UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL SCIENCES
SCHOOL OF ECONOMIC, SOCIAL AND POLITICAL SCIENCES
SOCIAL STATISTICS AND DEMOGRAPHY

Automated probabilistic record linkage without classification for dual system population size estimation

Viktor Račinskij

Thesis for the Degree of Doctor of Philosophy
April 2024

Abstract

Population size estimation from two incomplete surveys, known as dual system estimation, requires to know which of the population elements are simultaneously captured in both of the surveys, the task imperfectly accomplished by the means of record linkage. In this thesis we explore the conceptual closeness of the fields of probabilistic record linkage and dual system estimation, and develop methods for the population size estimation, called linkage free dual system estimation, that seamlessly integrate probabilistic record linkage and dual system estimation. Unlike many existing record linkage approaches, the one developed in this thesis is purely estimation-based and does not classify records into links and non-links. It also does not require clerical resolution of possible links.

In order to theoretically justify the linkage free dual system estimation method, we revisited certain problematic aspects of probabilistic record linkage and proposed a different approach conceptualizing record linkage models. This conceptualization takes into account a very specific sampling mechanism behind record linkage tasks. It also allows analysis of certain properties and limitations of parameter estimation in linkage models. We also introduce a special case of data blocking that bridges the gap between record linkage data and estimation with these data. Special attention is paid to between-variables associations in the outcomes obtained by comparing the values of linkage variables. We also assess linkage models for identifiability using a variety of methods from the field of algebraic statistics.

We demonstrate that in situations where the data in both surveys are collected for the same geographical clusters, the linkage free dual system estimation is feasible and can yield outputs of similar quality to the regular classification approaches that involve clerical interventions. We also develop accompanying variance estimation methods, and these methods rely on less restrictive assumptions than existing methods. All developments are undertaken within the frequentist paradigm.

Keywords: dual system estimation, probabilistic record linkage, justification of probabilistic record linkage, identifiability, simulated annealing, Taylor series approximation, within and between linkage variables associations, variance estimation, census and census coverage survey, simulations, no-classification record linkage, linkage free dual system estimation.

Contents

List of tables **v**

1 Introduction **1**

1.1 Automated probabilistic record linkage without classification for dual system population size estimation . . . 1

1.2 Scope and limits of this work . . . 2

1.3 Structure of the thesis . . . 4

2 Literature review and preliminaries **5**

2.1 Capture-recapture and dual system estimation . . . 5

2.1.1 Setup . . . 6

2.1.2 Dual system estimator . . . 7

2.1.3 Variance estimation . . . 13

2.2 Record linkage . . . 13

2.2.1 Preparing for linkage . . . 14

2.2.2 Fellegi-Sunter approach . . . 16

2.2.3 Formulation of the Fellegi-Sunter approach in terms of mixture models . . . 21

2.2.4 One-to-one linkage . . . 25

2.2.5 A linkage experiment and the invalidity of mixture models for record linkage . . . 25

2.2.6 Notation in a context . . . 32

2.3 An overview of census coverage estimation . . . 35

2.4 Simulated annealing . . . 39

2.5 Identifiability . . . 40

2.5.1 Types of identifiability . . . 41

2.5.2 Number of observables and number of parameters . . . 43

2.5.3 Assessing local identifiability . . . 43

2.5.4 Tensor methods to assess identifiability . . . 44

2.5.5 Gröbner basis based methods . . . 48

2.5.6 Assessing global and generic identifiability with Gröbner basis based methods . . . 48

2.5.7 Assessing rational identifiability with Gröbner basis based methods . . . 49

3 Parameter estimation in mixture-like models **50**

3.1 Mixture-like model . . . 50

3.2 Justification of the mixture-like model for record linkage . . . 54

3.3 Some issues related to parameter estimation of a mixture-like model . . . 65

3.4 Constructing a data-conforming estimator and averaging blocking . . . 66

3.5 Parameter estimation using Markov chain Monte Carlo methods . . . 68

4 Connection between record linkage and dual system estimation **70**
4.1 Linkage free dual system estimator . . . 70
4.2 Modified linkage free dual system estimator to reflect 1-to-1 matches . . . 73
4.3 Quality measures . . . 80
4.4 Heuristics in model specification and goodness-of-fit . . . 80

5 Checking identifiability of certain linkage models with four linkage variables **81**
5.1 Linkage models and identifiability . . . 81
5.2 Checking identifiability for a selection of models . . . 83
5.2.1 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ . . . 83
5.2.2 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ . . . 85
5.2.3 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{3,4})$ . . . 88
5.2.4 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3},\gamma_4)$ . . . 89
5.2.5 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_4)$ . . . 90
5.2.6 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_{2,3},\gamma_4)$ . . . 91
5.2.7 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ . . . 92

6 Estimating the variance of the linkage free dual system estimator **93**
6.1 General approach . . . 93
6.2 Practical approach: no auxiliary data . . . 95
6.3 Practical approach: with auxiliary data . . . 96
6.4 Why not just use bootstrap? . . . 97

7 Simulation study for point estimation **100**
7.1 Aims of the simulation study . . . 100
7.2 Simulating data . . . 101
7.3 Measures of performance and tuning parameters . . . 107
7.4 Worked-out examples of no-classification estimators . . . 109
7.4.1 Between-variables independence in both sets of matches and non-matches . . . 110
7.4.2 Between-variables independence in the set of matches, association between v_1 and v_2 in the set of non-matches . . . 112
7.5 Simulations assessing practical performance . . . 113
7.5.1 Between-variables independence in both sets of matches and non-matches . . . 114
7.5.2 Between-variables independence in the set of matches, association between v_1 and v_2 in the set of non-matches . . . 119
7.5.3 Association between v_2 and v_3 in the set of matches, between-variables independence in the set of non-matches . . . 123
7.5.4 Association between v_2 and v_3 in the set of matches, association between v_1 and v_2 in the set of non-matches . . . 125
7.5.5 Between-variables independence in the set of matches, three-way association between v_1, v_2 and v_3 in the set of non-matches . . . 129

7.5.6 Conclusions . . . 132
7.6 Simulations verifying theoretical results . . . 133
7.6.1 Conclusions . . . 137
7.7 Comparing the data generated according to the linkage experiment against a parametric approach . . . 137

8 Simulation study for variance estimation **139**

9 Summary, conclusions and future work **143**

References **148**

Appendices **158**

A Primer on Gröbner basis **158**

B Extra simulation results **169**
B.1 Additional scenarios for main results . . . 169
B.2 Results with very strict acceptance thresholds for the classical approach . . . 172

C Sample code to check identifiability **187**

List of Tables

1 Probabilities of outcomes for a person e_i . . . 8
2 Probabilities of outcomes for a person e_i under the dual system estimation assumptions 8
3 Aggregate outcomes under the dual system estimation assumptions . . . 8
4 Population $\mathcal{P}$ . . . 32
5 Survey S_1 . . . 33
6 Survey S_2 . . . 33
7 Example of frequencies of the comparison patterns . . . 35
8 Between-variables independence in both the set of matches and the set of non-matches: observed data, true and estimated frequencies of comparison patterns in the set of matches and the set of non-matches . . . 111
9 Between-variables independence in both sets of matches and non-matches: outputs for the modified linkage free dual system estimator and classification-based linkage . . . 111
10 Between-variables independence in the set of matches, association between the first and second variable in the set of non-matches: observed data, true and estimated frequencies of comparison patterns in the set of matches and the set of non-matches . . . 112
11 Observed data, true and estimated frequencies of comparison patterns in the sets of matches and non-matches: outputs for the modified linkage free dual system estimator and classification-based linkage . . . 113
12 Simulated data: between-variables independence in sets of both matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ 115
13 Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$. 117
14 Single block vs averaged blocking: between-variables independence in both sets of matches and non-matches . . . 118
15 Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ . . . 120
16 Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ . . . 121
17 Single block vs averaged blocking: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches . . . 122
18 Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ . . . 124
19 Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ . . . 125

20 Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ 126
21 Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ 127
22 Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ 128
23 Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ 129
24 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3},\gamma_4)$ 130
25 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_{2,3},\gamma_4)$ 131
26 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ 132
27 Population attributes as in the main simulations, Section 7.5 134
28 One binary population attribute 135
29 Population attributes with excessively many uniformly distributed values 136
30 Standard deviations of the simulated data cells vs parametric approach with within-variables independence: $\tau = 500, \pi_1 = \pi_2 = 0.9$ 138
31 Variance estimation: between-variables independence in both the set of matches and the set of non-matches 140
32 Variance estimation: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches 141
33 Variance estimation: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches 142
34 Variance estimation: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches . . 143
B1 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_4)$ 169
B2 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{2,3},\gamma_4)$ 170

B3 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ 171
B4 Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 172
B5 Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$. 173
B6 Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ 174
B7 Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 175
B8 Simulated data: between-variables independence in the set of matches, association between the second and third variable in the set of matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 176
B9 Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 177
B10 Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ 178
B11 Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ 179
B12 Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 180
B13 Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 181
B14 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$ 182
B15 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$ 183
B16 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_4)$ 184

B17 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{2,3}, \gamma_4)$ 185
B18 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ 186
B19 Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ 187

Declaration of authorship

I, Viktor Račinskij, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Title of thesis: Automated probabilistic record linkage without classification for dual system population size estimation

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Račinskij V., Smith P. A. & van der Heijden P. G. M. (2019). Linkage free dual system estimation. arXiv:1903.10894.

Signed:
Date:

Acknowledgements

I would like to thank my supervisors, Professor Paul A. Smith and Professor Peter G. M. van der Heijden, for guiding me patiently through the chaos of ideas for such a long time.

I am grateful to the Office for National Statistics for funding this research project. I am especially grateful to Marie Cruddas for having motivation and strength to organize this studentship. Owen Abbott, Jane Naylor and Peter Brodie were all very helpful in dealing with administrative hurdles; thank you for your help.

Certain important ideas presented in this thesis were shaped by discussions with Professor Li-Chun Zhang, Professor Dankmar Böhning and Professor Peter W. F. Smith during my progression reviews. I appreciate your insights very much.

I acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Unsurprisingly, the heaviest burden of me working on this dissertation fell on my family, Olga and Jurgis. Thank you for not stopping to support me and sorry for the time not spent together.

1 Introduction

In this chapter we explain motivation for this work, its objectives and provide a high-level overview of problems we need to address along the way. We also outline the scope and limits of the ideas developed in this thesis. Finally, we present the structure of the thesis.

1.1 Automated probabilistic record linkage without classification for dual system population size estimation

The population size of a certain domain, be it a country, a district of a town or individuals with some specific demographic characteristic, is one of the most basic and frequently used statistics describing the population of interest. It is not often the case, however, that a population count gives the exact figure of the population size. This is true even for the observed population count in a census or a population register. Therefore, one must resort to population size estimation from multiple incomplete surveys of the population. The term 'survey' does not always strictly mean 'sample survey' in this thesis. Instead, it is a convenient general designation of different data and data collection types relevant to the discussed topics.

Methods for population size estimation from several incomplete surveys of the target population in the absence of a perfect sampling frame are known as capture-recapture methods. The simplest case of capture-recapture methods uses two samples to produce a population size estimate and is usually called dual system estimation. The key factor enabling dual system estimation, as well as other types of capture-recapture approaches, is the ability to determine, in the absence of unique identifiers, whether a member of the population captured in the first survey was or was not captured in the second survey. The area of record linkage comprises the theory and practice of classifying pairs, or n-tuples in general, of records from two or more sources as referring to the same unit in the target population or not. Therefore, dual system population size estimation is infeasible without record linkage and linkage is a prerequisite for estimation.

This thesis revolves around connections between record linkage and dual system estimation. There are studies, discussed in literature review chapter, of the effect of linkage errors on dual system estimates and ways to adjust for biases incurred by imperfect linkage. However, while dual system estimation is always preceded by record linkage, there was surprisingly little research until recent on how the two are related from the conceptual standpoint. In this thesis several closely interrelated tasks are pursued. First and foremost, a more holistic approach to record linkage for dual system estimation is proposed. It is shown that for a certain probabilistic record linkage model the dual system estimator is a function of one of the parameters of the linkage model. The fact that the dual system estimator is embedded in the linkage model has several implications. On the one hand, it motivates a purely estimation-focused approach to record linkage. An approach that does not have a strict classification of record pairs as links and non-links, but only high quality estimates of the linkage model parameters. The dual system estimate would be simply an outcome of the estimation of linkage model parameters without classification. On the other hand, knowing the dual system estimator is a transformation of the record linkage model parameter, allows the properties of the dual system estimator to be used to

study the properties of record linkage.

Going to the pure estimation-based approach for record linkage entails several problems and challenges on its own. The biggest one is that it effectively means no clerical review, resolution and classification of record pairs that are half-way between likely links and likely non-links. This is because of the difficulties in formalizing an estimation approach with clerical interventions. Also, it is some sort of contradiction to develop a pure estimation approach without classification and have clerical decisions at the same time. On the contrary, having a method that does not require clerical review means a substantial reduction of the cost associated with the record linkage task. Hence, the problem of fully automated probabilistic record linkage is also present and discussed in this work. The absence of clerical resolution is challenging in a number of ways. Apart from the obvious contribution to the final quality of the linkage exercise, clerical review, in combination with certain constraints, can mitigate the effect of model misspecification and non-identifiability.

Therefore, the no-classification approach to record linkage requires greater care with the linkage model compared to the linkage where clerical review is available. First, it requires a good understanding of the data generating mechanism associated with a record linkage exercise. Second, it requires a useful and well-justified representation of the data generating mechanism in the form of a statistical model and study of the properties of the model. Third, it requires understanding of how to parameterize the corresponding model given the set of linkage variables. Fourth, it requires a parameter estimation approach that reflects the peculiarities of the linkage model. Fifth, the study of identifiability of the linkage model is required. Once the estimation, rather than classification, becomes the ultimate goal of the linkage exercise, the question of measuring the related uncertainty in such estimation emerges naturally.

In this thesis, a model which is parameterized as a finite mixture model but does not conform with the sampling mechanism of a mixture is introduced and all the above tasks are addressed to enable the no-classification linkage and the corresponding no-classification dual system estimation, referred to as linkage free dual system estimation. Variance estimation for the linkage free dual system estimator is discussed and an estimation approach is developed. All theoretical developments are assessed in simulations.

1.2 Scope and limits of this work

There is a wide range of applications of dual system estimation. The method developed here is not necessarily universally applicable for any type of a population or situation. Similarly, record linkage can be seen as an entire branch not only of statistics, but also of machine learning and computer science, where it is more commonly known under the name entity resolution. This means that it is impossible to cover all aspects and ramifications of record linkage in this work. Below is a short list of the key aspects that define the scope of this work. Other limitations will be mentioned in the following chapters as the discussion evolves.

1. Only human populations in a stable modern country with individuals nested in households or addresses are considered in this work.

2. The data collection for both surveys is done in at least a reasonably controlled and well-designed manner. For instance, not necessarily perfect but decent quality sampling frames and address lists are used. There are similar, well-defined variables to be purposefully collected in both of the surveys. Ideally, data are specifically collected for the population size estimation.

3. The performance of capture-recapture methods depends greatly on the data collection design and strategies of carrying out estimation. Statistically efficient designs are often very expensive and difficult to implement in large human populations. Many existing designs employ some form of a cluster sampling and dual system population size estimation within those clusters. In other words, for a selection of geographical clusters an attempt to capture as many individuals as possible within those particular clusters is made on each capture (sampling) occasion. Dual system estimation is carried out for such individual clusters or aggregations of neighbouring clusters. These individual dual system estimates are used as an input to other estimators that allow high quality population totals to be achieved. A real example of such data collection and estimation exercises can be a decennial census of a population with a census post-enumeration survey of a sample of postcodes. A potential example could be a population register or a nearly exhaustive administrative data set, such as a health system registration, with an appropriate coverage survey. No-classification record linkage and dual system estimation provide the most benefits in large linkage and estimation exercises, where minimization of clerical resolutions results in a substantial cost reduction. It also turns out that the no-classification methods as developed in this thesis have the best performance when applied to estimate small subpopulations with imperfect, but reasonably high coverage in both surveys. This is a set-up found in census coverage estimation. Hence, we focus on applications where the no-classification methods would be most needed and where the best performance can be achieved. This means that in general we are interested in census coverage-like situations in this thesis. Specifically, we assume that data for a certain selection of geographical clusters, such as a selection of postcodes in the UK, can be collected in both surveys. We are exploring how well no-classification dual system estimation can estimate the size of such a relatively small individual subpopulation. It does not mean that the methods developed here cannot in principle be used with data collected by two surveys not purposefully designed for the coverage estimation. However, such a general design does not necessarily guarantee good performance of the no-classification methods and is not explored in this thesis.

4. High-quality data pre-processing for analysis, linkage and estimation is assumed. This includes tasks such as data coding, cleaning, standardization of the values of the linkage variables, etc. Such tasks do not constitute the discussion in this work and are only briefly reviewed.

5. Linkage variables such as full address, first name, second name, date of birth are available on both of the surveys.

6. There are several reasons to focus on linkage with exactly four linkage variables. Under the linkage model considered in this work, the three linkage variable case has only one available

parameterization, whereas the four variable case allows substantially more parameterizations. This in turn makes it possible to deal with some cases of the dependence between the comparison outcomes of certain combinations of linkage variables. Also, since there are multiple model specifications for the four variable linkage models, establishing which model is identifiable and which is not becomes important. Overall, the four variable case gives the sufficient material for developing and testing the many general ideas. Moreover, in many situations similar to census coverage estimation, the five and more variable cases are likely to result in non-identifiable models because available linkage variables often lack independence leading to multi-way between-variables associations.

1.3 Structure of the thesis

Chapter 2 is a combination of a literature review and all the preliminaries needed for the main developments of the thesis. Most of the notation used throughout the work is also introduced in this chapter. The topics discussed include: the dual system estimator; record linkage; the Fellegi–Sunter and related mixture model-based approaches to record linkage; conceptual issues of mixture model-based approaches and unfeasibility of the maximum likelihood approach; the data generating mechanism behind the record linkage exercises; overview of the context of the census coverage estimation; the simulated annealing algorithm; and approaches to establish model identifiability using algebraic statistics.

Chapter 3 introduces the mixture-like model and justifies the use of such a model in record linkage. Some consequences of the above modelling on properties of the related estimators are discussed. An outline of constructing an estimator that agrees well with the mixture-like conceptualization is given. A special case of blocking, called averaging blocking, is introduced and its importance for record linkage and the linkage free dual system estimator is explained. Finally, we discuss the parameter estimation of mixture-like models using the Markov chain Monte Carlo method known as the simulated annealing algorithm.

Chapter 4 demonstrates how the dual system estimation follows from the estimation of parameters of a mixture-like record linkage model. The classification or linkage free dual system estimator is introduced. Several variants of the linkage free estimator are presented. Finally, the modified linkage free estimator utilizing the 1-to-1 match constraint is introduced.

In Chapter 5 a variety of methods are used to establish identifiability of certain linkage models with four linkage variables.

Chapter 6 develops a method for variance estimation of the linkage free dual system estimator. A number of variants of this method are presented.

Chapter 7 describes the design, the purpose and the performance criteria of a simulation study. The simulation results assessing the point estimation are presented for a range of scenarios and settings.

Chapter 8 is a continuation of the simulation study, but the focus is on the variance estimation.

Chapter 9 contains a summary and conclusions of the thesis along the outline of the further work.

Finally, there is additional material in the appendix. Since the discussion of identifiability refers to algebraic statistics and specifically uses Gröbner bases, a topic which not every member of the statistical community is familiar with, there is a section with the prerequisites needed to understand

the topic. Additional simulation results and sample code for checking identifiability are also contained in the appendix.

2 Literature review and preliminaries

We start with a combination of a literature review and some introductory ideas that are needed for the discussion that will follow. There is a rather wide range of topics covered in this chapter: dual system estimation, standard classification-based probabilistic record linkage, mixture-model representation of record linkage including the criticism of such representation, an overview of census coverage estimation, a summary of the simulated annealing algorithm and a brief introduction to algebraic statistics methods that are used to assess identifiability of statistical models.

2.1 Capture-recapture and dual system estimation

Capture-recapture methods provide a way to estimate the size of a population from multiple but incomplete surveys of that population, even when a union of surveys is still incomplete. One of the first documented applications of what nowadays could be called a capture-recapture approach considered estimation of the total population of France by Laplace (1783). A century later Petersen (1896) used capture-recapture method with two samples to estimate the population of plaice in the Limfjord, Denmark, while Lincoln (1930) used a similar method to estimate the abundance of waterfowl. Geiger & Werner (1924) used capture-recapture to estimate the number of flashes on a zinc sulphide screen. The first use of the maximum likelihood method in the context of capture-recapture and estimation with multiple lists is discussed in Schnabel (1938), yet again in a fishery application. Sekar & Deming (1949) focused on the estimation of births and deaths in human population. Starting from 1950–1960s, the interest in capture-recapture grows rapidly and the number of research papers in this area becomes vast. Therefore, only those papers and monographs that are the most relevant to this work will be referred to as the discussion progresses.

It is astounding to see capture-recapture methods finding their application and generating new research questions in so many areas. These methods have particularly, but not exclusively, been used to estimate the size of animal populations (Seber, 1982; McCrea & Morgan, 2015), the under-enumeration of a human population census (Wolter, 1986; Fienberg, 1992; Brown et al., 2019), the number of war casualties (Ball et al., 2002), the number of duplicates in a database (Herzog et al., 2007, chap. 14), the extent of human trafficking (Silverman, 2020), the number of homeless people (Coumans et al., 2017) as well many other populations of interest in social and medical sciences (Böhning et al., 2018). All items referenced here contain rich bibliographies relevant to particular areas of application.

In this thesis we are considering the special case of capture-recapture methods with two surveys, which is often referred to as the dual system estimator. Moreover, we are dealing with the simple or basic dual system estimator, to contrast it with more elaborate log-linear or logistic regression based approaches (Alho, 1990). Essentially, we are dealing with the Petersen model in Wolter's (Wolter, 1986) classification.

2.1.1 Setup

We start with the description of our target population and the two data sets used in the dual system estimation. Some of the notation which is used throughout this work is introduced as well.

The target population is $\mathcal{P} = \{e_i : i = 1, \ldots, \tau\}$. Here e_i is an element of the population, which is an individual person in this thesis. Index i can be regarded as the hypothetical unique identifier, which is unobserved in reality. The population has a fixed size $\tau \in \mathbb{Z}^+$. Every element e_i has a number of associated attributes such as first name, surname, date of birth, address, etc. More detail related to these attributes will be provided later.

There are two independent sample surveys, S_1 and S_2, of the population $\mathcal{P}$. Since our goal is a unified framework combining record linkage and dual system estimation, we choose to use a generic term 'sample survey' instead of the terms such as 'list' or 'file' often used in record linkage literature. 'Sample survey' is not always technically accurate in the discussion that follows. For instance, dual system estimation is often used to adjust censuses for the coverage errors and while the census is a survey it is not a sample survey. However, it seems that there is no terminology general enough to cover all the different nuances of the topics under discussion, hence we chose to use an imperfect term 'sample survey'. We often use 'sample' and 'survey' interchangeably as a short version of 'sample survey', or whichever suits better a certain context. Each e_i has the probability $\pi_{1,i}$ to be selected, or counted in capture-recapture terminology, in S_1 and the probability $\pi_{2,i}$ to be selected in S_2. If the selection probabilities are constant within each of the surveys, then the probabilities are π_1 and π_2 for S_1 and S_2, respectively. In general, $\pi_1 \neq \pi_2$. In the context of dual system estimation, the selection probabilities are also often called the inclusion or count probabilities. These terms, as well as selection, count and inclusion, will be used interchangeably. Observe, that in real-life applications the target population is finite and sampling is carried out without replacement. Therefore, the inclusion probabilities cannot be constant. However, it is more convenient to deal with constant inclusion probabilities, as done in the majority of statistical models for capture-recapture. We discuss such a model in the next session. The size of S_1 is $n_1 \in \mathbb{Z}^+$, while the size of S_2 is $n_2 \in \mathbb{Z}^+$. It is often convenient to think about n_j, $j = \{1, 2\}$, as the realization of the random variable N_j. In this case N_j maps the outcome of drawing the sample S_j from $\mathcal{P}$ to the size of this sample. We write $s_{1,a}$ and $s_{2,b}$ to denote a^{th} and b^{th} elements of S_1 and S_2, respectively. We can write $S_1 = \{s_{1,a} : a = 1, \ldots, n_1\}$ and $S_2 = \{s_{2,b} : b = 1, \ldots, n_2\}$.

Now let $\text{id}(x)$ be the function that takes a record in a given survey and returns its corresponding hypothetical unique identifier. Say, for a record $s_{1,a}$ of the first survey $\text{id}(s_{1,a}) = i$, meaning that the record corresponds to the element e_i in $\mathcal{P}$. Consider the set $\mathcal{M}$ of the population elements that are sampled in both S_1 and S_2. This set is called the match-set or the set of matches and can be written as $\mathcal{M} = \{(s_{1,a}, s_{2,b}) : s_{1,a} \in S_1, s_{2,b} \in S_2, \text{id}(s_{1,a}) = \text{id}(s_{2,b})\}$, where $(s_{1,a}, s_{2,b})$ is a record pair created by binning one record from S_1 and one record from S_2 in a tuple. The size of the set of matches is $m \in \mathbb{Z}^+$. Again, in the case of repeated sampling, m is the realization of the random variable M that maps the outcomes of draws of two samples to the size of the resulting set of matches.

2.1.2 Dual system estimator

With the set up as in Section 2.1.1, the goal is to estimate the population size τ having two incomplete surveys S_1 and S_2 of the population $\mathcal{P}$, such that $S_1 \cup S_2 \neq \mathcal{P}$. Several assumptions must be satisfied in order to obtain the dual system estimator of τ (Seber, 1982; Wolter, 1986). Throughout this work all these assumptions are considered to hold unless stated otherwise.

The first assumption is that the target population $\mathcal{P}$ is closed. That is to say that both surveys S_1 and S_2 are taken on the exactly the same population, no elements are taken out or added between those two sampling occasions. In practice it means that there are no births, deaths or migrations in $\mathcal{P}$ between the surveys.

The second assumption is that there is no overcount in the form of duplicated counting of e_i or counting e_i in the wrong location. Duplication in the survey $S_j, j = \{1, 2\}$, means that there are $s_{j,x}$ and $s_{j,y}$ such that $\mathrm{id}(s_{j,x}) = \mathrm{id}(s_{j,y})$. In practice it means that a certain population element was counted twice by one of the surveys. To explain counting in the wrong location, consider the geographical location attribute l_i of e_i. For instance, the value of l_i can be a local authority where e_i is located in the population. Let $\mathrm{loc}(x)$ be the function that takes a record on a given survey and returns the recorded location attribute. We say that $s_{1,a}$ is counted (overcounted) in the wrong location whenever $\mathrm{id}(s_{1,a}) = i$, but $\mathrm{loc}(s_{1,a}) \neq l_i$. In practice it means that the recorded location of an element is different from the true location.

The third assumption is that for every e_i, being captured (or missed) in the first survey does not affect the chance of being captured (or missed) in the second survey. In other words, the joint probability $\pi_{11,i}$ of e_i being captured in both surveys is $\pi_{11,i} = \pi_{1,i}\pi_{2,i}$.

The fourth assumption is of constant inclusion probability at least in one of the surveys. That is, for every e_i in $\mathcal{P}$, $\pi_{j,i} = \pi_j$ at least for one $j = \{1, 2\}$. Again, $\pi_1 \neq \pi_2$ in general.

Finally, perfect matching is assumed. Perfect matching means that all pairs $(s_{1,a}, s_{2,b}) \in \mathcal{M}$ can be identified. Our main development will be around replacing this assumption with either the assertion that the size M of the set of matches can be estimated, or that the estimate of τ can be obtained directly from the estimate of a certain linkage model parameter. In any case, there is no ultimate classification of pairs $(s_{1,a}, s_{2,b})$ in this work.

We will revisit some of these assumptions in more detail later. We can now introduce the dual system estimator. Among the most widely used statistical models for the dual system estimation of the population size τ is the multinomial model (Bishop et al., 1975; Pollock, 1976; Wolter, 1986). While this model corresponds to the above setup, this set up and the model itself are not exact, since in the majority of real-life applications sampling is carried out without replacement. However, such a model is a good approximation and it is very useful in subsequent discussion. Therefore, the multinomial model is assumed for dual system estimation throughout this work unless stated otherwise. Under this model every e_i in the population has four possible outcomes as the result of the two surveys taking place. The element can be either captured by both surveys with the probability $\pi_{11,i}$, captured in the first survey only with the probability $\pi_{10,i}$, captured in the second survey only with the probability $\pi_{01,i}$ or not captured in either of the survey with the probability $\pi_{00,i} = 1 - \pi_{11,i} - \pi_{10,i} - \pi_{01,i}$. The marginal capture probabilities for S_1 and S_2 are $\pi_{1,i}$ and $\pi_{2,i}$, respectively. It is often convenient to

summarize the resulting probabilities in a 2×2 table, as shown in Table 1.

Table 1: Probabilities of outcomes for a person e_i

		Counted in S_2	Missed in S_2
		$\pi_{2,i}$	$1 - \pi_{2,i}$
Counted in S_1	$\pi_{1,i}$	$\pi_{11,i}$	$\pi_{10,i}$
Missed in S_1	$1 - \pi_{1,i}$	$\pi_{01,i}$	$\pi_{00,i}$

Using the assumptions of independence between two surveys and constant selection probabilities allows writing Table 1 as Table 2.

Table 2: Probabilities of outcomes for a person e_i under the dual system estimation assumptions

		Counted in S_2	Missed in S_2
		π_2	$1 - \pi_2$
Counted in S_1	π_1	$\pi_1\pi_2$	$\pi_1(1 - \pi_2)$
Missed in S_1	$1 - \pi_1$	$(1 - \pi_1)\pi_2$	$(1 - \pi_1)(1 - \pi_2)$

Random variables $M, N_{10}, N_{01}, N_{00}$, which map the outcome of drawing two samples to the number of people counted in both surveys, counted in the first only, counted in the second only and missed from both, respectively, follow the multinomial distribution with the probabilities as in Table 2. The corresponding realisations of these random variables, sometimes referred as cell counts, are m, n_{10}, n_{01} and n_{00}. Note, that in dual system estimation the cell value n_{00} is unobservable and unknown. The sum of the cell counts is equal to the population size τ. Table 3 is the related contingency table, we treat these quantities as the objective truth, which not necessarily can be accurately established in practice.

Table 3: Aggregate outcomes under the dual system estimation assumptions

		Counted in S_2	Missed in S_2
		N_2	$\tau - N_2$
Counted in S_1	N_1	M	N_{10}
Missed in S_1	$\tau - N_1$	N_{01}	N_{00}

Using the multinomial distribution has a number of implications both for the dual system estimation and our discussion of the no-classification linkage and population size estimation. Two important properties of the multinomial distribution are: (1) if $(X_1, \ldots, X_z)^T$ have a multinomial distribution, then $(\sum_{\alpha_1} X_i, \ldots, \sum_{\alpha_j} X_i)^T$, where $\alpha_1, \ldots, \alpha_j$ are partitions of $1, \ldots, z$ with each X_i belonging to only one partition, also have a multinomial distribution; (2) the conditional distribution of a subset of the $(X_1, \ldots, X_z)^T$ given the values of the remaining subset also has a multinomial distribution (Bishop et al., 1975, chap. 13)

The first property means that the marginal distributions of $N_1, \tau - N_1$ and $N_2, \tau - N_2$ are also multinomial. Since there are only two outcomes, and the binomial distribution is the special case of the multinomial with only two outcomes, it follows that $N_1 \sim Bin(\tau, \pi_1)$ and $N_2 \sim Bin(\tau, \pi_2)$. Similarly, $M \sim Bin(\tau, \pi_1\pi_2)$.

The second property allows the maximum likelihood estimates of π_1, π_2 and τ to be obtained (Bishop et al., 1975, chap. 6). Let $n = m + n_{10} + n_{01}$ be the total number of observed cases after drawing both samples. Treating n as fixed and using the second property of a multinomial distribution mentioned above, the distribution of (m, n_{10}, n_{01}) is multinomial with the corresponding probability mass function

$$\binom{n}{m, n_{10}, n_{01}} \frac{\pi_1^{n_1} \pi_2^{n_2} (1-\pi_1)^{n_{01}} (1-\pi_2)^{n_{10}}}{[1-(1-\pi_1)(1-\pi_2)]^n}, \tag{1}$$

where

$$\binom{n}{m, n_{10}, n_{01}} = \frac{n!}{m! n_{10}! n_{01}!}$$

is the multinomial coefficient. The denominator of (1) comes from conditioning on the event that an element e_i is observed in at least one of the surveys, an event that has the probability $1-(1-\pi_1)(1-\pi_2)$.

Finding the values of π_1 and π_2 that maximise (1) given n, m, n_{10} and n_{01}, yields the maximum likelihood estimates for these parameters, which are

$$\widehat{\pi}_1 = \frac{m}{n_2}, \widehat{\pi}_2 = \frac{m}{n_1}. \tag{2}$$

Note that the capture probability for the first survey is estimated using the size of the second and vice versa.

Using the first property from the above, we have that the sum $n = m + n_{10} + n_{01}$ has the binomial distribution with the probability mass function

$$\binom{\tau}{n} [1-(1-\pi_1)(1-\pi_2)]^n [(1-\pi_1)(1-\pi_2)]^{\tau-n}. \tag{3}$$

If the population size τ is fixed and the probability of success, pr(event occurs), is known (in this context, success means being counted at least in one of the surveys), then the maximum likelihood estimate of τ given the probability of success is $\lfloor n/\text{pr(event occurs)} \rfloor$ (Bishop et al., 1975, chap. 13). Here $\lfloor x \rfloor$ is the greatest integer function, that is the function that returns the greatest integer smaller than or equal to x for $x \in \mathbb{R}$. Therefore, given π_1 and π_2 the maximum likelihood estimator of τ is

$$\widehat{\tau}^* = \left\lfloor \frac{n}{1-(1-\pi_1)(1-\pi_2)} \right\rfloor.$$

The greatest integer function is usually ignored in the maximum likelihood estimation of the population size as it has little effect for a reasonably large τ.

