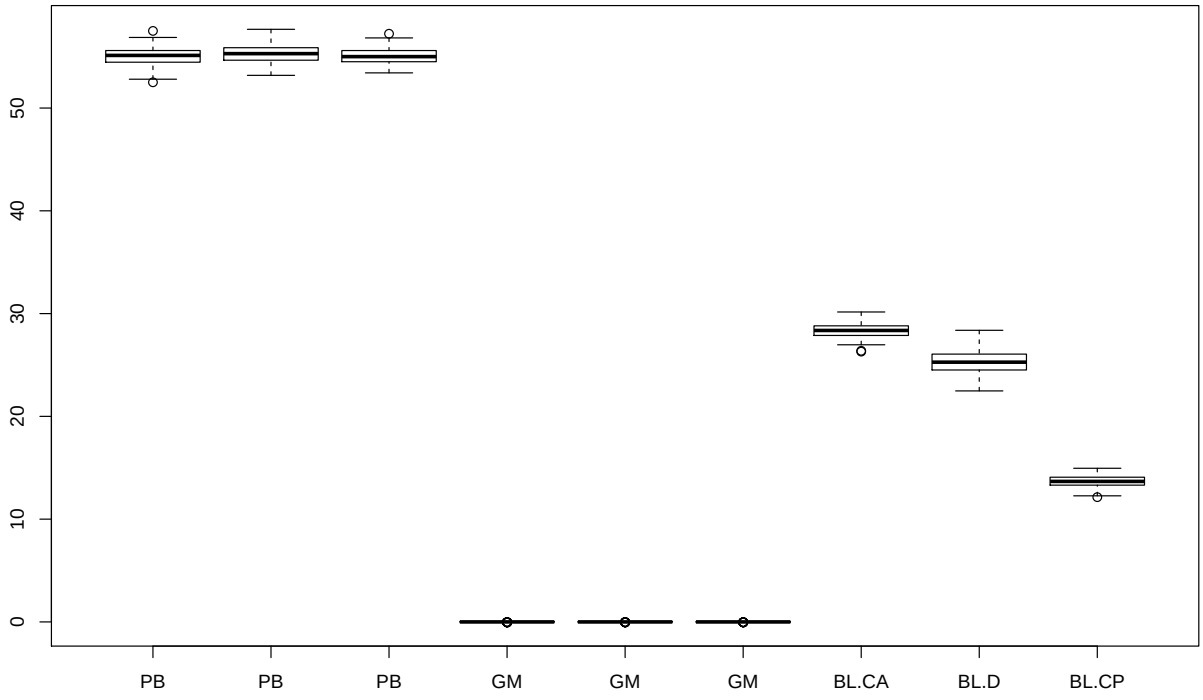


FIGURE B.1: HSHF Posterior Match Rate: Average PMR for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these PMR values.

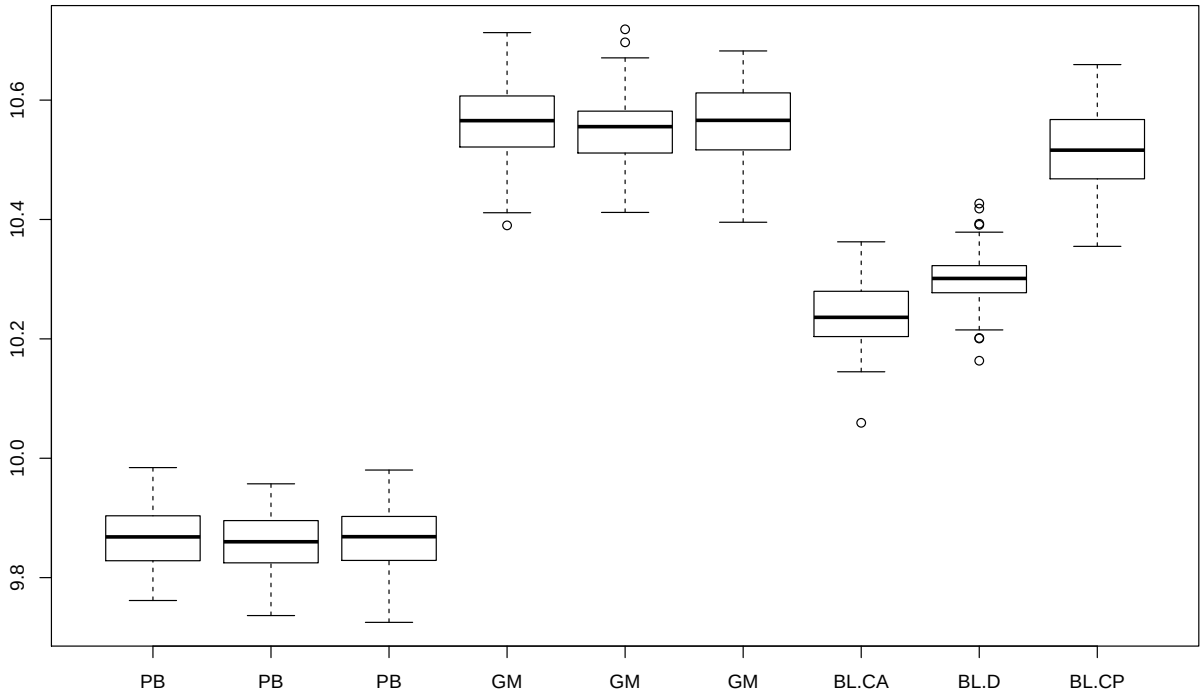
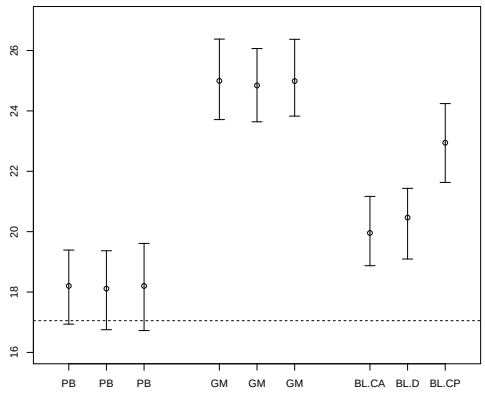
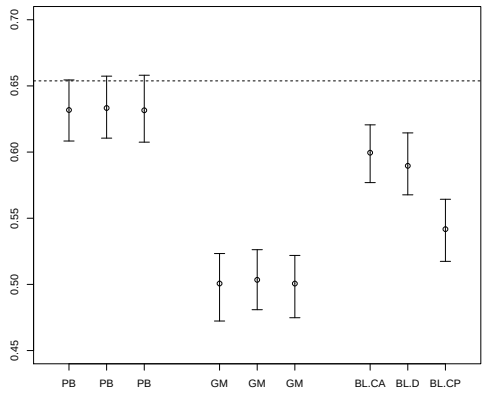


