Statistical Methods for Human Rights and Child Mortality Estimation

Serge Aleshin-Guendel

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Jon Wakefield, Chair
Mauricio Sadinle, Chair
Abel Rodriguez

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Statistical Methods for Human Rights and Child Mortality Estimation

Serge Aleshin-Guendel

Co-Chairs of the Supervisory Committee:
Jon Wakefield
Departments of Statistics and Biostatistics

Mauricio Sadinle
Department of Biostatistics

This dissertation addresses statistical methodology commonly used in human rights research and child mortality estimation. We first consider two related problems, record linkage and multiple-systems estimation, typically used to estimate the number of civilian casualties in the wake of a conflict when probability surveys are not available, and then consider the problem of estimating child mortality over time in a country that has experienced conflict. In Chapter 2, we propose a novel Bayesian approach for record linkage in the general setting where there may be any number of files, with arbitrary patterns of duplication across files. In Chapter 3, we present a re-framing of multiple-systems estimation which places identifying assumptions front and center in the multiple-systems estimation workflow, and examine how common models fit into this framing. In Chapter 4, we develop spatial and temporal smoothing models which incorporate knowledge of expected shocks in child mortality, such as the timing of a conflict, leading to estimates of child mortality which are not oversmoothed. Finally, we conclude with discussion of future work in Chapter 5.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables vii

Chapter 1: Introduction 1

Chapter 2: Multifile Partitioning for Record Linkage and Duplicate Detection . . . 3
- 2.1 Introduction 3
- 2.2 Multifile Partitioning 5
- 2.3 A Structured Prior for Multifile Partitions 8
- 2.4 A Model for Comparison Data 13
- 2.5 Bayesian Estimation of Multifile Partitions 15
- 2.6 Simulation Studies 19
- 2.7 Discussion and Future Work 23

Chapter 3: The Central Role of the Identifying Assumption in Population Size Estimation 25
- 3.1 Introduction 25
- 3.2 Multiple-Systems Estimation as a Missing Data Problem 26
- 3.3 Log-Linear and Latent Class Models 32
- 3.4 Revisiting Log-Linear Models and Their Identifying Assumptions 35
- 3.5 Civilian Casualties in the Kosovo War 38
- 3.6 Discussion 44

Chapter 4: Adaptive Gaussian Markov Random Fields for Child Mortality Estimation 45
- 4.1 Introduction 45
- 4.2 Background and Motivating Application 46
- 4.3 Adaptive Gaussian Markov Random Fields 54

4.4 Scaling, Reparameterizations, Prior Choice, and Computation 60
4.5 Simulations 64
4.6 Estimation of U5MR at the National Level in Rwanda 66
4.7 Estimation of U5MR across Multiple Countries 71
4.8 Conclusions 81

Chapter 5: Discussion and Future Work 85

Appendix A: Appendix for Chapter 1 104
A.1 Structured Prior Appendix 104
A.2 Posterior Inference Appendix 108
A.3 Point Estimation Appendix 114
A.4 Simulation Appendix 116
A.5 Colombia Application 139

Appendix B: Appendix for Chapter 2 146
B.1 Conditional Identifiability in Models for Heterogeneity 146
B.2 Computation for Conditionally Identified Models 152
B.3 Identifying Assumption Derivations 164
B.4 Latent Class Model Simulations 165
B.5 Kosovo Analysis Appendix 176

LIST OF FIGURES

Figure Number Page

2.1 A toy example of the multifile record linkage and duplicate detection problem. 5

2.2 An illustration of a multifile partition of [12], where $\boldsymbol{X}_1$ contains records $1-5$ and $\boldsymbol{X}_2$ contains records $6-12$. 9

2.3 Performance comparison for simulation with equal measurement error across files. Black lines refer to results under our structured prior, grey lines to results under the flat prior, solid lines show medians, and dashed lines show 2nd and 98th percentiles. 22

2.4 Performance comparison for simulation with unequal measurement error across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings. Dots show medians, and bars show 2nd and 98th percentiles. 24

3.1 Likelihood surface of L_1 when $n = 100$. 29

4.1 Top Panel: Direct estimates of U5MR for Rwanda from six DHS surveys, for the 15 years prior to each survey. Bottom Panel: U5MR estimates from IGME and a meta-analysis estimator of U5MR based on the direct estimates in the top panel. 52

4.2 Left Panel: Direct estimates of U5MR for Burundi, Ethiopia, Kenya, Rwanda, Tanzania, and Uganda. Right Panel: Zoomed in direct estimates of U5MR for Burundi and Rwanda. 53

4.3 The three trends used for μ_i to simulate data. 66

4.4 RMSE for simulation settings where τ_i is the same for all time points. 67

4.5 RMSE for simulation settings where τ_i is not the same for all time points. . . 68

4.6 DIC for simulation settings where τ_i is the same for all time points. 69

4.7 DIC for simulation settings where τ_i is not the same for all time points. . . . 70

4.8 LS for simulation settings where τ_i is the same for all time points. 71

4.9 LS for simulation settings where τ_i is not the same for all time points. 72

4.10 Comparison of prior and posterior density for θ in the Rwanda application. . 74

4.11 Comparison of U5MR estimates from the smoothed direct model, the proposed model, IGME, and meta-analysis estimator of U5MR based on the direct estimates. 75

4.12 Comparison of prior and posterior density for θ the proposed model for the multi-country application. 78

4.13 Maps of U5MR estimates from the smoothed direct model and the proposed model. 79

4.14 Comparison of U5MR estimates from the smoothed direct model and the proposed model, in addition to the country-specific smoothed direct model fits, direct estimates for each Admin1 region in black, and direct estimates for each country in red. 80

4.15 Comparison of prior and posterior density for θ the proposed country-intercept model for the multi-country application. 82

4.16 Maps of U5MR estimates from the smoothed direct and proposed country-intercept models. 83

4.17 Comparison of U5MR estimates from the smoothed direct and proposed country-intercept models, in addition to the country-specific smoothed direct model fits, direct estimates for each Admin1 region in black, and direct estimates for each country in red. 84

A.1 Performance comparison for no-three-file overlap simulation with more informative settings of $\boldsymbol{\alpha}$. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ with $\kappa \in \{49, 99\}$. 118

A.2 Performance comparison for simulation with datafiles with duplicates and full estimates. Black lines refer to results under our structured prior, grey lines refer to the approach of [114], solid lines show medians, and dashed lines show 2nd and 98th percentiles. 119

A.3 Performance comparison for simulation with datafiles with duplicates and full estimates, with varying priors for the within-file cluster sizes. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. 121

A.4 Performance comparison for simulation with datafiles with duplicates and partial estimates. Black solid and dashed lines refer to precision for partial estimates, grey solid and dashed lines refer to precision for full estimates, and dot-dashed and dotted lines refer to the abstention rate for partial estimates. Solid and dot-dashed lines show medians, and dashed and dotted lines show 2nd and 98th percentiles. 122

A.5 Bias estimates for simulation with duplicate-free files and equal errors across files. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ discussed in Appendix A.4.2. 123

A.6 Mean-squared error estimates for simulation with duplicate-free files and equal errors across files. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ discussed in Appendix A.4.2. 124

A.7 Bias estimates for simulation with duplicate-free files and unequal errors across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings. 125

A.8 Mean-squared error estimates for simulation with duplicate-free files and unequal errors across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings. . . . 126

A.9 Bias estimates for simulation with files with duplicates and equal errors across files. "Sadinle (2014)" refers to the approach of [114] and "lambda=..." refers to the proposed approach, varying the prior over within-file cluster sizes. . . 126

A.10 Mean-squared error estimates for simulation with files with duplicates and equal errors across files. "Sadinle (2014)" refers to the approach of [114] and "lambda=..." refers to the proposed approach, varying the prior over within-file cluster sizes. 127

A.11 Average running time for simulation varying the number of latent entities. . 129

A.12 Performance comparison across different loss function specifications for simulation with duplicate-free files and equal measurement error across files. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. 132

A.13 Performance comparison across different loss function specifications for simulation with duplicate-free files and unequal measurement error across files. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. . 133

A.14 Performance comparison across different loss function specifications for simulation with files with duplicates and equal measurement error across files, when using full estimates. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. 134

A.15 Performance comparison across different loss function specifications for simulation with files with duplicates and equal measurement error across files, when using partial estimates. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. 135

A.16 Trace plots for n for the last 5 of 100 runs for the simulation with duplicate-free files and equal measurement error across files. "Proposed" refers to the proposed approach and "Flat" refers to the approach using a flat prior for tripartite matchings. 136

A.17 Trace plots for n for the last 5 of 100 runs for the simulation with duplicate-free files and unequal measurement error across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings. 137

A.18 Trace plots for n for the last 5 of 100 runs for the simulation with files with duplicates and equal measurement error across files, when using full estimates. "Proposed" refers to the proposed approach and "Sadinle (2014)" refers to the approach of [114]. 138

A.19 Trace plots for n in Colombia application. 145

B.1 Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption. 179

B.2 Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption. 181

LIST OF TABLES

Table Number Page

3.1 Kosovo dataset, reproduced from Section 6 of [9]. 38

3.2 Point estimates and 95% uncertainty intervals for N under the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. 41

3.3 Point estimates and 95% uncertainty intervals for N under the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. . . 41

3.4 Point estimates and 95% uncertainty intervals for sensitivity analysis probing the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a marginal odds ratio, as described in Section 3.4.3. 43

4.1 Comparison of parameter estimates for the smoothed direct and proposed models in the Rwanda application. 73

4.2 Comparison of parameter estimates from fitting the smoothed direct model to each country separately. 76

4.3 Comparison of parameter estimates for the smoothed direct and proposed models in the multi-country application. 77

4.4 Comparison of parameter estimates for the smoothed direct and proposed country-intercept models in the multi-country application. 81

A.1 Types of errors per field in the simulation studies. 116

A.2 Construction of levels of disagreement for the simulation studies. 117

A.3 Average running time in seconds for proposed approach in simulations with duplicate-free files and equal errors across files. 127

A.4 Average running time in seconds for proposed approach in with files with duplicates and equal errors across files. 128

A.5 Construction of levels of disagreement for the Colombian homocide record systems. 140

A.6 Posterior distribution of the overlap table for the Colombian record systems, under the informative prior specification. Black lines indicate the ground truth, dotted lines indicate quantities derived from the full estimate of the tripartite matching. 143

A.7 Posterior distribution of the overlap table for the Colombian record systems, under the default prior specification. Black lines indicate the ground truth, dotted lines indicate quantities derived from the full estimate of the tripartite matching. 145

B.1 Catalog of $p(N \mid n, \pi_0)$ and $p(n \mid \pi_0)$ under common priors for N. 161

B.2 Parameters of two latent class models, Q_{1a} and Q_{1b} (rounded for presentation) 166

B.3 Results of the simulation study where data was generated from the two-class latent class model Q_{1a}. Truth is $\pi_{Q_{1a},0} = 0.316$. 167

B.4 Parameters of latent class model which generated data in simulation of [87]. 167

B.5 Parameters of latent class model Q_2 168

B.6 Results of the simulation study where data was generated from the two-class latent class model Q_2. Truth is $\pi_{Q_2,0} = 0.704$. 168

B.7 Parameters of latent class model Q_{3a} 169

B.8 Results of the simulation study where data was generated from the two-class latent class model Q_{3a}. Truth is $\pi_{Q_{3a},0} = 0.681$. 169

B.9 Parameters of latent class model Q_{3b} 170

B.10 Results of the simulation study where data was generated from the two-class latent class model Q_{3b}. Truth is $\pi_{Q_{3b},0} = 0.658$. 170

B.11 Parameters of latent class model which generated data in simulation of [87], with a third class added. 171

B.12 Parameters of latent class model Q_{4a} 172

B.13 Results of the simulation study where data was generated from the two-class latent class model Q_{4a}. Truth is $\pi_{Q_{4a},0} = 0.613$. 172

B.14 Parameters of latent class model Q_{4b} 173

B.15 Results of the simulation study where data was generated from the two-class latent class model Q_{4b}. Truth is $\pi_{Q_{4b},0} = 0.569$. 173

B.16 Results of the simulation study where data was generated from the two-class latent class model Q_{4c}. Truth is $\pi_{Q_{4c},0} = 0.536$. 174

B.17 Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption. 178

B.18 Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption. 180
B.19 Point estimates and 95% uncertainty intervals for sensitivity analysis probing the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a ratio of ratios of odds ratios, as described in Section 4.2 of Chapter 3 and Appendix B.5.1. 180

ACKNOWLEDGMENTS

Thank you to my advisors Mauricio Sadinle and Jon Wakefield for all of your mentorship and allowing me to work on interesting projects. Thank you to all of the students in the program that made grad school possible. Thank you to Ryan for always being a familiar face in a new city. And finally, thank you to Taylor and Alice for all of your love and support.

DEDICATION

In memory of Ken Guendel.

Chapter 1

INTRODUCTION

Assessing the extent of mortality in the wake of a conflict is an important task in human rights research and demography. This dissertation addresses the statistical methodology used in such settings. Chapters 2 and 3 address the methodology of two related problems, record linkage and multiple-systems estimation. When probability surveys are not available in the wake of a conflict, record linkage and multiple-systems estimation are typically used in conjunction to estimate the number of civilian casualties from convenience samples. Chapter 4 addresses the methodology used to estimate child mortality from probability surveys over time in a country that has experienced a conflict.

Merging datafiles containing information on overlapping sets of entities is a challenging task in the absence of unique identifiers, and is further complicated when some entities are duplicated in the datafiles. Most approaches to this problem have focused on linking two files assumed to be free of duplicates, or on detecting which records in a single file are duplicates. However, it is common in practice to encounter scenarios that fit somewhere in between or beyond these two settings. In Chapter 2 we propose a Bayesian approach for the general setting of what we refer to as multifile record linkage and duplicate detection.

The problem of estimating the size of a population based on a subset of individuals observed across multiple data sources is often referred to as capture-recapture or multiple-systems estimation (MSE). This is fundamentally a missing data problem, where the number of unobserved individuals represents the missing data. As with any missing data problem, MSE requires users to make an untestable identifying assumption in order to estimate the population size from the observed data. In Chapter 3 we present a re-framing of the MSE problem that leads to an approach which places the identifying assumption front and center

in the MSE workflow, and examine how common MSE models fit into this approach.

The under-5 mortality rate (U5MR) is an important statistic in understanding the health of a country. In lower and middle income countries, estimates of U5MR are typically based on household surveys. Reliable estimation from such surveys at fine spatio-temporal scales require the usage of smoothing models which borrow information across space and time. The assumptions of these smoothing models may not be realistic when certain time periods or regions are expected to have shocks in mortality relative to their neighbors, such as when a conflict occurs. In such settings, these models can lead to oversmoothing of U5MR estimates. In Chapter 4 we develop spatial and temporal smoothing models which incorporate knowledge of these expected shocks in mortality, leading to estimates of U5MR which are not oversmoothed.

We conclude with discussion of future work in Chapter 5.

Chapter 2

MULTIFILE PARTITIONING FOR RECORD LINKAGE AND DUPLICATE DETECTION

2.1 Introduction

When information on individuals is collected across multiple datafiles, it is natural to merge these datafiles to harness all available information. This merging requires identifying *coreferent* records, i.e., records that refer to the same entity, which is not trivial in the absence of unique identifiers. This problem arises in many fields, including public health [59], official statistics [67], political science [36], and human rights [114, 115, 10].

Most approaches in this area have thus far focused on one of two settings. *Record linkage* has traditionally referred to the setting where the goal is to find coreferent records across two datafiles, where the files are assumed to be free of duplicates. *Duplicate detection* has traditionally referred to the setting where the goal is to find coreferent records within a single file. In practice, however, it is common to encounter problems that fit somewhere in between or beyond these two settings. For example, we could have multiple datafiles that are all assumed to be free of duplicates, or we might have duplicates in some files but not in others. In these general settings, the data collection processes for the different datafiles possibly introduce different patterns of duplication, measurement error, and missingness into the records. Further, dependencies among these data collection processes determine which specific subsets of files contain records of the same entity. We refer to this general setting as *multifile record linkage and duplicate detection*.

Traditional approaches to record linkage and duplicate detection have mainly followed the seminal work of [42], by modeling comparisons of fields between pairs of records in a mixture model framework [141, 67, 75]. These approaches work under, and take advantage

of, the intuitive assumption that coreferent records will look similar, and non-coreferent records will look dissimilar. However, these approaches output independent decisions for the coreference status of each pair of records, necessitating the use of ad hoc post-processing steps to reconcile incompatible decisions that ignore the logical constraints of the problem.

Our approach to multifile record linkage and duplicate detection builds on previous Bayesian approaches where the parameter of interest is defined as a partition of the records. These Bayesian approaches have been carried out in two frameworks. In the *direct-modeling* framework, one directly models the fields of information contained in the records [91, 131, 82, 125, 126, 132, 90, 37], which requires a custom model for each type of field. While this framework can provide a plausible generative model for the records, it can be difficult to develop custom models for complicated fields like strings, so most approaches are limited to modeling categorical data, with some exceptions [82, 125]. In the *comparison-based* framework, following the traditional approaches, one models comparisons of fields between pairs of records [48, 74, 114, 115]. By modeling comparisons of fields, instead of the fields directly, a generic modeling approach can be taken for any field type, as long as there is a meaningful measure of similarity for that field type.

[117] generalized [42] by linking $K > 2$ files with no duplicates. However, in addition to inheriting the issues of traditional approaches, their approach does not scale well in the number of files or the file sizes encountered in practice. [126] presented a Bayesian approach in the direct-modeling framework for the general setting of multifile record linkage and duplicate detection, which has been extended by [125] and [90]. This approach uses a flat prior on arbitrary labels of partitions, which incorporates unintended prior information.

In light of the shortcomings of existing approaches, we propose an extension of Bayesian comparison-based models that explicitly handles the setting of multifile record linkage and duplicate detection. We first present in Section 2.2 a parameterization of partitions specific to the context of multifile record linkage and duplicate detection. Building on this parameterization, in Section 2.3 we construct a structured prior for partitions that can incorporate prior information about the data collection processes of the files in a flexible manner. As

a by-product, a family of priors for K-partite matchings is constructed. In Section 2.4 we construct a likelihood function for comparisons of fields between pairs of records that accommodates possible differences in the datafile collection processes. In Section 2.5 we present a family of loss functions that we use to derive Bayes estimates of partitions. These loss functions have an *abstain option* which allow portions of the partition with large amounts of uncertainty to be left unresolved. Finally, we explore the performance of our proposed methodology through simulation studies in Section 2.6. In Appendix A we present an application of our proposed approach to link three Colombian homicide record systems.

2.2 Multifile Partitioning

Consider K files $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$, each containing information on possibly overlapping subsets of a population of entities. The goal of multifile record linkage and duplicate detection is to identify the sets of records in $\boldsymbol{X}_1$, $\cdots$, $\boldsymbol{X}_K$ that are coreferent, as illustrated in Figure 2.1. Identifying coreferent records across datafiles represents the goal of record linkage, and identifying coreferent records within each file represents the goal of duplicate detection.

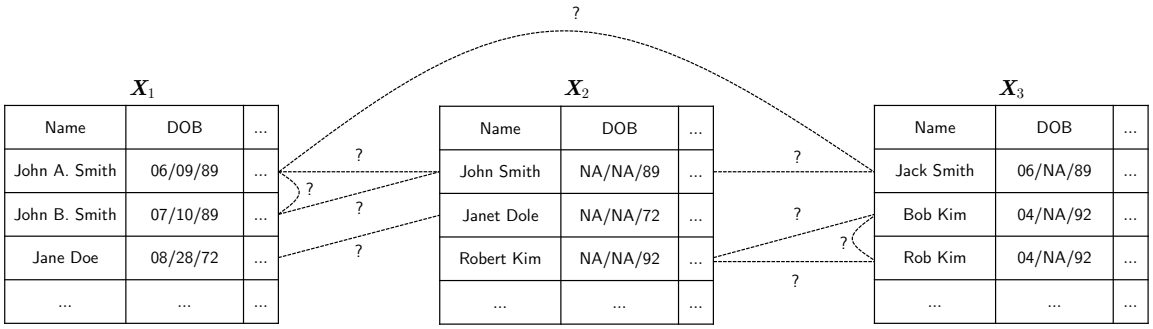


Figure 2.1: A toy example of the multifile record linkage and duplicate detection problem.

We denote the number of records contained in datafile $\boldsymbol{X}_k$ as r_k, and the total number of records across all files as $r = \sum_{k=1}^{K} r_k$. We label the records in all datafiles in a consecutive order, that is, those in $\boldsymbol{X}_1$ as $R_1 = (1, \cdots, r_1)$, those in $\boldsymbol{X}_2$ as $R_2 = (r_1 + 1, \cdots, r_1 + r_2)$, and so on, finally labeling the records in $\boldsymbol{X}_K$ as $R_K = (\sum_{k=1}^{K-1} r_k + 1, \cdots, r)$. We denote

$[r] = (1, \cdots, r)$, where it is clear that $[r] = (R_1, \cdots, R_K)$, which represents all the records coming from all datafiles.

Formally, multifile record linkage and duplicate detection is a partitioning problem. A partition of a set is a collection of disjoint subsets, called clusters, whose union is the original set. In this context, the term *coreference partition* refers to a partition $\mathcal{C}$ of all the records in $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$, or equivalently a partition $\mathcal{C}$ of $[r]$, such that each cluster $c \in \mathcal{C}$ is exclusively composed of all the records generated by a single entity [91, 114]. This implies that there is a one-to-one correspondence between the clusters in $\mathcal{C}$ and the entities represented in at least one of the datafiles. Estimating $\mathcal{C}$ is the goal of multifile record linkage and duplicate detection.

2.2.1 Multifile Coreference Partitions

In the setting of multifile record linkage and duplicate detection, the datafiles are the product of K data collection processes, which possibly introduce different patterns of duplication, measurement error, and missingness. This indicates that records coming from different datafiles should be treated differently. To take this into account, we introduce the concept of a *multifile coreference partition* by endowing a coreference partition $\mathcal{C}$ with additional structure to preserve the information on where records come from. Each cluster $c \in \mathcal{C}$ can be decomposed as $c = c_1 \cup \cdots \cup c_k \cup \cdots \cup c_K$, where c_k is the subset of records in cluster c that belong to datafile $\boldsymbol{X}_k$, which leads us to the following definition.

Definition 2.1. *The* multifile coreference partition *of datafiles $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$ is obtained from the coreference partition $\mathcal{C}$ by expressing each cluster $c \in \mathcal{C}$ as a K-tuple $(c_1, \cdots, c_K)$, where c_k represents the records of c that come from datafile $\boldsymbol{X}_k$.*

For simplicity we will continue using the notation $\mathcal{C}$ to denote a multifile coreference partition, although technically this new structure is richer and therefore different from a coreference partition that does not preserve the datafile membership of the records. The multifile representation of partitions is useful for decoupling the features that are important

for within-file duplicate detection or for across-files record linkage.

For duplicate detection, the goal is to identify coreferent records within each datafile. This can be phrased as estimating the *within-file coreference partition* $\mathcal{C}_k$ of each datafile $\boldsymbol{X}_k$. Clearly, these $\mathcal{C}_k$ can be obtained from the multifile partition $\mathcal{C}$ by extracting the kth entry of each cluster $c = (c_1, \ldots, c_K) \in \mathcal{C}$. Two useful summaries of a given within-file partition $\mathcal{C}_k$ are the number of within-file clusters $n_k = |\mathcal{C}_k|$, which is the number of unique entities represented in datafile $\boldsymbol{X}_k$, and the within-file cluster sizes $\boldsymbol{d}_k = \{|c_k| : c_k \in \mathcal{C}_k\}$, which represent the number of records associated with each entity in datafile $\boldsymbol{X}_k$.

On the other hand, in record linkage the goal is to identify coreferent records across datafiles. Given the within-file partitions, $\mathcal{C}_1, \cdots, \mathcal{C}_K$, the goal can be phrased as identifying which clusters across these partitions represent the same entities. This across-datafiles structure can be formally represented by a *K-partite matching*. Given K sets $V_1, \cdots, V_K$, a K-partite matching $\mathcal{M}$ is a collection of subsets from $\cup_{k=1}^{K} V_k$ such that each $m \in \mathcal{M}$ contains maximum one element from each V_k. If we think of each V_k as the set of clusters $\mathcal{C}_k$ representing the entities in datafile $\boldsymbol{X}_k$, then it is clear that the goal is to identify the K-partite matching $\mathcal{M}$ that puts together the clusters that refer to the same entities across datafiles. This structure can be extracted from a multifile coreference partition $\mathcal{C}$, given that each element $c = (c_1, \ldots, c_K) \in \mathcal{C}$ contains the coreferent clusters across all within-file partitions. Indeed, a multifile coreference partition can be thought of as a K-partite matching of within-file coreference partitions.

A useful summary of the across-datafile structure is the amount of entity-overlap between datafiles, represented by the number of clusters $c = (c_1, \ldots, c_K) \in \mathcal{C}$ with records in specific subsets of the files. We can concisely summarize the entity-overlap of the datafiles through a contingency table. In particular, consider a 2^K contingency table with cells indexed by $\boldsymbol{h} \in \{0, 1\}^K$ and corresponding cell counts $n_{\boldsymbol{h}}$. Here, $\boldsymbol{h}$ represents a pattern of inclusion of an entity in the datafiles, where a 1 indicates inclusion and a 0 exclusion. For instance, if $K = 3$, n_{011} is the number of clusters $c = (c_1, c_2, c_3) \in \mathcal{C}$ representing entities with records in datafiles 2 and 3 but without records in datafile 1. We let $\mathcal{H} = \{0, 1\}^K \setminus \{0\}^K$ and denote the

(incomplete) contingency table of counts as $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$, which we refer to as the *overlap table*. We ignore the cell $\{0\}^K$ which would represent entities that are not recorded in any of the K files. This cell is not of interest in this chapter, although it is the parameter of interest in population size estimation [see e.g. 20].

Example. To illustrate the concept of a multifile partition, consider two files with five and seven records respectively, so that $\boldsymbol{X}_1$ contains records $1-5$ and $\boldsymbol{X}_2$ contains records $6-12$. Suppose the coreference partition is $\{\{1,9\},\{2\},\{3,8,10,11\},\{4,5,7\},\{6\},\{12\}\}$. The corresponding multifile partition is $\mathcal{C} = \{(\{1\},\{9\}),(\{2\},\emptyset),(\{3\},\{8,10,11\}),(\{4,5\},\{7\}),(\emptyset,\{6\}),(\emptyset,\{12\})\}$. As illustrated in Figure 2.2, the within-file partitions can be extracted as $\mathcal{C}_1 = \{\{1\},\{2\},\{3\},\{4,5\}\}$ and $\mathcal{C}_2 = \{\{6\},\{7\},\{9\},\{8,10,11\},\{12\}\}$, and the within-file cluster sizes are $\boldsymbol{d}_1 = (1,1,1,2)$ and $\boldsymbol{d}_2 = (1,1,1,3,1)$. The overlap table in this case is $\{n_{11}, n_{10}, n_{01}\}$, indicating that $n_{11} = 3$ entities are represented in both datafiles, $n_{10} = 1$ entity is represented only in the first datafile, and $n_{01} = 2$ entities are represented only in the second datafile. In total, there are $n_1 = |\mathcal{C}_1| = n_{11}+n_{10} = 4$ unique entities represented in $\boldsymbol{X}_1$, $n_2 = |\mathcal{C}_2| = n_{11}+n_{01} = 5$ unique entities represented in $\boldsymbol{X}_2$, and $n = |\mathcal{C}| = n_{11}+n_{10}+n_{01} = 6$ entities among both datafiles.

2.3 A Structured Prior for Multifile Partitions

Bayesian approaches to multifile record linkage and duplicate detection require prior distributions on multifile coreference partitions. We present a generative process for multifile partitions, building on our representation introduced in Section 2.2.1. The idea is to generate a multifile partition by first generating summaries that characterize it, as follows:

1. Generate the number of unique entities n represented in the datafiles, which also corresponds to the number of clusters of the multifile partition.
2. Given n, generate an overlap table $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ so that $n = \sum_{\boldsymbol{h}\in\mathcal{H}} n_{\boldsymbol{h}}$, where $\mathcal{H} = \{0,1\}^K \setminus \{0\}^K$. From $\boldsymbol{n}$ we can derive the number of entities in datafile $\boldsymbol{X}_k$ as $n_k = \sum_{\boldsymbol{h}\in\mathcal{H}} h_k n_{\boldsymbol{h}}$, where h_k is the kth entry of $\boldsymbol{h}$.

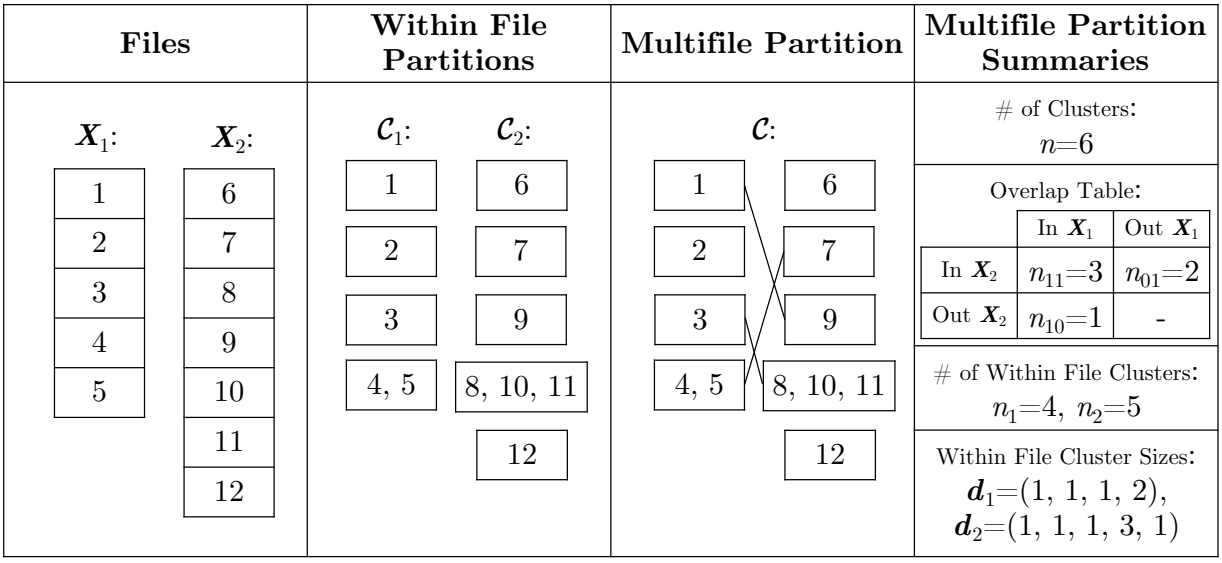


Figure 2.2: An illustration of a multifile partition of [12], where $\boldsymbol{X}_1$ contains records $1-5$ and $\boldsymbol{X}_2$ contains records $6-12$.

3. For each $k = 1, \cdots, K$, given n_k, independently generate a set of counts $\boldsymbol{d}_k = \{d_{ki}\}_{i=1}^{n_k}$, representing the number of records associated with each entity in file $\boldsymbol{X}_k$. From $\boldsymbol{d}_k$, we can derive the number of records in file $\boldsymbol{X}_k$ as $r'_k = \sum_{i=1}^{n_k} d_{ki}$. Index the r'_k records as $R'_k = (\sum_{l=1}^{k-1} r'_l + 1, \cdots, \sum_{l=1}^{k} r'_l)$.

4. For each $k = 1, \cdots, K$, given $\boldsymbol{d}_k$, induce a within-file partition $\mathcal{C}_k$ by randomly allocating R'_k into n_k clusters of sizes $d_{k1}, \cdots, d_{kn_k}$.

5. Given the overlap table $\boldsymbol{n}$ and within-file partitions $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$, generate a K-partite matching of the within-file partitions by selecting uniformly at random from the set of all K-partite matchings with overlap table $\boldsymbol{n}$. By definition, the result is a multifile coreference partition.

By parameterizing each step of this generative process, we can construct a prior distribution for multifile partitions, as we now show.

2.3.1 Parameterizing the Generative Process

Prior for the Number of Entities or Clusters. In the absence of substantial prior information, we follow a simple choice for the prior on the number of clusters, by taking a uniform distribution over the integers less than some upper bound, U, i.e. $\mathbb{P}(n) = U^{-1} I(n \in \{1, \cdots, U\})$. In practice, we set U to be the actual number of records across all datafiles r, which is observed. More informative specifications are discussed in Appendix A.

Prior for the Overlap Table. Conditional on n, we use a Dirichlet-multinomial distribution as our prior on the overlap table $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h} \in \mathcal{H}}$. Given a collection of positive hyperparameters for each cell of the overlap table, $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{h}}\}_{\boldsymbol{h} \in \mathcal{H}}$, and letting $\alpha_0 = \sum_{\boldsymbol{h} \in \mathcal{H}} \alpha_{\boldsymbol{h}}$, the prior for the overlap table under this choice is $\mathbb{P}(\boldsymbol{n} \mid n) = [(n!)\Gamma(\alpha_0)/\Gamma(n + \alpha_0)] \prod_{\boldsymbol{h} \in \mathcal{H}} [\Gamma(n_{\boldsymbol{h}} + \alpha_{\boldsymbol{h}})/(n_{\boldsymbol{h}}!)\Gamma(\alpha_{\boldsymbol{h}})]$. Due to conjugacy, $\boldsymbol{\alpha}$ can be interpreted as prior cell counts, which can be used to incorporate prior information about the overlap between datafiles. In the absence of substantial prior information, when the number of files is not too large and the overlap table is not expected to be sparse, we recommend setting $\boldsymbol{\alpha} = (1, \cdots, 1)$. In Appendix A we discuss alternative specifications when the overlap table is expected to be sparse. *Prior for the Within-File Cluster Sizes.* Given the number of entities in datafile $\boldsymbol{X}_k$, $n_k = \sum_{\boldsymbol{h} \in \mathcal{H}} h_k n_{\boldsymbol{h}}$, we generate the within-file cluster sizes $\boldsymbol{d}_k = \{d_{ki}\}_{i=1}^{n_k}$ assuming that $d_{k1}, \cdots, d_{kn_k} \mid n_k \overset{iid}{\sim} p_k(\cdot)$. Here $p_k(\cdot)$ represents the probability mass function of a distribution on the positive integers, so that $\mathbb{P}(\boldsymbol{d}_k \mid n_k) = \prod_{i=1}^{n_k} p_k(d_{ki})$. We do not expect a priori many duplicates per entity, and therefore we expect the counts in $\boldsymbol{d}_k$ to be mostly ones or to be very small [95, 144]. We therefore use a similar approach to [69], and use distributions truncated to the range $\{1, \cdots, U_k\}$, where U_k is a file-specific upper bound on cluster sizes. We further use distributions where prior mass is concentrated at small values. A default specification is to use a Poisson distribution with mean 1 truncated to $\{1, \cdots, U_k\}$, i.e. $p_k(d_{ki}) \propto (d_{ki}!)^{-1} I(d_{ki} \in \{1, \cdots, U_k\})$. More informative options could be used for $p_k(\cdot)$ by using any distribution on $\{1, \cdots, U_k\}$, where this could vary from file to file if some files were known to have more or less duplication. See Appendix A for more discussion.

Prior for the Within-File Partitions. Given the within-file cluster sizes $\boldsymbol{d}_k$, the number of ways of assigning $d_{k1}, \cdots, d_{kn_k}$ records to clusters $1, \cdots, n_k$, respectively, is given by the multinomial coefficient $r'_k! / \prod_{i=1}^{n_k} d_{ki}!$, with $r'_k = \sum_{i=1}^{n_k} d_{ki}$. However, the ordering of the clusters is irrelevant for constructing the within-file partition $\mathcal{C}_k$ of R'_k. There are $n_k!$ ways of ordering the n_k clusters of $\mathcal{C}_k$, which leads to $r'_k!/(n_k! \prod_{i=1}^{n_k} d_{ki}!)$ partitions of R'_k into clusters of sizes $d_{k1}, \cdots, d_{kn_k}$. We then have $\mathbb{P}(\mathcal{C}_k \mid \boldsymbol{d}_k) = (n_k!/r'_k!) \prod_{i=1}^{n_k} d_{ki}!$.

Prior for the K-Partite Matching. Given the overlap table $\boldsymbol{n}$ and the within-file partitions $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$, our prior over K-partite matchings of the within-file partitions is uniform. Thus we just need to count the number of K-partite matchings with overlap table $\boldsymbol{n}$. This is taken care of by Proposition 2.1, proven in Appendix A.

Proposition 2.1. *The number of K-partite matchings that have the same overlap table,* $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h} \in \mathcal{H}}$, *is* $\prod_{k=1}^{K} n_k! / \prod_{\boldsymbol{h} \in \mathcal{H}} n_{\boldsymbol{h}}!$. *Thus* $\mathbb{P}(\mathcal{C} \mid \{\mathcal{C}_k\}_{k=1}^{K}, \boldsymbol{n}) = \prod_{\boldsymbol{h} \in \mathcal{H}} n_{\boldsymbol{h}}! / \prod_{k=1}^{K} n_k!$.

The Structured Prior for Multifile Partitions. Letting quantities followed by $(\mathcal{C})$ mean they are computable from $\mathcal{C}$, the density of our structured prior for multifile partitions is

$$\begin{aligned}
\mathbb{P}(\mathcal{C}) &= \mathbb{P}(n)\mathbb{P}(\boldsymbol{n} \mid n) \prod_{k=1}^{K} \left[\mathbb{P}(\boldsymbol{d}_k \mid n_k)\mathbb{P}(\mathcal{C}_k \mid \boldsymbol{d}_k)\right] \mathbb{P}(\mathcal{C} \mid \{\mathcal{C}_k\}_{k=1}^{K}, \boldsymbol{n}) \\
&= \mathbb{P}(n(\mathcal{C})) \frac{n(\mathcal{C})!\, \Gamma(\alpha_0)}{\Gamma(n(\mathcal{C}) + \alpha_0)} \prod_{\boldsymbol{h} \in \mathcal{H}} \left[\frac{\Gamma(n_{\boldsymbol{h}}(\mathcal{C}) + \alpha_{\boldsymbol{h}})}{\Gamma(\alpha_{\boldsymbol{h}})}\right] \prod_{k=1}^{K} \left[\frac{1}{r'_k(\mathcal{C})!} \prod_{c_k \in \mathcal{C}_k} |c_k|!\; p_k(|c_k|)\right]. \quad (2.1)
\end{aligned}$$

2.3.2 Comments and Related Literature

The structured prior for multifile partitions allows us to incorporate prior information about the total number of clusters, the overlap between files, and the amount of duplication in each file. If we restrict the prior for the within-file cluster sizes to be $p_k(d_{ki}) = I(d_{ki} = 1)$ for a given datafile $\boldsymbol{X}_k$, then we enforce the assumption that there are no duplicates in that file. Imposing this restriction for all datafiles leads to the special case of a prior for K-partite matchings, which is of independent interest, as we are not aware of any such constructions outside of the bipartite case [48, 74, 115].

Our prior construction, where priors are first placed on interpretable summaries of a partition and then a uniform prior is placed on partitions which have those summaries, mimics the construction of the priors on bipartite matchings of [48, 74] and [115], and the Kolchin and allelic partition priors of [144] and [15]. While the Kolchin and allelic partition priors could both be used as priors for multifile partitions, these do not incorporate the datafile membership of records. Using these priors in the multifile setting would imply that the sizes of clusters containing records from only one file have the same prior distribution as the sizes of clusters containing records from two files, which should not be true in general.

[95] and [144] proposed the *microclustering property* as a desirable requirement for partition priors in the context of duplicate detection: denoting the size of the largest cluster in a partition of $[r]$ by M_r, a prior satisfies the microclustering property if $M_r/r \to 0$ in probability as $r \to \infty$. A downside of priors with this property is that they can still allow the size of the largest cluster to go to ∞ as r increases. For this reason [15] introduced the stronger *bounded microclustering property*, which we believe is more practically important: for any r, M_r is finite with probability 1. Our prior satisfies the bounded microclustering property as $M_r \leq \sum_{k=1}^{K} U_k$.

While our parameter of interest is a partition $\mathcal{C}$ of r records, the prior developed in this section is a prior for a partition of a random number of records. In practice we condition on the file sizes, $\{r_k\}_{k=1}^K$, and use the prior $\mathbb{P}(\mathcal{C} \mid \{r_k\}_{k=1}^K) \propto \mathbb{P}(\mathcal{C}) I(r'_k(\mathcal{C}) = r_k(\mathcal{C}) \text{ for all } k)$, which alters the interpretation of the prior. A similar problem occurs for the Kolchin partition priors of [144]. This motivated the exchangeable sequences of clusters priors of [16], which are similar to Kolchin partition priors, but lead to a directly interpretable prior specification. It would be interesting in future work to see if an analogous prior could be developed for our structured prior for multifile partitions. Despite this limitation, we demonstrate in simulations in Section 2.6 that incorporating strong prior information into our structured prior for multifile partitions can lead to improved frequentist performance over a default specification.

2.4 A Model for Comparison Data

We now introduce a comparison-based modeling approach to multifile record linkage and duplicate detection, building on the work of [42, 67, 141, 75, 48, 74] and [114, 115]. Working under the intuitive assumption that coreferent records will look similar, and non-coreferent records will look dissimilar, these approaches construct statistical models for comparisons computed between each pair of records.

There are two implications of the multifile setting described in Section 2.2 that are important to consider when constructing a model for the comparison data. First, models for the comparison data should account for the fact that the distribution of the comparisons between record pairs might potentially change across different pairs of files. For example, if files $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$ are not accurate, whereas files $\boldsymbol{X}_q$ and $\boldsymbol{X}_{q'}$ are, then the distribution of comparisons between $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$ will look very different compared with the distribution of comparisons between $\boldsymbol{X}_q$ and $\boldsymbol{X}_{q'}$. Second, the fields available for comparison will vary across pairs of files. For example, files $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$ may have collected information on a field that file $\boldsymbol{X}_q$ did not. In this scenario, we would like a model that is able to utilize this extra field when linking $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$, even though it is not available in $\boldsymbol{X}_q$. In this section we introduce a Bayesian comparison-based model that explicitly handles the multifile setting by constructing a likelihood function that models comparisons of fields between different pairs of files separately. The separate models are able to adapt to the level of noise of each file pair, and the maximal number of fields are able to be compared for each file pair.

2.4.1 Comparison Data

We construct comparison vectors for pairs of records to provide evidence for whether they correspond to the same entity. For $k \leq k'$, let $\mathcal{P}_{kk'} = \{(i,j) : i < j, i \in \boldsymbol{X}_k, j \in \boldsymbol{X}_{k'}\}$ denote the set of all record pairs between files $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$, and let F be the total number of different fields available from the K files. For each file pair (k, k'), $k \leq k'$, and record pair $(i,j) \in \mathcal{P}_{kk'}$, we compare each field $f = 1, \ldots, F$ using a similarity measure $\mathcal{S}_f(i,j)$,

which will depend on the data type of field f. For unstructured categorical fields such as race, $\mathcal{S}_f$ can be a binary comparison which checks for agreement. For more structured fields containing strings or numbers, $\mathcal{S}_f$ should be able to capture partial agreements. For example, string fields can be compared using a string metric like the Levenshtein edit distance [see e.g. 17], and numeric fields can be compared using absolute differences. Comparison $\mathcal{S}_f(i,j)$ will be missing if field f is not recorded in record i or record j, which includes the case where field f is not recorded in datafiles $\boldsymbol{X}_k$ or $\boldsymbol{X}_{k'}$.

While we could directly model the similarity measures $\mathcal{S}_f(i,j)$, this would require a custom model for each type of comparison, which inherits similar problems to the direct modeling of the fields themselves. Instead, we follow [140] and [114, 115] in dividing the range of $\mathcal{S}_f$ into L_f+1 intervals $I_{f0}, I_{f1}, \cdots, I_{fL_f}$ that represent varying levels of agreement, with I_{f0} representing the highest level of agreement, and I_{fL_f} representing the lowest level of agreement. We then let $\gamma_{ij}^f = l$ if $\mathcal{S}_f(i,j) \in I_{fl}$, where larger values of γ_{ij}^f represent larger disagreements between records i and j in field f. Finally, we form the comparison vector $\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \cdots, \gamma_{ij}^F)$. Constructing the comparison data this way allows us to build a generic modeling approach. In particular, extending [48, 74] and [114, 115], our model for the comparison data is

$$\begin{aligned}
\boldsymbol{\gamma}_{ij} \mid \mathcal{C}(i) = \mathcal{C}(j), (i,j) \in \mathcal{P}_{kk'} &\overset{iid}{\sim} \mathsf{M}_{kk'}(\boldsymbol{m}_{kk'}), \\
\boldsymbol{\gamma}_{ij} \mid \mathcal{C}(i) \neq \mathcal{C}(j), (i,j) \in \mathcal{P}_{kk'} &\overset{iid}{\sim} \mathsf{U}_{kk'}(\boldsymbol{u}_{kk'}),
\end{aligned}$$

where $\mathcal{C}$ is a multifile partition, $\mathcal{C}(i)$ denotes record i's cluster in $\mathcal{C}$, $\mathcal{C}(i) = \mathcal{C}(j)$ indicates that records i and j are coreferent, $\mathsf{M}_{kk'}(\boldsymbol{m}_{kk'})$ is a model for the comparison data among coreferent record pairs from the file pair $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$, $\mathsf{U}_{kk'}(\boldsymbol{u}_{kk'})$ is a model for the comparison data among non-coreferent record pairs from datafile pair $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$, and $\boldsymbol{m}_{kk'}$ and $\boldsymbol{u}_{kk'}$ are vectors of parameters.

In the next section we make two further assumptions that simplify the model parameterization. Before doing so, we note a few limitations of our comparison-based model. First, computing comparison vectors scales quadratically in the number of records. Second, com-

parison vectors for different record pairs are not actually independent conditional on the partition [see Section 2 of 131]. Third, modeling discretized comparisons of record fields represents a loss of information. While the first limitation is computational and unavoidable in the absence of blocking (see Appendix A), the other two limitations are inferential. Despite these limitations, we find in Section 2.6 that the combination of our structured prior for multifile partitions and our comparison-based model can produce linkage estimates with satisfactory frequentist performance.

2.4.2 Conditional Independence and Missing Data

Under the assumptions that the fields in the comparison vectors are conditionally independent given the multifile partition of the records and that missing comparisons are ignorable [114, 115], the likelihood of the observed comparison data, $\boldsymbol{\gamma}^{obs}$, becomes

$$\mathcal{L}(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs}) = \prod_{k \leq k'} \prod_{f=1}^{F} \prod_{l=0}^{L_f} (m_{kk'}^{fl})^{a_{kk'}^{fl}(\mathcal{C})} (u_{kk'}^{fl})^{b_{kk'}^{fl}(\mathcal{C})}. \tag{2.2}$$

Here $m_{kk'}^{fl} = \mathbb{P}(\gamma_{ij}^{f} = l \mid \mathcal{C}(i) = \mathcal{C}(j), (i,j) \in \mathcal{P}_{kk'})$, $u_{kk'}^{fl} = \mathbb{P}(\gamma_{ij}^{f} = l \mid \mathcal{C}(i) \neq \mathcal{C}(j), (i,j) \in \mathcal{P}_{kk'})$, $a_{kk'}^{fl}(\mathcal{C}) = \sum_{(i,j) \in \mathcal{P}_{kk'}} I_{obs}(\gamma_{ij}^{f}) I(\gamma_{ij}^{f} = l) I(\mathcal{C}(i) = \mathcal{C}(j))$, $b_{kk'}^{fl}(\mathcal{C}) = \sum_{(i,j) \in \mathcal{P}_{kk'}} I_{obs}(\gamma_{ij}^{f}) I(\gamma_{ij}^{f} = l) I(\mathcal{C}(i) \neq \mathcal{C}(j))$, $I_{obs}(\cdot)$ is an indicator of whether its argument was observed, and $\Phi = (\boldsymbol{m}, \boldsymbol{u})$ where $\boldsymbol{m}$ collects all of the $\boldsymbol{m}_{kk'}^{f} = (m_{kk'}^{f0}, \cdots, m_{kk'}^{fL_f})$ and $\boldsymbol{u}$ collects all of the $\boldsymbol{u}_{kk'}^{f} = (u_{kk'}^{f0}, \cdots, u_{kk'}^{fL_f})$. For a given multifile partition $\mathcal{C}$, $a_{kk'}^{fl}(\mathcal{C})$ represents the number of record pairs in $\mathcal{P}_{kk'}$ that belong to the same cluster with observed agreement at level l in field f, and $b_{kk'}^{fl}(\mathcal{C})$ represents the number of record pairs in $\mathcal{P}_{kk'}$ that do not belong to the same cluster with observed agreement at level l in field f.

2.5 Bayesian Estimation of Multifile Partitions

Bayesian estimation of the multifile coreference partition $\mathcal{C}$ is based on the posterior distribution $p(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs}) \propto \mathbb{P}(\mathcal{C}) p(\Phi) \mathcal{L}(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs})$, where $\mathbb{P}(\mathcal{C})$ is our structured prior for multifile partitions (2.1), $\mathcal{L}(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs})$ is the likelihood from our model for comparison data (2.2),

and $p(\Phi)$ represents a prior distribution for the $\Phi = (\boldsymbol{m}, \boldsymbol{u})$ model parameters. We now specify this prior $p(\Phi)$, outline a Gibbs sampler to sample from $p(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs})$, and present a strategy to obtain point estimates of the multifile partition $\mathcal{C}$.

2.5.1 Priors for $\boldsymbol{m}$ and $\boldsymbol{u}$

We will use independent, conditionally conjugate priors for $\boldsymbol{m}^f_{kk'}$ and $\boldsymbol{u}^f_{kk'}$, namely $\boldsymbol{m}^f_{kk'} \sim \mathsf{Dirichlet}(\mu^{f0}_{kk'}, \cdots, \mu^{fL_f}_{kk'})$ and $\boldsymbol{u}^f_{kk'} \sim \mathsf{Dirichlet}(\nu^{f0}_{kk'}, \cdots, \nu^{fL_f}_{kk'})$. In this chapter we will use a default specification of $(\mu^{f0}_{kk'}, \cdots, \mu^{fL_f}_{kk'}) = (\nu^{f0}_{kk'}, \cdots, \nu^{fL_f}_{kk'}) = (1, \cdots, 1)$. We believe this prior specification is sensible for the $\boldsymbol{u}$ parameters, following the discussion in Section 3.2 of [114], as comparisons amongst non-coreferent records are likely to be highly variable and it is more likely than not that eliciting meaningful priors for them is too difficult. For the $\boldsymbol{m}$ parameters, it might be desirable in certain applications to introduce more information into the prior. For example, one could set $\mu^{f0}_{kk'} > \cdots > \mu^{fL_f}_{kk'}$ to incorporate the prior belief that higher levels of agreement should have larger prior probability than lower levels of agreement. Another route would be to use the sequential parameterization of the $\boldsymbol{m}$ parameters, and the associated prior recommendations, described in [114].

2.5.2 Posterior Sampling

In Appendix A we outline a Gibbs sampler that produces a sequence of samples $\{\mathcal{C}^{[t]}, \Phi^{[t]}\}_{t=1}^{T}$ from the posterior distribution $p(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs})$, which we will use to obtain Monte Carlo approximations of posterior expectations involved in the derivation of point estimates $\hat{\mathcal{C}}$, as presented in the next section. In Appendix A, we discuss the computational complexity of the Gibbs sampler, how computational performance can be improved through the usage of indexing techniques, and the initialization of the Gibbs sampler.

2.5.3 Point Estimation

In a Bayesian setting, one can obtain a point estimate $\hat{\mathcal{C}}$ of the multifile partition using the posterior $\mathbb{P}(\mathcal{C} \mid \boldsymbol{\gamma}^{obs}) = \int p(\mathcal{C}, \Phi \mid \boldsymbol{\gamma}^{obs}) d\Phi$ and a loss function $L(\mathcal{C}, \hat{\mathcal{C}})$. The Bayes estimate is the multifile partition $\hat{\mathcal{C}}$ that minimizes the expected posterior loss $\mathbb{E}[L(\mathcal{C}, \hat{\mathcal{C}}) \mid \boldsymbol{\gamma}^{obs}] = \sum_{\mathcal{C}} L(\mathcal{C}, \hat{\mathcal{C}}) \mathbb{P}(\mathcal{C} \mid \boldsymbol{\gamma}^{obs})$, although in practice such expectations are approximated using posterior samples. Previous examples of loss functions for partitions included Binder's loss [18] and the variation of information [93], both recently surveyed in [137]. The quadratic and absolute losses presented in [131] are special cases of Binder's loss, and [126] drew connections between their proposed maximal matching sets and the losses of [131].