Plugging the maximum likelihood estimates (2) into the above expression and ignoring the greatest integer function, we obtain the simple or basic dual system estimator of the population size

$$\widehat{\tau} = \frac{n_1 n_2}{m}. \tag{4}$$

There are several alternative ways to obtain the same estimator. Some of the derivations have merits making them worth discussing here. All of the setup and assumptions are as before. The first alternative is based on the expectation of the ratio of the number of matches to the number of elements

in the second survey, M/N_2. We have

$$\mathbb{E}\left(\frac{M}{N_2}\right) \approx \frac{\mathbb{E}(M)}{\mathbb{E}(N_2)} = \frac{\pi_1\pi_2\tau}{\pi_2\tau} = \pi_1 = \mathbb{E}\left(\frac{N_1}{\tau}\right) = \frac{\mathbb{E}(N_1)}{\tau},$$

where the approximate equality of the expectation of the ratio and the ratio of expectations can be demonstrated using Taylor series approximation. We do not include the related computations here, but Taylor linearisation is used extensively in a similar context later in Section 3.2. From the above expression it follows that

$$\tau \approx \frac{\mathbb{E}(N_1)\mathbb{E}(N_2)}{\mathbb{E}(M)}, \tag{5}$$

and replacing the expectations with the observed values m, n_1 and n_2, gives the dual system estimator (4). The importance of this derivation is that it makes explicitly visible that the dual system estimator is a special case of the ratio estimator; see Cochran (1977, chap. 6). Inherently, the dual system estimator has the properties of the ratio estimator. For instance, it is biased, but the bias decreases as the sample size increases.

Another alternative is related to the theory of complete and incomplete contingency tables; see Fienberg (1972), Bishop et al. (1975, chap. 6.2.3) and literature therein. Independence between the two surveys is equivalent to the cross-product ratio of the expected values of the cells of the contingency Table 3 to be equal to 1. Let the expected values of $M, N_{10}, N_{01}, N_{00}$ be $\overline{m}, \overline{n}_{10}, \overline{n}_{01}, \overline{n}_{00}$, respectively. The cross-product ratio under the independence is

$$\frac{\overline{n}_{00}\overline{m}}{\overline{n}_{10}\overline{n}_{01}} = 1.$$

When estimating the population total, the missed from both cell is unobserved and the contingency table is incomplete. The assumption of independence expressed as the cross-product ratio means that

$$\overline{n}_{00} = \frac{\overline{n}_{10}\overline{n}_{01}}{\overline{m}}.$$

Hence, the estimate for the non-observed cell is

$$\widehat{n}_{00} = \frac{\widehat{\overline{n}}_{10}\widehat{\overline{n}}_{01}}{\widehat{\overline{m}}} = \frac{n_{10}n_{01}}{m},$$

and combining with the rest of the observed cells gives the basic dual system estimator (4):

$$\widehat{\tau} = m + n_{10} + n_{01} + \frac{n_{10}n_{01}}{m} = \frac{n_1 n_2}{m}.$$

It is possible to use sampling weights within the dual system estimator if needed. The simplest way is to up-weight each term of the estimator (4) by the appropriate sampling weight. However, when using the dual system estimator in census coverage situations, a census provides population counts for all clusters in the population. Hence, the census data can be used as auxiliary information, which enables a substantially more efficient estimation by means of ratio estimation (Wolter, 1986; Brown, 2000; Brown et al., 2019). Since the coverage survey samples clusters within a sample stratum,

all elements within the cluster or neighbouring clusters have the same sampling weights. Therefore, there is no need for sampling weights when computing dual system estimates for each individual cluster. Sampling weights, if needed, are used within the ratio estimator. This usually happens when individual dual system estimates from different strata are pooled together in the ratio estimator. In this thesis we are focusing on application of the no-classification methods for the census coverage or similar problems. Therefore, there is no need to use sampling weights when estimating the size of small subpopulations with the no-classification methods and we leave the development of the weighted no-classification approaches for future research.

No matter which way the dual system estimator is derived, what remains unchanged is that the only two directly observed random variables are N_1 and N_2, the sizes of two surveys. The rest of the variables needed for the estimation, such as M, come from a linkage process. Therefore, perfect or at least very high quality linkage is needed in order to produce a reliable estimate for τ and the dual system estimator is sensitive to errors in linkage; see for instance Biemer (1988); Tancredi & Liseo (2011).

We will discuss record linkage in more detail in Section 2.2. For the time being we only need to be aware of the useful distinction between matches and links. If a record pair represents the same element in a population, it is said that the pair is a match. Otherwise, the pair is a non-match. If a record pair is classified as a match, but its true match / non-match status is unknown, then it is said that the pair is a link. If a pair is classified as a non-match, but its true match / non-match status is unknown, then it is said that the pair is a non-link (Larsen & Rubin, 2001).

There are two possible errors in the linkage process. A pair $(s_{1,a}, s_{2,b})$ such that $\mathrm{id}(s_{1,a}) = \mathrm{id}(s_{2,b})$ can be erroneously classified as a non-link. This type of error is known as a false negative. In this case, the number of declared links will be smaller than the realisation m of M. If the error is systematic, then the declared number of links is systematically smaller than the true number of links which leads to overestimation of τ. By contrast, a pair $(s_{1,a}, s_{2,b})$ such that $\mathrm{id}(s_{1,a}) \neq \mathrm{id}(s_{2,b})$ can be erroneously classified as a link. This type of error is known as a false positive and in this case the number of declared links is larger than the realisation m of M. If such errors are systematic, the number of declared links is inflated relative to the true number of links which leads to underestimation of τ. It is easy to see that the overall effect of the linkage errors on the dual system estimator is determined by the total linkage error which is the sum of all false positive minus the sum of all false negative errors. More on linkage errors is presented in Section 2.2. Usually, the practical difficulty of achieving perfect linkage is recognized by setting some admissible level for linkage errors and accepting the consequent small bias. In addition, linkage, and probabilistic linkage in particular, involves a trade-off between the level of admissible errors and the amount of clerical resolution needed to resolve the set of potential links. Essentially, it is a trade off between the quality of classification and the time and cost of processing. The lower the level of admissible errors, the more clerical resolution is generally needed.

There was some interest in linkage error in relation to the dual system estimation, including adjustment of the dual system estimates for linkage errors (Biemer, 1988; Ding & Fienberg, 1994; Di Consiglio & Tuoto, 2015; Tuoto, 2016; de Wolf et al., 2019). The conceptual framework for linkage error adjustment is set out in Ding & Fienberg (1994) with the majority of later research building on

it. These methods require an external estimate of the linkage error rates. The estimation of these error rates often depends on a supplementary data collection and processing exercise.

In this thesis we are focusing on linkage without actual classification of record pairs into links and non-links. The no-classification methods we develop aim on producing an accurate estimate of M or τ with a carefully specified linkage model. As a result, the notions of false negative and false positive errors are irrelevant for such methods. Instead, the error in estimation of M or τ becomes important. We will show that the dual system estimation follows from a certain linkage model, called here a mixture-like model. Other work that discuss the close relationship of dual system estimation and / or fully automated record linkage are Tancredi & Liseo (2011); Johndrow et al. (2018); Tancredi et al. (2020); Lee et al. (2022).

Among the largest population size estimation exercises that employ dual system estimation are censuses of human populations. Despite the aim to survey every member of a population, no census is perfect. Therefore, an independent data collection exercise is carried out for a sample of the target population and dual system estimation using the census and survey data is carried out to estimate the size of the missed population. While the goal is to reliably estimate the population total of a large human population, several practical considerations influence the data collection design. This, in turn, affects how the dual system estimator is applied. In order to keep two data collection exercises operationally independent, different data collection modes are employed for each of the data sources. One of them, usually a smaller sample-based survey, is conducted as face-to-face interviews. It is operationally easier to achieve a high response rate and cheaper to run a survey with a cluster design. In addition, since no perfect sampling frame of ultimate sampling units (which are usually households in coverage surveys) exists, clusters, such as postcodes, allow the capture and use in estimation elements that could not otherwise be captured due to the frame imperfections. With such a cluster design it is convenient to produce separate dual system estimates of the population size of each such geographical cluster (or aggregation of neighbouring clusters), usually post-stratifying the population by some demographic variable. For example, by age-sex group. These separate dual system estimates can then be used as input for other estimators, that produce estimates for the large domains, such as local authority by age-sex group. A more detailed overview of the census coverage estimation is presented in Section 2.3. What is really important for the methods discussed in this thesis, is that in the census coverage context the dual system estimator is applied to estimate the population size of relatively small and to a degree homogeneous groups. In addition, the coverage of the clusters in each survey is quite high in such situations. The no-classification approaches developed in this thesis are best suited for the size estimation of such small populations (reasons are explained in Section 3.3). Therefore, the main focus is on the census coverage-like situations, where individual population size estimates for small groups can be obtained. Hence, we explore how well the no-classification methods can estimate a domain with size varying between 250 and 1000 individuals, with coverage probabilities ranging between 0.7 and 0.9. These configurations are very common for censuses and coverage (post-enumeration) surveys. Similar data collection design may be used with an incomplete population register or a nearly exhaustive administrative data set instead of a census. The success of the no-classification methods is not guaranteed with an arbitrary data collection design and is not explored in this thesis, but left for

future development.

2.1.3 Variance estimation

The dual system estimator is a non-linear function of M, N_1 and N_2. Therefore, it is hard to obtain an exact variance estimator for $\widehat{\tau}$. The linearisation approach based on the Taylor series is usually used in this case. Sekar & Deming (1949) have shown that the approximate variance estimator is

$$\widehat{\mathrm{Var}}(\widehat{\tau}) \approx \frac{n_1 n_2 n_{10} n_{01}}{m^3}, \tag{6}$$

and this result is generally accepted since (Wolter, 1986; Bishop et al., 1975, chap. 6.2.2).

In the capture-recapture literature, there exists a tradition of distinguishing two types of variance estimators. The first estimator is conditional on the observed sample sizes n_1 and n_2, the second one is unconditional on the observed sample sizes (Seber, 1982; Buckland & Garthwaite, 1991). Some researchers also talk about variance estimation conditional or unconditional on the unique number of captures among all the samples (Norris & Pollock, 1996). Such a distinction stems from the fact that there are several ways to represent the data generating mechanism for the data used in dual system estimation. The first way, discussed in Section 2.1.2, uses the multinomial distribution for the four outcomes of a capture-recapture data collection. In this case, the population size and coverage probabilities are fixed, but the realized sample sizes are random variables. The number of matches in this case is binomially distributed. An alternative way to conceptualize the data generating mechanism, is to fix the population size as well as the sizes of sample. Then the hypergeometric distribution is a convenient one for the number of matches. The true variance of the dual system estimator differs under these two generating models. Whenever variance estimation of the dual system estimator is considered, variance unconditional on the observed sample size is related to the multinomial generating mechanism with fixed coverage probabilities but varying (random) sample sizes, whereas variance conditional on the observed sample sizes is related to the hypergeometric generating mechanism with the fixed sample sizes. When using the variance approximation presented above, there is often little difference between these two views (Seber, 1982). However, resampling methods, in particular the parametric bootstrap, allow the distinction between the two generating mechanisms to be reflected (Buckland & Garthwaite, 1991). Therefore, one often uses the terms 'conditional' and 'unconditional' variance in this case.

2.2 Record linkage

In a narrow sense, record linkage is a process of classifying n-tuples, where each entry of the tuple is an observation from a particular one of the n datasets, as either referring to the same entity in the target population or not, given no unique identifiers of the population entities are available. An absence of unique identifiers means that attributes of the population elements used in linkage are not necessarily unique for each of the elements. Examples of such attributes can be names, surnames, age, marital status, ethnicity and many others. In addition to non-uniqueness of the attributes, record linkage often needs to perform the mentioned classification in the situation when attributes for some of the elements in certain data sources are recorded with errors. In a wider sense, record linkage comprises

the theory and practice of such classification as well as the related fields of data preparation, indexing and comparison. See Herzog et al. (2007); Christen (2012) for detailed overviews.

The simplest, arguably the most frequent and the most researched case of record linkage concerns linking two data sets or surveys related to the same population. In this work, only the case of two-survey linkage is considered. Whenever two data sources are involved, a linkage process concerns classification of 2-tuples or ordered record pairs, which we for simplicity call just record pairs.

The modern theory of record linkage and the term itself goes back to work by Dunn (1946). It was followed by several papers by Newcombe and co-authors (Newcombe et al., 1959; Newcombe & Kennedy, 1961) where a general conceptual framework still in use today was developed, though informally. Note that, despite a crucial role played by clerical revision of potential links, this early research was aimed at a computer assisted record linkage and therefore frequently was called automatic record linkage. Nowadays, when pretty much any statistical or data processing task is done using a computer, automatic record linkage is a more suitable name for a process where there is no clerical involvement beyond some general assessment of a record linkage task and response to initial results by a statistician undertaking the linkage.

A seminal paper by Fellegi & Sunter (1969) formalised the earlier work and became the departure point for many later variants of record linkage. Methods developed in this thesis, while having some substantial deviations from the Fellegi–Sunter approach, derive from their work. Below in Sections 2.2.2 and 2.2.3 we discuss the Fellegi–Sunter approach and some important descendants of it in more detail.

Record linkage can be approached from various perspectives. It can be deterministic or probabilistic. Dependent on the data available, record linkage can employ supervised or unsupervised statistical learning methods. It can be viewed as a clustering or microclustering task. The problem of record linkage can be treated either within the frequentist or Bayesian paradigm. Unfortunately, we cannot discuss all the aspects of record linkage in this thesis. A very good overview of these aspects is presented in Binette & Steorts (2022).

The linkage method presented in this thesis has population size estimation as its ultimate goal and differs from many existing approaches by not producing classified pairs as links and non-links. Instead, this method targets accurate estimation of certain linkage model parameters that in turn lead to dual system estimates. While many existing probabilistic record linkage techniques aim at estimating either probabilities or scores associated with record pairs, the majority of methods produce classification as an output. In a sense, the approach we developed in this thesis, being purely estimation focused, contradicts the definition of record linkage. However, as will be shown, in a problem like dual system estimation, there is no strict necessity in classification and accurate estimation is sufficient.

2.2.1 Preparing for linkage

While the main focus of this thesis is the estimation of the linkage model and related parameters, it is important to bear in mind that substantial effort is needed when preparing the data for these tasks. This section offers a very brief outline of the processes preceding linkage and estimation. For an in-depth overview see again Herzog et al. (2007); Christen (2012). Typically, there are three preparatory

stages for record linkage once the data exist in the digital form: data cleaning and standardisation, data indexing or blocking, comparison of values of the linkage variables.

Data cleaning and standardisation tasks are needed in order to correct the collected data for nuisance errors and to put both data sets into the common format used in the processing pipeline. This is especially important when dealing with data related to human populations and linkage variables such as name, surname, address and date of birth. Data cleaning usually involves removing special characters, symbols or words from the recorded values of linkage variables. This is followed by correcting obvious or easy to detect misspellings, expanding abbreviations and looking up nicknames. The set of linkage variables chosen for a linkage task may not correspond to the variables collected in the raw input data and values of the variables may be collected differently in each of the surveys. Therefore, the values of linkage variables should be populated with the relevant attribute values collected on both surveys. Moreover, these attributes should be formatted or standardized in the same way for both surveys. For instance, one of the surveys may record first name as a separate variable and surname as a separate variable, while another survey may collect both within a single 'name' variable. If the linkage process uses the first name and surname variables, the 'name' variable on the second survey should be parsed in order to correctly populate the relevant linkage variables. Standardisation of address attributes may be particularly difficult as it requires parsing of several attributes such as street or locality name, house number or name and apartment number into a standard format. If circumstances allow, the verification of certain attributes can be performed. For instance, if both street name and postcode are collected and good quality data mapping streets to postcodes exist, typos in the street name can be corrected based on the value of the postcode. Ideally, both surveys will be collected in a controlled way having record linkage and other statistical usage of the data in mind. In this case, either the linkage variables or parts of the linkage variables can be collected in an explicit and standardised manner. For instance, both surveys may explicitly collect the 'name' and 'surname' variables that are used in linkage. Attributes constituting the ultimate address linkage variable, such as street name, house name, etc., can also be directly collected. Moreover, if the data collection is digital in the first place, such as online self-completion questionnaire or a tablet / smartphone assisted face-to-face interview, verification of certain attributes, such as addresses, can be performed at the time the data are collected. In general, a well-designed data collection can enhance the quality of linkage and estimation substantially. But even when data collection is planned and carefully thought about up-front, the tasks of cleaning and standardisation require non-trivial effort. More on designed data collection for population size estimation will be covered in Section 2.3.

Data indexing or blocking is used to reduce the number of record tuples to process. For instance, linking two data sets aims to find all members of the set of matches $\mathcal{M}$, as discussed in 2.1.1. In order to do it, $n_1 n_2$ pairs must be considered. The number of pairs to process grows approximately quadratically in τ, while the size of the set of matches grows linearly. Hence, as the population size gets larger, the number of pairs to process becomes very large with the majority of pairs being non-matches. To make the linkage process more efficient and reduce the computational burden, only records that satisfy certain criteria in both surveys are compared. For instance, only records within the same postcode or records with surnames starting with a certain letter are compared. Such splitting

of records is called blocking and the resulting groups of records satisfying the given criteria are called blocks. When blocking by first name or surname, phonetic encoding algorithms, such as Soundex or Phonex may be used. Those algorithms convert strings into codes and allow certain simple and common typographical errors to be bypassed. Detailed overviews of blocking can be found in Herzog et al. (2007, chap. 12) and Christen (2012, chap. 4). In Section 3.4 we introduce a special case of blocking that plays an important role beyond reducing the computational burden in the linkage model.

Finally, the values of linkage variables must be compared. The comparison can be exact, that is two values of a given variable must be exactly the same for a certain pair of records for a comparison to be declared as agreement. In this case, the outcome of comparison is in the set $\{0, 1\}$. In reality, however, even when the high quality data collection and processing practices are followed, the values of a variable of a record pair belonging to the set of matches may not agree exactly due to small typographical differences. These differences can come from scanning errors, typos, interviewer mistakes and many other reasons. Therefore, an approximate or fuzzy comparison is employed. This type of comparisons uses various edit distance functions that measure the similarity or dissimilarity of two strings. The definition of similarity varies from function to function. Among the most commonly used edit distance functions for names and surnames are Levenshtein distance, the Jaro function and the Winkler function. When comparing dates or age, methods based on the absolute difference are common. Again, a great overview of approximate comparison can be found in Christen (2012, chap. 5). The majority of edit distance functions return a normalised value within the $[0, 1]$ interval. Not all linkage models can accept values different from 0 and 1. Often, when a linkage model requires the binary comparison outcomes, an edit distance function is used to produce the comparison value first. Then this value is converted into 0 or 1 using some threshold for the comparison value above which the outcome of the comparison is treated as agreement, while treated as disagreement below that threshold.

2.2.2 Fellegi-Sunter approach

The majority of probabilistic linkage methods stem from the model proposed by Fellegi & Sunter (1969). In their paper, important record linkage concepts were introduced and an entirely formal approach for linkage was developed for the first time. While later developments in this area brought new ideas and ways of performing probabilistic linkage, very few of the innovations broke completely away from that classical work. The method developed in this thesis is no exception: while being quite different in many respects, certain basic premises remain the same.

Since this thesis aims at a more holistic approach viewing the areas of record linkage and dual system estimation as interrelated, we deliberately use a unified notation across these two areas. Recall that in Section 2.1.1 we introduced the population of elements of interest $\mathcal{P} = \{e_i : i = 1, \ldots, \tau\}$ and two surveys, $S_1 = \{s_{1,a} : a = 1, \ldots, n_1\}$ and $S_2 = \{s_{2,b} : b = 1, \ldots, n_2\}$, of this population. Linking S_1 and S_2 in the absence of unique identifiers concerns the classification of ordered pairs of records $(s_{1,a}, s_{2,b})$. Let $\mathcal{W} = S_1 \times S_2 = \{(s_{1,a}, s_{2,b}) : s_{1,a} \in S_1, s_{2,b} \in S_2, a = 1, \ldots, n_1, b = 1, \ldots, n_2\}$ be the set of all ordered pairs, where $\times$ is the Cartesian product. The set $\mathcal{W}$ can be partitioned into two disjoint sets. The set of matches, $\mathcal{M} = \{(s_{1,a}, s_{2,b}) : s_{1,a} \in S_1, s_{2,b} \in S_2, \mathrm{id}(s_{1,a}) = \mathrm{id}(s_{2,b})\}$, was

already introduced when discussing the dual system estimator. The second is the set of non-matches $\mathcal{U} = \{(s_{1,a}, s_{2,b}) : s_{1,a} \in S_1, s_{2,b} \in S_2, \mathrm{id}(s_{1,a}) \neq \mathrm{id}(s_{2,b})\}$ which consists of all pairs of records that do not represent the same element in the population. Recall, that M is the random variable that maps the outcome of undertaking two surveys to the size of the resulting set of matches and the realization of this random variable is m. Put simply, m is the corresponding size of the set of matches $\mathcal{M}$ for a given linkage exercise. The random variables U and W are defined in a similar way: they map the outcome of undertaking two surveys to the sizes of the resulting set of non-matches and the total number of resulting pairs, respectively. The realization of U is u and the realization of W is w.

As already mentioned in Section 2.1.1, every element e_i has a number of attributes associated with it. These attributes, like name, surname, address, are not necessarily unique for a particular element and there may be other elements in the population $\mathcal{P}$ with some or even all attributes being the same as those of e_i. Such non-uniqueness of attributes can also be called the absence of unique identifiers. Each of the surveys S_1 and S_2 collects or records the values of some of these attributes and these values are recorded as the values of linkage variables on these surveys. There are K linkage variables in common in the two surveys. We denote a specific linkage variable as $v_k, k = 1, \ldots, K$, and we use subscripts $k = 1, \ldots, K$ when dealing with functions or random variables associated with the k^{th} linkage variable. For some elements, the recorded values of the attributes may contain errors in one or both surveys. These errors may be misspellings, missing values, scanning errors or errors of any other nature. A further discussion of errors can be found in Section 2.2.5.

Recall that if a record pair represents the same element in a population, it is said that the pair is a match. Otherwise, the pair is a non-match. If a record pair is classified as a match (or estimated to be a match), but its true match / non-match status is unknown, then it is said that the pair is a link. If a pair is classified as a non-match, but its true match / non-match status is unknown, then it is said that the pair is a non-link (Larsen & Rubin, 2001).

Classification of record pairs into links and non-links is carried out based on the comparison outcomes of the values of linkage variables. Consider a record pair $(s_{1,a}, s_{2,b})$. Then we write $\gamma_k(s_{1,a}, s_{2,b})$ to denote the outcome of comparing the values of the k^{th} linkage variable for the pair. Whenever there is no risk of ambiguity, we use the shorter notation $\gamma_k(a, b)$ instead of $\gamma_k(s_{1,a}, s_{2,b})$. In principle, $\gamma_k(a, b)$ can be in any range, but in this work we are dealing with a linkage model that only allows the binary outcome $\gamma_k(a, b) \in \{0, 1\}$. This outcome may be a result of an exact comparison of the values of the k^{th} linkage variable of records $s_{1,a}$ and $s_{2,b}$, denoted $s_{1,a,k}$ and $s_{2,b,k}$ respectively. So that $\gamma_k(a, b) = 1$ if $s_{1,a,k} = s_{2,b,k}$, and 0 otherwise. Alternatively, if some edit distance function f is used, then $\gamma_k(a, b) = 1$ if edit distance $f(s_{1,a,k}, s_{2,b,k})$ is above a certain threshold, and 0 otherwise. For some of the records, certain variables may have missing values, which is treated as error, and the comparison outcome is set to 0 in that case in all approaches discussed in this work. The outcome of comparisons of all linkage variables of a record pair $(s_{1,a}, s_{2,b})$ is denoted $\boldsymbol{\gamma}(s_{1,a}, s_{2,b})$ or simply $\boldsymbol{\gamma}(a, b)$. This outcome is a vector of individual comparisons, $\boldsymbol{\gamma}(a, b) = (\gamma_1(a, b), \ldots, \gamma_K(a, b))^T$. There are no restriction on the number of comparison outcomes in the original paper by Fellegi & Sunter (1969) and all the general results in this chapter hold for arbitrary comparisons. Nevertheless, since the models developed in this work are exclusively for binary comparisons, we prefer to introduce notation

with these comparisons in mind. For binary comparison outcomes and K linkage variables, there are 2^K different combinations of comparison outcomes. Each such combination is called a comparison pattern, and is denoted $\boldsymbol{\gamma}_p, p = 1, \ldots, 2^K$. For instance, with $K = 4$, the comparison patterns are $\boldsymbol{\gamma}_1 = (1,1,1,1)^T, \boldsymbol{\gamma}_2 = (0,1,1,1)^T, \boldsymbol{\gamma}_3 = (1,0,1,1)^T, \ldots, \boldsymbol{\gamma}_{15} = (0,0,0,1)^T, \boldsymbol{\gamma}_{16} = (0,0,0,0)^T$. Whenever we want to refer to the k^{th} entry of a comparison pattern $\boldsymbol{\gamma}_p$, we use γ_p^k. For instance, the first and fourth entries of the pattern $\boldsymbol{\gamma}_2 = (0,1,1,1)^T$ are $\gamma_2^1 = 0$ and $\gamma_2^4 = 1$, respectively. The reason that entries in comparison patters are not all 1's or all 0's is the non-uniqueness of attributes, on the one hand, and error recording these attributes, on the other hand. So that if the linkage variables are address, surname, first name and date of birth, the pattern $(1,1,0,0)^T$ when comparing two pairs may, for instance, mean that two members of the same household / address with the same surname but different names and dates of birth are comprising the pair $(s_{1,a}, s_{2,b})$. Alternatively, $(s_{1,a}, s_{2,b})$ may refer to the same individual, but on one or both sources the person's name and date of birth were incorrectly recorded. In Section 2.2.6 we give a more detailed example of this notation.

Often blocking is employed to reduce the number of pairs to compare and classify. Blocking ensures that only those records satisfying certain condition or multiple conditions, such as agreeing on a particular variable, are considered for comparison. The above notation still applies when blocking is used, but now each block has the corresponding set $\mathcal{W}_b$, for a block b, etc. Linkage is either carried out on the pairs pooled from different blocks once the blocking has been applied, or sometimes in each block separately.

The collection of all possible comparisons $\boldsymbol{\gamma}_p$ is called the space of all comparisons in the original paper by Fellegi & Sunter (1969). The comparison outcomes for each pair are observable, but their match status is not. Each comparison outcome $\boldsymbol{\gamma}_p$ can be classified as either a link, possible link or non-link.

Linkage rule $d(\boldsymbol{\gamma}_p)$ associates each of the three possible classifications with the probability of making that classification. So that for a given $\boldsymbol{\gamma}_p$, $d(\boldsymbol{\gamma}_p) = \{\text{pr}(\text{link} \mid \boldsymbol{\gamma}_p), \text{pr}(\text{possible link} \mid \boldsymbol{\gamma}_p), \text{pr}(\text{non-link} \mid \boldsymbol{\gamma}_p)\}$.

In order to formalize the errors associated with the linkage rule, the following conditional probabilities are defined. The probability that comparison of a pair $(s_{1,a}, s_{2,b})$ will produce a pattern $\boldsymbol{\gamma}_p$ given the pair is a match: $\mu(\boldsymbol{\gamma}_p) = \text{pr}(\boldsymbol{\gamma}(a,b) = \boldsymbol{\gamma}_p \mid (a,b) \in \mathcal{M})$. The probability that comparison of a pair $(s_{1,a}, s_{2,b})$ will produce a pattern $\boldsymbol{\gamma}_p$ given the pair is a non-match: $\nu(\boldsymbol{\gamma}_p) = \text{pr}(\boldsymbol{\gamma}(a,b) = \boldsymbol{\gamma}_p \mid (a,b) \in \mathcal{U})$.

There are two errors related to the linkage rule $d(\gamma)$. The first is the false positive error which occurs when a pair from the non-matched set is classified as a link. The corresponding probability is

$$\pi_{\text{fp}} = \text{pr}(\text{declare } (a,b) \text{ a link} \mid (a,b) \in \mathcal{U}) = \sum_{p=1}^{2^K} \nu(\boldsymbol{\gamma}_p) \text{pr}(\text{declare } (a,b) \text{ a link} \mid \boldsymbol{\gamma}(a,b) = \boldsymbol{\gamma}_p).$$

The second error is the false negative which occurs when a pair from the matched set is classified as a non-link. The corresponding probability is

$$\pi_{\text{fn}} = \text{pr}(\text{declare } (a,b) \text{ a non-link} \mid (a,b) \in \mathcal{M}) = \sum_{p=1}^{2^K} \mu(\boldsymbol{\gamma}_p) \text{pr}(\text{declare } (a,b) \text{ a non-link} \mid \boldsymbol{\gamma}(a,b) = \boldsymbol{\gamma}_p).$$

If a linkage rule leads to π_{fp} and π_{fn}, then it is said that the linkage rule is at the levels $\pi_{\text{fp}}, \pi_{\text{fn}}$. The optimal linkage rule at some fixed admissible levels $\pi_{\text{fp}}, \pi_{\text{fn}}$ is defined as the rule that minimizes the probability of possible links. Hence, in this classical approach, the optimal linkage rule minimizes the amount of clerical review needed to resolve possible links while ensuring that errors made in classification are at the admissible level.

Fellegi & Sunter (1969) demonstrated that for an admissible pair of error levels $\pi_{\text{fp}}, \pi_{\text{fn}}$ the optimal linkage rule is

$$d(\boldsymbol{\gamma}_p) = \begin{cases} (1,0,0) \text{ if } T_{\text{fp}} \leq \mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p) \\ (0,1,0) \text{ if } T_{\text{fn}} < \mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p) < T_{\text{fp}} \\ (0,0,1) \text{ if } \mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p) \leq T_{\text{fn}} \end{cases}$$

where $T_{\text{fp}} = \mu(\boldsymbol{\gamma}_{p=x})/\nu(\boldsymbol{\gamma}_{p=x})$, $T_{\text{fn}} = \mu(\boldsymbol{\gamma}_{p=y})/\nu(\boldsymbol{\gamma}_{p=y})$ are decision thresholds obtained at the comparison patterns x and y, respectively. Here $(1,0,0)$ means that a link is declared with probability 1, $(0,1,0)$ that a possible link that requires clerical review is declared with probability 1, $(0,0,1)$ that a non-link is declared with the probability 1. Hence, all patterns satisfying $T_{\text{fp}} \leq \mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$ are declared links, while all patterns satisfying $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p) \leq T_{\text{fn}}$ are declared non-links.

In practical terms, employing the optimal linkage rule involves the estimation of $\mu(\boldsymbol{\gamma}_p)$ and $\nu(\boldsymbol{\gamma}_p)$ followed by ordering of all comparison patterns by the decreasing 'likelihood' or weight of a pair having this patterns. This 'likelihood' is based either directly on the ratio $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$ or some monotone increasing function of the ratio. For given admissible error levels $\pi_{\text{fp}}, \pi_{\text{fn}}$, the thresholds T_{fn} and T_{fp} can be determined by cumulative sums of $\mu(\boldsymbol{\gamma}_p)$ and $\nu(\boldsymbol{\gamma}_p)$ of the ordered patterns. That is, $\pi_{\text{fp}} <= \sum_{p=x(1)}^{x} \nu(\boldsymbol{\gamma}_p)$, where $x(1)$ is the index of pattern with the highest 'likelihood' or weight determined by $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$ or its function as above and x is the pattern for which the cumulative sum becomes less or equal than the admissible false positive error. Similarly, $\pi_{\text{fn}} <= 1 - \sum_{p=y}^{y(1)} \mu(\boldsymbol{\gamma}_p)$, where $y(1)$ is the pattern with the smallest weight, and y is the pattern that satisfies the given inequality for π_{fn} and such that the corresponding weight is smaller than the weight of any other pattern also satisfying the inequality. A worked-out example is provided later in Section 7.4.

Fellegi & Sunter (1969) proposed two approaches to estimate $\mu(\boldsymbol{\gamma}_p)$ and $\nu(\boldsymbol{\gamma}_p)$. We focus on the one that bears most resemblance to the approach we develop in this work. First, to make the estimation tractable, Fellegi and Sunter proposed a linkage model with binary comparisons of the values of linkage variables. Second, given the set of matches, agreements / disagreements on each of the K linkage variables are assumed to be independent. The same holds for the set of non-matches. This assumption is known as the conditional independence assumption given the match status. That is, the probability of agreement on the k^{th} linkage variable given a record pair belongs to the set of matches is μ_k, and the probability of disagreement is $1 - \mu_k$. Thus the probability of observing a certain pattern given that a record pair $(s_{1,a}, s_{2,b})$ belongs to the set of matches is $\prod_{k=1}^{K} \mu_k^{\gamma_k(a,b)}(1-\mu_k)^{1-\gamma_k(a,b)}$. Similarly, the probability of agreement on the k^{th} linkage variable given a record pair belongs to the set of non-matches is ν_k, and the probability of disagreement is $1-\nu_k$. The joint probability of a given comparison pattern given that a record pair $(s_{1,a}, s_{2,b})$ belongs to the set of non-matches is $\prod_{k=1}^{K} \nu_k^{\gamma_k(a,b)}(1-\nu_k)^{1-\gamma_k(a,b)}$. Let w be as above in this chapter and $\overline{m}$ denote the mean of M, then for the case of three linkage

variables, we have $k = 1, 2, 3$ and the following system of 7 equations can be solved to produce the estimates of μ_k, ν_k and the mean number of matches $\overline{m}$:

$$\begin{cases} w\mathbb{E}\left(\begin{smallmatrix}\text{proportion of agreements}\\ \text{on all variables except the } k^{th}\end{smallmatrix}\right) = \overline{m}\prod_{j=1,j\neq k}^{3}\mu_j + (w-\overline{m})\prod_{j=1,j\neq k}^{3}\nu_j \\ w\mathbb{E}\left(\text{proportion of agreements on the } k^{th}\right) = \overline{m}\mu_k + (w-\overline{m})\nu_k \\ w\mathbb{E}\left(\text{proportion of agreements on all variables}\right) = \overline{m}\prod_{k=1}^{3}\mu_k + (w-\overline{m})\prod_{k=1}^{3}\nu_k. \end{cases} \tag{7}$$

There are many other topics discussed in the paper by Fellegi and Sunter. What is of interest for this thesis is that in their paper Fellegi and Sunter explicitly state that the estimate of the number of matches (or the mean number of matches) is available upon solving the equation (7). Nevertheless, their approach and the majority of the approaches that evolved from their approaches remained classification-based and aimed to follow the optimal linkage rule by minimising the number of possible links to be resolved by clerical review.