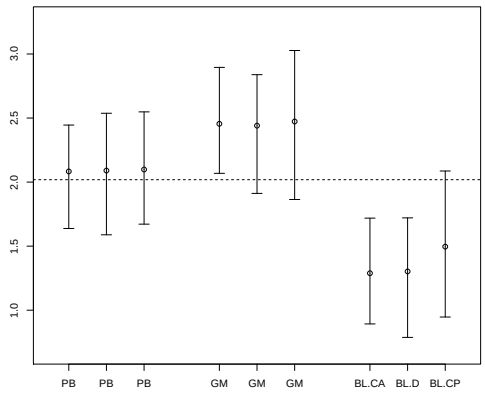
FIGURE B.2: HSHF RMSE: RMSE for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these RMSE values.



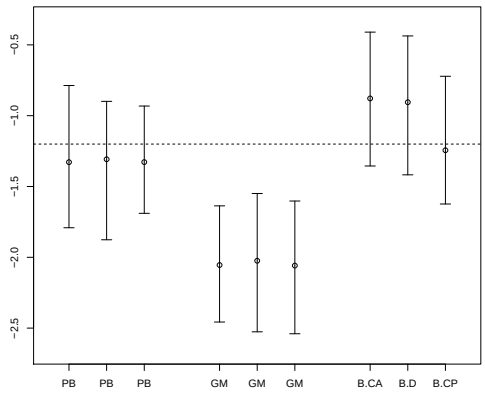
(a) Intercept



(b) Math



(c) Academic



(d) Vocational

FIGURE B.3: HSHF empirical distributions of posterior means: Posterior means for coefficients for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these posterior means. Dashed horizontal lines indicate true values of coefficients.

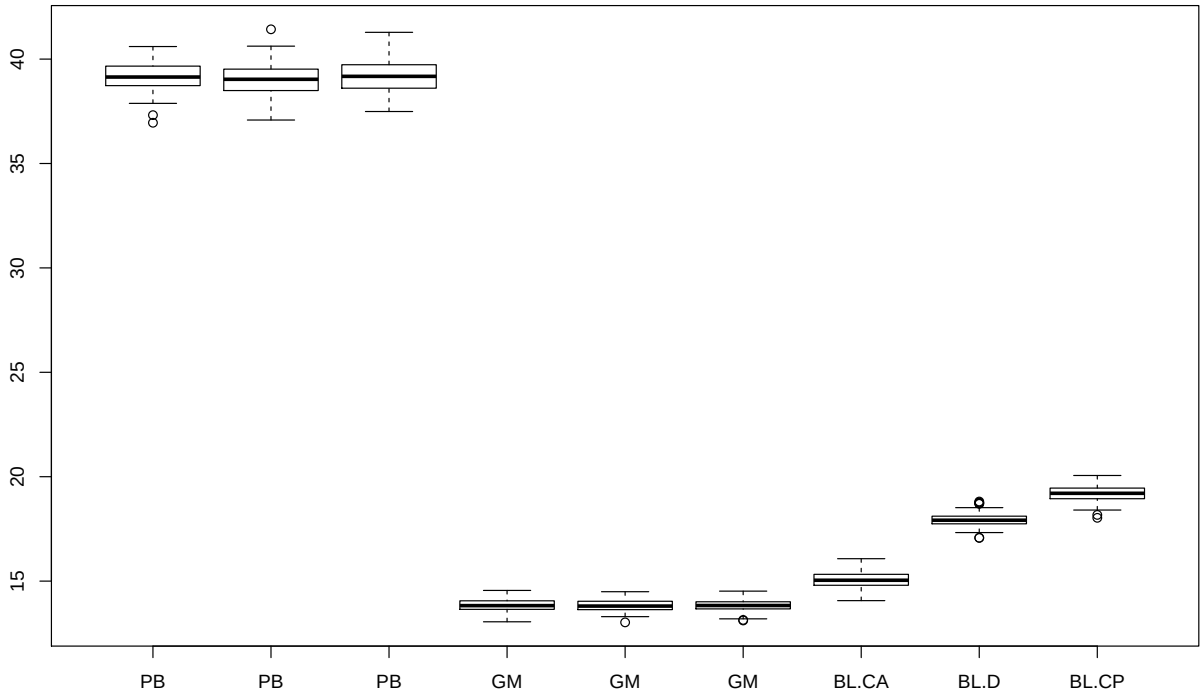


FIGURE B.4: LSHF Posterior Match Rate: Average PMR for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these PMR values.