In many applications there may be much uncertainty on the linkage decision for some records in the datafiles. For example, in Figure 2.1 it is unclear which of the records with last name "Smith" are coreferent. It is thus desirable to leave decisions for some records unresolved, so that the records can be hand-checked during a clerical review, which is common in practice [see e.g. 10]. In the classification literature, leaving some decisions unresolved is done through a *reject option* [see e.g. 57], which here we will refer to as an *abstain option*. We will refer to point estimates with and without an abstain option as *partial estimates* and *full estimates*, respectively. We now present a family of loss functions for multifile partitions which incorporate an abstain option, building upon the family of loss functions for bipartite matchings presented in [115].

A Family of Loss Functions with an Abstain Option

For the purpose of this section we will represent a multifile partition $\mathcal{C}$ as a vector $\boldsymbol{Z} = (Z_1, \cdots, Z_r)$ of labels, where $Z_i \in \{1, \ldots, r\}$, such that $Z_i = Z_j$ if $\mathcal{C}(i) = \mathcal{C}(j)$. We represent a Bayes estimate here as a vector $\hat{\boldsymbol{Z}} = (\hat{Z}_1, \ldots, \hat{Z}_r)$, where $\hat{Z}_i \in \{1, \ldots, r, A\}$, with A representing an abstain option intended for records whose linkage decisions are not clear and need further review. We assign different losses to using the abstain option and to different types of matching errors. We propose to compute the overall loss additively, as $L(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) =$

$\sum_{i=1}^{r} L_i(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$. To introduce the expression for the ith-record-specific loss $L_i(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$, we use the notation $\Delta_{ij} = I(Z_i = Z_j)$, and likewise $\hat{\Delta}_{ij} = I(\hat{Z}_i = \hat{Z}_j)$.

The proposed individual loss for record i is

$$L_i(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = \begin{cases} \lambda_A, & \text{if } \hat{Z}_i = A, \\ 0, & \text{if } \Delta_{ij} = \hat{\Delta}_{ij} \text{ for all } j \text{ where } \hat{Z}_j \neq A, \\ \lambda_{\text{FNM}}, & \text{if } \hat{Z}_i \neq A, \quad \sum_{j \neq i} \hat{\Delta}_{ij} = 0, \quad \sum_{j \neq i} \Delta_{ij} > 0, \\ \lambda_{\text{FM1}}, & \text{if } \hat{Z}_i \neq A, \quad \sum_{j \neq i} \hat{\Delta}_{ij} > 0, \quad \sum_{j \neq i} \Delta_{ij} = 0, \\ \lambda_{\text{FM2}}, & \text{if } \hat{Z}_i \neq A, \quad \sum_{j \neq i} \hat{\Delta}_{ij} > 0, \quad \sum_{j \neq i} (1 - \hat{\Delta}_{ij}) \Delta_{ij} > 0. \end{cases} \tag{2.3}$$

That is, λ_A represents the loss from abstaining from making a decision; λ_{FNM} is the loss from a false non-match (FNM) decision, that is, deciding that record i does not match any other record ($\sum_{j \neq i} \hat{\Delta}_{ij} = 0$) when in fact it does ($\sum_{j \neq i} \Delta_{ij} > 0$); λ_{FM1} is the loss from a type 1 false match (FM1) decision, that is, deciding that record i matches other records ($\sum_{j \neq i} \hat{\Delta}_{ij} > 0$) when it does not actually match any other record ($\sum_{j \neq i} \Delta_{ij} = 0$); and λ_{FM2} is the loss from a type 2 false match (FM2), that is, a false match decision when record i is matched to other records ($\sum_{j \neq i} \hat{\Delta}_{ij} > 0$) but it does not match all of the records it should be matching ($\sum_{j \neq i} (1 - \hat{\Delta}_{ij}) \Delta_{ij} > 0$).

The posterior expected loss is $\mathbb{R}(\hat{\boldsymbol{Z}}) = \sum_{i=1}^{r} \mathbb{E}[L_i(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) \mid \boldsymbol{\gamma}^{obs}]$, where

$$\mathbb{E}[L_i(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) \mid \boldsymbol{\gamma}^{obs}] = \begin{cases} \lambda_A, & \text{if } \hat{Z}_i = A, \\ \lambda_{\text{FNM}} \, \mathbb{P}(\sum_{j \neq i} \Delta_{ij} > 0 \mid \boldsymbol{\gamma}^{obs}), & \text{if } \hat{Z}_i \neq A, \sum_{j \neq i} \hat{\Delta}_{ij} = 0, \\ \lambda_{\text{FM1}} \, \mathbb{P}(\sum_{j \neq i} \Delta_{ij} = 0 \mid \boldsymbol{\gamma}^{obs}) + & \\ \lambda_{\text{FM2}} \, \mathbb{P}(\sum_{j \neq i} (1 - \hat{\Delta}_{ij}) \Delta_{ij} > 0 \mid \boldsymbol{\gamma}^{obs}), & \text{if } \hat{Z}_i \neq A, \sum_{j \neq i} \hat{\Delta}_{ij} > 0, \end{cases} \tag{2.4}$$

and quantities computed with respect to the posterior distribution, $\mathbb{P}(\boldsymbol{Z} \mid \boldsymbol{\gamma}^{obs})$, can all be approximated using posterior samples. While this presentation is for general positive losses $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}, \lambda_{\text{FM2}}$ and λ_A, these only have to be specified up to a proportionality constant [115]. If we do not want to allow the abstain option, then we can set $\lambda_A = \infty$ and the derived full estimate $\hat{\boldsymbol{Z}}$ will have a linkage decision for all records. Although we have been using partition labelings $\boldsymbol{Z}$, the expressions in (2.3) and (2.4) are invariant to different labelings of the same

partition. In the two-file case, [115] provided guidance on how to specify the individual losses $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}, \lambda_{\text{FM2}}$ and λ_A in cases where there is a notion of false matches being worse than false non-matches or vise versa. [115] also gave recommendations for default values of these losses that lead to good frequentist performance in terms not over- or under-matching across repeated samples. In Appendix A, we discuss how our proposed loss function differs from the loss function of [115] and propose a strategy for approximating the Bayes estimate.

2.6 Simulation Studies

To explore the performance of our proposed approach for linking three duplicate-free files, as in the application to the Colombian homicide record systems of Appendix A, we present two simulation studies under varying scenarios of measurement error and datafile overlap. The two studies correspond to scenarios with equal and unequal measurement error across files, respectively. Both studies present results based on full estimates. In Appendix A we further explore the performance of our proposed approach for linking three files with duplicates, with results based on full and partial estimates.

2.6.1 General Setup

We start by describing the general characteristics of the simulations. For each of the simulation scenarios we conduct 100 replications, for each of which we generate three files as follows. For each of $n = 500$ entities, $\boldsymbol{h} \in \mathcal{H}$ is drawn from a categorical distribution with probabilities $\{p_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$, where $\boldsymbol{h}$ represents the subset of files the entity appears in, and so we change the values of $\{p_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ across simulation scenarios to represent varying amounts of file overlap. Files are then created by generating the implied number of records for each entity. In the additional simulations considered in Appendix A, the generated number of records for each entity depends not only on $\boldsymbol{h}$, but also on the duplication mechanism.

All records are generated using a synthetic data generator developed in [133], which allows for the incorporation of different forms of measurement error in individual fields, along with dependencies between fields we would expect in applications. The data generator first

generates clean records before distorting them to create the observed records. In particular, each observed record will have a fixed number of erroneous fields, where errors selected uniformly at random from a set of field dependent errors displayed in Table 3 of [114] (reproduced in Appendix A), with a maximum of two errors per field. We generate records with seven fields of information: sex, given name, family name, age, occupation, postal code, and phone number.

For each simulation replicate, we construct comparison vectors as given in Table 4 of [114] (reproduced in Appendix A). We use the model for comparison data proposed in Section 2.4 with flat priors on $\boldsymbol{m}$ and $\boldsymbol{u}$ as discussed in Section 2.5.1, and the structured prior proposed in Section 2.3 with a uniform prior on the number of clusters and $\boldsymbol{\alpha} = (1, \cdots, 1)$ as described in Section 2.3.1. Using the Gibbs sampler presented in Appendix A we obtain $1,000$ samples from the posterior distribution of multifile partitions, and discard the first 100 as burn-in. In Appendix A we discuss convergence of the Gibbs sampler, present running times of the proposed approach, and present an extra simulation exploring the running time of the approach with a larger number of records. We then approximate the Bayes estimate $\hat{\boldsymbol{Z}}$ for multifile partitions using the loss function described in Section 2.5.3 as described in Appendix A. For full estimates, we use the default values of $\lambda_{\text{FNM}} = \lambda_{\text{FM1}} = 1$ and $\lambda_{\text{FM2}} = 2$ recommended by [115]. In Appendix A we explore alternative specifications of the loss function.

We will assess the performance of the Bayes estimate using *precision* and *recall* with respect to the true coreference partition $\boldsymbol{Z}$. Let $\mathcal{P}$ be the set of all record pairs. Using notation from Section 2.5.3, let $TM(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = \sum_{(i,j)\in\mathcal{P}} \Delta_{ij}\hat{\Delta}_{ij}$ be the number of true matches (record pairs correctly declared coreferent), $FM(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = \sum_{(i,j)\in\mathcal{P}}(1-\Delta_{ij})\hat{\Delta}_{ij}$ be the number of false matches (record pairs incorrectly declared coreferent), and $FNM(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = \sum_{(i,j)\in\mathcal{P}} \Delta_{ij}(1 - \hat{\Delta}_{ij})$ the number of false non-matches (record pairs incorrectly declared non-coreferent). Then *precision* is $TM(\boldsymbol{Z}, \hat{\boldsymbol{Z}})/[TM(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) + FM(\boldsymbol{Z}, \hat{\boldsymbol{Z}})]$, the proportion of record pairs declared as coreferent that were truly coreferent, and *recall* is $TM(\boldsymbol{Z}, \hat{\boldsymbol{Z}})/[TM(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) + FNM(\boldsymbol{Z}, \hat{\boldsymbol{Z}})]$, the proportion of record pairs that were truly coreferent that were correctly declared as

coreferent. Perfect performance corresponds to precision and recall both being 1. In the simulations, we computed the median, 2nd, and 98th percentiles of these measures over the 100 replicate data sets. Additionally, in Appendix A, we assess the performance of the Bayes estimate when estimating the number of entities, n.

2.6.2 Duplicate-Free Files, Equal Errors Across Files

In this simulation study we explore the performance of our methodology by varying the number of erroneous fields per record over $\{1, 2, 3, 5\}$, and also varying $\{p_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$, which determines the amount of overlap, over four scenarios:

- High Overlap: $p_{001} = p_{010} = p_{100} = 0.4/3, p_{011} = p_{101} = p_{110} = 0.15, p_{111} = 0.15$,
- Medium Overlap: $p_{001} = p_{010} = p_{100} = 0.7/3, p_{011} = p_{101} = p_{110} = 0.05, p_{111} = 0.15$,
- Low Overlap: $p_{001} = p_{010} = p_{100} = 0.8/3, p_{011} = p_{101} = p_{110} = 0.05/3, p_{111} = 0.15$,
- No-Three-File Overlap: $p_{001} = p_{010} = p_{100} = 0.55/3, p_{011} = p_{101} = p_{110} = 0.15, p_{111} = 0$.

These are intended to represent a range of scenarios that could occur in practice. In the high overlap scenario 60% of the entities are expected to be in more than one datafile, in the low overlap scenario 80% of the entities are expected to be represented in a single datafile, and in the no-three-file overlap scenario no entities are represented in all datafiles.

To implement our methodology, in addition to the general set-up described in Section 2.6.1, we restrict the prior for the within-file cluster sizes so that they have size one with probability one, incorporating the assumption of no-duplication within files (see Section 2.3.2). Imposing this restriction for all datafiles leads to a prior for tripartite matchings. To illustrate the impact of using our structured prior, we compare with the results obtained using our model for comparison data with a flat prior on tripartite matchings.

The results of the simulation are seen in Figure 2.3. We see that our proposed approach performs consistently well across different settings, with the exception of the no-three-file overlap setting under high measurement error, where the precision decreases dramatically.

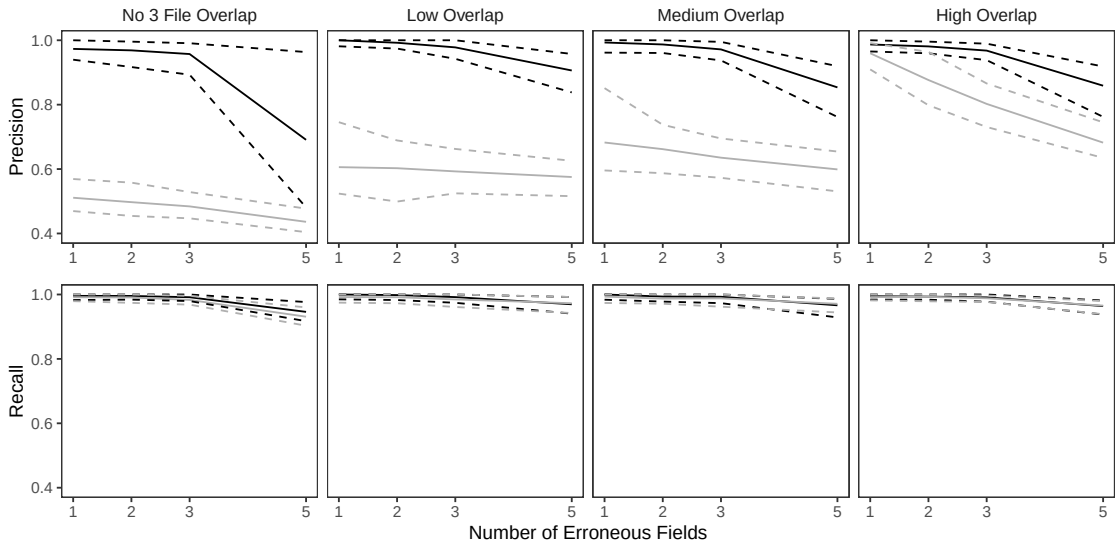


Figure 2.3: Performance comparison for simulation with equal measurement error across files. Black lines refer to results under our structured prior, grey lines to results under the flat prior, solid lines show medians, and dashed lines show 2nd and 98th percentiles.

The approach using a flat prior on tripartite matchings has poor precision in comparison, and it is particularly low when the amount of overlap is low. This suggests that our structured prior improves upon the flat prior by protecting against over-matching (declaring noncoreferent record pairs as coreferent). In Appendix A we demonstrate how the performance in the no-three-file overlap setting can be improved through the incorporation of an informative prior for the overlap table through $\boldsymbol{\alpha}$.

2.6.3 Duplicate-Free Files, Unequal Errors Across Files

In this simulation study we have different patterns of measurement error across the three files. Rather than each field in each record having a chance of being erroneous according to Table 3 of [114], we will use the following measurement error mechanism to generate the data. For each record in the first file, age is missing, given name has up to seven errors, and all other fields are error free. For each record in the second file, sex and occupation are

missing, last name has up to seven errors, and all other fields are error free. For each record in the third file, phone number and postal code have up to seven errors and all other fields are error free. Under this measurement error mechanism, there is enough information in the error free fields to inform pairwise linkage of the files. We further vary $\{p_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ over the no-three-file and high overlap settings from Section 2.6.2.

Our goal in this study is to demonstrate that having both the structured prior for partitions and the separate models for comparison data from each file-pair can lead to better performance than not having these components. We will compare our model as described in Section 2.6.2 to both our model for comparison data with a flat prior on tripartite matchings (as in Section 2.6.2) and a simplification of our model for comparison data using a single model for all file-pairs but with our structured prior for partitions.

The results of the simulation are given in Figure 2.4. We see that our proposed approach outperforms both alternative approaches in both precision and recall in both overlap settings. This suggests that both the structured prior for tripartite matchings and the separate models for comparison data from each file-pair can help improve performance over alternative approaches. We note that in the no-three-file (high) overlap setting the precision of the proposed approach is greater than or equal to the precision of the approach using a single model for all file-pairs in 98 (100) of the 100 replications.

2.7 Discussion and Future Work

The methodology proposed in this chapter makes three contributions. First, the multifile partition parameterization, specific to the context of multifile record linkage and duplicate detection, allows for the construction of our structured prior for partitions, which provides a flexible mechanism for incorporating prior information about the data collection processes of the files. This prior is applicable to any Bayesian approach which requires a prior on partitions, including direct-modeling approaches such as [126]. We are not aware of any priors for K-partite matchings when $K > 2$, so we hope our construction will lead to more development in this area. The second contribution is an extension of previous comparison-

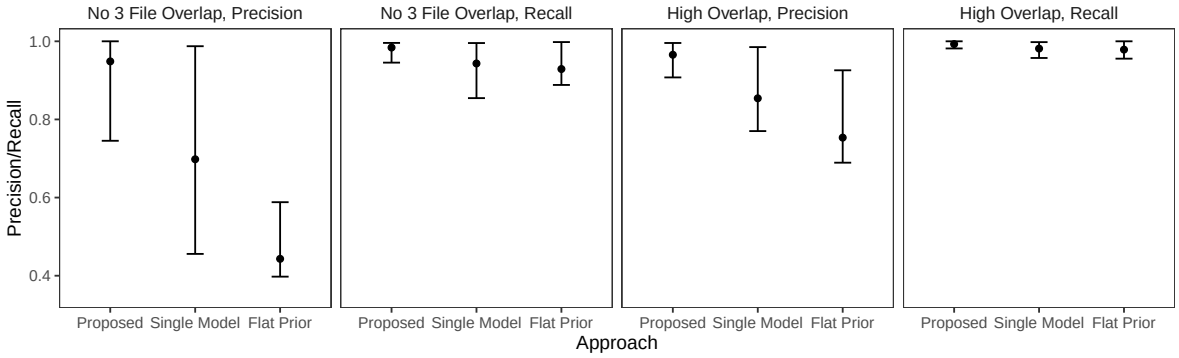


Figure 2.4: Performance comparison for simulation with unequal measurement error across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings. Dots show medians, and bars show 2nd and 98th percentiles.

based models that explicitly handles the multifile setting. Allowing separate models for comparison data from each file pair leads to higher quality linkage. The third is a novel loss function for multifile partitions which can be used to derive Bayes estimates with good frequentist properties. Importantly, the loss function allows for linkage decisions to be left unresolved for records with large matching uncertainty. As with our structured prior on partitions, the loss function is applicable to any Bayesian approach which requires point estimates of partitions, including direct-modeling approaches.

There are a number of directions for future work. One direction is the modeling of dependencies between the comparison fields [see e.g. 75], which should further improve the quality of the linkage. Another direction is the development of approaches to jointly link records and perform a downstream analysis, thereby propagating the uncertainty from the linkage. See Section 7.2 of [19] for a recent review of such joint models. In this direction, a natural task to consider next is population size estimation, where the linkage of the datafiles plays a central role [131, 132].

Chapter 3

THE CENTRAL ROLE OF THE IDENTIFYING ASSUMPTION IN POPULATION SIZE ESTIMATION

3.1 Introduction

Estimating the size of a closed population is a common problem in many fields, including ecology [101], epidemiology [62], official statistics [7], and human rights [9]. The available data typically take the form of multiple lists which record information on a subset of individuals in a population. When there exists a mechanism to identify which individuals are the same across lists, multiple-systems estimation (MSE), also known as capture-recapture, provides an approach to estimating the population size based on the overlap of the lists [20].

MSE is at its heart a missing data problem, as we do not observe all individuals in the population of interest [see e.g. 47, 87]. As in any missing data problem, MSE requires users to make an untestable identifying assumption about how the observed individuals relate to the unobserved individuals in order to estimate the population size from the observed data. In practice, this means that models with different identifying assumptions can produce arbitrarily different population size estimates, even when the models have identical fits to the observed data. Thus, any identifying assumptions used in an analysis need to be appropriately justified based on the context of the data. If an appropriate identifying assumption can not be found for a data set, no estimate of the population size should be produced based on that data set.

We believe that the central role of specifying the identifying assumption is not sufficiently appreciated, as it is usually conflated with model specification, which involves both making an identifying assumption *and* specifying a model for the observed data. See for example [45] who wrote "... we are assuming that the model which describes the observed data

also describes the count of the unobserved individuals. We have no way of checking this assumption," and [89] who wrote "The arguably most basic assumption in MSE is that the noninclusion of the fully unobserved individuals ... can be represented by the same model that represents the inclusion (and noninclusion) of those we can observe in at least one list. This is a strong and untestable condition."

This conflation of identifying assumption specification and model specification has led practitioners to perform model evaluation by comparing a suite of model fits that are the results of both fundamentally different identifying assumptions and different model specifications for the observed data [see e.g. 116, 88, 119]. This makes it essentially impossible to disentangle whether differences in inferences are due to differences in identifying assumptions, model specifications for the observed data, or some combination. More importantly, it is rare in these instances for practitioners to provide justification for any of the identifying assumptions being used.

In this chapter, we propose an approach for MSE that places the identifying assumption front and center in the MSE workflow. We first revisit the framing of MSE as a missing data problem and describe our approach in Section 3.2. Section 3.3 reviews two common MSE models—log-linear and latent class models—through our missing data framing. In Section 3.4 we focus on the identifying assumption associated with log-linear models, and describe how it can be used as a building block for alternative identifying assumptions and sensitivity analyses that examine the impact of the identifying assumption. Finally, in Section 3.5 we illustrate our approach in a case study of estimating the number of civilian casualties in the Kosovo war.

3.2 Multiple-Systems Estimation as a Missing Data Problem

3.2.1 The Data

Suppose we have a closed population of N individuals, of which $n < N$ are observed by one or more of K lists. Let $H = \{0, 1\}^K$ denote the possible patterns of inclusion of the

individuals in the lists, $H^* = H \setminus \{0\}^K$ denote the possible subsets of lists in which each of the n observed individuals could have been observed, and let $\boldsymbol{x}_i \in H$ denote the subset of lists in which individual i was included. For example, with $K = 3$, $\boldsymbol{x}_i = (0, 1, 1)$ indicates that individual i was observed in lists 2 and 3, but not list 1.

These data for the N individuals can be gathered into a 2^K contingency table of list overlap, where the cells of the table are indexed by $\boldsymbol{h} \in H$, with counts $n_{\boldsymbol{h}} = \sum_{i=1}^{N} I(\boldsymbol{x}_i = \boldsymbol{h})$. We do not observe the count for cell $\{0\}^K$, $n_0 := n_{(0,\cdots,0)} = N - n$, which records the number of individuals missing from all lists, so the observed contingency table is incomplete. Let $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H^*}$ denote the counts of the incomplete contingency table. The unobserved cell count n_0, or equivalently the population size N, is the target of inference.

3.2.2 The Complete-Data Distribution

Under independent and identically distributed (i.i.d.) sampling of individuals by the lists, the 2^K contingency table of counts is multinomially distributed, i.e.

$$\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi} \sim \textsc{Multinomial}(N, \boldsymbol{\pi}), \tag{3.1}$$

where $\boldsymbol{\pi} = \{\pi_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H} \in \mathbb{S}^{2^K - 1}$ is a set of cell probabilities, and $\mathbb{S}^d = \{(a_1, \cdots, a_{d+1}) \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} a_i = 1, a_i > 0 \ \forall i\}$ denotes the d-dimensional probability simplex. We note that this multinomial model, introduced as early as [29], is a possible simplification of reality, as it does not allow for correlation of individuals' inclusion patterns. We will refer to the model in (3.1) as the *complete-data distribution*, for which the evaluation relies on knowing the complete 2^K contingency table of counts. In general, the parameter space for this model will be some subset of $\Theta = \{N, \boldsymbol{\pi} \mid N \in \mathbb{N}, \boldsymbol{\pi} \in \mathbb{S}^{2^K - 1}\}$, which we will refer to as the *complete-data parameterization*. As shown in Appendix B, when individuals are not i.i.d. sampled, but are sampled independently with cell probabilities drawn i.i.d. from some mixing distribution on $\mathbb{S}^{2^K - 1}$, we also arrive at the model in (3.1). This is the case for common models for heterogeneity such as the M_h and M_{th} models of [101]. Because common models for heterogeneity reduce to (3.1), in the rest of this chapter we will view the cell probabilities

as being marginal of any possible heterogeneity mechanisms.

3.2.3 Decomposing the Complete-Data Distribution

It is instructive to decompose the complete-data distribution as

$$p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) = N! \prod_{\boldsymbol{h} \in H} \frac{\pi_{\boldsymbol{h}}^{n_{\boldsymbol{h}}}}{n_{\boldsymbol{h}}!} = L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}), \tag{3.2}$$

with $L_1(N, \pi_0 \mid n) = \binom{N}{n} \pi_0^{N-n} (1-\pi_0)^n$ and $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = n! \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{n_{\boldsymbol{h}}} / n_{\boldsymbol{h}}!$, where $\pi_0 := \pi_{(0,\cdots,0)} = 1 - \sum_{\boldsymbol{h} \in H^*} \pi_{\boldsymbol{h}}$ is the probability of being missing from every list, and $\tilde{\pi}_{\boldsymbol{h}} = \frac{\pi_{\boldsymbol{h}}}{1-\pi_0} = \frac{\pi_{\boldsymbol{h}}}{\sum_{\boldsymbol{h}' \in H^*} \pi_{\boldsymbol{h}'}}$ is the probability of being observed in the subset of the lists $\boldsymbol{h}$ conditional on being observed in at least one list. L_1 is a binomial likelihood for n, which has been well studied in the related binomial N problem literature [see e.g. 113]. L_2 is a multinomial likelihood for the observed data $\boldsymbol{n}$ conditional on their sum n, referred to as the *conditional likelihood* [45]. We will refer to π_0 as the *unobserved cell probability* and to $\tilde{\boldsymbol{\pi}}$ as the *observed cell probabilities*. This decomposition hints at an alternative to the complete-data parameterization Θ, $\Theta^* = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 \in (0,1), \tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K - 2}\}$, which we will refer to as the *observed-data parameterization*. The two parameterizations are equivalent, so we will work with whichever is more convenient for exposition.

3.2.4 Identifiability

Before performing inference in a statistical model, it is important to check that the model is identifiable. For $\theta \in \Theta^*$, let P_θ denote the complete-data distribution at the set of parameters θ. Consider the following standard definition of identifiability:

Definition 3.1. *The statistical model $\mathcal{P}_\Omega = \{P_\theta \mid \theta \in \Omega \subset \Theta^*\}$ is **identifiable** if $\forall \theta_1, \theta_2 \in \Omega$, $P_{\theta_1} = P_{\theta_2}$ implies that $\theta_1 = \theta_2$. Equivalently, $\mathcal{P}_\Omega$ is identifiable if $\forall \theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}, \theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = L_1(N', \pi_0' \mid n) L_2(\tilde{\boldsymbol{\pi}}' \mid \boldsymbol{n})$ $\forall \boldsymbol{n}$ implies that $\theta_1 = \theta_2$.*

One can show that the unrestricted model $\mathcal{P}_{\Theta^*}$ is identifiable. Since the goal is to estimate N, sufficiency might lead one to try to estimate N and π_0 in the unrestricted model based

solely on the binomial likelihood for n. Examining the likelihood surface for a given n, one finds a maximum at $N = n$ and $\pi_0 = 0$, with a ridge centered along the set $\{N \in \mathbb{N}, \pi_0 \in (0,1) \mid N(1-\pi_0) \approx n\}$ that monotonically decreases as N increases. In Figure 3.1 we plot this surface when $n = 100$. There is a fundamental problem in that two parameters are being estimated with one data point, which makes it impossible to construct an unbiased or consistent estimator of either N or π_0 [30, 38]. Thus the standard definition of identifiability is misleading in this setting, as it does not necessarily imply that the parameters are estimable in any traditional sense.

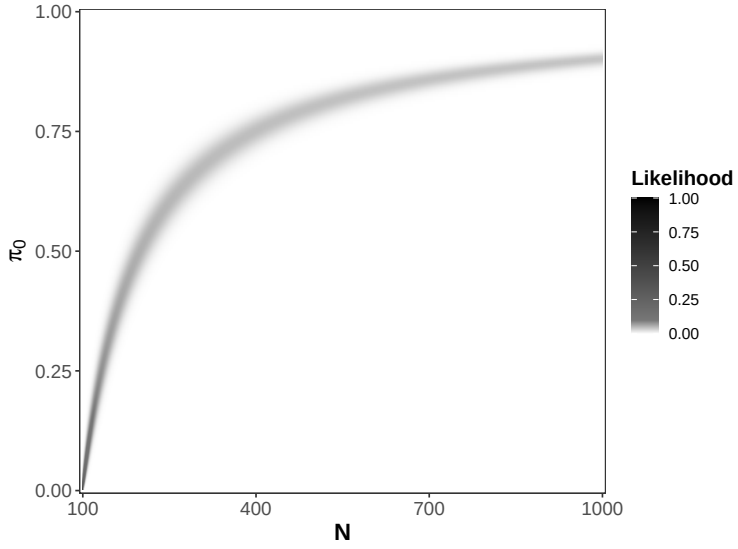


Figure 3.1: Likelihood surface of L_1 when $n = 100$.

We will instead use the following alternative definition of identifiability specific to MSE [79, 61]:

Definition 3.2. *The statistical model* $\mathcal{P}_\Omega$ *is* ***conditionally identifiable*** *if* $\forall \theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}$, $\theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = L_2(\tilde{\boldsymbol{\pi}}' \mid \boldsymbol{n})\ \forall \boldsymbol{n}$ *implies that* $\pi_0 = \pi_0'$.

In a conditionally identifiable model, the conditional likelihood, L_2, identifies the unobserved cell probability, π_0. Clearly the unrestricted model $\mathcal{P}_{\Theta^*}$ is not conditionally identifi-

able. Standard identifiability of the multinomial conditional likelihood tells us that we can equivalently state Definition 3.2 as follows: the statistical model $\mathcal{P}_\Omega$ is conditionally identifiable if $\forall\theta_1 = \{N, \pi_0, \tilde{\boldsymbol{\pi}}\}, \theta_2 = \{N', \pi_0', \tilde{\boldsymbol{\pi}}'\} \in \Omega$, $\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}'$ implies that $\pi_0 = \pi_0'$. Thus for a conditionally identifiable model, there exists a function $\mathcal{T}\colon \tilde{T} \to (0,1)$ that maps observed cell probabilities, $\tilde{\boldsymbol{\pi}}$, to unobserved cell probabilities, π_0, where $\tilde{T} \subset \mathbb{S}^{2^K-2}$. When the domain $\tilde{T}$ of this function is not equal to $\mathbb{S}^{2^K-2}$, this restricts the set of possible values for $\tilde{\boldsymbol{\pi}}$ in the model to $\tilde{T}$. Any extra assumptions in the model involving $\tilde{\boldsymbol{\pi}}$ can then further restrict the set of possible values for $\tilde{\boldsymbol{\pi}}$ in the model to a set $\tilde{S} \subset \tilde{T}$. Thus conditionally identifiable models take the form $\mathcal{P}_\Omega$, where $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$. When a model is not conditionally identifiable, we have no guarantees for when the parameters are estimable in any traditional sense. In particular, non-identifiability precludes consistent estimation as "there will be uncertainty in parameter estimates that is not washed out as more data are collected" [78]. If a model $\mathcal{P}_\Omega$ is conditionally identifiable, all parameters of the model can be consistently estimated [118]. However, we emphasize that the data needs to have been generated by a distribution in the model $\mathcal{P}_\Omega$ for the parameters to be consistently estimable. In other words, in order to estimate the population size N, we need to assume a functional relationship, $\mathcal{T}$, between the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ and the unobserved cell probability π_0. This is the main idea behind MSE.

3.2.5 Missing Data

The framing in the previous section is motivated by our treatment of MSE as a missing data problem. The decomposition in (3.2) is related to the decomposition in the missing data literature of the complete-data distribution into the *extrapolation distribution* and the *observed-data distribution* [60]. The extrapolation distribution captures how to extrapolate to the missing data given the observed data, which in our context corresponds to L_1. The observed-data distribution, as the name indicates, is the distribution of the observed data, which in this context corresponds to L_2. Following the analogy of the missing data literature, by restricting ourselves to models of the form $\mathcal{P}_\Omega$, where $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 =$

$\mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$, one is making an *identifying assumption*, $\mathcal{T}$, about how $\tilde{\boldsymbol{\pi}}$ relates to π_0 in order to identify π_0.

The observed-data distribution is restricted when the set of possible values for the observed cell probabilities, $\tilde{S}$, is not equal to $\mathbb{S}^{2^K-2}$. Based on standard properties of the multinomial conditional likelihood, restrictions on the observed-data distribution are assumptions that are testable from the data. As noted in the previous section, these restrictions could be due to the domain, $\tilde{T}$, of the identifying assumption (see Section 3.4.3 for an example), or due to extra modeling assumptions for the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$ (see Section 3.3.1 for an example). This motivates the following definition [see Chapter 8 of 60]:

Definition 3.3. *A model $\mathcal{P}_\Omega$, where $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$, is nonparametric identified when $\tilde{S} = \tilde{T} = \mathbb{S}^{2^K-2}$, i.e. the observed-data distribution is not restricted by the model.*

3.2.6 Our Approach to Multiple-Systems Estimation

In the MSE literature, previous work has been concerned with determining *when* certain models are conditionally identified [see e.g. 79, 61]. Here we are concerned with determining both when *and how* models are conditionally identified. Since the validity of our inferences rests on the untestable identifying assumption and any restrictions on the observed-data distribution being correct, we would like to know what identifying assumption we are actually making so we can determine whether or not the assumption is plausible in a given context. Thus, in this chapter our approach to MSE will be to use conditionally identified models that are based on *explicitly specified* identifying assumptions. Additionally, to make as few testable assumptions as possible, we will use models where the observed-data distribution is only possibly restricted by the identifying assumption (i.e. $\tilde{S} = \tilde{T}$).

Given such a conditionally identified model, our approach to MSE is agnostic to the inferential framework used, so one can perform inference for N in a frequentist or Bayesian framework. In Appendix B, we outline how computation, including sensitivity analyses

probing the identifying assumption as we will describe in Section 3.4, can be carried out in either framework using existing software.

In the rest of this chapter, we examine the identifying assumptions (and sometimes lack thereof) associated with commonly used MSE models, and propose a new family of identifying assumptions. While these identifying assumptions may be useful in some applications, *there is no one-size-fits-all solution.* In practice, the use of identifying assumptions should be accompanied by appropriate justification based on the context of the data. However, in some applications none of the identifying assumptions discussed in this chapter will be appropriate for the data at hand. There is no default identifying assumption that practitioners can fall back on, and so in these scenarios *no estimate of the population size should be produced based on the data at hand.* Such a scenario is clearly unsatisfactory, and thus it is an important task for researchers in the field of MSE to develop new explicit identifying assumptions, so that practitioners are able to select identifying assumptions appropriate for their applications.

3.3 Log-Linear and Latent Class Models

In this section we describe two commonly used models, which we use to demonstrate the drawbacks of using models that either place unnecessary restrictions on the observed-data distribution or that are not based on explicit identifying assumptions.

3.3.1 Log-Linear Models

For $\boldsymbol{h} \in H^*$, let h_k denote the kth element of $\boldsymbol{h}$. Any set of cell probabilities, $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, can be represented as $\pi_{\boldsymbol{h}} = \mu_{\boldsymbol{h}} / \sum_{\boldsymbol{h}' \in H} \mu_{\boldsymbol{h}'}$, where $\log(\mu_{\boldsymbol{h}}) = \sum_{\boldsymbol{h}' \in H^*} \lambda_{\boldsymbol{h}'} \prod_{k=1}^{K} h_k^{h'_k}$, for some set of log-linear parameters $\boldsymbol{\lambda} = \{\lambda_{\boldsymbol{h}}\}_{\boldsymbol{h} \in H^*} \in \mathbb{R}^{2^K-1}$. This leads to the *log-linear parameterization* $\Theta_{LL} = \{N, \boldsymbol{\lambda} \mid N \in \mathbb{N}, \boldsymbol{\lambda} \in \mathbb{R}^{2^K-1}\}$. Note that under this parameterization, there is no $\lambda_{(0,\cdots,0)}$, so that $\mu_{(0,\cdots,0)} = 1$.

For cells in the incomplete table $\boldsymbol{h} \in H^*$ such that $\sum_{k=1}^{K} h_k = 1$ we refer to $\lambda_{\boldsymbol{h}}$ as a main effect; for $\boldsymbol{h} \in H^*$ such that $\sum_{k=1}^{K} h_k = \ell > 1$ we refer to $\lambda_{\boldsymbol{h}}$ as an ℓ-way interaction. The main effects and interactions all have interpretations as log ratios of certain cross-product ratios

[see e.g. Chapter 2 of 21]. Of particular interest is the K-way, or highest-order, interaction $\lambda_{\mathbf{1}}$, where $\mathbf{1} := (1, \cdots, 1)$, for which we have the relationship $\prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})} / \prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = \exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\}$, where $I_{odd}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is odd})$ and $I_{even}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is even})$, using the convention that 0 is even. This notation differs from [21] as we index the complete table by $H = \{0, 1\}^K$ rather than $\{2, 1\}^K$.

The model $\mathcal{P}_{\Theta_{LL}}$ is equivalent to the unrestricted model $\mathcal{P}_{\Theta}$, so we need to restrict Θ_{LL} to identify the unobserved cell probability π_0. It is standard in this scenario to set $\lambda_{\mathbf{1}} = 0$, so that there is no highest-order interaction in the model. Referring to the resulting parameter space as Ω_{LL}, we would like to understand the identifying assumption made by the *saturated model* $\mathcal{P}_{\Omega_{LL}}$. In Appendix B, we show $\mathcal{P}_{\Omega_{LL}}$ is nonparametric identified and that this no-highest-order interaction (NHOI) assumption corresponds to the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})/(1 + \tilde{\Pi}_{odd}/\tilde{\Pi}_{even})$, where $\tilde{\Pi}_{odd} = \prod_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}$ and $\tilde{\Pi}_{even} = \prod_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}$, which we discuss in more detail in Section 3.4.

In practice there is an emphasis on achieving low variance estimates of the log-linear parameters and, consequentially, N. To this end, rather than just setting the highest-order interaction to zero and using the saturated model, it is common to further restrict the model and set other interactions to zero. This is the case, for example, when restricting to decomposable graphical models [86], or when only including main effects and 2-way interactions [119], which can be hard to justify in practice [see e.g. 32, 139]. This restricts the observed-data distribution, so that we are making a testable assumption that, in addition to the untestable identifying assumption, must be correct in order for inferences to be valid. The hope is that by specifying a model with fewer parameters, the resulting estimates will have lower variance if the chosen restricted model generated the data. However, if the chosen restricted model did not generate the data, estimates of N can be arbitrarily biased, and more generally can have arbitrarily poor frequentist properties [106, 139].

This is a classic bias-variance trade off, which has been acknowledged since the seminal work of [45] (edited to match our notation): "In analyzing multiple recapture census data our aim is to fit the incomplete 2^K table by a log linear model with the fewest possible

parameters, since the fewer parameters in an 'appropriate' model for estimating n_0, the smaller the variance of the estimate. Thus it is not a good practice simply to use the saturated model. On the other hand, if we use a model with too few parameters, we introduce a bias into our estimate of population size that can possibly render the variance formulae of the next section meaningless." Unlike [45], we believe there is a clear route to take if one is using the NHOI assumption, in line with our approach described in Section 3.2.6: make as few testable assumptions as possible (i.e. use the saturated model $\mathcal{P}_{\Omega_{LL}}$) in the hopes of not being arbitrarily biased because of incorrect restrictions on the observed data distribution. If one does wish to produce lower variance estimators, we discuss in Appendix B how regularization can be used to reduce the variance of estimates, at the cost of increasing the bias of estimates, and some difficulties associated with using regularized estimators.

3.3.2 Latent Class models

Latent class models (LCMs) are typically motivated as models of multivariate categorical data that capture individual heterogeneity when the population can be stratified into J classes, where lists sample individuals independently within each class [55, 87]. Thus they are so-called M_{th} models as described in Appendix B [101]. Corollary 1 of [35] shows that for any set of cell probabilities $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, there exists some $J < \infty$ such that $\boldsymbol{\pi}$ can be represented as a J-class latent class model, i.e. $\pi_{\boldsymbol{h}} = \sum_{j=1}^{J} \nu_j \prod_{k=1}^{K} q_{jk}^{h_k}(1-q_{jk})^{1-h_k}$, where $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_J)$ are class membership probabilities, and $\boldsymbol{q} = \{q_{jk}\}_{j=1,k=1}^{J,K}$ are class specific observation probabilities for each list. This leads to the *latent class model parameterization* $\Theta_{LCM} = \{N, \boldsymbol{\nu}, \boldsymbol{q}, J \mid N \in \mathbb{N}, \boldsymbol{\nu} \in \mathbb{S}^{J-1}, \boldsymbol{q} \in (0,1)^{J\times K}, J \in \mathbb{N}\}$. As $\mathcal{P}_{\Theta_{LCM}}$ is equivalent to the unrestricted model $\mathcal{P}_{\Theta}$, we need to restrict Θ_{LCM} to identify the unobserved cell probability π_0. It is common to fix the number of latent classes, J, in advance, to arrive at the the restricted parameterization $\Omega_{LCM,J} = \{N, \boldsymbol{\nu}, \boldsymbol{q} \mid N \in \mathbb{N}, \boldsymbol{\nu} \in \mathbb{S}^{J-1}, \boldsymbol{q} \in (0,1)^{J\times K}\}$.

In Appendix B we show that $\mathcal{P}_{\Omega_{LCM,J}}$ is conditionally identified if and only if $2J \leq K$. However, when $\mathcal{P}_{\Omega_{LCM,J}}$ *is* conditionally identified we do not know what explicit identifying assumption is being made or whether the model is nonparametric identified. A recent de-

velopment in MSE is the use of LCMs with J large enough that $2J > K$ [87]. Such LCMs with too many latent classes (i.e. $2J > K$) suffer from the opposite problem of log-linear models: rather than making too many assumptions, and hence restricting the observed-data distribution, so few assumptions are being made that the model is not conditionally identified. In Appendix B we show through a variety simulations that this is a practically relevant problem, as we have no guarantees for when estimates based on non-identified models are going to be accurate.

3.4 Revisiting Log-Linear Models and Their Identifying Assumptions

In this section we revisit the NHOI identifying assumption associated with log-linear models and discuss its role in our framing of MSE. We then describe how this assumption can be used as a building block for alternative identifying assumptions.

3.4.1 The No-Highest-Order Interaction Assumption

The NHOI assumption introduced in Section 3.3.1 can be interpreted as follows: for any given subset of $K-1$ lists, appearing in all $K-1$ lists is not associated with appearing or not appearing in the Kth list. Here the meaning of "associated with" changes as the number of lists K changes. When $K = 2$ we are assuming that the odds of appearing in list 1 conditional on appearing in list 2 is equal to the odds of appearing in list 1 conditional on not appearing in list 2, and thus the lists are independent: $\pi_{(1,0)}/\pi_{(0,0)} = \pi_{(1,1)}/\pi_{(0,1)}$. When $K = 3$ we are assuming that the odds ratio for lists 1 and 2 conditional on appearing in list 3 is equal to the odds ratio for lists 1 and 2 conditional on not appearing in list 3: $\pi_{(1,1,1)}\pi_{(0,0,1)}/(\pi_{(1,0,1)}\pi_{(0,1,1)}) = \pi_{(1,1,0)}\pi_{(0,0,0)}/(\pi_{(1,0,0)}\pi_{(0,1,0)})$. When $K = 4$ we assume that certain ratios of odds ratios are equal, and so on for larger K.

As discussed in Section 3.2.6, in order to use the NHOI assumption in a given application, we need to be able to determine whether or not it is plausible. Odds and odds ratios are commonly used in statistics [21], and thus the NHOI assumption may be of use when there are $K = 2$ or $K = 3$ lists. However, higher order measures of association like ratios of odds

ratio are more obscure and hard to interpret, which makes the NHOI assumption difficult to use when there are more than $K = 3$ lists. This difficulty compounds when considering sensitivity analyses as we explain in the next section.

3.4.2 Sensitivity Analyses for the No-Highest-Order Interaction Assumption

Sensitivity analyses aim to gauge how sensitive inferences are to untestable assumptions, and are an important part of missing data workflows [see Chapter 9 of 60]. The NHOI assumption facilitates sensitivity analyses based on varying the highest-order interaction across a range of non-zero values. In particular, when fixing $\xi = \exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\} \in \mathbb{R}^+$, we show in Appendix B that we arrive at the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})/(\xi+\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})$. This generalizes the two list sensitivity analyses of [85] and [51]. Under this identifying assumption, rather than assuming certain measures of association are equal, we are assuming one measure is ξ times another. For example, when $K = 2$ we are assuming that the odds of appearing in list 1 conditional on not appearing in list 2 is ξ times the odds of appearing in list 1 conditional on appearing in list 2: $\pi_{(1,0)}/\pi_{(0,0)} = \xi\pi_{(1,1)}/\pi_{(0,1)}$.

In order to perform a meaningful sensitivity analysis, one needs to be able to specify a range of values for the highest-order interaction that are plausible for a given application. Due to our understanding of odds and odds ratios, performing this sort of sensitivity analysis may be possible when there are $K = 2$ or $K = 3$ lists. When considering more than $K = 3$ lists, it can become difficult to even start thinking about whether it is plausible that ξ is less than or greater than 1, let alone determine specific values of ξ that are plausible.

3.4.3 K'-List Marginal No-Highest-Order Interaction Assumptions

The NHOI assumption can be used as a building block to generate other identifying assumptions. Suppose we can assume that, without loss of generality, the NHOI assumption holds for the first $1 < K' < K$ lists, marginal of the remaining $K - K'$ lists. This leads to a new identifying assumption which in general does not imply that there is no highest-order interaction for all K lists. To introduce this assumption formally we need to introduce

some notation. Let $G = \{0,1\}^{K'}$ index the marginal $2^{K'}$ contingency table for the first K' lists and $G^* = G \setminus \{0\}^{K'}$. For a set of cell probabilities, $\boldsymbol{\pi} \in \mathbb{S}^{2^K-1}$, and a given cell in the marginal table, $\boldsymbol{g} \in G$, let $\pi_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$ denote the probability of being observed in cell $\boldsymbol{g}$ of the marginal table implied by $\boldsymbol{\pi}$. Similarly let $\tilde{\pi}_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$ and $\tilde{\pi}_{0+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = (0,\cdots,0)\}$.

Assuming that the NHOI assumption holds for the first $1 < K' < K$ lists, marginal of the remaining $K - K'$ lists, is equivalent to assuming $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = 1$. In Appendix B we show that this K'-list marginal no-highest-order interaction assumption corresponds to the explicit identifying assumption $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$, where $\tilde{\Pi}_{odd,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. Further, we can perform sensitivity analyses for this assumption by fixing $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi \in \mathbb{R}^+$. As we show in Appendix B, this leads to the explicit identifying assumption

$$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = \frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+}}{\xi + (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+})}. \tag{3.3}$$

Models that use the assumption that $\prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})} / \prod_{\boldsymbol{g}\in G} \pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi \in \mathbb{R}^+$ are not non-parametric identified, as the domain of the identifying assumption is $\tilde{T} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > \xi\}$.

A special case of this identifying assumption was originally suggested in [107] as an alternative to the NHOI assumption. They considered a data set consisting of $K = 3$ lists recording cases of spina bifida in upstate New York, where they believed that the assumption that two of the lists were marginally independent (i.e., using the 2-list marginal NHOI assumption) was more plausible than the NHOI assumption. This illustrates that there may be applications where one may be more willing to make marginal assumptions about a subset of K' lists, rather than an assumption involving all K lists. Additionally when there are $K > 3$ lists and $K' = 2$ or $K' = 3$, the K'-list marginal NHOI assumption and its sensitivity analyses are much more straightforward to interpret than the highest-order interaction and its sensitivity analyses, as discussed in Sections 3.4.1 and 3.4.2.

For these reasons, we believe that the K'-list marginal NHOI assumption can be useful

as an explicit identifying assumption in the toolbox of the MSE practitioner. However, we emphasize here our message from Section 3.2.6: there are no one-size-fits-all identifying assumptions. Specification of identifying assumptions in practice should be accompanied with appropriate justification based on the context of the data. In Section 3.5.1 we attempt to provide such a justification for our use of the 2-list marginal NHOI assumption in an application estimating the number of civilian casualties in the Kosovo war.

3.5 Civilian Casualties in the Kosovo War

In this section we estimate the number of civilian casualties in the Kosovo war between March 20 and June 22, 1999, using data originally analyzed in [9]. The data consist of $K = 4$ lists with $n = 4400$ observed casualties, and are presented in Table 3.1, reproduced from Section 6 of [9]. Three of the lists were constructed from refugee interviews conducted separately by the American Bar Association Central and East European Law Initiative (ABA), Human Rights Watch (HRW), and the Organization for Security and Cooperation in Europe (OSCE). The fourth list was constructed from exhumation reports conducted on behalf of the International Criminal Tribunal for the Former Yugoslavia (EXH). We refer the reader to Appendix 1 of [9] for a detailed description of each list.

Table 3.1: Kosovo dataset, reproduced from Section 6 of [9].

	ABA	yes	yes	no	no
	EXH	yes	no	yes	no
HRW	OSCE				
yes	yes	27	32	42	123
yes	no	18	31	106	306
no	yes	181	217	228	936
no	no	177	845	1131	n_0

The Kosovo data was originally analyzed in [9] under the NHOI assumption, but as

we discuss in the next section, we believe the K'-list marginal NHOI assumption is more appropriate. We will analyze the Kosovo data under both assumptions, highlighting the importance of careful specification of the identifying assumption.

3.5.1 Choice of Identifying Assumption

For our main analysis we will consider two identifying assumptions. The first assumption is the 2-list marginal NHOI assumption described in Section 3.4.3, where we will assume that the ABA and HRW lists are marginally independent. We believe this assumption is plausible given that "there were no overt efforts by any of the researchers to exclude or include witnesses who had participated in another data collection project" [1, p. 40] and that the two lists had similarly extensive geographic reach in their interviews. In particular, ABA conducted interviews in Albania, Macedonia, Kosovo, the United States, and Poland, while HRW conducted interviews in Albania, Macedonia, Kosovo, and Montenegro. ABA only conducted around 10% of its interviews in the United States and Poland, and HRW only conducted 3% of its interviews in Montenegro. Further, within Kosovo, ABA and HRW conducted interviews in similar geographic regions. For more information on where the lists conducted interviews see Appendix 1 of [9].

The original analysis of the Kosovo data set in [9] used the NHOI assumption described in Section 3.3.1. To justify this assumption for the Kosovo data, as we have $K = 4$ lists, we would need to reason about certain ratios of odds ratios being equal, which can be difficult, as discussed in Section 3.4.1 and further explained in Appendix B. Nevertheless, we will also analyze the Kosovo data using the NHOI assumption to highlight the importance of careful specification of the identifying assumption.

3.5.2 Inference

For each identifying assumption, our main analysis will present both a frequentist analysis and a Bayesian analysis, using the methods discussed in Appendix B, to demonstrate how our proposed approach to MSE is agnostic to the inferential framework used. The Bayesian

analysis will use a negative-binomial prior for N and the prior induced for the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ from using the Dirichlet process prior of [87] for the J class LCM $\Omega_{LCM,J}$, with $J = 10$ and default hyperparameters, as implemented in the `R` package `LCMCR` (see Appendix B for further details). In Appendix B we perform a prior sensitivity analysis for the Bayesian analyses, exploring the impact of the priors for N and $\tilde{\boldsymbol{\pi}}$ on our estimates of N.