While not wording it as the reason for not pursuing a pure estimation-based approach to linkage, Fellegi and Sunter give some explanations that probably influenced the classification-based approaches and discouraged estimation-focused ones. They warned that the conditional independence assumption given the match status is unlikely to hold in practice. They argued, nevertheless, that failure of this assumption, while resulting in inaccurate estimates of the linkage model parameters, will not affect substantially the ordering of comparison patterns based on a monotone increasing function of $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$. This is indeed one of the reasons why such a classification approach works in practice (see simulation results in Section 7.5). However, in this case the probabilistic nature of the linkage is in doubt since the estimates of π_{fp} and π_{fn} are incorrect. There exist models that allow the specification of models with dependencies between outcomes of linkage variables; see for instance Winkler (1993); Armstrong & Mayda (1993); Thibaudeau (1993); Belin & Rubin (1995); Larsen & Rubin (2001). Some of these approaches require training data. In this thesis, we will discuss the nature of these dependencies and will use models allowing for between linkage variables dependence without any training data and will show the effect of not accounting for this kind of dependence. In our view, accounting for between-variables dependencies is among the most important features of any approach that avoids classification and clerical resolution.

Another reason implied by Fellegi and Sunter is the excessive variability in the estimation approach presented in this chapter. There are two points regarding the excessive variability. First, while there may be indeed a substantial variability in the estimation of record linkage parameters, there is an effective way of dealing with it. Its use in the classical classification-based linkage is discussed in Section 2.2.4 and we show how to use it within the no-classification methods in Section 4.2. Second, the approaches to assess the variability suggested by Fellegi and Sunter are not quite adequate in the context of record linkage. This is discussed in Section 2.2.5 and Chapter 6.

There are possibly a few more reasons why a purely estimation-focused approach to linkage was not initially accepted. We will discuss these reasons as the development of the classification free approach progresses. One of the goals of this work is to re-ignite the interest and prospect of the solely estimation-based probabilistic record linkage.

2.2.3 Formulation of the Fellegi-Sunter approach in terms of mixture models

Since $\mathcal{W}$ is partitioned into two non-overlapping sets $\mathcal{M}$ and $\mathcal{U}$, it is sensible to think that the observed comparison outcomes $\boldsymbol{\gamma}(a, b)$ follow some mixture distribution with two mixture components. The corresponding probability mass or density function is

$$\text{pr}(\boldsymbol{\gamma}(a,b); \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi \text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M}; \boldsymbol{\mu}) + (1-\pi)\text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U}; \boldsymbol{\nu}), \tag{8}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are the vectors of probabilities of agreements on K linkage variables (or their combinations) given the set of matches and the set of non-matches, respectively. For instance, in the case of conditional independence given the match status presented in the previous section, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^T$. These vectors may contain probabilities that take into account dependencies between outcomes of linkage variables; see Section 2.2.5. The parameter π is the mixing proportion in mixture models language. In a record linkage model this is the proportion of matches among all record pairs. Note, the mixing proportion π should not be confused with the coverage probabilities π_1 and π_2. In principle, $\mathcal{W}$ can be partitioned into more than two non-overlapping sets if it is practical to do so (Winkler, 1993; Larsen & Rubin, 2001). However, it is easy to show that the parameters of a two component mixture can be related to dual system estimation and therefore only two component cases are considered in this thesis.

A mixture model is implicitly present in the system of equations (7). Say, the last equation when divided by w is the same as (8) for $\boldsymbol{\gamma}(a, b) = (1, 1, 1)^T$, and the rest of the equations can be viewed as different marginalisations that involve only μ_k and ν_k, but not $1-\mu_k$ and $1-\nu_k$. The first known explicit formulation of a record linkage model in terms of mixtures is apparently in Jaro (1989). Formulating record linkage using a mixture models approach may seem very alluring as it places the problem at hand into the vast area of results and estimation methods available for mixture models (Everitt & Hand, 1981; Titterington et al., 1985; McLachlan & Basford, 1988; McLachlan & Peel, 2000; Böhning, 2005; Frühwirth-Schnatter et al., 2019). There is a wide range of approaches to estimate parameters of mixture models, including method of moments, maximum likelihood estimation via the expectation-maximization and related algorithms, minimum distance methods, Bayesian methods, spectral methods (Frühwirth-Schnatter et al., 2019) and many more. It is unsurprising that researchers turned their attention to mixture models and maximum likelihood estimation (Jaro, 1989; Larsen & Rubin, 2001; Herzog et al., 2007, chap. 9.4), given all the important properties of maximum likelihood estimation (Zehna, 1966; Lehmann & Casella, 1998, chap. 6.3; Pawitan, 2013, chap. 9). Many Bayesian approaches also adopted a mixture model perspective in one way or another (Fortini et al., 2001; Larsen, 2005; Sadinle, 2017).

The problem with employing mixture models for record linkage tasks is that the data generating mechanism associated with these tasks is not a mixture in its usual sense. Therefore, certain estimation methods, including the method of maximum likelihood estimation, are not strictly valid in the form they are usually applied, even though these methods are capable of producing acceptable point estimates. Specifically for the maximum likelihood estimation this means that there are no desired properties of estimates when dealing with linkage. Yet, approaches with mixture models may perform better

than other methods (Jaro, 1989) and be easier to work with. More on the data generating mechanism in record linkage is presented in Section 2.2.5. Nevertheless, the mixture model-based perspective is worth discussing since it motivates the linkage model proposed later in this thesis. The proposed model is not a mixture in an ordinary sense, but is parametrised as a mixture, therefore we will call it the mixture-like model.

Most mixture model non-Bayesian record linkage approaches rely on the maximum likelihood estimation via the expectation-maximization algorithm (Dempster et al., 1977). This algorithm is an iterative technique for obtaining maximum likelihood estimates. It is primarily used in the situations where the maximum likelihood estimation by other means is either difficult or even impossible (McLachlan & Krishnan, 2008), for instance, in missing data situations. Record linkage can be viewed as a missing data problem. In this case, the membership indicator $g(a, b)$, that is whether a record pair belongs to the $\mathcal{M}$ or $\mathcal{U}$, is not observable:

$$g(a,b) = \begin{cases} 1 \text{ if } (a,b) \in \mathcal{M} \\ 0 \text{ if } (a,b) \in \mathcal{U}. \end{cases}$$

The complete data are not observable and comprised of record pairs membership of one of $\mathcal{M}$, $\mathcal{U}$ as well as their comparison outcomes: $(g(a,b), \ldots, \boldsymbol{\gamma}(a,b), \ldots)^T, a = 1, \ldots, n_1, b = 1, \ldots, n_2$. The expectation-maximization maximizes the observed or incomplete data likelihood

$$\prod_{\substack{(s_{1,a}, s_{2,b}),\\ a=1,\ldots,n_1,\\ b=1,\ldots,n_2}} \left[\pi \text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M}; \boldsymbol{\mu}) + (1-\pi)\text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U}; \boldsymbol{\nu})\right], \tag{9}$$

by maximizing the complete data likelihood

$$\prod_{\substack{(s_{1,a}, s_{2,b}),\\ a=1,\ldots,n_1,\\ b=1,\ldots,n_2}} \{\pi \text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M}; \boldsymbol{\mu})\}^{g(a,b)} \{(1-\pi)\text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U}; \boldsymbol{\nu})\}^{1-g(a,b)}. \tag{10}$$

Similarly to the presentation in Section 2.2.2, we consider the case with binary comparisons of the values of linkage variables and the conditional independence of comparisons given the match status, but this can be extended to more complex cases. Note that such conditional independence given the match status is often referred to as the latent class model (Vermunt & Magidson, 2004) outside the record linkage community. Under the conditional independence given a record pair is either in the set of matches $\mathcal{M}$ or the set of non-matches $\mathcal{U}$ with binary comparison outcomes, we have

$$\text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M}) = \prod_{k=1}^{K} \mu_k^{\gamma_k(a,b)} (1-\mu_k)^{1-\gamma_k(a,b)} \tag{11}$$

and

$$\text{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U}) = \prod_{k=1}^{K} \nu_k^{\gamma_k(a,b)} (1-\nu_k)^{1-\gamma_k(a,b)}.$$

Then the expectation-maximization algorithm is run to find solutions for $\boldsymbol{\mu}, \boldsymbol{\nu}$ and π. At the first step of the algorithm we use some initial guess for the values of parameters, and at iteration ι we use the estimates obtained at iteration $\iota - 1$. At each iteration, the algorithm has expectation and maximization step.

At the expectation step of iteration ι we compute the conditional expectation of the indicator function $\widehat{g}(a,b)^{(\iota-1)} = g\left(a, b \mid \boldsymbol{\gamma}(a,b), \widehat{\boldsymbol{\mu}}^{(\iota-1)}, \widehat{\boldsymbol{\nu}}^{(\iota-1)}, \widehat{\pi}^{(\iota-1)}\right)$ given the observed comparison pattern and the estimates $\widehat{\boldsymbol{\mu}}^{(\iota-1)}, \widehat{\boldsymbol{\nu}}^{(\iota-1)}$ and $\widehat{\pi}^{(\iota-1)}$ from the previous step

$$\widehat{g}(a,b)^{(\iota-1)} = \frac{\widehat{\pi} \prod_{k=1}^{K} \widehat{\mu}_k^{\gamma_k(a,b)} (1-\widehat{\mu}_k)^{1-\gamma_k(a,b)}}{\widehat{\pi} \prod_{k=1}^{K} \widehat{\mu}_k^{\gamma_k(a,b)} (1-\widehat{\mu}_k)^{1-\gamma_k(a,b)} + (1-\widehat{\pi}) \prod_{k=1}^{K} \widehat{\nu}_k^{\gamma_k(a,b)} (1-\widehat{\nu}_k)^{1-\gamma_k(a,b)}}. \tag{12}$$

Note, we omitted the step indicator $\iota - 1$ on the right-hand side of the above expression as it is already clunky; however, all the parameters are from the step $\iota - 1$. In fact, (12) is just the conditional expectation of a Bernoulli random variable $g(a,b)$ given that we observe the pattern $\boldsymbol{\gamma}(a,b)$ out of 2^K possible patterns. The numerator is the probability of a pair belonging to the set of matches (a success) and the denominator is the probability of observing the pattern $\boldsymbol{\gamma}(a,b)$.

At the maximization step, we obtain $\widehat{\mu}_k^{(\iota)}, \widehat{\nu}_k^{(\iota)}$ and $\widehat{\pi}^{(\iota)}$ that maximize $\widehat{g}(a,b)^{(\iota-1)}$. This is achieved by the usual means of maximization: differentiating the conditional expectation (12) with respect to each of the parameter of interest, equating the partial derivatives to 0 and solving the corresponding equations. The estimates are

$$\widehat{\mu}_k^{(\iota)} = \frac{\sum_{(a,b)} \widehat{g}(a,b)^{(\iota-1)} \gamma_k(a,b)}{\sum_{(a,b)} \widehat{g}(a,b)^{(\iota-1)}},\ \widehat{\nu}_k^{(\iota)} = \frac{\sum_{(a,b)} (1-\widehat{g}(a,b)^{(\iota-1)}) \gamma_k(a,b)}{\sum_{(a,b)} (1-\widehat{g}(a,b)^{(\iota-1)})},\ \widehat{\pi}^{(\iota)} = \frac{\sum_{(a,b)} \widehat{g}(a,b)^{(\iota-1)}}{w}.$$

The algorithm runs until a pre-specified level of tolerance achieved. Like many other optimization techniques, the expectation-maximization algorithm does not guarantee convergence to the global maximum, but almost always converges to a local maximum. Procedures like simulated annealing are required when searching for the global maximum.

Once the parameter estimates are obtained, the rest of the linkage usually follows the approach of Fellegi and Sunter presented above. That is, comparison patterns are ordered by a monotone function of $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$, which in the case of the conditional independence given the match status is often the natural logarithm. Then these patterns are classified as links and non-links to achieve the admissible level of errors and clerical resolution of possible links is undertaken. Ignoring that the maximum likelihood estimation of mixture model parameters is not the most statistically justified approach, it only simplifies the estimation of linkage model parameters in comparison to the original Fellegi and Sunter approach. It neither attempts to depart from the classification paradigm, nor circumvents the obstacles that prevented a pure estimation-focused approach in the first place, though, it is possible to specify a more complex model that accounts for between-variables associations of comparison outcomes (Larsen & Rubin, 2001). As before, it is not meant to produce good parameter estimates as such, but is meant to provide a good ordering of comparison patterns.

While we will not directly use proper mixture models, it is useful to be aware of some complexity of

mixtures that make dealing with this kind of model very challenging at times. At least some of these challenges are relevant for the mixture-like models. Most of the difficulties arise from several closely related concepts.

The first challenge is the identifiability of mixtures. Put simply, a model is said to be identifiable if all its parameters can be uniquely estimated from the observed data. In the case of mixture models, this definition is slightly relaxed to allow for label permutation in mixing proportions (Frühwirth-Schnatter et al., 2019). Since there exist many types of identifiability, such as global, generic, rational and local (Sullivant, 2018), there may sometimes be conflicting accounts on identifiability of certain types of mixture models. As an illustration, it has been proved that the mixtures of Bernoulli random variables such as (9) discussed above are non-identifiable (Gyllenberg et al., 1994). However, this model type is only globally non-identifiable and in work by Allman et al. (2009) it was established that the parameters of the finite mixtures of multivariate Bernoulli distributions are generically identifiable in the case of the latent class model, and conditions for such identifiability were provided. Practically, it means that such a model is identifiable with the exception of a few parameter values. We provide an overview of the methods to establish identifiability in discrete mixture and mixture-like models using the methods from computational commutative algebra and algebraic statistics in Section 2.5. In Chapter 5 we consider a set of useful record linkage models with four linkage variables and establish which of them are identifiable and which are not.

Another closely related issue is that the mixtures of binomial and multinomial distributions belong to the class of singular models (Watanabe, 2009). A model is regular if it is identifiable and its Fisher information matrix is positive definite. Models that are not regular are strictly singular (Watanabe, 2009). Note, that according to Watanabe 'identifiability is neither necessary not a sufficient condition of positive definiteness of the Fisher information matrix' (Watanabe, 2009, p. 10). Strictly singular models contain singularities that cannot be resolved by transformation or by restrictions made on the parameter space. As a result of singularity, certain properties of the maximum likelihood estimators may not hold. For instance, the goodness of fit cannot be reliably assessed by information criteria such as the Akaike information criterion or Bayesian information criterion.

Yet another layer of complexity is due to the fact that mixture (8) does not belong to the exponential family, but to the so-called stratified exponential family (Geiger et al., 2001). As a result, many nice properties of the exponential family do not hold in the case of mixtures. For instance, it is impossible to reduce information in the mixture through sufficiency since the only sufficient statistic is the vector of observed counts (Fienberg et al., 2009). In general, many properties of such mixtures are unknown.

Another issue with mixtures is that the data may be easily generated by a model with more parameters than it is possible to estimate. As an example, suppose we are dealing with a two-component mixture with K binary variables. The observed distribution of the model has 2^K observations, and it lies in the space of dimension $2^K - 1$. However, each of the components of the mixture on its own may have up to $2^K - 1$ parameters, plus the mixture parameter. This means that the most complex estimable model may be not complex enough to adequately model the distribution of interest and there is no easy way of testing whether this occurs.

There is also some confusion around whether the intercept terms in each mixture component should

be counted as parameters, when the mixture is represented as a log-linear model (Haberman, 1974; Vermunt & Magidson, 2004). From the theory of log-linear models, it is known that a log-linear model for Poisson sampling gives the same parameter estimates as the log-linear model for multinomial sampling (Agresti, 2002). In the case of mixtures, the equivalence is not that obvious. However, the probability product parameterization (8) suggests that the intercept terms are redundant.

2.2.4 One-to-one linkage

In many record linkage applications, especially in linking surveys of human populations, it is expected to have 1-to-1 matches underpinning the data. That is, assuming there are no duplicates in the surveys or duplicates were removed, every *record* on the first survey either matches to one and only one *record* on the second survey, or does not have a corresponding match. Recall, that the term match refers to the true status. The record linkage approaches presented above and those that will be developed later in this thesis are dealing primarily with record pairs, not the individual records. Therefore, it is often the case that two record pairs involving the same observation from one of the surveys, say pairs (a, b) and (a, c), have the same comparison outcome $\boldsymbol{\gamma}(a, b) = \boldsymbol{\gamma}_p$ and $\boldsymbol{\gamma}(a, c) = \boldsymbol{\gamma}_p$ with $\boldsymbol{\gamma}_p$ such that it should be declared a link under chosen linkage rule. Under a 1-to-1 match constraint and again assuming no duplications, or resolving them beforehand, $s_{1,a}$ can match only to one of $s_{2,b}$ or $s_{2,c}$. Hence, it is often desirable to enforce the 1-to-1 linkage when doing an actual record linkage exercise, which enforces the above condition on the classification outcomes.

Such 1-to-1 linkage can be achieved either naïvely or through formulation and solution of an optimisation problem. A naïve classification may simply involve declaring a link for the pair whose comparison pattern has the highest weight based on $\mu(\boldsymbol{\gamma}_p)/\nu(\boldsymbol{\gamma}_p)$. If both pairs have the same comparison pattern, then either some clerical involvement may be needed to resolve the situation, or a random decision may be made. Alternatively, a more sophisticated approach to 1-to-1 linkage is to formulate this task as an optimization problem. Jaro (1989) proposed to see the 1-to-1 linkage as the assignment problem that achieves the task in an optimal way. We discuss the assignment problem in more detail and employ it in the context of no-classification record linkage and related dual system estimation in Section 4.2. In no-classification record linkage the 1-to-1 constraint provides a useful source of information allowing the variability of no-linkage dual system estimates to be reduced substantially. The approach of Jaro (1989) with 1-to-1 linkage is essentially the Fellegi-Sunter approach with a mixture model, but after the parameter estimates are obtained, the assignment problem is solved and then the remaining pairs are declared as links or non-links based on the desired thresholds. All possible links are then clerically reviewed.

Bayesian (Fortini et al., 2001; Larsen, 2005; Sadinle, 2017) and maximum entropy-based (Lee et al., 2022) estimation approaches allow incorporation of the 1-to-1 constraint into the main model instead of doing it after the parameter estimation.

2.2.5 A linkage experiment and the invalidity of mixture models for record linkage

Statistical formulation of a record linkage task is not a trivial problem. It is not surprising that the approaches presented in Sections 2.2.2 and 2.2.3 have compromised certain aspects of the structure of

the data underlying record linkage. In this section we discuss why the mixture model representation of record linkage is not an appropriate representation. As a result, the maximum likelihood approach is not well justified and alternative model specifications and estimation methods are required. In Section 3.2 we will show that a mixture-like model, that is a model parametrized as a mixture but conceptually not being a mixture, is a more suitable model candidate for record linkage. Section 3.5 presents a well justified and flexible estimation method. For now, it is going to be demonstrated that the record pairs cannot be drawn independently of each other as the methods in Section 2.2.3 imply. In fact, the notion of drawing pairs does not seem applicable for record linkage in general. The insights on which this chapter is based are not new. At the dawn of automated record linkage, Newcombe et al. (1959) observed that record pairs obtained by comparing all the records of the first survey to all the records of the second survey are not necessarily independent, though the issue was not explicitly formulated as the lack of independence. More recently, Tancredi & Liseo (2011); Lee et al. (2022) turned their attention to this issue and proposed methods for record linkage that do not assume the independence of record pairs.

First, recall that a standard mixture model (8), with say two mixture components, implies the following sampling or data generating scheme of n_{mix} observations; see for instance McLachlan & Peel (2000, chap. 1). For the first observation a random draw of a mixing component is carried out. In the context of record linkage, this is drawing a pair's membership in either of $\mathcal{M}$ or $\mathcal{U}$. Next, the outcomes for the selected component are generated, that is the vector $\boldsymbol{\gamma}(a, b)$ given a probability mass or density function associated with the selected component in the previous step. Then, for the second and all the remaining observations, the same two-step procedure is repeated independently from any other draws. Having a closer look at the underlying sampling or data generating mechanism in the case of record linkage reveals that it cannot be represented by such a standard mixture.

In order to see it, we present a possible view on how the data used in record linkage are generated. We call this generating mechanism the record linkage experiment. The linkage experiment is a sort of thought experiment or an instruction for simulation. In this experiment we have total control over its course and observe every outcome, even those that are not observable in real situations. This is not a formal discussion, as making it formal would be complicated and would not have much benefit beyond what is provided. We keep using all the notation introduced up to this point. Our discussion is partly based on Tancredi & Liseo (2011).

While the presented experiment is applicable for many types of population, our focus is a human population $\mathcal{P} = \{e_i : i = 1, \ldots, \tau\}$, as everywhere else in this work, where e_i are elements of population, that is individuals. Discussing a human population will allow us to touch an important topic of why the comparison outcomes of different linkage variables may be associated in certain situations. The population $\mathcal{P}$ has a habitual organisation in a modern society. That is, individuals are nested in households or addresses. In addition, at least a reasonable proportion of households in the population have individuals that tend to share, or have very similar, attributes, such as surname, due to relationships between the members of a household. For simplicity, in this discussion we assume that households and addresses exactly coincide.

In accordance with the above discussion of the dual system estimation and record linkage, in

our linkage experiment the population size is a fixed parameter τ. Consider $k = 1, \ldots, K', K' \geq K$ attributes associated with every element e_i. Each of these attributes has a fixed corresponding set $\mathcal{R}_k$ of distinct and genuine values that this attribute can take. The size of this set, also called the range, is a fixed parameter $\rho_k \in \mathbb{Z}^+$. For instance, if attribute is the 'month of birth', then the corresponding set of possible values is $\mathcal{R} = \{\text{'January'}, \text{'February'}, \ldots, \text{'December'}\}$ and the corresponding range is $\rho = 12$. Some attributes may be uniformly or nearly uniformly distributed, while other are not and certain values from $\mathcal{R}_k$ are more prevalent in the population than others. Examples of such non-uniformly distributed attributes can be names and surnames – some are a lot more common than others. We are not specifying any particular probability distribution of members of $\mathcal{R}_k$, but we think that there is some distribution $\text{Dist}(\boldsymbol{\theta}_k, \mathcal{R}_k)$, where $\boldsymbol{\theta}_k$ is an unspecified vector of parameters. Some attributes, such as the full address, are 'parent' attributes to other attributes and the corresponding distribution $\text{Dist}(\boldsymbol{\theta}_k, \mathcal{R}_k)$ is then bivariate: one variate is for the attribute's value and another for the number of 'children' in the 'parent'. In this example, the number of 'children' would be the household size associated with the house number.

The linkage experiment starts by randomly generating values of the attributes for the population elements. The data generating mechanism is multistage, nested and complex, with intricate dependencies between certain attributes. What attribute we start with may depend on the population and geographical structure. Suppose we start with the full address attribute, which is a combination of a street name, house number or name and apartment number or name if applicable. A value is drawn from the corresponding $\text{Dist}(\boldsymbol{\theta}_k, \mathcal{R}_k)$. Since this is a 'parent' attribute, the size of the household related to the address is also drawn. Then the attributes of the first individual are drawn from the corresponding distributions. For instance, the surname value is drawn from the distribution of surnames. Some attribute values are associated with each other: the date of birth can be associated with the marital status or the highest degree achieved. Likewise, the surname can be associated with the ethnicity or country of birth attributes. Once all attributes for the first individual in the household are drawn and there is another individual in the household, their attributes are drawn next. Now values for certain attributes of the second individual may be associated with the values of the first one. If association exists, it can be such that certain attributes are exactly the same or are variants of the attribute of the first person. For instance, the surname may be exactly the same, or be a male, female or child variant of the surname's value of the first element. Examples of other associated attributes include relationship, marital status, tenure, ethnicity. Once the household is populated in the described way the next full address is generated. The value of the next full address may be strongly associated with the preceding value, so that it is a consecutive / preceding value. The process continues until all τ individual values are generated.

The next stage is to draw two samples S_1 and S_2 from the population $\mathcal{P}$, as presented in Sections 2.1.1 and 2.1.2, so that each element e_i has an equal probability of selection π_1 in the first survey, and π_2 of being selected in the second survey. The two selection processes are completely independent. This selection results in n_1 elements being selected in the first survey and n_2 in the second. Recall that in Section 2.1.2 we proposed to use the multinomial distribution associated with the sampling, which also leads to the survey sizes N_1 and N_2 being binomially distributed. We also said that this is not an

exact model for the given situation: the sampling in both surveys in the linkage experiment is without replacement. Hence, it is impossible to maintain the same π_j, $j = \{1, 2\}$ for every e_i, and N_1 and N_2 are not exactly binomially distributed. An alternative could be the hypergeometric distribution for M (Seber, 1982) conditional on fixed n_1 and n_2. However, it is often difficult to work analytically with the hypergeometric distribution. Also, in our sampling mechanism we want N_1 and N_2 to be random variables. For human populations in general, and particularly in the census coverage situations, it is virtually impossible to achieve a fixed sample size, that is fix the number of individuals observed in a study. This is not only due to the fact that it is hard to plan and control the 'catch' effort. The main reason stems from dealing with large human populations where the primary sampling units are some geographical areas and the ultimate sampling units are households. Therefore, there is an inevitable variation in the sizes of samples. We accept this discrepancy between the theoretical conceptualization presented earlier and the sampling mechanism used in the linkage experiment, but for large τ there should be no substantial differences between the two. In fact, in the simulation work in Chapter 7, sampling without replacement is used and the theoretical and empirical results agree well.

The third stage of the linkage experiment is to record the values of attributes of interest into the values of record linkage variables $k = 1, \ldots, K$ on each of the surveys. This process of recording is subject to errors such as missingness, mistyping, mishearing, scanning, etc. Errors occur randomly where the probability of error occurring in the k^{th} variable of S_j is $\xi_{j,k}$, $j = \{1, 2\}$.

Once the surveys are drawn and the values of attributes are recorded with some errors, the Cartesian product of S_1 and S_2 is taken that produces the set $\mathcal{W}$ of all possible record pairings. Every record pair $(a, b) \in \mathcal{W}$ belongs to one and only one subset of $\mathcal{W}$: ether the set of matches $\mathcal{M}$ or the set of non-matches $\mathcal{U}$. Here $\text{id}(s_{1,a})$, $\text{id}(s_{2,b})$ are unique identifier attributes of the corresponding elements in the population. These attributes are never observed on the actual surveys, but are known in the linkage experiment.

For every record pair $(a, b) \in \mathcal{W}$ the value of the linkage variable as recorded on the first survey is compared with the value of the same variable as recorded on the second survey for all K variables in turn. Comparison on each of the variables gives a binary outcome $\gamma_k(a, b)$ of agreement of the values, $\gamma_k(a, b) = 1$, or disagreement of values / missing value on one or both surveys, $\gamma_k(a, b) = 0$. The agreement of the values need not be exact, but based on some accepted score of a specific distance function. One can think that if the comparisons are not exact and some threshold exists for binary classification of outcomes, then the use of fuzzy comparisons changes the error probabilities $\xi_{j,k}$, $j = \{1, 2\}$ as well as the size of the set of possible values ρ_k. The vector of comparisons is $\boldsymbol{\gamma}(a, b) = (\gamma_1(a, b), \ldots, \gamma_K(a, b))^T$.

Now for a pair in the set of matches, $(a, b) \in \mathcal{M}$, whether the comparison on the k^{th} variable yields $\gamma_k(a, b) = 1$ or $\gamma_k(a, b) = 0$, depends on whether an error occurred in the k^{th} variable in either of the surveys or simultaneously on both of them. That is, whether one or both of $s_{1,a,k}$, $s_{1,b,k}$ captured the attribute's value with an error. Let $\boldsymbol{\gamma}_{k,l}(a, b) = (\gamma_k(a, b), \gamma_l(a, b))^T$ be the comparison outcome on two variables, $(a, b) \in \mathcal{M}$. If for one of the surveys $S_j, j = \{1, 2\}$, an error made in v_k affects whether or not the error was made in v_l, or an error present in variable v_k in survey S_j affects the probability of an error in variable v_l in S_o, $j \neq o$, then the outcomes of $\gamma_k(a, b)$ and $\gamma_l(a, b)$ will be dependent. We

write the corresponding conditional probabilities

$$\mu_{l|k(1)}(a,b) = \text{pr}(\gamma_l(a,b) = 1 \mid \mathcal{M}, \gamma_k(a,b) = 1), \tag{13}$$

$$\mu_{l|k(0)}(a,b) = \text{pr}(\gamma_l(a,b) = 1 \mid \mathcal{M}, \gamma_k(a,b) = 0),$$

for binary outcomes we have

$$1 - \mu_{l|k(1)}(a,b) = \text{pr}(\gamma_l(a,b) = 0 \mid \mathcal{M}, \gamma_k(a,b) = 1),$$

$$1 - \mu_{l|k(0)}(a,b) = \text{pr}(\gamma_l(a,b) = 0 \mid \mathcal{M}, \gamma_k(a,b) = 0).$$

On the other hand, for a pair in the set of non-matches, $(a,b) \in \mathcal{U}$, the comparison on the k^{th} variable yields $\gamma_k(a,b) = 1$ if by chance $s_{1,a,k} = s_{1,b,k}$ or the value of $f(s_{1,a,k}, s_{1,b,k})$ is above a chosen acceptance thresholds for some edit distance function f. Say, two common first names agree for two different individuals in the population. Consider $\boldsymbol{\gamma}_{k,l}(a,b) = (\gamma_k(a,b), \gamma_l(a,b))^T$, $(a,b) \in \mathcal{U}$. Now if the values of v_l do not depend on the values of v_k in the population, then the comparison outcomes $\gamma_k(a,b)$ and $\gamma_l(a,b)$ will be independent. However, if the values of v_l depend on the values of v_k in the population, then the comparison outcomes $\gamma_k(a,b)$ and $\gamma_l(a,b)$ will be associated. From the data generating mechanism of the attributes of $e_i \in \mathcal{P}$, we see that this association happens mainly as the result of nested data structure. That is, if v_k is the 'parent' attribute relative to v_l, and the attribute v_l is such that the 'children' tend to share the value of this attribute, then the comparisons $\gamma_l(a,b)$ will tend to be equal to 1 given $\gamma_k(a,b) = 1$. For example, if v_k is the full address and v_l is the surname, than given that the address agrees, there will be more agreements on the surname for the pairs that belong to the set of non-matches than for the members of two different addresses in general. Therefore, outcomes for $\gamma_k(a,b)$ and $\gamma_l(a,b)$ are likely to be dependent for the 'parent'–'child' variables and we expect to see the excess of $\boldsymbol{\gamma}_{k,l}(a,b) = (1,1)^T$. We write the corresponding conditional probabilities

$$\nu_{l|k(1)}(a,b) = \text{pr}(\gamma_l(a,b) = 1 \mid \mathcal{U}, \gamma_k(a,b) = 1), \tag{14}$$

$$\nu_{l|k(0)}(a,b) = \text{pr}(\gamma_l(a,b) = 1 \mid \mathcal{U}, \gamma_k(a,b) = 0).$$

Again, binary outcomes mean that

$$1 - \nu_{l|k(1)}(a,b) = \text{pr}(\gamma_l(a,b) = 0 \mid \mathcal{U}, \gamma_k(a,b) = 1),$$

$$1 - \nu_{l|k(0)}(a,b) = \text{pr}(\gamma_l(a,b) = 0 \mid \mathcal{U}, \gamma_k(a,b) = 0).$$

This illustrates how the dependencies between comparison outcomes of different linkage variables originate. We will usually refer to this dependence as a between-variables dependence or between-variables association. The between-variables dependencies given the set of matches will stem from the error generating mechanism. It is reasonable to expect that in high quality surveys and with good edit distance functions, the extent of such a dependence may be minimised. Alternatively, between-variables dependencies in the set of matches should be dealt with by appropriate model specification.

Whereas the between-variables dependencies given the set of non-matches have a different nature and will stem from the population structure. To avoid these dependencies, one may choose linkage variables that evade the situation described above. If a choice leading to the between-variables dependencies in the set of non-matches is unavoidable, then appropriate identifiable model specifications must be used.

Going back to the linkage experiment, it is important to note that apart from the generation of population values, sampling survey and making errors when recording the values of linkage variables there is no randomness in the experiment. So the process of comparing the values of linkage variables does not have any randomness in itself and is completely determined by the three earlier stages. Therefore, comparison outcomes are in a sense produced by a deterministic process. In addition, we observe that each of n_1 elements of the first survey is paired with every one of n_2 elements of the second survey. Because the comparison outcomes are deterministic given the sampling of population elements and the errors made in recording the values of the linkage variables, any two record pairs that share the same element of either of the lists also share some common information. This leads to a certain restriction on the comparison outcomes that any two pairs sharing the same element can take from the space of all possible comparison values. The deterministic nature of comparisons and the fact that the vast majority of record pairs share information imply that in general the observed comparison outcomes for a given variable k cannot be treated as independent. Note that in this case there is a lack of independence in outcomes for a given linkage variable, so we are dealing with a sort of within-variable dependence. This should not be confused with the dependence of comparison outcomes between several linkage variables (between-variables dependence) discussed earlier in this section. To illustrate the within-variable dependence, consider a simple example with two records on the first survey, $s_{1,a}, s_{1,d}$ and two records on the second survey $s_{2,b}, s_{2,c}$ and let v_k be the linkage variable 'month of birth'. There are four record pairs in this case: $(a,b),(a,c),(d,b),(d,c)$. Suppose, the recorded value of $s_{1,a}$ is $s_{1,a,k}$ = 'January'. Then, assuming exact comparison for simplicity, $\gamma_k(a,b) = 1$ if $s_{2,b,k}$ = 'January'. Then $\gamma_k(a,c) = 1$ also if $s_{2,c,k}$ = 'January'. Now if $\gamma_k(d,b) = 1$ then $s_{1,d,k}$ = 'January' and $\gamma_k(d,c)$ must be 1. Contrarily, if $\gamma_k(d,b) = 0$ then $s_{1,d,k}$ is any month except 'January' and $\gamma_k(d,c)$ must be 0. Overall, there are only 12 unique outcomes of comparisons for these four pairs:

$$
\begin{aligned}
&\gamma_k(a,b) = 1, \gamma_k(a,c) = 1, \gamma_k(d,b) = 1, \gamma_k(d,c) = 1,\\
&\gamma_k(a,b) = 1, \gamma_k(a,c) = 1, \gamma_k(d,b) = 0, \gamma_k(d,c) = 0,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 0, \gamma_k(d,b) = 1, \gamma_k(d,c) = 1,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 0, \gamma_k(d,b) = 0, \gamma_k(d,c) = 0,\\
&\gamma_k(a,b) = 1, \gamma_k(a,c) = 0, \gamma_k(d,b) = 1, \gamma_k(d,c) = 0,\\
&\gamma_k(a,b) = 1, \gamma_k(a,c) = 0, \gamma_k(d,b) = 0, \gamma_k(d,c) = 1,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 1, \gamma_k(d,b) = 1, \gamma_k(d,c) = 0,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 1, \gamma_k(d,b) = 0, \gamma_k(d,c) = 1,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 0, \gamma_k(d,b) = 0, \gamma_k(d,c) = 1,\\
&\gamma_k(a,b) = 0, \gamma_k(a,c) = 0, \gamma_k(d,b) = 1, \gamma_k(d,c) = 0,
\end{aligned}
$$

$$\gamma_k(a,b)=0,\gamma_k(a,c)=1,\gamma_k(d,b)=0,\gamma_k(d,c)=0,$$
$$\gamma_k(a,b)=1,\gamma_k(a,c)=0,\gamma_k(d,b)=0,\gamma_k(d,c)=0,$$

while none of the combinations of three agreements and one disagreement is possible.