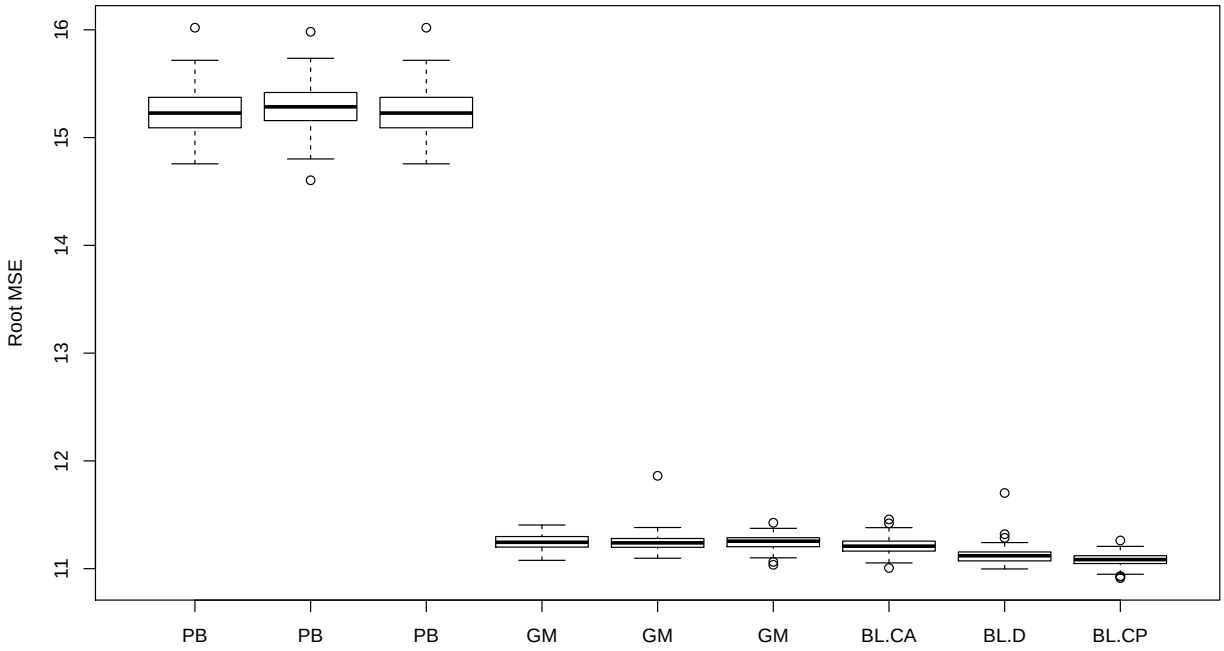
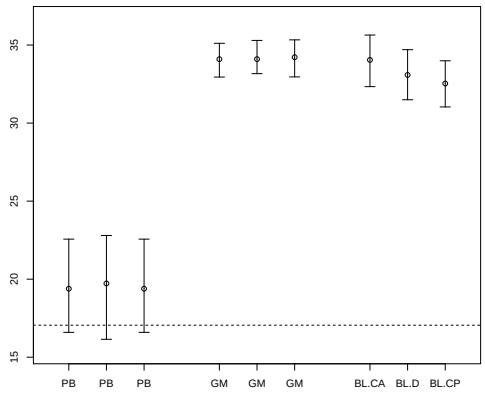
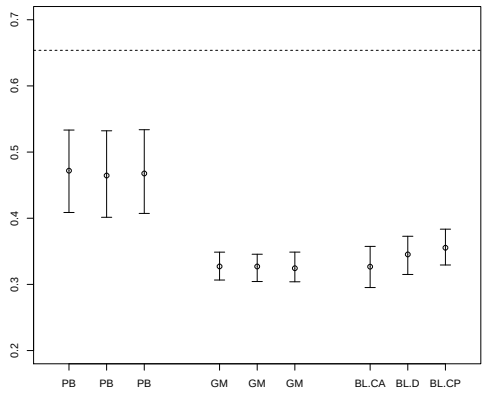


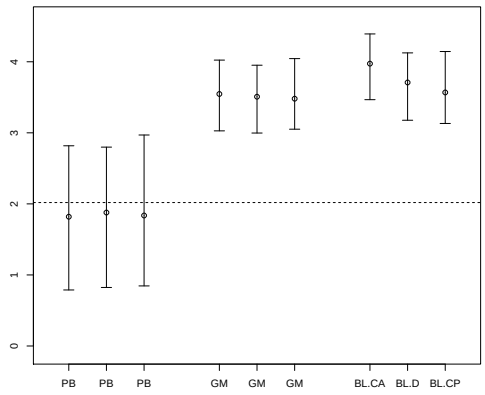
FIGURE B.5: LSHF RMSE: RMSE for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these RMSE values.