To inform the negative-binomial prior for N, we will rely on two studies that attempted to estimate the number of casualties in the Kosovo war using different data sources than [9]. [123] estimated there were 12000 casualties with a 95% confidence interval of $[5500, 18300]$. [65] estimated there were 8000 casualties with a 95% confidence interval of $[5800, 10200]$. Using the negative-binomial parameterization given in Table 1 of Appendix B, we will use a specification with mean $M = 10000$ (the average of the estimates from the two studies) and overdispersion parameter $a = 1.6$, which places 95% of the prior mass on $[818, 30371]$. This specification is meant to be weakly informative in the sense that the information it incorporates is intentionally weaker than what is available to us, so as to provide a proper alternative to the "noninformative" improper scale prior $p(N) \propto 1/N$ discussed in Appendix B [see e.g. 50]. This prior places mass below the observed sample size of $n = 4400$, as we are not attempting to use the observed data to inform our prior. Practically speaking this does not make a difference, as the prior is effectively truncated to $[n, \infty)$ when performing posterior inference.

3.5.3 Main Analysis

In Table 3.2 we present the results from our frequentist and Bayesian analyses under the 2-list marginal NHOI assumption, i.e. assuming marginal independence of the ABA and HRW lists. Assuming marginal independence of the ABA and HRW lists, under a frequentist analysis we estimate there were 9691 civilian casualties, with a 95% confidence interval of $[8074, 11308]$, and under a Bayesian analysis with the chosen priors we estimate there were 9359 civilian casualties, with a 95% credible interval of $[7967, 11059]$. These point

estimates and uncertainty intervals from these two analyses are in close agreement. Both of the uncertainty intervals include the point estimate from [65], but not from [123], and fall within the confidence interval of [123]. Based on the results of the prior sensitivity analysis in Appendix B, the Bayesian analysis is not sensitive to the prior choices for N and $\tilde{\boldsymbol{\pi}}$.

Table 3.2: Point estimates and 95% uncertainty intervals for N under the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean.

	Point Estimate	Uncertainty Interval
Frequentist	9691	[8074, 11308]
Bayesian	9359	[7967, 11059]

In Table 3.3 we present the results from our frequentist and Bayesian analyses under the NHOI assumption. Under the NHOI assumption, under a frequentist analysis we estimate there were 16941 civilian casualties, with a 95% confidence interval of $[5304, 28579]$, and under a Bayesian analysis with the chosen priors we estimate there were 14071 civilian casualties, with a 95% credible interval of $[9321, 21604]$. The point estimates and uncertainty intervals from these two analyses are in relative agreement. Both of the uncertainty intervals include the point estimate from [123], and the frequentist confidence interval includes the point estimate. Based on the results of the prior sensitivity analysis in Appendix B, the Bayesian analysis is fairly sensitive to the prior choices for N and $\tilde{\boldsymbol{\pi}}$.

Table 3.3: Point estimates and 95% uncertainty intervals for N under the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean.

	Point Estimate	Uncertainty Interval
Frequentist	16941	[5304, 28579]
Bayesian	14071	[9321, 21604]

Focusing on point estimates, we see a large difference between the analyses under the two

identifying assumptions (besides the uncertainty interval widths being considerably larger under the NHOI assumption). The point estimates under the NHOI assumption are 75% larger for the frequentist analyses (50% larger for the Bayesian analyses) than the point estimates under the 2-list marginal NHOI assumption. If the 2-list marginal NHOI assumption truly holds, as we are inclined to believe based on the justification provided in Section 3.5.1, an analysis based on using the NHOI assumption produces estimates with a large positive bias for the Kosovo data. This should serve as an illustration of the dangers of using the NHOI assumption (or any other identifying assumption) that can not be justified based on the context of the data. If a practitioner can not find an identifying assumption that is appropriate for their data, no estimate of the population size should be produced based on their data, as there is no one-size-fits-all or default identifying assumption to fall back on. There is a need for researchers to develop new explicit identifying assumptions, so that practitioners do not find themselves in such a scenario.

3.5.4 A Sensitivity Analysis Probing the 2-List Marginal NHOI Assumption

While we believe that it is plausible that the ABA and HRW lists are marginally independent, we would also like to understand how sensitive our resulting estimates are to realistic violations of the assumption. If this marginal independence was violated, it would likely be the case that the lists are positively dependent and thus population size estimates under marginal independence are downward biased, as is common in human rights applications [see e.g. the discussion in Section 5 of 85]. In particular, HRW selected regions in Kosovo to conduct interviews based on reports of human rights violations from refugees and other sources [1]. Thus it seems possible that a casualty appearing in HRW could be more likely to appear in ABA than a casualty that did not appear in HRW.

We now perform a sensitivity analysis probing the 2-list marginal NHOI assumption. In Appendix B, we provide a similar sensitivity analysis probing the NHOI assumption. We will consider models with the identifying assumption (3.3), varying ξ over $\{0.7, 0.8, 0.9, 1\}$. Thus in each case we are assuming that the odds of appearing in ABA conditional on not

appearing in HRW is ξ times the odds of appearing in ABA conditional on appearing in HRW, with $\xi = 1$ corresponding to the 2-list marginal NHOI assumption. For each value of ξ, we will present both a frequentist analysis and a Bayesian analysis, with the Bayesian analysis using the same priors from the main analysis as presented in Section 3.5.2. In Table 3.4 we present the results from our frequentist and Bayesian analyses under each identifying assumption.

Table 3.4: Point estimates and 95% uncertainty intervals for sensitivity analysis probing the 2-list marginal NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a marginal odds ratio, as described in Section 3.4.3.

	$\xi = 1$	$\xi = 0.9$	$\xi = 0.8$	$\xi = 0.7$
Frequentist	9691 [8074, 11308]	10534 [8738, 12330]	11588 [9568, 13607]	12942 [10636, 15249]
Bayesian	9359 [7967, 11059]	10155 [8607, 12038]	11147 [9419, 13258]	12419 [10451, 14816]

The estimates of the number of casualties N increase as the amount of assumed positive dependence increases, i.e. as ξ decreases, as expected. When $\xi = 0.9$, the point estimates and uncertainty intervals are still largely compatible with the point estimates and uncertainty intervals under marginal independence. Thus our estimates under marginal independence are not sensitive to this small amount of positive dependence. However, this is not still the case under stronger positive dependence. When $\xi = 0.7$, the uncertainty intervals barely overlap with the uncertainty intervals under marginal independence, and further they do not contain the point estimates under marginal independence. While this may seem like cause for concern, we note that these estimates under stronger positive dependence are still within an order of magnitude of the estimates under independence, and all uncertainty intervals in this sensitivity analysis fall within the confidence interval of [123]. We note that the frequentist analysis requires a marginal odds ratio of $\xi \approx 0.51$ to produce a point estimate as large as the point estimate under the NHOI assumption. This is a large amount of positive dependence which casts further doubt on the plausibility of the NHOI assumption.

3.6 Discussion

In this chapter we revisited the framing of MSE as a missing data problem and proposed an approach for MSE that places the identifying assumption front and center in the MSE workflow. As we have emphasized throughout this chapter, a natural next step is to develop new explicit identifying assumptions, for situations where the identifying assumptions described in Section 3.4 can not be justified in the context of a given data set. We believe that this is an extremely under-researched problem that will hopefully gain attention with the re-framing of MSE we present in this chapter.

The presentation of MSE in this chapter was focused on estimating the size of a single population. When the population can be stratified based on observed covariates, such as location or time, it may be desirable to estimate the population sizes within each strata. In theory, the methodology developed in this chapter could be applied independently to each strata. However, stratification can lead to sparse contingency tables, which need significant regularization when estimating $\tilde{\boldsymbol{\pi}}$. In this case, it would be desirable to develop observed data models that borrow strength across strata.

Chapter 4

ADAPTIVE GAUSSIAN MARKOV RANDOM FIELDS FOR CHILD MORTALITY ESTIMATION

4.1 Introduction

The under-5 mortality rate (U5MR) is an important statistic in understanding the health of a country. This is highlighted by the United Nations (UN) Sustainable Development Goals (SDGs) in SDG 3.2, which states "By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births" [134]. Due to a lack of vital registration systems in many lower and middle income countries, U5MR is typically estimated from household surveys, like the Demographic and Health Surveys (DHS). The reliable estimation of U5MR from such household surveys at fine spatio-temporal scales require the usage of smoothing models which borrow information across space and time.

The statistical methods for estimating U5MR over space and/or time typically accomplish spatio-temporal smoothing through the use of random effects based on Gaussian Markov random fields (GMRFs) or closely related models. In particular, the United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME) produces national level estimates of U5MR yearly using the Bayesian B-spline bias-reduction (B3) method [3, 4], which uses smoothing splines that have well-known connections to GMRFs [see e.g. 143], and has supported and produced subnational estimates of U5MR using GMRFs [76, 136]. The assumptions underlying these GMRF-based smoothing models may not be realistic for estimating U5MR when certain time periods or regions are expected to have shocks in mortality relative to their neighbors. We are motivated in this paper by U5MR estimation in two contexts

where such shocks in mortality could occur: 1) Rwanda, where a civil war and genocide took place in the mid 1990's, and 2) multi-country models, which have become more common in the global health literature [24], where we simultaneously estimate U5MR subnationally across multiple countries. In such scenarios, GMRF-based smoothing models may lead to oversmoothing of U5MR estimates in certain time periods or regions.

In this chapter, we develop smoothing models which incorporate knowledge of expected shocks in mortality, but still allow information to be borrowed across space and time. We first discuss our motivating applications and review GMRFs and U5MR estimation in Section 4.2. In Section 4.3, we extend commonly used GMRFs to allow the incorporation of knowledge of expected shocks in mortality, which we call adaptive Gaussian Markov random fields (AGMRFs). Section 4.4 provides details of implementing our AGMRFs in practice. Section 4.5 presents a simulation study assessing how AGMRFs can improve the performance of a model used to estimate U5MR. Finally, we apply our AGMRFs to estimate U5MR in Rwanda in Section 4.6 and in a multicountry setting in Section 4.7.

4.2 Background and Motivating Application

In this section we first review GMRFs and U5MR estimation using the smoothed direct model of [94], and then provide two motivating applications involving U5MR.

4.2.1 Gaussian Markov Random Fields

Statistical models for child mortality estimation typically involve random effects in space and/or time, as is the case for the smoothed direct model of [94] we review in the following section. In most cases, these random effects are Gaussian Markov random fields (GMRFs), Gaussian random vectors defined on labelled graphs where sparsity in the precision matrix implies certain conditional independence properties [111]. Typically these spatial and temporal random effects are improper GMRFs. In this section we review the random walk of first order temporal random effect and the intrinsic conditional autoregression spatial random effect. For a full review of GMRFs, improper GMRFs, and their various properties, we

refer the reader to Chapters 2 and 3 of [111].

Random Walk of First Order

Suppose we have N time periods and we are specifying a structured temporal random effect $\boldsymbol{x} = (x_1, \ldots, x_N)$. Suppose for $i = 1, \ldots, N-1$ we specify

$$x_{i+1} \mid x_i, \tau \sim \mathsf{Normal}(x_i, \tau^{-1})$$

where τ is the precision for each transition. This is referred to as a random walk of first order or random walk 1 (RW1), which we denote as $\boldsymbol{x} \sim \mathsf{RW1}(\tau)$ [see e.g. page 95 of 111]. This leads to a density for $\boldsymbol{x}$ of

$$p(\boldsymbol{x} \mid \tau) \propto \tau^{(N-1)/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N-1}(x_{i+1} - x_i)^2\right\} = \tau^{(N-1)/2} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x}\right\},$$

where $Q = \tau R$ is a precision matrix determined by the structure matrix R with

$$R_{ij} = \begin{cases} 1, & \text{if } i = j = 1 \text{ or } i = j = N \\ 2, & \text{if } i = j = k \text{ and } k \notin \{1, N\} \\ -1, & \text{if } j = i + 1 \text{ or } i = j + 1 \\ 0, & \text{if } |i - j| > 1. \end{cases}$$

It can be verified that $Q\mathbf{1} = \mathbf{0}$, and so Q is rank $N-1$ and thus RW1s are improper GMRFs. It follows that $p(\boldsymbol{x} \mid \tau)$ is invariant to an addition of a constant vector to $\boldsymbol{x}$, thus when a RW1 is included in a model with an intercept we enforce the sum-to-zero constraint $\sum_{i=1}^{N} x_i = 0$.

Intrinsic Conditional Autoregression

Suppose we have N areal units and we are specifying a structured spatial random effect $\boldsymbol{x} = (x_1, \ldots, x_N)$. We will assume there are no islands, i.e. we assume the graph of the areal units is connected, see [49] for more details when there are islands. Suppose for neighboring regions $i \sim j$ we specify

$$x_i - x_j \mid \tau \sim \mathsf{Normal}(0, \tau^{-1})$$

where τ is the precision for each difference. This is referred to as an intrinsic conditional autoregression (ICAR), which we denote as $\boldsymbol{x} \sim \mathsf{ICAR}(\tau)$ [see e.g. page 101 of 111]. This leads to a density for $\boldsymbol{x}$ of

$$p(\boldsymbol{x} \mid \tau) \propto \tau^{(N-1)/2} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x}\right\}$$

where $Q = \tau R$ is a precision matrix determined by the structure matrix R with

$$R_{ij} = \begin{cases} n_i, & \text{if } i = j \\ -1, & \text{if } i \sim j \\ 0, & \text{otherwise,} \end{cases}$$

where n_i is the number of regions neighbouring i. It can be verified that $Q\mathbf{1} = \mathbf{0}$, and so Q is rank $N - 1$ and thus ICARs are improper GMRFs. It follows that $p(\boldsymbol{x} \mid \tau)$ is invariant to an addition of a constant vector to $\boldsymbol{x}$, thus when a ICAR is included in a model with an intercept we enforce the sum-to-zero constraint $\sum_{i=1}^{N} x_i = 0$. As time periods can be viewed as areal units with a specific neighborhood structure, RW1s are special cases of ICARs.

4.2.2 The Smoothed Direct Model

In this section we will review the smoothed direct model of [94] for child mortality estimation, a recent extension of the seminal Fay–Herriot model [41]. Suppose for a country of interest we have S surveys of full birth histories, each of which can be used to produce direct estimates of U5MR, i.e. survey weighted estimates of U5MR with associated design-based standard errors which use the full birth histories of children [94]. For ease of exposition we will focus on estimating U5MR either at the national level over multiple time periods, or at the subnational level in one time period. While one could use these direct estimates as estimates of U5MR, the temporal or spatial disaggregation of the data can lead to noisy estimates with large standard errors. It is thus desirable to borrow information across time or space to smooth these estimates.

Let N denote either the number of time periods in a national level model or the number of administrative regions in a subnational level model, which we will refer to generically as subdivisions. For $s \in [S]$ and $i \in [N]$, let $\hat{p}_{is}$ denote the direct estimate of U5MR in subdivision i from survey s. Let $y_{is} = \text{logit}\,(\hat{p}_{is})$, and let $\hat{V}_{is}$ denote the design-based standard error associated with y_{is}. [94] used the asymptotic distribution of y_{is} as a working likelihood:

$$y_{is} \mid \eta_{is} \sim \mathsf{Normal}(\eta_{is}, \hat{V}_{is}).$$

Smoothing of the direct estimates over the subdivisions i is accomplished through a prior model for η_{is},

$$\eta_{is} = \mu + v_i + x_i + \nu_s, \tag{4.1}$$

where $v_i \overset{iid}{\sim} \mathsf{Normal}(0, \sigma_v^2)$ is an unstructured subdivision-level random effect, $\nu_s \overset{iid}{\sim} \mathsf{Normal}(0, \sigma_\nu^2)$ is a survey random effect (if S is small this can be replaced by a fixed effect with a sum-to-zero constraint), and $\boldsymbol{x}$ is a structured subdivision-level random effect, either $\boldsymbol{x} \sim \mathsf{RW1}(\tau_x)$ if we are working with the national level model or $\boldsymbol{x} \sim \mathsf{ICAR}(\tau_x)$ if we are working with the subnational level model. Priors are then set on the intercept μ and all of the precision parameters. Smoothed estimates of U5MR in subdivision i are then based on the posterior $p(p_i \mid \{y_{is}\}_{is})$, where

$$p_i = \text{expit}(\mu + v_i + x_i).$$

The total subdivision-level random effect in the smoothed direct model, $\boldsymbol{b} = \boldsymbol{v} + \boldsymbol{x}$, which is the sum of an unstructured random effect and a structured random effect, is due to Besag, York, and Mollié (BYM) [14]. We will reparameterize the total subdivision-level random effect, $\boldsymbol{b}$, as a so-called BYM2, following [110] and [120]. In this parameterization, rather then representing the total subdivision-level random effect $\boldsymbol{b}$ as a a sum of an unstructured random effect, $\boldsymbol{v}$, and a structured random effect, $\boldsymbol{x}$, with independent precision parameters, we will represent the total subdivision-level random effect as $\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}(\sqrt{1-\phi}\boldsymbol{v} + \sqrt{\phi}\boldsymbol{x}^*)$, where $v_i \overset{iid}{\sim} \mathsf{Normal}(0, 1)$ and $\boldsymbol{x}^* \sim \mathsf{Normal}(0, R_\star^-)$. Here $R_\star$ is the structure matrix of the RW1 or ICAR scaled as in [122], and $R_\star^-$ is the generalized inverse of $R_\star$. $1/\tau_b$ can then be

interpreted as the marginal variance of $\boldsymbol{b}$, and ϕ $(1-\phi)$ can be interpreted as the fraction of the variance explained by the structured (unstructured) random effect. Penalized complexity priors [120] are then adopted for τ_b and ϕ.

The smoothed direct model is a latent Gaussian model, i.e. the data, y_{is}, are conditionally independent given a latent Gaussian vector, $\{\eta_{is}\}_{is}$, with a small number of hyperparameters. Thus, the smoothed direct model can be fit using the Integrated Nested Laplace Approximation (INLA) [112], a deterministic alternative to Markov chain Monte Carlo (MCMC) methods for fitting Bayesian hierarchical models, using the `R` package `R-INLA`. In the case of latent Gaussian models, INLA is typically as fast or faster than MCMC for model fitting, while also being essentially fully automated. It can difficult to fully automate MCMC do to the need to diagnose convergence. Fast and automated computation is important in the context of child mortality estimation, as it would be ideal for individuals in settings with minimal computational resources to be able to fit the models being used. The `R` package `SUMMER` [77] contains implementations of the smoothed direct model and has been used to produce estimates of U5MR supported by the UN IGME [76].

4.2.3 Child Mortality Estimation: Two Motivating Examples

In this section we provide two motivating child mortality applications where the assumptions underlying the RW1 and ICAR random effects used in the smoothed direct model are not realistic.

A National Model for Rwanda

Suppose we would like to estimate U5MR at the national level in Rwanda from 1985 through 2015, the last year with a DHS survey, and predict U5MR from 2016 through 2021. Rwanda experienced a civil war from 1990-1994, culminating in the Rwandan genocide in 1994 [31, 100]. The civil war, the genoicde, and their aftermath produced a shock in child mortality in Rwanda in the 90s.

In Figure 4.1, we plot the direct estimates of U5MR for Rwanda from the six DHS surveys between 1992 and 2015, for the 15 years prior to each survey. The shock to child mortality in the mid 90s is clear from this plot. Further, we plot the estimates of U5MR from UN IGME, which are based on the B3 model, and overlay a "meta-analysis" estimator of U5MR based on the direct estimates from the six DHS surveys, i.e., the precision-weighted average of each survey's estimate in each year. We note here that the B3 model uses an ad-hoc approach to produce estimates for what it deems to be conflict years, which are 1993-1999 for Rwanda [135]. Rather than fitting a model to all of the available data, it fits a model leaving out data from conflict years. From the model, it predicts U5MR for the conflict years, and then adds on a separate conflict-specific mortality rate to the estimate for conflict years. This conflict-specific mortality rate may be based on the data used to fit the rest of the model, or based on other outside sources, and does not include any uncertainty.

We would like to modify the national level smoothed direct model so that it does not oversmooth U5MR during the years in which we expect shocks in mortality, which we will refer to as conflict years. In more detail, when using a RW1 prior for the structured temporal effect in the smoothed direct model, we do not believe it is likely that transitions not involving conflict years will have the same variance as transitions involving conflict years. We expect transitions not involving conflict years to have a smaller variance than transitions involving conflict years. In other words, the transitions of the structured temporal effect should not be exchangeable as is assumed in the RW1.

A Multi-country Subnational Model

Suppose we would like to simultaneously estimate U5MR subnationally at the Admin1 level across multiple countries during the 2010-2014 time period. Specifically, we will consider Burundi, Ethiopia, Kenya, Rwanda, Tanzania, and Uganda using DHS surveys from 2016, 2016, 2014, 2015, 2015, and 2016 respectively. Due to the country boundaries, Admin1 regions within the same country are likely to have more similar outcomes than Admin1 regions not within the same country. In Figure 4.2, we plot the direct estimates of U5MR

Figure 4.1: Top Panel: Direct estimates of U5MR for Rwanda from six DHS surveys, for the 15 years prior to each survey. Bottom Panel: U5MR estimates from IGME and a meta-analysis estimator of U5MR based on the direct estimates in the top panel.

at the Admin1 level for Burundi, Ethiopia, Kenya, Rwanda, Tanzania, and Uganda. We see that along some borders, the Admin1 U5MR direct estimates are fairly similar (e.g. Uganda and Tanzania), and along other borders there's a large difference (e.g. Kenya and Ethiopia).

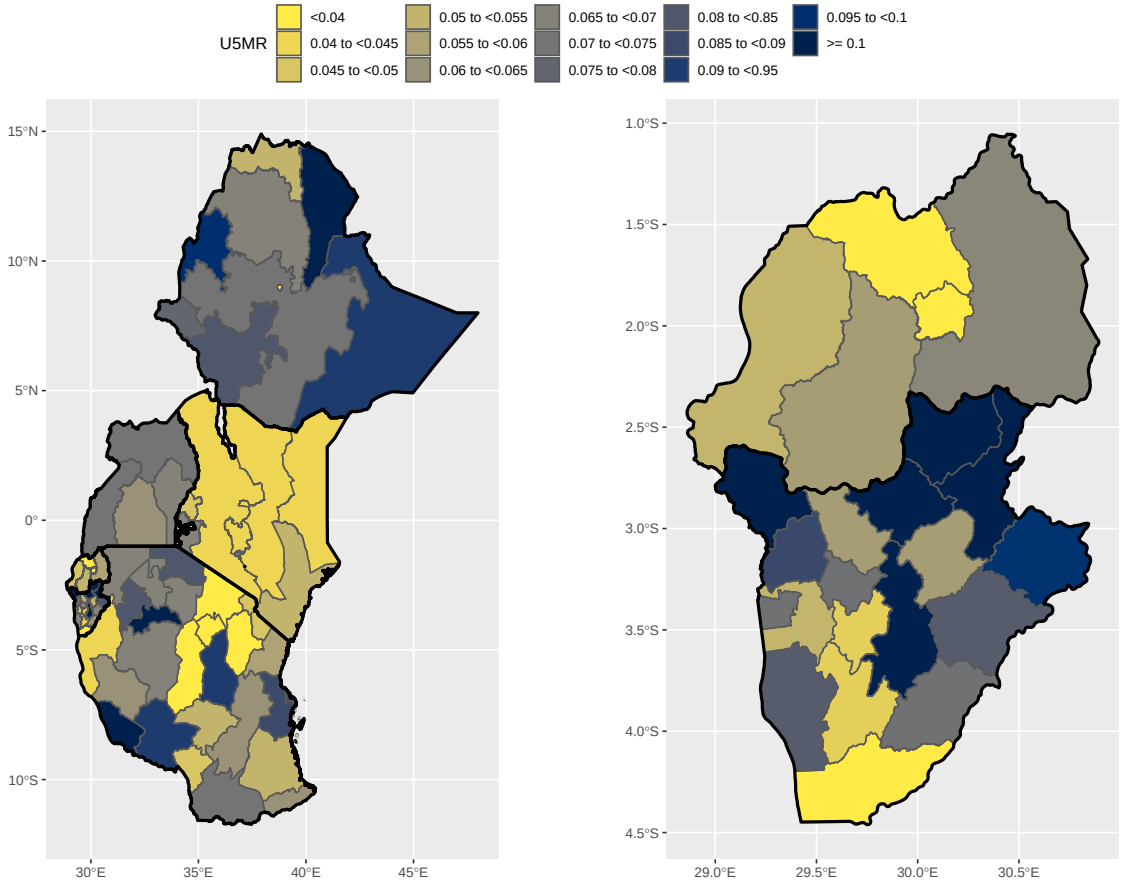


Figure 4.2: Left Panel: Direct estimates of U5MR for Burundi, Ethiopia, Kenya, Rwanda, Tanzania, and Uganda. Right Panel: Zoomed in direct estimates of U5MR for Burundi and Rwanda.

We would like to modify the subnational level smoothed direct model so that it does not oversmooth U5MR over country boundaries. In more detail, when using an ICAR prior for the structured spatial effect in the smoothed direct model, we do not believe it is likely that differences involving neighboring Admin1 regions within the same country will have the same

variance as differences involving neighboring Admin1 regions not within the same country. We expect differences involving neighboring Admin1 regions within the same country to have a smaller variance than differences involving neighboring Admin1 regions not within the same country. In other words, the differences of the structured temporal effect should not be exchangeable as is assumed in the ICAR.

4.3 Adaptive Gaussian Markov Random Fields

Motivated by the child mortality examples described in the previous section, in this section we will develop what we call adaptive Gaussian Markov random fields (AGMRFs). We will first develop general first order adaptive Gaussian Markov random fields, before focusing on two specific use cases for the Rwandan and Kenyan child mortality applications. We conclude by considering possible extensions to higher order AGMRFs and connections to previous work.

4.3.1 General First Order Adaptive Gaussian Markov Random Fields

Adaptive Random Walk of First Order

Suppose we have N time periods and we are specifying a structured temporal random effect $\boldsymbol{x} = (x_1, \ldots, x_N)$. Suppose for $i = 1, \ldots, N-1$ we specify

$$x_{i+1} \mid x_i, \tau_i \sim \mathsf{Normal}(x_i, \tau_i^{-1}) \tag{4.2}$$

where τ_i is the precision when moving from time i to time $i+1$. We will refer to this as an adaptive random walk of first order or an adaptive random walk 1 (ARW1). This leads to a density for $\boldsymbol{x}$ of

$$p(\boldsymbol{x} \mid \tau_1, \ldots, \tau_{N-1}) \propto \left[\prod_{i=1}^{N-1} \tau_i^{1/2}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{N-1} \tau_i (x_{i+1} - x_i)^2\right\} = \left[\prod_{i=1}^{N-1} \tau_i^{1/2}\right] \exp\left\{-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x}\right\},$$

where Q is a precision matrix with the same sparsity structure as the precision matrix of a RW1, such that

$$Q_{ij} = \begin{cases} \tau_1, & \text{if } i = j = 1 \\ \tau_{N-1}, & \text{if } i = j = N \\ \tau_{i-1} + \tau_i, & \text{if } i = j = k \text{ and } k \notin \{1, N\} \\ -\tau_i, & \text{if } j = i + 1 \text{ or } i = j + 1 \\ 0, & \text{if } |i - j| > 1. \end{cases}$$

It can be verified that $Q\mathbf{1} = \mathbf{0}$, and so Q is rank $N - 1$ and thus ARW1s are improper GMRFs. It follows that $p(\boldsymbol{x} \mid \tau)$ is invariant to an addition of a constant vector to $\boldsymbol{x}$, thus when an ARW1 is included in a model with an intercept, we enforce the sum-to-zero constraint $\sum_{i=1}^{N} x_i = 0$.

Adaptive Intrinsic Conditional Autoregression

Suppose we have N areal units with no islands and we are specifying a spatial random effect $\boldsymbol{x} = (x_1, \ldots, x_N)$. Suppose for neighboring regions $i \sim j$ we specify

$$x_i - x_j \mid \tau_{ij} \sim \mathsf{Normal}(0, \tau_{ij}^{-1}) \tag{4.3}$$

where τ_{ij} is the precision for the difference between units i and j. We will refer to this as an adaptive intrinsic conditional autoregression (AICAR). This leads to a density for $\boldsymbol{x}$ of

$$p(\boldsymbol{x} \mid \{\tau_{ij}\}_{i\sim j}) \propto (|Q|^*)^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x}\right\}$$

where $|\cdot|^*$ denotes the generalized determinant of a matrix, defined as the product of its non-zero eigenvalues, and Q is a precision matrix with the same sparsity structure as the precision matrix of an ICAR, such that

$$Q_{ij} = \begin{cases} \sum_{k|i\sim k} \tau_{ik}, & \text{if } i = j \\ -\tau_{ij}, & \text{if } i \sim j \\ 0, & \text{otherwise.} \end{cases}$$

It can be verified that $Q\mathbf{1} = \mathbf{0}$, and so Q is rank $N-1$ and thus AICARs are improper GMRFs. It follows that $p(\boldsymbol{x} \mid \tau)$ is invariant to an addition of a constant vector to $\boldsymbol{x}$, thus when an AICAR is included in a model with an intercept, we enforce the sum-to-zero constraint $\sum_{i=1}^{N} x_i = 0$.

4.3.2 Two Specific First Order Adaptive Gaussian Markov Random Fields

The general AGMRFs introduced in the previous section have a large number of precision parameters, which makes them very flexible. However, this flexibility can difficulties for prior specification and computation. It is difficult in practice to specify substantive priors for a large number of hyperparameters. Similarly, one needs a small number of hyperparameters if INLA is to be used for computation. In this section, we will specialize these general AGMRFs to our motivating child mortality applications from Section 4.2.3, drastically reducing the number of precision parameters. This specialization sacrifices flexibility in our model so that it is more amenable to substantive prior specification and fitting in INLA.

A Conflict Adaptive RW1

Suppose we are specifying a structured temporal random effect for N years, where for some subset of the N years are conflict years. In the case of Rwanda, the UN IGME categorizes 1993-1999 as conflict years. Let C denote the subset of conflict years. Suppose for $i = 1, \ldots, N-1$ we specify

$$\begin{aligned} x_{i+1} \mid x_i, \tau_1, \tau_2 &\sim \mathsf{Normal}(x_i, \tau_1^{-1}) && \text{if } \{i, i+1\} \notin C, \\ x_{i+1} \mid x_i, \tau_1, \tau_2 &\sim \mathsf{Normal}(x_i, \tau_2^{-1}) && \text{if } i \in C \text{ or } i+1 \in C. \end{aligned}$$

This is a simplified ARW1 with only two precisions: one precision for transitions not involving conflict years, τ_1, and one precision for transitions involving conflict years, τ_2. As discussed in Section 4.2.3, we expect that $\tau_1 > \tau_2$. It follows that $\boldsymbol{x} \mid \tau_1, \tau_2 \sim \mathsf{Normal}(0, Q^{-})$, where Q is a special case of the general ARW1 precision matrix. We will refer to this as a conflict adaptive random walk 1.

For this conflict ARW1 we can simplify Q to $Q = \sum_{l=1}^{2} \tau_l R_l$, where for $l \in \{1, 2\}$, $R_l = D_1 - W_l$ where

$$W_{l,ij} = \begin{cases} I(|i-j| = 1 \text{ and } \{i,j\} \notin C), & \text{if } l = 1 \\ I(|i-j| = 1 \text{ and } i \in C \text{ or } j \in C), & \text{if } l = 2, \end{cases}$$

and $D_l = \text{diag}\left(\sum_{j=1}^{N} W_{l,1j}, \ldots, \sum_{j=1}^{N} W_{l,Nj}\right)$. Note that when $\tau_1 = \tau_2$ this random effect specification reduces to a RW1 as presented in Section 4.2.1. In particular we have in this case that $Q = \tau_1[R_1 + R_2]$ where $R_1 + R_2$ is the structure matrix of a random walk 1.

A Multi-Country Adaptive ICAR

Suppose we are specifying a structured spatial random effect for N regions at the Admin1 level with no islands, which are nested within M countries. For neighboring Admin1 regions $i \sim j$, let A_{ij} be an indicator that i and j are nested within the same country. Suppose for neighboring Admin1 regions $i \sim j$ we specify

$$\begin{aligned} x_i - x_j \mid \tau_1, \tau_2 &\sim \mathsf{Normal}(0, \tau_1^{-1}) && \text{if } A_{ij} = 1, \\ x_i - x_j \mid \tau_1, \tau_2 &\sim \mathsf{Normal}(0, \tau_2^{-1}) && \text{if } A_{ij} = 0. \end{aligned}$$

This is a simplified AICAR with only two precisions: one precision for neighboring Admin1 regions within the same country, τ_1, and one precision for neighboring Admin1 regions between different countries, τ_2. As discussed in Section 4.2.3, we expect that $\tau_1 > \tau_2$. It follows that $\boldsymbol{x} \mid \tau_1, \tau_2 \sim \mathsf{Normal}(0, Q^-)$, where Q is a special case of the general AICAR precision matrix. We will refer to this as a multi-country adaptive intrinsic conditional autoregression.

For this multi-country AICAR we can simplify Q to $Q = \sum_{l=1}^{2} \tau_l R_l$, where for $l \in \{1, 2\}$, $R_l = D_1 - W_l$ where

$$W_{l,ij} = \begin{cases} I(i \sim j \text{ and } A_{ij} = 1), & \text{if } l = 1 \\ I(i \sim j \text{ and } A_{ij} = 0), & \text{if } l = 2, \end{cases}$$

and $D_l = \text{diag}\left(\sum_{j=1}^{N} W_{l,1j}, \ldots, \sum_{j=1}^{N} W_{l,Nj}\right)$. Note that when $\tau_1 = \tau_2$ this random effect specification reduces to an ICAR as presented in Section 4.2.1. In particular we have in this case that $Q = \tau_1[R_1 + R_2]$ where $R_1 + R_2$ is the structure matrix of an ICAR.

4.3.3 Higher Order Adaptive Gaussian Markov Random Fields

Extensions to higher order AGMRFs are straightforward mathematically, but not so much conceptually, as we will now briefly illustrate with an adaptive random walk of second order. Suppose we have N time periods and we are specifying a structured temporal random effect $\boldsymbol{x} = (x_1, \ldots, x_N)$. Suppose for $i = 1, \ldots, N-2$ we specify

$$x_{i+2} - x_{i+1} \mid x_{i+1}, x_i, \tau_i \sim \mathsf{Normal}(x_{i+1} - x_i, \tau_i^{-1})$$

where τ_i is the precision for the ith second order difference. This leads to an adaptive version of the second-order random walk [see e.g. page 110 of 111].

As with the ARW1, as is this model has a large number of precision parameters, which makes it too flexible and computationally difficult to fit. However, because the model is defined using second order differences, rather than first order differences, it is not clear how one should go about specializing the model as in Section 4.3.2. In particular, suppose we were trying to specialize this adaptive second-order random walk to the Rwanda application as in Section 4.3.2. In Section 4.3.2, we were able to use the interpretation of first order differences to reduce the parameter space down to two precisions for differences involving conflicts and not involving conflicts. It is not clear to the authors how to use the interpretation of second order differences to reduce the parameter space. This difficulty compounds when moving to even higher order GMRFs.

4.3.4 Connections to Previous Work

Variants of the general first order AGMRFs introduced in Section 4.3.1 have been used before, typically for the purpose of constructing flexible and locally adaptive curve fitting methods in applications where shocks in the curves being fit are *unknown* [73, 142, 111, 39, 40, 23, 108].

We note that this is closely related to change point detection in the time series literature [6] and wombling in the spatial statistics literature [11, 83, 84, 25, 56], where the goal is to *identify* unknown shocks or regions of rapid change in curves. In the previous approaches, the models in Equations 4.2 and 4.3 are used directly, with priors placed directly on all of the precision parameters, with the exception of [23] and [108]. The works of [23] and [108], restrict the AICAR in Equation 4.3 by reparameterizing the precision parameters as $\tau_{ij} = \tau_i \tau_j$, where τ_i, τ_j are region-level precision parameters. While this parameterization reduces the number of precision parameters in the AICAR down to N, this is still a large number which requires the precision parameters to have smoothing priors of their own.

[73, 142] induce dependence between these precision parameters by letting them follow GMRFs on the log scale. [111], [39], and [40] place priors on the precision parameters with the intention of marginalizing them out to produce non-normal differences. In particular, [39] and [40] focus on the case where differences marginally have horseshoe priors [27]. Due to the large number of precision parameters, these models are in most cases not amenable to fitting with INLA. Further, using MCMC to fit the models in [39] and [40] inherits the difficulties of using MCMC to sample from models with horseshoe priors [104], although this can be avoided by using priors with lighter tails.

In contrast, in this work, we restrict the general first order AGMRFs as we have a priori knowledge of the location of the shocks in the curves we are fitting. This lets us work with a small number of precision parameters, allowing us to specify more substantive priors and fit the model in INLA, as we discuss in the following section. However, the prior information available in some applications may not always be enough to restrict an AGMRF to a small number of precision parameters. In such cases, MCMC or alternative deterministic Bayesian approximations, such as Template Model Builder [71], would be necessarry for model fitting.

Another approach to developing adaptive random effects models would be to consider adaptive generalizations of other spatial or temporal random effects. For areal spatial data, another common spatial random effect is the simultaneous autoregressive (SAR) model [see e.g. Chpater 4 of 11]. Compared to the adaptive generalizations of RWs and ICARs which

have been proposed in the literature, there has been minimal work proposing adaptive generalizations of the SAR model [96].

In the Rwanda application, we have changed the smoothed direct model to account for events that have caused a shock in child mortality. This is related to the problem of incorporating information on feed-back interventions into forecasts in the dynamic linear models literature [138]. The dynamic linear models literature is motivated by providing more accurate forecasts in the future, whereas we are motivated by providing more accurate retrospective estimates of U5MR in the Rwanda application.

In the multi-country application, we estimate U5MR at the subnational level, while leveraging the fact that subnational regions are nested within countries. However, it is sometimes desirable to simultaneously provide smoothed estimates of U5MR at the subnational and national level that are coherent to some extent. Methods from multiresolution modeling could be used for this purpose [43]; in particular multiscale random field models [44] would naturally fit into the smoothed direct modelling framework of [94].

4.4 Scaling, Reparameterizations, Prior Choice, and Computation

In this section, we describe various considerations for the use of our adaptive GMRFS in practice, including: scaling as in [122], reparameterizations for interpretability, the choice of priors for hyperparameters, and computation.

4.4.1 Scaling

Let $Q = \sum_{l=1}^{2} \tau_l R_l$ denote the precision matrix of one of the AGMRFs described in Section 4.3.2. As these AGMRFs are improper, we need to worry about the interpretation of the precision parameters, which are dependent on the structure matrix $R_1 + R_2$. In particular, we will scale the precisions as in [122]. Let $\sigma^2(\boldsymbol{x})$ denote the geometric mean of the marginal variances of the elements of $\boldsymbol{x}$ when setting $\tau_1 = \tau_2 = 1$ (i.e. so that $Q = R_1 + R_2$ is the structure matrix of a RW1 or ICAR). We will work with the following scaled precision matrix $Q^* = \sum_{l=1}^{2} \tau_l R_l^*$, where for $l \in \{1, 2\}$, $R_l^* = R_l / \sigma^2(\boldsymbol{x})$. It then follows that when $\tau_1 = \tau_2$,

$1/\tau_1$ represents the approximate marginal variance of $\boldsymbol{x}$, independent of the structure matrix $R_1 + R_2$.

4.4.2 Reparameterizing for Interpretability

Reparameterization of AGMRFs

Let $Q^* = \tau_1 R_1^* + \tau_2 R_2^*$ denote the scaled precision matrix of one of the AGMRFs described in Section 4.3.2. For both of the AGMRFs described in Section 4.3.2, there is one precision, τ_1, which we expect to be larger than the other, τ_2: we expect the non-conflict precision to be larger than the conflict precision, and we expect the within country precision to be larger than the between country precision. We will reparameterize τ_2 such that $\tau_2 = \tau_1\theta$, where $\theta \in (0, 1]$, so that $Q^* = \tau_1 R_1^* + \tau_1\theta R_2^* = \tau_1[R_1^* + \theta R_2^*]$.

A BYM2-Like Parameterization

In the smoothed direct model introduced in Section 4.2.2, when the structured subdivision-level random effect was a RW1 or an ICAR, we reparameterized the total subdivision-level random effect, $\boldsymbol{b}$, as a BYM2 following [110]. Suppose instead our total subdivision-level random effect is given by $\boldsymbol{b} = \boldsymbol{v} + \boldsymbol{x}$, where $\boldsymbol{v}$ is an unstructured random effect, and $\boldsymbol{x}$ is one of the AGMRFs described in Section 4.3.2, scaled as described in Section 4.4.1. We will now develop a BYM2-like parameterization of this total subdivision-level random effect. Let

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left[\sqrt{1-\phi}\boldsymbol{v} + \sqrt{\phi}\boldsymbol{x}^*\right]$$

where $v_i \overset{iid}{\sim} \mathsf{Normal}(0, 1)$, $\boldsymbol{x}^* \sim \mathsf{Normal}(0, [R_1^* + \theta R_2^*]^-)$, and $\phi \in [0, 1]$. Then $1/\tau_b$ represents the approximate marginal variance of the total subdivision-level random effect $\boldsymbol{b}$, and ϕ represents the proportion of this approximate variance attributed to the structured component when $\theta = 1$.

For computational purposes, we will then reparameterize the $\boldsymbol{b}$ to preserve sparsity as in [110]. Let $\boldsymbol{w} = (\boldsymbol{w}_1^T \ \boldsymbol{w}_2^T)^T$, where $\boldsymbol{w}_1 = \boldsymbol{b}$ and $\boldsymbol{w}_2 = \boldsymbol{x}^*$. Then it follows that $\boldsymbol{w} \sim$

$\mathsf{Normal}(0, S^-)$ where

$$S = \begin{pmatrix} \frac{\tau_1}{1-\phi} I & -\frac{\sqrt{\phi\tau_1}}{1-\phi} I \\ -\frac{\sqrt{\phi\tau_1}}{1-\phi} I & R_1^* + \theta R_2^* + \frac{\phi}{1-\phi} I \end{pmatrix}.$$

4.4.3 Prior Choice

When using the AGMRFs described in Section 4.3.2 in a BYM2 like parameterization as outlined in Section 4.4.2, we must specify priors for the hyperparameters τ_b, θ, and ϕ. We will use the penalized complexity (PC) prior framework of [120] to specify these priors.

PC Prior for τ_b

We will first consider the PC prior for τ_b, where we are shrinking to $\tau_b = \infty$, i.e. no subdivision-level random effect. Conditional on θ and ϕ, the PC prior for τ_b is the PC prior for a normal precision as derived in [120]

$$p(\tau_b \mid \theta) = \frac{\lambda}{2} \tau_b^{-3/2} \exp(-\lambda \tau_b^{-1/2}).$$

In this paper, we will specify the PC prior for τ_b such that $P(1/\sqrt{\tau_b} > 1) = 0.01$.

PC Prior for ϕ

We will now consider the PC prior for ϕ, where we are shrinking to $\phi = 0$, i.e. no structured subdivision-level random effect. For ease of implementation, we will specify the PC prior for ϕ conditional on $\theta = 1$. Conditional on $\theta = 1$, the PC prior for ϕ is derived in Appendix C of [110], although it is not difficult to re-derive the PC prior for ϕ under a different value of θ if desired. In this paper, we will specify the PC prior for ϕ such that $P(\phi < 0.5) = 2/3$.

PC Prior for θ

We will now consider the PC prior for θ, where we are shrinking to $\theta = 1$, i.e. $\tau_1 = \tau_2$ in the original parameterization of the AGMRFs described in Section 4.3.2. We will now derive an analytical formula for the PC prior for θ that shrinks to $\theta = 1$. This involves: 1) deriving

the KL divergence between our "flexible" model, where $\theta \in (0, 1]$, and our "base" model, where $\theta = 1$, 2) placing an exponential prior on a transformation of the KL divergence, and 3) transforming the prior on the KL divergence to a prior on θ.

We first need to derive the KL divergence between our flexible model and this base model. As both the base and flexible models are intrinsic, we will instead calculate the KL divergence in the $n-1$ subspace where we have taken away the singular portion of the Gaussians. In particular, let $\hat{R}_1^*$ and $\hat{R}_2^*$ denote R_1^* and R_2^* after removing the first row and column, although we could remove in general the rth row and column and the result would stay the same. Then we can now derive the KL divergence between our flexible model and base model:

$$\begin{aligned} d(\theta) &= \sqrt{2KL(\mathsf{Normal}(0, (\tau_1[\hat{R}_1^* + \theta\hat{R}_2^*])^{-1}) || \mathsf{Normal}(0, (\tau_1[\hat{R}_1^* + \hat{R}_2^*])^{-1}))} \\ &= \sqrt{\text{trace}\left([\hat{R}_1^* + \hat{R}_2^*][\hat{R}_1^* + \theta\hat{R}_2^*]^{-1}\right) - (n-1) - \log\left(\frac{|\hat{R}_1^* + \hat{R}_2^*|}{|\hat{R}_1^* + \theta\hat{R}_2^*|}\right)}. \end{aligned}$$

Note that we can write $\hat{R}_1^* + \theta\hat{R}_2^* = \hat{R}_1^* + \hat{R}_2^* + \hat{R}_2^*(\theta - 1) = [\hat{R}_1^* + \hat{R}_2^*][I + (\hat{R}_1^* + \hat{R}_2^*)^{-1}\hat{R}_2^*(\theta - 1)]$. Let $\varepsilon_1, \ldots, \varepsilon_{n-1}$ denote the $n-1$ eigenvalues of $(\hat{R}_1^* + \hat{R}_2^*)^{-1}\hat{R}_2^*$. It follows that we can simplify the calculation of $d(\theta)$ as follows:

$$\begin{aligned} d(\theta) &= \sqrt{trace([\hat{R}_1^* + \hat{R}_2^*][\hat{R}_1^* + \theta\hat{R}_2^*]^{-1}) - (n-1) - \log\left(\frac{|\hat{R}_1^* + \hat{R}_2^*|}{|\hat{R}_1^* + \theta\hat{R}_2^*|}\right)} \\ &= \sqrt{trace([I + (\hat{R}_1^* + \hat{R}_2^*)^{-1}\hat{R}_2^*(\theta - 1)]^{-1}) - (n-1) + \log\left(|I + (\hat{R}_1^* + \hat{R}_2^*)^{-1}\hat{R}_2^*(\theta - 1)|\right)} \\ &= \sqrt{\sum_{i=1}^{n-1} \frac{1}{1 + (\theta - 1)\varepsilon_i} - (n-1) + \log\left(\prod_{i=1}^{n-1} 1 + (\theta - 1)\varepsilon_i\right)} \\ &= \sqrt{\sum_{i=1}^{n-1} \frac{1}{1 + (\theta - 1)\varepsilon_i} - (n-1) + \sum_{i=1}^{n-1} \log\left(1 + (\theta - 1)\varepsilon_i\right)}. \end{aligned}$$

Placing an exponential prior on $d(\theta)$ and transforming to a prior on θ, we find that the PC prior for θ is given by

$$p(\theta) = \lambda \exp\{-\lambda d(\theta)\} \left|\frac{\partial d(\theta)}{\partial \theta}\right|,$$

where we can calculate the Jacobian as

$$\begin{aligned}\frac{\partial d(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\sqrt{d^2(\theta)} \\ &= \frac{1}{2}\left[d(\theta)\right]^{-1}\frac{\partial d^2(\theta)}{\partial \theta} \\ &= \frac{1}{2}\left[d(\theta)\right]^{-1}\sum_{i=1}^{n-1}\frac{(\theta-1)\varepsilon_i^2}{[1+(\theta-1)\varepsilon_i]^2}.\end{aligned}$$

Thus the PC prior for θ can be written as

$$p(\theta) = \frac{\lambda(1-\theta)}{2d(\theta)}\left[\sum_{i=1}^{n-1}\frac{\varepsilon_i^2}{[1+(\theta-1)\varepsilon_i]^2}\right]\exp\{-\lambda d(\theta)\}.$$

The user can specify the parameter λ using a prior probability statement of the form $P(\theta < U) = \alpha$. As $P(\theta < U) = P(d(\theta) > d(U)) = \exp\{-\lambda d(U)\} = \alpha$, since $d(\theta)$ is a decreasing function of θ on $(0, 1]$, this corresponds to using $\lambda = -\log(\alpha)/d(U)$. In this paper, we will specify the PC prior for θ such that $P(\theta < 0.75) = 0.75$.

4.4.4 Computation

As the AGMRFs described in Section 4.3.2 are Gaussian conditional on a small number of hyperparameters, they can be used as random effects in models fit using INLA. In particular, we implemented the AGMRFs through the `rgeneric` functionality in the `R-INLA` package. Implementing our AGMRFs using INLA allows them to be readily be incorporated into the `R` package `SUMMER` [77].

4.5 Simulations

Before applying our AGMRFs to the motivating applications, we will perform a simulation study to assess whether the smoothed direct model using our AGMRFs can improve upon the performance of the smoothed direct model described in Section 4.2.2. This simulation study is designed to mimic the structure of the Rwanda data which we analyze in Section 4.6.

We will simulate data from the model

$$y_i \mid \eta_i \sim \mathsf{Normal}(\eta_i, \hat{V})$$

where $\eta_i = \mu_i + b_i$ and $b_i \sim \mathsf{Normal}(0, \tau_i^{-1})$. There will be $N = 30$ time points, with time points $9 - 15$ designated as conflict time points, imitating the structure of the Rwanda data. We will vary μ_i, τ_i, and $\hat{V}$:

- $\hat{V}$ will take values in $\{1/75, 1/150, 1/300\}$,
- τ_i will either be 20 for all time points, or 20 for non-conflict time points and 10 for conflict time points,
- μ_i will take on one of three trends, as seen in Figure 4.3. We refer to the trends as constant, level change, and triangle, based on their shapes.

We will simulate 100 data from this model for each parameter setting and fit two models to the simulated data: the smoothed direct model with a RW1 for the structured temporal random effect, which we will refer to as the smoothed direct model, and the smoothed direct model with our proposed conflict ARW1 for the structured temporal random effect, which we will refer to as the proposed model. Both models are mispecified except in the case where μ_i is constant in time and the random effect precision is constant in time. We will evaluate the two model fits with RMSE $\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \hat{\eta}_i)^2}\right)$, DIC [124], and average proper logarithmic scoring rule (LS) [52]. For each metric, a lower value represents better model performance.

In Figures 4.4 and 4.5 we plot RMSE for the simulation settings where τ_i is the same for all time points and where τ_i is not the same for all time points, respectively. The models have nearly identical performance under RMSE. In Figures 4.6 and 4.7 we plot DIC for the simulation settings where τ_i is the same for all time points and where τ_i is not the same for all time points, respectively. The proposed model outperforms the smoothed direct model under DIC when $\hat{V}$ is larger, and the trend is non-constant. In Figures 4.8 and 4.9 we plot LS for the simulation settings where τ_i is the same for all time points and where τ_i is not the

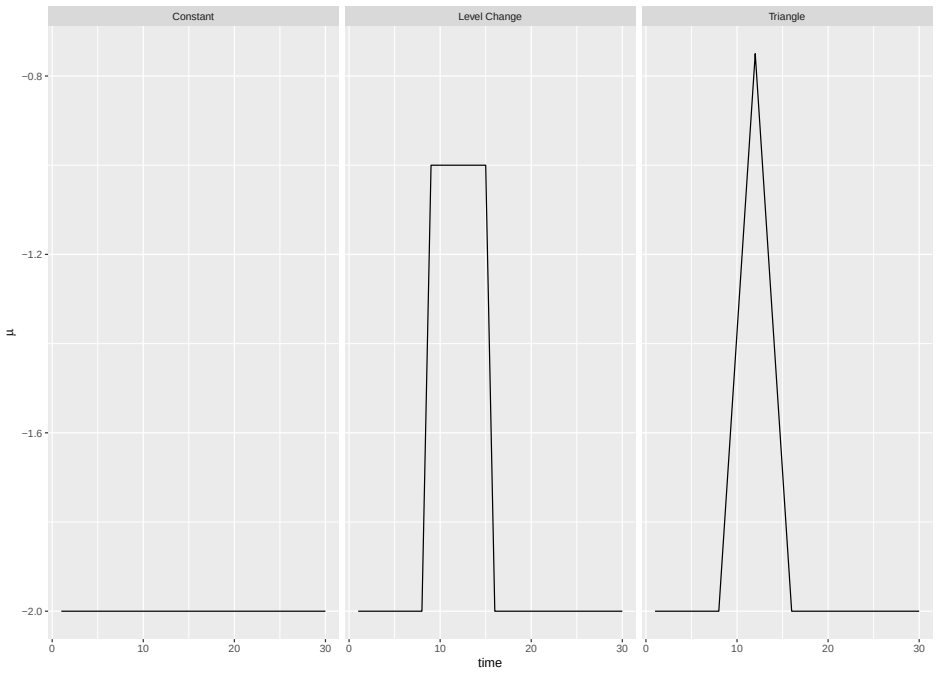


Figure 4.3: The three trends used for μ_i to simulate data.

same for all time points, respectively. The proposed model outperforms the smoothed direct model under LS when the trend is non-constant.