Even more important, there is no sequential independent draws of observations as described in the beginning of this section. Since our ultimate observations are comparison patterns for record pairs, and record pairs are generated in bulk by taking the Cartesian product, there is no decision on which component the current observation will come from. Instead, every record from the first survey is necessarily paired with n_2 observations from the second survey. Also, there is no usual way of controlling how many observations to draw. For instance, the only way to achieve a prime number of overall record pairs w, is to have $n_1 = 1$ and n_2 a prime number, or vice versa. Another example of a peculiar behaviour around the sample size is the impossibility to increase in general the number of record pairs, w, by 1. Sampling an additional record in S_1 results in w increasing by n_2, and sampling an additional record in S_2 results in w increasing by n_1. In addition, under the 1-to-1 match restriction, only a single pair may belong to the set of matches after sampling a single additional record in S_1 or S_2. If a particular record on S_1 is compared to n_2 records of S_2, it is necessary that there are either n_2 or $n_2 - 1$ corresponding pairs in the set of non-matches $\mathcal{U}$.

All in all, these two points demonstrate that the process of generating the ultimate observations, that is comparison outcomes, in the linkage experiment does not follow a regular mixture distribution. It also means that the observed joint probability (or corresponding likelihood) cannot be factorised as in (9):

$$\begin{aligned}\mathrm{pr}(\boldsymbol{\gamma}(a,b),\boldsymbol{\gamma}(a,c),\dots) \neq [\pi \mathrm{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M};\boldsymbol{\mu}) + (1-\pi)\mathrm{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U};\boldsymbol{\nu})] \times \\ [\pi \mathrm{pr}(\boldsymbol{\gamma}(a,c) \mid \mathcal{M};\boldsymbol{\mu}) + (1-\pi)\mathrm{pr}(\boldsymbol{\gamma}(a,c) \mid \mathcal{U};\boldsymbol{\nu})] \times \dots .\end{aligned} \tag{15}$$

Also, the complete joint probability (or corresponding likelihood) cannot be factorized as in (10):

$$\begin{aligned}\mathrm{pr}(g(a,b),g(a,c),\dots) \neq &\{\mathrm{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{M};\boldsymbol{\mu}\}^{g(a,b)}\{(1-\pi)\mathrm{pr}(\boldsymbol{\gamma}(a,b) \mid \mathcal{U};\boldsymbol{\nu})\}^{1-g(a,b)} \times \\ &\{\mathrm{pr}(\boldsymbol{\gamma}(a,c) \mid \mathcal{M};\boldsymbol{\mu}\}^{g(a,c)}\{(1-\pi)\mathrm{pr}(\boldsymbol{\gamma}(a,c) \mid \mathcal{U};\boldsymbol{\nu})\}^{1-g(a,c)} \times \dots .\end{aligned} \tag{16}$$

Hence, the maximum likelihood estimation approach presented in Section 2.2.3, being based on maximisation of the observed data likelihood via the maximisation of the complete data likelihood is, strictly speaking, not correct. Despite theoretically not being the most appropriate approach, the maximum likelihood estimation using the expectation-maximization algorithm (whether referred to as the maximum likelihood or not) in record linkage applications has been in use for a long time and has not been invalidated (Winkler, 2002; Herzog et al., 2007, chap. 9; Christen, 2012, chap. 6.3). Among explanations for why it might still work see Lee et al. (2022).

While the maximum likelihood approach is capable of producing reasonable and practically useful estimates of the parameters of the record linkage model specified as a regular mixture, without the mixture being a good representation of the data generating mechanism, there are several issues with this approach. First, we cannot be sure that the maximum value of the likelihood attained gives the best solution. Second, with a regular mixture any attempts at variance estimation using the mixture

representation will be wrong. In fact, multiple simulated datasets generated according to the linkage experiment show that the data do not follow a multinomial distribution, something that is be expected in a regular mixture; results are presented in Section 7.7. Some research in record linkage assumes a multinomial distribution for the outcomes of a linkage exercise, for instance Chipperfield & Chambers (2015), which is unlikely to be the case in reality.

In this work we will show how a mixture-like parameterization makes sense in the case of record linkage, will resort to Markov chain Monte Carlo based methods for parameter estimation that are more in line with the nature of the problem and will show how valid variance estimates can be obtained.

2.2.6 Notation in a context

In this section through an artificial example of a human population $\mathcal{P}$ we show how notation introduced earlier corresponds to data encountered in record linkage tasks. A small sample of observations from this population is displayed in Table 4. The true population size is $\tau = 250$ individuals in some geographical domain, but τ is unknown to the subject conducting the record linkage exercise. The columns of the table include a unique index i of the corresponding element e_i (this unique identifying index only exists in an abstract sense, there are no real unique identifiers in the population) and four attributes associated with the population elements. For simplicity, we ignore the data preparation process and these attributes are the same as the linkage variables used in this record linkage exercise. The variable v_1 is the true standardized address, v_2 is the true surname, v_3 is the true first name and v_4 is the true full date of birth. Here, individuals are nested within households and each address for simplicity contains exactly one household. The values as presented in Table 4 are the true values and any attempt to collect these values may results in some errors. The column with dots indicate that there exist other population attributes.

Table 4: Population $\mathcal{P}$

i	e_i	v_1	v_2	v_3	v_4	...
1	e_1	1 Census road	Miller	John	23-07-1989	...
2	e_2	1 Census road	Miller	George	03-01-2020	...
3	e_3	1 Census road	Miller	Mary	15-08-1989	...
4	e_4	2 Census road	Lee	Andrew	21-04-1991	...
5	e_5	2 Census road	McGann	Clare	11-02-1993	...
6	e_6	3 Census road	Conrad	George	03-01-2020	...
7	e_7	3 Census road	Conrad	Susana	27-10-1990	...
8	e_8	4 Census road	Anderson	Mike	15-05-1985	...
9	e_9	5 Census road	Clark	Jane	01-08-1992	...
10	e_{10}	5 Census road	Keegan	Peter	13-01-1993	...
...	...	...	...	...	...	...
250	e_{250}	...	...	...	...	...

Two independent surveys of the population are S_1 and S_2. Both surveys attempt to collect information on all occupants of all households in the population. The first survey, S_1, captures individuals $e_1, e_2, e_3, e_6, e_7, e_8, e_9, e_{10}, \ldots$. The size of S_1 is $n_1 = 231$. The second survey, S_2, captures individuals $e_1, e_2, e_3, e_4, e_6, e_7, \ldots$. The size of this survey is $n_2 = 222$. The data as collected in these surveys are displayed in Table 5 and Table 6. The columns of these tables are: index a of a given record $s_{1,a}$ on

survey S_1 and b of a given record $s_{2,b}$ on survey S_2; the corresponding identifying index, $\mathrm{id}(s_{1,a}) = i$ or $\mathrm{id}(s_{2,b}) = i$, of the population element e_i followed by four linkage variables which are exactly the same as in the population table. Clearly, n_1 and n_2 are smaller than the population size τ due to both the entire household non-response and the within household non-response. An example of the former is the entire household with individuals e_4 and e_5 missing in S_1, an example of the latter is the missing individual e_5 in S_2.

Some of the values of the population attributes get recorded incorrectly or are not recorded at all. In our example we have the following errors in S_1: the survey record $s_{1,3}$ that captures the population element e_3 gets the date of birth recorded incorrectly, the survey record $s_{1,5}$ that captures element e_7 has the name recorded as 'Susan' instead of 'Susana' and the survey record $s_{1,8}$ contains missing information on the date of birth of the individual e_{10}. The errors in S_2 are: a survey record $s_{2,2}$ misses the surname and date of birth of the individual e_2, the record $s_{2,3}$ incorrectly records the name of individual e_3, and records $s_{2,5}$ and $s_{2,6}$ captures the address of the individuals e_6 and e_7 incorrectly.

Table 5: Survey S_1

a	$s_{1,a}$	$\mathrm{id}(s_{1,a})$	v_1	v_2	v_3	v_4
1	$s_{1,1}$	1	1 Census road	Miller	John	23-07-1989
2	$s_{1,2}$	2	1 Census road	Miller	George	03-01-2020
3	$s_{1,3}$	3	1 Census road	Miller	Mary	15-08-1999
4	$s_{1,4}$	6	3 Census road	Conrad	George	03-01-2020
5	$s_{1,5}$	7	3 Census road	Conrad	Susan	27-10-1990
6	$s_{1,6}$	8	4 Census road	Anderson	Mike	15-05-1985
7	$s_{1,7}$	9	5 Census road	Clark	Jane	01-08-1992
8	$s_{1,8}$	10	5 Census road	Keegan	Peter	NA
...	...	...	...	...	...	...
231	$s_{1,231}$	...	...	...	...	...

Table 6: Survey S_2

b	$s_{2,b}$	$\mathrm{id}(s_{2,b})$	v_1	v_2	v_3	v_4
1	$s_{2,1}$	1	1 Census road	Miller	John	23-07-1989
2	$s_{2,2}$	2	1 Census road	NA	George	NA
3	$s_{2,3}$	3	1 Census road	Miller	Carey	15-08-1989
4	$s_{2,4}$	4	2 Census road	Lee	Andrew	21-04-1991
5	$s_{2,5}$	6	1 Census road	Conrad	George	03-01-2020
6	$s_{2,6}$	7	1 Census road	Conrad	Susana	27-10-1990
...	...	...	...	...	...	...
222	$s_{2,222}$	...	...	...	...	...

A record pair is a tuple like $(s_{1,4}, s_{2,3})$ and the classical record linkage attempts to classify these pairs. The observed set of all ordered pairs is $\mathcal{W} = \{(s_{1,1}, s_{2,1}), (s_{1,1}, s_{2,2}), (s_{1,1}, s_{2,3}), \dots\}$. An example of a match is the record pair $(s_{1,4}, s_{2,5})$ as both records refer to the same population element e_6. The set of matches $\mathcal{M} = \{(s_{1,1}, s_{2,1}), (s_{1,2}, s_{2,2}), (s_{1,3}, s_{2,3}), (s_{1,4}, s_{2,5}), \dots\}$ is unobservable as is the set of non-matches $\mathcal{U} = \{(s_{1,1}, s_{2,2}), (s_{1,2}, s_{2,3}), (s_{1,2}, s_{2,4}), \dots\}$. An example of a non-match is the record pair $(s_{1,4}, s_{2,3})$, since $s_{1,4}$ refers to the population element e_6, whereas $s_{2,3}$ refers to the population element

e_3.

The following is an example of exact comparison of the values of a linkage variable k for a given record pair (a, b), $\gamma_k(s_{1,a}, s_{2,b}) = \gamma_k(a, b)$: comparing the values of the surname variable, v_2, of a record pair $(s_{1,1}, s_{2,1})$ with the binary comparison outcome. Here, we are comparing 'Miller' with 'Miller' and $\gamma_2(s_{1,1}, s_{2,1}) = \gamma_2(1, 1) = 1$. Another example, the exact comparison of the values of the first name linkage variable, v_3, of a pair $(s_{1,5}, s_{2,6})$. Here we are comparing the values 'Susan' and 'Susana' and $\gamma_3(s_{1,5}, s_{2,6}) = \gamma_3(5, 6) = 0$.

Comparing all variables of some record pair produces a certain comparison pattern, $\boldsymbol{\gamma}(s_{1,a}, s_{2,b}) = \boldsymbol{\gamma}(a, b) = (\gamma_1(a, b), \ldots, \gamma_K(a, b))^T$. For instance, when performing exact comparison of the values of linkage variables for the pair $(s_{1,5}, s_{2,6})$ with the binary comparison outcome, we get the following comparison patter $\boldsymbol{\gamma}(s_{1,5}, s_{2,6}) = \boldsymbol{\gamma}(5, 6) = (\gamma_1(5, 6), \gamma_2(5, 6), \gamma_3(5, 6), \gamma_4(5, 6))^T = (0, 1, 0, 1)^T$.

Because of the errors, a matching record pair can have a comparison patterns with some (or even all) disagreements. Say, in the above example, a pair $(s_{1,5}, s_{2,6}) = (5, 6) \in \mathcal{M}$, but there are disagreements on several linkage variables in the corresponding comparison outcome $(0, 1, 0, 1)^T$. On the other hand, a non-matching pair can have agreements on some (or even all) linkage variables by chance or due to errors or missingness. Say, a pair $(s_{1,2}, s_{2,5}) = (2, 5) \in \mathcal{U}$, but the corresponding comparison pattern is $(1, 0, 1, 1)^T$.

Comparisons need not to be exact and some distance or similarity functions may be helpful. The resulting comparison score can be then used to decide whether values agree or disagree based on some chosen acceptance threshold. For instance, we have already seen an example of a pair $(s_{1,5}, s_{2,6})$ where the values 'Susan' and 'Susana' were compared leading to $\gamma_3(s_{1,5}, s_{2,6}) = \gamma_3(5, 6) = 0$. Alternatively, we could try, say, the Jaro-Winkler similarity function, f_{JW}, and treat scores equal or greater than 0.9 as agreements and as disagreements otherwise. The Jaro-Winkler similarity in this case is f_{JW}('Susan', 'Susana') ≈ 0.96, and this comparison is regarded as agreement, so that $\gamma_{3,JW,0.9}(5, 6) = 1$.

For a given record pair, say, $(s_{1,5}, s_{2,6}) = (5, 6)$ using a mixture model (8) for record linkage gives

$$\begin{aligned}\mathrm{pr}(\boldsymbol{\gamma}(5, 6) =& (0, 1, 0, 1)^T; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \\ & \pi \mathrm{pr}(\boldsymbol{\gamma}(5, 6) = (0, 1, 0, 1)^T \mid \mathcal{M}; \boldsymbol{\mu}) + (1 - \pi)\mathrm{pr}(\boldsymbol{\gamma}(5, 6) = (0, 1, 0, 1)^T \mid \mathcal{U}; \boldsymbol{\nu}),\end{aligned}$$

with the vector of probabilities $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ reflecting any possible between-variables associations of comparisons. Suppose, we have a reason to believe, that within the set of matches all comparisons between variables are independent. Then $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_4)^T$ with $\mu_k(5, 6) = \mathrm{pr}(\gamma_k(5, 6) = 1 \mid (5, 6) \in \mathcal{M})$. At the same time, the population structure is such that the comparison outcomes of the address and surname variables must be associated. In this case, the conditional probability that the surname agrees if the address agrees for a non-matching pair $(a, b) = (1, 3)$ is $\nu_{l|k(1)}(a, b) = \nu_{2|1(1)}(1, 3) = \mathrm{pr}(\gamma_2(1, 3) = 1 \mid \gamma_1(1, 3) = 1, (1, 3) \in \mathcal{U})$, and the conditional probability that the surname agrees if the address disagrees is $\nu_{l|k(0)}(a, b) = \nu_{2|1(0)}(1, 3) = \mathrm{pr}(\gamma_2(1, 3) = 1 \mid \gamma_1(1, 3) = 0, (1, 3) \in \mathcal{U})$. The mixture-like models for record linkage that are introduced and discussed in Sections 3.1 and 3.2 are conceptually different from regular mixtures, but the notation is very similar.

Most often, we are interested in the frequencies, f_p, of the comparison patterns $\boldsymbol{\gamma}_p$. The estimation

methods employed in this thesis use such frequencies as an input after the comparison outcomes of the individual record pairs are established. Individual pair information is only required by the methods that take into account 1-to-1 constraint. In fact, the expectation-maximization approach presented in Section 2.2.3 can be run very efficiently when the frequencies of comparison patterns are obtained first. An example of the frequencies generated by comparing records of S_1 and S_2 is provided Table 7. Note that in this thesis there is no fixed correspondence between the index p and the value taken by a comparison pattern $\boldsymbol{\gamma}_p$. In most cases, we are just interested in an arbitrary pattern $\boldsymbol{\gamma}_p$. Whenever we are dealing with examples where we are interested in a few specific $\boldsymbol{\gamma}_p$, the correspondence between the index and the patterns is evident from the context.

Table 7: Example of frequencies of the comparison patterns

p	$\boldsymbol{\gamma}_p$	f_p
1	0000	24626
2	0001	243
3	0010	94
4	0011	2
5	0100	144
6	0101	1
7	0110	3
8	0111	10
9	1000	209
10	1001	7
11	1010	5
12	1011	14
13	1100	272
14	1101	18
15	1110	14
16	1111	134

Finally, suppose some classification-based approach is used and all pairs $(a, b) \in \mathcal{W}$ are classified into links and non-links. The examples of true positive, true negative, false positive and false negative could be as follows. If a pair $(s_{1,1}, s_{2,1})$ is classified as a link, then it is a true positive since this pair is a match. If a pair $(s_{1,7}, s_{2,6})$ is classified as a non-link, then it is a true negative since this pair is a non-match. If a pair $(s_{1,3}, s_{2,3})$ is classified as a non-link, this is a false negative since this pair is a match. If a pair $(s_{1,2}, s_{2,5})$ is classified as a link, this is a false positive since this pair is a non-match.

2.3 An overview of census coverage estimation

Among the most remarkable applications of dual system estimation and record linkage is census coverage estimation and adjustment. In fact, many important developments in the theory and practice of dual system estimation of human populations and record linkage originated or matured within the field of census coverage estimation (Wolter, 1986; Jaro, 1989; Winkler & Thibaudeau, 1991; Hogan, 1992, 1993; Anderson & Fienberg, 1999; Brown, 2000; US Census Bureau, 2008; Brown et al., 2019). It may seem at first that there is a contradiction between the definition of a census as a complete survey of a population and two incomplete surveys of the population that are used in the dual system estimation of the total. In fact, any census of a large human population is imperfect and contains coverage

errors such as undercoverage (missingness of population elements) and overcoverage (elements being counted more than once or counted in a wrong location). Furthermore, these errors are not always small enough to be ignored and vary substantially by domains in the population; see for instance ONS (2012). Hence, often the extent of the census coverage error needs to be estimated. In the case of the United Kingdom countries, the census is actually adjusted for the coverage errors (Brown et al., 1999; Cabinet Office, 2008, chap. 5; Cabinet Office, 2018, chap. 4). While coverage estimation combines together undercoverage and overcoverage estimation as well as adjustment for biases, the dual system estimation itself deals specifically with the undercoverage error. The first survey is the census itself, and the second is the post-enumeration survey, which conceptually has its origins in the 1950 Census of the United States (Marks et al., 1953). This survey is usually called the Census coverage survey in the Census of England and Wales and other United Kingdom's countries. In this chapter we are mostly describing the Census of England and Wales.

Census coverage estimation is a fine example of how a careful design and implementation of data collection can substantially facilitate successful record linkage and dual system estimation. The main goal of this section is to overview these design features and how they fit together, rather than provide a full account of the intricacies of the coverage estimation. Certain of these design features are important for making the classification free linkage and related dual system estimation feasible in practice. On the other hand, such classification free approaches for linkage and estimation seem very suitable for the census-like estimation. In Section 2.2.1 we stressed how much effort is needed when preparing the data for record linkage, in this section we want to stress how much a structured and planned data collection can contribute to high quality outputs or enable a higher degree of automated processing at the later stages of data analysis.

Many modern censuses use the post-out model: either a paper questionnaire or an access code to an electronic questionnaire is posted to all the addresses on the address frame first. At the core of the census data collection lies an address frame which attempts to compile all the addresses in the population of interest as best as possible (ONS, 2010, 2023). Usually, some response chasing schedule exists that involves sending a reminder letter to the non-responding addresses after a certain amount of time. If no return is made after a reminder, the field interviewer may attempt to remind or collect the information from the address (ONS, 2015, chap. 3; Cabinet Office, 2018, chap. 4, ONS, 2021a, chap. 3). In any case, the vast majority of census returns are made for the addresses on the address frame and a very good quality frame is essential. In addition, the questionnaires are designed to confirm or rectify the addresses they were sent to. For example, in the 2021 Census of England and Wales the online questionnaire had an address look-up functionality (ONS, 2021b). The paper questionnaire had an address, to which it was posted, printed on the front page and fields for a corrected address were provided, in case the address was not accurate (ONS, 2021c).

Household and person attributes collected on the census questionnaire enable good quality linkage and coverage estimation. These attributes include explicitly collected name, middle name and surname as well as the full date of birth (ONS, 2021c). Such explicit collection simplifies data cleaning and standardization, but even in this case the preprocessing is far from simple.

Another key success factor in the census coverage estimation is the Census coverage survey (Brown

et al., 2011; Castaldo, 2018; Burke & Račinskij, 2020). This is a re-count of a population within a large sample (approximately 340,000 households in England and Wales are sampled) carried out independently of the census. The coverage survey takes place several weeks after the census reference date (census day) in the form of face-to-face interviews. The fact that the survey starts soon after the census day ensures that the target population remains nearly closed. This survey data collection is operationally as independent of the census data collection as possible. Such operational independence contributes to statistical dependence minimization between the two data. Therefore, no census address frame is used for the coverage survey. Instead, the survey's sampling frame is based on the postcode directory. The survey has a stratified multistage cluster design. It is stratified by local authority by the hard-to-count index. The hard-to-count index is a variable specifically created for the coverage survey; it had five levels in the 2021 and 2011 Censuses and three levels in the 2001 Census (Brown, 2000; Brown et al., 2011; Dini, 2018). The hard-to-count index is derived at the lower super output area level, a unit of geography for statistical reporting with the size ranging from 400 to 1200 households (ONS, 2021d), and indicates relative ease or difficulty of obtaining a census return in that area. An area that has the hard-to-count '1' is the easiest to enumerate in a census and an area that belongs to the hard-to-count '5' is the hardest to enumerate. The sampling probabilities are unequal, reflecting the fact that the sample is allocated disproportionally to the harder to count areas relatively to the sample allocated to those easier to count areas. The primary sampling units are output areas, another geography for reporting census outputs with size ranging from 40 to 250 households (ONS, 2021d), and the secondary sampling units are postcodes. At least one output area is sampled from a local authority by hard-to-count combination; therefore, the sample is spread across the entire target population. Before the interviews begin, a field interviewer creates an address listing for the sampled postcodes by physically inspecting the area. This address listing is thus independent of the census address frame. After that, the attempt to collect information about all members (usually, interviewing only a single person within a household) of each of the households in the address list is made. The coverage survey questionnaire collects many attributes similar to those collected in the census and in a similar format. The address is always checked with the householder and names, surnames, date of birth and other attributes are collected in a structured way for all members of the household.

There are many stages of processing before the collected data reach record linkage and estimation (ONS, 2021a, chap. 4). Among those relevant for our discussion is the process of resolving multiple responses. This process aims to deduplicate cases within the same household / address and facilitates later linkage and estimation. Another important process is the imputation (ONS, 2018a). This process imputes the missing values of variables for the collected population units in the coverage survey and census data. Imputation allows the use of all the valid observations in estimation even if the originally collected values of certain variables are missing. Note that record linkage is performed on the non-imputed data. Outputs of record linkage and imputation are combined together for the estimation.

Record linkage itself is a combination of deterministic and probabilistic methods similar to those described in Section 2.2.3 with a substantial amount of clerical review (ONS, 2018b,c). While the combination of linkage methods and actual implementation may vary from census to census, what is important for us is that estimation-wise there is a need for two linkage exercises (Račinskij & Hammond,

2018). The first one is done for undercoverage estimation, while the second is done for overcoverage estimation. The correct location of an element is defined first and it is both assumed and ensured during the data collection that the coverage survey captures elements in the correct location. In principle, the correct location is a relative notion and depends on the goals of estimation and various practical considerations. In the census coverage estimation, being captured in the correct location usually means being captured within the postcode of the true location or in a neighbouring postcode (so that even if the reported address is not exactly correct, but a link is establishing within the postcode or neighbouring postcode, the linked case counts as being captured in the correct location). Therefore, in the undercoverage estimation, record linkage aims to determine whether a given element captured in the survey in a particular postcode was captured in the census in the same postcode or in neighbouring postcodes. On the contrary, in the overcoverage estimation, for a given element captured in the survey in the given postcode, record linkage aims to determine whether the element was captured in the census outside the given or neighbouring postcodes.

Coverage estimation itself is done in several steps. Undercoverage is estimated first, then overcoverage is estimated, and these estimates are combined to provide the net coverage adjusted estimate. Dual system estimation, our topic of interest, deals with the undercoverage and we are not discussing the overcoverage estimation in this thesis. There are two broad strategies for applying the dual system estimation in order to mitigate the bias due to heterogeneity. The first strategy is to use logistic or mixed effects logistic regression (Alho, 1990; US Census Bureau, 2008; Račinskij, 2018, 2020). This approach usually pools the entire sample data together and fits one undercoverage model. This undercoverage model reflects how the coverage probability varies by demographic and geographic variables as well as their interactions, which allows us to tackle heterogeneity. As we are dealing with the simple dual system estimator in this thesis, we do not delve into the logistic regression-based approaches here. The second approach is post-stratification, which involves obtaining the individual dual system estimates for every post-stratum in turn (Brown, 2000; Brown et al., 2019). In the 2011 Census of England and Wales such post-strata were defined to be a variable of interest, with the primary variable of interest being quinary age-sex groups, by aggregation of postcodes within sampled output area. Therefore, such post-stratification implicitly reflected stratification by local authority by hard-to-count index, due to the sampling design. The resulting post-strata sizes can be quite small in this case, and the dual system estimator with the Chapman correction is often used to avoid small sample bias (Seber, 1982):

$$\widehat{\tau}_{cc} = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1.$$

Note that the dual system estimation post-stratum is constructed by aggregation of postcodes which also happen to be linkage blocks. Such individual dual system estimates were then summed across sample clusters and used with the ratio estimator to produce the estimation area (combination of local authorities) by hard-to-count by the variable of interest domain totals at the population level (Brown et al., 2019). The local authority by hard-to-count by the variable of interest estimates were produced using the simple synthetic estimator. Such individual estimates are then summed together to produce the higher level totals. Another post-stratification approach was used in the United States Census prior to 2010 (Hogan, 1992, 1993). The post-stratification was by region by age group by ethnicity by

tenure and possibly by other variables if the size of the corresponding post-stratum was large enough. Again, individual dual system estimates were produced for these post-strata. One way or another, the important fact for us is that often in the large estimation exercises multiple dual system estimates are produced, rather than pooling the data and then estimating.

Estimation is followed by a dependence bias adjustment (Brown et al., 2006; Račinskij, 2022), but for our discussion it is sufficient to know that there is often this additional processing step. It is based on either demographic analysis data or some additional information collected during the census field follow up.

Finally, variance estimation is carried out. Unlike the variance estimator (6) for the simple case of dual system estimation, variance estimation for the coverage error corrected census population totals needs to take into account the design variance associated with the coverage survey. Since the design of the coverage survey is complex and multiple estimators are involved, Taylor series approximation becomes impractical and resampling methods, such as the jackknife or bootstrap, are used. For instance, in the case of the post-stratified dual system and ratio estimation the bootstrap variance estimation is as follows. The output areas with the corresponding parent sample dual system estimates are resampled according to the sampling design of the parent sample. Then the rest of the estimation process is carried out for every bootstrap resample. It is important to see that there is no recalculation of the dual system estimates, just the sampling of the output areas with their original estimates. Note that unless the coverage probabilities are small, the component of variability due to dual system estimation in each dual system post-stratum is small relative to the overall sampling variability.

In summary, the following features of the census coverage estimation are noteworthy for the discussion of the classification free record linkage and related dual system estimation. Existence of the address frame and / or address listings; a careful and well designed collection of attributes used in linkage and estimation with a special attention to addresses; blocking by low level geography; applying the dual system estimator at the level that aggregates several linkage blocks; sequential application of the dual system estimator in each post-stratum; carrying out undercoverage and overcoverage estimation separately; existence of bias adjustment procedures.

2.4 Simulated annealing

Simulated annealing is an optimization method originally proposed by Kirkpatrick et al. (1983). In this brief discussion of the algorithm we follow closely Liu (2004).

The task is to find the minimum of a target function $h(\boldsymbol{x})$ which is the same as finding a maximum of $\exp(-h(\boldsymbol{x})/T)$ at any value ot the artificial temperature parameter T. We then consider a sequence of monotone decreasing temperatures $T_1 > T_2 > \ldots T_t > \ldots$ with T_1 being sufficiently large and $T_t \to 0$ as $t \to \infty$. At each T_t, N_t iterations of Metropolis-Hastings sampling are run with $\Pi_t(\boldsymbol{x}) \propto \exp(-h(\boldsymbol{x})/T_t)$ as the equilibrium distribution. The simulated annealing algorithm uses the fact that in any system satisfying $\int \exp(-h(\boldsymbol{x})/T_t)d\boldsymbol{x} < \infty, T > 0$ the distribution Π_t puts more and more of its probability mass close to the global maximum of h as t keeps increasing. Hence, when T_t is close to 0, the sample drawn from Π_t is in vicinity of the global minimum of the target function.

Therefore, at least in principle the global minimum of $h(\boldsymbol{x})$ can be reached if the number N_t is

sufficiently large and the temperature T_t decreases sufficiently slowly.

The simulated annealing algorithm consists of those steps

- Initialize an arbitrary solution $\boldsymbol{x}^{(0)}$ and temperature T_1;
- For each t run N_t iterations of a Markov chain Monte Carlo scheme with $\Pi_t(\boldsymbol{x})$ being its target distribution. The final solution at a step t is used at the initial solution for a step $t+1$;
- Increment t by 1.

In this work we use the Metropolis algorithm as a Markov chain Monte Carlo scheme which proceeds in the following way; see Liu (2004) for more details:

1. Start with a random configuration $\boldsymbol{x}^{(0)}$;
2. At a state $\boldsymbol{x}^{(t)}$ make a random 'unbiased' perturbation $\boldsymbol{x}'$ and compute the change $\Delta h = h(\boldsymbol{x}') - h(\boldsymbol{x}^{(t)})$;
3. Generate $Y \sim \text{Uniform}\,[0,1]$ and make a decision regarding the next configuration $\boldsymbol{x}^{(t+1)}$:

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{x}' \text{ if } Y \leq \frac{\Pi(\boldsymbol{x}')}{\Pi(\boldsymbol{x}^{(t+1)})} = \exp(-\Delta h) \\ \boldsymbol{x}^{(t)} \text{ otherwise.} \end{cases}$$

2.5 Identifiability

Identifiability is a property of a statistical model allowing non-ambiguous recovery of model parameters from the observed data generated by the model (Allman et al., 2009, Sullivant, 2018, chap. 16.1). A statistical model $p : \boldsymbol{\Theta} \to P_\Theta, \boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}$ is globally or strictly identifiable if any two parameters $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ in $\boldsymbol{\Theta}$ produce different probability distributions $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}'}$. In other words, the mapping is one-to-one. However, in certain cases, finitely many-to-one mappings can also be regarded as globally identifiable. Moreover, the global identifiability is not the only type of identifiability. We will discuss different cases of identifiability later in this chapter.

Identifiability is crucial in statistical applications. If a parameter of interest is used for inference, it is desirable that a chosen statistical model produces a unique estimate of the parameter given the data. If a model is not identifiable, then information about the parameter of interest is not recoverable and the model is not practically useful. Therefore, establishing identifiability or non-identifiability of a model, despite often being a difficult exercise, is generally worth the effort.

We focus only on identifiability of record linkage models that can be represented or parameterized as finite mixtures of discrete probability distributions:

$$\text{pr}(x_j; \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i \text{pr}(x_j; \boldsymbol{\theta}_i), \tag{17}$$

where $\pi_i = \text{pr}(\boldsymbol{\theta} = \boldsymbol{\theta}_i)$ are mixing proportions, and $\text{pr}(x_j; \boldsymbol{\theta}_i)$ are probability mass functions that usually belong to the same family of distributions (but in principle may belong to different families).

We already encountered two-component mixtures in Section 2.2.3. For a detailed discussion of finite mixture models see, for instance, McLachlan & Peel (2000). Note that we are deliberately saying that a model of interest can be *represented* or *parameterized* rather than that the model follows some finite mixture of discrete distributions. This is because in record linkage problems, the sampling is not carried out from a mixture distribution in the usual sense; see the presentation of the linkage experiment in Section 2.2.5. Nevertheless, generated data can be meaningfully summarized in the form of (17) and provide a means of estimating the first moments of the parameters of interest, but not higher moments. For more details see Section 3.2 of this thesis.

For the regular mixtures, classic presentations of identifiability are Teicher (1961, 1963, 1967) and Yakowitz & Spragins (1968). A good discussion of the problem can be found in Titterington et al. (1985, chap. 3.1). Also, Rao (1992, chap. 8) is dedicated to identifiability of finite mixtures which, among other methods, reviews and consolidates methods presented in the above papers.

In this thesis, we are relying on the methods from algebraic statistics (Drton et al., 2009; Sullivant, 2018). Algebraic statistics provides methods from computational algebra and algebraic geometry that can be used to establish identifiability / non-identifiability of statistical models of interest. There are many other areas where algebraic statistics can be applied or used to study statistical properties, such as the maximum likelihood estimation (Hosten et al., 2005; Allman et al., 2019) or model selection (Krampe & Kuhnt, 2010). In this thesis, we only focus on use of algebraic statistics to study identifiability of the models of interest. Algebraic methods for identifiability are viable since the problem of identifiability of a statistical model can be regarded as equivalent to injectivity of a polynomial or rational map (Sullivant, 2018, chap. 16). The algebraic approach to identifiability offers a variety of conceptual and computational methods to tackle the problem as well as a rich selection of tools to study related properties. These are more flexible than those methods presented in Rao (1992, chap. 8) since these classical results are mainly concerned with the families of distributions. In our case, we quite often deal with models within the same family, but some of the models are identifiable, while others are not. Being able to tackle this case by case is important. Here we will consider the following algebraic approaches: checking the dimension of the image of a rational map, tensor methods and Gröbner basis based methods. Note that these algebraic methods allow identifiability for a general model specification to be established, without any restrictions on how the actual sampling works. Therefore, they are as applicable for regular mixtures as for the mixture-like models of our interest.