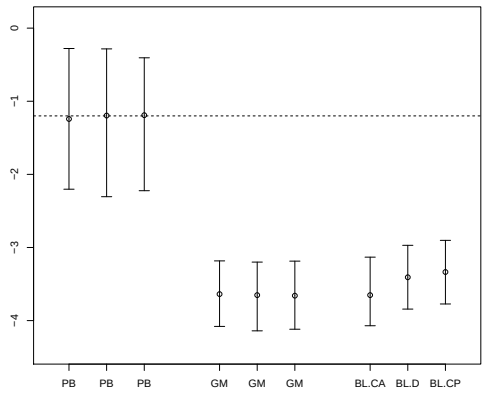
(a) Intercept



(b) Math



(c) Academic



(d) Vocational

FIGURE B.6: LSHF empirical distributions of posterior means: Posterior means for coefficients for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these posterior means. Dashed horizontal lines indicate true values of coefficients.

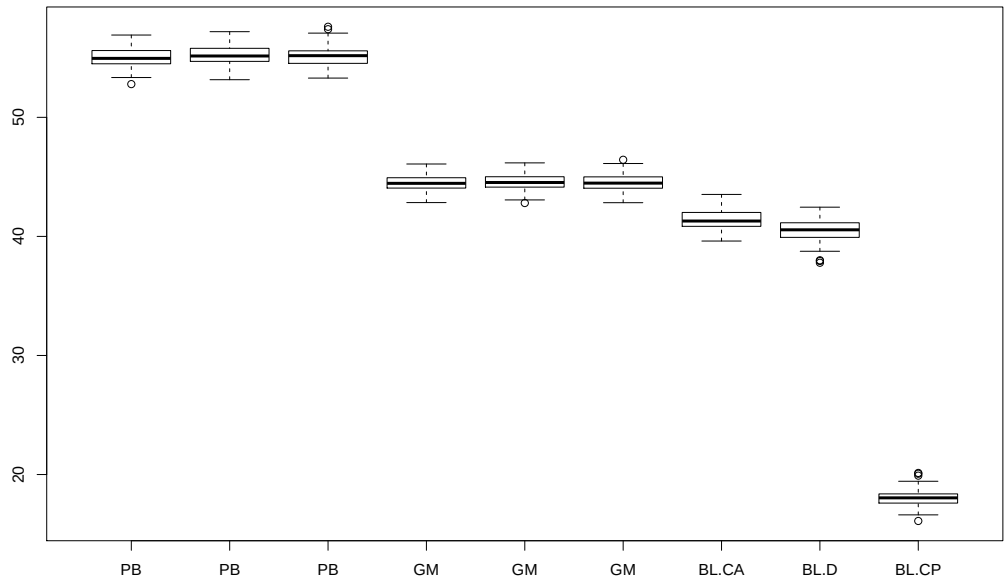


FIGURE B.7: HSLF Posterior Match Rate: Average PMR for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these PMR values.

B.2 Low fault

The two Low Fault (LF) simulation groups provide the least opportunity for BLASE to show improvement over GM. Indeed, the HSLF scenario is the closest of the four groups to be ideal for GM. Accordingly, Figure B.7 indicates a GM PMR which is much more similar to the PB results than those displayed in the HSHF runs of Figure B.1. For the CA and D prior settings, this corresponds to an increase in PMR relative to BLASE. GM and BLASE RMSE results are comparable with reasonable prior choice (Figure B.8). This suggests that while records may be moved to blocks which do not contain true matches, these moves preserve the relationships in the linking analysis as informed by the seeds. Figure B.9 suggests that at the W and CA settings, BLASE and GM perform similarly in estimating the coefficients, with BLASE obtaining a slight (but significant) improvement.