Across the different metrics, we see that the potential for the proposed model to provide improvements over the smoothed direct model lies in the underlying curve having abrupt shocks, as in the case of the level change and triangle trends, as desired.

4.6 Estimation of U5MR at the National Level in Rwanda

In this section we will estimate U5MR at the national level in Rwanda from 1985 through 2015, the last year with a DHS survey, and predict U5MR from 2016 through 2021. We will use two models two estimate U5MR, as in the simulations: the smoothed direct model with a RW1 for the structured temporal random effect, which we will refer to as the smoothed direct model, and the smoothed direct model with our proposed conflict ARW1 for the structured temporal random effect, which we will refer to as the proposed model. For the

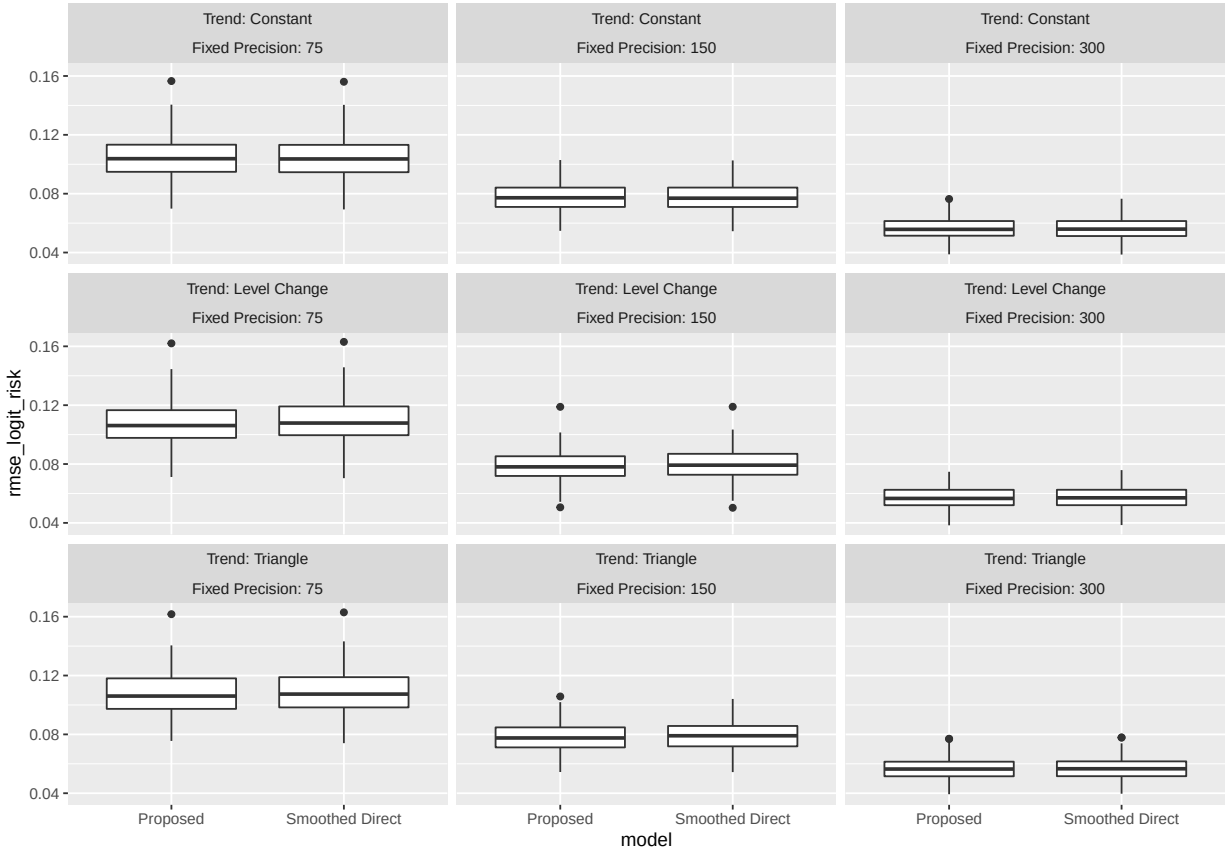


Figure 4.4: RMSE for simulation settings where τ_i is the same for all time points.

conflict ARW1, we will categorize 1993-1999 as conflict years following UN IGME. We will limit the direct estimates for each survey to those from the 15 years prior to when the survey was conducted. We note here that we explored fitting variants of both the smoothed direct model and the proposed model with separate intercepts for conflict and non-conflict periods, and the results did not substantively change.

In Table 4.1 we display summaries of the posterior for the various parameters in each model. We see that the posterior summaries for μ and ϕ are comparable between models. The posterior summaries for ν_s, the survey random effects, are also comparable between models, and are all centered around 0 except for the effect for the 2008 survey. Looking at

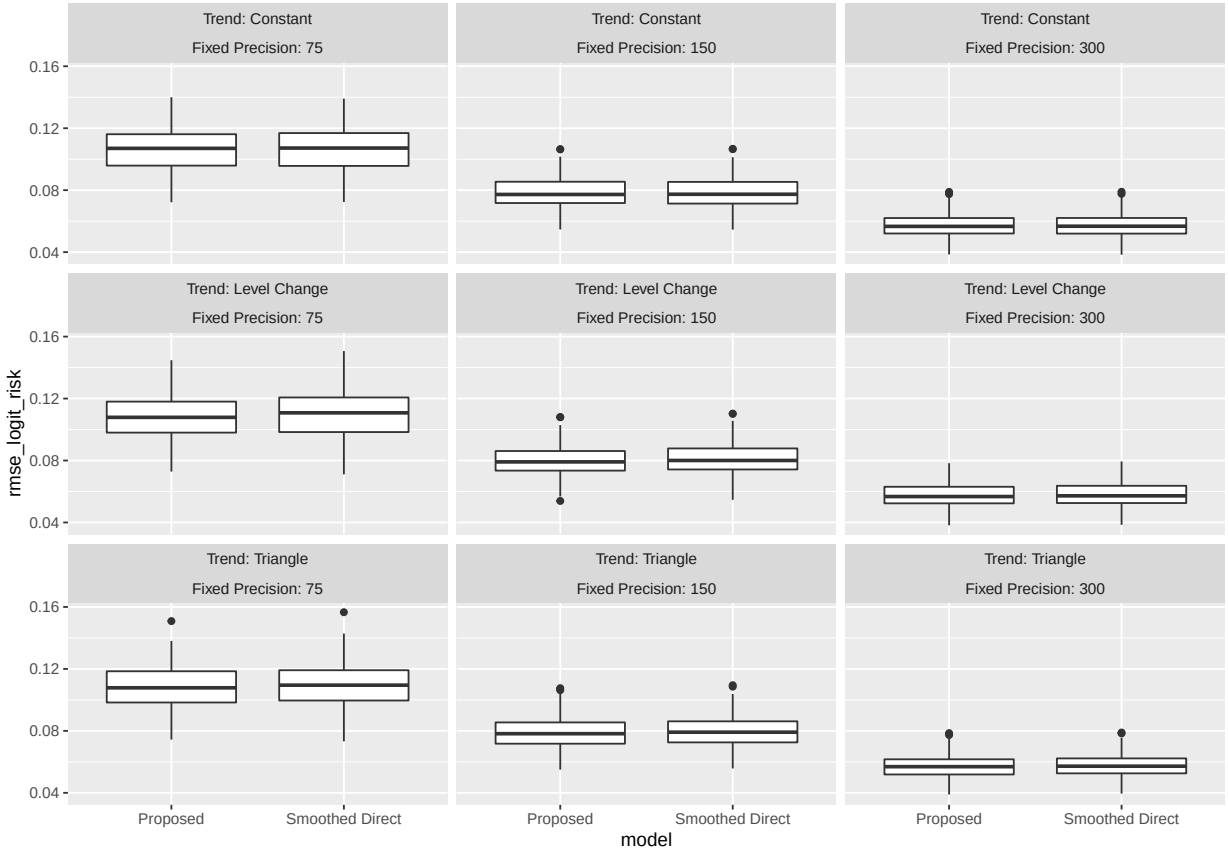


Figure 4.5: RMSE for simulation settings where τ_i is not the same for all time points.

figure 4.1, we see that the direct estimates in the 90s from the 2008 survey are a bit lower than those from the 2000 and 2005 surveys, which explains the posterior for ν_{2008} being negative across the two models.

The posterior summaries for τ are all larger under the proposed model. As ϕ/τ represents the approximate marginal variance of the total temporal random effect attributed to the structured component when $\theta = 1$, a larger τ means that the marginal variance of the structured component for non-conflict time periods in the proposed model is smaller than the marginal variance of the structured component in the smoothed direct model.

Focusing on θ, in Figure 4.10 we plot the prior and posterior for θ from the proposed

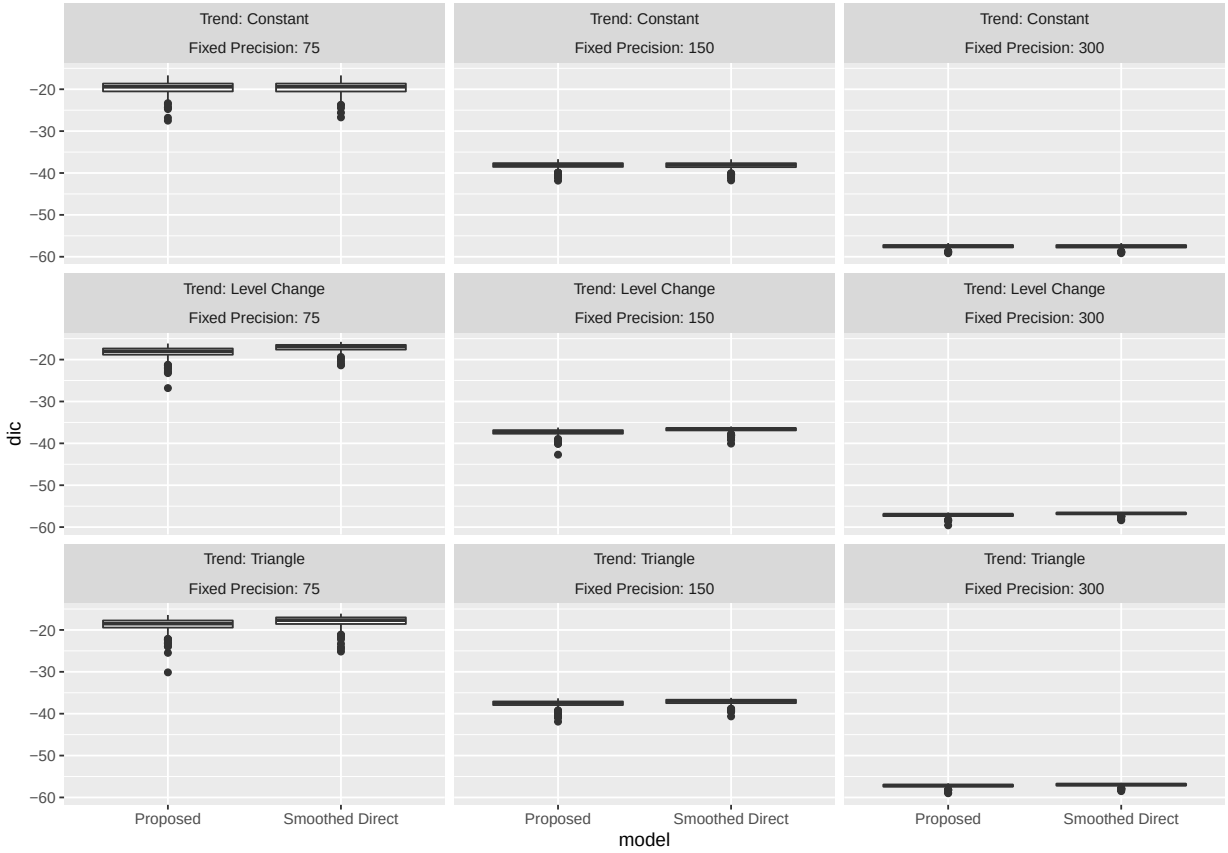


Figure 4.6: DIC for simulation settings where τ_i is the same for all time points.

model. We see that the prior is relatively flat from 0.2 to 1, with a mode at 0.35, and 95% of the prior mass lying in $[0.09, 0.97]$. The posterior has a mode at 0.32, close to the mode of the prior, and the 95% credible interval for θ is $[0.13, 0.82]$. The posterior places very little mass close to 1, but is not heavily concentrated around the mode. We can interpret this behavior as follows: the data has informed the posterior that θ is less than 1, i.e. that $\tau_1 > \tau_2$, but the data is not informative enough to hone in on a specific value of θ.

In Figure 4.11 we plot U5MR estimates from the smoothed direct model and the proposed model, in addition to the UN IGME and meta-analysis estimates discussed in Section 4.2.3 for reference. The estimates from the smoothed direct and proposed models differ the most

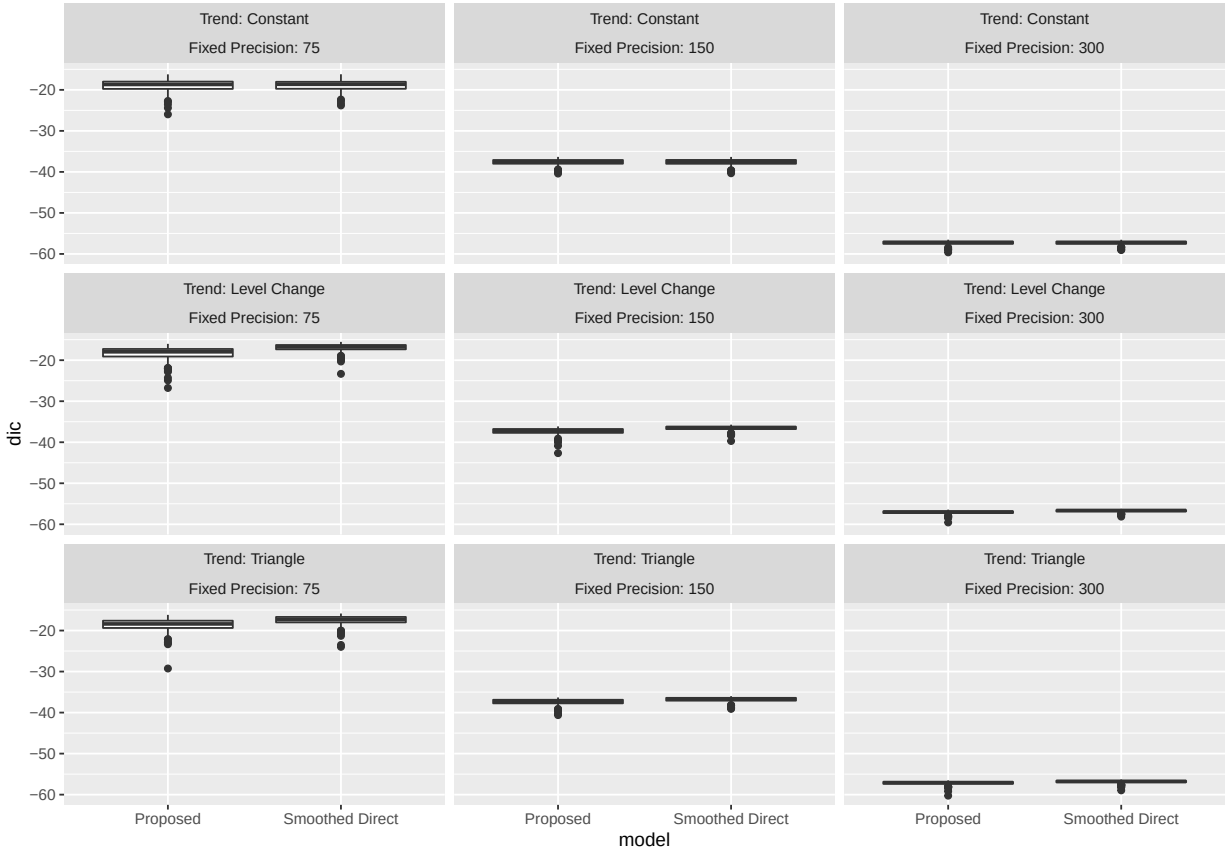


Figure 4.7: DIC for simulation settings where τ_i is not the same for all time points.

at two time periods: 1994 and 2016-2019. 1994 was the peak of the Rwandan civil war, when the Rwandan genocide occurred, which lead to a shock child mortality. The proposed model is able to avoid oversmoothing this time period, compared to the smoothed direct model, leading a higher estimate of U5MR that is closer to the meta analysis estimate.

The last year in which a survey was conducted was 2015, so we are predicting U5MR from 2016-2019. We see that the proposed model has much narrower prediction intervals over this time period than the smoothed direct model. The smoothed direct model uses a RW1 for the structured temporal random effect, which has a single variance for all temporal transitions. Thus, if transitions not involving conflicts truly have a smaller variance than transitions

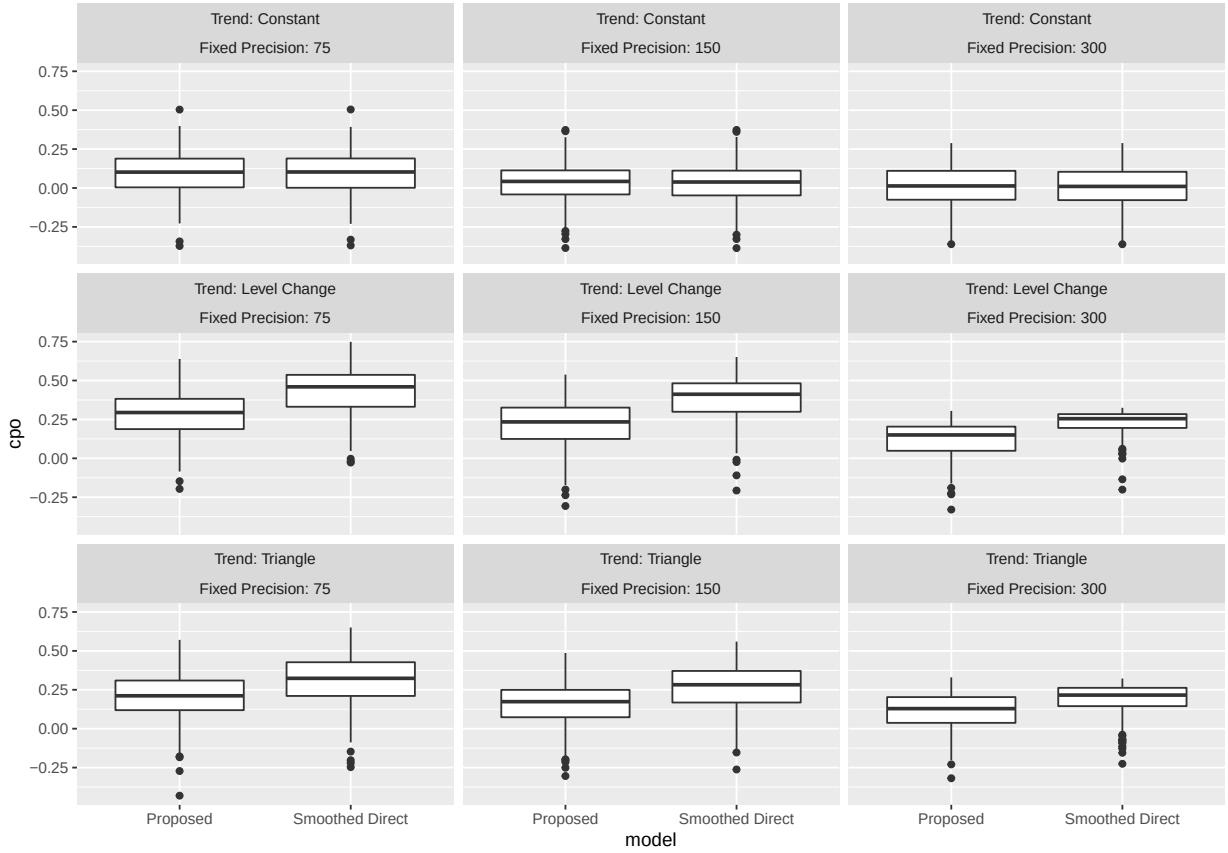


Figure 4.8: LS for simulation settings where τ_i is the same for all time points.

involving conflicts, the single variance parameter in the smoothed direct model will have to deal with this behavior by assuming a value somewhere in between the variance of the two types of transitions. This means that transitions not involving conflicts, which includes 2016-2019, will have a larger variance than under the smoothed direct model than under the proposed model, leading to narrower prediction intervals under the proposed model.

4.7 Estimation of U5MR across Multiple Countries

In this section we will simultaneously estimate U5MR subnationally at the Admin1 level across multiple countries during the 2010-2014 time period. Specifically, we will consider

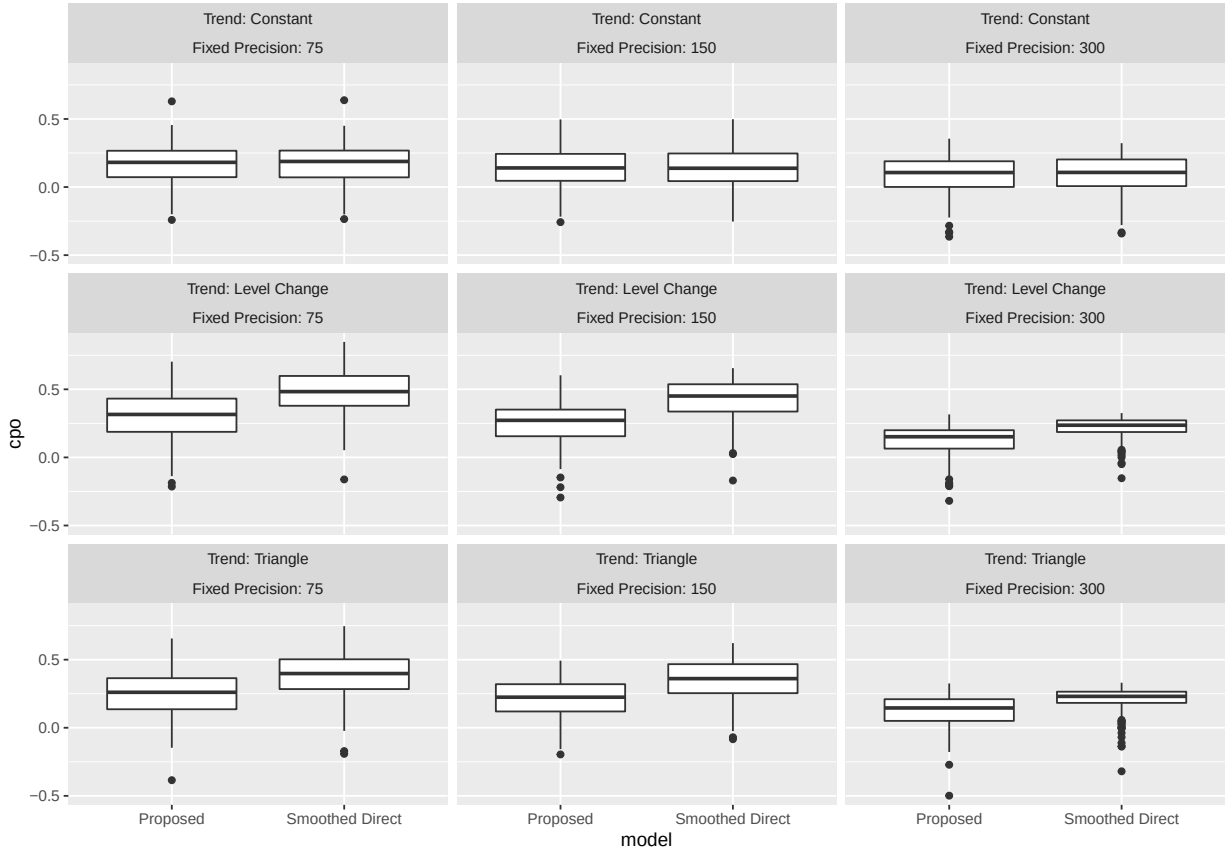


Figure 4.9: LS for simulation settings where τ_i is not the same for all time points.

Burundi, Ethiopia, Kenya, Rwanda, Tanzania, and Uganda using DHS surveys from 2016, 2016, 2014, 2015, 2015, and 2016 respectively. First, to understand between-country variation, we will fit separate smoothed direct models to each country, with ICARs for the structured spatial random effect. We will then consider two sets of models to compare:

1. The smoothed direct model with an ICAR for the structured spatial random effect, which we will refer to as the smoothed direct model, and the smoothed direct model with our proposed multi-country AICAR for the structured spatial random effect, which we will refer to as the proposed model.

Table 4.1: Comparison of parameter estimates for the smoothed direct and proposed models in the Rwanda application.

Model	Parameter	Mean	SD	2.5% Quantile	Median	97.5% Quantile	Mode
Smoothed Direct	μ	-2.08	0.05	-2.18	-2.08	-1.99	-2.08
Smoothed Direct	τ	7.05	2.23	3.49	6.79	12.17	6.28
Smoothed Direct	ϕ	0.96	0.04	0.85	0.98	1.00	1.00
Smoothed Direct	ν_{1992}	0.02	0.04	-0.05	0.02	0.11	0.01
Smoothed Direct	ν_{2000}	0.00	0.03	-0.06	0.00	0.07	0.00
Smoothed Direct	ν_{2005}	0.04	0.03	-0.02	0.04	0.12	0.04
Smoothed Direct	ν_{2008}	-0.07	0.04	-0.16	-0.07	-0.00	-0.06
Smoothed Direct	ν_{2010}	0.00	0.03	-0.07	0.00	0.07	0.00
Smoothed Direct	ν_{2015}	-0.00	0.04	-0.08	-0.00	0.07	0.00
Proposed	μ	-2.08	0.05	-2.17	-2.08	-1.99	-2.08
Proposed	τ	9.62	3.27	4.55	9.18	17.24	8.32
Proposed	ϕ	0.98	0.03	0.89	0.99	1.00	1.00
Proposed	θ	0.43	0.19	0.13	0.41	0.82	0.32
Proposed	ν_{1992}	0.02	0.04	-0.05	0.02	0.12	0.02
Proposed	ν_{2000}	0.00	0.03	-0.06	0.00	0.07	-0.00
Proposed	ν_{2005}	0.05	0.03	-0.02	0.04	0.12	0.04
Proposed	ν_{2008}	-0.07	0.04	-0.16	-0.07	-0.00	-0.06
Proposed	ν_{2010}	0.00	0.03	-0.07	0.00	0.07	0.00
Proposed	ν_{2015}	-0.00	0.04	-0.08	-0.00	0.07	0.00

2. The smoothed direct model and proposed model, but replacing the intercept μ in (4.1) with country-specific intercepts $\mu_{c[i]}$, where $c[i]$ denotes the country in which region i resides. We will refer to these models as the smoothed direct country-intercept model and proposed country-intercept model.

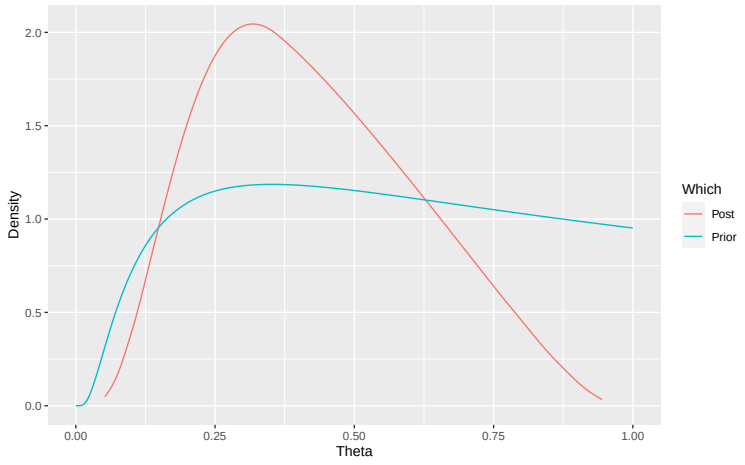


Figure 4.10: Comparison of prior and posterior density for θ in the Rwanda application.

4.7.1 Country-Specific Models

In Table 4.2 we display summaries of the posterior for the various parameters from fitting the smoothed direct model to each country separately. The posterior summaries for the intercepts for Kenya and Rwanda are smaller than intercepts for the rest of the countries, indicating that U5MR is on average smaller in these countries across Admin1 regions. All countries except Burundi have roughly comparable posterior summaries for ϕ: a posterior mode close to 0 and a posterior median close to 0.3. This indicates for these countries that there is not much weight being placed on the structured spatial random effect. Burundi has a posterior mode at 0.23 and a posterior median of 0.43, so while again there is not much weight being placed on the structured spatial random effect, there is still more spatial smoothing occurring than in the other five countries. The parameter with the most heterogeneity across countries is τ. Besides Uganda, the countries have posterior medians from around 10-30, with varying posterior standard deviations from around 5 to 70. Uganda however has a very large posterior summaries for precision, with posterior median of 457. This indicates that there is a large amount of heterogeneity in how the smoothed direct models for each country weight the total region-level random effect. Uganda in particular has such large posterior summaries

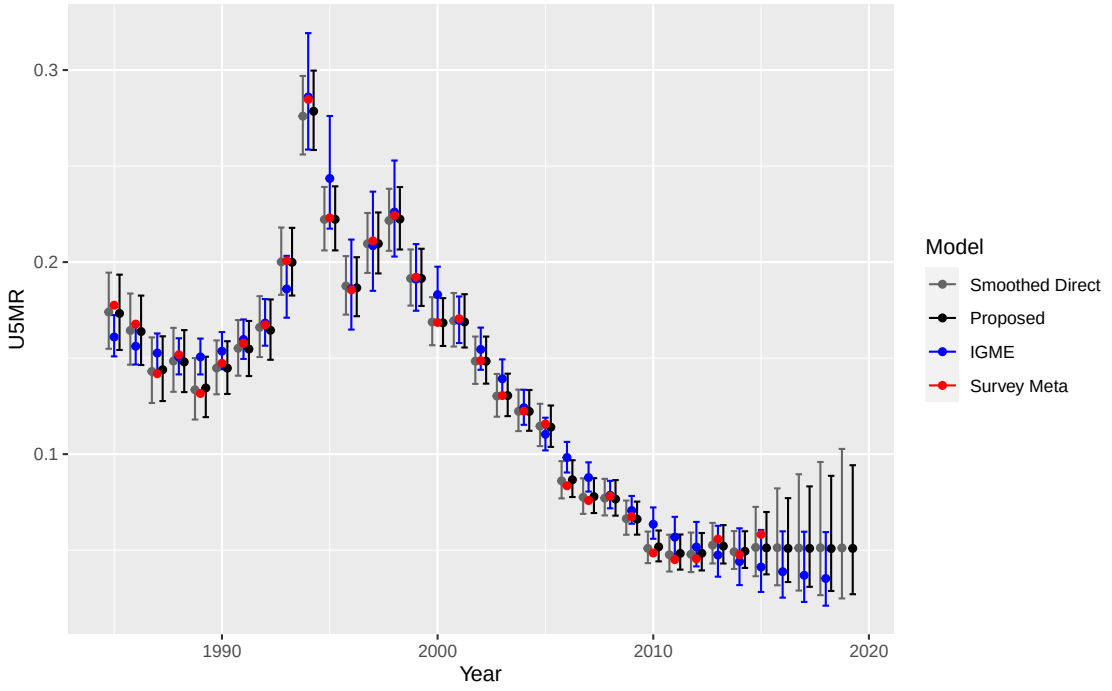


Figure 4.11: Comparison of U5MR estimates from the smoothed direct model, the proposed model, IGME, and meta-analysis estimator of U5MR based on the direct estimates.

for τ that the Admin1 region estimates are heavily smoothed towards the country's direct estimate.

4.7.2 Smoothed Direct and Proposed Models

In Table 4.3 we display summaries of the posterior for the various parameters from fitting the smoothed direct and proposed models. We see that the posterior summaries for μ is comparable between models. However, the posterior summaries for τ and ϕ are all larger under the proposed model. In particular, the larger posterior summaries of ϕ in the proposed model indicate that it is placing more weight on the structured spatial random effect.

Focusing on θ, in Figure 4.12 we plot the prior and posterior for θ from the proposed model. We see that the prior is relatively flat from 0.2 to 1, with a mode at 0.23, and 95%

Table 4.2: Comparison of parameter estimates from fitting the smoothed direct model to each country separately.

Country	Parameter	Mean	SD	2.5% Quantile	Median	97.5% Quantile	Mode
Burundi	μ	-2.60	0.07	-2.74	-2.60	-2.46	-2.60
Burundi	τ	11.58	5.36	4.37	10.53	24.96	8.69
Burundi	ϕ	0.45	0.25	0.06	0.43	0.91	0.23
Ethiopia	μ	-2.49	0.08	-2.65	-2.48	-2.33	-2.48
Ethiopia	τ	21.66	15.80	5.06	17.44	63.38	11.56
Ethiopia	ϕ	0.31	0.25	0.02	0.24	0.86	0.03
Kenya	μ	-2.94	0.07	-3.09	-2.95	-2.79	-2.95
Kenya	τ	47.57	47.91	7.16	33.45	173.17	17.61
Kenya	ϕ	0.31	0.25	0.01	0.24	0.87	0.03
Rwanda	μ	-3.04	0.11	-3.28	-3.04	-2.83	-3.03
Rwanda	τ	44.53	70.11	3.64	24.17	211.64	9.09
Rwanda	ϕ	0.35	0.26	0.02	0.29	0.91	0.04
Tanzania	μ	-2.63	0.07	-2.77	-2.63	-2.50	-2.63
Tanzania	τ	18.64	12.02	5.63	15.46	50.22	11.17
Tanzania	ϕ	0.32	0.26	0.01	0.25	0.90	0.02
Uganda	μ	-2.61	0.05	-2.71	-2.61	-2.52	-2.61
Uganda	τ	3783.69	37696.07	18.76	457.11	25111.91	34.55
Uganda	ϕ	0.37	0.27	0.02	0.31	0.91	0.05

of the prior mass lying in $[0.07, 0.97]$. The posterior has a mode at 0.11, which it is heavily concentrated around, and the 95% credible interval for θ is $[0.04, 0.65]$. Combining this with the posterior summaries for ϕ, we find that the proposed model is performing more spatial smoothing than the smoothed direct model, as ϕ is larger in the proposed model, but there is not much spatial smoothing occurring on the borders of different countries, as θ is close

Table 4.3: Comparison of parameter estimates for the smoothed direct and proposed models in the multi-country application.

Model	Parameter	Mean	SD	2.5% Quantile	Median	97.5% Quantile	Mode
Smoothed Direct	μ	-2.67	0.04	-2.75	-2.67	-2.59	-2.67
Smoothed Direct	τ	10.23	2.39	6.33	9.96	15.67	9.45
Smoothed Direct	ϕ	0.23	0.22	0.01	0.15	0.80	0.01
Proposed	μ	-2.67	0.03	-2.74	-2.67	-2.60	-2.67
Proposed	τ	12.42	3.48	6.89	11.99	20.45	11.17
Proposed	ϕ	0.49	0.23	0.09	0.49	0.90	0.49
Proposed	θ	0.24	0.16	0.04	0.19	0.65	0.11

to 0 in the proposed model.

In Figure 4.13 we map U5MR estimates from the smoothed direct model and the proposed model. To get a better grasp on how the estimates differ between the two models, in Figure 4.14 we plot U5MR estimates from the smoothed direct model and the proposed model, in addition to the country-specific smoothed direct model fits, direct estimates for each Admin1 region, and direct estimates for each country. We also include whether each Admin1 region borders a different country. The estimates from the smoothed direct and proposed models differ the most among Admin1 regions that do not border a different country. In particular, in these regions, the estimates from the proposed model are pulled towards their country level direct estimate. This can be explained by the proposed model estimating larger values of ϕ. This places more weight on the structured spatial random effect, leading to more spatial smoothing within countries.

4.7.3 Smoothed Direct and Proposed Country-Intercept Models

In Table 4.4 we display summaries of the posterior for the various parameters from fitting the smoothed direct and proposed country-intercept models. We see that the posterior esti-

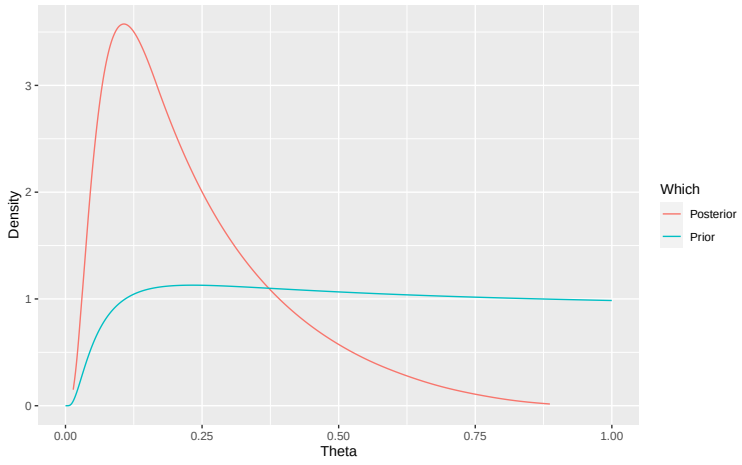


Figure 4.12: Comparison of prior and posterior density for θ the proposed model for the multi-country application.

mates for the country-specific intercepts are comparable between models, whereas the posterior standard deviations are larger in the proposed country-intercept model. The posterior summaries for roughly comparable for τ and ϕ, with the smoothed direct country-intercept model having slightly larger summaries for ϕ and the proposed country-intercept model having slightly larger summaries for τ.

Focusing on θ, in Figure 4.15 we plot the prior and posterior for θ from the proposed model. The posterior has a mode at 0.29, which it is not heavily concentrated around, and the 95% credible interval for θ is $[0.06, 0.95]$. We see that the posterior is very close to the prior, with the main difference being that the posterior does not place much mass immediately around 1. Thus the data did not inform the posterior for θ beyond the fact that it is not 1.

In Figure 4.16 we map U5MR estimates from the smoothed direct and proposed country-intercept models. To get a better grasp on how the estimates differ between the two models, in Figure 4.17 we plot U5MR estimates from the smoothed direct and proposed country-intercept models, in addition to the country-specific smoothed direct model fits, direct es-

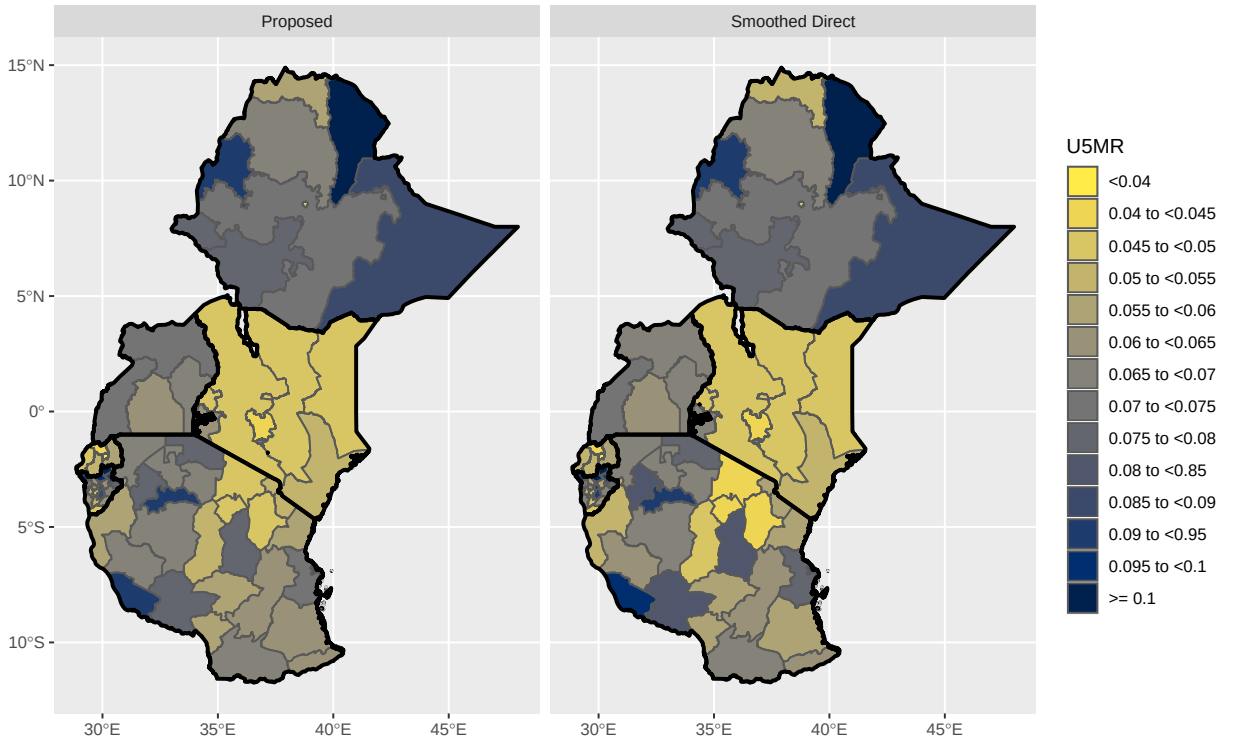


Figure 4.13: Maps of U5MR estimates from the smoothed direct model and the proposed model.

timates for each Admin1 region, and direct estimates for each country. We also include whether each Admin1 region borders a different country. The smoothed direct and proposed country-intercept models have close to identical estimates in nearly all regions.

In this multi-country application we have found that while our AICAR improves estimate of U5MR when country-specific intercepts are not used, it does not provide benefit for estimation of U5MR in the presence of country-specific intercepts. Country-specific intercepts are not always used in such mutli-country settings [see e.g. 24], but we believe that they are natural to consider. In contrast, in the temporal Rwanda application in the previous Section, the results did not substantively change when conflict-period specific intercepts were

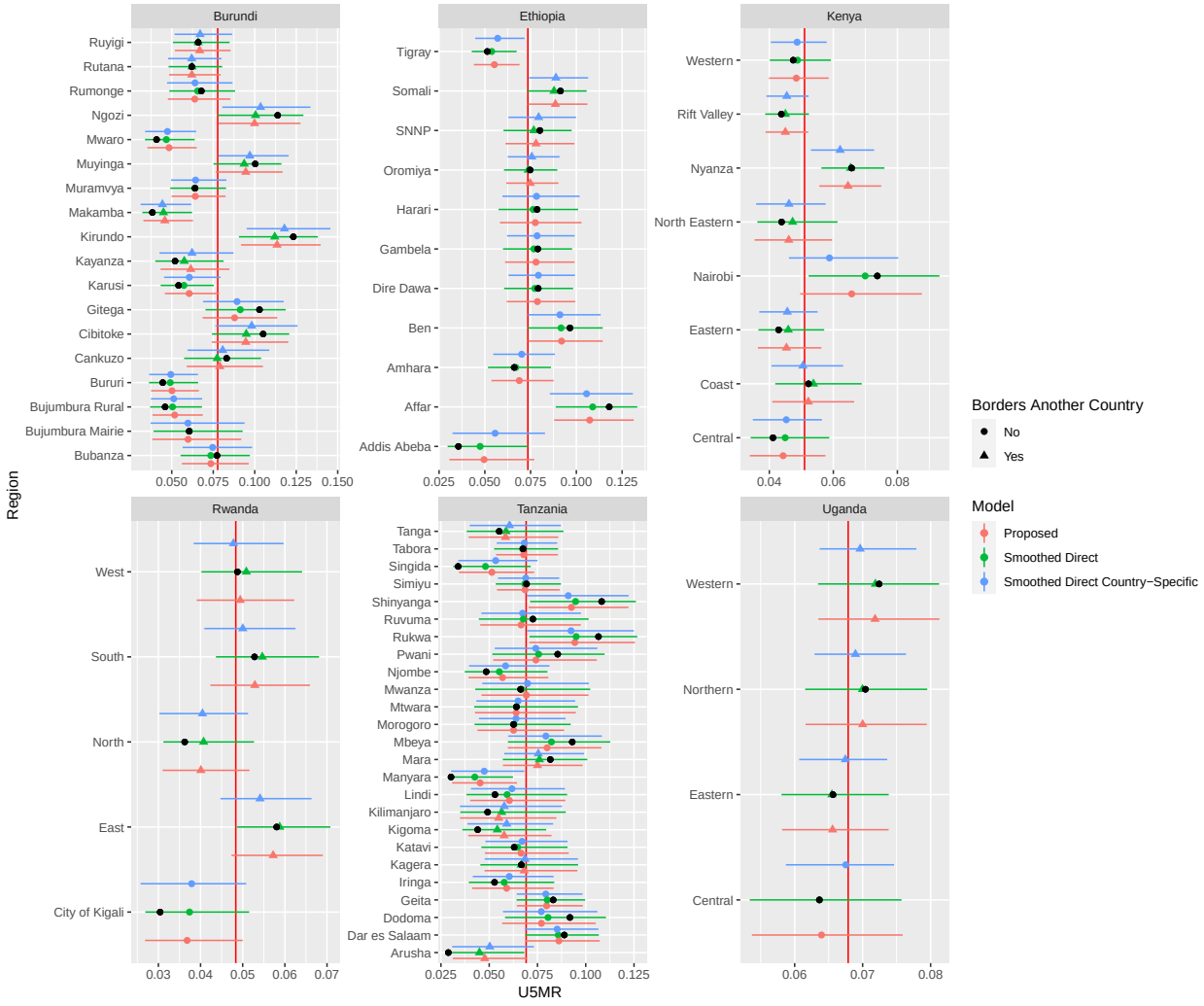


Figure 4.14: Comparison of U5MR estimates from the smoothed direct model and the proposed model, in addition to the country-specific smoothed direct model fits, direct estimates for each Admin1 region in black, and direct estimates for each country in red.

introduced. Thus, it would seem that the benefit of our AGMRFs is application dependent. We conjecture that our AICAR would improve estimation of U5MR, when country-specific intercepts are used, in settings where some of the countries have much less data than other countries. This data scarcity would make the structured spatial random effect more influen-

Table 4.4: Comparison of parameter estimates for the smoothed direct and proposed country-intercept models in the multi-country application.

Model	Country	Param.	Mean	SD	2.5% Quantile	Median	97.5% Quantile	Mode
Smoothed Direct C-Int.	Burundi	μ	-2.56	0.13	-2.83	-2.56	-2.29	-2.57
Smoothed Direct C-Int.	Ethiopia	μ	-2.36	0.21	-2.75	-2.37	-1.91	-2.41
Smoothed Direct C-Int.	Kenya	μ	-2.93	0.14	-3.21	-2.93	-2.65	-2.93
Smoothed Direct C-Int.	Rwanda	μ	-3.17	0.18	-3.54	-3.16	-2.82	-3.16
Smoothed Direct C-Int.	Tanzania	μ	-2.69	0.11	-2.92	-2.68	-2.49	-2.67
Smoothed Direct C-Int.	Uganda	μ	-2.67	0.18	-3.03	-2.67	-2.32	-2.66
Smoothed Direct C-Int.		τ	12.80	3.96	6.78	12.21	22.19	11.12
Smoothed Direct C-Int.		ϕ	0.57	0.26	0.08	0.60	0.96	0.86
Proposed C-Int.	Burundi	μ	-2.57	0.18	-2.94	-2.57	-2.20	-2.57
Proposed C-Int.	Ethiopia	μ	-2.36	0.30	-2.94	-2.38	-1.73	-2.41
Proposed C-Int.	Kenya	μ	-2.93	0.19	-3.31	-2.93	-2.54	-2.93
Proposed C-Int.	Rwanda	μ	-3.17	0.24	-3.68	-3.17	-2.71	-3.15
Proposed C-Int.	Tanzania	μ	-2.68	0.14	-2.97	-2.68	-2.42	-2.67
Proposed C-Int.	Uganda	μ	-2.67	0.26	-3.20	-2.67	-2.15	-2.66
Proposed C-Int.		τ	13.50	3.97	7.39	12.92	22.86	11.84
Proposed C-Int.		ϕ	0.53	0.25	0.08	0.54	0.94	0.71
Proposed C-Int.		θ	0.51	0.27	0.06	0.50	0.95	0.29

tial, and it would be interesting to see in such a setting whether the AICAR would improve over the performance of an ICAR for the structured spatial random effect.

4.8 Conclusions

In this chapter, we developed a class of adaptive GMRFs which can incorporate knowledge of expected shocks in mortality, while still allowing information to be borrowed across space and time. Methodologically, an important next step will be to develop spatio-temporal models that incorporate our adaptive GMRFs. A natural first step would be to utilize the spatio-temporal interaction framework of [70]. For example, if one wanted to estimate U5MR across multiple countries over multiple time periods, one could use a type IV interaction from [70],

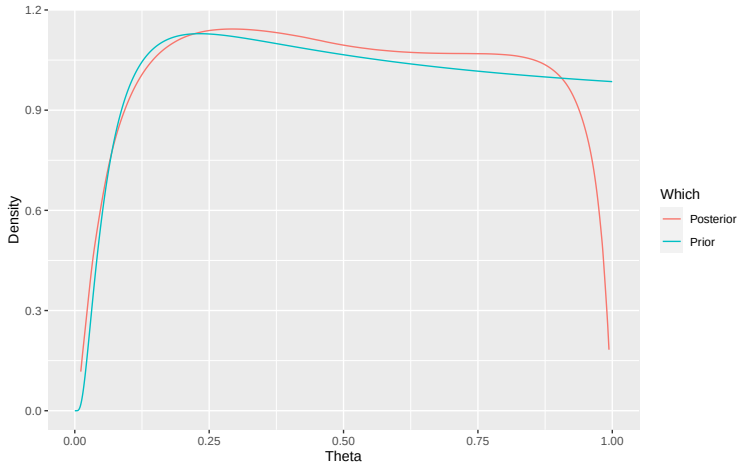


Figure 4.15: Comparison of prior and posterior density for θ the proposed country-intercept model for the multi-country application.

where the spatial component is a multi-country AICAR and the temporal component is a non-adaptive GMRF, like a RW1.

In the Rwanda application, we focused in this chapter on developing a national model for U5MR from 1985-2015 that did not oversmooth during conflict years. However, we would ultimately like to develop a *subnational* model for U5MR from 1985-2015. This poses challenges, as earlier surveys that provide the most information for conflict years do not have GPS information available for sampled households, unlike later surveys. The only geographic information these earlier surveys have for sampled households is administrative region at the time of the survey, and the administrative division of Rwanda changed in 2006. Thus, future work will need to develop a spatial model that harmonizes these different administrative divisions. This new spatial model could then be combined with the conflict ARW1 to develop a subnational model for U5MR from 1985-2015.

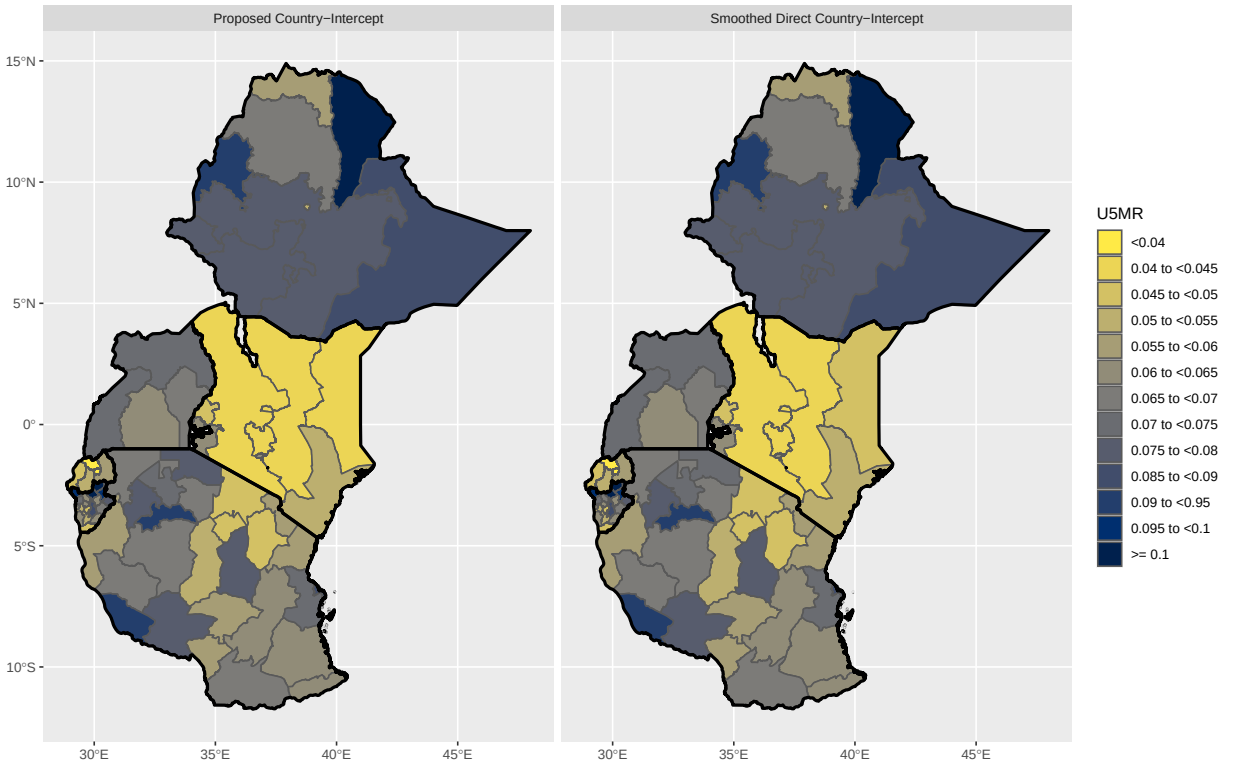


Figure 4.16: Maps of U5MR estimates from the smoothed direct and proposed country-intercept models.