2.5.1 Types of identifiability

While identifiability of a model in general means that the parameters of the model can be recovered from the data in a non-ambiguous way, there exist several types of identifiability. It is not always made clear in the literature which type of identifiability is being established. We start our discussion by providing the classification of identifiability types. Below we largely follow the presentation by Sullivant (2018, chap. 16.1).

The theory as presented here and below assumes that we are dealing with an extension of the regular exponential family called the algebraic exponential family (Sullivant, 2018, chap. 6.5). We have a rational map $\phi : \boldsymbol{\Theta} \to \boldsymbol{\eta}$. A rational map is a map $\phi : \mathbb{C}^m \to \mathbb{C}^n$ with $\phi_i = f_i/g_i$, where f_i, g_i

are polynomials and $g_i \neq 0$. Here $\boldsymbol{\Theta}$ is a parameter space and $\boldsymbol{\eta}$ is the natural parameter space of an exponential family or its transformation. Recall, that the natural parameter space is the set of all values $\boldsymbol{\eta}$ where the function $f_X(x; \boldsymbol{\theta})$ is finite. Also, it is assumed that $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ is a semialgebraic set. Informally, a semialgebraic set is a subset of $\mathbb{R}^d$ given by polynomial equations and inequalities. For more formal definition of a semialgebraic set and its relation to statistical models; see Zwiernik (2016, chap. 2.2.2) and Sullivant (2018, chap. 6.4).

The following is based on Godfrey & DiStefano (1987) and Definition 16.1.1 of Sullivant (2018, chap. 16.1): let $\phi : \boldsymbol{\Theta} \to \boldsymbol{\eta}$ with model $\mathcal{Q} = \text{im } \phi$. Here, im ϕ means the image of map ϕ, that is the set of all possible outputs of the map. The model $\mathcal{Q}$ is

- globally identifiable if ϕ is a one-to-one map on $\boldsymbol{\Theta}$;
- generically identifiable (also known as almost everywhere identifiable) if $\phi^{-1}(\phi(\boldsymbol{\theta})) = \boldsymbol{\theta}$ for almost all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (that is, the probability of drawing data points that lead to non-identifiable parameter equals zero in this case);
- rationally identifiable if there is a dense open subset of $\boldsymbol{Y} \subseteq \boldsymbol{\Theta}$ (informally, it means that every member of $\boldsymbol{Y}$ is either in $\boldsymbol{\Theta}$ or arbitrarily close to a member of $\boldsymbol{\Theta}$) and a rational function $\psi : \boldsymbol{\eta} \to \boldsymbol{\Theta}$ such that $\psi \circ \phi(\boldsymbol{\theta}) = \boldsymbol{\theta}$ on $\boldsymbol{Y}$;
- locally identifiable if there exists an open neighbourhood $\boldsymbol{Y}_\theta$ around a generic point $\boldsymbol{\theta}$ such that ϕ is identifiable;
- nonidentifiable if for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the set of values $\phi^{-1}(\phi(\boldsymbol{\theta}))$ is greater than 1;
- generically nonidentifiable if almost for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the set of values $\phi^{-1}(\phi(\boldsymbol{\theta}))$ is infinite.

In certain situations, a model may be generically nonidentifiable, but some individual parameters or functions of individual parameters may be identifiable. Types of identifiability for individual parameters are similar to the types of a model identifiability presented above.

Practically, one aims to establish whether a model or a parameter of interest is identifiable and what the corresponding type of identifiability is. First, an attempt to determine local identifiability is made. If the model is locally non-identifiable, no further computations are needed and the model is declared non-identifiable. If the model is locally identifiable, then generic or rational identifiability is checked dependent of what is feasible. As we will see, quite often it is computationally too difficult to work out if a model is generically or rationally identifiable, so there may be cases where no result is obtained. Informally, global identifiability means that there is a unique parameter $\boldsymbol{\theta}$ that results in output $\phi(\boldsymbol{\theta})$. Generic identifiability means that there are some points $\boldsymbol{\theta}$ for which the output is not unique, but those points have corresponding zero probabilities of being selected. For instance, given the mixing proportion is exactly 0.5, component parameters can be swapped with other component parameters without affecting the value of $\phi(\boldsymbol{\theta})$. However, the probability of the mixture proportion being exactly equal to 0.5 is 0. Rational identifiability means that every $\theta_i \in \boldsymbol{\theta}$ can be expressed as the rational function ψ of $\phi(\boldsymbol{\theta})$. Finally, local identifiability means that a model is identified in the neighbourhood of $\boldsymbol{\theta}$, but there may in general be several $\boldsymbol{\theta}$ that produce the same $\phi(\boldsymbol{\theta})$. For example, in mixture

models swapping the components will lead to the same model output, the phenomena known as label switching. Say, in the two component mixture, we have $\text{pr}(x;\boldsymbol{\theta}) = \pi\text{pr}(x;\boldsymbol{\theta}_1) + (1-\pi)\text{pr}(x;\boldsymbol{\theta}_2) = \pi'\text{pr}(x;\boldsymbol{\theta}'_1) + (1-\pi')\text{pr}(x;\boldsymbol{\theta}'_2)$ with $\pi' = (1-\pi), \boldsymbol{\theta}'_1 = \boldsymbol{\theta}_2, \boldsymbol{\theta}'_2 = \boldsymbol{\theta}_1$. Still the model is identifiable in the neighbourhood of π.

2.5.2 Number of observables and number of parameters

Before starting to assess identifiability, the basic check of whether the number of observables does not exceed the number of parameters that need to be estimated is needed. In the linkage model with binary comparisons and K linkage variables there are 2^K observed quantities that correspond to unique comparison patterns $\boldsymbol{\gamma}_p$. Hence, we cannot estimate more than 2^K parameters. The parameters are π, μ_k, ν_k and, when between-variables dependencies are present, $\mu_{l|k(1)}, \mu_{l|k(0)}, \nu_{l|k(1)}, \nu_{l|k(0)}$.

2.5.3 Assessing local identifiability

Local identifiability of a model or a single parameter is the easiest type of identifiability to check. The approach is based on checking the dimension of the image of a rational map ϕ. Sullivant (2018, chap. 16.1) states, that if the dimension of the image of ϕ equals the dimension of the parameter space $\boldsymbol{\Theta}$ then a generic point has a fixed finite number of preimages (where a preimage is the inverse of an image). A fixed finite number of preimages means that there are finitely many $\boldsymbol{\theta}$ that produce the same $\phi(\boldsymbol{\theta})$. For instance, in the example of label swapping mentioned in Section 2.5.1, for any $\phi(\boldsymbol{\theta})$, we have two preimages: $\{\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ and $\{\pi', \boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2\}$.

The following is Proposition 16.1.7 from Sullivant (2018, chap. 16.1): let $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ with $\dim \boldsymbol{\Theta} = d$ and suppose that ϕ is a rational map. Then $\dim \text{im } \phi$ is equal to the rank of the Jacobian matrix evaluated at a generic point:

$$J(\phi) = \begin{pmatrix} \frac{\partial \phi_1}{\partial \theta_1} & \cdots & \frac{\partial \phi_1}{\partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_r}{\partial \theta_1} & \cdots & \frac{\partial \phi_r}{\partial \theta_d} \end{pmatrix}.$$

In particular, the parameter vector $\boldsymbol{\theta}$ is locally identifiable if $\text{rank } J(\phi) = d$ and the parameter vector is generically nonidentifiable if $\text{rank } J(\phi) < d$. For the precise definition of a generic point, see Hartshorne (1977, chap. 2), but loosely speaking it is a point at which all generic properties of a set hold. In this particular case, there may be some points that lead to the components of the Jacobian matrix being undefined, say, due to division by 0, but for any 'non-extreme' point, the property that the rank of $J(\phi)$ is d holds.

Given a model that is parameterized in a similar way to the models presented in 2.2.3, one can use symbolic computation software like Mathematica (Wolfram Research, Inc., 2022) or Maple (Maple, 2021) to compute the Jacobian matrix and find its rank. There are several examples of checking local identifiability in Chapter 5.

2.5.4 Tensor methods to assess identifiability

Identifiability of mixtures of discrete probability distributions can often be checked using tensor methods. Tensors are multidimensional arrays describing multilinear relationships. They are generalizations of vectors and matrices. While a vector has only one index and a matrix has two indices, a tensor has an arbitrary number of indices. Our discussion of the tensor methods requires only superficial knowledge of tensors and operations with them. We will only be dealing with the tensor product, denoted $\otimes$. The tensor product is a generalization of the outer product. For instance, given two arrays $\boldsymbol{a} = (a_1, a_2)^T$ and $\boldsymbol{b} = (b_1, b_2)^T$, the corresponding tensor product is

$$\boldsymbol{a} \otimes \boldsymbol{b} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \otimes \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_1 b_1 & a_1 b_2 \\ a_2 b_1 & a_2 b_2 \end{pmatrix}.$$

In the case of the two above arrays and an additional one $\boldsymbol{c} = (c_1, c_2)^T$, the product of these three arrays is

$$\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \otimes \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \otimes \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} a_1 b_1 c_1 & a_1 b_2 c_1 \\ a_2 b_1 c_1 & a_2 b_2 c_1 \\ a_1 b_1 c_2 & a_1 b_2 c_2 \\ a_2 b_1 c_2 & a_2 b_2 c_2 \end{pmatrix}.$$

The tensor product is associative: $(\boldsymbol{a} \otimes \boldsymbol{b}) \otimes \boldsymbol{c} = \boldsymbol{a} \otimes (\boldsymbol{b} \otimes \boldsymbol{c})$.

Tensor methods for checking identifiability are based on the results first obtained in Kruskal (1976) and Kruskal (1977). This approach was refined and extended in Allman et al. (2009). As before, we follow Sullivant (2018) in our discussion.

Key to our discussion is that 3-way tensors can be related to a g-component mixture model with 3 discrete variables X_1, X_2, X_3 that are independent given a mixing component. Each variable has r_k levels. With g and r_k as defined above, let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_g \in \mathbb{K}^{r_1}$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_g \in \mathbb{K}^{r_2}$, and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_g \in \mathbb{K}^{r_3}$, where $\mathbb{K}$ is a field such as the field of the rational numbers $\mathbb{R}$, or the field of the complex numbers $\mathbb{C}$. Those vectors can be arranged into matrices $\boldsymbol{A} \in \mathbb{K}^{r_1 \times g}$, $\boldsymbol{B} \in \mathbb{K}^{r_2 \times g}$, and $\boldsymbol{C} \in \mathbb{K}^{r_3 \times g}$ with columns of these matrices being $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_g \in \mathbb{K}^{r_1}$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_g \in \mathbb{K}^{r_2}$, and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_g \in \mathbb{K}^{r_3}$. Note that in Allman et al. (2009) the above arrays are rows of the matrices and the algebra is different from the one used here.

The following is the Definition 16.3.1 from Sullivant (2018, chap. 16.3): let $\boldsymbol{A} \in \mathbb{K}^{r \times g}$ be a matrix. The *Kruskal rank* of $\boldsymbol{A}$, denoted $\text{rank}_K(\boldsymbol{A})$, is the largest number l_c such that any l_c columns of $\boldsymbol{A}$ are linearly independent.

A tensor of the form $\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$ is called a *rank one tensor* or *decomposable tensor*. A tensor $\boldsymbol{M} \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}$ has rank g if it can be written as

$$\boldsymbol{M} = \sum_{i=1}^{g} \boldsymbol{a}_i \otimes \boldsymbol{b}_i \otimes \boldsymbol{c}_i \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3} \tag{18}$$

for some $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_g \in \mathbb{K}^{r_1}$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_g \in \mathbb{K}^{r_2}$, and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_g \in \mathbb{K}^{r_3}$ but cannot be written as a sum of fewer rank one tensors.

An important result regarding the uniqueness of the decomposition of a tensor into rank one tensors is Theorem 16.3.2 in Sullivant (2018, chap. 16.3), known as Kruskal's theorem: let $\boldsymbol{M} \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}$ be a tensor, and let

$$\boldsymbol{M} = \sum_{i=1}^{g} \boldsymbol{a}_i \otimes \boldsymbol{b}_i \otimes \boldsymbol{c}_i \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3},$$

$$\boldsymbol{M} = \sum_{i=1}^{g} \boldsymbol{a}'_i \otimes \boldsymbol{b}'_i \otimes \boldsymbol{c}'_i \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}$$

be two decompositions of $\boldsymbol{M}$ into rank one tensors. Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ be the matrices whose columns are $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_g$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_g$, and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_g$, respectively. Matrices $\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}'$ are defined in a similar way. If $\text{rank}_K(\boldsymbol{A}) + \text{rank}_K(\boldsymbol{B}) + \text{rank}_K(\boldsymbol{C}) \geq 2g + 2$, then there exists a permutation $\sigma \in S_g$ and non-zero $\lambda_1, \ldots, \lambda_g, \gamma_1, \ldots, \gamma_g \in \mathbb{K}$ such that

$$\boldsymbol{a}_{\sigma(i)} = \lambda_i \boldsymbol{a}'_i, \boldsymbol{b}_{\sigma(i)} = \gamma_i \boldsymbol{b}'_i \text{ and } \boldsymbol{c}_{\sigma(i)} = \lambda_i^{-1}\gamma_i^{-1}\boldsymbol{c}'_i \text{ for all } i \in [g] = \{1, \ldots, g\}.$$

In particular, $\boldsymbol{M}$ has rank g.

A toy example related to the Kruskal theorem will be presented after we introduce the triple product notation which simplifies the presentation.

The triple product notation allows us to write $\boldsymbol{M}$ compactly as $\boldsymbol{M} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]$ to denote the rank one decomposition of (18). The Kruskal theorem can also be more compactly stated that if $\text{rank}_K(\boldsymbol{A}) + \text{rank}_K(\boldsymbol{B}) + \text{rank}_K(\boldsymbol{C}) \geq 2g+2$ and $\boldsymbol{M} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}] = [\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}']$, then there is a permutation matrix $\boldsymbol{P}$ and invertible diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3$ with $\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{D}_3 = \boldsymbol{I}$ such that

$$\boldsymbol{D}_1\boldsymbol{P}\boldsymbol{A} = \boldsymbol{A}', \boldsymbol{D}_2\boldsymbol{P}\boldsymbol{B} = \boldsymbol{B}', \boldsymbol{D}_3\boldsymbol{P}\boldsymbol{C} = \boldsymbol{C}'.$$

The Kruskal theorem gives us conditions under which a decomposition of tensors is unique as much as it is possible. This general result can be applied to the mixtures of three discrete variables that are independent given a mixing component (the conditional independence model). Such a model is denoted here as $\text{Mixt}^g(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$. The following is Corollary 16.3.3 in Sullivant (2018, chap. 16.3): consider the mixture model $\text{Mixt}^g(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$, where X_k has r_k levels for $k = 1, 2, 3$. This model is generically identifiable up to relabelling of mixture components (up to label switching) if

$$\min(r_1, g) + \min(r_2, g) + \min(r_3, g) \geq 2g + 2.$$

It is worth replicating the proof of this corollary given in Sullivant (2018, chap. 16.3), since the main argument of this proof is used when applying the Kruskal theorem in practice. We use the familiar parameterization of $\text{Mixt}^g(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$ with some slight changes in notation to facilitate the discussion:

$$\text{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \sum_{i=1}^{g} \pi_i \alpha_i(x_1)\beta_i(x_2)\gamma_i(x_3),$$

where $\alpha_i(x_1), \beta_i(x_2)$, and $\gamma_i(x_3)$ represent conditional probabilities of $X_1 = x_1, X_2 = x_2, X_3 = x_3$ given a mixture component i, respectively. The parameterization is similar to the one of rank g tensors.

However, there are additional restrictions on the matrices $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ as their columns must add to one. Also, we have an additional parameter π. We let Π denote the diagonal matrix of mixing proportions $\text{diag}(\pi_1, \ldots, \pi_g)$. Let $\boldsymbol{A} = \boldsymbol{\alpha\Pi}, \boldsymbol{B} = \boldsymbol{\beta}, \boldsymbol{C} = \boldsymbol{\gamma}$, so that the above distribution can be written as the triple product $\boldsymbol{M} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]$. For generic choices of the conditional distributions and generic choices of r_k, g we have

$$\text{rank}_K(\boldsymbol{A}) + \text{rank}_K(\boldsymbol{B}) + \text{rank}_K(\boldsymbol{C}) = \min(r_1, g) + \min(r_2, g) + \min(r_3, g)$$

and we can apply Kruskal's theorem. Suppose that there is another decomposition of the above distribution, that is

$$\left[\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}'\right] = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\,.$$

By Kruskal's theorem, there must exist a permutation matrix $\boldsymbol{P}$ and invertible diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3$ with $\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{D}_3 = \boldsymbol{I}$ such that

$$\boldsymbol{D}_1\boldsymbol{P}\boldsymbol{A} = \boldsymbol{A}', \boldsymbol{D}_2\boldsymbol{P}\boldsymbol{B} = \boldsymbol{B}', \boldsymbol{D}_3\boldsymbol{P}\boldsymbol{C} = \boldsymbol{C}'.$$

However, since we are dealing with the probability distributions, the columns of $\boldsymbol{B}'$ and $\boldsymbol{C}'$ sum to one. This forces $\boldsymbol{D}_2 = \boldsymbol{I}$ and $\boldsymbol{D}_3 = \boldsymbol{I}$, which in turn forces $\boldsymbol{D}_1 = \boldsymbol{I}$. This means that the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}'$ only differ by simultaneous permutations of the columns. So up to those permutations, $\boldsymbol{A} = \boldsymbol{A}', \boldsymbol{B} = \boldsymbol{B}', \boldsymbol{C} = \boldsymbol{C}'$. This means that both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be recovered up to permutation of the labels of mixing components. Finally, π and $\boldsymbol{\alpha}$ can be recovered from $\boldsymbol{A}$ noting that π is the vector of column sums of $\boldsymbol{A}$, and then $\boldsymbol{\alpha} = \boldsymbol{A\Pi}^{-1}$. This finishes the proof.

The proof of the above corollary gives us a good opportunity to give an example of how the Kruskal theorem works. Consider a two component mixture with binary variables, so that we are dealing with the following:

$$\boldsymbol{\alpha} = \begin{pmatrix} \mu_1 & \nu_1 \\ 1-\mu_1 & 1-\nu_1 \end{pmatrix}, \alpha_1 = \begin{pmatrix} \mu_1 \\ 1-\mu_1 \end{pmatrix}, \alpha_2 = \begin{pmatrix} \nu_1 \\ 1-\nu_1 \end{pmatrix},$$

$$\boldsymbol{B} = \boldsymbol{\beta} = \begin{pmatrix} \mu_2 & \nu_2 \\ 1-\mu_2 & 1-\nu_2 \end{pmatrix}, \beta_1 = \begin{pmatrix} \mu_2 \\ 1-\mu_2 \end{pmatrix}, \beta_2 = \begin{pmatrix} \nu_2 \\ 1-\nu_2 \end{pmatrix},$$

$$\boldsymbol{C} = \boldsymbol{\gamma} = \begin{pmatrix} \mu_3 & \nu_3 \\ 1-\mu_3 & 1-\nu_3 \end{pmatrix}, \gamma_1 = \begin{pmatrix} \mu_3 \\ 1-\mu_3 \end{pmatrix}, \gamma_2 = \begin{pmatrix} \nu_3 \\ 1-\nu_3 \end{pmatrix},$$

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi & 0 \\ 0 & 1-\pi \end{pmatrix}.$$

Kruskal's theorem above is formulated for 3-way tensors, but the above problem is of higher order, so in order to apply it we define

$$\boldsymbol{A} = \boldsymbol{\alpha}\Pi = \begin{pmatrix} \mu_1 & \nu_1 \\ 1-\mu_1 & 1-\nu_1 \end{pmatrix} \begin{pmatrix} \pi & 0 \\ 0 & 1-\pi \end{pmatrix} = \begin{pmatrix} \pi\mu_1 & (1-\pi)\nu_1 \\ \pi(1-\mu_1) & (1-\pi)(1-\nu_1) \end{pmatrix}$$

and now we are dealing with a 3-way tensor and can apply the triple product notation.

First, we check that $\text{rank}_K(\boldsymbol{A}) = 2$, because 2 is the largest number of columns that can be linearly independent. Kruskal ranks of $\boldsymbol{B}$ and $\boldsymbol{C}$ are also 2. Hence, $\text{rank}_K(\boldsymbol{A}) + \text{rank}_K(\boldsymbol{B}) + \text{rank}_K(\boldsymbol{C}) = 6 \geq 2g + 2 = 6$ and Kruskal's theorem tells us that all rank one decompositions of $\boldsymbol{M} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]$ differ only up to permutation of the elements in the columns of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. It is easy to find invertible diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3$ and a permutation matrix $\boldsymbol{P}$ in this case. Let

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and

$$\boldsymbol{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This gives

$$\boldsymbol{D_1 P A} = \boldsymbol{A}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \pi\mu_1 & (1-\pi)\nu_1 \\ \pi(1-\mu_1) & (1-\pi)(1-\nu_1) \end{pmatrix} = \begin{pmatrix} \pi(1-\mu_1) & (1-\pi)(1-\nu_1) \\ \pi\mu_1 & (1-\pi)\nu_1 \end{pmatrix}.$$

Also, letting $\boldsymbol{D}_2 = \boldsymbol{D}_3 = \boldsymbol{D}_1$, we have

$$\boldsymbol{D_2 P B} = \boldsymbol{B}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_2 & \nu_2 \\ 1-\mu_2 & 1-\nu_2 \end{pmatrix} = \begin{pmatrix} 1-\mu_2 & 1-\nu_2 \\ \mu_2 & \nu_2 \end{pmatrix},$$

$$\boldsymbol{D_3 P C} = \boldsymbol{C}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_3 & \nu_3 \\ 1-\mu_3 & 1-\nu_3 \end{pmatrix} = \begin{pmatrix} 1-\mu_3 & 1-\nu_3 \\ \mu_3 & \nu_3 \end{pmatrix}.$$

Indeed, these matrices give the same rank one tensors, but differ only by permutation of the elements in columns and hence can be identified up to these permutations. The parameters $\mu_2, \mu_3, \nu_2, \nu_3$ are recovered directly alongside the functions $s_1(\boldsymbol{\theta}) = \pi\mu_1, s_2(\boldsymbol{\theta}) = \pi(1-\mu_1), s_3(\boldsymbol{\theta}) = (1-\pi)\nu_1, s_4(\boldsymbol{\theta}) = (1-\pi)(1-\nu_1)$. Then π can be recovered using $s_1(\boldsymbol{\theta}) + s_2(\boldsymbol{\theta}) = \pi\mu_1 + \pi(1-\mu_1) = \pi$ and μ_1 recovered using $\mu_1 = s_1(\boldsymbol{\theta})/\pi$ and finally ν_1 recovered using $\nu_1 = s_3(\boldsymbol{\theta})/(1-\pi)$.

Kruskal's theorem can be generalized to higher order tensors. However, the theorem for 3-way tensors can be used to determine identifiability for models with an arbitrary finite number of variables. This result is given as Corollary 16.3.4. in Sullivant (2018, chap. 16.3): let $X_1, \ldots, X_K$ be discrete random variables, where X_k has r_k levels. Let $(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_K)$ denote the complete independence model. Let $\boldsymbol{A} \mid \boldsymbol{B} \mid \boldsymbol{C}$ be a tripartition of $[K] = \{1, \ldots, K\}$ with no empty parts. Then the mixture model $\text{Mixt}^g(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_K)$ is generically identifiable up to label swapping if

$$\min(\prod_{a \in \boldsymbol{A}} r_a, g) + \min(\prod_{b \in \boldsymbol{B}} r_b, g) + \min(\prod_{c \in \boldsymbol{C}} r_c, g) \geq 2g + 2.$$

We do not replicate the proof here. It is based on 'clumping' random variables into three blocks and flattening the array into a 3-way array.

A related important and apparently easier to use result regarding the models of conditional inde-

pendence given the component membership with arbitrary number of variables K, $\text{Mixt}^g(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_K)$, is given and proved in Allman et al. (2009). The model $\text{Mixt}^g(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_K)$ is generically identifiable up to label switching if

$$K \geq 2\lceil \log_2 g \rceil + 1, \tag{19}$$

where $\lceil x \rceil$ is the smallest integer at least as large as x.

Whenever dealing with models having some association between the variables, tensor methods need to be applied for each case individually. However, one always tries to 'clump' variables to arrive to $\text{Mixt}^g(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$ as the first step. Some examples can be found in Allman et al. (2015) and some of our later results will be based on their work.

2.5.5 Gröbner basis based methods

Another class of methods for checking identifiability is based on Gröbner bases. These methods utilise computational algebra and algebraic geometry. There are several advantages of those methods. First, identifiability of a statistical model is largely checked in an algorithmic way. Second, these methods allow checking identifiability of an entire model or just a single parameter of interest. Third, rational, generic and global identifiability can be checked. Finally, it is very flexible and any model that is parameterized as a polynomial map can be assessed for identifiability. Among disadvantages we can mention computational complexity associated with computing the Gröbner basis as well as the relative unfamiliarity of many statisticians with this concept. To address this unfamiliarity, we included a small chapter in Appendix A with a short primer on basic concepts from computational commutative algebra and algebraic geometry needed to understand the notion of a Gröbner basis. A detailed discussion of the ideas can be found in Becker & Weispfenning (1993); Cox et al. (2004, 2015). Usage of these key concepts in algebraic statistics is covered in Sullivant (2018). Most of the definitions in the appendix are from Cox et al. (2015) and Sullivant (2018).

2.5.6 Assessing global and generic identifiability with Gröbner basis based methods

We assume that at this point we are equipped for presentation of the Gröbner basis based methods of establishing identifiability. This method was introduced by García-Puente et al. (2010) and is also discussed in Sullivant (2018) and corresponding proofs of the results can be found in the references.

Below is the Proposition 16.1.8 from Sullivant (2018, chap. 16.1). For a given rational map ϕ and parameter s, let $\tilde{\phi}$ be augmented rational map $\tilde{\phi} : \boldsymbol{\Theta} \to \mathbb{R}^{n+1}, \boldsymbol{\theta} \mapsto (s(\boldsymbol{\theta}), \phi(\boldsymbol{\theta}))$. We denote the coordinates of $\mathbb{R}$ by $q, x_1, \ldots, x_n$. Suppose that $g(q, x_1, \ldots, x_n) \in I(\tilde{\phi}(\boldsymbol{\Theta})) \subseteq \mathbb{R}[q, x_1, \ldots, x_n]$ is a polynomial such that q appears in this polynomial, $g(q, x_1, \ldots, x_n) = \sum_{i=0}^{d} g_i(x_1, \ldots, x_n) q^i$, and $g_0(x_1, \ldots, x_n)$ does not belong to $I(\phi(\boldsymbol{\Theta}))$.

1. If g is linear in q, $g = g_1(x_1, \ldots, x_n)q - g_0(x_1, \ldots, x_n)$, then s is generically identifiable by the rational formula $s = \frac{g_0(x_1,\ldots,x_n)}{g_1(x_1,\ldots,x_n)}$. If, in addition, $g_1(x_1, \ldots, x_n) \neq 0$ for $x_1, \ldots, x_m \in \phi(\boldsymbol{\Theta})$, then s is globally identifiable.

2. If g has higher degree d in q, then s may or may not be generically identifiable. Generically, there are at most d possible choices for the parameter $s(\boldsymbol{\theta})$ given $\phi(\boldsymbol{\theta})$.

3. If no such polynomial g exists, then the parameter s is not generically identifiable.

It is possible to provide an intuitive and informal explanation of why the above proposition works. We are dealing with vanishing ideals $I(\phi(\boldsymbol{\Theta}))$ and $I(\tilde{\phi}(\boldsymbol{\Theta}))$. A vanishing ideal of a set of points is the set of polynomials that vanish (equals to zero) on this set of points; see Appendix A for more details. If a polynomial $g(q, x_1, \ldots, x_n) \in I(\tilde{\phi}(\boldsymbol{\Theta}))$ then it means that for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ this polynomial must evaluate to zero. In addition, if $g(q, x_1, \ldots, x_n) \in I(\tilde{\phi}(\boldsymbol{\Theta}))$ is linear, then $g(q, x_1, \ldots, x_n)$ may be written as $g = g_1(x_1, \ldots, x_n)q - g_0(x_1, \ldots, x_n) = g_1(x_1, \ldots, x_n)s(\boldsymbol{\theta}) - g_0(x_1, \ldots, x_n)$. Note that $q = s(\boldsymbol{\theta})$ by construction of the augmented map $\tilde{\phi} : \boldsymbol{\Theta} \to \mathbb{R}^{n+1}$. Since $g \in I(\tilde{\phi}(\boldsymbol{\Theta}))$, we have $g_1(x_1, \ldots, x_n)s(\boldsymbol{\theta}) - g_0(x_1, \ldots, x_n) = 0$, and the fact that $g_1 \notin I(\phi(\boldsymbol{\Theta}))$ allows this linear equation to be solved for $s(\boldsymbol{\theta})$.

If $g(q, x_1, \ldots, x_n)$ is of some higher degree d, one solves the corresponding equation of this degree to find $s(\boldsymbol{\theta})$. If multiple solutions exist this may mean that $s(\boldsymbol{\theta})$ is at least locally identifiable. Note, that unlike the method of checking for local identifiability based on the rank of the Jacobian matrix presented in Section 2.5.3, the current approach not only tells us whether $s(\boldsymbol{\theta})$ is identifiable or not, but also provides additional information in the form of the polynomial g. Thus, if we know that $s(\boldsymbol{\theta})$ has certain constraints, for instance is a real number in the interval $(0, 1)$ and there is only one solution that satisfies this condition, then $s(\boldsymbol{\theta})$ is globally identifiable.

The above proposition does not say how to find such a polynomial $g(q, x_1, \ldots, x_n)$ or show that it does not exist. However, it is possible to solve this problem using Gröbner bases. Below is the Proposition 16.1.9 from Sullivant (2018, chap. 16.1).

Let G be a reduced Gröbner basis for $I(\tilde{\phi}(\boldsymbol{\Theta})) \in \mathbb{R}[q, x_1, \ldots, x_n]$ with respect to an elimination ordering such that $q \succ x_i$ for all i. Suppose that there is a polynomial $g(q, x_1, \ldots, x_n) = \sum_{i=0}^{d} g_i(x_1, \ldots, x_n)q^i \in I(\tilde{\phi}(\boldsymbol{\Theta}))$ with $g_0(x_1, \ldots, x_n) \notin I(\phi(\boldsymbol{\Theta}))$ that has non-zero degree in q. Then a polynomial of lowest non-zero degree in q of this form appears in G. If no such polynomial exist in $I(\tilde{\phi}(\boldsymbol{\Theta}))$, then G does not contain any polynomial involving the indeterminate q.

In other words, if we have a model parameterized as a rational map ϕ and we are interested in checking if a parameter or a function of certain parameters of this map is identifiable, we augment the map ϕ with an additional parameter q. In this case, q equals to the parameter (or function of parameters) which identifiability we are interested in. Then we find a reduced Gröbner basis of the augmented map with respect to any elimination order and check if it contains a polynomial g in q as described above.

2.5.7 Assessing rational identifiability with Gröbner basis based methods

In practice, assessing global and generic identifiability using Gröbner bases may yield no results in real time due to computational complexity issues. However, Gröbner bases can be used to check rational identifiability. In many cases we have considered so far, computations can be carried out in real time.

The following is the Proposition 16.4.8 from Sullivant (2018, chap. 16.4). Let $c : \mathbb{R}^d \to \mathbb{R}^n$ be a polynomial map, and let I_c be the ideal

$$I_c = \langle c_1(\boldsymbol{t}) - c_1(\boldsymbol{\theta}), \ldots, c_n(\boldsymbol{t}) - c_n(\boldsymbol{\theta}) \rangle \subseteq \mathbb{R}(\boldsymbol{t})[\boldsymbol{\theta}].$$

Let $f \in \mathbb{R}[\boldsymbol{\theta}]$. Then f is rationally identifiable if and only if $f(\boldsymbol{t}) - f(\boldsymbol{\theta}) \in I_c$.

Here, $\mathbb{R}(\boldsymbol{t})[\boldsymbol{\theta}]$ means that we are dealing with polynomials in indeterminates $\boldsymbol{\theta}$ with symbolic parameters in the coefficients $\boldsymbol{t} \in \mathbb{R}^d$. For example, consider a quadratic polynomial $f = ax^2 + bx + c \in \mathbb{R}(a, b, c)[x]$. This polynomial is in indeterminate x while a, b, c are some unspecified (varying) symbolic parameters.

Sullivant (2018, chap. 16.4) suggests one way of using the above proposition for checking rational identifiability of $f(\boldsymbol{\theta})$. First, compute a Gröbner bases of the ideal I_c with respect to a random lexicographic order and inspect the results for polynomials in the form $f(\boldsymbol{t}) - f(\boldsymbol{\theta})$. Some examples of this approach in the case of determining identifiability of ordinary differential equations can be found in Meshkat et al. (2009).

We note, that checking $f(\boldsymbol{t}) - f(\boldsymbol{\theta}) \in I_c$ means solving the ideal membership problem. Since we are interested in a particular parameter or a function of parameters, we can make some simple guesses of f. Suppose we are interested in the mixing proportion, $\theta = \pi$, and we expect the model to have label switching. Then we can guess $f(\theta) = \theta(1 - \theta)$ as $\theta(1 - \theta) - t(1 - t) = 0$ has exactly two solutions, so we check whether $\theta(1 - \theta) - t(1 - t) \in I_c$. If it is, then $f(\theta) = \theta(1 - \theta)$ is rationally identifiable and it means that we also can find solutions $f(\theta) = 0$ and identify θ. Interestingly, that in this case θ itself is not rationally but rather generically identifiable, since the solution involves square roots. If no label switching is expected, then our simplest guess is $f(\theta) = \theta$ and we are solving the ideal membership problem for $\theta - t \in I_c$ to check if θ is rationally identifiable.