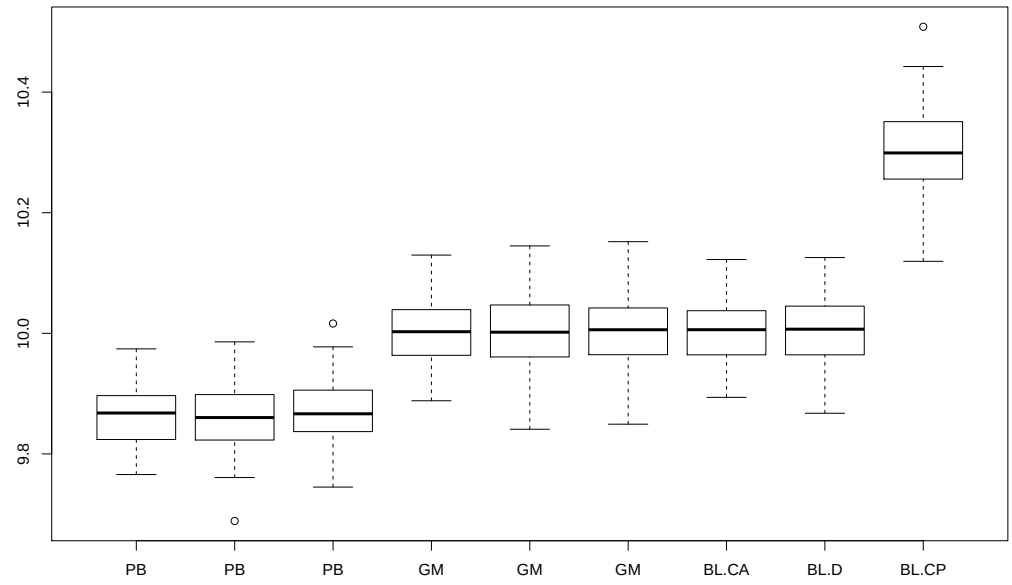
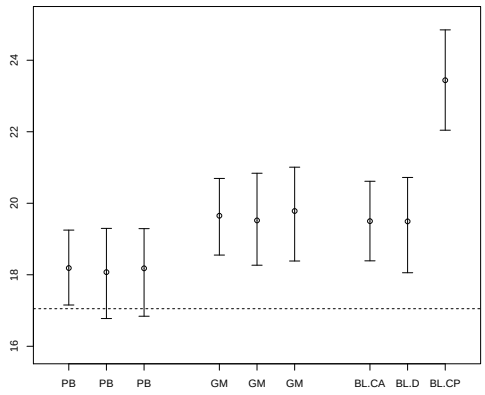
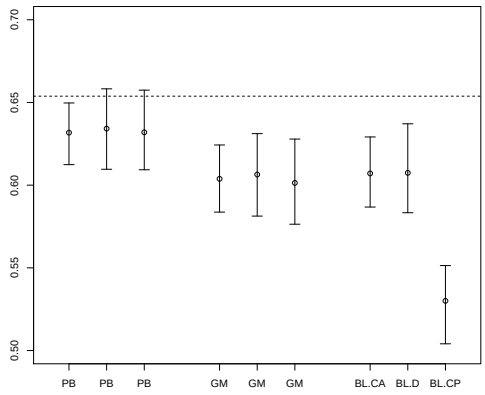


FIGURE B.8: HSLF RMSE: RMSE for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these RMSE values.

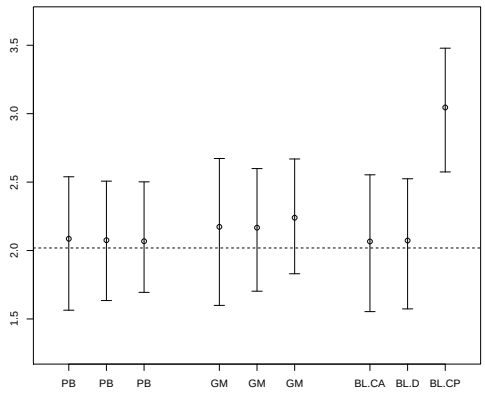
The final simulation group, LSLF, is designed to illustrate the performance of GM and BLASE in a situation not ideal for either method. Accordingly, both methods perform poorly as compared to the HSLF simulations. Here, applying BLASE as opposed to GM results in lower PMR (Figure B.10), increased RMSE (Figure B.11), and poor estimation of the parameters (Figure B.12).



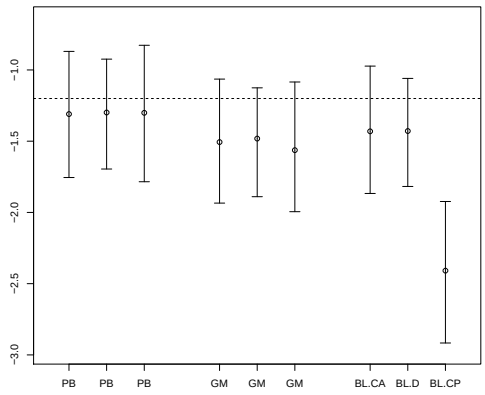
(a) Intercept



(b) Math



(c) Academic



(d) Vocational

FIGURE B.9: HSLF empirical distributions of posterior means: Posterior means for coefficients for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these posterior means. Dashed horizontal lines indicate true values of coefficients.

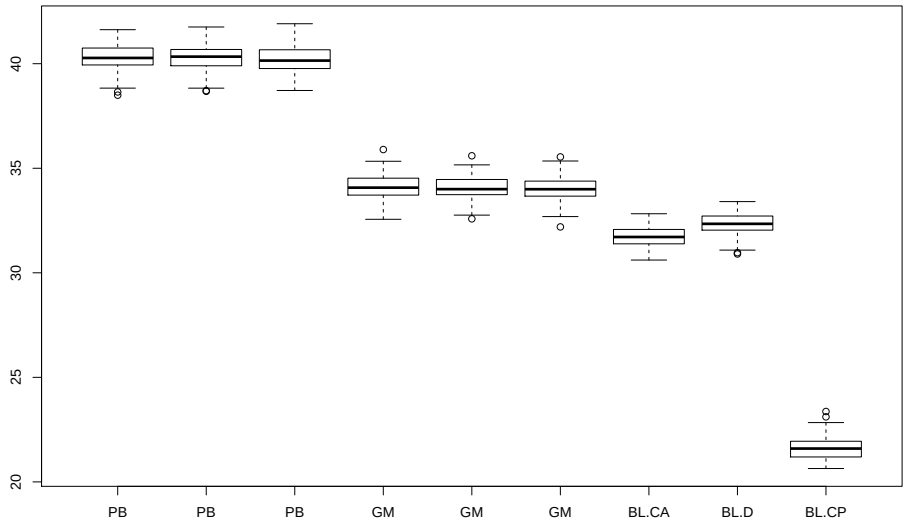


FIGURE B.10: LSLF Posterior Match Rate: Average posterior match rate for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these PMR values.