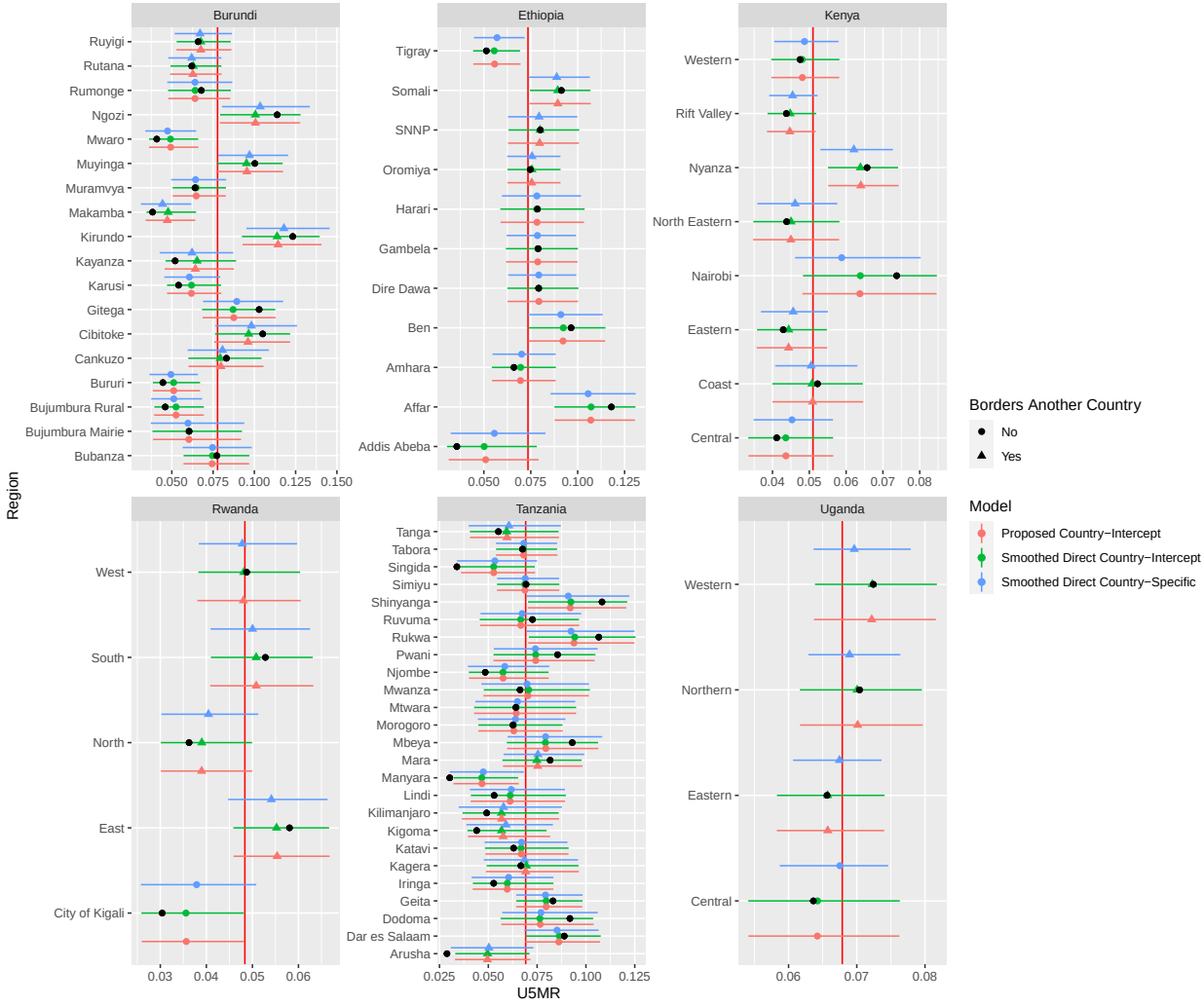


Figure 4.17: Comparison of U5MR estimates from the smoothed direct and proposed country-intercept models, in addition to the country-specific smoothed direct model fits, direct estimates for each Admin1 region in black, and direct estimates for each country in red.

Chapter 5

DISCUSSION AND FUTURE WORK

In this dissertation, we have addressed three problems related to the statistical methodology used to assess the extent of mortality in the wake of a conflict. Chapter 2 proposed a novel Bayesian approach to the general problem of multifile record linkage and duplicate detection. However, there is much work that needs to be done to get modern Bayesian approaches used regularly in practice for record linkage. There is currently no guidance in the literature for how practitioners are supposed to choose amongst the many competing methods available for Bayesian record linkage. Comparison studies of competing methods need to be undertaken to help practitioners understand the relative computational and statistical performance of approaches in the direct-modeling and comparison-based frameworks. In order to get practitioners to actually use some of the competing methods, general purpose software needs to be developed for Bayesian approaches. The current software for Bayesian record linkage is disjoint and bespoke, so a practitioner familiar with the software of one method is not necessarily prepared to use the software of a different method. Developing general purpose software capable of implementing a wide array of methods would allow for greater usability, and would additionally help facilitate methodological work and comparison studies. Further, graphical user interfaces need to be built on top of general purpose software to further expand the possible user base for Bayesian approaches outside of those comfortable with coding in statistical programming languages like `R`. An important part of such general purpose software would be standardized Markov chain Monte Carlo convergence diagnostics. Essentially all Bayesian approaches to record linkage currently use Markov chain Monte Carlo samplers to explore the posterior distribution over the space of partitions, a high-dimensional discrete parameter space. However, there is no guidance in the literature for how to tune

these samplers in practice. Good default settings are necessary for non-statistical experts to reliably use these approaches.

Chapter 3 presented a re-framing of the MSE problem that leads to an approach which places the identifying assumption front and center in the MSE workflow. It is important to continue research in this vein to make MSE approaches that center the identifying assumption common place in practice. For example, it may be the case in a given application that no available identifying assumptions are appropriate based on the context of the data. Thus it is important to develop new interpretable and explicitly specified identifying assumptions that practitioners can choose from and interpretable sensitivity analyses that allow practitioners to examine impact of these new identifying assumptions on population size estimates. It may also be useful to develop explicit *partial* identifying assumptions that set identify, rather than point identify, MSE models. The assumptions that a user is required to make to set identify a model can be much weaker than the assumptions required to achieve point identification. This can allow more leeway for users when attempting to justify an assumption in the context of their data.

While we separately considered the problems of record linkage and MSE in this dissertation, they are closely related in practice. When record linkage is used to identify which individuals are the same across data sources, it is desirable to propagate the uncertainty from the linkage to the MSE method. It would be interesting in the future to combine the techniques from Chapters 2 and 3 to provide a joint approach to record linkage and MSE that allows the usage of modern Bayesian record linkage models with MSE approaches that use explicitly specified identifying assumptions. Current joint approaches [82, 131, 132] do not allow the usage of common MSE approaches, nor would they be amenable to MSE approaches that use explicitly specified identifying assumptions. However, the structured prior for multifile partitions in Chapter 2 provides a natural route to incorporate arbitrary MSE models into a Bayesian record linkage model, as it places a prior directly on the overlap table which is viewed as data in MSE methods.

Chapter 4 developed spatial and temporal smoothing models which incorporate knowl-

edge of expected shocks in mortality. The main motivation for this chapter was the development of a subnational model for U5MR in Rwanda from 1985-2015 that does not oversmooth during conflict years. Earlier surveys in Rwanda provide the most information for conflict years, but do not have GPS information available for sampled households unlike later surveys. Instead, these earlier surveys contain administrative region at the time of the survey for sampled households, and the administrative division of Rwanda changed in 2006. While it would be straightforward to produce U5MR estimates for these out of date administrative regions, we unfortunately would like to produce U5MR estimates for the current administrative regions. We thus need to develop a spatial model that is able to harmonize the data from these different administrative divisions. This is known as the spatial misalignment problem [see e.g. Chapter 7 of 11]. This new spatial model that takes into account the spatial misalignment could then be combined with the conflict ARW1 to form a type IV spatio-temporal interaction [70] that can be used as the backbone of a subnational model for U5MR in Rwanda from 1985-2015 that does not oversmooth during conflict years.

BIBLIOGRAPHY

[1] ABA/AAAS. Political killings in Kosova/Kosovo, March-June 1999. Technical report, American Bar Association Central and East European Law Initiative and the American Association for the Advancement of Science, 2000.

[2] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[3] Leontine Alkema and Jin Rou New. Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, pages 2122–2149, 2014.

[4] Leontine Alkema, Jin Rou New, Jon Pedersen, Danzhen You, all members of the UN Inter-agency Group for Child Mortality Estimation, and its Technical Advisory Group. Child mortality estimation 2013: an overview of updates in estimation methods by the United Nations Inter-agency Group for Child Mortality Estimation. *PloS one*, 9(7):e101112, 2014.

[5] Elizabeth S Allman, Catherine Matias, John A Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[6] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[7] Margo Anderson and Stephen E Fienberg. *Who Counts?: The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation, 1999.

[8] Sophie Baillargeon, Louis-Paul Rivest, et al. Rcapture: loglinear models for capture-recapture in R. *Journal of Statistical Software*, 19(5):1–31, 2007.

[9] Patrick Ball, Wendy Betts, Fritz Scheuren, Jana Dudukovich, and Jana Asher. *Killings and Refugee Flow in Kosovo March-June 1999*. American Association for the Advancement of Science and American Bar Association Central and East European Law Initiative, 2002.

[10] Patrick Ball and Megan Price. Using statistics to assess lethal violence in civil and inter-state war. *Annual Review of Statistics and its Application*, 6:63–84, 2019.

[11] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[12] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. *Machine Learning*, 56(1):89–113, Jul 2004.

[13] James O Berger, Jose M Bernardo, and Dongchu Sun. Overall objective priors. *Bayesian Analysis*, 10(1):189–221, 2015.

[14] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

[15] Brenda Betancourt, Juan Sosa, and Abel Rodríguez. A Prior for Record Linkage Based on Allelic Partitions. *arXiv preprint arXiv:2008.10118*, 2020.

[16] Brenda Betancourt, Giacomo Zanella, and Rebecca C Steorts. Random Partition Models for Microclustering Tasks. *Journal of the American Statistical Association*, pages 1–13, 2020.

[17] M. Bilenko, R. J. Mooney, W. W. Cohen, P. Ravikumar, and S. E. Fienberg. Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems*, 18(5):16–23, October 2003.

[18] David A Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978.

[19] Olivier Binette and Rebecca C Steorts. (Almost) All of Entity Resolution. *arXiv preprint arXiv:2008.04443*, 2020.

[20] Sheila M Bird and Ruth King. Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and its Application*, 5:95–118, 2018.

[21] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press. Reprinted in 2007 by Springer, New York, 1975.

[22] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Finding Graph Clusterings with Maximum Modularity. In *Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer Berlin Heidelberg, 2007.

[23] Mark J Brewer and Andrew J Nolan. Variable smoothing in Bayesian intrinsic autoregressions. *Environmetrics: The official journal of the International Environmetrics Society*, 18(8):841–857, 2007.

[24] Roy Burstein et al. Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature*, 574(7778):353–358, 2019.

[25] Bradley P Carlin and Haijun Ma. Bayesian multivariate areal wombling for multiple disease boundary analysis. *Bayesian analysis*, 2(2):281–302, 2007.

[26] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[27] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[28] Peter Christen. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555, September 2012.

[29] John N Darroch. The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, 45(3/4):343–359, 1958.

[30] A DasGupta and Herman Rubin. Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference*, 130(1-2):391–404, 2005.

[31] Damien De Walque and Philip Verwimp. The demographic and socio-economic distribution of excess mortality during the 1994 genocide in Rwanda. *Journal of African Economies*, 19(2):141–162, 2010.

[32] Petros Dellaportas and Jonathan J Forster. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633, 1999.

[33] Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2):172 – 187, 2006.

[34] Departamento Administrativo Nacional de Estadisticas, DANE. Metodologia Estadisticas Vitales. Technical Report 82, Direccion de Censos y Demografia, 2009.

[35] David B Dunson and Chuanhua Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

[36] Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371, 2019.

[37] Ted Enamorado and Rebecca C Steorts. Probabilistic Blocking and Distributed Bayesian Entity Resolution. In *International Conference on Privacy in Statistical Databases*, pages 224–239. Springer, 2020.

[38] Alessio Farcomeni and Luca Tardella. Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics*, 6:2602–2626, 2012.

[39] James R Faulkner and Vladimir N Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian analysis*, 13(1):225, 2018.

[40] James Robert Faulkner. *Adaptive Bayesian Nonparametric Smoothing with Markov Random Fields and Shrinkage Priors*. PhD thesis, University of Washington, 2019.

[41] Robert E Fay and Roger A Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

[42] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[43] Marco Antonio Rosa Ferreira and Herbert KH Lee. *Multiscale modeling: a Bayesian perspective*, volume 2. Springer, 2007.

[44] Marco AR Ferreira, David Higdon, Herbert KH Lee, and Mike West. Multi-scale random field models. Technical report, Technical Report, UFRJ-DME, 2005.

[45] Stephen E Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591–603, 1972.

[46] Stephen E Fienberg, Matthew S Johnson, and Brian W Junker. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3):383–405, 1999.

[47] Stephen E Fienberg and Daniel Manrique-Vallier. Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *AStA Advances in Statistical Analysis*, 93(1):49–60, 2009.

[48] Marco Fortini, Brunero Liseo, Alessandra Nuccitelli, and Mauro Scanu. On Bayesian record linkage. *Research in Official Statistics*, 4(1):185–198, 2001.

[49] Anna Freni-Sterrantino, Massimo Ventrucci, and Håvard Rue. A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34, 2018.

[50] Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

[51] Susanna C Gerritse, Peter GM van der Heijden, and Bart FM Bakker. Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of Official Statistics*, 31(3):357–379, 2015.

[52] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[53] IJ Good and Ray A Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[54] Paul Gustafson. Bayesian inference for partially identified models. *The International Journal of Biostatistics*, 6(2), 2010.

[55] Shelby J Haberman. *Analysis of Qualitative Data. Volume 2*. Academic Press, 1979.

[56] Matthew J Heaton. Wombling analysis of childhood tumor rates in Florida. *Statistics and Public Policy*, 1(1):60–67, 2014.

[57] Radu Herbei and Marten H Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721, 2006.

[58] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[59] Michel H Hof, Anita C Ravelli, and Aeilko H Zwinderman. A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515, 2017.

[60] Joseph W Hogan and Michael J Daniels. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC, 2008.

[61] Hajo Holzmann, Axel Munk, and Walter Zucchini. On identifiability in capture–recapture models. *Biometrics*, 62(3):934–936, 2006.

[62] Ernest B Hook and Ronald R Regal. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews*, 17(2):243–264, 1995.

[63] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[64] Richard Huggins. A note on the difficulties associated with the analysis of capture–recapture experiments with heterogeneous capture probabilities. *Statistics & probability letters*, 54(2):147–152, 2001.

[65] Vincent Iacopino, Martina W Frank, Heidi M Bauer, Allen S Keller, Sheri L Fink, Doug Ford, Daniel J Pallin, and Ronald Waldman. A population-based assessment of human rights abuses committed against ethnic Albanian refugees from Kosovo. *American Journal of Public Health*, 91(12):2013–2018, 2001.

[66] Christopher Jackson. Multi-state models for panel data: the msm package for R. *Journal of statistical software*, 38(1):1–28, 2011.

[67] Matthew A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, June 1989.

[68] Ruth King and SP Brooks. On the Bayesian analysis of population size. *Biometrika*, 88(2):317–336, 2001.

[69] Arto Klami and Aditya Jitta. Probabilistic size-constrained microclustering. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 329–338. AUAI Press, 2016.

[70] Leonhard Knorr-Held. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567, 2000.

[71] Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Bradley M Bell. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70:1–21, 2016.

[72] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117, Nov 2009.

[73] Stefan Lang, Eva-Maria Pronk, and Ludwig Fahrmeir. Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17(4):479–499, 2002.

[74] Michael D. Larsen. Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory. In *Proceedings of the Section on Survey Research Methods*, pages 3277–3284. American Statistical Association, 2005.

[75] Michael D. Larsen and Donald B. Rubin. Iterative Automated Record Linkage Using Mixture Models. *Journal of the American Statistical Association*, 96(453):32–41, March 2001.

[76] Zehang Li, Yuan Hsiao, Jessica Godwin, Bryan D Martin, Jon Wakefield, Samuel J Clark, with support from the United Nations Inter-agency Group for Child Mortality Estimation, and its technical advisory group. Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PloS one*, 14(1):e0210645, 2019.

[77] Zehang R Li, Bryan D Martin, Tracy Q Dong, Geir-Arne Fuglstad, Jessica Godwin, John Paige, Andrea Riebler, Samuel Clark, and Jon Wakefield. *Space-Time Smoothing of Demographic and Health Indicators using the R Package SUMMER*, 2020.

[78] Antonio R Linero. Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika*, 104(2):327–341, 2017.

[79] William A Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003.

[80] William A Link. Rejoinder to "On Identifiability in Capture-Recapture Models". *Biometrics*, 62(3):936–939, 2006.

[81] William A Link. A cautionary note on the discrete uniform prior for the binomial N. *Ecology*, 94(10):2173–2179, 2013.

[82] Brunero Liseo and Andrea Tancredi. Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets. *Journal of Official Statistics*, 27(3):491–505, 2011.

[83] Haolan Lu and Bradley P Carlin. Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285, 2005.

[84] Haolan Lu, Cavan S Reilly, Sudipto Banerjee, and Bradley P Carlin. Bayesian areal wombling via adjacency modeling. *Environmental and ecological statistics*, 14(4):433–452, 2007.

[85] Kristian Lum and Patrick Ball. Estimating undocumented homicides with two lists and list dependence. Technical report, Human Rights Data Analysis Group, 2015.

[86] David Madigan and Jeremy C York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.

[87] Daniel Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.

[88] Daniel Manrique-Vallier, Patrick Ball, and David Sulmont. Estimating the Number of Fatal Victims of the Peruvian Internal Armed Conflict, 1980-2000: an application of modern multi-list Capture-Recapture techniques. *arXiv preprint arXiv:1906.04763*, 2019.

[89] Daniel Manrique-Vallier, Megan E Price, and Anita Gohdes. Multiple systems estimation techniques for estimating casualties in armed conflicts. *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*, pages 165–182, 2013.

[90] Neil G Marchant, Andee Kaplan, Daniel N Elazar, Benjamin IP Rubinstein, and Rebecca C Steorts. d-blink: Distributed end-to-end Bayesian entity resolution. *Journal of Computational and Graphical Statistics*, pages 1–16, 2021.

[91] Nicholas Elias Matsakis. *Active Duplicate Detection with Bayesian Nonparametric Models*. PhD thesis, Massachusetts Institute of Technology, 2010.

[92] Brendan S McVeigh, Bradley T Spahn, and Jared S Murray. Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking: An Application to the California Great Registers. *arXiv preprint arXiv:1905.05337*, 2019.

[93] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[94] Laina D Mercer, Jon Wakefield, Athena Pantazis, Angelina M Lutambi, Honorati Masanja, and Samuel Clark. Space-time smoothing of complex survey data: small area estimation for child mortality. *The annals of applied statistics*, 9(4):1889, 2015.

[95] Jeffrey Miller, Brenda Betancourt, Abbas Zaidi, Hanna Wallach, and Rebecca C Steorts. Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *arXiv preprint arXiv:1512.00792*, 2015.

[96] Chiranjit Mukherjee, Prasad S Kasibhatla, and Mike West. Spatially varying SAR models and Bayesian inference for high-resolution lattice data. *Annals of the Institute of Statistical Mathematics*, 66(3):473–494, 2014.

[97] Jared S Murray. Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *Journal of Privacy and Confidentiality*, 7(1), 2015.

[98] Yuval Nardi and Alessandro Rinaldo. The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli*, 18(3):945–974, 2012.

[99] M. E. J. Newman. Spectral methods for community detection and graph partitioning. *Phys. Rev. E*, 88:042822, Oct 2013.

[100] Hollie Nyseth Brehm. Subnational determinants of killing in Rwanda. *Criminology*, 55(1):5–31, 2017.

[101] David L Otis, Kenneth P Burnham, Gary C White, and David R Anderson. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135, 1978.

[102] Antony Overstall and Ruth King. conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58(7):1–27, 2014.

[103] George Lucas Moraes Pezzott, Luis Ernesto Bueno Salasar, José Galvão Leite, and Francisco Louzada-Neto. A note on identifiability and maximum likelihood estimation for a heterogeneous capture-recapture model. *Communications in Statistics-Theory and Methods*, pages 1–21, 2019.

[104] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[105] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006.

[106] Ronald R Regal and Ernest B Hook. The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, 10(5):717–721, 1991.

[107] Ronald R Regal and Ernest B Hook. Marginal versus conditional versus 'structural source' models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Statistics in Medicine*, 17(1):69–74, 1998.

[108] Brian J Reich and James S Hodges. Modeling longitudinal spatial periodontal data: A spatially adaptive model with tools for specifying priors and checking fit. *Biometrics*, 64(3):790–799, 2008.

[109] Jorge A. Restrepo and Katherine Aguirre. Homicidios y Muertes Violentas: Un Analisis Comparativo de las Fuentes en Colombia. *Forensis: Datos Para La Vida*, 2007.

[110] Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165, 2016.

[111] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[112] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[113] AL Rukhin. Statistical decision about the total number of observable objects. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 514–522, 1975.

[114] Mauricio Sadinle. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, 8(4):2404–2434, 2014.

[115] Mauricio Sadinle. Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017.

[116] Mauricio Sadinle. Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics*, 12(2):1013–1038, 2018.

[117] Mauricio Sadinle and Stephen E. Fienberg. A Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems. *Journal of the American Statistical Association*, 108(502):385–397, June 2013.

[118] Lalitha Sanathanan. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, pages 142–152, 1972.

[119] BW Silverman. Multiple systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *Journal of the Royal Statistical Society: Series A*, 183(3):691–736, 2020.

[120] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28, 2017.

[121] Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

[122] Sigrunn Holbek Sørbye and Håvard Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

[123] Paul B Spiegel and Peter Salama. War and mortality in Kosovo, 1998–99: an epidemiological testimony. *The Lancet*, 355(9222):2204–2209, 2000.

[124] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[125] Rebecca C. Steorts. Entity Resolution with Empirically Motivated Priors. *Bayesian Analysis*, 10(4):849–875, 2015.

[126] Rebecca C. Steorts, Rob Hall, and Stephen E. Fienberg. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.

[127] Rebecca C Steorts, Samuel L Ventura, Mauricio Sadinle, and Stephen E Fienberg. A comparison of blocking methods for record linkage. In *International Conference on Privacy in Statistical Databases*, pages 253–268. Springer, 2014.

[128] Jinghao Sun, Luk Van Baelen, Els Plettinckx, and Forrest W Crawford. Partial identification and dependence-robust confidence intervals for capture-recapture surveys. *arXiv preprint arXiv:2008.00127*, 2020.

[129] Behrooz Tahmasebi, Seyed Abolfazl Motahari, and Mohammad Ali Maddah-Ali. On the Identifiability of Finite Mixtures of Finite Product Measures. *arXiv preprint arXiv:1807.05444*, 2018.

[130] Elie Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

[131] Andrea Tancredi and Brunero Liseo. A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems. *Annals of Applied Statistics*, 5(2B):1553–1585, 2011.

[132] Andrea Tancredi, Rebecca Steorts, and Brunero Liseo. A Unified Framework for De-Duplication and Population Size Estimation (with Discussion). *Bayesian Analysis*, 15(2):633–682, 2020.

[133] Khoi-Nguyen Tran, Dinusha Vatsalan, and Peter Christen. GeCo: an online personal data generator and corruptor. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 2473–2476, 2013.

[134] United Nations General Assembly. Transforming our world: the 2030 agenda for sustainable development. Resolution Adopted by the General Assembly on 25 September 2015: 70/1., 2015.

[135] United Nations Inter-agency Group for Child Mortality Estimation. Levels & Trends in Child Mortality: Report 2019, Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation, 2019.

[136] United Nations Inter-agency Group for Child Mortality Estimation. Subnational Under-five Mortality Estimates, 1990–2019: Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation, 2021.

[137] Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

[138] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[139] John Whitehead, James Jackson, Alex Balch, and Brian Francis. On the Unreliability of Multiple Systems Estimation for Estimating the Number of Potential Victims of Modern Slavery in the UK. *Journal of Human Trafficking*, pages 1–13, 2019.

[140] W. E. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Association, 1990.

[141] W. E. Winkler. Advanced Methods for Record Linkage. In *Proceedings of the Section on Survey Research Methods*, pages 467–472. American Statistical Association, 1994.

[142] Yu Yue and Paul L Speckman. Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 19(1):96–116, 2010.

[143] Yu Ryan Yue, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis*, 9(2):397–424, 2014.

[144] Giacomo Zanella, Brenda Betancourt, Jeffrey W. Miller, Hanna Wallach, Abbas Zaidi, and Rebecca C. Steorts. Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems*, pages 1417–1425, 2016.

[145] Jing Zhou, Anirban Bhattacharya, Amy H Herring, and David B Dunson. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512):1562–1576, 2015.

Appendix A

APPENDIX FOR CHAPTER 1

A.1 Structured Prior Appendix

In this appendix, we prove Proposition 1 from Chapter 2 and provide additional guidance for the specification of the structured prior for multifile partitions.

A.1.1 Proof of Proposition 1

In this section, we restate and prove Proposition 1 from Chapter 2.

Proposition A.1. *The number of K-partite matchings that have the same overlap table, $\boldsymbol{n} = \{n_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$, is $\prod_{k=1}^{K} n_k! / \prod_{\boldsymbol{h}\in\mathcal{H}} n_{\boldsymbol{h}}!$, where $n_k = \sum_{\boldsymbol{h}\in\mathcal{H}} h_k n_{\boldsymbol{h}}$ is the number of entities in datafile $\boldsymbol{X}_k$. Thus $\mathbb{P}(\mathcal{C} \mid \{\mathcal{C}_k\}_{k=1}^{K}, \boldsymbol{n}) = \prod_{\boldsymbol{h}\in\mathcal{H}} n_{\boldsymbol{h}}! / \prod_{k=1}^{K} n_k!$.*

Proof. Let us first count all of the ways that we can place the clusters in file $\boldsymbol{X}_k$ into the overlap table cells that $\boldsymbol{X}_k$ is included in, $\mathcal{H}_k = \{\boldsymbol{h} \in \mathcal{H} : h_k = 1\}$. This is just a multinomial coefficient, $n_k!/(\prod_{\boldsymbol{h}\in\mathcal{H}_k} n_{\boldsymbol{h}}!)$. Thus the number of ways we can place all of the clusters from all of the files into the cells in $\mathcal{H}$ is $\prod_{k=1}^{K} n_k!/(\prod_{\boldsymbol{h}\in\mathcal{H}_k} n_{\boldsymbol{h}}!) = (\prod_{k=1}^{K} n_k!)/[\prod_{h\in\mathcal{H}}(n_{\boldsymbol{h}}!)^{\sum_{k=1}^{K} h_k}]$. Given that there are $n_{\boldsymbol{h}}$ clusters from each file with $h_k = 1$ in cell $\boldsymbol{h}$, now all we have to count is how many distinct complete matchings are possible between them, which is just $(n_{\boldsymbol{h}}!)^{(\sum_{k=1}^{K} h_k)-1}$. Thus the number of K-partite matchings is $[(\prod_{k=1}^{K} n_k!)/(\prod_{\boldsymbol{h}\in\mathcal{H}}(n_{\boldsymbol{h}}!)^{\sum_{k=1}^{K} h_k})][\prod_{\boldsymbol{h}\in\mathcal{H}}(n_{\boldsymbol{h}}!)^{\sum_{k=1}^{K} h_k}/n_{\boldsymbol{h}}!] = \prod_{k=1}^{K} n_k! / \prod_{\boldsymbol{h}\in\mathcal{H}} n_{\boldsymbol{h}}!$. □

A.1.2 Prior Specification Guidance

In this section we provide additional guidance for the specification of the structured prior for multifile partitions described in Section 3 of Chapter 2. In particular, we further discuss

the priors for the number of clusters, the overlap tables, and the within-file cluster sizes.

Prior for the Number of Entities or Clusters. In Chapter 2 we recommended, in the absence of substantial prior information, to use a uniform prior on $\{1, \cdots, U\}$ for the number of clusters, for some upper bound U. Our default recommendation was to set $U = r$, i.e. the total number of records. If one has substantive prior information about the number of clusters, this could instead be incorporated using other distributions on the positive integers. [95] and [144] both suggest to use a negative-binomial distribution with parameters $a > 0$ and $q \in (0, 1)$ truncated to the positive integers, i.e. $\mathbb{P}(n) \propto \frac{\Gamma(n+a)}{(n!)\Gamma(a)}(1-q)^a q^n I(n \in \mathbb{N})$. [144] further suggest a weakly informative specification for this Negative-binomial prior where a and q are selected such that $\mathrm{E}[n] = \sqrt{\mathrm{Var}(n)} = r/2$. We follow [95] and [144] and suggest a negative-binomial prior for the number of clusters n when incorporating substantive prior information.

Prior for the Overlap Table. In Chapter 2 we recommended using a Dirichlet-multinomial prior for the overlap table, specified by a collection of positive hyperparameters. In the absence of substantial prior information we recommended setting $\boldsymbol{\alpha} = (1, \cdots, 1)$. Due to conjugacy of the Dirichlet distribution with the multinomial, $\boldsymbol{\alpha}$ can be interpreted as prior cell counts, and thus our recommendation amounts to incorporating a prior count of 1 to each cell and an overall prior sample size of $2^K - 1$.

In our simulations and application, we found across a variety of overlap settings that this default prior performed satisfactorily. However, in the no-three-file-overlap setting, our approach struggled when there was a large amount of measurement error. We find in Appendix D.2 that we can improve performance in this setting by using informative prior cell counts, rather than using our default specification. What sets this no-three-file-overlap simulation setting apart from the other settings is that it is sparse, i.e. in this setting the count for the three-file-overlap cell of the overlap table is truly 0.

When linking a large number of files K, the size of the overlap table, $2^K - 1$, becomes large very quickly, which makes it likely that the true overlap table is sparse. Our default prior specification may not be appropriate in these settings as using a prior cell count of

1 for each cell may be incorporating prior information that is too strong, as illustrated in Example 1.4 of [13]. One possible alternative as a default specification when the overlap table is potentially sparse, would be to set $\boldsymbol{\alpha} = [1/(2^K - 1), \cdots, 1/(2^K - 1)]$ (see Section 3.2 of [13] for justification). If one has prior information concerning which cells of the overlap table are likely to be sparse, based on the results in Appendix D.2, we recommend attempting to incorporate this information into the prior. For example, if it is believed that some combination of files are likely not to have collected information on the same set of entities, one can incorporate this information by making the corresponding prior cell counts close to 0.

Another route one could take would be to replace the Dirichlet-multinomial prior on $\boldsymbol{n}$ with a multinomial prior on $\boldsymbol{n}$, with a non-Dirichlet prior on the multinomial cell probabilities. For example, one could use the tensor-factorization priors of [35] and [145], which have been shown to have lead to estimates of cell probabilities with good performance in large sparse contingency tables.

Prior for the Within-File Cluster Sizes. Given that in a multifile record linkage and duplicate detection scenario we do not expect there to be many duplicates per entity in any given file, in Chapter 2 we recommended specifying i.i.d. priors for the sizes of the within-file clusters, i.e. $d_{k1}, \cdots, d_{kn_k} \mid n_k \overset{iid}{\sim} p_k(\cdot)$ for a given file $\boldsymbol{X}_k$. Here $p_k(\cdot)$ represents the probability mass function of a distribution on $\{1, \cdots, U_k\}$, where U_k is a file-specific upper bound on cluster sizes.

When a given file $\boldsymbol{X}_k$ is assumed to have no duplicates, in Chapter 2 we recommended enforcing this restriction that there are no duplicates in that file by setting $U_k = 1$ and $p_k(d_{ki}) = I(d_{ki} = 1)$. When a given file $\boldsymbol{X}_k$ is assumed to have duplicates, in Chapter 2 we recommended a Poisson distribution with parameter $\lambda = 1$, i.e. $p_k(d_{ki}) \propto (d_{ki}!)^{-1} I(d_{ki} \in \{1, \cdots, U_k\})$. This prior places most of the prior mass close to 1, where various properties such as prior mean and standard deviation can be computed numerically (e.g. when $U_k = 10$ the prior mean is 1.58). If one has information on the average amount of duplication in file k, given a specified upper bound U_k, one could specify the parameter λ of the Poisson

prior such that the prior mean is equal to the average amount of duplication. Alternatively, one could place a hyperprior on λ. This is similar approach to [144], who used a negative-binomial distribution with parameters $r > 0$ and $p \in (0, 1)$ truncated to the positive integers, with a gamma hyperprior for r and a beta hyperprior for p. Indeed, the negative-binomial prior of [144] can be seen as a generalization of our Poisson prior, based on well-known connections between the Poisson and negative-binomial, and could be used instead of our Poisson recommendation if desired.

In the simulations presented in Appendix D.3 we explore using a Poisson prior with parameter λ varying over $\{0.1, 1, 2\}$, when the within-file cluster sizes are generated from a Poisson with parameter λ varying over $\{0.1, 1, 2\}$. We find in these simulations that when there is medium or high duplication (i.e. the within-file cluster sizes are generated from a Poisson with mean in $\{1, 2\}$), the results are not sensitive to λ, whereas when there is low duplication (i.e. the within-file cluster sizes are generated from a Poisson with mean 0.1), the results are sensitive to λ. This suggests that model performance is more sensitive to the specification of the prior distribution for within-file cluster sizes when there is a low amount of duplication, and that care should be taken when specifying the prior for the within-file cluster sizes for files which are expected to have very little duplication.

A.2 Posterior Inference Appendix

In this appendix, we first derive full conditional distributions of our structured prior for partitions, and then use the full conditional distributions to derive a Gibbs sampler for posterior inference in our model. We then discuss the computational complexity of our approach to posterior inference, how computational performance can be improved through the usage of indexing techniques, and the initialization of the Gibbs sampler.

A.2.1 Conditional Assignment Probabilities

In this section we use the form of the prior distribution in (2.1) to derive the conditional probability for assigning a record j from file $\boldsymbol{X}_k$ to a given cluster of an existing multifile partition $\mathcal{C}_{-j}$ of the other records. Specifically, we derive $\mathbb{P}(j \to c \mid \mathcal{C}_{-j})$, where $j \to c$ denotes adding record j to a cluster $c \in \mathcal{C}_{-j}$ or to an empty cluster. Let a quantity followed by $(\mathcal{C}_{-j})$ denote that it is derived from $\mathcal{C}_{-j}$ analogously to in Section 3.1 of Chapter 2. Let $\mathbf{1}_k$ denote the inclusion pattern indicating inclusion only in file $\boldsymbol{X}_k$, that is, $\mathbf{1}_k$ is a vector of zeroes except for its kth entry which equals 1. Further, let $\mathbf{1}_c$ denote the inclusion pattern of the cluster $c \in \mathcal{C}_{-j}$, that is, the lth entry of $\mathbf{1}_c$ is $I(c_l \neq \emptyset)$. Finally, let $\mathbf{1}_{c\cup j}$ denote the inclusion pattern of the cluster $c \in \mathcal{C}_{-j}$ after adding record j to it. Then the conditional assignment probability is

$$
\begin{aligned}
&\mathbb{P}(j \to c \mid \mathcal{C}_{-j}) \propto \\
&\begin{cases}
\left[\dfrac{\mathbb{P}(n(\mathcal{C}_{-j})+1)}{\mathbb{P}(n(\mathcal{C}_{-j}))}\right]\left[\dfrac{(n(\mathcal{C}_{-j})+1)(n_{\mathbf{1}_k}(\mathcal{C}_{-j})+\alpha_{\mathbf{1}_k})}{n(\mathcal{C}_{-j})+\alpha_0}\right] p_k(1) & , \text{ if } c = (\emptyset, \cdots, \emptyset) \\
\left[\dfrac{n_{\mathbf{1}_{c\cup j}}(\mathcal{C}_{-j})+\alpha_{\mathbf{1}_{c\cup j}}}{n_{\mathbf{1}_c}(\mathcal{C}_{-j})+\alpha_{\mathbf{1}_c}-1}\right] p_k(1) & , \text{ if } c \neq (\emptyset, \cdots, \emptyset), |c_k| = 0 \\
(|c_k|+1)\left[\dfrac{p_k(|c_k|+1)}{p_k(|c_k|)}\right] & , \text{ if } |c_k| > 0.
\end{cases}
\end{aligned}
$$

A.2.2 Gibbs Sampler

We will now derive a Gibbs sampler to explore the posterior of Φ and $\mathcal{C}$. Suppose we are at iteration $t+1$ of the sampler, with current samples $\Phi^{[t]} = (\boldsymbol{m}^{[t]}, \boldsymbol{u}^{[t]})$ and $\mathcal{C}^{[t]}$. Then we obtain the samples for iteration $t+1$ through the following steps:

1. For $k \leq k'$ and $f \in \{1, \ldots, F\}$, sample

$$\boldsymbol{m}_{kk'}^{f[t+1]} \mid \mathcal{C}^{[t]}, \boldsymbol{\gamma}^{obs} \sim \mathsf{Dirichlet}(a_{kk'}^{f0}(\mathcal{C}^{[t]}) + \mu_{kk'}^{f0}, \cdots, a_{kk'}^{fL_f}(\mathcal{C}^{[t]}) + \mu_{kk'}^{fL_f})$$

   and

$$\boldsymbol{u}_{kk'}^{f[t+1]} \mid \mathcal{C}^{[t]}, \boldsymbol{\gamma}^{obs} \sim \mathsf{Dirichlet}(b_{kk'}^{f0}(\mathcal{C}^{[t]}) + \nu_{kk'}^{f0}, \cdots, b_{kk'}^{fL_f}(\mathcal{C}^{[t]}) + \nu_{kk'}^{fL_f}).$$

   Call these samples $\Phi^{[t+1]}$.

2. We now sample the cluster assignment for each record $j \in [r]$ sequentially. Suppose we have sampled the first $j-1$ records, and are sampling the cluster assignment for record j from file $\boldsymbol{X}_k$. Let $\mathcal{C}_{-j}^{[t]}$ denote the current partition of $[r]$, without record j, after sampling the first $j-1$ records. Then we sample the cluster assignment for record j according to the following probabilities:

$$\mathbb{P}(j \to c \mid \mathcal{C}_{-j}^{[t]}, \Phi^{[t+1]}, \boldsymbol{\gamma}^{obs}) \propto$$
$$\begin{cases} \left[\dfrac{\mathbb{P}(n(\mathcal{C}_{-j}^{[t]}) + 1)}{\mathbb{P}(n(\mathcal{C}_{-j}^{[t]}))}\right] \left[\dfrac{(n(\mathcal{C}_{-j}^{[t]}) + 1)(n_{\mathbf{1}_k}(\mathcal{C}_{-j}^{[t]}) + \alpha_{\mathbf{1}_k})}{n(\mathcal{C}_{-j}^{[t]}) + \alpha_0}\right] p_k(1) & , \text{ if } c = (\emptyset, \cdots, \emptyset) \\ \left[\prod_{k'=1}^{K} \prod_{i \in c_{k'}} \mathcal{L}_{ij}^{[t+1]}\right] \left[\dfrac{n_{\mathbf{1}_{c \cup j}}(\mathcal{C}_{-j}^{[t]}) + \alpha_{\mathbf{1}_{c \cup j}}}{n_{\mathbf{1}_c}(\mathcal{C}_{-j}^{[t]}) + \alpha_{\mathbf{1}_c} - 1}\right] p_k(1) & , \text{ if } |c_k| = 0, c \neq (\emptyset, \cdots, \emptyset) \\ \left[\prod_{k'=1}^{K} \prod_{i \in c_{k'}} \mathcal{L}_{ij}^{[t+1]}\right] (|c_k| + 1) \left[\dfrac{p_k(|c_k| + 1)}{p_k(|c_k|)}\right] & , \text{ if } |c_k| > 0, \end{cases}$$

where, letting k' denote the file that record i is in,

$$\begin{aligned}\mathcal{L}_{ij}^{[t+1]} &= \prod_{f=1}^{F}\left[\prod_{l=0}^{L_f}\left(\frac{m_{kk'}^{fl[t+1]}}{u_{kk'}^{fl[t+1]}}\right)^{I(\gamma_{ij}^f=l)}\right]^{I_{obs}(\gamma_{ij}^f)} \\ &= \exp\left[\sum_{f=1}^{F} I_{obs}(\gamma_{ij}^f)\sum_{l=0}^{L_f}\log\left(\frac{m_{kk'}^{fl[t+1]}}{u_{kk'}^{fl[t+1]}}\right)I(\gamma_{ij}^f=l)\right].\end{aligned}$$

A.2.3 Computational Complexity

The computational complexity of posterior inference in our proposed approach can be broken up into the complexity of pre-computing comparison vectors, and the complexity of individual steps of the Gibbs sampler presented in Appendix A.2.2.

- The computational complexity of pre-computing comparison vectors is $\mathcal{O}(\texttt{rp}*F)$, where $\texttt{rp}$ is the number of valid record pairs. To be more specific, when we assume there are duplicates in every file, $\texttt{rp} = r(r-1)/2$, and when we assume there are no duplicates in each file, $\texttt{rp} = \sum_{k<k'} r_k r_{k'}$. For in between situations where we assume there are no duplicates in some files and duplicates in the remaining files, it can be shown that $\sum_{k<k'} r_k r_{k'} < \texttt{rp} < r(r-1)/2$. Thus in the most general case, pre-computing comparison vectors scales quadratically in the number of records. We discuss in the following section how the cost of this step can be reduced through the usage of blocking.

- The computational complexity of step 1 of the Gibbs sampler presented in Appendix A.2.2, i.e. sampling the $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters, is $\mathcal{O}(\texttt{rp}*\texttt{fl}+\texttt{fp}*\texttt{fl})$, where $\texttt{fp}$ is the number of valid file pairs, and $\texttt{fl} = \sum_{f=1}^{F}(L_f+1)$ is the total number of agreement levels across all fields. We have that $\texttt{fp} = \binom{K}{2} + K_d$, where K_d is the number of files that are assumed to have duplicates. This follows as the a and b summaries of the partition can be calculated from a matrix multiplication of a $\texttt{fl} \times \texttt{rp}$ matrix and a $\texttt{rp} \times 1$ matrix, and given these a and b summaries the complexity of sampling the $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters from their full conditionals is $\mathcal{O}(\texttt{fp}*\texttt{fl})$.

- The computational complexity of step 2 of the Gibbs sampler presented in Appendix A.2.2, i.e. sampling the partition $\mathcal{C}$, is difficult to analyze in general. In the most general case, where we assume there are duplicates in each file, in the worst case scenario, each record could be placed in its own cluster. The complexity of sampling the cluster assignment for a single record would then be $\mathcal{O}(r)$, and the complexity of sampling the cluster assignment for all records would then be $\mathcal{O}(r^2)$. However, the number of clusters potentially changes whenever a new cluster assignment is sampled, which complicates this analysis. Further, once introduces constraints on the partition space, either through assuming there are no duplicates in some files, or using indexing as described in the next section, the number of clusters available for a specific record's cluster assignment step will depend on these constraints. In general the best we can say is that this step will be faster when each record has on average (with respect to the posterior) a small number of clusters to which it can be assigned, and slower when each record has on average a large number of clusters to which it can be assigned.

In our current implementation of the proposed approach, we have found that even though both steps of the Gibbs sampler scale quadratically in the number of records in the worst case, the cost of sampling the partition generally dominates the cost of sampling the $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters, and is the main bottleneck of our approach. We note that the sampling of the partition will essentially have the same computational complexity regardless of whether one uses a comparison-based model for records, as we have proposed in Chapter 2, or one uses a direct-modeling approach, as in [126].

A.2.4 Blocking, Indexing, and Scalability

As described in the previous section, there are two main bottlenecks to scalability in our proposed approach: pre-computing the comparison vectors and sampling the partition from its full conditional in the Gibbs sampler presented in Appendix A.2.2. Both of these bottlenecks can be sped up through the use of indexing techniques, which declare certain pairs of

records non-coreferent a priori based on comparisons of a small number of fields [28, 127, 97]. This both reduces the number of record pairs under consideration, and reduces on average the number of clusters to which each record can be assigned in the Gibbs sampler presented in Appendix A.2.2.

In particular, if $\mathcal{P} = \cup_{k \leq k'} \mathcal{P}_{kk'}$ is the set of all possible record pairs, indexing techniques generate a set $\mathcal{P}^* \subset \mathcal{P}$, such that $|\mathcal{P}^*| \ll |\mathcal{P}|$, where record pairs in $\mathcal{P}^*$ are candidate coreferent pairs, and record pairs in $\mathcal{P} \backslash \mathcal{P}^*$ are fixed as non-coreferent. Thus when performing posterior inference, this truncates our prior on multifile partitions to the set $\{\mathcal{C} : \mathcal{C}(i) \neq \mathcal{C}(j), \forall (i,j) \in \mathcal{P} \setminus \mathcal{P}^*\}$.

We briefly review two common indexing techniques from [97], blocking and indexing by disjunction. Blocking declares pairs of records to be non-coreferent when they disagree on a set of error-free fields. The use of error-free fields guarantees that the candidate coreferent pairs output from blocking are transitive, so that $\mathcal{P}^*$ forms a partition of the records. Indexing by disjunction declares pairs of records to be non-coreferent when they disagree at a certain threshold for each field in a given set of reliable fields. Candidate coreferent pairs output from indexing by disjunction are not guaranteed to be transitive.

When a set of error-free fields are available, we recommend blocking. Blocking schemes can be implemented without constructing comparison vectors for each record pair, thus reducing the cost of pre-computing comparison vectors for all record pairs to just the cost of pre-computing comparison vectors for record pairs within each block. Our proposed approach can then be run independently in each block, drastically reducing on average the number of clusters to which each record can be assigned in the Gibbs sampler presented in Appendix A.2.2.

Within blocks, there is no further way to reduce cost of the pre-computing the comparison vectors. However, it is still possible to reduce the cost of sampling the partition from its full conditional in the Gibbs sampler presented in Appendix A.2.2 through the use of indexing by disjunction. By fixing certain pairs of records to be non-coreferent, one reduces on average the number of clusters to which each record can be assigned in the Gibbs sampler presented

in Appendix A.2.2. Note that the comparisons for record pairs fixed as non-coreferent in $\mathcal{P} \setminus \mathcal{P}^*$ still contribute to the model through the $b_{kk'}^{fl}$ term in the likelihood in (2.2), avoiding many of the issues presented in [97].

However, the non-transitivity of $\mathcal{P}^*$ output from indexing by disjunction can be problematic, as it suggests that the thresholds used in indexing by disjunction are too stringent, and that they may be excluding true coreferent pairs. Non-transitivity can also cause problems for Markov chain Monte Carlo samplers (like our Gibbs sampler in Appendix A.2.2), as it can make traversing the constrained space of multifile partitions difficult. To avoid the issue of non-transitivity, we propose to use the transitive closure of the candidate coreferent pairs, $\mathcal{P}^*$, generated by indexing by disjunction, which we refer to as *transitive indexing*. Transitive indexing has been used before in the post-hoc blocking methodology of [92] for two-file record linkage.

A.2.5 Initialization

Due to the nature of the Gibbs sampler in Appendix A.2.2, we can initialize the multifile partition $\mathcal{C}$ without needing to initialize Φ. A simple initialization for $\mathcal{C}$ is to let each record belong to its own cluster, which works well when indexing is used. However, we observed during some preliminary simulations that when sampling K-partite matchings without using indexing, the sampler can take a large number of iterations to mix if we initialize $\mathcal{C}$ in this way, where the number of iterations depends on the partition the data was simulated from. This problem can not be avoided as in Appendix A.2.4, as the constraints on the space of K-partite matchings cannot be relaxed. In this case, we constructed a simple alternative initialization. The idea is to use an indexing scheme for initialization, even if indexing is not being used to reduce the number of candidate coreferent pairs. In particular, we first generate a set of record pairs $\mathcal{P}^* \subset \mathcal{P}$ through transitive indexing as described in Appendix A.2.4. For each block of records in $\mathcal{P}^*$, we sample a random K-way matching of records in that block. We then initialize $\mathcal{C}$ such that each record belongs to its own cluster, except for the sampled K-way matchings.

A.3 Point Estimation Appendix

In this appendix we discuss how our proposed loss function differs from the loss function of [115], and propose a strategy for approximating the Bayes estimate under our proposed loss function.

A.3.1 Comparison to [115]

Unlike our loss function construction, in the two-file set-up [115] constructed the loss function from individual losses for the records in the smaller datafile only. Such construction however leads to an asymmetry in the loss function that is arbitrary. Consider an example of two datafiles, where the first file has records a and b, and the second file has records c and d. In that case the role of the datafiles can be arbitrarily interchanged. If the true matching has a link between a and c but the matching estimate has a link between b and c, the loss will be $\lambda_{\text{FNM}} + \lambda_{\text{FM1}}$ if file two is chosen not to contribute to the loss, but it will be λ_{FM2} if file one is chosen not to contribute to the loss. Our new construction presented in Chapter 2 does not lead to such issues.

A.3.2 Approximating the Bayes Estimate

Finding a partition $\hat{\boldsymbol{Z}}$ such that $\mathbb{R}(\hat{\boldsymbol{Z}})$ is minimized corresponds to an optimization problem closely related to graph partitioning problems [e.g., 22, 72, 99] or correlation clustering [e.g., 12, 33], both of which are known to be NP-complete. Thus, unlike in [115], we cannot minimize $\mathbb{R}(\hat{\boldsymbol{Z}})$ exactly in general. All approaches for graph partitioning problems or for correlation clustering instead rely on heuristic algorithms whose performance is evaluated empirically via simulation studies and benchmark datasets. We will follow a similar approach.

We take advantage of the fact that in practice a large number of record pairs will have zero or close to zero posterior probability of matching $\mathbb{P}(\Delta_{ij} = 1 \mid \boldsymbol{\gamma}^{obs})$. Based on this, we propose to threshold $\mathbb{P}(\Delta_{ij} = 1 \mid \boldsymbol{\gamma}^{obs})$ at a small value δ to create a graph where an edge represents a non-negligible probability of matching between two records. We then break the records

up into connected components of this graph, each component representing groups of records that are more likely to be coreferent. We then find the Bayes estimate by minimizing $\mathbb{R}(\hat{\boldsymbol{Z}})$ separately within each of these connected components. We can think of δ as a way of trading-off between accuracy of the Bayes estimate and computational tractability: larger values of δ decrease the size of the resulting connected components, making the minimization more tractable within each component, while smaller δ make the resulting approximation more accurate as using the threshold $\delta = 0$ is no longer an approximation. We recommend setting δ as the smallest probability such that the largest connected component is smaller than some pre-specified upper bound that captures a computational budget. To minimize $\mathbb{R}(\hat{\boldsymbol{Z}})$ within the connected components, we propose to do so over posterior samples, $\{\boldsymbol{Z}^{[t]}\}_{t=1}^{T}$, and find the sample which minimizes $\mathbb{R}(\boldsymbol{Z}^{[t]})$. As this minimization is happening separately within each connected component, the final Bayes estimate of the partition of all r records does not itself have to be a posterior sample.