3 Parameter estimation in mixture-like models

In this chapter we begin developing the key ideas of this thesis. We start by introducing a mixture-like model which, despite being parameterized in a similar way, suits record linkage data better than a regular two component mixture of discrete random variables. We provide a justification of model's suitability for record linkage and demonstrate how the model relates to the underlying data and redefine the parameters introduced in Section 2.2 according to the mixture-like conceptualization. This type of modelling does not require knowledge of the joint distribution of the comparisons vectors. We also explore some cases where the mixture-like approach may be inaccurate or even fail altogether when applied to record linkage tasks. As already discussed, the sampling or data generation mechanism behind the record linkage data is quite unusual and complicated. In this chapter we show that this mechanism has important consequences for the properties of parameter estimates of the mixture-like model (and regular mixtures as well). Nevertheless, at least in theory, an estimator that has a better agreement with the record linkage data generating mechanism can be obtained. To construct such an estimator, a special case of blocking is introduced. Finally, we discuss a method for estimating parameters of a mixture-like model.

3.1 Mixture-like model

We start discussing how record linkage and dual system estimation are connected by a more formal presentation of what we refer to as a mixture-like model. Suppose a vector $\boldsymbol{X}$ that can take

finitely many values $\boldsymbol{x}_j$, $j = 1, \ldots, J$. For every value $\boldsymbol{x}_j$ of $\boldsymbol{X}$ and some finite vector of parameters $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_g)^T, \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)^T$, a mixture-like model puts into correspondence a value $\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta}) \in [0, 1]$. Furthermore, in the case of the mixture-like model, it is possible to write

$$\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta}) = \sum_{i=1}^{g} \upsilon_i p_i(\boldsymbol{x}_j; \boldsymbol{\Theta}_i) \tag{20}$$

where $0 \leq \upsilon_i \leq 1$ and the following equality holds:

$$\sum_{i=1}^{g} \upsilon_i = 1;$$

as well as for all i, $0 \leq p_i(\boldsymbol{x}_j; \boldsymbol{\Theta}_i) \leq 1$ and the following holds

$$\sum_{j=1}^{J} p_i(\boldsymbol{x}_j; \boldsymbol{\Theta}_i) = 1.$$

Clearly, any discrete mixture model is mixture-like. For instance, mixture model (8) from Section 2.2.3 is such that $\boldsymbol{X}$ is the comparison outcome on K linkage variables, with possible values $\boldsymbol{x}_j$ being $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p, p = 1, \ldots, 2^K$, where $\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta}) = \text{pr}(\boldsymbol{\gamma}(a, b); \pi, \boldsymbol{\mu}, \boldsymbol{\nu})$, $\upsilon_1 p_1(\boldsymbol{x}_j; \boldsymbol{\Theta}_1) = \pi \text{pr}(\boldsymbol{\gamma}(a, b) \mid \mathcal{M}; \boldsymbol{\mu})$ and $\upsilon_2 p_2(\boldsymbol{x}_j; \boldsymbol{\Theta}_2) = (1 - \pi)\text{pr}(\boldsymbol{\gamma}(a, b) \mid \mathcal{U}; \boldsymbol{\nu})$.

Now $\boldsymbol{X}$ can be partitioned into several vectors, or scalars, or a combination of both and there is a corresponding factorization $p_i(\boldsymbol{x}_j; \boldsymbol{\Theta}_i) = p_{i,1}(\boldsymbol{x}_{j,1}) p_{i,2}(\boldsymbol{x}_{j,2}) \ldots$ which for all i satisfies

$$\sum_{\boldsymbol{x}_{j,n}} p_{i,n}(\boldsymbol{x}_{j,n}) = 1,$$

when summing across unique values of $\boldsymbol{x}_{j,n}$ and $0 \leq p_{i,n}(\boldsymbol{x}_{j,n}) \leq 1$.

In terms of partitioning of $\boldsymbol{X}$, one example could be the case of conditional independence between linkage variables given the match status. Then, with $K = 4$ linkage variables, $\boldsymbol{X}$ is partitioned into four scalars where each $x_{j,k}$ is either 0 or 1. Here $x_{j,k}$ is the k^{th} entry of $\boldsymbol{x}_j$, which corresponds to a comparison pattern $\boldsymbol{\gamma}_j$ in the case of the linkage model. For instance, if $\boldsymbol{x}_1 = \boldsymbol{\gamma}_1 = (1, 1, 1, 1)^T$ and $\boldsymbol{x}_2 = \boldsymbol{\gamma}_2 = (0, 1, 1, 1)^T$, then $x_{1,1} = 1$ and $x_{2,1} = 0$. Hence, in our record linkage example $p_{1,k}(x_{j,k}) = \mu_k$, $p_{2,k}(x_{j,k}) = \nu_k$ if $x_{j,k} = 1$ and $p_{1,k}(x_{j,k}) = 1 - \mu_k$, $p_{2,k}(x_{j,k} = 0) = 1 - \nu_k$ if $x_{j,k} = 0$ for $j = 1, \ldots, J$.

Another example of partitioning of $\boldsymbol{X}$ is when there is association between comparisons on a second linkage variable given the comparison outcome of the first linkage variable within both the sets of matches and non-matches. Then $\boldsymbol{X}$ is partitioned into three components, one is vector-valued and the remaining ones are scalars. Possible values of $x_{j,k}$ are $\boldsymbol{x}_{j,1} \in \{(1, 1)^T, (1, 0)^T, (0, 1)^T, (0, 0)^T\}$ for simultaneous agreements and disagreements on the first and second variables; $x_{j,2} \in \{0, 1\}$ and $x_{j,3} \in \{0, 1\}$.

Mixture-like models are more general than regular mixtures. In certain situations they are a more appropriate representation of a problem of interest. Specifically, $\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta})$, υ_i and $p_i(\boldsymbol{x}_j; \boldsymbol{\Theta}_i)$ need not be probabilities. Instead, they can be expectations of random variables or functions of random

variables that happen to satisfy the above conditions. Therefore, mixture-like models can be used in the situations where (20) models well the first moments, but not the higher moments.

In this thesis we are only considering two component variant of the mixture-like model (20) with $\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta}) = \pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu})$, $\upsilon_1 p_1(\boldsymbol{x}_j; \boldsymbol{\Theta}_1) = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$ and $\upsilon_2 p_2(\boldsymbol{x}_j; \boldsymbol{\Theta}_2) = (1-\pi)\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$:

$$\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu}) + (1-\pi)\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu}), \tag{21}$$

where $\boldsymbol{\gamma}_p$ is the comparison pattern indexed $p = 1, \ldots, 2^K$, $\boldsymbol{\mu}$ is the vector of parameters related to the outcomes in the set of matches, $\boldsymbol{\nu}$ is the vector of parameters related to the outcomes in the set of non-matches and the factorization of $\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$, $\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$ depends on the model specification used. Often, a model for the set of matches and the set of non-matches is not the same and therefore factorizations differ.

In order to see more clearly the difference between mixture models and mixture-like models, recall that by definition mixtures are suitable to model independent and identically distributed random variables or vectors (McLachlan & Peel, 2000, chap. 1.9). In other words, if we have a sample of w random vectors, every one of them has the common mass (or density) $\pi(\boldsymbol{X} = \boldsymbol{x}_j; \boldsymbol{\upsilon}, \boldsymbol{\Theta})$. In particular, the component memberships are also independent and identically distributed. Specifically, in the case of a finite number of components, these components are multinomially distributed (McLachlan & Peel, 2000, chap. 1.9). In record linkage models, we are dealing with two possible memberships. Therefore, in a regular mixture, the component membership is binomially distributed and the parameter π is the probability parameter of such a distribution. Also, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in (21) would be probabilities in a regular record linkage mixture. In principle, the component memberships may be associated in mixture models. If this is the case, hidden Markov models are used to estimate the parameters of interest (McLachlan & Peel, 2000, chap. 13). However, then the model is not parameterized exactly as (21) any more since it needs to reflect the transition probabilities between the components.

Recall the discussion of the record linkage experiment where we demonstrated that both the observed comparisons and component memberships lack independence. Particularly, we cannot think about record pairs, which constitute the data points used in parameter estimation, as being drawn independently. As an illustration, imagine a small population of size $\tau = 3$, with the population elements $\{e_1, e_2, e_3\}$. As everywhere else in this thesis, we assume no duplication. Therefore, we have 1-to-1 matching. The first survey is a sample of n_1 records from $\{s_{1,1}, s_{1,2}, s_{1,3}\}$ and the second survey is a sample of n_2 records from $\{s_{2,1}, s_{2,2}, s_{2,3}\}$. Hence, there are 9 possible record pairs. Without loss of generality, let $\mathrm{id}(s_{1,i}) = \mathrm{id}(s_{2,i}) = e_i$, so that the pairs $(1,1), (2,2), (3,3)$ are matches. Let $n_1 = n_2 = 2$, so we observe four record pairs. Since the record pairs are produced by taking the Cartesian product of $s_{1,a}$ and $s_{2,b}$, it is impossible, for instance, to observe the following sample of pairs: $\{(1,1), (2,2), (2,3), (3,3)\}$. On the other hand, having sampled, say, $s_{1,1}, s_{1,2}$ in the first sample and $s_{2,1}, s_{2,2}$ in the second implies the observed sample of pairs $\{(1,1), (1,2), (2,1), (2,2)\}$. Provided the pairs $(1,1)$ and $(2,2)$ are in a sample, $(2,1)$ and $(1,2)$ also must be in the sample. Furthermore, if $(1,1)$ and $(2,2)$ are matches, then $(2,1)$ and $(1,2)$ must be non-matches with probability 1, rather than with probability $1-\pi$. Since in a regular record linkage mixture the component membership is determined for each individual *pair* by the probability π which is *same* for each record pair, such a regular mixture

is not a correct model for the data in the linkage experiment. Nevertheless, in certain situation it is possible to usefully model the data by (21), but to do it at the aggregate rather than at the individual level and with expectations of certain ratios instead of the probabilities as the parameters.

More generally, for any population size τ there will be τ matches in the population and up to τ^2 record pairs can be observed, depending on the obtained sample sizes. The difference between mixture and mixture-like record linkage models, with respect to the mixing proportion π, is then as follows. In a regular mixture inspired by the above setting, the probability of getting a success (a pair being a match in the record linkage situation) is $\pi = 1/\tau$ and the total number of successes m (the total number of matches) out of $w = n_1 n_2 \leq \tau^2$ trials is binomially distributed with $\text{pr}(M = m) = \binom{w}{m}(1/\tau)^m(1 - 1/\tau)^{w-m}$. Each success (match) on a given trial (drawing a single record pair) occurs independently of the outcomes of other trials with the same fixed probability $\pi = 1/\tau$. There is no theoretical restriction on how many successes will be observed, as long as the number ranges between 0 and w. The expected number of successes is $w/\tau = n_1 n_2/\tau$. There are many situations where such a model is a good representation of the data generating mechanism. For instance, the cheating coin flipper in Drton et al. (2009, chap. 2.2) is such an example. This binomial model is not appropriate for record linkage, since the selections of the component memberships of record pairs are correlated. In other words, there is no fixed probability $\pi = 1/\tau$ of getting a match (success) in a sequence of draws. Hence, (21) cannot be a correct probability mass function of each record pair having a comparison $\boldsymbol{\gamma}$. It is possible to derive the probability mass function for the number of matches in a record linkage situations by using a combinatorial argument. This probability mass function accounts for correlated draws. The assumptions outlined in Section 2.1.2 are relevant here. We have $\binom{\tau}{n_1}$ and $\binom{\tau}{n_2}$ ways to draw the first and second samples, respectively. Then the total number of samples of record pairs that takes into account the constraints on the combinations of pairs that can be drawn in a particular sample is $\binom{\tau}{n_1}\binom{\tau}{n_2}$. There are $\binom{\tau}{m}$ ways to select exactly m matches out of the total τ matches. Note again, that we are choosing from τ, not from w because of the restrictions on what combinations of pairs can be selected. Then there are $\binom{\tau-m}{n_1-m}$ ways to select the remaining non-matches in the first sample and there are $\binom{\tau-n_1}{n_2-m}$ ways to select the remaining non-matches in the second sample. Provided $n_1 - m \geq 0$ and $n_2 - m \geq 0$ the probability of observing exactly m matches (matching pairs) is

$$\text{pr}(M = m) = \frac{\binom{\tau}{m}\binom{\tau-m}{n_1-m}\binom{\tau-n_1}{n_2-m}}{\binom{\tau}{n_1}\binom{\tau}{n_2}}.$$

After some manipulations, we can show that $\text{pr}(M = m) = \frac{\binom{\tau}{m}\binom{\tau-m}{n_1-m}\binom{\tau-n_1}{n_2-m}}{\binom{\tau}{n_1}\binom{\tau}{n_2}} = \frac{\binom{n_1}{m}\binom{\tau-n_1}{n_2-m}}{\binom{\tau}{n_2}}$, which is just a hypergeometric distribution where τ is the population size, n_1 is the number of successes in the population, n_2 is the number of draws and m is the number of observed successes. This is the same distribution which is frequently used in the simple dual system estimator to model the number of matches. What matters to our discussion on the distinction between mixture and mixture-like models, is the fact that in a sequence of draws from a hypergeometric distribution, the probability of getting, say, a second success is not the same as getting the first success (as the population size and the number of successes change as draws are carried out and successes are observed). Which shows that we are not dealing with a constant match (success) probability $\pi = 1/\tau$. Actually, a hypergeometric

distribution has a certain constant probability associated with it. It is the probability of observing a success on the i^{th} draw (which is obtained by summing up the probabilities of all possible sequences with a success on the i^{th} draw). This probability equals n_1/τ. Again, this is not $\pi = 1/\tau$, which we dealt with in the above example of a standard mixture. Since we are dealing with a hypergeometric distribution, the expected number of matches is $n_1 n_2/\tau$, the same as in the binomial model. Now we observe, that for fixed n_1, n_2 the expectation of the ratio of matches to the number of observed record pairs is $\mathbb{E}(M/W) = \mathbb{E}(M)/w = (n_1 n_2/\tau)(1/n_1 n_2) = 1/\tau = \pi$. Therefore, instead of postulating a regular mixture model which treats each observation as being drawn or generated independently, in the situation where it is not the case, we can postulate a mixture-like model. This model is parameterized in exactly the same way as a regular mixture, but its parameters are the expected values of the ratios of random variables rather than probabilities, as in the example above, where $\pi = \mathbb{E}(M/W)$. Such a postulation of course requires the remaining parameters $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ also to be meaningful expectations. It also requires ensuring that a mixture-like model for a given problem gives a reasonably accurate factorized approximation of the expected value of the random variable of interest. In Section 3.2 we demonstrate the suitability of a mixture-like approximation of the expectation of the ratio of a comparison pattern $\boldsymbol{\gamma}_p$ to the total number of record pairs.

For a given record linkage problem, there would be no differences between the parameterization of a regular mixture and mixture-like model. Also, at least with the parameter estimation methods discussed in this thesis, the parameter estimates would be exactly the same for these two types of models. The differences are conceptual and affect the way we think about record linkage and what can be deduced from each of the conceptualizations. A standard mixture is not a quite correct probability model. It implies incorrect variance estimation and cannot explain why the record linkage model works. More importantly, it does not offer the means to study properties of record linkage models and determine where such a model becomes less reliable. On the contrary, a mixture-like model is a well-defined approximation of the expectation of a random variable or vector, it prompts a careful consideration of variance estimation (though does not offer a straightforward solution), allows estimation of useful parameters and clarifies the nature of these parameters. It also provides a better explanation for why record linkage models work with the data generated by such a specific mechanism. Finally, as will be demonstrated in Section 3.2, it allows certain properties of the model to be studied.

An analogy from dual system estimation can be useful here. The estimators (4) and (5) give the same estimate of the population size. The latter, however, gives us not only the means to discover that the dual system estimator is biased whenever the inclusion probabilities are heterogeneous, but also provides a theoretical approximation of the extent of this bias (Wolter, 1986). Moreover, a post-stratification solution to reduce the heterogeneity bias is obvious from the analysis of (5), but not from (4).

3.2 Justification of the mixture-like model for record linkage

It was discussed earlier in Section 2.2.5 that, despite widespread applications to record linkage tasks, a finite mixture is not a valid statistical model for the problem. Yet, in this section we demonstrate that a mixture-like model (21) is a meaningful model for record linkage and the linkage free dual sys-

tem estimation. As already claimed, a mixture-like model means that a statistical model is genuinely parameterized as a mixture, but such a model, among other things, does not impose that every observation is drawn in turn from a certain component independently of other observations, but allow useful description of the record linkage data. The demonstration below is not a rigorous proof.

As far as parameter estimation approaches that do not require or assume the knowledge of joint distribution of comparison outcomes are available and a model in consideration is identifiable, meaningful parameter estimates of the mixture-like model can be obtained. However, as it is going to be shown, unlike well-defined probability parameters of a regular mixture, the parameters of the mixture-like model are ratios of the expectations of certain events. It is helpful to use the law of averages (Grimmet & Stirzaker, 2001, chap. 2.2) as an illustration of the difference here. If the law of averages allows us to naïvely regard the probability of an event A as the ratio of occurrences of this event to the number of trials, that is $\mathrm{pr}(A) = N(A)/N$, then the corresponding ratio of expectations is $R(A) = \mathbb{E}(N(A))/\mathbb{E}(N)$. We can think that a series of experiments are run under similar conditions with variable number of trials for each member of the series, rather than a single experiment with N trials. While it may be hard to see the reason for considering such a ratio here, it will be important in the later development.

It will be demonstrated that in the mixture-like record linkage model, instead of the probability π of a record pair being in the set of matches $\mathcal{M}$, the corresponding parameter is the ratio of the expected number of matches to the expected number of record pairs in repeated linkage experiments with a fixed setup. Also, instead of the probability $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu})$ of observing a comparison pattern $\boldsymbol{\gamma}_p$, the parameter of the mixture-like linkage model is the ratio of the expected number of the p^{th} pattern to the expected number of all record pairs. Similarly, instead of the probability $\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$ of the joint agreements and disagreements associated with a comparison pattern $\boldsymbol{\gamma}_p$ for a record pair $(s_{1,a}, s_{2,b}) = (a, b)$ belonging to the set of matches $\mathcal{M}$, the parameter of the mixture-like linkage model is the ratio of the expected number of matching pairs where comparison results in the pattern $\boldsymbol{\gamma}_p$ to the expected number of matched pairs. Finally, instead of the probability $\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$ of the joint agreements and disagreements in the set of non-matches $\mathcal{U}$ for the p^{th} pattern, the actual parameter we are dealing with in linkage is the ratio of the expected number of non-matching pairs where comparison results in the pattern $\boldsymbol{\gamma}_p$ to the expected number of non-matched pairs, and so on for any parameter of interest.

Recall from the previous discussion of the linkage model that we have a random variable N_1 that maps the outcome of drawing the first sample from the population to the size of the sample. Similarly N_2 maps the outcome of drawing the second sample to the size of this sample. The two samples are drawn independently. The number of record pairs is $W = N_1 N_2$, while the number of matches is M. None of the issues arising in the linkage model is related to W and M. For a long sequence of repeated linkage experiments with the fixed conditions, the expectations are $\overline{w}$ and $\overline{m}$, respectively. The number of non-matches is $U = W - M$, its expected values is $\overline{u}$. Unless stated otherwise, all expectations in this and following sections are obtained over a long sequence of repeated linkage experiments with the fixed set-up. It will be necessary later to know some distributional properties of W and M, but for the time being we do not assume any specific distribution.

Let an indicator function be $I\{(a, b) : \text{some property}\} = 1$ if a record pair (a, b) satisfies this

property, and $I\{(a,b) : \text{some property}\} = 0$ otherwise. As before, we are dealing with K linkage variables. For the record pairs in the set of non-matches $\mathcal{U}$, let $U_{k(1)} = \sum I\{(a,b) : s_{1,a,k} = s_{2,b,k}, (a,b) \in \mathcal{U}\}$ be the number of agreements and $U_{k(0)} = U - U_{k(1)} = \sum I\{(a,b) : s_{1,a,k} \neq s_{2,b,k}, (a,b) \in \mathcal{U}\}$ be the number of disagreements on the k^{th} variable. There are no assumptions on how either of $U_{k(1)}$, $U_{k(0)}$ is distributed and the numbers of agreements and disagreements are sums of outcomes that are in general correlated. The expected values are $\overline{u}_{k(1)}$ and $\overline{u}_{k(0)}$, respectively. In order to avoid overloading notation, we choose to abuse it slightly and, whenever it is possible, do not explicitly distinguish between $U_{k(1)}$ and $U_{k(0)}$, simply writing U_k in both cases. The corresponding expectation of U_k is $\overline{u}_k$. It will be clear whether agreements or disagreements are considered based on a given comparison pattern $\boldsymbol{\gamma}_p$. Also, define $U_{k(1),j(1)} = \sum I\{(a,b) : s_{1,a,k} = s_{2,b,k}, s_{1,a,j} = s_{2,b,j}, (a,b) \in \mathcal{U}\}$, $U_{k(1),j(0)} = \sum I\{(a,b) : s_{1,a,k} = s_{2,b,k}, s_{1,a,j} \neq s_{2,b,j}, (a,b) \in \mathcal{U}\}$ for a combination of simultaneous agreements and disagreements on two linkage variables. The expected values are $\overline{u}_{k(1),j(1)}$ and $\overline{u}_{k(1),j(0)}$, respectively. Then $U_{k(0),j(1)} = U_{j(1)} - U_{k(1),j(1)}$ and $U_{k(0),j(0)} = U_{j(0)} - U_{k(1),j(0)}$. Whenever distinction between agreements and disagreements is not needed, or a given comparison pattern indicates which one of the events is meant, we write $U_{k,j}$ which has the expected value $\overline{u}_{k,j}$. No assumptions about the distribution of these random variables are made.

In the same vein, variables $M_{k(1)} = \sum I\{(a,b) : s_{1,a,k} = s_{2,b,k}, (a,b) \in \mathcal{M}\}$, $M_{k(0)} = M - M_{k(1)}$, $M_{k(1),j(1)}$ and $M_{k(1),j(0)}$ are defined for the matching pairs. Since each comparison patterns $\boldsymbol{\gamma}_p$ gives information on whether agreements or disagreements are of interest for the k^{th} linkage variable, we again abuse the notation by simply using M_k. The expected values of M_k, $M_{k(1),j(1)}$ and $M_{k(1),j(0)}$ are $\overline{m}_k$, $\overline{m}_{k(1),j(1)}$ and $\overline{m}_{k(1),j(0)}$, respectively. As before, no distributional assumptions are made.

We can now use the above definition to facilitate understanding of the reasoning behind the parameterization of the linkage model as a mixture-like. Since some technical details related to the discussion become tedious as the complexity of models increases, we give a generalised presentation for the case of the conditional independence of comparison outcomes between linkage variables given the match status. After, we look only at several special cases of the models with dependence in comparison outcomes between linkage variables.

We revisit the record linkage experiment introduced earlier (Section 2.2.5). Suppose we run a record linkage data generating experiment in which we can observe all outcomes of interest, including those that are not directly observed in real linkage exercises. The conditions of this particular experiment are such that neither errors of recording values of the linkage variables are correlated between any variables, nor the values of any linkage variable, given the value of another linkage variable, tend to get selected substantially more frequently compared to what would be obtained by simple random sampling of the values of the linkage variable. Such conditions lead to the case of conditional independence in comparisons between the linkage variables given the match status. For any given linkage variable, individual agreement and disagreement outcomes are correlated in the set of non-matches $\mathcal{U}$.

Since every record pair is either a match or a non-match, once the linkage experiment is run and the binary comparison outcomes are summed for each of the comparison pattern within matches and non-matches, the result can be viewed as two contingency tables with 2^K cells in each of the table. We can call the table related to the matching pairs the *match table*, and the table related to the non-matching

pairs the *non-match table*. Each of the cells of both tables corresponds to the comparison pattern $\boldsymbol{\gamma}_p$ and contains the number of cases of this particular pattern in the $\mathcal{M}$ or $\mathcal{U}$-sets. The marginal counts of the first table are $M_{k(1)}$ and $M_{k(0)}$, while the marginal counts of the second table are $U_{k(1)}$ and $U_{k(0)}$. Since these random variable are sums of correlated outcomes in general and there are many parameters in play (such as the range of the possible values of linkage variables), the distributions of M_k, U_k are unknown. It is only known that, in this setup, the margins are independent.

Summing two cells for a given comparison pattern gives the frequency, f_p, of the pattern which is observed in a linkage exercise. Dividing this number by the number of pairs w gives the overall relative frequency of the pattern, $\pi(\boldsymbol{\gamma}_p)$.

There can be between $2K+1$ to $2(2^K-1)+1 = 2\cdot 2^K - 1$ parameters needed to parameterize the model (two tables and the mixing proportion), but only 2^K observables. We attempt to achieve some reduction in parameters and model $\pi(\boldsymbol{\gamma}_p) = f_p/w$ with $2^K - 1$ or fewer parameters. In the case of the conditional between-variables independence given the match status we are aiming for the model with the minimum possible number of parameters $2K+1$. Estimate the ratio of the number of matches to the overall number of pairs by the realization of M/W. For each linkage variable in the matching set estimate the ratio of agreements or disagreements to the number of matching pairs by the realization of M_k/M. Now fix an arbitrary comparison pattern $\boldsymbol{\gamma}_p$, which is the equivalent of fixing a corresponding cell in the contingency tables. Estimate the relative frequency of the comparison pattern (or cell) in the set of matches to the number of matches by the realization of $(M_1/M)(M_2/M)$.... This is similar to estimating the probability of a particular cell in a $2\times 2\times\ldots$ contingency table under the mutual independence assumption. The key difference from modelling a regular contingency table is that the 'number of trials' M or U are random variables and no assumption about the distribution of a cell count is made. Then we can estimate the relative frequency of the cell in the match table to the overall number of record pairs by $(M/W)(M_1/M)(M_2/M)$.... The same is done for the set of non-matches and the results are summed to produce the estimate of $\pi(\boldsymbol{\gamma}_p)$. Overall, for a given comparison pattern $\boldsymbol{\gamma}_p$ we are dealing with the following random variable

$$\Pi_p = g_p(W, M, U, M_1, \ldots, M_K, U_1, \ldots, U_K) = \frac{M}{W}\frac{M_1}{M}\cdots\frac{M_K}{M} + \frac{U}{W}\frac{U_1}{U}\cdots\frac{U_K}{U}, \tag{22}$$

with deliberately avoided cancellation of M and U. Note that it is the pattern $\boldsymbol{\gamma}_p$ that tells us for which k we have $M_{k(1)}$, $U_{k(1)}$ and for which we have $M_{k(0)}$, $U_{k(0)}$. For instance, if $K=4$ and, say, the pattern $\boldsymbol{\gamma}_2 = (1,1,1,0)^T$, then in this case we are dealing with $M_1 = M_{1(1)}, M_2 = M_{2(1)}, M_3 = M_{3(1)}, M_4 = M_{4(0)}$ and $U_1 = U_{1(1)}, U_2 = U_{2(1)}, U_3 = U_{3(1)}, U_4 = U_{4(0)}$.

Suppose we run the linkage experiment with a fixed set-up repeatedly and we are interested in approximating the expected value $\mathbb{E}(\Pi_p)$ using the second order Taylor expansion around the point $(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_K, \overline{u}_1, \ldots, \overline{u}_K)$ to see if

$$\mathbb{E}(\Pi_p) = \mathbb{E}\left(\frac{M}{W}\frac{M_1}{M}\cdots\frac{M_K}{M} + \frac{U}{W}\frac{U_1}{U}\cdots\frac{U_K}{U}\right) \approx \frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}}\cdots\frac{\overline{m}_K}{\overline{m}} + \frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}}\cdots\frac{\overline{u}_K}{\overline{u}}. \tag{23}$$

This will allow us to answer the question of whether and under what conditions we can achieve the desired reduction of the number of parameters in a repeated linkage exercise using the mixture-like

parameterization.

Recall, that for a function $g(x_1, \ldots, x_K)$ of K variables, the second order Taylor series g_T around the point $(a_1, \ldots, a_K)$ is given by

$$\begin{aligned} g_T(a_1, \ldots, a_K) = g(a_1, \ldots, a_K) + \sum_{k=1}^{K} \frac{\partial g(a_1, \ldots, a_K)}{\partial x_k}(x_k - a_k) \\ + \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{K} \frac{\partial^2 g(a_1, \ldots, a_K)}{\partial x_k \partial x_j}(x_k - a_k)(x_j - a_j), \end{aligned} \tag{24}$$

where $g(a_1, \ldots, a_K)$ is function g evaluated at the point $(a_1, \ldots, a_K)$, $\partial g(a_1, \ldots, a_K)/\partial x_k$ is the partial derivative of g with respect to x_k evaluated at the same point, and $\partial^2 g(a_1, \ldots, a_K)/\partial x_k \partial x_j$ is the second partial derivative of g with respect to x_k and x_j also evaluated at the above point.

When taking the expectation of the Taylor expansion around the points that are the expected values of the random variables, the following expectations arise: $\mathbb{E}(x_k - \overline{x}_k) = 0$, $\mathbb{E}[(x_k - \overline{x}_k)^2] = \mathrm{Var}(x_k)$ and $\mathbb{E}[(x_k - \overline{x}_k)(x_j - \overline{x}_j)] = \mathrm{Cov}(x_k, x_j)$. It means that all terms related to the first derivatives can be disregarded.

The first term in the approximation is g_p evaluated at $(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_K, \overline{u}_1, \ldots, \overline{u}_K)$:

$$g_p(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_K, \overline{u}_1, \ldots, \overline{u}_K) = \frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}} \ldots \frac{\overline{m}_K}{\overline{m}} + \frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}} \ldots \frac{\overline{u}_K}{\overline{u}}.$$

Our strategy to check if this expression is a good approximation for $\mathbb{E}(\Pi_p)$ is going to consist in working out the order of $g_p(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_K, \overline{u}_1, \ldots, \overline{u}_K)$ as well as the order of the remaining terms and checking whether the remaining terms have substantially smaller order. Here by order $O(x^n)$ of a function $g(x)$ we mean that $|g(x)| \leq Cx^n$ for some constant C. Since we are interested in order, we can disregard the sign and a numeric coefficient of each term in the expansion. That is the reason why we do not need to distinguish at this point between $M_{k(1)}$ and $M_{k(0)}$ as defined above. The derivative of the corresponding terms will only differ in sign.

Note that the first term on the right hand side of (22) is structurally the same as the second term. Therefore, it is sufficient to work out the relevant set of derivatives for one of the terms and then re-use them for the second term.

Working out all the derivatives and considering only those that are unique up to some constant multiple, hence the use proportionality rather than equality, we have to deal with 5 terms:

$$\mathbb{E}\left(\frac{\partial^2 g_p(\overline{w}, \overline{m}, \ldots, \overline{m}_k, \ldots \overline{u}_k, \ldots)}{\partial W^2}(W - \overline{w})(W - \overline{w})\right) \propto \frac{\prod_{k=1}^{K} \overline{m}_k}{\overline{w}^3 \overline{m}^{K-1}}\mathrm{Var}(W), \tag{25}$$

$$\mathbb{E}\left(\frac{\partial^2 g_p(\overline{w}, \overline{m}, \ldots, \overline{m}_k, \ldots \overline{u}_k, \ldots)}{\partial M^2}(M - \overline{m})(M - \overline{m})\right) \propto \frac{\prod_{k=1}^{K} \overline{m}_k}{\overline{w} \cdot \overline{m}^{K+1}}\mathrm{Var}(M), \tag{26}$$

$$\mathbb{E}\left(\frac{\partial^2 g_p(\overline{w}, \overline{m}, \ldots, \overline{m}_k, \ldots \overline{u}_k, \ldots)}{\partial M \partial W}(M - \overline{m})(W - \overline{w})\right) \propto \frac{\prod_{k=1}^{K} \overline{m}_k}{\overline{w}^2 \overline{m}^{K}}\mathrm{Cov}(M, W), \tag{27}$$

$$\mathbb{E}\left(\frac{\partial^2 g_p(\overline{w},\overline{m},\ldots,\overline{m}_k,\ldots\overline{u}_k,\ldots)}{\partial M\partial M_j}(M-\overline{m})(M_j-\overline{m}_j)\right)\propto\frac{\prod_{k=1,k\neq j}^{K}\overline{m}_k}{\overline{w}\cdot\overline{m}^K}\mathrm{Cov}(M,M_j),\tag{28}$$

$$\mathbb{E}\left(\frac{\partial^2 g_p(\overline{w},\overline{m},\ldots,\overline{m}_k,\ldots\overline{u}_k,\ldots)}{\partial W\partial M_j}(W-\overline{w})(M_j-\overline{m}_j)\right)\propto\frac{\prod_{k=1,k\neq j}^{K}\overline{m}_k}{\overline{w}^2\overline{m}^{K-1}}\mathrm{Cov}(W,M_j).\tag{29}$$

The derivatives for the second term of the right-hand side of (22) are as above with U and U_k instead of M and M_k.

Now we focus on the order of $\overline{w},\overline{m}$ and $\overline{m}_k$. Since the population size is τ, neither $\overline{m}$ nor $\overline{m}_k$ can exceed it. Therefore we conclude that the $O(\overline{m})$ is τ. Since $\overline{w}$ involves the product of sizes of two samples from a population of size τ, $O(\overline{w})$ is τ^2. Now if dealing with the agreements on the k^{th} linkage variable, the order of $\overline{m}_k$ is also τ. If dealing with disagreements, let the positive integer ϵ be some number of disagreements we observe on the k^{th} variable, $\epsilon <= \tau$. For simplicity, we use a common ϵ for all of the linkage variables. Then, assuming that the number of disagreements is small relatively to the number of agreements, $O(\overline{m}_k)$ (for the number of disagreements on the k^{th} variable) is ϵ. In addition, let z be the number of disagreements in the pattern $\boldsymbol{\gamma}_p$. Then the order

$$O\left(\frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}}\cdots\frac{\overline{m}_K}{\overline{m}}\right)\text{ is }\frac{\tau}{\tau^2}\frac{\tau}{\tau}\cdots\frac{\epsilon}{\tau}\cdots=\frac{\epsilon^z}{\tau^{z+1}},\tag{30}$$

so the order of the above term is between ϵ^K/τ^{K+1} and $1/\tau$.