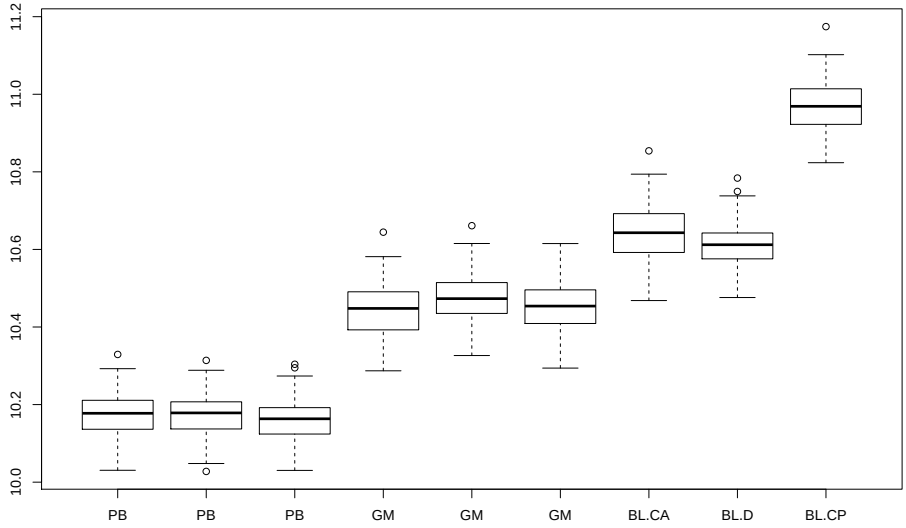
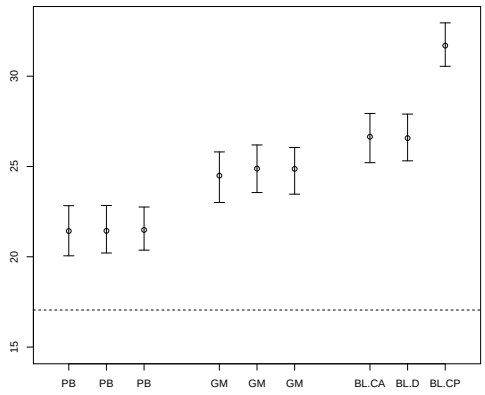
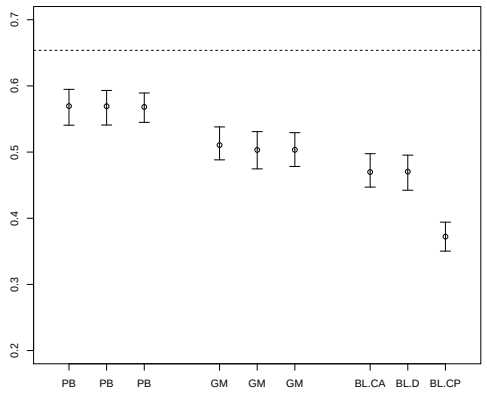


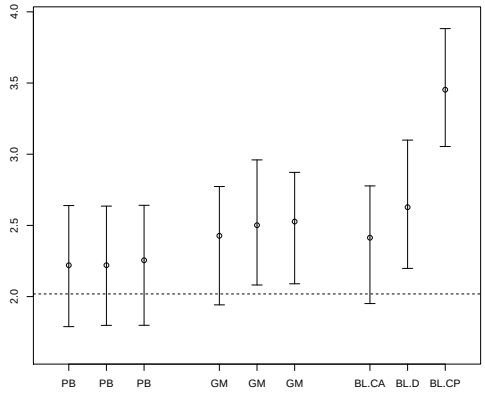
FIGURE B.11: LSLF RMSE: RMSE for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these RMSE values.



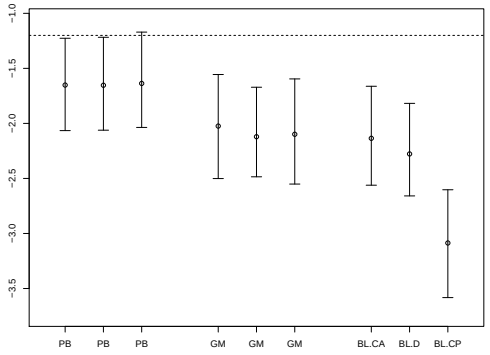
(a) Intercept



(b) Math



(c) Academic



(d) Vocational

FIGURE B.12: LSLF empirical distributions of posterior means: Posterior means for coefficients for each of 100 replications at each prior setting (Concentrated/Appropriate, Diffuse, Concentrated/Poor). Displayed intervals contain the central 95% of these posterior means. Dashed horizontal lines indicate true values of coefficients.

Bibliography

Allcott, H. and Greenstone, M. (2012), "Is there an energy efficiency gap?" *Journal of Economic Perspectives*, 26, 3–28.

Belin, T. and Rubin, D. (1995), "A method for calibrating false-match rates in record linkage," *Journal of the American Statistical Association*, 90, 694–707.

Bohensky, M. (2015), "Bias in data linkage studies," in Harron et al. (2015), pp. 63–82.

Burgette, L. F. and Reiter, J. P. (2010), "Multiple imputation for missing data via sequential regression trees," *American Journal of Epidemiology*, p. 260.