To minimize $\mathbb{R}(\hat{\boldsymbol{Z}})$ when searching for partial estimates, let Ω denote the power set of $[r]$, and let $\boldsymbol{Z}_{\omega}^{[t]}$ denote the posterior draw $\boldsymbol{Z}^{[t]}$ where the records in $\omega \in \Omega$ are set to abstain, A. Then in order to accommodate partial estimates, we can minimize $\mathbb{R}(\hat{\boldsymbol{Z}})$ over $\{\boldsymbol{Z}_{\omega}^{[t]} \mid t \in [T], \omega \in \Omega\}$.

Unless stated otherwise, in all simulations and in the application, for full (partial) estimates, we find the Bayes estimate separately within connected components of records with posterior probability larger than δ of matching, where δ is the smallest probability such that the largest connected component is smaller than 50 (12).

A.4 Simulation Appendix

A.4.1 Tables 3 and 4 from [114]

Table 3 of [114] is reproduced in Table A.1. In this table, edit errors are insertions, deletions, or substitutions of characters in a string, OCR errors are optical character recognition errors, keyboard errors are typing errors that rely on a certain keyboard layout, and phonetic errors are errors using a list of predefined phonetic rules. Table 4 of [114] is reproduced in Table A.2.

Table A.1: Types of errors per field in the simulation studies.

	Type of error					
Field	**Missing values**	**Edits**	**OCR**	**Keyboard**	**Phonetic**	**Misspelling**
Given name		✓	✓	✓	✓	
Family name		✓	✓	✓	✓	✓
Age interval	✓					
Sex	✓					
Occupation	✓					
Phone number	✓	✓	✓	✓		
Postal code	✓	✓	✓	✓		

A.4.2 Prior Sensitivity Analysis for Simulation with Duplicate-Free Files, Equal Errors Across Files

In Section 6.2 of Chapter 2, we saw that our proposed approach struggled in the no-three-file overlap setting when there was high measurement error. In practice, if we knew that no entity is represented in all three datafiles we could enforce that restriction just like we enforce that there are no duplicates in given files, which would likely lead to better performance. While it is reasonable to assume in some applications that there are no duplicates in a given

Table A.2: Construction of levels of disagreement for the simulation studies.

		Levels of disagreement			
Field	**Similarity measure**	**0**	**1**	**2**	**3**
Given name	Levenshtein	0	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 1]$
Family name	Levenshtein	0	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 1]$
Age interval	Binary comparison	Agree	Disagree		
Sex	Binary comparison	Agree	Disagree		
Occupation	Binary comparison	Agree	Disagree		
Phone number	Levenshtein	0	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 1]$
Postal code	Levenshtein	0	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 1]$

file (for example the application considered in Section 7 of Chapter 2), it is less reasonable to assume with absolute certainty that there is no entity represented in all three datafiles. Thus we want to instead incorporate the weaker information that there is a low amount of three way overlap. We can achieve this through an informative specification of $\boldsymbol{\alpha}$.

In the no-three-file overlap setting, the overlap table was generated from a multinomial distribution with probability vector $\boldsymbol{p} = (p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111})$ where $p_{001} = p_{010} = p_{100} = 0.55/3, p_{011} = p_{101} = p_{110} = 0.15, p_{111} = 0$. Note that our Dirichlet-multinomial prior can be motivated as the result of first drawing $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ from a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}$, then drawing $\boldsymbol{n}$ from a multinomial distribution of size n with probabilities $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$. As discussed in Appendix A.1.2, $\boldsymbol{\alpha}$ can be interpreted as prior cell counts, which can be used to incorporate prior information about the amount of overlap between datafiles. Thus when specifying an informative $\boldsymbol{\alpha}$, we want the Dirichlet prior for $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ to be centered roughly around $\boldsymbol{p}$. We can accomplish this by setting $\boldsymbol{\alpha} = \kappa \times (p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, 1/\kappa)$. $\kappa + 1$ represents the sum of prior cell counts. As κ increases, the Dirichlet prior for $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ becomes more concentrated near $\boldsymbol{p}$.

We repeated the no-three-file overlap setting simulation from Section 6.2 of Chapter 2

using this informative specification with $\kappa \in \{49, 99\}$. The results are presented in Figure A.1. We see that when there is high measurement error, these more informative specifications improve upon the performance of the default specification of $\boldsymbol{\alpha} = (1, \cdots, 1)$.

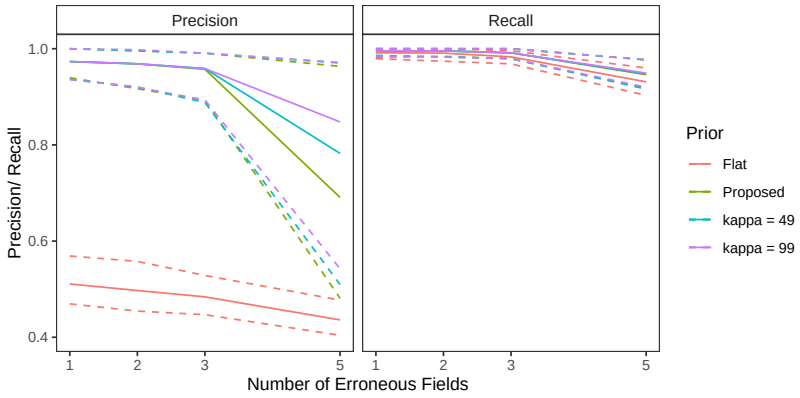


Figure A.1: Performance comparison for no-three-file overlap simulation with more informative settings of $\boldsymbol{\alpha}$. Solid lines show medians, and dashed lines show 2nd and 98th percentiles. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ with $\kappa \in \{49, 99\}$.

A.4.3 Files with Duplicates, Full Estimates

This simulation study consists of linkage and duplicate detection for three datafiles, so that the target of inference is a general multifile partition. We conduct this study with probabilities fixed at $p_{001} = p_{010} = p_{100} = 0.3, p_{011} = p_{101} = p_{110} = 0.025, p_{111} = 0.025$, representing a very low overlap setting, which can be challenging as seen in Section 6.2 of Chapter 2. For each entity represented in the datafiles we generated a within-file cluster size from a Poisson distribution with mean λ truncated to $\{1, \cdots, 5\}$. In this study, in addition to varying the number of erroneous fields per record over $\{1, 2, 3, 5\}$ to explore different amounts of measurement error, and we vary λ over $\{0.1, 1, 2\}$ to explore low, medium, and high amounts of duplication.

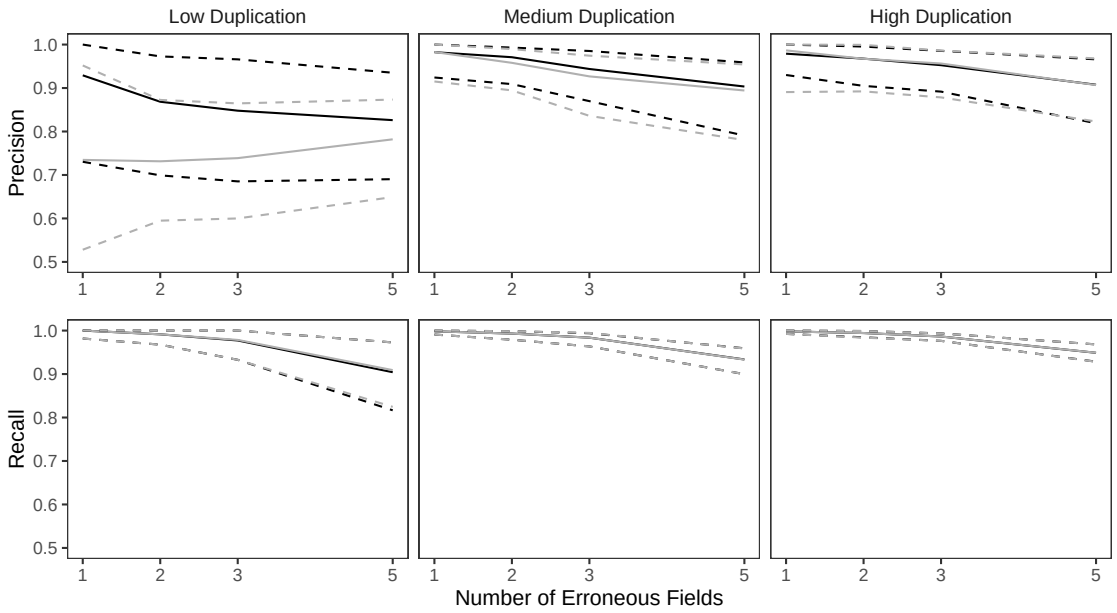


Figure A.2: Performance comparison for simulation with datafiles with duplicates and full estimates. Black lines refer to results under our structured prior, grey lines refer to the approach of [114], solid lines show medians, and dashed lines show 2nd and 98th percentiles.

To implement our methodology, in addition to the general set-up described in Section 6.1 of Chapter 2, we use a Poisson prior with mean 1 truncated to $\{1, \cdots, 10\}$ on the within-file cluster sizes. For comparison we use the comparison-based model of [114], which treats all of the records as coming from one file and uses a flat prior on partitions. For both models we use transitive indexing as in described Appendix A.2.4 to reduce the number of comparisons, where the initial indexing scheme declares record pairs as non-coreferent if they disagree in either given or family name at the highest level (according to Table 4 of [114]).

The results of the simulation are seen in Figure A.2. In the medium and high duplication settings, the models have similarly good performance. We believe that the similar performance between models in these settings is due to the use of the indexing, which significantly reduces the size of the space of possible multifile partitions, so that the influence of the structured prior is minimized. However, in the low duplication setting we see that the precision of

the proposed model is better across the varying measurement error settings than the model of [114]. This suggests that in low duplication settings, our approach once again improves upon an approach that uses flat priors for partitions by protecting against over-matching.

We now explore the sensitivity of our approach to changes in the prior for the number within-file cluster sizes, and demonstrate how the performance in the low duplication setting can be further improved through the incorporation of an informative prior for the within-file cluster sizes. In the simulation that was just described, we used a Poisson prior with mean $\lambda = 1$ truncated to $\{1, \cdots, 10\}$ for the within-file cluster sizes. In the simulation, the within-file cluster sizes were generated from Poisson distributions with mean λ truncated to $\{1, \cdots, 5\}$, where λ varied over $\{0.1, 1, 2\}$. We now repeat the same simulation using Poisson priors with mean λ truncated to $\{1, \cdots, 10\}$ for the within-file cluster sizes, where $\lambda \in \{0.1, 1, 2\}$. The results are presented in Figure A.3. The results for the medium and high duplication settings are very robust to the within-file cluster size prior specification. In the low duplication setting we see that the performance among the different within-file cluster size prior specifications is best when $\lambda = 0.1$, and worst when $\lambda = 2$ (but still better than the approach of [114]). This behavior is expected as the within-file cluster size prior with $\lambda = 0.1$ is informative in the low duplication setting.

A.4.4 Files with Duplicates, Partial Estimates

We now examine the performance of partial estimates in the low duplication setting of the simulation presented in the previous section, where both the proposed approach and the approach of [114] struggled the most. For partial estimates, we use $\lambda_{\text{FNM}} = \lambda_{\text{FM1}} = 1$, $\lambda_{\text{FM2}} = 2$, and $\lambda_A = 0.1$, so that abstaining from making a linkage decision is 10% as costly as making a false non-match. We will assess the performance of the Bayes estimate using precision and the *abstention rate*, $\sum_{i=1}^{r} I(\hat{Z}_i = A)/r$, the proportion of records which the Bayes estimate abstained from making a linkage decision. Recall is no longer useful when using partial estimates, as we are not trying to find all true matches.

In Figure A.4 we see that, for both approaches, using partial estimates leads to improved

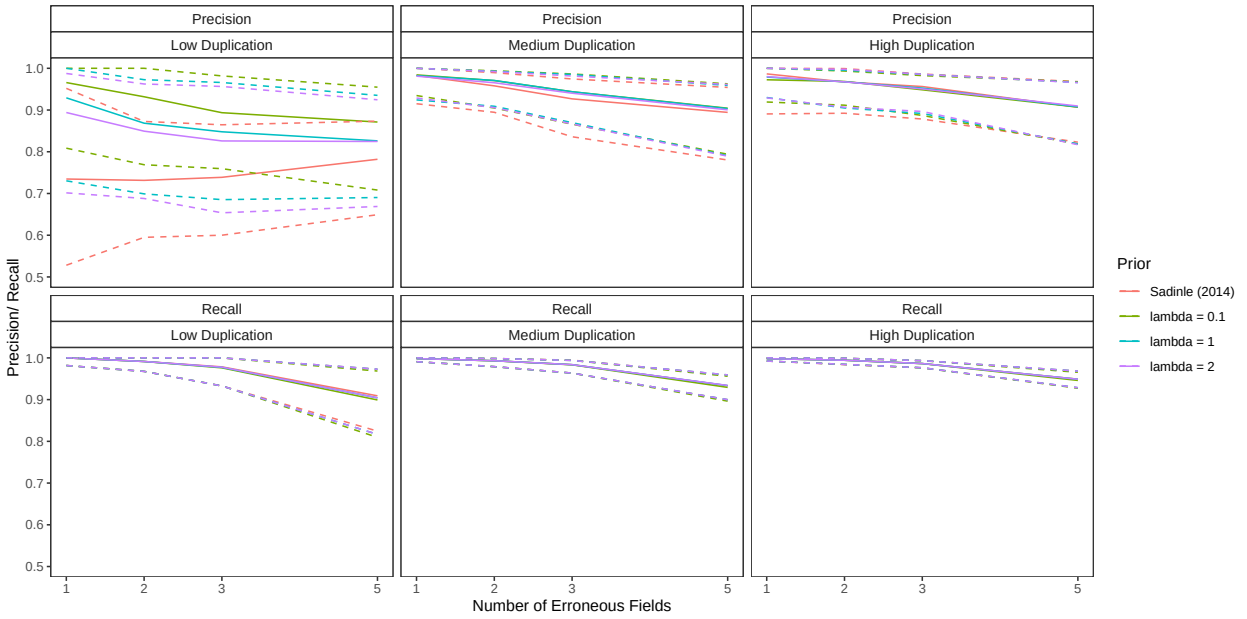


Figure A.3: Performance comparison for simulation with datafiles with duplicates and full estimates, with varying priors for the within-file cluster sizes. Solid lines show medians, and dashed lines show 2nd and 98th percentiles.

precision in comparison with full estimates, while maintaining a relatively low abstention rate. This result is expected, as the records to which our partial estimate assigns the abstain option are the most ambiguous in terms of which records they should be linked to, and therefore they are the most likely to lead to false matches which decrease the precision. Using partial estimates with an abstain option is therefore a good way of compromising between automated and manual linkage: records for which linkage decisions are difficult are left to be handled via clerical review.

A.4.5 Results for an Alternative Metric

When evaluating the performance of the full estimates in the simulations thus far, we have focused on the metrics of precision and recall, which are global measures of how well the true partition is being estimated. One could also be interested in how well other summaries of

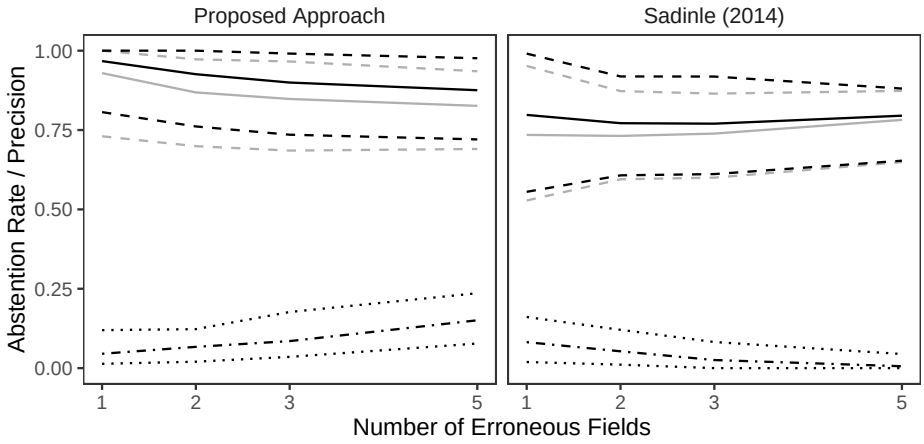


Figure A.4: Performance comparison for simulation with datafiles with duplicates and partial estimates. Black solid and dashed lines refer to precision for partial estimates, grey solid and dashed lines refer to precision for full estimates, and dot-dashed and dotted lines refer to the abstention rate for partial estimates. Solid and dot-dashed lines show medians, and dashed and dotted lines show 2nd and 98th percentiles.

the partition are being estimated, e.g. the number of entities (i.e. the number of clusters), the sizes of the clusters, the overlap table, etc. In this section we report the performance of the full estimates in the previous simulations when estimating the number of entities.

For each replicate data set in each simulation scenario considered thus far, we obtained a full estimate of the partition, which can be used to derive an estimate of the number of entities. For a given simulation scenario, let n_0 denote the true number of entities (in all the scenarios considered thus far, $n_0 = 500$), and let $\hat{n}_s$ denote the estimate of the number of entities based on the full estimate of the partition for replicate data set $s \in \{1, \cdots, 100\}$. For each simulation scenario, we can thus estimate the bias of these estimates, $\sum_{s=1}^{100} \hat{n}_s - n_0$, and the mean-squared error of these estimates, $\sum_{s=1}^{100} (\hat{n}_s - n_0)^2$.

Duplicate-Free Files, Equal Errors Across Files

The results for estimating the number of latent entities in the simulations conducted in Section 6.2 of Chapter 2 and Appendix A.4.2 are seen in Figures A.5 and A.6. We see that that across the different simulation settings the proposed approach has a slight negative bias, and the approach using a flat prior has a very large negative bias. In the no-three-file overlap settings, we see that the more informative prior specifications are less biased than the proposed approach when there are a larger number of erroneous fields, which mirrors the results from Appendix A.4.2. Across all approaches, the bias increases as the number of erroneous fields increases. The results for the mean-squared error estimates are very similar to the results for the bias estimates.

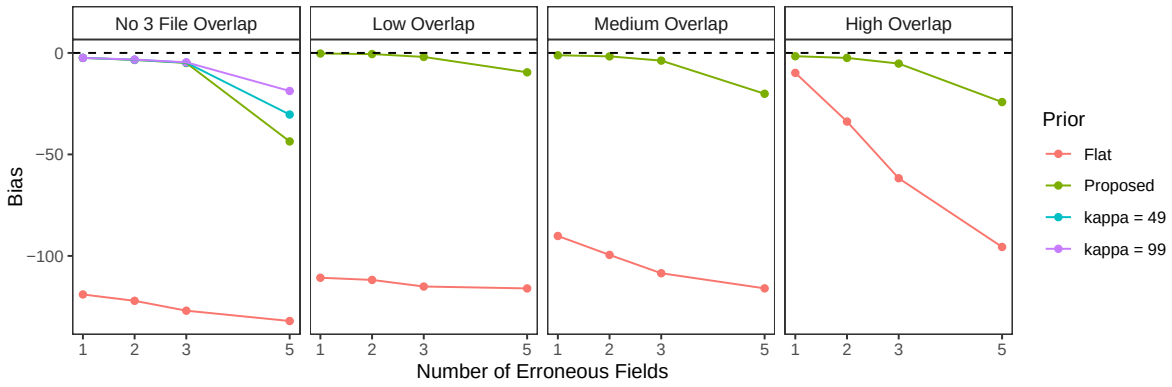


Figure A.5: Bias estimates for simulation with duplicate-free files and equal errors across files. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ discussed in Appendix A.4.2.

Duplicate-Free Files, Unequal Errors Across Files

The results for estimating the number of latent entities in the simulation conducted in Section 6.3 of Chapter 2 are seen in Figures A.5 and A.6. We see that that across the two simulation

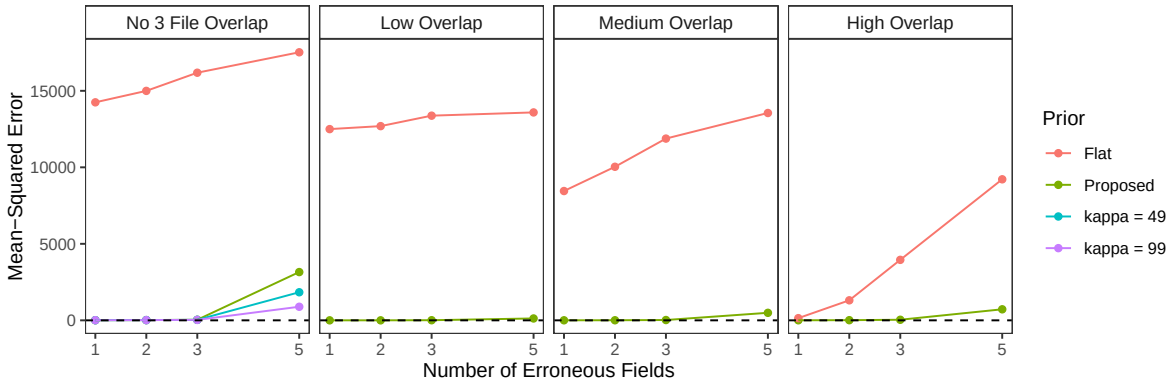


Figure A.6: Mean-squared error estimates for simulation with duplicate-free files and equal errors across files. "Flat" refers to a flat prior on tripartite matchings, "Proposed" refers to our structured prior for partitions when $\boldsymbol{\alpha} = (1, \cdots, 1)$, and "kappa = 49" and "kappa = 99" refer to the more informative specifications of $\boldsymbol{\alpha}$ discussed in Appendix A.4.2.

settings all approaches have a negative bias, with the proposed approach having the smallest bias and the approach using a flat prior having the largest bias in both scenarios. The results for the mean-squared error estimates are very similar to the results for the bias estimates.

Files with Duplicates, Full Estimates

The results for estimating the number of latent entities in the simulation conducted in Appendix A.4.3 are seen in Figures A.9 and A.10. In the low duplication setting, we see that the proposed approach with $\lambda = 0.1$ has the smallest bias across error settings, followed by the proposed approach with $\lambda = 1$, then the proposed approach with $\lambda = 2$, and then the approach of [114] with the largest bias across error settings. This mirrors the results from Appendix A.4.3. In the medium and high duplication settings we see that all approaches have a slight negative bias when there are a small number of erroneous fields, and a slight positive bias when there are a large number of erroneous fields. The different variants of the proposed approach all have similar performance in these settings. The proposed approach performs best when there are a small number of erroneous fields, and the approach of [114] performs

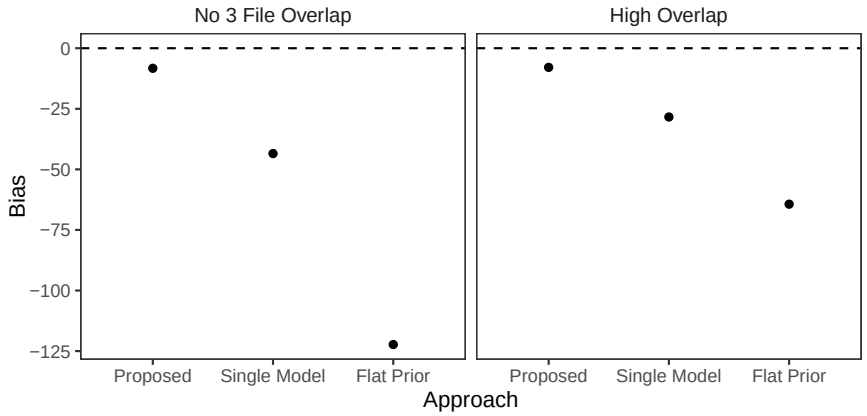


Figure A.7: Bias estimates for simulation with duplicate-free files and unequal errors across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings.

best when there are a large number of erroneous fields. The results for the mean-squared error estimates are very similar to the results for the bias estimates.

A.4.6 Simulation Running Times

In this section we present the average running time of our proposed Gibbs sampler across the various simulations settings (i.e. the average time it takes to draw 1000 samples for each simulation setting). The running times are based on the implementation in the R package multilink provided in the supplementary materials, with the Gibbs sampler described in Appendix A.2.2 written in C++, running on a laptop with a 3.1 GHz processor. We are working on improving this implementation, and the current version can be downloaded on GitHub at the following link: `https://github.com/aleshing/multilink`.

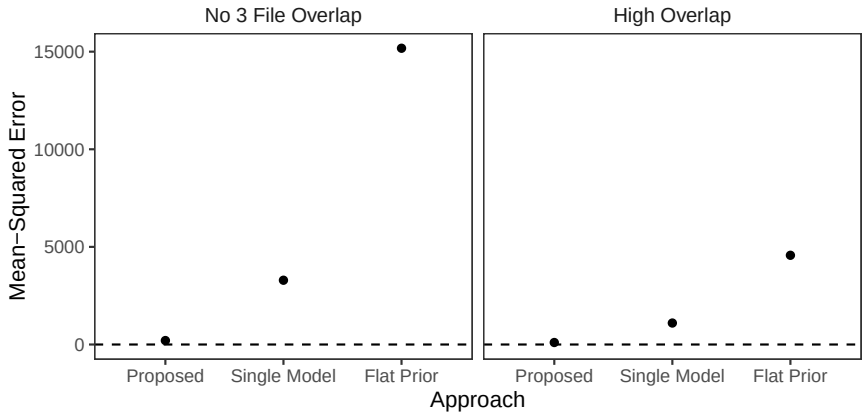


Figure A.8: Mean-squared error estimates for simulation with duplicate-free files and unequal errors across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings.

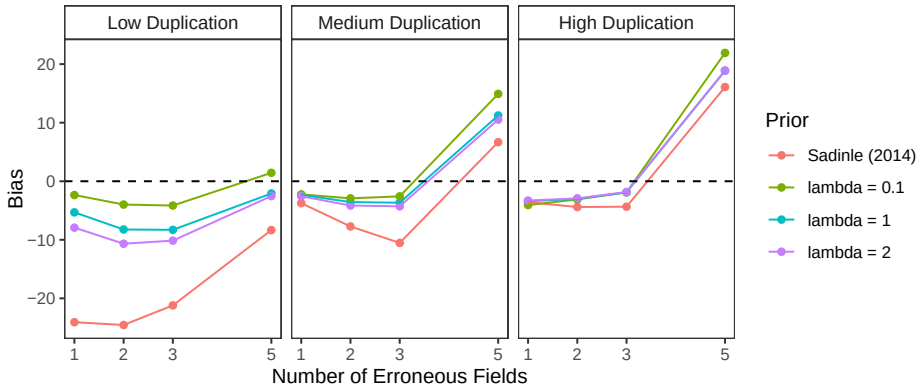


Figure A.9: Bias estimates for simulation with files with duplicates and equal errors across files. "Sadinle (2014)" refers to the approach of [114] and "lambda=..." refers to the proposed approach, varying the prior over within-file cluster sizes.

Duplicate-Free Files, Equal Errors Across Files

The average running times of our approach in the simulations conducted in Section 6.2 of Chapter 2 are presented in Table A.3. The average number of records was 725 in the

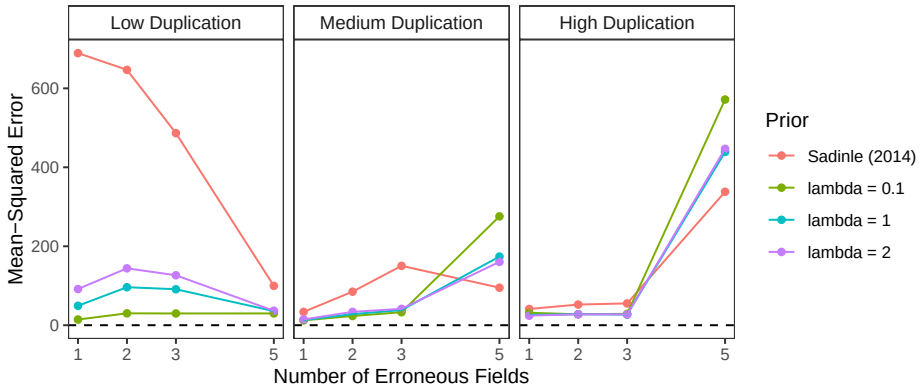


Figure A.10: Mean-squared error estimates for simulation with files with duplicates and equal errors across files. "Sadinle (2014)" refers to the approach of [114] and "lambda=..." refers to the proposed approach, varying the prior over within-file cluster sizes.

settings with no-three-file overlap, 676 in the settings with low overlap, 725 in the settings with medium overlap, and 875 in the settings with high overlap.

Table A.3: Average running time in seconds for proposed approach in simulations with duplicate-free files and equal errors across files.

Number of Erroneous Fields	No 3 File Overlap	Low Overlap	Medium Overlap	High Overlap
1	111.0	77.6	83.7	108.6
2	112.9	77.7	83.7	109.2
3	111.8	77.4	83.0	109.4
5	95.7	73.8	77.3	101.1

Duplicate-Free Files, Unequal Errors Across Files

The average running time of our proposed approach in the simulations conducted in Section 6.3 of Chapter 2 was 121.8 seconds in the no-three-file overlap setting and 104.1 seconds in the high overlap setting. The average number of records was 725 in the settings with

no-three-file overlap and 875 in the settings with high overlap.

Files with Duplicates, Full Estimates

The average running times of our approach in the simulations conducted in Appendix A.4.3 are presented in Table A.4. The average number of records was 590 in the settings with low duplication, and 889 in the settings with medium duplication, and 1260 in the settings with high duplication. We note here that in this simulation, compared to the simulations with duplicate-free files, we used indexing, which sped up the running time.

Table A.4: Average running time in seconds for proposed approach in with files with duplicates and equal errors across files.

Number of Erroneous Fields	Low Duplication	Medium Duplication	High Duplication
1	1.5	8.0	20.9
2	1.1	7.7	20.8
3	1.0	7.4	21.0
5	0.9	6.6	19.2

A.4.7 Larger Sample Size Simulation

All of the simulations thus far had fixed the true number of latent entities, n, to 500. To explore the running time of our proposed approach further, we now present an additional set of simulations where the number of latent entities varies over $\{100, 500, 1000, 2500\}$. For concreteness we will focus on the simulation setting with duplicate-free files, equal errors across files, medium overlap, and 1 erroneous field per record (i.e. one of the settings from the simulation conducted in Section 6.2 of Chapter 2). For this chosen simulation setting, we repeat the simulation as conducted in Section 6.2 of Chapter 2, varying the number of latent entities over $\{100, 500, 1000, 2500\}$. For the $n = 2500$ setting we conduct 25, rather than

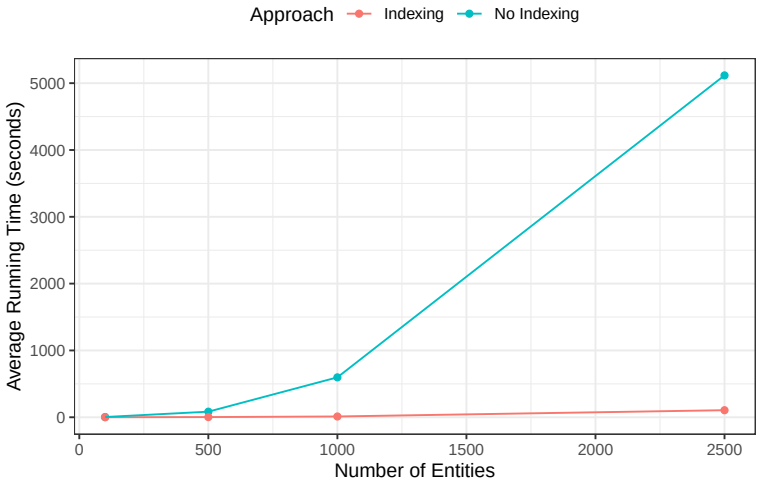


Figure A.11: Average running time for simulation varying the number of latent entities.

100, replications (due to how long each replication takes). The average number of records was 146 when $n = 100$, 725 when $n = 500$, 1452 when $n = 1000$, and 3629 when $n = 2500$. In addition to fitting the proposed approach without indexing as in Section 6.2 of Chapter 2, we additionally fit the proposed approach using the indexing scheme described in Appendix A.4.3 to demonstrate the utility of indexing for improving the running time or our proposed approach, as discussed in Appendix A.2.4.

The average running time for each setting of the number of entities, n, is presented in Figure A.11. Our proposed approach without indexing runs relatively quickly in the settings with $n \in \{100, 500, 1000\}$, for example it only takes around 10 minutes on average to draw 1000 samples when $n = 1000$. However, when $n = 2500$ our proposed approach without indexing runs takes roughly an hour and a half on average to draw 1000 samples. While this running time is manageable, it indicates that the running time in settings with more records than this would prohibitively slow. When looking at the running time of our proposed approach using indexing, we see that the average running is reduced drastically. For example, when $n = 2500$, the average running time is only around 100 seconds.

These results suggest that our approach without indexing can be run in a manageable

amount of time for thousands of records, but would be prohibitively slow when the number of records is much larger than a few thousand. Our proposed approach with indexing however can scale to larger file sizes, potentially in the tens of thousands. Using our approach, with or without indexing, in conjunction with blocking, as suggested in Appendix A.2.4, can help to further scale our approach to large file sizes.

We note that the precision and recall in these simulations, across the different settings of the number of entities, were comparable to the results for the medium overlap, 1 erroneous field per record setting from the simulation results in Section 6.2 of Chapter 2.

A.4.8 Alternative Loss Function Specifications

In this section explore the impact of varying the specification of the losses $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}, \lambda_{\text{FM2}}$, and λ_A on the performance of our proposed approach across the different simulation settings. For full estimates, we follow [115] and consider the following specifications of the losses $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}$, and λ_{FM2} (with $\lambda_A = \infty$).

A) $\lambda_{\text{FNM}} = 1, \lambda_{\text{FM1}} = 1, \lambda_{\text{FM2}} = 2$. This is the specification used in all of the simulations thus far.

B) $\lambda_{\text{FNM}} = \lambda_{\text{FM1}} = \lambda_{\text{FM2}} = 1$. Compared to specification A, this specification does not penalize type 2 false matches as heavily.

C) $\lambda_{\text{FNM}} = 4, \lambda_{\text{FM1}} = \lambda_{\text{FM2}} = 1$. This specification penalizes false non-matches more heavily than false matches.

D) $\lambda_{\text{FNM}} = 1, \lambda_{\text{FM1}} = 3, \lambda_{\text{FM2}} = 5$. This specification penalizes false matches more heavily than false non-matches, and type 2 false matches more heavily than type 1 false matches.

E) $\lambda_{\text{FNM}} = 1, \lambda_{\text{FM1}} = 2, \lambda_{\text{FM2}} = 3$. Compared to specification E, this specification does not penalizes false matches as heavily.

F) $\lambda_{\text{FNM}} = \lambda_{\text{FM1}} = 1, \lambda_{\text{FM2}} = 4$. Compared to specification A, this specification penalizes type 2 false matches more heavily.

For partial estimates, we consider combining $\lambda_A \in \{0.05, 0.1, 0.25\}$ with the six specifications of $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}$, and λ_{FM2} that we have just introduced.

Duplicate-Free Files, Equal Errors Across Files

The performance of our proposed approach in the simulations conducted in Section 6.2 of Chapter 2, using the different loss function specifications, are seen in Figure A.12. So that the figure is easier to scrutinize, we do not plot the results under specifications E and F, as the results under specifications E are very similar to the results under specification D, and the results under specifications F are very similar to the results under specifications A and B.

We see that when there are a low number of erroneous fields, the results are fairly robust to the loss function specification. When there are a high number of erroneous fields, we see that specification D leads to the highest precision and the lowest recall, specification C leads to the lowest precision and the highest recall, and specifications A and B are somewhere in between. These results are expected, as specification C penalizes false non-matches much more than false matches, and will thus decide to match records more often than the other specifications, and specification D penalizes false matches much more than false non-matches, and thus will decide to match records less often than the other specifications.

Duplicate-Free Files, Unequal Errors Across Files

The performance of our proposed approach in the simulations conducted in Section 6.3 of Chapter 2, using the different loss function specifications, are seen in Figure A.13. Overall the results are fairly robust to the loss function specification. The results under specifications A, B and F are very similar and the results under specifications D and E are very similar. Similar to the last section, specifications D and E lead to the highest precision and the lowest

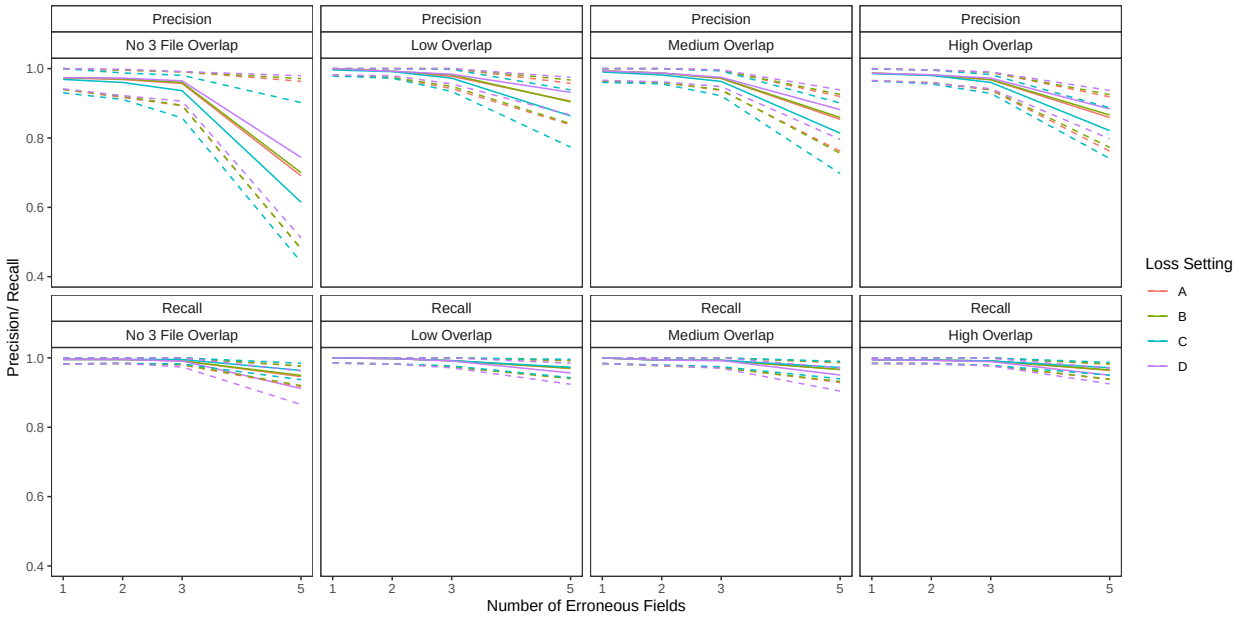


Figure A.12: Performance comparison across different loss function specifications for simulation with duplicate-free files and equal measurement error across files. Solid lines show medians, and dashed lines show 2nd and 98th percentiles.

recall, specification C leads to the lowest precision and the highest recall, and specifications A, B, and F are somewhere in between.

Files with Duplicates, Full Estimates

The performance of our proposed approach in the simulations conducted in Appendix A.4.3, using the different loss function specifications, are seen in Figure A.12. So that the figure is easier to scrutinize, we do not plot the results under specifications E and F, as the results under specifications E are very similar to the results under specification D, and the results under specifications F are very similar to the results under specifications A and B.

We see that when there is medium and high duplication, the results are fairly robust to the loss function specification. When there is low duplication, we see that specification D leads to the highest precision and the lowest recall, specification C leads to the lowest precision

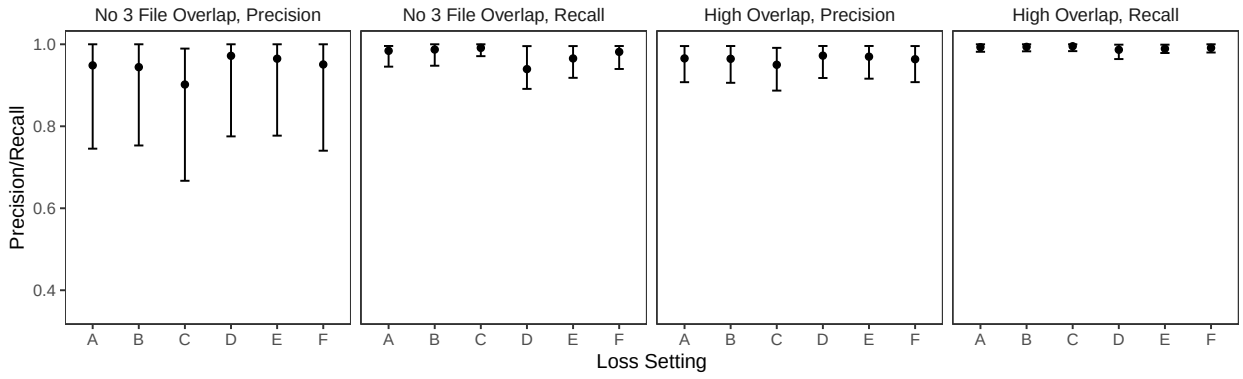


Figure A.13: Performance comparison across different loss function specifications for simulation with duplicate-free files and unequal measurement error across files. Solid lines show medians, and dashed lines show 2nd and 98th percentiles.

and the highest recall, and specifications A and B are somewhere in between. These results are expected, as described in the previous sections.

Files with Duplicates, Partial Estimates

The performance of our proposed approach in the simulations conducted in Appendix A.4.4, using the different loss function specifications, are seen in Figure A.15. So that the figure is easier to scrutinize, we do not plot the results under specifications F, as the results under specifications F are very similar to the results under specifications A and B.

We see that as we increase λ_A, the abstention rate decreases and the precision decreases across the different specifications of $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}$, and λ_{FM2}. Further, for each loss specification the abstention rate increases as the number of erroneous fields increases, as there is more uncertainty in the linkage. For a given setting of λ_A, the precision is fairly robust to the different specifications of $\lambda_{\text{FNM}}, \lambda_{\text{FM1}}$, and λ_{FM2}, with specifications D and E having slightly higher precision that specifications A, B and C. For a given setting of λ_A, the abstention rate is highest under specifications D and E, lowest under specifications A and B, with specification C somewhere in between.

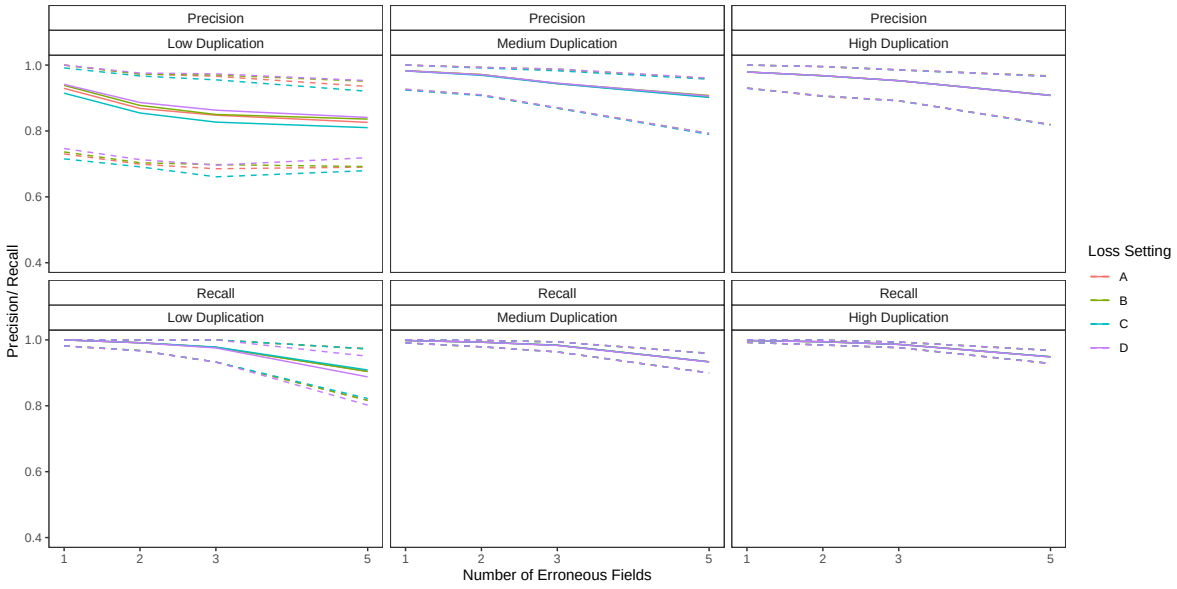


Figure A.14: Performance comparison across different loss function specifications for simulation with files with duplicates and equal measurement error across files, when using full estimates. Solid lines show medians, and dashed lines show 2nd and 98th percentiles.

A.4.9 Convergence Diagnostics

For all simulations we ran the Gibbs sampler described in Appendix A.2.2 for $1,000$ iterations, discarding the first 100 samples as burn-in. We initially came up with these sampling and burn-in lengths based on a small number of test runs for each simulation scenario. In particular, for each simulation scenario we ran a small number of test runs and examined the trace plots for the number of entities, n. As the chains for n appeared to converge quickly based on these trace plots, we determined that a burn-in period of 100 samples was appropriate. To illustrate this procedure, for each simulation scenario we now present trace plots for the number of entities, n, for a small number of runs.

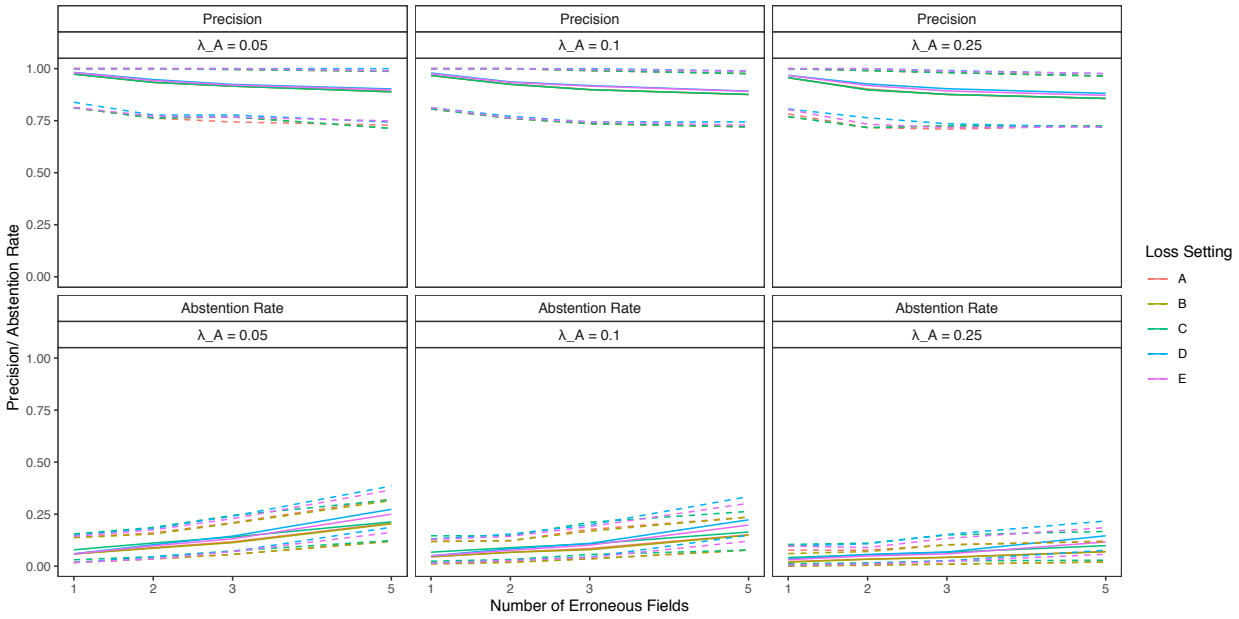


Figure A.15: Performance comparison across different loss function specifications for simulation with files with duplicates and equal measurement error across files, when using partial estimates. Solid lines show medians, and dashed lines show 2nd and 98th percentiles.

Duplicate-Free Files, Equal Errors Across Files

For the simulations conducted in Section 6.2 of Chapter 2 with 3 erroneous fields per record, for each overlap setting we present the trace plots for n for the last 5 of the 100 simulation runs in Figure A.16. The chains for the other scenarios with $1, 2$, and 5 erroneous fields per record converged similarly quickly.

Duplicate-Free Files, Unequal Errors Across Files

For the simulations conducted in Section 6.3 of Chapter 2, for each overlap setting we present the trace plots for n for the last 5 of the 100 simulation runs in Figure A.17.

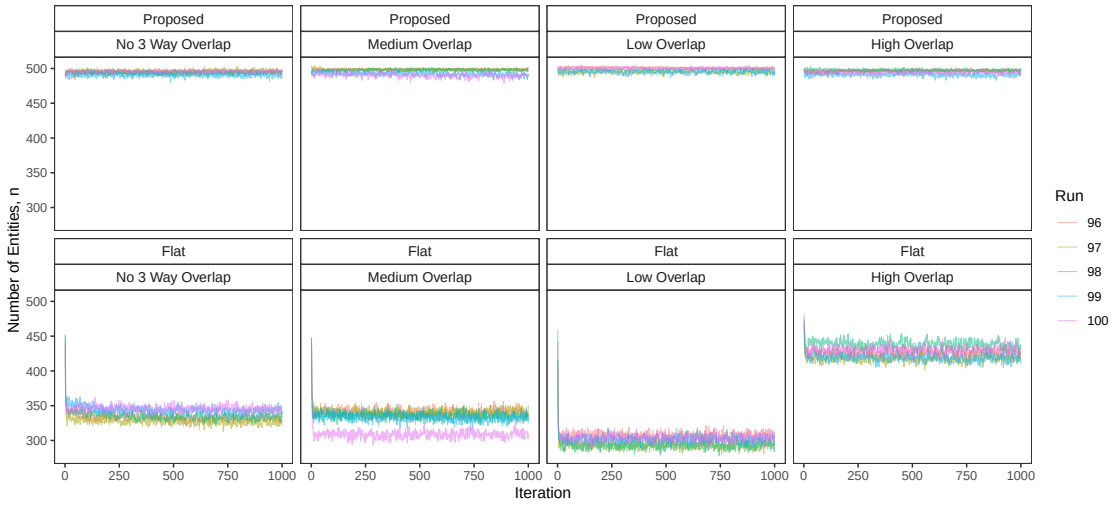


Figure A.16: Trace plots for n for the last 5 of 100 runs for the simulation with duplicate-free files and equal measurement error across files. "Proposed" refers to the proposed approach and "Flat" refers to the approach using a flat prior for tripartite matchings.

Files with Duplicates, Full Estimates

For the simulations conducted in Appendix A.4.3 with 3 erroneous fields per record, for each overlap setting we present the trace plots for n for the last 5 of the 100 simulation runs in Figure A.18. The chains for the other scenarios with $1, 2$, and 5 erroneous fields per record converged similarly quickly.

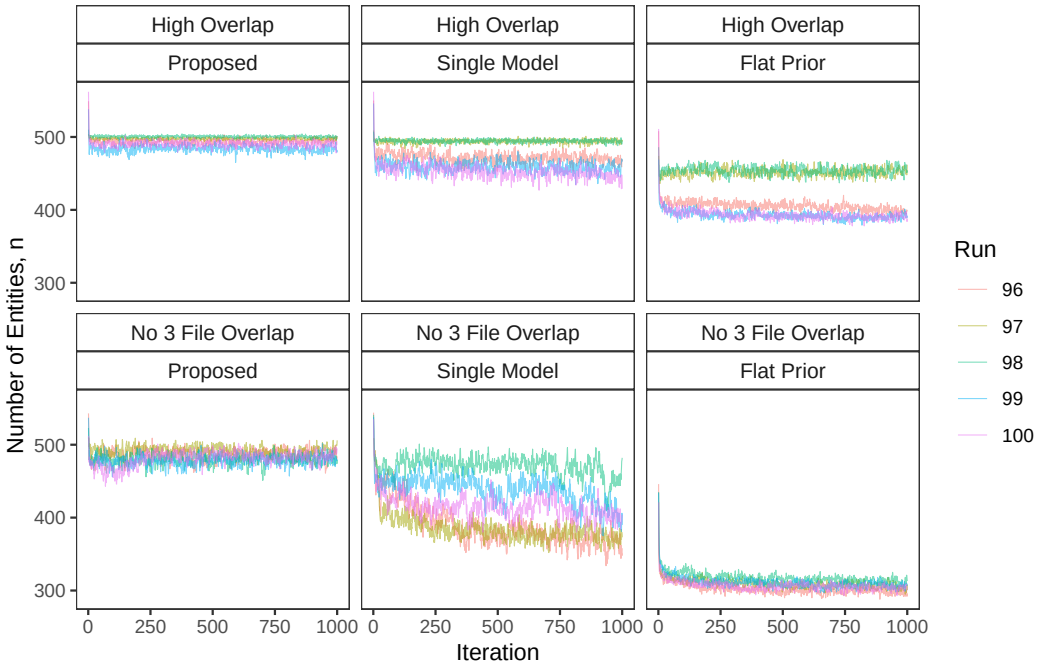


Figure A.17: Trace plots for n for the last 5 of 100 runs for the simulation with duplicate-free files and unequal measurement error across files. "Proposed" refers to our proposed approach, "Single Model" refers to the approach using a single model for all file-pairs and our structured prior for partitions, and "Flat Prior" refers to the approach using our model for comparison data with a flat prior on tripartite matchings.