Regarding the order of $\overline{w},\overline{u}$ and $\overline{u}_k$ for the second term of the right-hand side of (22), we have the following. Clearly $O(\overline{w})$ is τ^2. Now in majority of the cases $\overline{u}$ is close to $\overline{w}$, hence we can think that $O(\overline{u})$ is also τ^2. We expect a fair number of disagreements on the linkage variable v_k among the non-matches. Hence, in case of disagreements we are safe to think that $O(\overline{u}_k)$ is again τ^2. In the case of agreements, recall that ρ_k is the number of unique values the k^{th} linkage variable can take. Assume for simplicity the same ρ for all of the linkage variables. Then roughly, the number of agreements on a certain linkage variable among non-matches is τ^2/ρ. This figure is not necessarily very accurate, but it should be sufficient when discussing the orders. Again, let z be the number of disagreements and the order of the term

$$O\left(\frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}}\cdots\frac{\overline{u}_K}{\overline{u}}\right)\text{ is }\frac{\tau^2}{\tau^2}\frac{\tau^2/\rho}{\tau^2}\cdots\frac{\tau^2}{\tau^2}\cdots=\frac{1}{\rho^{K-z}}.\tag{31}$$

Overall we have a tuple of orders

$$\left(\frac{\epsilon^z}{\tau^{z+1}},\frac{1}{\rho^{K-z}}\right)\tag{32}$$

associated with $g_p(\overline{w},\overline{m},\overline{m}_1,\ldots,\overline{m}_K,\overline{u}_1,\ldots,\overline{u}_K)$.

We can work out the order for each variance and covariance in (25) – (29). In several instances, we will be using the fact that

$$-\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}\leq\mathrm{Cov}(X,Y)\leq\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.\tag{33}$$

Fix τ,π_1 and π_2. Suppose that $N_1\sim Bin(\tau,\pi_1)$, $N_2\sim Bin(\tau,\pi_2)$ and N_1, N_2 are independent.

Then $M \sim Bin(\tau, \pi_1\pi_2)$. Note that these distributions are the consequence of assumption that the cells of a contingency table are distributed according to the multinomial distribution, when odds ratio is 1 (see Section 2.1.2). As we have seen earlier, the multinomial distribution is frequently used when discussing the dual system estimation. There is no particular reason confining to the binomial distribution except the fact that the orders of variances and covariances can be easily obtained for in this case. In principle, any distributions can be used, as long as such orders are similar to the orders under the binomial distribution. At the end of this section, we will visit a distribution which can lead to a different order of variance.

Since the variance of M is $\text{Var}(M) = \pi_1\pi_2(1-\pi_1\pi_2)\tau$, the order $O(\pi_1\pi_2(1-\pi_1\pi_2)\tau)$ is τ.

Aiming to work out the order of $\text{Var}(W)$, we use $W = N_1N_2$ and the assumption that two samples are independent, so that $\mathbb{E}(W) = \mathbb{E}(N_1N_2) = \mathbb{E}(N_1)\mathbb{E}(N_2)$. Now, $\mathbb{E}(N_1) = \pi_1\tau$ and $\text{Var}(N_1) = \mathbb{E}(N_1^2) - \mathbb{E}(N_1)^2 = \pi_1(1-\pi_1)\tau$ imply that $\mathbb{E}(N_1^2) = \pi_1\tau(1-\pi_1+\pi_1\tau)$. Similarly, $\mathbb{E}(N_2^2) = \pi_2\tau(1-\pi_2+\pi_2\tau)$. Then

$$\begin{aligned}\text{Var}(W) &= \mathbb{E}(W^2) - \mathbb{E}(W)^2 = \mathbb{E}(N_1^2N_2^2) - \mathbb{E}(N_1N_2)^2 = \mathbb{E}(N_1^2)\mathbb{E}(N_2^2) - \mathbb{E}(N_1N_2)^2 \\ &= \pi_1\tau(1-\pi_1+\pi_1\tau)\pi_2\tau(1-\pi_2+\pi_2\tau) - \pi_1^2\pi_2^2\tau^4 \\ &= -2\pi_1^2\pi_2^2\tau^3 + \pi_1^2\pi_2\tau^3 + \pi_1\pi_2^2\tau^3 + \pi_1^2\pi_2^2\tau^2 - \pi_1^2\pi_2\tau^2 - \pi_1\pi_2^2\tau^2 + \pi_1\pi_2\tau^2\end{aligned}$$

and the order of $\mathbb{E}(W)$ is τ^3.

We use (33) to work out conservatively the order of covariance terms:

$$O(\text{Cov}(W,M)) \leq O\left(\sqrt{\text{Var}(W)\text{Var}(M)}\right) \text{ is } \tau^{1.5}\tau^{0.5} = \tau^2,$$

$$O(\text{Cov}(W,M_k)) \text{ is } \begin{cases}\tau^{1.5}\tau^{0.5} = \tau^2 & \text{if } M_k \text{ is the number of agreements} \\ \tau^{1.5}\epsilon^{0.5} \leq \tau^2 & \text{if } M_k \text{ is the number of disagreements,}\end{cases}$$

$$O(\text{Cov}(M,M_k)) \text{ is } \begin{cases}\tau^{0.5}\tau^{0.5} = \tau & \text{if } M_k \text{ is the number of agreements} \\ \tau^{0.5}\epsilon^{0.5} \leq \tau & \text{if } M_k \text{ is the number of disagreements.}\end{cases}$$

In most situations it is the case that $U \approx W$, hence the order of $\text{Var}(U)$ is the same as the order of $\text{Var}(W)$, which is τ^3.

With the help of (33) it is possible to determine the orders of covariance terms associated with the non-match table:

$$O(\text{Cov}(W,U)) \leq O\left(\sqrt{\text{Var}(W)\text{Var}(U)}\right) \text{ is } \tau^{1.5}\tau^{1.5} = \tau^3,$$

$$O(\text{Cov}(W,U_k)) \approx O(\text{Cov}(U,U_k)) \text{ is } \begin{cases}\tau^{1.5}\tau/\rho^{0.5} = \tau^{2.5}/\rho^{0.5} & \text{if } U_k \text{ is the number of agreements} \\ \tau^{1.5}\tau = \tau^{2.5} & \text{if } U_k \text{ is the number of disagreements.}\end{cases}$$

Now it is possible to determine the order of (25) – (29) associated with both the match and non-match tables by plugging in the orders of individual terms. For the match table the orders of terms

are following:

$$O\left(\frac{\prod_{k=1}^{K}\overline{m}_k}{\overline{w}^3\overline{m}^{K-1}}\text{Var}(W)\right) \text{ is } \frac{\epsilon^z}{\tau^{z+2}},$$

$$O\left(\frac{\prod_{k=1}^{K}\overline{m}_k}{\overline{w}\cdot\overline{m}^{K+1}}\text{Var}(M)\right) \text{ is } \frac{\epsilon^z}{\tau^{z+2}},$$

$$O\left(\frac{\prod_{k=1}^{K}\overline{m}_k}{\overline{w}^2\overline{m}^{K}}\text{Cov}(M,W)\right) \text{ is } \frac{\epsilon^z}{\tau^{z+2}},$$

$$O\left(\frac{\prod_{k=1,k\neq j}^{K}\overline{m}_k}{\overline{w}\cdot\overline{m}^{K}}\text{Cov}(M,M_j)\right) \text{ is } \begin{cases}\frac{\epsilon^{z^*}}{\tau^{z^*+2}} & \text{if } M_j \text{ is the number of agreements}\\ \frac{\epsilon^{z^*+0.5}}{\tau^{z^*+2.5}} & \text{if } M_j \text{ is the number of disagreements,}\end{cases}$$

$$O\left(\frac{\prod_{k=1,k\neq j}^{K}\overline{m}_k}{\overline{w}^2\overline{m}^{K-1}}\text{Cov}(W,M_j)\right) \text{ is } \begin{cases}\frac{\epsilon^{z^*}}{\tau^{z^*+2}} & \text{if } M_j \text{ is the number of agreements}\\ \frac{\epsilon^{z^*+0.5}}{\tau^{z^*+2.5}} & \text{if } M_j \text{ is the number of disagreements.}\end{cases}$$

We use z^* in the terms that already incorporate an agreement or a disagreement. This means that there may be some restrictions on what values z^* can take. For instance, if M_j is the number of agreements, then $z^* \neq K$. It reflect the fact, that there are situations, where the order of some terms remains unchanged for the different numbers of the overall disagreements.

For the non-match table the orders are following:

$$O\left(\frac{\prod_{k=1}^{K}\overline{u}_k}{\overline{w}^3\overline{u}^{K-1}}\text{Var}(W)\right) \text{ is } \frac{1}{\rho^{K-z}\tau},$$

$$O\left(\frac{\prod_{k=1}^{K}\overline{u}_k}{\overline{w}\cdot\overline{u}^{K+1}}\text{Var}(U)\right) \text{ is } \frac{1}{\rho^{K-z}\tau},$$

$$O\left(\frac{\prod_{k=1}^{K}\overline{u}_k}{\overline{w}^2\overline{u}^{K}}\text{Cov}(U,W)\right) \text{ is } \frac{1}{\rho^{K-z}\tau},$$

$$O\left(\frac{\prod_{k=1,k\neq j}^{K}\overline{u}_k}{\overline{w}\cdot\overline{u}^{K}}\text{Cov}(U,U_j)\right) \text{ is } \begin{cases}\frac{1}{\rho^{K-z^*-0.5}\tau^{1.5}} & \text{if } U_j \text{ is the number of agreements}\\ \frac{1}{\rho^{K-z^*-1}\tau^{1.5}} & \text{if } U_j \text{ is the number of disagreements,}\end{cases}$$

$$O\left(\frac{\prod_{k=1,k\neq j}^{K}\overline{u}_k}{\overline{w}^2\overline{u}^{K-1}}\text{Cov}(W,U_j)\right) \text{ is } \begin{cases}\frac{1}{\rho^{K-z^*-0.5}\tau^{1.5}} & \text{if } U_j \text{ is the number of agreements}\\ \frac{1}{\rho^{K-z^*-1}\tau^{1.5}} & \text{if } U_j \text{ is the number of disagreements.}\end{cases}$$

It can be seen that the above terms generally have substantially smaller order than the terms of interest (32). There may be several cases when this is not true. First, when ρ is small, that is all or just certain linkage variable have a few possible values they can take. The extreme examples would

be binary variables like sex, student indicator, born in the country of residence, etc. In this case the term $1/\rho^{K-z}\tau$ from the non-match table need not be substantially smaller than the main term of the match table ϵ^z/τ^{z+1}. This is later confirmed in simulation work, results are presented in Section 7.6. On the other hand, if ρ is substantially larger than τ, then the order ϵ^z/τ^{z+2} of the term (25) from the match table may be larger than the main term of the non-match table $1/\rho^{K-z}$. This is also confirmed in simulations. However, the effect of having variables with a small number of levels appears to have a bigger effect on the performance of the methods and is arguably more likely to occur in real applications.

Hence, as long as we are dealing with W and M of similar order as above and ρ is neither very small, say binary or ternary, nor very large relatively to τ for all or most variables, the following approximation

$$\mathbb{E}(\Pi_p) = \mathbb{E}\left(\frac{M}{W}\frac{M_1}{M}\cdots\frac{M_K}{M} + \frac{U}{W}\frac{U_1}{U}\cdots\frac{U_K}{U}\right) \approx \frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}}\cdots\frac{\overline{m}_K}{\overline{m}} + \frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}}\cdots\frac{\overline{u}_K}{\overline{u}}$$

holds.

The beta-binomial would be an example of a distribution that may result in orders of the above terms being different from those considered in our discussion. If N_1 follows the beta-binomial distribution with the number of trials τ, and parameters α, β, then the variance of this random variable is $(\tau\alpha\beta(\alpha+\beta+\tau))/((\alpha+\beta)^2(\alpha+\beta+1))$. If N_2 also follows the beta-binomial with the same number of trials, then for some choices of parameters α and β the order of $\mathrm{Var}(W)$ can be τ^4 and some of $(25)-(29)$ may be of order similar to the order of the terms (32). It is not clear if such an extreme situation can occur in real applications since it requires extraordinary heterogeneous response probabilities.

With those caveats in mind, we we can now redefine the parameters of the linkage model as

$$\pi = \frac{\overline{m}}{\overline{w}},$$

$$\mu_k = \frac{\overline{m}_k}{\overline{m}},$$

$$1-\mu_k = 1 - \frac{\overline{m}_k}{\overline{m}},$$

$$\nu_k = \frac{\overline{u}_k}{\overline{u}},$$

$$1-\nu_k = 1 - \frac{\overline{u}_k}{\overline{u}},$$

and replace the ratios in (23) with the above parameters to get the mixture-like model of independence between the outcomes of comparisons on different linkage variables given the match status. In this case $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)^T$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3, \nu_4)^T$, and the relative frequency of each pattern γ_p can be written as

$$\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_2\nu_3\nu_4,$$
$$\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_2\nu_3\nu_4,$$
$$\pi(\boldsymbol{\gamma}_3;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_1(1-\nu_2)\nu_3\nu_4,$$
$$\dots$$

From now on we will often be using the compact notation for mixture-like models. In this notation the between-variables associations of the comparison outcomes will be given in brackets after μ_p and ν_p corresponding to the set of matches and non-matches, respectively. A single γ_k means that the comparison on the k^{th} variable is independent of comparisons on any other variables, and $\gamma_{k,l}$ means that there is association between comparisons on the k^{th} and l^{th} linkage variables. So the compact form for the above independence model is $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$.

It is possible to use the above approach to show that dependencies between comparison outcomes of the linkage variables can be similarly taken into account. We consider here only two models with between-variables dependencies for the case of $K = 4$ linkage variables. The first model is considered throughout this work because it is likely to appear in practical applications. The second model is less useful in practice, but serves as a good example of what happens to the approximation like (23) when the complexity of a model of interest increases. Both models considered are identifiable. The above approach can be applied for any parameterization of a mixture-like model if required.

The first model is of the conditional between-variables independence given the set of matches and dependence between v_k and v_j, say between v_1 and v_2, in the set of non-matches. We write this model in a compact form as

$$\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$$

and it is parameterized as

$$\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4,$$
$$\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4,$$
$$\pi(\boldsymbol{\gamma}_3;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_1(1-\nu_{2|1(1)})\nu_3\nu_4,$$
$$\dots$$

for the comparison patterns $\gamma_1 = (1,1,1,1)^T, \gamma_2 = (0,1,1,1)^T, \gamma_3 = (1,0,1,1)^T, \dots, \gamma_{16} = (0,0,0,0)^T$.

The terms related to the match table are the same as in the discussion above. The ratio of the interest when the dependence between two variables is present is estimated as $(U_1/U)(U_{2,1}/U_1)$. So this time we are aiming to show that the following approximation holds:

$$\begin{aligned}\mathbb{E}(\Pi_{p,2|1}) &= g_{p,2|1}(W,M,U,M_1,\dots,M_4,U_1,U_{2,1},\dots,U_4)\\ &= \mathbb{E}\left(\frac{M}{W}\frac{M_1}{M}\frac{M_2}{M}\frac{M_3}{M}\frac{M_4}{M} + \frac{U}{W}\frac{U_1}{U}\frac{U_{2,1}}{U_1}\frac{U_3}{U}\frac{U_4}{U}\right) \approx \frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}}\cdots\frac{\overline{m}_4}{\overline{m}} + \frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}}\frac{\overline{u}_{2,1}}{\overline{u}_1}\frac{\overline{u}_3}{\overline{u}}\frac{\overline{u}_4}{\overline{u}}.\end{aligned} \tag{34}$$

First, we determine the order of $(U_1/U)(U_{2,1}/U_1)$. Note, that if U_1 is the number of agreements,

then $U_{2,1}/U_1$ is of order $1/\rho$, no matter whether $U_{2,1}$ is the number of agreements or disagreements. If U_1 is the number of disagreements, then the order of $U_{2,1}/U_1$ is $1/\rho$ and 1 for agreements and disagreements, respectively. Then the tuple of orders associated with $g_{p,2|1}(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_4, \overline{u}_1, \overline{u}_{2,1}, \ldots, \overline{u}_4)$ is

$$\begin{cases} \left(\frac{\epsilon^z}{\tau^{z+1}}, \frac{1}{\rho^{2-z^*}}\right) & \text{if } U_1, U_{2,1} \text{ are the numbers of disagreements} \\ \left(\frac{\epsilon^z}{\tau^{z+1}}, \frac{1}{\rho^{3-z^*}}\right) & \text{otherwise.} \end{cases} \tag{35}$$

Note that z in this case is the number of disagreements in a comparison pattern related to two remaining independent variables. So z can take values of $0, 1$ or 2 in this model.

Accounting for between-variables dependence in this model results in the order having a smaller power of ρ compared to the case when comparisons are independent between linkage variables in the non-match table. A similar reduction of the power of ρ occurs in the remaining terms $(25) - (29)$ of the approximation. It is sufficient to check whether the term with $\text{Var}(W)$ is substantially smaller than the terms (35). For the model of interest, this term is

$$\frac{\overline{u}_{2,1}\overline{u}_3\overline{u}_4}{\overline{w}^3\overline{u}^2}\text{Var}(W)$$

and the corresponding order is

$$\begin{cases} \frac{1}{\rho^{2-z^*}\tau} & \text{if } U_1, U_{2,1} \text{ are the numbers of disagreements} \\ \frac{1}{\rho^{3-z^*}\tau} & \text{if otherwise.} \end{cases}$$

Again, the approximation (34) holds as long as the ρ not too small compared to τ. Since having between-variables dependency reduces the power of ρ, one can anticipate that for a given ρ with a few levels the approximation is less accurate when the above dependence is present compared to the accuracy in the independence between linkage variables case.

We can now define the parameters of the linkage model similarly to the model of conditional independence between variables given the match status presented above, but with $\nu_{2,1} = \overline{u}_{2,1}/\overline{u}_1$.

The second model is of the conditional independence between linkage variables given the set of matches and dependence in the set of non-matches of variables v_j, v_l on the value of v_k. For instance, v_2 and v_3 depend on the value of v_1. A compact form is

$$\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_4)$$

and is parameterized as

$$\begin{aligned} \pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_{3|1(1)}\nu_4, \\ \pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_{3|1(0)}\nu_4, \\ \pi(\boldsymbol{\gamma}_3; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_1(1-\nu_{2|1(1)})\nu_{3|1(1)}\nu_4, \\ &\ldots \end{aligned}$$

for the patterns $\gamma_1 = (1,1,1,1)^T, \gamma_2 = (0,1,1,1)^T, \gamma_3 = (1,0,1,1)^T, \ldots, \gamma_{16} = (0,0,0,0)^T$.

We want to show that the following approximation holds:

$$\begin{aligned}\mathbb{E}(\Pi_{p,2|1,3|1}) &= g_{p,2|1,3|1}(W, M, U, M_1, \ldots, M_4, U_1, U_{2,1}, U_{3,1}, U_4) \\ &= \mathbb{E}\left(\frac{M}{W}\frac{M_1}{M}\frac{M_2}{M}\frac{M_3}{M}\frac{M_4}{M} + \frac{U}{W}\frac{U_1}{U}\frac{U_{2,1}}{U_1}\frac{U_{3,1}}{U_1}\frac{U_4}{U}\right) \approx \frac{\overline{m}}{\overline{w}}\frac{\overline{m}_1}{\overline{m}}\cdots\frac{\overline{m}_4}{\overline{m}} + \frac{\overline{u}}{\overline{w}}\frac{\overline{u}_1}{\overline{u}}\frac{\overline{u}_{2,1}}{\overline{u}_1}\frac{\overline{u}_{3,1}}{\overline{u}_1}\frac{\overline{u}_4}{\overline{u}}.\end{aligned} \tag{36}$$

The tuple of orders associated with $g_{p,2|1}(\overline{w}, \overline{m}, \overline{m}_1, \ldots, \overline{m}_4, \overline{u}_1, \overline{u}_{2,1}, \overline{u}_{3,1}\overline{u}_4)$ is

$$\begin{cases}\left(\frac{\epsilon^z}{\tau^{z+1}}, \frac{1}{\rho^{1-z^*}}\right) & \text{if } U_1, U_{2,1}, U_{3,1} \text{ are the numbers of disagreements} \\ \left(\frac{\epsilon^z}{\tau^{z+1}}, \frac{1}{\rho^{3-z^*}}\right) & \text{if } U_1 \text{ is the number of disagreements, } U_{2,1}, U_{3,1} \text{ are the number of agreements} \\ \left(\frac{\epsilon^z}{\tau^{z+1}}, \frac{1}{\rho^{2-z^*}}\right) & \text{otherwise.}\end{cases} \tag{37}$$

Again, we need only to check the order of the term with $\mathrm{Var}(W)$, which is either $\frac{1}{\rho^{3-z^*}\tau}$, or $\frac{1}{\rho^{2-z^*}\tau}$, or $\frac{1}{\rho^{1-z^*}\tau}$. As in the previous model, z is the number of disagreements on the remaining variable not involved in dependencies, so that z can take values of $0, 1$ in this model.

The desired approximation holds under the same conditions as before, but accuracy deteriorates even faster than in all models considered so far if the number of levels in ρ decreases. Again, we define the parameters of the corresponding mixture-like model as above, this time with $\nu_{2,1} = \overline{u}_{2,1}/\overline{u}_1$ and $\nu_{3,1} = \overline{u}_{3,1}/\overline{u}_1$.

3.3 Some issues related to parameter estimation of a mixture-like model

The previous section discussed the nature of the parameters of a mixture-like linkage model. It also highlighted several cases, where the parameterization may fail to provide an accurate reflection of the outcomes of the linkage experiment. Such failure results from a mixture-like model's drastic reduction of the number of original parameters involved in the linkage experiment and associated loss of the important information required for accurate representation of the outcomes. In this section, a few more issues with the model will be looked at, this time in relation to parameter estimation of mixture-like models.

Note again that the parameters of interest are the ratios of the means. Take for instance, $\pi = \overline{m}/\overline{w}$, where $\overline{m} = \mathbb{E}(M), \overline{w} = \mathbb{E}(W)$. However, in any real application, we are dealing with a single observable realisation of W, and a single unobservable realization of M. Essentially, we have a single observation available for estimation, which is different from the most cases of estimation or statistical analysis where we have a sequence of observation $x_1, \ldots, x_n$ that are, say, generated from the same distribution or are drawn according to the same sampling design. Dealing with a single observation has several implications both for the point and variance estimation.

One way to see the issues related to the point estimation is to observe that estimating the above linkage model parameter does not conform to the law of large numbers. Recall, that according to the law of large numbers if random variables $X_1, X_2, \ldots$ are independent and identically distributed and have the mean μ_x, then $\sum_{i=1}^{n} X_i/n \to \mu_x$ as $n \to \infty$. However, in the linkage exercise n corresponds to the number of repetitions of the linkage experiment, rather than a sample size, and $n = 1$ in practice.

With a single observable realisation of W and a single unobserved realisation of M, we can obtain neither the estimate of $\overline{m}/\overline{w}$, nor $\overline{w}$, nor $\overline{m}$ with the above properties. While we can anticipate, and it is confirmed with the simulation work (Section 7.5), $\widehat{\pi}$ to be in general quite close to $\overline{m}/\overline{w}$ to be a useful estimator, this estimator does not tend to $\overline{m}/\overline{w}$ with the increase of τ and therefore is not consistent.

The impact on variance estimation of essentially dealing with a single observation is rather obvious: there is no information available to make inference of uncertainty associated with the estimation of the linkage parameters.

Another complication associated with a mixture-like (and regular mixture) linkage model is that the principal parameter of interest π tends to 0 as the population size τ tends to infinity. This follows from $\mathbb{E}(M) = \pi_1\pi_2\tau$, $W = \pi_1\pi_2\tau^2$ so that

$$\frac{\mathbb{E}(M)}{\mathbb{E}(W)} = \frac{\pi_1\pi_2\tau}{\pi_1\pi_2\tau^2} = \frac{1}{\tau} \to 0 \text{ as } \tau \to \infty. \tag{38}$$

Therefore, there is none of the asymptotic behaviour one usually expects from a well-defined statistical estimator for the estimator of the linkage parameter π and its functions.

In many practical applications, blocking is used to reduce the number of record pairs W and the above limit may not hold. However, a blocking simply aiming at a reduction of the number of pairs may have the adverse effect on the accuracy of the approximation of the mixture-like record linkage model since the orders of the terms involving W may be different to what we used in Section 3.2. Note also that an arbitrary blocking makes analysis similar to the one in previous section more difficult. Finally, being able to estimate the parameter π at the population or estimation stratum level is crucial for the no-classification dual system estimation as will be shown in Section 4.1.

In the next section, we will discuss how the issues described in this section can, at least in theory, be overcome. We will also see that a certain special case of blocking not only reduces the number of record pairs in processing, but has a more profound impact on the parameter estimation. This in turn opens opportunities for variance estimation.

In spite of these drawbacks, a mixture-like model is a useful approach to linkage and related parameters estimation as demonstrated by simulation work. It also allows some possibilities for studying certain properties of the model and the corresponding estimators. Such theoretical tractability should not be taken for granted given the complexity of record linkage problem from the statistical point of view.

3.4 Constructing a data-conforming estimator and averaging blocking

An approach to constructing an estimator for linkage and related parameters that corresponds or conforms to the data as generated in the repeated record linkage experiment is outlined in this section. Again, it is not a rigorous demonstration, which is left for the future work. However, the conceptual development presented here agrees well with the simulations (Section 7.6) and paves the way for variance estimation.

Suppose we are performing a single iteration of the linkage experiment with the population size

τ and the coverage probabilities of two data samples π_1, π_2. We stress again, that each element of the population has a constant or near constant probability π_1 to be selected in the first sample, and constant or near constant probability π_2 to be selected in the second sample. We also assume that all remaining parameters of the experiment, such as error rates and ranges of the values of linkage variables, do not vary substantially between possible subsets of the population. This experiment gives rise to two data sets S_1, S_2 of the size N_1 and N_2, respectively.

As always, comparing every record of S_1 with every record of S_2 results in the $W = N_1 N_2$ record pairs, M of which are matches. Suppose that the set-up of the experiment remains unchanged, but the population is split into B non-overlapping groups, $G_1, \ldots, G_\beta, \ldots, G_B$, of equal or nearly equal size τ_β, so that $\tau = \sum_{\beta=1}^{B} \tau_\beta$. In this case we have several non-overlapping samples $S_{1,1}, \ldots, S_{1,B}$, such that $S_1 = \cup_{\beta=1}^{B} S_{1,\beta}$, and non-overlapping samples $S_{2,1}, \ldots, S_{2,B}$, such that $S_2 = \cup_{\beta=1}^{B} S_{2,\beta}$. Each $S_{1,\beta}, S_{2,\beta}$ is drawn from G_β. The corresponding random variables that map the outcome of sampling to the sample sizes are $N_{1,1}, \ldots, N_{1,B}$, with $N_1 = \sum_{\beta=1}^{B} N_{1,\beta}$, and $N_{2,1}, \ldots, N_{2,B}$, with $N_2 = \sum_{\beta=1}^{B} N_{2,\beta}$. Note, that while $N_{1,\beta}, N_{1,\alpha}, \beta \neq \alpha$ are generated from the same model, they are not in general equal; the same applies for $N_{2,\beta}, N_{2,\alpha}$.

Instead of making comparisons between every record of S_1 with every record of S_2, every record of $S_{1,\beta}$ is compared to every record of $S_{2,\beta}$, for $\beta = 1, \ldots, B$. For each group β, the number of resulting record pairs is W_β and the number of matches is M_β. Since we assume the absence of overcount, it follows that $M = \sum_{\beta=1}^{B} M_\beta$, where M is the number of matches resulting from comparing the entire S_1 to the entire S_2. However, because no comparisons between $S_{1,\beta}$ and $S_{2,\alpha}, \beta \neq \alpha$ are made, $W \neq \sum_{\beta=1}^{B} W_\beta$. For the p^{th} comparison pattern in the group β, the corresponding frequency is $f_{p,\beta}$ and the relative frequency is $\pi_\beta(\boldsymbol{\gamma}_p) = f_{p,\beta}/w_\beta$. Consider the average $\overline{\pi}_p$ of $\pi_\beta(\boldsymbol{\gamma}_p), \beta = 1, \ldots, B$. This average is an empirical version of (22), with the W_β and $\overline{w}_\beta$ instead of W and $\overline{w}$. Estimation of $\overline{\pi}_p$ allows estimation of $\overline{m}$, which then allows estimation of the remaining parameters of interest. What is really important here is that having a single iteration of the linkage experiment, we constructed a situation that mimics multiple runs of the linkage experiment, corresponds to the conceptualisation (22) of a mixture-like model and also conforms with the law of large numbers in the way explained in the previous section. This addresses the first issue with the estimation of linkage parameters and allows, at least in theory, obtaining an estimate of the actual parameter $\pi = \overline{m}/\overline{w}$.

Grouping and making comparisons as described above can also be used to avoid convergence to 0 as shown in (38). Let the population size τ increase. While it is increasing, gradually increase B so that $\tau_1, \ldots, \tau_\beta, \ldots \tau_B$ are approximately of the same size that allows stable estimation and do not increase as τ increases. Then $\tau \to \infty$ causes $B \to \infty$, but the ratio of expectations (38) remains $1/\tau_\beta$.

Splitting the population or stratum of interest into non overlapping groups generated by the same or very similar mechanisms $G_1, \ldots, G_\beta, \ldots, G_B$, of equal or nearly equal size τ_β, and making comparisons of records in $S_{1,\beta}$ to the records in $S_{2,\beta}$ is a special case of blocking. So that G_β is essentially a block. This blocking differs from standard blocking approaches used in record linkage in a number of ways. Usually, a classical blocking is done to reduce the computational burden by discarding vast numbers of pairs that most likely are not matches. Such blocking is deemed efficient as long it does not lead to matching pairs being missed while substantially reducing the overall number of pairs. Also, in a

classical blocking, comparisons are made within blocks and then the comparison results are either pooled together for parameter estimation or analysis or estimation is done for each block separately. The blocking we are presenting here requires data points in blocks to be generated by the same or very similar models and be of the same or very similar size. The goal of such blocking is not only to reduce the number of record pairs to process, but also to get the data into a format that corresponds to the conceptualisation of a mixture-like model under repetitive linkage experiment. To achieve the latter goal, comparisons are not pulled together, but averaged out for each of the comparison pattern $\boldsymbol{\gamma}_p$, to satisfy the nature of the record linkage model. In order to distinguish this blocking from any classical blocking approach, we call such blocking the *averaging blocking*. In fact, outcomes averaged across the blocks and those pooled across the blocks are proportional to each other, since $\sum f_{p,\beta} = B\overline{f}_p$. Hence, we can see that this certain type of blocking sets the data in accordance with the model if the conditions of the same block generating model and block size are satisfied. Nevertheless, the averaging blocking does not mean that the corresponding estimator is consistent. This is due to the fact that a mixture-like model is an approximation that depends on the properties of the population attributes.

Averaging blocking has been considered from the purely theoretical viewpoint. But is it possible in any real applications where nearly equal size blocks are required? This seems difficult and even may look like a contradiction if the population size estimation is ones goal. However, in applications like the census, address frames or address listings are available for both data sources and postcodes rather than individuals or households are the sampling units. So it is possible to aggregate postcodes or output areas together using the address frame information into blocks with approximately similar numbers of households. This still may lead to a reasonably high variation of τ_β. The reason for τ_β to be of equal or nearly equal sizes is that it guarantees orders of W and M for which a mixture-like model holds. Some variation in τ_β does not affect the model itself. Some simulations related to the practical application of averaging blocking are presented in Section 7.5.

The development in this section has an important consequence for variance estimation as it allows outcomes in each block obtained by averaging blocking to be treated as a realization of a random vector. With several blocks constructed as above, one has a few realisations of the random vector and can proceed to variance estimation. There are certain problems associated with the variation of τ_β and the fact that we will be interested in the estimate of π, or τ rather than π_β or τ_β. Solutions of these problems and variance estimation for the linkage free dual system estimator will be presented in Chapter 6.

3.5 Parameter estimation using Markov chain Monte Carlo methods

Complex data generating mechanism encountered in the case of record linkage (see Section 2.2.5) leads to modelling the resulting data by mixture-like models (see Section 3.2). Such models are models for expectations of ratios of random variables, rather than regular well-defined probability models. These data complexities and the nature of linkage model do limit estimation methods that can be justified for parameter estimation. Earlier, it was demonstrated, that the maximum likelihood method is not fully justifiable in the record linkage setting. Note, we are not saying that the maximum likelihood approach for record linkage is not feasible in general, but an underlying statistical model cannot be as

simple as a two-component mixture of discrete random variables (8).

The approach chosen for the parameter estimation of the record linkage model is to minimize the distance between the observed ratios $\pi(\boldsymbol{\gamma}_p)$ of the number of cases in a given comparison pattern $\boldsymbol{\gamma}_p$ to the number of record pairs (which would be probabilities in a regular probability model) and the estimated ratios $\pi(\boldsymbol{\gamma}_p; \widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}) = \left[\widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}}) + (1-\widehat{\pi})\nu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\nu}})\right] / f_p$, where $\boldsymbol{\mu}, \boldsymbol{\nu}$ are parameters for a given specification of the model. Such a specification may take into account between-variables associations of the comparison outcomes either in the set of non-matches, or the set of matches, or both. For instance, a model with the conditional independence between linkage variables given the set of matches and association between first and second linkage variables given the set of non-matches is $\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4, \pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4, \ldots$ and so on.