Chambers, R. (2009), "Regression analysis of probability-linked data," Statisphere 4, Official Statistics Research Series, Statistics New Zealand.

Chambers, R. and Kim, G. (2015), "Secondary analysis of linked data," in Harron et al. (2015), pp. 83–108.

Dalzell, N. M. and Reiter, J. P. (2016), "Regression modeling and file matching using possibly erroneous matching variables," *ArXiv e-prints*.

Domingo-Ferrer, J. and Torra, V. (2002a), "Distance-based and probabilistic record linkage for re-identification of records with categorical variables," *Butlletí de lA-CIA, Associació Catalana dIntelligència Artificial*, pp. 243–250.

Domingo-Ferrer, J. and Torra, V. (2002b), "Validating distance-based record linkage with probabilistic record linkage," in *Topics in Artificial Intelligence*, pp. 207–215, Springer.

Domingo-Ferrer, J. and Torra, V. (2003), "Disclosure risk assessment in statistical microdata protection via advanced record linkage," *Statistics and Computing*, 13, 343–354.

Dunson, D. B. and Xing, C. (2009), "Nonparametric bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Escobar, M. and West, M. (1995), "Estimating normal means with a dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Fellegi, I. P. and Sunter, A. B. (1969), "A theory for record linkage," *Journal of The American Statistical Association*, 64, 1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001), "On Bayesian record linkage," *Research in Official Statistics*, 4, 185–198.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian data analysis*, Chapman & Hill, London.

Gerarden, T. D., Newell, R. G., and Stavins, R. N. (2015), "Assessing the energy-efficiency gap," Working Paper 20904, National Bureau of Economic Research.

Gutman, R., Afendulis, C., and Zaslavsky, A. (2013), "A Bayesian procedure for file linking to analyze end-of-life medical costs," *Journal of the American Statistical Association*, 18, 34–47.

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

Harron, K., Goldstein, H., and Dibben, C. (2015), *Methodological developments in data linkage*, Wiley Series in Probability and Statistics, Wiley.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Jaro, M. A. (1989), "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Karakasidis, A. and Verykios, V. S. (2012), "Reference table based K-anonymous private blocking," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pp. 859–864, New York, NY, USA, ACM.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple imputation of missing or faulty values under linear constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015), "Simultaneous edit-imputation for continuous microdata," *Journal of the American Statistical Association*, 110, 987–999.

Kim, H. S. and Lee, D. (2010), "HARRA: fast iterative hashed record linkage for large-scale data collections," in *Proceedings of the 13th International Conference on Extending Database Technology*, EDBT '10, pp. 525–536, New York, NY, USA, ACM.

Kohnen, C. N. and Reiter, J. P. (2009), "Multiple imputation for combining confidential data owned by two agencies," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 172, 511–528.

Lahiri, P. and Larsen, M. (2005), "Regression analysis with linked data," *Journal of the American Statistical Association*, 100, 222–230.

Larsen, M. (2004), "Record linkage using finite mixture models," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, eds. A. Gelman and X. Meng, pp. 309–318, Chichester, UK: Wiley.

Larsen, M. and Rubin, D. (2001), "Iterative automated record linkage using mixture models," *Journal of the American Statistical Association*, 96, 32–41.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, John Wiley & Sons, Inc.

Manrique-Vallier, D. and Reiter, J. P. (2016), "Bayesian simultaneous edit and imputation for multivariate categorical data," Tech. rep., Dept. of Statistical Science, Duke University.

McGlincy, M. H. (2004), "A Bayesian record linkage methodology for multiple imputation of missing links," in *ASA Proceedings of the Joint Statistical Meetings*, pp. 4001–4008, American Statistical Association.

Meng, X. and Rubin, D. (1993), "Maximum likelihood via the EXM algorithm: a general framework," *Biometrika*, 80, 267–278.

Muller, M. B. (2001), "IAC database manual," Tech. rep., U.S. Department of Energy, Center for Advanced Energy Studies, Rutgers University.

NCERDC (2013), "North Carolina Education Research Data Center: Duke Center for Child and Family Policy," `https://childandfamilypolicy.duke.edu/pdfs/projects/NCERDC_DataHoused.pdf`.

Neter, J., Maynes, E., and Ramanthan, R. (1965), "The effect of mismatching on the measurement of response errors," *Journal of the American Statistical Association*, 90, 1005–1027.

Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959), "Automatic linkage of vital records," *Science*, 130, 954–959.

Newcombe, H. B. (1988), *Handbook of record linkage: methods for health and statistical studies, administration, and business*, Oxford University Press, Inc., New York, NY, USA.

Pagliuca, D. and Seri, G. (1999), "Some results of individual ranking method on the system of enterprise accounts annual survey," *Esprit SDC Project, Deliverable MI-3/D2*.

Reiter, J. P. (2004), "Simultaneous use of multiple imputation for missing data and disclosure limitation," *Survey Methodology*, 30, l242.

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

Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, Wiley, New York, USA.

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

Rubin, D. B. (2003), "Nested multiple imputation of NMES via partially incompatible MCMC," *Statistica Neerlandica*, 57, 3–18.

Sadinle, M. (2016), "Bayesian estimation of bipartite matchings for record linkage," *Journal of the American Statistical Association*, p. forthcoming.