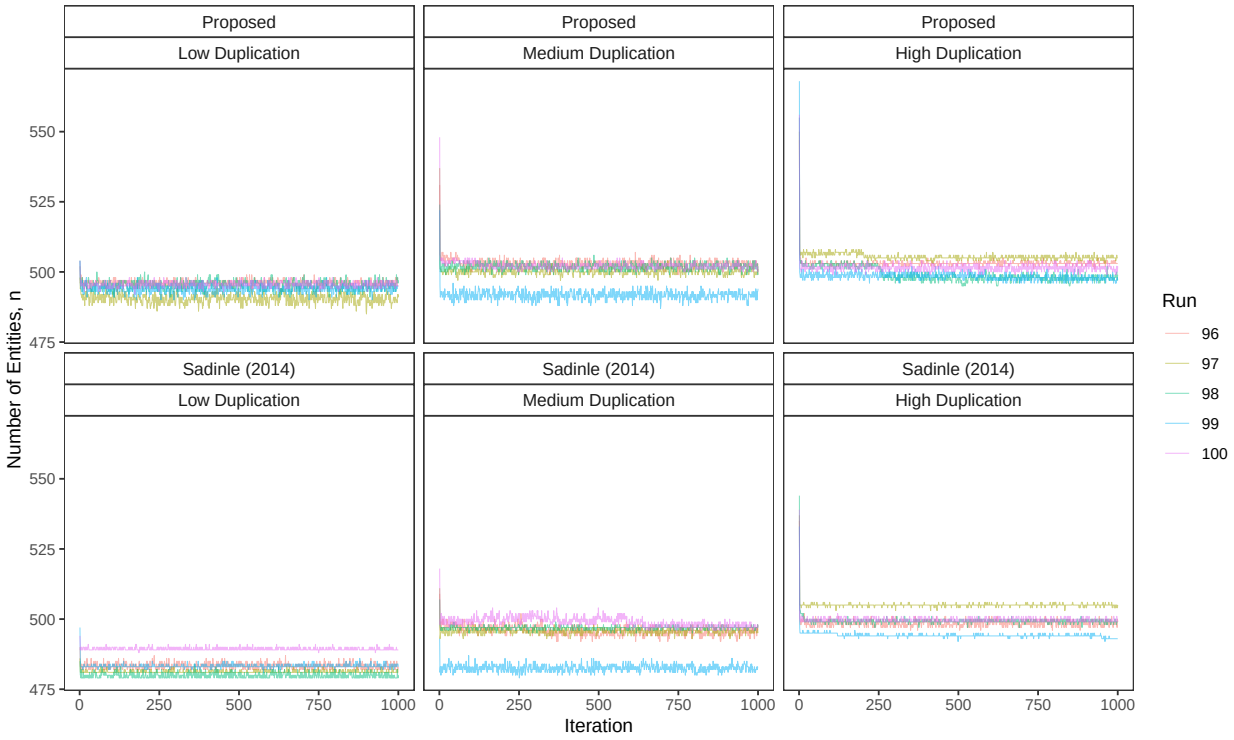


Figure A.18: Trace plots for n for the last 5 of 100 runs for the simulation with files with duplicates and equal measurement error across files, when using full estimates. "Proposed" refers to the proposed approach and "Sadinle (2014)" refers to the approach of [114].

A.5 Colombia Application

In this appendix we re-examine the three record systems containing information on homicides from 2004 in the Quindio province of Colombia, previously studied in [117]. These record systems, provided by the Conflict Analysis Resource Center (CERAC), were maintained by the Colombian National Statistics Office (Departamento Administrativo Nacional de Estadistica, DANE), the Colombian National Police (Policia Nacional de Colombia, PN), and the Colombian Forensics Institute (Instituto Nacional de Medicina Legal y Ciencias Forenses, ML). While the purpose of DANE is to record all homicides occurring in Colombia, PN and ML only record homicides obtained from their daily activities [34, 109]. Linking DANE to PN and ML is thus an important step in assessing the coverage of DANE and arriving at better estimates of the number of homicides in Colombia. Previously the records were linked by hand, which gives us a ground truth to assess the performance of our proposed approach. The linkage methodology of [117] did not scale to a large number of records, so the authors restricted their analysis to the 162 records from the last three months of 2004. We will now use our proposed approach to link all the 769 records from 2004.

A.5.1 Implementation Details

The three record systems are believed to be free of duplicates, so the target of inference is a tripartite matching. The fields available from all three record systems are municipality and date of the homicide, whether the location of the homicide was urban or rural, and the age, sex, and marital status of the victim. Additionally, educational status of the victim is available in DANE and ML, which we are able to use despite it being missing in PN, as explained in Section 4. Although we have seven fields available for the linkage, none of them provide a large amount of discriminative information, which comparison-based approaches rely on. Thus we expect there to be significant uncertainty in the linkage, making the proposed approach particularly relevant.

There are $r_1 = 323$ records in DANE, $r_2 = 157$ records in ML, and $r_3 = 289$ records in

Table A.5: Construction of levels of disagreement for the Colombian homocide record systems.

		Levels of disagreement			
Field	**Similarity measure**	**0**	**1**	**2**	**3**
Date	Absolute Difference	0	$1-2$	$3-7$	8+
Age	Absolute Difference	0	$1-2$	$3-9$	10+
Other Fields	Binary comparison	Agree	Disagree		

PN, so there are $189,431$ record pairs for which we construct comparison data, according to Table A.5. We use transitive indexing as described in Appendix B, where the initial indexing scheme declares record pairs as non-coreferent if they disagree in either municipality, sex, date by more than 60 days, or age by more than 9 years. This reduces the number of candidate coreferent record pairs down to $60,324$.

We present results from our approach under two prior specifications. The first is the default specification used in the simulations in Sections 6.2 and 6.3. The second specification differs from the default by placing a more informative prior on the overlap table through $\boldsymbol{\alpha}$. In particular, based on characteristics of the record systems described in [109], we specify a prior that captures the beliefs that: 1) if a homicide is recorded in PN or ML, it is highly likely to also be recorded in DANE, and 2) DANE and PN are expected to have a high coverage of homicides. This prior is described in more detail in the following section. We used the same loss function specification as outlined in Section 6.1 and Appendix D. We ran $3,000$ iterations of the Gibbs sampler presented in Appendix B, discarding the first $1,000$ as burn-in. In Section A.5.5 we discuss convergence of the Gibbs sampler for this application.

A.5.2 An Informative Prior Specification

We will guide our prior specification using the following two passages from [109] (translated to English):

- "According to DANE, 'The differences in the Legal Medicine and DANE data are mainly due to the fact that the latter organization receives, in addition to the death certificates sent by Legal Medicine (which are sent to DANE after a technical examination), the homicide reports made by police inspectors, nurses or health promoters - in places where there are no legal doctors - who arrive at the site where the body is found and register the cases as homicide without a technical examination and according to the picture that presents the corpse'. So we find here a reason for the increased coverage of vital statistics" [109, p. 331].

- "The National Police assures to have coverage at the national level, making an institutional presence in all the municipalities of the country: 'We have a presence throughout the country and we know all the cases; Legal Medicine does not have the same coverage and thus their data is not exact' " [109, p. 330].

Note that our Dirichlet-Multinomial prior for the overlap table can be motivated as the result of first drawing $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$ from a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}$, then drawing $\boldsymbol{n}$ from a multinomial distribution of size n with probabilities $\{q_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}$. Based on these passages we specify $\boldsymbol{\alpha}$ as follows, referring in our notation to DANE as list 1, ML as list 2, and PN as list 3:

- If a homicide is going to be recorded by one of the three record systems, it is very likely that it will be known by DANE. Therefore, we choose $\alpha_{1++} = \alpha_{100} + \alpha_{101} + \alpha_{110} + \alpha_{111}$ and $\alpha_{0++} = \alpha_{001} + \alpha_{010} + \alpha_{011}$ such that $\text{mode}(q_{1++}) = 0.95$ and $\mathbb{P}(q_{1++} > 0.9) = 0.95$, where $q_{1++} = q_{100} + q_{101} + q_{110} + q_{111} \sim \mathsf{Beta}(\alpha_{1++}, \alpha_{0++})$ is the prior probability of a homicide being recorded in DANE given it is recorded in one of the three systems.

- If a homicide is recorded by PN, then it is very likely that it will be recorded by DANE. Therefore, we choose $\alpha_{1+1} = \alpha_{101} + \alpha_{111}$ and $\alpha_{0+1} = \alpha_{001} + \alpha_{011}$ such that $\text{mode}(q_{1+1}) = 0.95$ and $P(q_{1+1} > 0.9) = 0.9$, where $q_{1+1} = q_{101} + q_{111} \sim \mathsf{Beta}(\alpha_{1+1}, \alpha_{0+1})$

is the prior probability of a homicide being recorded in DANE given it is recorded in PN.

- If a homicide is recorded by ML, then it is very likely that it will be known by DANE. Therefore, we choose $\alpha_{11+} = \alpha_{110} + \alpha_{111}$ and $\alpha_{01+} = \alpha_{010} + \alpha_{011}$ such that $\text{mode}(q_{11+}) = 0.95$ and $P(q_{11+} > 0.9) = 0.9$, where $q_{11+} = q_{110} + q_{111} \sim \mathsf{Beta}(\alpha_{11+}, \alpha_{01+})$ is the prior probability of a homicide being recorded in DANE given it is recorded in ML.

- The above induce six constraints, which determine α_{011}, α_{001}, and α_{010}, but we need one extra constraint to determine the remaining α. We thus choose the configuration of α_{100}, α_{101}, α_{110}, and α_{111} that maximizes α_{101}, which controls the prior probability of a homicide being jointly recorded by DANE and PN, given that these systems that are supposed to have the largest coverage of homicides.

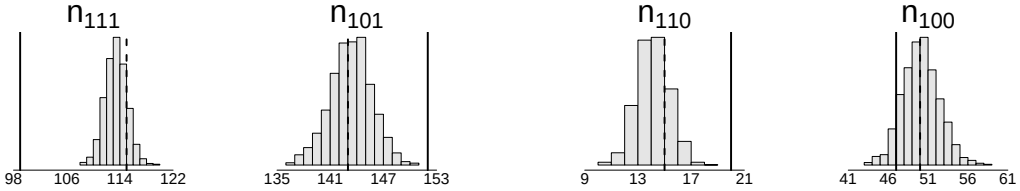
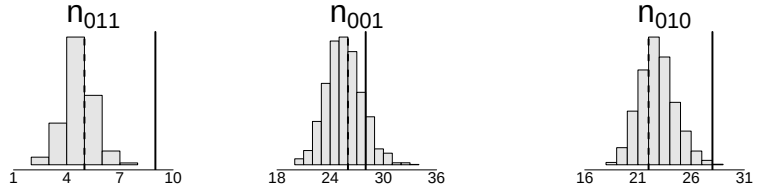
Under this specification we arrive at the setting $\alpha_{001} = 1.63, \alpha_{010} = 1.63, \alpha_{011} = 2.94, \alpha_{100} = 0, \alpha_{101} = 30.96, \alpha_{110} = 30.96, \alpha_{111} = 37.78$. For the sake of propriety, we set $\alpha_{100} = 0.1$.

A.5.3 Results

Under the default prior specification the precision and recall of the full estimate are 0.90 and 0.93 respectively. Recall is no longer useful when using partial estimates, as we are not trying to find all true matches. Thus we will assess the performance of the partial estimates using precision and the *abstention rate*, the proportion of records which the Bayes estimate abstained from making a linkage decision. The partial estimate has an abstention rate of 11%, and improves the precision of the estimate to 0.93. Under the informative prior specification the precision and recall of the full estimate are 0.93 and 0.96 respectively. The partial estimate has an abstention rate of 11%, and improves the precision of the estimate to 0.95. Due to the performance difference, we will focus on the results under the informative prior specification for the rest of this section. Analogous results under the default specification are provided in the next section.

The total number of homicides based on the hand labelling is 383. Under the informative prior specification a 95% credible interval for the number of unique homicides n is $[372, 383]$, with an estimate based on the full estimate of the tripartite matching of 376. In Table A.6 we display the posterior distribution for the overlap table and the overlap table derived from the full estimate, along with the overlap table derived from the ground truth hand labelling. We can see that n_{111}, the number of homicides recorded in all three files, and n_{100}, the number of homicides recorded in just DANE, are overestimated, and the remaining cells of the overlap table (and n) are underestimated.

Table A.6: Posterior distribution of the overlap table for the Colombian record systems, under the informative prior specification. Black lines indicate the ground truth, dotted lines indicate quantities derived from the full estimate of the tripartite matching.

	In PN		Out PN	
DANE	In ML	Out ML	In ML	Out ML
In				
Out				–

While the performance of the estimated matching is fairly good, with precision of both the full and partial estimate (before clerical review) above 0.9, the overestimation of n_{111} and the underestimation of n is indicative of over-matching. We believe this over-matching occurrs due to the low amount of discriminative information provided by the fields. In particular, the only fields we believe can be fully trusted are municipality of the homicide (of which

there are 12) and sex of the victim (which is coded as binary), which can only partition the records into 22 blocks of candidate coreferent records (there are no records of female homicide victims in two municipalities). The remaining fields all have some amount of error and do not provide highly discriminative information since they are either low dimensional categorical fields (urban/rural location of the homicide has two categories, marital status has five categories, and educational status has six categories) or numeric fields that are essentially ordinal categorical (date of homicide and age of victim). Therefore, records of different homicides may look similar based on the comparisons of these fields, causing them to be mistakenly matched. In these low-information settings, clerical review becomes especially important for the record linkage workflow, which our proposed approach of using partial estimates incorporates by design.

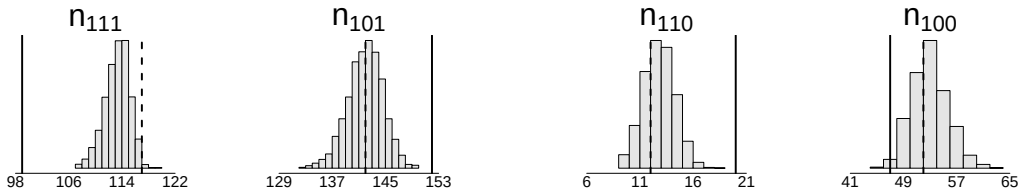
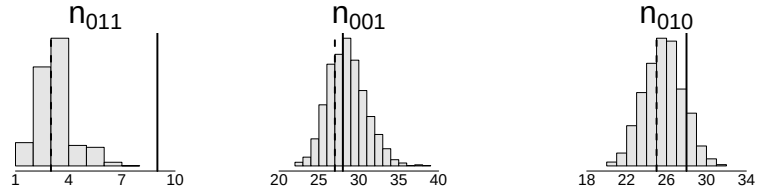
A.5.4 Results Under Default Prior

Under the default prior specification a 95% credible interval for the number of unique homicides n is $[376, 389]$, with an estimate based on the full estimate of the tripartite matching of 378. Thus we see that n is better estimated under the default prior compared to the informative prior (though the point estimate is still an underestimate). In Table A.7 we display the posterior distribution for the overlap table and the overlap table derived from the full estimate, along with the overlap table derived from the ground truth hand labelling. We can see that, as with the informative prior specification, n_{111} and n_{100} are overestimated, and the remaining cells of the overlap table (and n) are underestimated.

A.5.5 Convergence Diagnostics

In the application, we ran the the Gibbs sampler described in Appendix A.2.2 for $3,000$ iterations, discarding the first $1,000$ samples as burn-in, under a default and an informative prior specification. In Figure A.19 we present the the trace plots for the number of entities, n, under each of these prior specifications. The chains for n appear to converge quickly based on these trace plots. For each of these chains we computed Geweke's convergence

Table A.7: Posterior distribution of the overlap table for the Colombian record systems, under the default prior specification. Black lines indicate the ground truth, dotted lines indicate quantities derived from the full estimate of the tripartite matching.

	In PN		Out PN	
DANE	In ML	Out ML	In ML	Out ML
In				
Out				–

diagnostic as implemented in the `R` package `coda` [105]. The Geweke's Z-scores indicated it was reasonable to treat these chains as drawn from their stationary distributions.

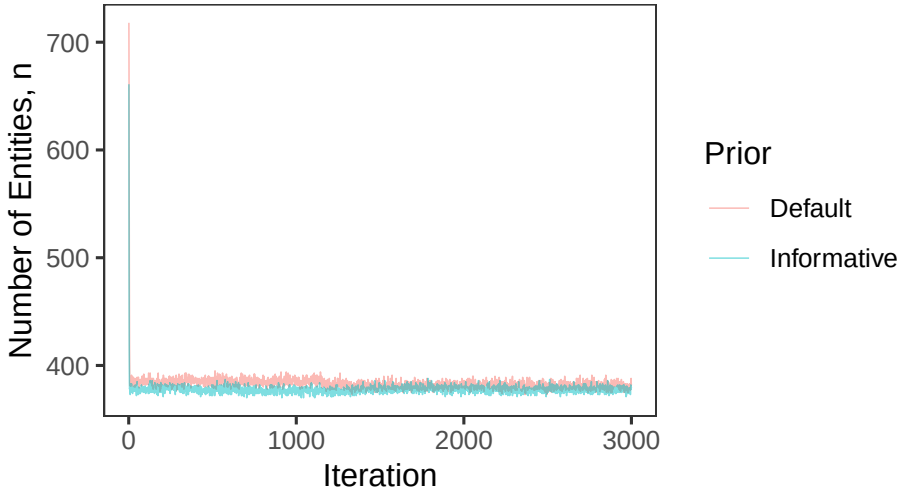


Figure A.19: Trace plots for n in Colombia application.

Appendix B

APPENDIX FOR CHAPTER 2

B.1 Conditional Identifiability in Models for Heterogeneity

The purpose of this appendix is to show how common models for heterogeneity fit into the model described in Section 2.2 of Chapter 3, and to provide results regarding conditional identifiability in a particular family of heterogeneous models.

B.1.1 Models for Heterogeneity

Consider the following heterogeneous model

$$\begin{aligned} \boldsymbol{\pi}^i &\overset{i.i.d.}{\sim} Q, \\ \boldsymbol{x}_i \mid \boldsymbol{\pi}^i &\overset{ind.}{\sim} \textsc{Categorical}(\boldsymbol{\pi}^i), \end{aligned} \tag{B.1}$$

where $\boldsymbol{\pi}^i = \{\pi^i_{\boldsymbol{h}}\}_{\boldsymbol{h}\in H} \in \mathbb{S}^{2^K-1}$ for $i = 1, \ldots, N$. Under this model each individual has its own set of cell probabilities, $\boldsymbol{\pi}^i$, drawn from some mixing distribution Q on $\mathbb{S}^{2^K-1}$. Working with the heterogeneous model in (B.1) is equivalent, after marginalizing out $\boldsymbol{\pi}^i$, to working with the complete-data distribution in (3.1), where $\boldsymbol{\pi} := \boldsymbol{\pi}_Q = E_Q(\boldsymbol{\pi}^i)$ and E_Q denotes the expectation with respect to the mixing distribution Q. This is a consequence of the data only providing information about the first moment of the mixing distribution. Suppose $\mathcal{Q}$ is a family of mixing distributions on $\mathbb{S}^{2^K-1}$. For $Q \in \mathcal{Q}$, let $\pi_{Q,0}$ denote the induced observed cell probability and $\tilde{\boldsymbol{\pi}}_Q$ denote the induced observed cell probabilities. The parameter space induced by the family $\mathcal{Q}$, as a subset of the observed-data parameterization, can then be written as $\Omega_{\mathcal{Q}} = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \pi_{Q,0} \text{ and } \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}_Q \text{ for some } Q \in \mathcal{Q}\}$.

The general heterogeneous model in (B.1) captures common models for heterogeneity, including the M_h and M_{th} models [101]. The M_{th} model assumes the individual cell probabilities take the form $\pi^i_{\boldsymbol{h}} = \prod_{k=1}^K (q^i_k)^{h_k}(1-q^i_k)^{1-h_k}$, where $(q^i_1, \cdots, q^i_K) \overset{i.i.d.}{\sim} Q$ and Q is a

mixing distribution on $(0,1)^K$. Under this model, conditional on an individual's sampling probabilities, $(q_1^i, \cdots, q_K^i)$, each individual is independently sampled by each list. The M_h model is a submodel of the M_{th} model that assumes that the individual sampling probabilities, $(q_1^i, \cdots, q_K^i)$, are the same for each list, i.e. $q_1^i = \cdots = q_K^i$. Thus the M_h model assumes individuals have the same probability of being sampled by each list. After marginalizing out $\boldsymbol{\pi}^i$, this enforces a symmetry where the probability of appearing in k lists is the same for each subset of k lists. We do not believe this is plausible in human population settings.

B.1.2 Conditional Identifiability in M_{th} Models

While there exists a literature characterizing identifiability in M_h models [64, 79, 61, 80], no such results exist for M_{th} models. The purpose of this section is to provide a mechanism for verifying whether the M_{th} model $\mathcal{P}_{\Omega_\mathcal{Q}}$ is conditionally identifiable based on moments of the mixing distributions $Q \in \mathcal{Q}$, analogously to the results for M_h models presented in [61].

Before proving the main theorem of this section, we have the following lemma, which tells us that for any mixing distribution Q on $(0,1)^K$, the induced cell probabilities, $\boldsymbol{\pi}_Q$, only depend on Q through its mixed moments.

Lemma B.1. *For any* $\boldsymbol{h} \in H^*$, $\pi_{Q,\boldsymbol{h}} = \sum_{\boldsymbol{h}' \in H^*} c_{\boldsymbol{h},\boldsymbol{h}'} m_{Q,\boldsymbol{h}'}$ *where* $c_{\boldsymbol{h},\boldsymbol{h}'} = (-1)^{\sum_{k=1}^K h'_k - h_k} \prod_{k=1}^K I(h_k \leq h'_k)$ *and* $m_{Q,\boldsymbol{h}'} = E_Q(\prod_{k=1}^K q_k^{h'_k})$.

Proof. For all $\boldsymbol{h} \in H^*$, $\prod_{k=1}^K q_k^{h_k}(1-q_k)^{1-h_k} = \sum_{\boldsymbol{h}' \in H^*} c_{\boldsymbol{h},\boldsymbol{h}'} \prod_{k=1}^K q_k^{h'_k}$ by an application of the multi-binomial theorem (a generalization of the binomial theorem). The result follows from taking the expectation over both sides with respect to Q. □

We can restate Lemma B.1 in matrix form. Letting $\boldsymbol{\pi}_Q^* = (\pi_{Q,\boldsymbol{h}})_{\boldsymbol{h} \in H^*}$ and $\boldsymbol{m}_Q = (m_{Q,\boldsymbol{h}})_{\boldsymbol{h} \in H^*}$, we have that $\boldsymbol{\pi}_Q^* = C\boldsymbol{m}_Q$, where $C = (c_{\boldsymbol{h},\boldsymbol{h}'})_{\boldsymbol{h} \in H^*, \boldsymbol{h}' \in H^*}$. C is invertible as it is upper triangular with non-zero diagonal entries. We are now ready to prove Theorem B.1.

Theorem B.1. *For any two distributions* Q, R *on* $(0,1)^K$, $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ *is equivalent to* $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ *for some* $A > 0$.

Proof. $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ is equivalent to $\boldsymbol{\pi}^*_Q/(1-\pi_{Q,0}) = \boldsymbol{\pi}^*_R/(1-\pi_{R,0})$. Rearranging terms we have that $\boldsymbol{\pi}^*_Q = \boldsymbol{\pi}^*_R(1-\pi_{Q,0})/(1-\pi_{R,0})$, and thus $\boldsymbol{\pi}^*_Q = A\boldsymbol{\pi}^*_R$, where $A = (1-\pi_{Q,0})/(1-\pi_{R,0}) > 0$. Using Lemma B.1, this is equivalent to $C\boldsymbol{m}_Q = AC\boldsymbol{m}_R$, and thus $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ due to the invertibility of C. □

The immediate consequence of Theorem B.1 is that to verify conditional identifiability of an M_{th} model $\mathcal{P}_{\Omega_\mathcal{Q}}$, one can demonstrate that if $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for some $Q, R \in \mathcal{Q}$, then $\pi_{Q,0} = \pi_{R,0}$. We use this mechanism in the next section to characterize when latent class models (LCMs) are conditionally identifiable.

B.1.3 Conditional Identifiability of Latent Class Models

We denote the family of mixing distributions corresponding to LCMs with J classes by $\mathcal{Q}_J = \{Q = \sum_{j=1}^J \nu_{Q,j} \prod_{k=1}^K \delta_{q_{Q,jk}} \mid \nu_{Q,j} \geq 0, \sum_{j=1}^J \nu_{Q,j} = 1, q_{Q,jk} \in (0,1)^K\}$, so that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is equivalent to $\mathcal{P}_{\Omega_{LCM,J}}$ from Chapter 3. To provide necessary and sufficient conditions for $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ to be conditionally identifiable, we restrict the family of mixing distributions to $\mathcal{Q}_J = \{Q = \sum_{j=1}^J \nu_{Q,j} \prod_{k=1}^K \delta_{q_{Q,jk}} \mid \nu_{Q,j} \geq 0, \sum_{j=1}^J \nu_{Q,j} = 1, q_{Q,jk} \in (0,1)^K, q_{Q,jk} \neq q_{Q,j'k} \text{ for } j \neq j'\}$. This restriction makes the mild assumption that each class' sampling probabilities are distinct, which simplifies the proof of Theorem B.2. Loosening this restriction could only make the conditions on J for $\mathcal{Q}_J$ to be identifiable stricter, and thus the conclusions we reach in Section B.1.6 would still stand for families where this restriction is violated.

There are $J(K+1)-1$ parameters in $\mathcal{Q}_J$, thus when $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable, J satisfies $J(K+1)-1 \leq 2^K-2$, as the observed cell probabilities, $\tilde{\boldsymbol{\pi}}_Q$, are 2^K-2 dimensional. However, we now prove that J must satisfy a stricter condition for $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ to be conditionally identifiable. In Section B.1.6 we discuss some limitations of this result.

Theorem B.2. *$\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable iff $2J \leq K$.*

Proof. We will first show that if $2J \leq K$, then $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identifiable. The proof of this direction is similar in spirit to the proofs of Theorem 2 in [61] and Theorem 1 in [103], which were both concerned with characterizing the identifiability of the M_h analogue

of $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. Assume $2J \leq K$, and let $Q, R \in \mathcal{Q}_J$ such that $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for some $A > 0$, so that we have the following system of equations:

$$\sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{j=1}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} = 0 \quad (\boldsymbol{h} \in H^*). \tag{B.2}$$

Let $\mathcal{I}_Q = \{j \mid q_{Q,j} \notin (q_{R,1}, \ldots, q_{R,J})\}$ and $\mathcal{I}_R = \{j \mid q_{R,j} \notin (q_{Q,1}, \ldots, q_{Q,J})\}$, where $q_{Q,j} = (q_{Q,j1}, \ldots, q_{Q,jK})$ and $q_{R,j} = (q_{R,j1}, \ldots, q_{R,jK})$. We can then rewrite (B.2) as

$$\sum_{j=1}^{J} y_j \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{i \in \mathcal{I}_R}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} = 0 \quad (\boldsymbol{h} \in H^*), \tag{B.3}$$

where $y_j = \nu_{Q,j}$ if $j \in \mathcal{I}_Q$ and $y_j = \nu_{Q,j} - A\nu_{R,j'}$ for some $j' \in \{1, \ldots, J\} \setminus \mathcal{I}_R$ otherwise. Letting $m = |\mathcal{I}_R| = |\mathcal{I}_Q|$ and labelling the elements of $\mathcal{I}_R$ as $i_1, \ldots, i_m$, the system of equations in (B.3) can be written in matrix form as $\Lambda \boldsymbol{y} = 0$, where

$$\Lambda = \begin{pmatrix} q_{Q,1K} & \cdots & q_{Q,JK} & q_{R,i_1K} & \cdots & q_{R,i_mK} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \prod_{k=1}^{K} q_{Q,1k}^{h_k} & \cdots & \prod_{k=1}^{K} q_{Q,Jk}^{h_k} & \prod_{k=1}^{K} q_{R,i_1k}^{h_k} & \cdots & \prod_{k=1}^{K} q_{R,i_mk}^{h_k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \prod_{k=1}^{K} q_{Q,1k} & \cdots & \prod_{k=1}^{K} q_{Q,Jk} & \prod_{k=1}^{K} q_{R,i_1k} & \cdots & \prod_{k=1}^{K} q_{R,i_mk} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_J \\ -A\nu_{R,i_1} \\ \vdots \\ -A\nu_{R,i_m} \end{pmatrix},$$

and the rows of Λ are indexed by $\boldsymbol{h} \in H^*$. In Section B.1.4, we prove that Λ is full rank, and thus $\boldsymbol{y} = 0$, for any $m \in \{0, \ldots, J\}$. The proof of this direction concludes by examining three possible cases.

Case 1. *Suppose $m = 0$, i.e. for each $j \in \{1, \ldots, J\}$, there exists some $j' \in \{1, \ldots, J\}$ such that $q_{Q,j} = q_{R,j'}$ and $\nu_{Q,j} = A\nu_{R,j'}$. As $\sum_{j=1}^{J} \nu_{Q,j} = \sum_{j=1}^{J} \nu_{R,j} = 1$, this implies that $A = 1$ and thus $\pi_{Q,0} = \pi_{R,0}$.*

Case 2. *Suppose $m \in \{1, \ldots, J-1\}$, i.e. for each $j \in \{1, \ldots, J\} \setminus \mathcal{I}_Q$, there exists some $j' \in \{1, \ldots, J\} \setminus \mathcal{I}_R$ such that $q_{Q,j} = q_{R,j'}$ and $\nu_{Q,j} = A\nu_{R,j'}$. Further, for each $j \in \mathcal{I}_Q$ and $j' \in \mathcal{I}_R$ $\nu_{Q,j} = \nu_{R,j'} = 0$. We can thus ignore the classes $j \in \mathcal{I}_Q$ and $j' \in \mathcal{I}_R$. As $\sum_{j=1}^{J} \nu_{Q,j} = \sum_{j=1}^{J} \nu_{R,j} = 1$, this implies that $A = 1$ and thus $\pi_{Q,0} = \pi_{R,0}$.*

Case 3. *Suppose $m = J$, i.e. for each $j \in \{1, \dots, J\}$, there exists no $j' \in \{1, \dots, J\}$ such that $q_{Q,j} = q_{R,j'}$. Then $\nu_{Q,j} = \nu_{R,j} = 0$ for $j \in \{1, \dots, J\}$, which is a contradiction.*

We will now show that if $2J > K$, then $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is not conditionally identifiable. To do so we will provide explicit $Q, R \in \mathcal{Q}_J$ such that $\pi_{Q,0} \neq \pi_{R,0}$, but $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for $A > 0$. This counterexample is modified from [129], who studied identifiability of families of LCMs outside of the multiple-systems estimation context where n_0 is observed. Choose J such that $2J > K$. For $j \in \{1, \dots, J\}$, let $\nu_{Q,j} = \binom{2J}{2j}/(2^{2J-1} - 1)$ and $\nu_{R,j} = \binom{2J}{2j-1}/(2^{2J-1})$. For $j \in \{1, \dots, J\}$ and $k \in \{1, \dots, K\}$, let $q_{Q,jk} = \alpha(2j)$ and $q_{R,jk} = \alpha(2j-1)$ where $0 < \alpha < 1/(2J)$. We thus have that $Q, R \in \mathcal{Q}_J$, where clearly $Q \neq R$. In Section B.1.5 we prove that for these choices of Q, R, $\boldsymbol{m}_Q = A\boldsymbol{m}_R$ for $A > 0$ such that $A \neq 1$, and thus $\pi_{Q,0} \neq \pi_{R,0}$. □

B.1.4 Proof that Λ is Full Rank

We will prove that Λ is full rank for any $m \in \{0, \dots, J\}$ by proving a stronger result. Recall that $K \geq 2$ and let $x_{\ell k} \in (0, 1)$ for $\ell \in \{1, \dots, K\}$ and $k \in \{1, \dots, K\}$, such that $x_{\ell k} \neq x_{\ell k'}$ for $k \neq k'$. Let

$$X^K = \begin{pmatrix} x_{1K} & \cdots & x_{KK} \\ \vdots & \ddots & \vdots \\ \prod_{k=1}^K x_{1k}^{h_k} & \cdots & \prod_{k=1}^K x_{Kk}^{h_k} \\ \vdots & \ddots & \vdots \\ \prod_{k=1}^K x_{1k} & \cdots & \prod_{k=1}^K x_{Kk} \end{pmatrix},$$

where the rows of X^K are indexed by $\boldsymbol{h} \in H^*$. We will show that X^K is full rank by induction on K. This implies that Λ is full rank, as $J + m \leq 2J \leq K$ by assumption for any $m \in \{0, \dots, J\}$.

For the base case when $K = 2$, verifying X^2 is full rank is straightforward. Assume that X^{K-1} is full rank. Let $\boldsymbol{v} \in \mathbb{R}^{K \times 1}$ be such that $X^K \boldsymbol{v} = 0$. For each $\boldsymbol{h} \in \{\boldsymbol{h}' \in H^* \mid h'_K = 0\}$ we have that $v_K \prod_{k=1}^{K-1} x_{Kk}^{h_k} = -\sum_{\ell=1}^{K-1} v_\ell \prod_{k=1}^{K-1} x_{\ell k}^{h_k}$, which implies that $\sum_{\ell=1}^{K-1} v_\ell (x_{\ell K} -$

$x_{KK}) \prod_{k=1}^{K-1} x_{\ell k}^{h_k} = 0$. For $\ell \in \{1, \ldots, K-1\}$, let $v'_\ell = v_\ell(x_{\ell K} - x_{KK})$ and $\boldsymbol{v}' = (v'_1, \ldots, v'_{K-1})$. This leads to the system of equations $X^{K-1}\boldsymbol{v}' = 0$. By the inductive assumption, $\boldsymbol{v}' = 0$. Since $x_{\ell K} \neq x_{KK}$ for $\ell \in \{1, \ldots, K-1\}$, we have that $v_\ell = 0$ for $\ell \in \{1, \ldots, K-1\}$, and thus $v_K = 0$.

B.1.5 Proof of Counterexample

We will now prove that $m_{Q,\boldsymbol{h}} = A m_{R,\boldsymbol{h}}$ for all $\boldsymbol{h} \in H^*$, where $A = (2^{2J-1})/(2^{2J-1} - 1) \neq 1$. Define the function $h(x) = (1 - e^{\alpha x})^{2J} = \sum_{i=0}^{2J} \binom{2J}{i} (-1)^i e^{\alpha i x}$. For $t \in \{1, \ldots, K\}$, we can differentiate the series representation of h to find that $h^{(t)}(x) = \sum_{i=0}^{2J} \binom{2J}{i} (-1)^i (\alpha i)^t e^{\alpha i x}$ and thus $h^{(t)}(x)|_{x=0} = \sum_{i=0}^{2J} \binom{2J}{i} (-1)^i (\alpha i)^t = \sum_{i=1}^{2J} \binom{2J}{i} (-1)^i (\alpha i)^t$. We can alternatively differentiate the non-series representation of h using the fact that $t \leq K < 2J$ and the chain rule for higher order derivatives to find that $h^{(t)}(x)|_{x=0} = 0$. Let $\boldsymbol{h} \in H^*$ and $t = \sum_{k=1}^{K} h_k \in \{1, \ldots, K\}$. The desired result follows as

$$
\begin{aligned}
m_{Q,\boldsymbol{h}} - A m_{R,\boldsymbol{h}} &= \sum_{j=1}^{J} \nu_{Q,j} \prod_{k=1}^{K} q_{Q,jk}^{h_k} - A \sum_{j=1}^{J} \nu_{R,j} \prod_{k=1}^{K} q_{R,jk}^{h_k} \\
&= \sum_{j=1}^{J} \binom{2J}{2j} (2^{2J-1} - 1)^{-1} \prod_{k=1}^{K} \{\alpha(2j)\}^{h_k} - A \sum_{j=1}^{J} \binom{2J}{2j-1} (2^{2J-1})^{-1} \prod_{k=1}^{K} \{\alpha(2j-1)\}^{h_k} \\
&= (2^{2J-1} - 1)^{-1} \sum_{i=1}^{2J} \binom{2J}{i} (-1)^i (\alpha i)^t = (2^{2J-1} - 1)^{-1} \{h^{(t)}(x)|_{x=0}\} = 0.
\end{aligned}
$$

B.1.6 Limitations of Theorem B.2

Theorem B.2 shows that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is not conditionally identifiable if $2J > K$ by counterexample, by demonstrating two mixing distributions $Q, R \in \mathcal{Q}_J$ where $\tilde{\boldsymbol{\pi}}_Q = \tilde{\boldsymbol{\pi}}_R$ but $\pi_{Q,0} \neq \pi_{R,0}$. Within each latent class of Q and R, the sampling probabilities were the same, meaning Q and R can be seen as mixing distributions of an M_h model. It would be interesting in future work to see whether further restrictions on $\Omega_{\mathcal{Q}_J}$, for example restrictions not allowing the sampling probabilities within latent classes to be equal, lead to different results concerning conditional identifiability. Another interesting route would be to see whether results concerning *generic*

identifiability of latent class models [5] could be applied to the multiple-systems estimation setting.

However, this does not mean Theorem B.2 is not a practically useful result. Theorem B.2 provides assumptions under which which we have formal statistical guarantees for when we can estimate the parameters in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$: the parameters of $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ can be consistently estimated if $2J \leq K$. When $2J > K$ we currently have no such guarantees. In Web Appendix B.4 we demonstrate this reality across a variety of simulation studies.

B.2 Computation for Conditionally Identified Models

The purpose of this appendix is to provide details of how computation for conditionally identified models can be carried out in both frequentist and Bayesian frameworks using existing software. Recall from Sections 2.3 and 2.4 of Chapter 3 that the complete-data distribution can be written as

$$p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) = N! \prod_{\boldsymbol{h} \in H} \frac{\pi_{\boldsymbol{h}}^{n_{\boldsymbol{h}}}}{n_{\boldsymbol{h}}!} = L_1(N, \pi_0 \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}), \tag{B.4}$$

with $L_1(N, \pi_0 \mid n) = \binom{N}{n} \pi_0^{N-n} (1 - \pi_0)^n$ and $L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) = n! \prod_{\boldsymbol{h} \in H^*} \tilde{\pi}_{\boldsymbol{h}}^{n_{\boldsymbol{h}}} / n_{\boldsymbol{h}}!$, and that conditionally identified models have parameter spaces of the form $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$.

B.2.1 Computation for Frequentist Multiple-Systems Estimation

In this section we will first describe an approach for frequentist inference in general conditionally identified models, followed by the specific cases of models using the NHOI and the K'-list marginal NHOI identifying assumptions.

Conditionally Identified Models in General

Suppose that we are using a conditionally identified model with parameter space $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$. Frequentist inference for this general condition-

ally identified model will follow from the conditional maximum likelihood approach outlined in [118] and [45]. In particular, this approach can be summarized in two steps:

1. Estimate the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ by maximizing the conditional likelihood over the set of possible observed cell probabilities $\tilde{S}$:

$$\hat{\boldsymbol{\pi}} = \arg\max_{\tilde{\boldsymbol{\pi}} \in \tilde{S}} L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}).$$

2. Estimate the population size N by maximizing the binomial likelihood for n conditional on the estimate of the observed cell probabilities, $\hat{\boldsymbol{\pi}}$:

$$\hat{N}(\hat{\boldsymbol{\pi}}) = \arg\max_{N \in \mathbb{N}} L_1(N, \mathcal{T}(\hat{\boldsymbol{\pi}}) \mid n) = \left\lfloor \frac{n}{1 - \mathcal{T}(\hat{\boldsymbol{\pi}})} \right\rfloor,$$

where $\lfloor \cdot \rfloor$ is the floor function. We will ignore the rounding and write the estimator of N as $\hat{N}(\hat{\boldsymbol{\pi}}) = n/\{1 - \mathcal{T}(\hat{\boldsymbol{\pi}})\}$. This is well known as the Horvitz-Thompson estimator [63].

We note here that $\hat{\boldsymbol{\pi}}$, and thus $\hat{N}(\hat{\boldsymbol{\pi}})$, may not exist in general, depending on the set of possible observed cell probabilities $\tilde{S}$. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ maximize the conditional likelihood over $\mathbb{S}^{2^K-2}$, so if the sample proportions lie in $\tilde{S}$, then they maximize the conditional likelihood over $\tilde{S}$. If the sample proportions do not lie in $\tilde{S}$, care must be taken to make sure that $\hat{\boldsymbol{\pi}}$ exists.

For the rest of this section we will assume that the model is correctly specified, and $\hat{\boldsymbol{\pi}}$ exists. Let $\tilde{\boldsymbol{\pi}}^*$ denote the true observed cell probabilities. Suppose it is true, for an estimator $\hat{\boldsymbol{\pi}}$ of $\tilde{\boldsymbol{\pi}}^*$, that $\sqrt{n}(\hat{\boldsymbol{\pi}} - \tilde{\boldsymbol{\pi}}^*) \mid n \to_d \textsc{Normal}(0, \Sigma(\tilde{\boldsymbol{\pi}}^*))$, where $\to_d$ denotes convergence in distribution and we are conditioning on n (i.e. ignoring binomial variation in n). For example, when the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ lie within $\tilde{S}$, we have that $\hat{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h} \in H^*}$ and $\Sigma(\tilde{\boldsymbol{\pi}}^*) = \operatorname{diag}(\tilde{\boldsymbol{\pi}}^*) - \tilde{\boldsymbol{\pi}}^*(\tilde{\boldsymbol{\pi}}^*)^T$ [see e.g. chapter 14 of 2]. For $\tilde{\boldsymbol{\pi}} \in \tilde{S}$, let $f(\tilde{\boldsymbol{\pi}}) = 1/(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}))$. From the delta method, it follows that $\sqrt{n}(f(\hat{\boldsymbol{\pi}}) - f(\tilde{\boldsymbol{\pi}}^*)) \mid n \to_d \textsc{Normal}(0, (\nabla f(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla f(\tilde{\boldsymbol{\pi}}^*))$. Thus for large n, $nf(\hat{\boldsymbol{\pi}}) = \hat{N}(\hat{\boldsymbol{\pi}}) \approx \textsc{Normal}(nf(\tilde{\boldsymbol{\pi}}^*), n(\nabla f(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla f(\tilde{\boldsymbol{\pi}}^*))$. We can then substitute our estimate

$\hat{\boldsymbol{\pi}}$ of the observed cell probabilities for $\tilde{\boldsymbol{\pi}}^*$, and use this large sample approximation to construct 95% confidence intervals for N of the form $\hat{N}(\hat{\boldsymbol{\pi}}) \pm 1.96 * \sqrt{n(\nabla f(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla f(\hat{\boldsymbol{\pi}})}$. The term $(\nabla f(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla f(\hat{\boldsymbol{\pi}})$ can be calculated automatically using e.g. the `delta.method` function in the `R` package `msm` [66].

The confidence interval construction in the last paragraph conditions on n, and thus does not incorporate the binomial variation of n. Let N^* denote the true population size. For $\tilde{\boldsymbol{\pi}} \in \tilde{S}$, let $g(\tilde{\boldsymbol{\pi}}) = \mathcal{T}(\tilde{\boldsymbol{\pi}})/(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}))$. Following [45], unconditional of n we have that $(N^*)^{-1/2}(\hat{N}(\hat{\boldsymbol{\pi}}) - N^*) \to_d \textsc{Normal}(0, g(\tilde{\boldsymbol{\pi}}^*) + (1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}^*))(\nabla g(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla g(\tilde{\boldsymbol{\pi}}^*))$. Thus for large N^*, $\hat{N}(\hat{\boldsymbol{\pi}}) \approx \textsc{Normal}(N^*, N^* g(\tilde{\boldsymbol{\pi}}^*) + N^*(1 - \mathcal{T}(\tilde{\boldsymbol{\pi}}^*))(\nabla g(\tilde{\boldsymbol{\pi}}^*))^T \Sigma(\tilde{\boldsymbol{\pi}}^*) \nabla g(\tilde{\boldsymbol{\pi}}^*))$. We can then substitute our estimate $\hat{\boldsymbol{\pi}}$ of the observed cell probabilities for $\tilde{\boldsymbol{\pi}}^*$ and our estimate $\hat{N}(\hat{\boldsymbol{\pi}})$ of the population size for N^*, and use this large sample approximation to construct 95% confidence intervals for N of the form $\hat{N}(\hat{\boldsymbol{\pi}}) \pm 1.96 * \sqrt{\hat{N}(\hat{\boldsymbol{\pi}}) g(\hat{\boldsymbol{\pi}}) + n(\nabla g(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla g(\hat{\boldsymbol{\pi}})}$. Again, the term $(\nabla g(\hat{\boldsymbol{\pi}}))^T \Sigma(\hat{\boldsymbol{\pi}}) \nabla g(\hat{\boldsymbol{\pi}})$ can be calculated automatically using e.g. the `delta.method` function in the `R` package `msm` [66].

Computation for the NHOI and K'-List Marginal NHOI Identifying Assumptions

In this section we will focus on frequentist inference in the specific cases of models using the NHOI and the K'-list marginal NHOI identifying assumptions. While one could construct estimators and confidence intervals for N, under these assumptions, by hand using the results from the previous section, software is already available which accomplishes these tasks.

NHOI Identifying Assumption

For the NHOI identifying assumption, there are many `R` packages which produce estimates and confidence intervals for the population size under this assumption. For example, in our Kosovo application we use the `Rcapture` package [8]. The function `closedpMS.t` produces estimates and standard errors for the population size under all hierarchical log-linear models, including the saturated log-linear model $\mathcal{P}_{\Omega_{LL}}$. These can then be used to construct confidence intervals for the population size.

K'-List Marginal NHOI Identifying Assumption

Recall from Section 4.3 of Chapter 3 that the K'-list marginal NHOI identifying assumption restricts the observed cell probabilities to lie in $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$. Thus there are two cases to consider when fitting a model in the frequentist framework using the K'-list marginal NHOI identifying assumption:

1. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie within $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$.

2. The sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ do not lie within $\tilde{S}$.

There is a simple way to verify for a given data set, which case one is in. Consider the restricted data set from just the first K' lists. In particular, using notation from Section 4.3 of Chapter 3, $\{n^\dagger_{\boldsymbol{g}}\}_{\boldsymbol{g}\in G^*}$ is the restricted data, where $n^\dagger_{\boldsymbol{g}} = \sum_{\boldsymbol{h}\in H^*} n_{\boldsymbol{h}} I\{(h_1, \cdots, h_{K'}) = \boldsymbol{g}\}$, and the restricted sample size is $n^\dagger = \sum_{\boldsymbol{g}\in G^*} n^\dagger_{\boldsymbol{g}}$. Using this restricted data set of K' lists, one could compute the frequentist population size estimator under the NHOI assumption (for K' lists), using standard software (e.g., the `Rcapture` package as just described). Call this estimate $\hat{N}^\dagger$. Then the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie within $\tilde{S}$ as long as $\hat{N}^\dagger > n$.

Suppose we are in the second case, i.e. the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ do not lie in $\tilde{S}$. In this case, $\hat{\boldsymbol{\pi}}$ may not exist. One needs to verify that $\hat{\boldsymbol{\pi}}$ exists, and if it does, compute it and derive its asymptotic distribution to compute confidence intervals for N as described in Appendix B.2.1. This could potentially be quite difficult technically, so we recommend if one truly believes that the K'-list marginal NHOI identifying assumption holds in this case, that they use a Bayesian estimator as described in Appendix B.2.2.

Suppose now we are in the first case, i.e. the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie in $\tilde{S}$. Then $\hat{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$, and thus we could then follow the details at the end of Appendix B.2.1 to arrive at a confidence interval for N. However, we want to take advantage of existing software in order to compute estimates and confidence intervals for N. The following theorem accomplishes this task:

Theorem B.3. *Let $\hat{N}$ denote the population size estimator under the K'-list marginal NHOI identifying assumption using the full K list data set. Let $\hat{N}^{\dagger}$ denote the population size estimator when restricting to data from just the first K' lists and using the NHOI assumption for K' lists. If the sample proportions $\{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$ lie in $\tilde{S} = \{\tilde{\boldsymbol{\pi}} \in \mathbb{S}^{2^K-2} \mid \tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1\}$, then $\hat{N} = \hat{N}^{\dagger}$.*

We prove Theorem B.3 in Appendix B.2.1. Theorem B.3 tells us that if we want to calculate estimates and confidence intervals for the population size under the K'-list marginal NHOI identifying assumption, we can accomplish this by restricting the data set to K' lists, and calculating estimates and confidence intervals for the population size estimate for just these K' lists under the NHOI assumption for K' lists. This can be accomplished using the function `closedpMS.t` in the `Rcapture` package.

Sensitivity Analyses

`Rcapture` does not support sensitivity analyses that examine the impact of the NHOI or K'-list marginal NHOI identifying assumptions, as described in Section 4.2 and 4.3 of Chapter 3. However, it is straightforward to use the `glm` function in `R` to perform these sensitivity analyses, which is what `Rcapture` uses under the hood. In the code accompanying this manuscript, available at `github.com/aleshing/central-role-of-identifying-assumptions`, we provide a function which performs these sensitivity analyses.

Proof of Theorem B.3

Proof. Suppose we have data from K lists, $\{n_{\boldsymbol{h}}\}_{\boldsymbol{h}\in H^*}$, with observed sample size n, and we are using the K'-list marginal NHOI assumption, for $1 < K' < K$. For this proof, denote the sample proportions by $\tilde{\boldsymbol{\pi}} = \{n_{\boldsymbol{h}}/n\}_{\boldsymbol{h}\in H^*}$.

We start by restating some notation from Section 4.3 of the main paper. Let $G = \{0,1\}^{K'}$ index the marginal $2^{K'}$ contingency table for the first K' lists and $G^* = G \setminus \{0\}^{K'}$. Let $\tilde{\pi}_{\boldsymbol{g}+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$ and $\tilde{\pi}_{0+} = \sum_{\boldsymbol{h}\in H^*} \tilde{\pi}_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = (0,\cdots,0)\}$. The K'-lists marginal NHOI assumption corresponds to the explicit identifying assumption

$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1+\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$, where $\tilde{\Pi}_{odd,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*} \tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. $\mathcal{T}(\tilde{\boldsymbol{\pi}}) \in (0,1)$ since we assume that $\tilde{\Pi}_{odd,+}/(\tilde{\Pi}_{even,+}\tilde{\pi}_{0+}) > 1$.

Now we introduce some new notation. Suppose we are restricted to just the data from the first K' lists. Let $\{n_{\boldsymbol{g}}^\dagger\}_{\boldsymbol{g}\in G^*}$ denote the restricted data, so that $n_{\boldsymbol{g}}^\dagger = \sum_{\boldsymbol{h}\in H^*} n_{\boldsymbol{h}} I\{(h_1,\cdots,h_{K'}) = \boldsymbol{g}\}$, and the restricted sample size is $n^\dagger$. Denote the restricted sample proportions by $\tilde{\boldsymbol{\pi}}^\dagger = \{n_{\boldsymbol{g}}^\dagger/n^\dagger\}_{\boldsymbol{g}\in G^*}$. Using this restricted K' list data set, the NHOI assumption corresponds to the explicit identifying assumption $\mathcal{T}^\dagger(\tilde{\boldsymbol{\pi}}^\dagger) = (\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger)/(1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger)$, where $\tilde{\Pi}_{odd}^\dagger = \prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}_{\boldsymbol{g}}^\dagger)^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even}^\dagger = \prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}_{\boldsymbol{g}}^\dagger)^{I_{even}(\boldsymbol{g})}$.