Samart, K. and Chambers, R. (2014), "Linear regression with nested errors using probability-linked data," *Australian & New Zealand Journal of Statistics*, 56, 27–46.

Schafer, J. L. and Graham, J. W. (2002), "Missing data: our view of the state of the art," *Psychological Methods*, 7, 147–177.

Scheuren, F. and Winkler, W. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39–58.

Scheuren, F. and Winkler, W. (1997), "Regression analysis of data files that are computer matched, part II," *Survey Methodology*, 23, 157–165.

Sethuraman, J. (1994), "A constructive definition of dirichlet priors," *Statistica Sinica*, 4, 639–650.

Shen, Z. (2000), "Nested imputation," Ph.D. thesis, Harvard University, Cambridge, MA.

Si, Y. and Reiter, J. P. (2013), "Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys," *Journal of Educational and Behavioral Statistics*, 38, 499–521.

Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. (2014a), "A comparison of blocking methods for record linkage," *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 8744, 253–268.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014b), "SMERED: a Bayesian approach to graphical record linkage and de-duplication," *Journal of Machine Learning Research*, 33, 922–930.

Tancredi, A. and Liseo, B. (2011), "A hierarchical bayesian approach to record linkage and population size problems," *The Annals of Applied Statistics*, 5, 1553–1585.

Thibaudeau, Y. (1993), "The discrimination power of dependency structures in record linkage," *Survey Methodology*, 19, 31–38.

Torra, V., Abowd, J. M., and Domingo-Ferrer, J. (2006), "Using mahalanobis distance-based record linkage for disclosure risk assessment," in *International Conference on Privacy in Statistical Databases*, pp. 233–242, Springer.

UCLA: Stat Consulting Group (2016), "R data analysis examples: multinomial logistic regression," `http://www.ats.ucla.edu/stat/r/dae/mlogit.htm`, Accessed: 2015-09-02.

U.S. Census Bureau (2016), "North American Industry Classification System," `http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007`.

U.S. Department of Energy (2016), "Industrial Assessment Centers (IAC)," `https://energy.gov/eere/amo/industrial-assessment-centers-iacs`.

Vatsalan, D. and Christen, P. (2013), *Sorted nearest neighborhood clustering for efficient private blocking*, pp. 341–352, Springer Berlin Heidelberg, Berlin, Heidelberg.

Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A., and Sijtsma, K. (2008), "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, 38, 369–397.

White, T. K. (2014), "Recovering the item-level edit and imputation flags in the 1977-1997 Censuses of Manufactures," *US Census Bureau Center for Economic Studies Paper No. CES-WP-14-37*, pp. 14–37.

White, T. K., Reiter, J. P., and Petri, A. (2015), "Imputation in U.S. manufacturing data and implications for within-industry productivity dispersion," in *2015 Federal Committee on Statistical Methodology Research Conference, At Washington, DC.*

Winkler, W. (1993a), "Advances in record linkage," in *Proceedings of the Section on Survey Research Methods*, pp. 464–472, American Statistical Association.

Winkler, W. (1993b), "Improved decision rules in the Fellegi-Sunter model of record linkage," in *Proceedings of the Section on Survey Research Methods*, pp. 274–279, American Statistical Association.

Winkler, W. (2004), "Methods for evaluating and creating data quality," *Information systems*, 29, 531–550.

Winkler, W. E. (1988), "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667–671.

Winkler, W. E. (1999), "The state of record linkage and current research problems," Tech. rep., Statistical Research Division, U.S. Census Bureau.

Winkler, W. E. (2015), "Probabilistic linkage," in Harron et al. (2015), pp. 8–35.

Winkler, W. E. and Thibaudeau, Y. (1987), "An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Decennial Census," in *U.S. Decennial Census: Technical report, US Bureau of the Census*.

Wu, Y. (1995), "Random shuffling: a new approach to matching problems," in *ASA Proceedings of the Statistical Computing Section*, pp. 69–74, American Statistical Association.

Zhou, X. and Reiter, J. P. (2010), "A note on Bayesian inference after multiple imputation," *The American Statistician*, 64, 159–163.

Biography

Nicole Marie Dalzell was born in Florida. She attended the University of South Florida and received her Bachelor of Arts in Mathematics, Summa Cum Laude, with a minor in French, in 2012. She was a successful graduate of the University of South Florida Honors College in the same year. After that, she attended Duke University in Durham, North Carolina, where she received her Masters en route to her Ph.D. in December 2014. She plans to graduate with her Ph.D. in May 2017, under the supervision of Dr. Jerome P. Reiter.

Nicole was awarded a 2017 Dean's Award for Excellence in Teaching. Nicole won the Best Graduate Presentation Award at the AISC Conference in 2014. Dalzell has served as a teaching assistant for six classes, five in-person and one through an online Coursera course. Through a Bass Online Instructional Fellowship, she helped transform this Coursera course into an online continuing education course, which she also helped teach. She received the Teaching Assistant of the Year Award from the Statistical Science Department in 2016, with a unanimous vote from the selection committee. Nicole has also taught her own course, Data Analysis and Statistical Inference, to undergraduates at Duke for two summers. She is a participant in The Graduate Schools Certificate in College Teaching and Preparing Future Faculty Program. In addition to her classroom teaching, Nicole successfully co-organized the 2016 DataFest at Duke, a three-day data analysis competition for over 300 undergraduates.