In a frequentist framework, the population size estimate using the K'-list marginal NHOI assumption when the estimated observed cell probabilities are $\tilde{\boldsymbol{\pi}}$ is

$$\hat{N} = \frac{n}{1 - \frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+}}{1+\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+}}} = n\left[1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+}\right].$$

Similarly, in a frequentist framework the population size estimate using the NHOI assumption with the restricted K' list data set is

$$\hat{N}^\dagger = \frac{n^\dagger}{1 - \frac{\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger}{1+\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger}} = n^\dagger\left[1 + \tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right].$$

Our task is to prove that $\hat{N} = \hat{N}^\dagger$.

We list here two useful facts that can be verified through simple algebra:

1. $n^\dagger = n - n\tilde{\pi}_{0+} = n[1-\tilde{\pi}_{0+}]$.

2. $\tilde{\pi}_{\boldsymbol{g}} = \tilde{\pi}_{\boldsymbol{g}}^\dagger[1-\tilde{\pi}_{0+}]$.

Using the first fact, we can rewrite $\hat{N}^\dagger$:

$$\begin{aligned}
\hat{N}^\dagger &= n^\dagger\left[1 + \tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right] \\
&= n[1-\tilde{\pi}_{0+}]\left[1 + \tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger\right] \\
&= n\left[1 + (1-\tilde{\pi}_{0+})(\tilde{\Pi}_{odd}^\dagger/\tilde{\Pi}_{even}^\dagger) - \tilde{\pi}_{0+}\right].
\end{aligned}$$

Thus if we can show that $(1-\tilde{\pi}_{0+})(\tilde{\Pi}^{\dagger}_{odd}/\tilde{\Pi}^{\dagger}_{even}) = \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}$, the proof is complete. Using the second fact, we can rewrite $(1-\tilde{\pi}_{0+})(\tilde{\Pi}^{\dagger}_{odd}/\tilde{\Pi}^{\dagger}_{even})$:

$$\begin{aligned}(1-\tilde{\pi}_{0+})(\tilde{\Pi}^{\dagger}_{odd}/\tilde{\Pi}^{\dagger}_{even}) &= (1-\tilde{\pi}_{0+})\left[\frac{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{odd}(\boldsymbol{g})}}{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{even}(\boldsymbol{g})}}\right] \\ &= (1-\tilde{\pi}_{0+})\left[\frac{1-\tilde{\pi}_{0+}}{1-\tilde{\pi}_{0+}}\right]^{2^{K'-1}}\left[\frac{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{odd}(\boldsymbol{g})}}{\prod_{\boldsymbol{g}\in G^*}(\tilde{\pi}^{\dagger}_{\boldsymbol{g}})^{I_{even}(\boldsymbol{g})}}\right] \\ &= \left[\frac{\prod_{\boldsymbol{g}\in G^*}\tilde{\pi}^{I_{odd}(\boldsymbol{g})}_{\boldsymbol{g}+}}{\prod_{\boldsymbol{g}\in G^*}\tilde{\pi}^{I_{even}(\boldsymbol{g})}_{\boldsymbol{g}+}}\right] = \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+}.\end{aligned}$$

□

B.2.2 Computation for Bayesian Multiple-Systems Estimation

In this section we will describe a computational approach for Bayesian inference in general conditionally identified models, that allows any prior for the population size, N, and any prior for the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$. Various sensitivity analyses are facilitated from this approach. We further give some guidance to specification of the prior for N.

Bayesian Multiple-Systems Estimation

Suppose that we are using a conditionally identified model with parameter space $\Omega = \{N, \pi_0, \tilde{\boldsymbol{\pi}} \mid N \in \mathbb{N}, \pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}} \in \tilde{S}\}$, and we have specified independent prior distributions for N and $\tilde{\boldsymbol{\pi}}$, with densities $p(N)$ and $p(\tilde{\boldsymbol{\pi}})$. In this section, and the following two sections, we will let $p(\cdot)$ denote a density of a given random variable. The joint posterior of N and $\tilde{\boldsymbol{\pi}}$ can be written as $p(N, \tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_1(N, \mathcal{T}(\tilde{\boldsymbol{\pi}}) \mid n) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(N) p(\tilde{\boldsymbol{\pi}}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S})$. The marginal posteriors of $\tilde{\boldsymbol{\pi}}$ and N can be written as

$$p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}})) L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(\tilde{\boldsymbol{\pi}}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S}), \tag{B.5}$$

and $p(N \mid \boldsymbol{n}) = \int p(N \mid n, \mathcal{T}(\tilde{\boldsymbol{\pi}})) p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) d\tilde{\boldsymbol{\pi}}$, where $p(n \mid \pi_0) = \sum_{N=n}^{\infty} L_1(N, \pi_0 \mid n) p(N)$ and $p(N \mid n, \pi_0) = L_1(N, \pi_0 \mid n) p(N) / p(n \mid \pi_0)$, with $\pi_0 = \mathcal{T}(\tilde{\boldsymbol{\pi}})$. As we discuss in Section B.2.2, we can compute $p(n \mid \pi_0)$, and thus $p(N \mid n, \pi_0)$, analytically for common priors on N.

If one has access to Markov chain Monte Carlo (MCMC) samples $\{\tilde{\boldsymbol{\pi}}^{[t]}\}_{t=1}^{T}$ from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, one can then generate MCMC samples $\{N^{[t]}\}_{t=1}^{T}$ from $p(N \mid \boldsymbol{n})$ via $N^{[t]} \sim p(N \mid n, \mathcal{T}(\tilde{\boldsymbol{\pi}}^{[t]}))$. Summaries of the marginal posterior of N can then be calculated based on these samples.

Mixing and Matching Identifying Assumptions and Priors

While computation as described in the previous section may seem straightforward, the marginal posterior for the observed cell probabilities, $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, depends on the specific combination of priors for $\tilde{\boldsymbol{\pi}}$ and N and identifying assumption $\mathcal{T}$. Thus we need new MCMC samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for each new combination of priors and identifying assumption, which can be difficult both technically and computationally. Rather than develop new MCMC samplers for each combination, we will rely on a combination of existing software and a computationally cheap rejection sampler.

Let $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})p(\tilde{\boldsymbol{\pi}})$ denote the marginal "posterior" for the observed cell probabilities using just the conditional likelihood L_2. We use the subscript C (for "C"onditional) to denote that it is a special density that we are introducing for computational purposes. We can then rewrite the actual marginal posterior for the observed cell probabilities (B.5) as $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}}))I(\tilde{\boldsymbol{\pi}} \in \tilde{S})p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$. This suggests a computationally cheap rejection sampler to generate samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, if we have access to MCMC samples from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ [121]:

1. Generate $U \sim \text{Unif}(0, 1)$ and $\tilde{\boldsymbol{\pi}} \sim p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ independently.

2. If $U < p(n \mid \mathcal{T}(\tilde{\boldsymbol{\pi}}))I(\tilde{\boldsymbol{\pi}} \in \tilde{S})/\{\max_{\pi_0} p(n \mid \pi_0)\}$ accept $\tilde{\boldsymbol{\pi}}$. Else go back to (1).

Thus, for a given prior $p(\tilde{\boldsymbol{\pi}})$, if we want to perform prior sensitivity analyses for N and/or sensitivity analyses probing the identifying assumption as discussed in Sections 4.2 and 4.3 of Chapter 3, we can take a one time sample from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$, and then reuse this sample to generate samples from $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for each combination of prior for N and identifying assumption. The approach just described is only useful if we have access to MCMC samples

from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$. The rest of this section will describe how we can generate samples from the density $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ using existing software.

Previous work in Bayesian MSE specifies priors for $\tilde{\boldsymbol{\pi}}$ indirectly. In particular, most work specifies priors on reparametrizations of the cell probabilities $\boldsymbol{\pi}$, such as log-linear models or LCMs, which induce priors for $\boldsymbol{\pi}$, and thus for $\tilde{\boldsymbol{\pi}}$. Let $p^w(\boldsymbol{\pi})$ denote what we will call the "working" prior for $\boldsymbol{\pi}$, which induces the prior $p(\tilde{\boldsymbol{\pi}})$ we would like to use. We use the superscript w (for "w"orking) to denote that it is a special density that we are introducing for computational purposes. Consider the "working" posterior for $\boldsymbol{\pi}$, $p^w(\boldsymbol{\pi} \mid \boldsymbol{n}) \propto \sum_{N=n}^{\infty} p(\boldsymbol{n}, n_0 \mid N, \boldsymbol{\pi}) p^w(\boldsymbol{\pi})/N$, obtained using the "working" prior for N of $p^w(N) \propto 1/N$. The "posterior" for $\tilde{\boldsymbol{\pi}}$ under this working prior combination is equal to $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) p(\tilde{\boldsymbol{\pi}})$, as $p(n \mid \pi_0) \propto 1/n$ under the working prior for N (see Table B.1). Thus, given MCMC samples, $\{\boldsymbol{\pi}^{[t]}\}_{t=1}^{T}$, drawn from $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$, letting $\tilde{\pi}_{\boldsymbol{h}}^{[t]} = \pi_{\boldsymbol{h}}^{[t]}/(1 - \pi_0^{[t]})$, $\{\tilde{\boldsymbol{\pi}}^{[t]}\}_{t=1}^{T}$ are MCMC samples drawn from $p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$.

Thus if we want to use the prior $p(\tilde{\boldsymbol{\pi}})$ induced by a working prior $p^w(\boldsymbol{\pi})$, we can rely on a combination of existing software and a computationally cheap rejection sampler to generate draws from the posterior $p(N, \tilde{\boldsymbol{\pi}} \mid \boldsymbol{n})$ for any combination of prior for N and identifying assumptions, as long as the software uses the prior $p^w(N) \propto 1/N$. Note that our prior for N *does not* have to be $p^w(N)$. This is the case for most existing software, including the `R` package `conting` [102], which implements a reversible-jump MCMC sampler to target $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$ under a working prior $p^w(\boldsymbol{\pi})$ induced by a prior that averages over all hierarchical log-linear models [68], and the `R` package `LCMCR`, which implements a data augementation Gibbs sampler to target $p^w(\boldsymbol{\pi} \mid \boldsymbol{n})$ under a working prior $p^w(\boldsymbol{\pi})$ induced by a Dirichlet process prior for LCMs [87]. The steps of the MCMC samplers used in these packages are model specific and we would not be able to use them if we tried to create bespoke MCMC samplers targeting the marginal posterior in (B.5). We note that this approach is closely related to the working prior approach of [78], with some necessary modifications specific to MSE.

Recommended Priors for the Population Size, N

In Table B.1 we catalog $p(n \mid \pi_0) = \sum_{N=n}^{\infty} L_1(N, \pi_0 \mid n)p(N)$ and $p(N \mid n, \pi_0) = L_1(N, \pi_0 \mid n)p(N)/p(n \mid \pi_0)$ under Poisson, negative-binomial, and binomial priors for N, in addition to the class of priors $p(N) \propto (N-\ell)!/N!$, where $\ell \in \{0, 1, 2, \cdots\}$, suggested by [46]. This class of priors contains both the improper uniform prior, $p(N) \propto 1$, when $\ell = 0$, and the improper scale prior, $p(N) \propto 1/N$, when $\ell = 1$. If $p(n \mid \pi_0)$ is not available analytically, for example when $p(N)$ is beta-binomial, we recommend truncating the prior for N to the range $\{1, \cdots, N_{max}\}$ where N_{max} is an upper bound on the population size, in which case $p(n \mid \pi_0)$ can be computed numerically.

Table B.1: Catalog of $p(N \mid n, \pi_0)$ and $p(n \mid \pi_0)$ under common priors for N.

Prior	$p(N)$	$p(N \mid n, \pi_0)$	$p(n \mid \pi_0)$
$\textsc{Pois}(M)$	$(M)^N e^{-M}/N!$	$n + \textsc{Pois}(\pi_0 M)$	$\textsc{Pois}((1-\pi_0)M)$
$\text{NB}\left(a, \frac{M}{M+a}\right)$	$\binom{N+a-1}{N}(\frac{M}{M+a})^N(\frac{a}{M+a})^a$	$n + \text{NB}\left(n+a, \frac{M\pi_0}{M+a}\right)$	$\text{NB}\left(a, \frac{(1-\pi_0)M}{(1-\pi_0)M+a}\right)$
$\textsc{Bin}(M, q)$	$\binom{M}{N} q^N (1-q)^{M-N}$	$n + \textsc{Bin}\left(M-n, \frac{\pi_0 q}{\pi_0 q + 1 - q}\right)$	$\textsc{Bin}(M, (1-\pi_0)q)$
[46]	$\propto (N-\ell)!/N!$	$n + \text{NB}(n-\ell+1, \pi_0)$	$\propto \frac{(n-\ell)!}{n!}(1-\pi_0)^{\ell-1}$

The improper scale prior, under which $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) \propto p_C(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) I(\tilde{\boldsymbol{\pi}} \in \tilde{S})$, is a common "noninformative" prior for N and has the nice property that the posterior mean of N conditional on $\tilde{\boldsymbol{\pi}}$ is the Horvitz-Thompson estimator [63], $n/\{1 - \mathcal{T}(\tilde{\boldsymbol{\pi}})\}$, which is well understood in the present context [see e.g. 113]. Recall that the Horvitz-Thompson estimator also arose when considering frequentist inference in Appendix B.2.1. Following [81], we recommend using this prior in the absence of substantive knowledge about N.

When incorporating substantive knowledge about N into an informative prior for N we recommend using a negative-binomial or beta-binomial prior, as we have found Poisson and binomial priors to be more informative than we would usually like to use. For concreteness in Chapter 3 we focused on the negative-binomial prior. In Table B.1, we use a common

parameterization for the negative-binomial distribution in terms of the mean M and overdispersion parameter a. This parameterization arises from a Poisson-gamma mixture, where $N \mid \delta \sim \text{POISSON}(M\delta)$, $\delta \sim \text{GAMMA}(a, a)$. As $a \to \infty$ the prior approaches a Poisson prior with mean M, and as $a \to 0$ the prior approaches the improper scale prior.

B.2.3 Regularization and Data Sparsity

As discussed in Section 3.1 of Chapter 3, when one uses a model that places little to no restrictions on the observed data distribution (as we advocate for in Section 2.6 of Chapter 3), this can lead to population size estimates with large variances associated with them. This typically occurs when the data is sparse, i.e. when some cells of the observed contingency table are small (or even 0). Data sparsity can be a problem when conducting frequentist analyses, as the standard asymptotic arguments used in Appendix B.2.1 to derive standard errors and confidence intervals are generally not valid. This issue is secondary to our main focus of choosing the identifying assumption, in the sense that the amount of sparsity in the data should not affect the choice of identifying assumption.

We discuss here two possible routes to reduce the variance of population size estimators. The first route is to place restrictions on the observed data distribution, as advocated for by the quote of [45] in Section 3.1 of Chapter 3. This would require the restricted model to truly hold, otherwise the lower estimated variance would not be valid and the population size estimate could be arbitrarily biased. We would generally prefer not to take this route, as such restrictions are typically hard to justify in practice [see e.g. 32, 139]. Further, even if one places correct restrictions on the observed data distribution, in a frequentist analysis the standard errors and confidence intervals derived in Appendix B.2.1 can still be invalid when the data are sparse.

The second route is to use some form of regularization when estimating the observed cell probabilities within a model that places little to no restrictions on the observed data distribution. Regularization reduces the variances of estimates, at the cost of increasing the bias of estimates, by shrinking parameter estimates to a predetermined subset of parameter

space. We now briefly discuss how regularization can be incorporated into frequentist or Bayesian analyses:

- In a frequentist analysis, regularization can be incorporated through some form of penalized likelihood [53], where instead of estimating the observed cell probabilities $\tilde{\boldsymbol{\pi}}$ by maximizing the conditional likelihood as described in Appendix B.2.1, one would maximize the sum of the conditional likelihood and a penalty term

$$\hat{\boldsymbol{\pi}} = \arg\max_{\tilde{\boldsymbol{\pi}} \in \tilde{S}} L_2(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{n}) - cJ(\tilde{\boldsymbol{\pi}}). \tag{B.6}$$

  Here J is a penalty function and $c > 0$ is a regularization parameter. When $c = 0$ the estimate corresponds to the conditional maximum likelihood estimate, and as c increases the estimate gets shrunk to some subset of $\tilde{S}$ defined by the penalty function J. It will typically be feasible to obtain estimates by solving B.6. However, deriving standard errors and confidence intervals for these estimates can be difficult, especially when the data is sparse. Nonstandard asymptotic theory may be required [see e.g. 98].

- In a Bayesian analysis, regularization is inherent due to the prior distribution for $\tilde{\boldsymbol{\pi}}$. Here the prior serves a similar purpose to the penalty function J in a frequentist analysis, defining the subset of $\tilde{S}$ to which estimates of $\tilde{\boldsymbol{\pi}}$ are shrunk. Note that the computational techniques in Appendix B.2.2 are still valid even when the data are sparse.

We note that there is a common difficulty associated with regularizing estimates in a frequentist or Bayesian analysis: choosing where to shrink estimates of the observed cell probabilities, $\tilde{\boldsymbol{\pi}}$; i.e. choosing the penalty function, J, in a frequentist analysis or the prior in a Bayesian analysis. A fruitful direction for future research is to understand what are choices of penalty functions or priors that produce population size estimates with desirable properties when the data are sparse (e.g. good frequentist performance).

B.3 Identifying Assumption Derivations

The purpose of this appendix is to derive the identifying assumptions associated with no-highest-order interaction assumption and the K'-list marginal no-highest-order interaction assumption.

B.3.1 Derivation for No-Highest-Order Interaction Assumption

Recall from Section 3.1 of Chapter 3 that we have the following relationship between the cell probabilities and the highest order interaction, $\lambda_{\mathbf{1}}$: $\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}/\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = \exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\}$, where $I_{odd}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is odd})$ and $I_{even}(\boldsymbol{h}) = I(\sum_{k=1}^{K} h_k \text{ is even})$. Suppose we fix $\lambda_{\mathbf{1}} \in \mathbb{R}$, or equivalently $\xi = \exp\{(-1)^{K+1}\lambda_{\mathbf{1}}\} \in \mathbb{R}^+$. Under this assumption we have that $\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}/\prod_{\boldsymbol{h}\in H}\pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = \xi$. Multiplying the left-hand side by $1 = \left(\frac{1-\pi_0}{1-\pi_0}\right)^{2^{K-1}}$, we find that $\tilde{\Pi}_{odd}/\{[\pi_0/(1-\pi_0)]\tilde{\Pi}_{even}\} = \xi$, where $\tilde{\Pi}_{odd} = \prod_{\boldsymbol{h}\in H^*}\tilde{\pi}_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})}$ and $\tilde{\Pi}_{even} = \prod_{\boldsymbol{h}\in H^*}\tilde{\pi}_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})}$. Rearranging terms and solving for π_0, we find that the assumption that ξ is a fixed value corresponds to the explicit functional relationship

$$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = \frac{\tilde{\Pi}_{odd}/\tilde{\Pi}_{even}}{\xi + \tilde{\Pi}_{odd}/\tilde{\Pi}_{even}}. \tag{B.7}$$

The identifying assumption corresponding to the no-highest-order interaction assumption is recovered by setting $\lambda_{\mathbf{1}} = 0$, or equivalently $\xi = 1$: $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even})/(1 + \tilde{\Pi}_{odd}/\tilde{\Pi}_{even})$. The observed-data distribution is not restricted by the assumption that the highest-order interaction is fixed, and thus models that use this assumption without any extra assumptions regarding the observed cell probabilities are nonparametric identified.

B.3.2 Derivation for K′-list Marginal No-Highest-Order Interaction Assumption

Suppose we assume that $\prod_{\boldsymbol{g}\in G}\pi_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}/\prod_{\boldsymbol{g}\in G}\pi_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})} = \xi$, where $\xi \in \mathbb{R}^+$ is fixed. Multiplying the left-hand side by $1 = \left(\frac{1-\pi_0}{1-\pi_0}\right)^{2^{K'-1}}$, we find that $\tilde{\Pi}_{odd,+}/\{[\pi_0/(1-\pi_0) + \tilde{\pi}_{0+}]\tilde{\Pi}_{even,+}\} = \xi$, where $\tilde{\Pi}_{odd,+} = \prod_{\boldsymbol{g}\in G^*}\tilde{\pi}_{\boldsymbol{g}+}^{I_{odd}(\boldsymbol{g})}$ and $\tilde{\Pi}_{even,+} = \prod_{\boldsymbol{g}\in G^*}\tilde{\pi}_{\boldsymbol{g}+}^{I_{even}(\boldsymbol{g})}$. Rearranging terms and solving for π_0, we find that the assumption that ξ is a fixed value corresponds to the explicit

functional relationship

$$\mathcal{T}(\tilde{\boldsymbol{\pi}}) = \frac{\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+}}{\xi + (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \xi\tilde{\pi}_{0+})}. \tag{B.8}$$

The identifying assumption corresponding to the K'-list marginal no-highest-order interaction assumption is recovered by setting $\xi = 1$: $\mathcal{T}(\tilde{\boldsymbol{\pi}}) = (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})/(1 + \tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})$.

As noted in Section 4.3 of Chapter 3, the the K'-list marginal no-highest-order interaction assumption does not imply that there is no highest-order interaction for all K lists, as $\prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{odd}(\boldsymbol{h})} / \prod_{\boldsymbol{h}\in H} \pi_{\boldsymbol{h}}^{I_{even}(\boldsymbol{h})} = (\tilde{\Pi}_{odd}/\tilde{\Pi}_{even}) \times (\tilde{\Pi}_{odd,+}/\tilde{\Pi}_{even,+} - \tilde{\pi}_{0+})^{-1} \neq 1$ in general.

B.4 Latent Class Model Simulations

The purpose of this appendix is conduct simulation studies demonstrating the practical implications of Theorem B.2. In particular, we present a variety of simulations exploring the frequentist properties of the Bayesian LCM of [87]. In each example we generate 200 data sets from the model in (B.1) for a given number of lists K and a fixed parameter setting of $\theta \in \Omega_{\mathcal{Q}_J}$, i.e. a fixed population size N and a J-class LCM $Q \in \mathcal{Q}_J$. For all examples we will use $N \in \{2000, 10000, 100000\}$. For each simulated data set, we fit the Bayesian LCM of [87] as implemented in the `R` package `LCMCR`, using J latent classes (i.e. the same number that generated the data) and the default prior for $\boldsymbol{\nu}$, by running the Gibbs sampler implemented in `LCMCR` for $250,000$ iterations, with the first $50,000$ tossed for burn-in. We note that `LCMCR` uses the improper scale prior for N, i.e. $p(N) \propto 1/N$, and a flat prior for $\boldsymbol{q}$, i.e. $q_{jk} \overset{i.i.d.}{\sim} \text{UNIF}(0,1)$, which can not be changed. For each parameter setting of $\theta \in \Omega_{\mathcal{Q}_J}$ we examine the frequentist performance of the posterior median, 95% credible interval, and 50% credible interval for estimating the unobserved cell probability, π_0, through the sample mean of the posterior medians, the sample coverage of the 95% credible intervals, the sample mean of the 95% credible interval widths over the 200 replications, the sample coverage of the 50% credible intervals, and the sample mean of the 50% credible interval widths over the 200 replications.

B.4.1 Example 1

In this example we consider data from $K = 2$ lists generated from the two-class LCM Q_{1a} with parameters given in Table B.2. Under Q_{1a}, $\tilde{\pi}_{Q_{1a},(0,1)} = 0.276$, $\tilde{\pi}_{Q_{1a},(1,0)} = 0.276$, $\tilde{\pi}_{Q_{1a},(1,1)} = 0.448$, and $\pi_{Q_{1a},0} = 0.316$. There exists another two-class LCM Q_{1b}, with parameters given in Table B.2, such that $\tilde{\boldsymbol{\pi}}_{Q_{1a}} = \tilde{\boldsymbol{\pi}}_{Q_{1b}}$ but $\pi_{Q_{1b},0} = 0.219$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 2$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. This example was constructed using the counterexample used to prove Theorem B.2.

Table B.2: Parameters of two latent class models, Q_{1a} and Q_{1b} (rounded for presentation)

	ν_1	ν_2	q_{11}	q_{12}	q_{21}	q_{22}
Q_{1a}	0.500	0.500	0.248	0.248	0.743	0.743
Q_{1b}	0.857	0.143	0.495	0.495	0.990	0.990

The results of the simulation using data generated using the LCM Q_{1a} are presented in Table B.3. We see that the posterior median has a negative bias that does not vanish as N increases. One may have thought that the posterior median might possibly be a good estimator for $\pi_{Q_{1b},0} = 0.219$ since Q_{1a} and Q_{1b} induce the same observed-data distribution. However, the posterior median is also negatively biased for estimating $\pi_{Q_{1b},0}$, which suggests there are other LCMs in $\mathcal{Q}_2$ that induce very similar observed-data distributions to Q_{1a} and Q_{1b} but with different induced unobserved cell probabilities. While the 95% credible interval has nominal coverage when $N = 2000$, as N increases, coverage decreases and is no longer nominal. The 50% credible interval have essentially 0 coverage for settings of N, even for $N = 2000$ where the 95% credible interval has nominal coverage. This suggests the 95% credible interval only has nominal coverage at $N = 2000$ due to wide tails of the posterior for N.

Table B.3: Results of the simulation study where data was generated from the two-class latent class model Q_{1a}. Truth is $\pi_{Q_{1a},0} = 0.316$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.148	0.955	0.332	0.000	0.029
10000	0.146	0.730	0.316	0.000	0.023
100000	0.151	0.265	0.167	0.055	0.037

B.4.2 Example 2

One may object to the practicality of Example 1, as it examined a two class LCM constructed using the counterexample from the proof of Theorem B.2, and is thus an M_h LCM. So we now consider the following example. [87] presented a simulation study with $K = 5$ lists where data was generated from a LCM with $J = 2$ classes, which we reproduce in Table B.4. The parameters of this LCM were based on a hypothetical population where a small proportion of people have a high probability of being observed, and a large proportion of people have a small probability of being observed, which is plausible in some human rights applications.

Table B.4: Parameters of latent class model which generated data in simulation of [87].

		Sampling probabilities, $\boldsymbol{q}$				
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4	List 5
1	0.900	0.033	0.033	0.099	0.132	0.033
2	0.100	0.660	0.825	0.759	0.990	0.693

Suppose we only observed lists three and four, so that we have data from $K = 2$ lists generated from the two-class LCM Q_2 with parameters given in Table B.5. Under Q_2,

$\pi_{Q_2,0} = 0.704$. Just as in the previous example, because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 2$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_2 are presented in Table B.6. We see that the posterior median has a large negative bias that does not vanish as N increases, while the mean 95% and 50% credible interval widths decrease as N increases. Further, the 95% and 50% credible intervals have essentially 0 coverage across all N.

Table B.5: Parameters of latent class model Q_2

ν_1	ν_2	q_{11}	q_{12}	q_{21}	q_{22}
0.900	0.100	0.099	0.132	0.759	0.990

Table B.6: Results of the simulation study where data was generated from the two-class latent class model Q_2. Truth is $\pi_{Q_2,0} = 0.704$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.285	0.000	0.408	0.000	0.055
10000	0.283	0.010	0.401	0.000	0.036
100000	0.285	0.030	0.256	0.000	0.035

B.4.3 Example 3

In this example we present two more frequentist simulation studies based on only observing a subset of the five lists from the simulation of [87].

First suppose that we only observe lists two, three, and four from the simulation of [87], so that we have data from $K = 3$ lists generated from the two-class LCM Q_{3a} with parameters

given in Table B.7. Under Q_{3a}, $\pi_{Q_{3a},0} = 0.681$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is not conditionally identified when $K = 3$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{3a} are presented in Table B.8. We see that the posterior median has a slight negative bias that becomes negligible as N increases. The 95% credible intervals have over-coverage across the different settings of N. The 50% credible intervals have nominal coverage when $N = 2000$, but have over-coverage as N increases.

Table B.7: Parameters of latent class model Q_{3a}

		Sampling probabilities, $\boldsymbol{q}$		
Class	$\boldsymbol{\nu}$	List 2	List 3	List 4
1	0.900	0.033	0.099	0.132
2	0.100	0.825	0.759	0.990

Table B.8: Results of the simulation study where data was generated from the two-class latent class model Q_{3a}. Truth is $\pi_{Q_{3a},0} = 0.681$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.622	1.000	0.339	0.510	0.120
10000	0.667	1.000	0.274	0.800	0.091
100000	0.682	1.000	0.209	0.965	0.074

Suppose now we only observe lists two, three, four, and five from the simulation of [87], so that we have data from $K = 4$ lists generated from the two-class LCM Q_{3b} with parameters given in Table B.9. Under Q_{3b}, $\pi_{Q_{3b},0} = 0.658$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ is conditionally identified when

$K = 4$, we know that, since $\mathcal{P}_{\Omega_{\mathcal{Q}_2}}$ contains the true data generating model, we can consistently estimate the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{3a} are presented in Table B.10. We see that the posterior median has a negative bias that becomes negligible as N increases, as expected. The 95% and 50% credible intervals have slight under-coverage when $N = 2000$, which becomes nominal as N increases.

Table B.9: Parameters of latent class model Q_{3b}

		Sampling probabilities, $\boldsymbol{q}$			
Class	$\boldsymbol{\nu}$	List 2	List 3	List 4	List 5
1	0.900	0.033	0.099	0.132	0.033
2	0.100	0.825	0.759	0.990	0.693

Table B.10: Results of the simulation study where data was generated from the two-class latent class model Q_{3b}. Truth is $\pi_{Q_{3b},0} = 0.658$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.631	0.915	0.190	0.445	0.065
10000	0.653	0.940	0.089	0.505	0.031
100000	0.658	0.955	0.028	0.485	0.010

B.4.4 Example 4

In this example we present three more frequentist simulation studies based on adding a third class to the LCM from the simulation study of [87], representing a small proportion of the population having a probability of being observed somewhere between the other two classes. The parameters of this new LCM are given in Table B.11.

Table B.11: Parameters of latent class model which generated data in simulation of [87], with a third class added.

		Sampling probabilities, $\boldsymbol{q}$				
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4	List 5
1	0.700	0.033	0.033	0.099	0.132	0.033
2	0.200	0.275	0.250	0.200	0.300	0.325
3	0.100	0.660	0.825	0.759	0.990	0.693

First suppose that we only observe lists two, three, and four from the LCM in Table B.11, so that we have data from $K = 3$ lists generated from the three-class LCM Q_{4a} with parameters given in Table B.12. Under Q_{4a}, $\pi_{Q_{4a},0} = 0.613$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 3$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{4a} are presented in Table B.13. We see that the posterior median has a negative bias that does not vanish as N increases. The 95% credible intervals have over-coverage across the different settings of N, while the 50% credible intervals have under-coverage across the different settings of N. Similar to Example 1 in Section B.4.1, this suggests the 95% credible interval only has over-coverage due to wide tails of the posterior for N.

Next suppose that we only observe lists two, three, four, and five from the LCM in Table B.11, so that we have data from $K = 4$ lists generated from the three-class LCM Q_{4b} with parameters given in Table B.10. Under Q_{4b}, $\pi_{Q_{4b},0} = 0.569$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 4$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data

Table B.12: Parameters of latent class model Q_{4a}

		Sampling probabilities, $\boldsymbol{q}$		
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3
1	0.700	0.033	0.099	0.132
2	0.200	0.250	0.200	0.300
3	0.100	0.825	0.759	0.990

Table B.13: Results of the simulation study where data was generated from the two-class latent class model Q_{4a}. Truth is $\pi_{Q_{4a},0} = 0.613$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.524	1.000	0.387	0.210	0.119
10000	0.537	1.000	0.364	0.150	0.102
100000	0.538	1.000	0.323	0.175	0.096

generated using the LCM Q_{4b} are presented in Table B.15. We see that the posterior median has a negative bias that decreases as N increases. While the 95% and 50% credible intervals do not have nominal coverage, coverage improves as N increases (but is still far from nominal even when $N = 100000$).

Next suppose that we observe all five lists from the LCM in Table B.11, so that we have data from $K = 5$ lists generated from the three-class LCM which we will refer to as Q_{4c}. Under Q_{4c}, $\pi_{Q_{4c},0} = 0.536$. Because $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$ is not conditionally identified when $K = 5$, if we try to perform estimation within $\mathcal{P}_{\Omega_{\mathcal{Q}_3}}$, which contains the true data generating model, there is no guarantee that we can estimate well the cell probabilities and population size which generated the data. The results of the simulation using data generated using the LCM Q_{4c} are presented in Table B.16. We see that the posterior median has a negative bias that

Table B.14: Parameters of latent class model Q_{4b}

		Sampling probabilities, $\boldsymbol{q}$			
Class	$\boldsymbol{\nu}$	List 1	List 2	List 3	List 4
1	0.700	0.033	0.099	0.132	0.033
2	0.200	0.250	0.200	0.300	0.325
3	0.100	0.825	0.759	0.990	0.693

Table B.15: Results of the simulation study where data was generated from the two-class latent class model Q_{4b}. Truth is $\pi_{Q_{4b},0} = 0.569$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.469	0.525	0.199	0.090	0.065
10000	0.509	0.630	0.128	0.120	0.041
100000	0.519	0.695	0.066	0.290	0.023

decreases as N increases. While the 95% and 50% credible intervals do not have nominal coverage, coverage improves as N increases.

B.4.5 Takeaways

When using the model $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ for multiple-systems estimation, one is relying on the assumption that the data was generated from a distribution in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. If a practitioner is comfortable with the assumption that $2J \leq K$, then we know the model is conditionally identified, and thus this assumption is a combination of an explicit identifying assumption (which is currently unknown) and possibly some restrictions on the observed-data distribution. Due to conditional identification, the practitioners have guarantees under this assumption that they can estimate the population size, and other parameters, well if their observed sample

Table B.16: Results of the simulation study where data was generated from the two-class latent class model Q_{4c}. Truth is $\pi_{Q_{4c},0} = 0.536$.

N	Mean Posterior Median	95% CI Coverage	Mean 95% CI Width	50% CI Coverage	Mean 50% CI Width
2000	0.435	0.415	0.169	0.050	0.055
10000	0.500	0.875	0.132	0.370	0.044
100000	0.523	0.895	0.053	0.490	0.018

size n is large enough. However, if a practitioner is not comfortable with this assumption, and chooses to use $J > K/2$, they have no such guarantees as they are using a model that is not conditionally identified.

Through the four example simulation studies in this appendix we saw examples where models that were not conditionally identified had good frequentist performance (Q_{3a}, Q_{4a}) and bad frequentist performance (Q_1, Q_2, Q_{4b}, Q_{4c}) according to some of our simulation summary measures. The good and bad frequentist performances could have been due to

- where the prior of [87] places mass in the parameter space $\Omega_{\mathcal{Q}_J}$ (e.g. good frequentist performance if it places enough prior mass around the true data generating parameters),
- whether there actually exists other LCMs in $\mathcal{Q}_J$ that induce similar observed cell probabilities to the true data generating parameters but a different unobserved cell probability (e.g. good frequentist performance if other LCMs do not exist with these properties),
- or some combination of the two previous factors.

We currently have no way to tease apart these factors and tell when a model that is not conditionally identified will have good or bad performance. This is a problem for using these models in practice, as we have no way to tell practitioners "under these assumptions the model will perform well".

We believe there are two routes forward to combat this problem, if one wants to use LCMs for multiple-systems estimation. The first option is to further study technical results for conditional identification in LCMs. For example, as we discussed in Section B.1.6, suppose we can prove under further (practically relevant) restrictions on $\mathcal{Q}_J$ that $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ is conditionally identified for some $J > K/2$. We would then be able to expand the range of models we could fit under which we had guarantees that we could estimate well the parameters of the model.

The other option is to study LCMs through the framework of partial identification [130, 54], which was recently used in multiple-systems estimation by [128] for frequentist inference for partially-identified log-linear models. This would require both: 1) a better technical understanding of what parameters, or functions of parameters, of LCMs are not identified, and 2) placing substantively meaningful priors on the non-identified parameters (i.e. priors informed by substantive knowledge concerning the population of interest and how the data was collected) if a Bayesian approach is taken. Without 1), the best we can do in a Bayesian approach is to place substantively meaningful priors on all LCM parameters, i.e. on $\Omega_{\mathcal{Q}_J}$. The prior for $\Omega_{\mathcal{Q}_J}$ of [87] is based on the Dirichlet Process prior specification of [35], which is a prior of technical convenience. Specifying a substantively meaningful prior for $\Omega_{\mathcal{Q}_J}$ would require being able to specify a prior for the class membership probabilities $\boldsymbol{\nu}$ and for the class specific observation probabilities $\boldsymbol{q}$. It is difficult to imagine a scenario in which a practicioner would have knowledge of the population of interest and how the data was collected that could be incorporated into priors for all $J(K+1)-1$ of the parameters ($\boldsymbol{\nu}$ and $\boldsymbol{q}$).

While we do not believe that latent class models cannot be used for multiple-systems estimation (see our application in Section 5 of Chapter 3 where we use the LCM prior of [87] to induce a prior for the observed cell probabilities $\tilde{\boldsymbol{\pi}}$), we do believe that there needs to be further research to understand under what assumptions LCMs do and do not perform well in practice. We discuss one further area of research before concluding this section. The start of this section began by assuming that a practitioner assumed their data was generated by a distribution in $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$. It is not clear to the authors how in practice one would choose a

specific value of J. In practice, how would a practitioner choose between $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ and $\mathcal{P}_{\Omega_{\mathcal{Q}_{J'}}}$ for $J \neq J'$? What characteristics of the population being studied and the data collection process would allow one to differentiate between these two models? Research into understanding how to elicit plausible values of J would help to justify the use of the model $\mathcal{P}_{\Omega_{\mathcal{Q}_J}}$ in practice.

B.5 Kosovo Analysis Appendix

This appendix serves three purposes: 1) to describe the difficulty in justifying the NHOI assumption for the Kosovo data, 2) to describe a prior sensitivity analysis for the Bayesian analyses of the Kosovo data, and 3) to describe a sensitivity analysis for the Kosovo data probing the NHOI assumption.

B.5.1 The No-Highest-Order Interaction Assumption

The Kosovo data set has $K = 4$ lists, which we will order (without loss of generality) so that the American Bar Association Central and East European Law Initiative (ABA) list is first, the Human Rights Watch (HRW) list is second, the Organization for Security and Cooperation in Europe (OSCE) list is third, and the list constructed from exhumation reports conducted on behalf of the International Criminal Tribunal for the Former Yugoslavia (EXH) is fourth. Let $\text{Odds}(h_1 = 1 \mid h_2 = 1, h_3, h_4) = \pi_{(1,1,h_3,h_4)}/\pi_{(0,1,h_3,h_4)}$ denote the odds that an individual is observed in list 1, conditional on being observed in list 2 and the inclusion patterns h_3, h_4 for lists 3 and 4. For example, if $h_3 = 0$ and $h_4 = 1$, $\text{Odds}(h_1 = 1 \mid h_2 = 1, h_3 = 0, h_4 = 1)$ is the odds that an individual is observed in list 1, conditional on being observed in lists 2 and 4 and not being observed in list 3. Similarly let $\text{Odds}(h_1 = 1 \mid h_2 = 0, h_3, h_4) = \pi_{(1,0,h_3,h_4)}/\pi_{(0,0,h_3,h_4)}$ denote the odds that an individual is observed in list 1, conditional on not being observed in list 2 and the inclusion patterns h_3, h_4 for lists 3 and 4. We can then define $\text{OR}(h_3, h_4) = \text{Odds}(h_1 = 1 \mid h_2 = 1, h_3, h_4)/\text{Odds}(h_1 = 1 \mid h_2 = 0, h_3, h_4)$ as the odds ratio for lists 1 and 2, conditional on the inclusion patterns h_3, h_4 for lists 3 and 4. Following Section 4.1 of Chapter 3, the no-highest-order interaction assumption assumes that $\text{OR}(1, 0)/\text{OR}(0, 0) = \text{OR}(1, 1)/\text{OR}(0, 1)$, i.e. the highest-order interaction for the first

three lists, conditional on not being observed in list 4, $\text{OR}(1,0)/\text{OR}(0,0)$, is equal to the highest-order interaction for the first three lists, conditional on being observed in list 4, $\text{OR}(1,1)/\text{OR}(0,1)$.

This assumption is obscure and hard to justify based on our knowledge of how the four lists were generated. As the validity of our analysis rests on this assumption being correct, we stress that we are not confident that this assumption holds, and thus we are not confident in the validity of the analysis of the Kosovo data set using the NHOI assumption.

B.5.2 Prior Sensitivity Analyses

In this section we perform prior sensitivity analyses for the Bayesian analyses of the Kosovo data from Chapter 3. For N, we will consider the negative-binomial prior specification described in Chapter 3, in addition to the improper scale prior discussed in Appendix B.2.2. For the observed cell probabilities $\tilde{\boldsymbol{\pi}}$, we will consider four prior specifications : 1) the prior induced from using the Dirichlet process prior of [87] for the J class LCM $\Omega_{LCM,J}$, with $J = 10$ and default hyperparameters, as implemented in the `R` package `LCMCR` (i.e. the prior used in the main analyses), 2) a flat Dirichlet prior, i.e. $\tilde{\boldsymbol{\pi}} \sim \text{DIRICHLET}(1, \cdots, 1)$, 3) the prior induced from using $\text{NORMAL}(0, 5^2)$ priors for the log-linear parameters in the saturated log-linear model Ω_{LL}, fit using the `Stan` probabilistic programming language [26], and 4) the prior induced from using the Bayesian model averaging prior of [68] for the log-linear parameters in the saturated log-linear model Ω_{LL}, with the unit information prior on log-linear parameters, as implemented in the `R` package `conting` [102]. We note that `conting` uses an alternative log-linear parameterization based on sum to zero constraints rather than corner point constraints used in Section 3.1 of Chapter 3. For each combination of identifying assumption and priors for N and $\tilde{\boldsymbol{\pi}}$ we fit the corresponding model using the computational approach described in Appendix B.2.2.

In Table B.17 we present posterior means and 95% credible intervals for N under each prior combination under the 2-list marginal NHOI assumption, i.e. assuming marginal independence of the ABA and HRW lists. The posterior density for N under each prior

combination under the 2-list marginal NHOI assumption is displayed in Figure B.1. For each prior for $\tilde{\boldsymbol{\pi}}$, the posterior for N does not appear to be sensitive to the prior for N, as the point estimates and credible intervals are essentially the same between the two priors for N. Across the different priors for $\tilde{\boldsymbol{\pi}}$, the posterior summaries are fairly consistent, with the posterior summaries under the LCM prior for $\tilde{\boldsymbol{\pi}}$ being slightly lower than under the other priors. We note that all of the credible intervals fall within the confidence interval of [123].

Table B.17: Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption.

	Improper Scale Prior	Negative-Binomial
Conting	9618 [8224, 11195]	9621 [8232, 11191]
Dirichlet	9536 [8113, 11252]	9540 [8123, 11247]
LCMCR	9353 [7959, 11063]	9359 [7967, 11059]
Log-Linear	9764 [8277, 11549]	9766 [8288, 11550]

In Table B.18 we present posterior means and 95% credible intervals for N under each prior combination under the NHOI assumption. The posterior density for N under each prior combination under the NHOI assumption is displayed in Figure B.2. For each prior for $\tilde{\boldsymbol{\pi}}$, the posterior for N is somewhat sensitive to the prior for N, as the posterior mean and credible interval limits are always larger under the improper scale prior compared to the negative-binomial prior for N. The posterior for N appears to be the most sensitive to the prior for N under the Dirichlet and log-linear priors for $\tilde{\boldsymbol{\pi}}$, where the posterior means and upper credible interval limits increase by several thousand when using the improper scale prior for N instead of the negative-binomial prior. Across the different priors for $\tilde{\boldsymbol{\pi}}$, the posteriors corresponding to the Dirichlet prior, the log-linear model prior, and the LCM prior of [87] are in relative agreement. The posterior corresponding to the Dirichlet prior is the most diffuse of the three, and the posterior corresponding to the LCM prior of [87] is the most concentrated of the three. The posterior corresponding to the log-linear model prior

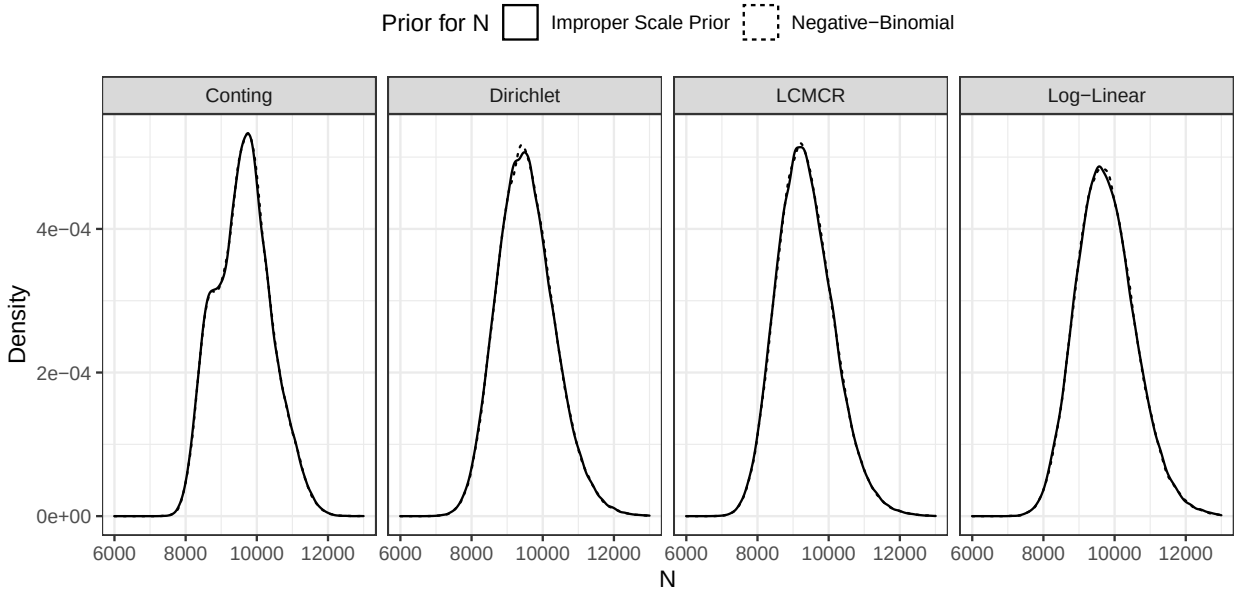


Figure B.1: Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the 2-list marginal NHOI assumption.

of [68], implemented in the `conting` package, is multimodal, which is not unexpected as it is performing Bayesian model averaging [58] over all hierarchical log-linear models. Due to this multimodality, point estimates (e.g. the posterior mean) may not be reliable summaries of the posterior distribution. We note that all of the credible intervals contain the point estimate of [123].

B.5.3 A Sensitivity Analysis Probing the NHOI Assumption

We now perform a sensitivity analysis probing the no-highest-order interaction assumption. We will consider models with the identifying assumption in Section 4.2 of Chapter 3, varying ξ over $\{1/2, 2/3, 1, 3/2, 2\}$ [following 51]. For each value of ξ, we will present both a frequentist analysis and a Bayesian analysis, with the Bayesian analysis using the same priors from the main analysis as presented in Section 5.1 of Chapter 3. This sensitivity analysis is limited in that we followed [51] and chose an arbitrary range of values for ξ around 1. Due to the

Table B.18: Posterior means and 95% credible intervals for N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption.

	Improper Scale Prior	Negative-Binomial
Conting	13000 [9202, 19971]	12694 [9175, 19299]
Dirichlet	18500 [9402, 35908]	16051 [9098, 27679]
LCMCR	14695 [9423, 23675]	14071 [9321, 21604]
Log-Linear	16209 [8731, 30025]	14719 [8579, 24878]

difficulty in interpreting the highest-order interaction when there are $K = 4$ lists, we are not able to say with confidence whether this range of values is meaningful or not. In Table B.19 we present the results from our frequentist and Bayesian analyses under each identifying assumption.

Table B.19: Point estimates and 95% uncertainty intervals for sensitivity analysis probing the NHOI assumption. For the Bayesian analysis the point estimate is the posterior mean. In this table ξ is a ratio of ratios of odds ratios, as described in Section 4.2 of Chapter 3 and Appendix B.5.1.

	$\xi = 1$ / 2	$\xi = 2$ / 3	$\xi = 1$	$\xi = 3$ / 2	$\xi = 2$
Frequentist	29483 [6210, 52757]	23212 [5757, 40668]	16941 [5304, 28579]	12761 [5002, 20520]	10670 [4851, 16490]
Bayesian	21476 [13518, 33507]	17983 [11492, 27987]	14071 [9321, 21604]	11121 [7766, 16564]	9538 [6943, 13821]

The results are not very robust to misspecification of ξ in the chosen range. The uncertainty intervals when $\xi = 1/2$ and $\xi = 2$ barely overlap. For the Bayesian analysis, the posterior mean when $\xi = 2$ is 32% lower than the posterior mean when $\xi = 1$ (i.e. under the no-highest-order interaction assumption), the posterior mean when $\xi = 1/2$ is 53% higher than the posterior mean when $\xi = 1$, and the posterior mean $\xi = 1/2$ is more than twice the posterior mean when $\xi = 2$. These differences are even more dramatic for the frequentist analysis. This lack of robustness to misspecification of ξ would be a cause for concern

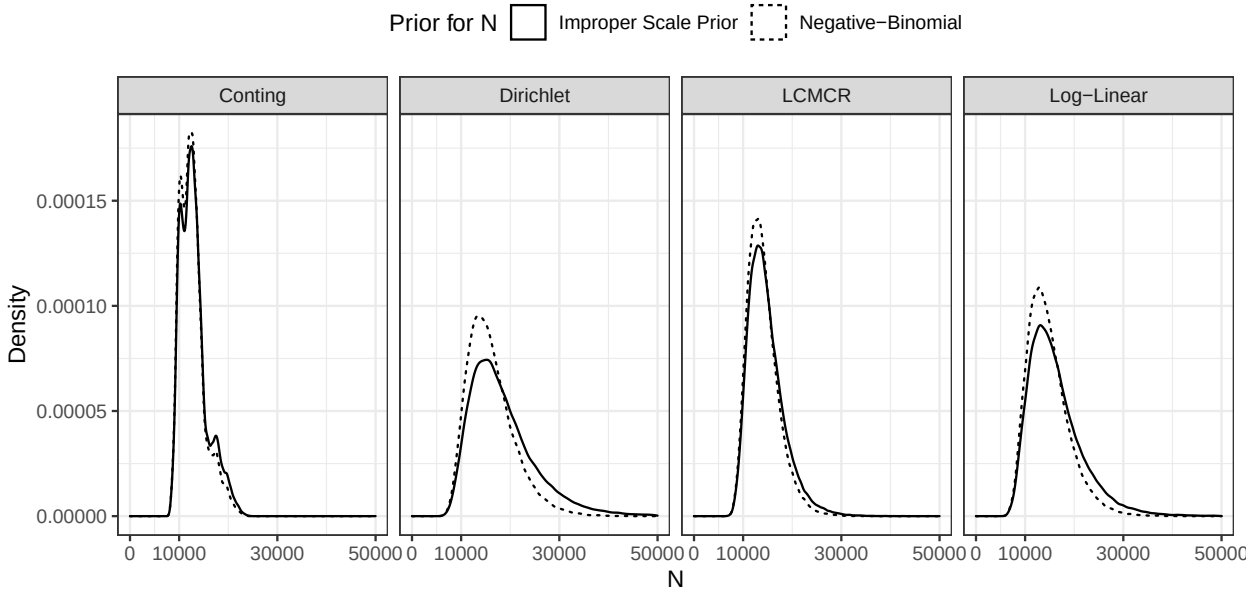


Figure B.2: Posterior density of N under each combination of prior for N and $\tilde{\boldsymbol{\pi}}$, under the NHOI assumption.

if the no-highest-order interaction assumption was plausible, and the deviations from the assumption in terms of ξ were also plausible, in the context of the Kosovo data set.