The distance minimization approach used in this thesis is the minimum modified chi-squared estimator (Agresti, 2002, chap. 15). In this case for given observables w and $\boldsymbol{\pi} = (\pi(\boldsymbol{\gamma}_1), \ldots, \pi(\boldsymbol{\gamma}_{2^K}))^T$, we are searching for $\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}$ which produce the estimated values $\boldsymbol{\pi}(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}) = (\pi(\boldsymbol{\gamma}_1; \widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}), \ldots, \pi(\boldsymbol{\gamma}_{2^K}; \widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}))^T$ that minimize the modified chi-squared statistic

$$\chi^2\left[\boldsymbol{\pi}, \boldsymbol{\pi}(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})\right] = w \sum_{p=1}^{2^K} \frac{\left[\pi(\boldsymbol{\gamma}_p) - \pi(\boldsymbol{\gamma}_p; \widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})\right]^2}{\pi(\boldsymbol{\gamma}_p)}. \tag{39}$$

The actual vector of parameter values $(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})^T$ that minimizes the modified chi-squared statistic is obtained by the Markov chain Monte Carlo algorithm called the simulated annealing, briefly presented in Section 2.4. In this case the target function is $\chi^2\left[\boldsymbol{\pi}, \boldsymbol{\pi}(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})\right]$ and the equilibrium distribution is $\Pi = (\pi(\boldsymbol{\gamma}_1), \ldots, \pi(\boldsymbol{\gamma}_{2^K}))^T$. For a given model specification, the first step is to randomly produce the vector of parameter values $(\pi, \boldsymbol{\mu}, \boldsymbol{\nu})^T$ and compute the corresponding value of the target function. Values are drawn from the admissible range $(0, 1)$ and for some parameters certain constraints may be imposed. For instance the maximum of π may be quite small, or at least smaller than 0.5 to prevent label switching in certain model specifications. For each temperature parameter t, n_t iterations of the Metropolis algorithm are run. At each iterations, one parameter from $(\pi, \boldsymbol{\mu}, \boldsymbol{\nu})^T$ is randomly chosen and the value for this parameter is drawn randomly from the uniform distribution of possible values of the parameter and the value of $\chi^2\left[\boldsymbol{\pi}, \boldsymbol{\pi}(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})\right]$ is recomputed. If this value is smaller than the value on the previous step, than the current vector of parameter values is accepted with some probability that depends on the temperature t. Otherwise, the vector of parameter values from the previous step is used. The next step repeats the above routine. As the temperature t decreases, the probability of rejecting a solution that is better on the current step than a solution on the previous, decreases. However, allowance to reject a better solution avoids being trapped about a local minimum of χ^2. The temperature decreases gradually using the parameter known as the cooling rate. Parameters such as temperature, cooling rate and the number of iterations of the Metropolis algorithm at each value of the temperature parameter are customized and need to be tuned. Once the algorithm terminates, the outcome is the vector $(\widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})^T$of estimated parameters of the linkage model that minimizes χ^2. As already mentioned, theoretically if the algorithm runs long enough, it should result in global minimum.

This estimation approach neither makes any assumptions about how the data are distributed, nor

postulates that $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu})$ should be a probability and not an expected ratio. The only condition is that the model is identifiable. This approach is very flexible and can easily work with various model specification so that it can be used with models that account for association between comparison outcomes of linkage variables. It works both without blocking and with the averaging blocking. In principle, such chi-squared estimators are the best asymptotic estimators. However, given that record-linkage is rather non-regular statistical problem, it is not known whether and in what sense the above property holds. This is left for future work. Note that in the case of regular mixtures, other methods to find the parameters that minimize the modified chi-squared statistic are suggested in the literature (Titterington et al., 1985, chap. 4.5.3), for instance, the Newton-Raphson method. But given our interest in estimating parameters of various parameterizations of a mixture-like model, the simulating annealing seems more flexible and reliable.

4 Connection between record linkage and dual system estimation

In this chapter we show that there is a close connection between certain parameters of probabilistic record linkage based on the mixture-like model and dual system estimation. More precisely, under a specific set-up, the dual system estimate of the population size follows from such a linkage model. This relationship implies that, at least in principle, population size estimation from two incomplete lists is feasible through the estimation of the linkage model parameters without the classification of the record pairs into links and non-links. More generally it means that some linkage requiring tasks do not need classification-based linkage and purely estimation-based linkage is sufficient. As a result, as long as an adequate linkage model can be specified, a fully automated record linkage without clerical reviews can be achieved; no matches are erroneously classified as non-links or non-matches classified as links; seamless dual system estimation and better uncertainty measures for the resulting population size estimates are possible. While our focus is on linkage for population size estimation, the ideas presented here can certainly be used for other applications as well.

4.1 Linkage free dual system estimator

We now show that the dual system estimator can be expressed as a function of one of the parameters of a mixture-like record linkage model. This relationship demonstrates that record linkage under the mixture-like model and dual system estimation are closely related. Since the population size estimate can be obtained using a parameter estimate of a linkage model rather than using a set of records classified as links, we will be calling the corresponding estimator the linkage free estimator or no-classification estimator. Also, given that estimation of linkage model parameters is sufficient for the task, we will refer to the corresponding linkage as no-classification linkage. While the motivation and meaning of 'no-classification linkage' is quite self-explanatory, despite contradicting the strict definition of record linkage, some clarification of why the name 'linkage free dual system estimation' was chosen is needed. The word 'linkage' in 'linkage free dual system estimation' is used in its narrow meaning as a process of establishing whether several records correspond to the same entity in a population or not manifested in the form of classification. Obviously, the linkage free dual system estimator requires

all the data preparation of any other linkage exercise. However, since there is no-classification, there is no linkage in the narrow sense of the term. Furthermore, anybody familiar with the dual system estimation knows that this estimator requires the number of linked records as an input. Therefore, saying 'no-classification dual system estimator' does not immediately tell us where to expect differences compared to the classical estimator.

The set up is the same as presented in Sections 2.1.1 and 2.2.2. We are interested in estimating the unobservable size τ of the population of interest $\mathcal{P}$. Two surveys of the population are available: S_1 and S_2 with n_1 and n_2 observations, respectively. We start with a situation where record linkage is carried out using S_1 and S_2 without blocking. It is important to note that in this case such entities as our population domain, estimation stratum and linkage block are equivalent and have the same number τ of the population elements belonging to them. The number of record pairs is $w = n_1 n_2$. Recall that n_1, n_2 and w are the realizations of the corresponding random variables N_1, N_2 and W. Assume also that all the additional parameters, such as sets of genuine values population attributes can take or probabilities of errors recording these attributes, that influence the results of linkage but are not directly reflected in the mixture-like record linkage model, are fixed.

Derivation of the linkage free dual system estimator is simple. Fix a particular set-up of a linkage experiment and suppose that the experiment is run repeatedly many times. Recall, that the dual system estimator itself can be derived using the relationship (5). Combining it with the relevant mixture-like model parameters we obtain

$$\tau \approx \frac{\mathbb{E}(N_1)\mathbb{E}(N_2)}{\mathbb{E}(M)} = \frac{\overline{n}_1\overline{n}_2}{\overline{m}} = \frac{\overline{n}_1\overline{n}_2}{\overline{m}\frac{\overline{w}}{\overline{w}}} = \frac{\overline{n}_1\overline{n}_2}{\pi\overline{w}} = \frac{\overline{w}}{\pi\overline{w}} = \frac{1}{\pi}. \tag{40}$$

This demonstrates how the dual system estimator of the population size τ and the parameter π of the record linkage model are related, given the assumptions of the dual system estimation, except the perfect linkage one, are satisfied and the estimation stratum and linkage block being the same. Certainly, both the dual system estimator and the parameters of the mixture-like linkage model are based on approximations, so that the relationship is also approximate.

The linkage free dual system estimator, sometimes referred to as the π-based estimator, is defined as

$$\widetilde{\tau} = \frac{1}{\widehat{\pi}} \tag{41}$$

where $\widehat{\pi}$ is the estimate of π obtained using the simulated annealing algorithm as presented in Sections 2.4 and 3.5. This estimator demonstrates, that in the case of population size estimation, linkage need not necessarily involve classification and can be entirely estimation-based. However, it requires an accurate estimate $\widehat{\pi}$ which requires specification of an identifiable model that adequately accounts for association between the linkage variables used. Note that while this estimator explicitly involves only one parameter, the simulated annealing approach cannot estimate π without estimating $\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$ and $\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$. Recall, however, that some identifiability checking methods allow the identifiability of a single parameter to be established. From that perspective, it can be sufficient to check if π is identifiable in a particular model for the linkage free dual system estimation. As a side note, the fact that identifiability of a single parameter can be checked also means that it is in principle possible to use a method similar

to the method of inversion (Everitt & Hand, 1981, chap. 1.4.3) in order to obtain an estimate of π only. In this case, however, one still needs to correctly specify the linkage model.

Another way to derive the linkage free dual system estimator is as follows. For a comparison pattern $\boldsymbol{\gamma}_p$ multiply both sides of a mixture-like record linkage model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu}) + (1-\pi)\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$ by the average number of record pairs $\overline{w}$ to obtain

$$\bar{f}_p = \pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu})\overline{w} = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})\overline{w} + (1-\pi)\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})\overline{w} = \overline{m}_p + \overline{u}_p,$$

where $\bar{f}_p$ is the mean number of record pairs in the p^{th} comparison pattern and

$$\overline{m}_p = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})\overline{w} \tag{42}$$

is the mean number of matches in the pattern $\boldsymbol{\gamma}_p$. Similarly, $\overline{u}_p$ is the mean number of non-matches in the same pattern. Since we are dealing with a mixture-like model, we have $\sum_{p=1}^{2^K} \mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu}) = 1$, and hence we can work out the mean for matches across all comparison patterns

$$\sum_{p=1}^{2^K} \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})\overline{w} = \pi\overline{w} = \frac{\overline{m}}{\overline{w}}\overline{w} = \overline{m}. \tag{43}$$

All these derivations come directly from the definitions of the parameters as presented in Section 3.2.

This allows a slightly different linkage free dual system estimator to be derived. We first use the simulated annealing method to obtain the parameter estimates for the appropriate record linkage model and use them alongside the observed number of record pairs, w to estimate of the number of matches

$$\widehat{m} = w\widehat{\pi}\sum_{p=1}^{2^K} \mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}}).$$

Now we can define the m-based linkage free dual system estimator

$$\widetilde{\tau}_m = \frac{n_1 n_2}{\widehat{m}}. \tag{44}$$

Such m-based estimator is more flexible than the π-based and can work when averaging blocking and for other blocking strategies, provided blocking does not result in some true matches being excluded from the input data and model specification adequately reflects between-variables associations of comparison outcomes. Suppose there we are dealing with B blocks. Then with the averaging blocking, the corresponding m-based linkage free dual system estimator of the population total τ is

$$\dot{\tau}_m = \frac{n_1 n_2}{B\widehat{\overline{m}}} = \frac{n_1 n_2}{B\overline{w}_b \sum_{p=1}^{2^K} \widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})}, \tag{45}$$

where $\widehat{\overline{m}}$ is the estimate of the number of matches based on the averaged observed frequencies, $\overline{f}_p$, of B blocks when the averaging blocking is employed with $\widehat{\pi}, \mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})$ being the corresponding parameter estimates of the mixture-like linkage model, and $\overline{w}_b$ is the average number of pairs in blocks. Based

on the discussion in Section 3.4, the above estimator has the best theoretical justification for practical usage among all linkage free dual system estimators. Simulations show, however, that in many practical situations the distinction in performance between (41) and (44) is not so prominent.

If an arbitrary blocking on the population $\mathcal{P}$ of the size τ is applied and blocks $\beta = 1, \ldots, B$ are not too small, then either (41) or (44) can be applied within each block to estimate the size of the block and then individual estimates can be summed:

$$\check{\tau} = \sum_{\beta=1}^{B} \frac{1}{\widehat{\pi}_\beta}, \tag{46}$$

$$\check{\tau}_m = \sum_{\beta=1}^{B} \frac{n_{1,\beta}, n_{2,\beta}}{\widehat{m}_\beta}, \tag{47}$$

where $\widehat{\pi}_\beta, \widehat{m}_\beta$ are estimates and $n_{1,\beta}, n_{2,\beta}$ are observed survey counts related to the block β. This approach can be used with the post-stratification as described in Section 2.3.

The main attractiveness of the linkage free estimation approach is that is does not require clerical resolution which leads to a higher level of automation of record linkage and seamless population size estimation. Taking out the clerical part of the record linkage process reduces the associated cost substantially. The linkage free dual system estimation may also be beneficial from the privacy perspective as it does not create links between records found on two surveys. The described approach, however, requires thoughtful model specification and checking of this model for identifiability. This may mean, that not all linkage variables available can be used in the linkage exercise. As discussed earlier, the simulated annealing-based parameter estimation approach neither requires knowledge of the joint distribution of the comparisons outcomes within each linkage variable, nor make any unrealistic assumptions about it. An estimation-based approach will generally have higher variance than the classification-based approach, even if the same linkage model and parameter estimation procedure were employed in both cases.

A worked-out example of how the application of the linkage free dual system estimator looks like is presented later in this work in Section 7.4.

4.2 Modified linkage free dual system estimator to reflect 1-to-1 matches

The linkage free dual system estimator seamlessly integrates record linkage and capture-recapture estimation. However, the absence of classification and clerical resolution of harder cases results in the higher variability of the linkage free estimates compared to the classical classification approaches involving some clerical resolution; see Section 7.5 with the simulation and estimation results. It is therefore important to explore ways of reducing variability of the linkage free estimator. Given the nature of mixture-like record linkage models it is difficult to establish the theoretical minimum of the variance of the linkage free estimator. After all, even for regular mixtures data reduction via sufficiency cannot be achieved (Fienberg et al., 2009). Nevertheless, it is possible to improve the efficiency of the estimator. In this section, a modified version of the linkage free estimator is presented. As shown by simulations, this modified estimator reduces the variability of the linkage free estimator by 5 to 35%.

The nature of linkage problem for population size estimation when undercoverage is present on both surveys, gives some additional information or some constraining conditions. This is known as the 1-to-1 match constraint and was briefly discussed in Section 2.2.4. Assuming that both surveys S_1 and S_2 either contain no duplicates, or were deduplicated prior to linkage, the 1-to-1 matching implies that every record on survey S_1 either matches to one and only one record on survey S_2, or has no match. In linkage terms it means that every record on survey S_1 can be either linked to one and only one record on survey S_2 or be classified as a non-link.

This constraint provides important and useful information about the data structure and one may expect that harnessing it will improve the precision of the estimation. Obviously, no-classification record linkage does not produce the classification in the first place. Therefore, the task is to integrate a classification constraint into an approach that avoids classification. The modified linkage free estimator presented in this section is one of the ways to achieve the task.

The modified linkage free dual system estimator has the same assumptions and set up as the standard linkage free estimator discussed in the previous section. In fact, the modified version uses the linkage free estimator as the basis, but has several additional steps. Recall, that the sizes of the surveys S_1, S_2 are n_1 and n_2, respectively. Carrying out the binary comparison of record pairs on K linkage variables yields 2^K comparison patterns $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p, \ldots, \boldsymbol{\gamma}_{2^K}$. The corresponding observed frequencies of comparison patterns are $f_1, \ldots, f_p, \ldots, f_{2^K}$. The modified estimator consists of the following steps.

1. The linkage free estimator (41) or (44) and related parameters of the mixture-like linkage model are obtained first using the simulated annealing approach. A model specification must take into account potential dependencies between linkage variables either in the set of non-matches, or the set of matches, or both. The estimated parameters are:
   $\widehat{\pi}$ is the estimate of the ratio of the mean number of matches to the mean number of record pairs, $\overline{m}/\overline{w}$;

   $\widetilde{\tau} = 1/\widehat{\pi}$ is the corresponding linkage free estimator;

   $\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})$ is the estimate of the ratio of the number of matches in a comparison pattern $\boldsymbol{\gamma}_p$ to the total number of matches. The estimate of the number of matches in the p^{th} pattern based on (42) is also available as the result of parameter estimation

   $$\widehat{m}_p = w\widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}}), \tag{48}$$

   where w is the observed number of record pairs;

   $\nu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\nu}})$ is the estimate of the ratio of the number of non-matches in a comparison pattern $\boldsymbol{\gamma}_p$ to the total number of non-matches;

2. Since the parameters of mixture-like models employed in record linkage are expectations rather than probabilities, we cannot use the definition of the conditional probability to obtain the

relationship $\text{pr}((a,b) \in \mathcal{M} \mid \boldsymbol{\gamma}(a,b) = \boldsymbol{\gamma}_p) = \pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu})/\left(\pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu}) + (1-\pi)\nu(\boldsymbol{\gamma}_p;\boldsymbol{\nu})\right)$. However, for each comparison pattern $\boldsymbol{\gamma}_p$ define the ratio of the contribution of the matches to the p^{th} comparison pattern to the proportion of the p^{th} pattern among all patterns:

$$r(\mathcal{M} \mid \boldsymbol{\gamma}_p) = \frac{\pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu})}{\pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu}) + (1-\pi)\nu(\boldsymbol{\gamma}_p;\boldsymbol{\nu})} = \frac{\pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu})}{\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu})}. \tag{49}$$

Clearly

$$r(\mathcal{U} \mid \boldsymbol{\gamma}_p) = 1 - r(\mathcal{M} \mid \boldsymbol{\gamma}_p) = \frac{(1-\pi)\nu(\boldsymbol{\gamma}_p;\boldsymbol{\nu})}{\pi\mu(\boldsymbol{\gamma}_p;\boldsymbol{\mu}) + (1-\pi)\nu(\boldsymbol{\gamma}_p;\boldsymbol{\nu})} = \frac{(1-\pi)\nu(\boldsymbol{\gamma}_p;\boldsymbol{\nu})}{\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu})},$$

and $r(\mathcal{M} \mid \boldsymbol{\gamma}_p) + (1 - r(\mathcal{M} \mid \boldsymbol{\gamma}_p)) = 1$. Again, we use the fact that the record linkage mixture-like model has relationships similar to those one would encounter in a regular mixture, but not necessarily with probabilities as parameters.

We obtain the estimates of the above ratios: $\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p) = \widehat{\pi}\mu(\boldsymbol{\gamma}_p;\widehat{\boldsymbol{\mu}})/\pi(\boldsymbol{\gamma}_p;\widehat{\pi},\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\nu}})$ and $\widehat{r}(\mathcal{U} \mid \boldsymbol{\gamma}_p) = (1-\widehat{\pi})\nu(\boldsymbol{\gamma}_p;\widehat{\boldsymbol{\nu}})/\pi(\boldsymbol{\gamma}_p;\widehat{\pi},\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\nu}})$.

3. For a pattern $\boldsymbol{\gamma}_p$ compute the following log-odds ratio (or a pseudo log-odds ratio since we are not dealing with probabilities):

$$\hat{l}_p = \log\frac{\hat{\text{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\left[1 - \hat{\text{r}}(\mathcal{U} \mid \boldsymbol{\gamma}_p)\right]}{\left[1 - \hat{\text{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\right]\hat{\text{r}}(\mathcal{U} \mid \boldsymbol{\gamma}_p)}. \tag{50}$$

Note that this log-odds ratio has a proper probabilistic counterpart in a univariate logistic regression (Agresti, 2002, chap. 5). If $x = \{0,1\}$ and a logistic model is $\text{logit}\left[p(x)\right] = \alpha + \beta x$, then

$$\beta = \text{logit}\left[p(1)\right] - \text{logit}\left[p(0)\right] = \log\frac{\widehat{\text{pr}}(1)\left[1 - \widehat{\text{pr}}(0)\right]}{\left[1 - \widehat{\text{pr}}(1)\right]\widehat{\text{pr}}(0)}.$$

In the case of linkage, for a comparison pattern $\boldsymbol{\gamma}_p$, $x_p = \{I_{\mathcal{M}|\boldsymbol{\gamma}_p}, I_{\mathcal{U}|\boldsymbol{\gamma}_p}\} = \{0,1\}$, where $\mathbb{I}_{\mathcal{M}|\boldsymbol{\gamma}_p}$ and $I_{\mathcal{U}|\boldsymbol{\gamma}_p}$ are mutually exclusive indicator functions. Using the relation above

$$\hat{l}_p = \log\frac{\hat{\text{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\left[1 - \hat{\text{r}}(\mathcal{U} \mid \boldsymbol{\gamma}_p)\right]}{\left[1 - \hat{\text{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\right]\hat{\text{r}}(\mathcal{U} \mid \boldsymbol{\gamma}_p)} = \widehat{\beta}_p$$

in $\text{logit}\left[\text{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\right] = \alpha + \widehat{\beta}_p x_p$.

4. Once $\hat{l}_p$ are obtained, they can be used as weights in solving an optimization problem called the assignment problem (Bertsekas, 1998, chaps. 1, 7). In our situation, solving the assignment problem will lead to an optimal pairing of each record in S_1 with one and only one record in S_2 based on the above weights. Overall, $\min(n_1, n_2)$ pairs are formed when solving the assignment problem. The actual algorithm for solving the assignment problem requires a square matrix of weights as an input. Therefore, a $\max(n_1, n_2) \times \max(n_1, n_2)$ matrix is used as an input with some of the 'redundant' entries set to 0. Note that comparison of an arbitrary record pair (a, b) results in one and only one comparison pattern $\boldsymbol{\gamma}_p$, so that it is always possible to find all a and

b that result in comparison pattern $\boldsymbol{\gamma}_p$. This is useful, because it is sometimes more convenient to write $\hat{l}_p = \hat{l}_{ab}$ for a certain pattern $\boldsymbol{\gamma}_p$ and a pair (a, b).

The assignment problem is formulated in the following way: for $X_{ab} \in \{0, 1\}$ and weights $\hat{l}_{ab}$ find

$$\max \left[Z = \sum_a \sum_b \hat{l}_{ab} X_{ab} \right]$$

subject to the constraint

$$\sum_a X_{ab} = 1, \ b = 1, 2, \ldots, \max(n_1, n_2)$$

and

$$\sum_b X_{ab} = 1, \ a = 1, 2, \ldots, \max(n_1, n_2).$$

The solution of the assignment problem finds exactly one record from the survey with the larger size to every record of the survey with the smaller size. No record on either of surveys can be paired more than once. Effectively, all records of the smaller survey are in 1-to-1 pairings with the records of the larger survey and some of the records of the larger survey have no pairings. In record linkage terms this is 1-to-1 linkage. In other words, for a particular specification of the record linkage model and the corresponding estimates of $\pi, \mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu}), \nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$ obtained by the simulated annealing, we obtained an optimal (in a sense presented in the formulation of the assignment problem) 1-to-1 pairing of observations of two surveys of the population of interest.

5. The solution of the assignment problem can be used to determine the number of 1-to-1 pairings for each comparison pattern $\boldsymbol{\gamma}_p$, denoted $\tilde{f}_1, \ldots, \tilde{f}_p, \ldots, \tilde{f}_{2^K}$. At this stage these counts cannot be used to estimate the number of matches since their sum equals $\min(n_1, n_2)$, which is the upper bound for the number of matching record pairs.

6. Our goal is no-classification linkage and linkage free dual system estimation which can be achieved by combining the results of 1-to-1 pairing obtained by solving the assignment problem and the initial linkage free estimates of the number of matches in each comparison pattern in a composite estimator. The weights of the composite estimator are $\hat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$ and $1 - \hat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$. Provided that the linkage model was carefully specified for the set of available linkage variables, we would like to rely more on the results of 1-to-1 pairings, $\tilde{f}_p$, for the comparison patterns with the large $\hat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$. This is because we expect the results of 1-to-1 pairing to be slightly more accurate than the original linkage free estimates in such cases. Meanwhile, for the situations where $\hat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$ is small, we would like to rely more on the original linkage free estimate. This is because we expect that accepting the results of 1-to-1 pairings will result in a larger error than the original linkage free estimate. The composite estimator (hence c in the subscript) of the number of matches in the p^{th} pattern is then

$$\widetilde{m}_{c,p} = \widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)\tilde{f}_p + (1 - \widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p))\widehat{m}_p. \tag{51}$$

The total number of estimated matches is obtained by summing each pattern's composite estimates

$$\widetilde{m}_c = \sum_{p=1}^{2^K} \widetilde{m}_{c,p}. \tag{52}$$

Finally, the modified linkage free (m-based) dual system estimator of the population total τ is

$$\widetilde{\tau}_c = \frac{n_1 n_2}{\widetilde{m}_c}. \tag{53}$$

Observe, that unlike any regular instance of composite estimation in which two or more estimates obtained by different procedures are combined, the case of (51) involves $\widehat{m}_p = w\widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})$, $\tilde{f}_p$ and $\widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p) = \widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})/\pi(\boldsymbol{\gamma}_p; \widehat{\pi}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}})$. These are all based on the same estimates of $\pi, \mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$ and $\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$. One way of thinking about it is that the modified linkage free dual system estimator is an attempt to improve the simple linkage free estimator with very little additional information about the data available. Obtaining the 1-to-1 pairings, $\tilde{f}_p$, for each comparison pattern is a combination of the linkage free estimator with the knowledge of 1-to-1 match constraints. These $\tilde{f}_p$ are an improvement of the original linkage free estimator for some comparison patterns as they incorporate additional information and reduce both the variability and bias. However, this improvement is not without flaws, since at this stage we only achieved 1-to-1 pairings, that is *every* record in the smallest survey is uniquely linked to one of the records on the lager survey. Clearly, there will be $\tilde{f}_p$ where some or even the majority of pairs obtained by 1-to-1 pairing will be the true non-matches, but there is no way of determining which ones are the true matches.

In the classification-based approaches, after 1-to-1 pairing, all the pairs with corresponding comparison patterns for which the estimated 'likelihood' is above a certain threshold would be classified as links, all the pairs below the certain threshold classified as non-links, and pairs trapped between the thresholds would be clerically reviewed and classified. On the one hand, such a method satisfies the definition of the 1-to-1 linkage constraint as it is formulated in Section 2.2.4: every record on the first survey either links to one record on the second survey, or has no link. On the other hand, accepting and rejecting pairs as they are leads to errors in the established number of matches as one accepts some false links and rejects some true links. One of the aims of the no-classification approaches is to circumvent such errors by estimating the number of matches in each of the comparison pattern. Therefore, even a pattern with very few or no agreements on linkage variables has some contribution to the estimated number of matches, reflecting the fact that there is a non-zero chance that such a pattern contains a true match. Also, a pattern where nearly all linkage variables agree may have a contribution to the number of matches that is smaller than the number of 1-to-1 constrained record pairs in this pattern. Therefore, in linkage free dual system-estimation we would like to rely more on the 1-to-1 constrained estimates where it is safe to do so, but rely on the pure linkage free estimates where it is not safe to use the constrained estimates.

Now there is no additional independent source of information that would tell us for which comparison patterns it is safe to use the constrained estimator and for which it is not. In this case $\widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$ is perhaps the one reasonable source of such information. So in a way we have two versions of the

estimator and we try to combine these two versions in the best way possible given very limited information. It is clear that such an estimator, while using the fact that linkage must obey the 1-to-1 constraint, does not produce the outcome which satisfies the definition of this constraint. In general, the estimate of the number of matches in each pattern is not an integer, but a real number. Hence, it is inappropriate to say that a record has a corresponding match or has no match at all. The modified linkage free estimator achieves a reduction in both variance and bias employing the knowledge that there is a 1-to-1 constraint, while remaining purely an estimation based approach, that is not imposing this constraint on the outcome.

If, for some reason, one wants to enforce the 1-to-1 constraint for such a no-classification approach, a possible solution would be to use the modified no-classification estimates for each pattern and integerize them using the modified no-classification estimate of the total number of matches as a benchmark. Note, that for such an approach, there would be comparison patterns for which we would not be able to determine which specific pair should be treated as a link. For instance, if there are 20 record pairs that result in no agreements and no records forming these pairs are forming other pairs (this is what the solution of the assignment problem produces), and a single pair is estimated to be a link, there is no way of telling which one of these 20 is a linking pair.

The modified estimator can also be used similarly to the estimator (47) by being applied in individual linkage blocks, and block estimates are then summed to estimate the population total τ.

A useful consequence of the modified linkage free dual system estimator, is that it permits a linkage weight to be associated with each observation on the survey with the smallest size. This in turn allows to work with the linked data at the individual level, but with each observation having some positive weight reflecting its contribution to the total number of estimated links, rather than a 0 or 1 classification to non-link and link. The smallest survey is used again due to the fact that the number of links cannot exceed $\min(n_1, n_2)$ and the assignment problem produces solution such that the sum of $\widetilde{f}_p$ equals the size of the smallest survey. However, this does not have any impact on the usage of the results of the modified estimator since all that matters is the estimate of the number of matches.

Each comparison pattern has an associated positive linkage weight derived as

$$\widehat{\omega}_p = \frac{\widehat{m}_{c,p}}{\widetilde{f}_p}. \tag{54}$$

Note that this weight is always greater than 0 and also could be greater than 1. In the case $\widetilde{f}_p = 0$ for some p the weight either can be set to 0 or the contribution of $(1 - \widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p))\widehat{m}_p$ can be redistributed across the non-zero frequency patterns, proportional to $\widehat{\mathrm{r}}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$.

Let $S_{j=x}$ be the survey with the smallest size among S_j, $j \in 1, 2$. The size of this survey is $n_{j=x}$, observations of it are indexed $a = 1, \ldots, n_{j=x}$ and after solving the assignment problem we have $\sum_{p=1}^{2^K} \widetilde{f}_p = n_{j=x}$. For every a we can determine an individual linkage weight as

$$\widehat{\omega}_{p,a} = \begin{cases} \widehat{\omega}_p \text{ if } a \text{ makes a pair } (a, b) \text{ such as } \boldsymbol{\gamma}(a, b) = \boldsymbol{\gamma}_p \\ 0, \text{ otherwise.} \end{cases} \tag{55}$$

When population size estimation is the goal, these weights can be used to produce the linkage free dual system estimates for the subdomains of the population $\mathcal{P}$ at which the modified linkage free estimator is initially applied. This, at least in principle, may allow the heterogeneity of the responses in different domains of the population to be taken into account, but also prompts an interesting question which will be discussed in Chapter 9. Let d be a domain of interest. Define the indicator function

$$I_{j,a,d} = \begin{cases} 1 \text{ if a record } a \text{ on the survey } S_j \text{ belongs to domain } d \\ 0 \text{ if a record } a \text{ on the survey } S_j \text{ does not belong to domain } d. \end{cases} \tag{56}$$

The estimate of the number of matches in domain d is then

$$\widehat{m}_{c,d} = \sum_{a=1}^{n_{j=x}} I_{j=x,a,d} \widehat{\omega}_{p,a}. \tag{57}$$

The linkage free dual system estimator for domain d can be obtained using estimator (57) and observed survey counts for the domain of interest $n_{1,d} = \sum_{a=1}^{n_1} I_{1,a,d}$, $n_{2,d} = \sum_{b=1}^{n_2} I_{2,b,d}$, as

$$\widetilde{\tau}_{c,d} = \frac{n_{1,d} n_{2,d}}{\widehat{m}_{c,d}}. \tag{58}$$

In this thesis we are assessing the performance of estimator (53), but assessing the performance of (58) is left for future research.

A worked-out example of how the application of the modified linkage free dual system estimator looks like is presented later in this work in Section 7.4.

Note that in the classification-based context it is in principle possible to estimate the linkage model parameters, apply the 1-to-1 constraint and then specify only a single threshold. In this case $\min(n_1, n_2)$ pairs would be classified into links and non-links only: all the pairs with corresponding comparison patterns for which the estimated 'likelihood' is above the threshold would be classified as links, while the rest of the pairs would be classified as non-links. There would be no clerical resolutions in such a scenario. Nevertheless, such an approach can hardly be more practical or outperform the modified linkage free dual system estimator in general. First, the classification with a single threshold would require careful model specification, since the incorrect ordering of the comparison patterns by the 'likelihood' of containing a link would have severe consequences for the outcome of classification. Hence, one would need to carry out the steps 1 to 5 of the modified linkage free dual system estimation. This would be followed by setting a threshold. It seems implausible that in general it is possible to select a single threshold with some near symmetric distribution of errors, so that the number of false positive errors would not substantially differ from the number of false negative errors. Recall, that this difference is the factor that drives the overall error in classification-based approaches. If one can somehow find such an 'optimal' threshold, it means that one is capable of obtaining a good estimate of the errors in each pattern after the 1-to-1 constraint was applied. Finding such estimates, however, is similar to modified linkage free estimation, with the difference that the latter is estimating the number of matches (instead of the number of errors) in each pattern after the 1-to-1 constraint was applied. In

a sense, such a hypothetical single threshold approach at best would perform as well as the modified no-classification method. In this thesis we are not considering how to develop a well-performing single threshold linkage approach. This simulation study demonstrates, that in fact the performance of the modified linkage free estimation is often as good or better than the classification-based approach with clerical resolution under the acceptance rate of of 10^{-3} for false negatives and 10^{-6} for false positives.

Since we are discussing how additional information can be used to improve the performance of the linkage free dual system estimator, there is a natural question of whether information from clerical revisions can be incorporated into the no-classification approaches. While this aspect is not developed and assessed in this thesis, clerical information can be fed into no-classification estimation in several ways. It is easiest to insert such information into the modified estimator. If a certain comparison pattern $\boldsymbol{\gamma}_p$ is fully clerically reviewed and the number of links is determined for this pattern, the estimate $\widetilde{m}_{c,p}$ can be replaced by the clerically established number of links. More generally, irrespective of whether clerical resolution of an entire pattern or just some individual pairs was carried out, one can shift the solution of the assignment problem to reflect the clerical reviews. So, if clerical review established that a pair (a, b) should be a link, but the solution of the assignment problem resulted in $X_{ab} = 0$, this solution can be forced to be 1. This requires revising the solutions for the remaining pairs. Thus, this approach would require some workaround to deal with such constraints of the assignment problem.

4.3 Quality measures

When assessing theoretical performance of record linkage approaches the following metrics are often used: precision and recall; false positive rate; accuracy; F-measure. In the situation where no-classification linkage is used to enable the linkage free dual system estimation, it is more convenient to use the estimates of m and τ straight away and compare them to the true values of $\overline{m}$ and the population size. Of course, such quality measures are only available in simulation studies, but such studies are indispensable in assessing properties of methods. In Chapter 7, where the performance of methods developed in this thesis is assessed, the only quality measures are via $\overline{m}$ and τ as those also allow easy comparison to the dual system estimation with perfect linkage and dual system estimation under classification-based linkage with clerical resolutions.

4.4 Heuristics in model specification and goodness-of-fit

The formal record linkage model specification and goodness-of-fit assessment approaches are beyond the scope of this thesis. Nevertheless, when linking two population surveys similar to a coverage survey and a census within specified geographies, a selection of linkage variables can be used to determine a suitable model specification, at least for the component corresponding to non-matches. This is due to between-variable dependencies being a consequence of the population structure, which is known in such applications.

Once a certain model is chosen and its parameters are estimated, an obvious, though informal, indicator of goodness-of-fit is the value of a modified chi-squared statistic. The results of simulations, not presented in this thesis, suggest that whenever an identifiable model is suitably parametrized or

is more complex than required, the modified chi-squared statistic is small. Whenever a parameterization misses some between linkage variable dependencies, the modified chi-square statistic becomes noticeably large.

5 Checking identifiability of certain linkage models with four linkage variables

In this chapter we will use several methods presented in Section 2.5 to check if certain linkage models with four linkage variables are identifiable or not. Our choice of models with four variables is motivated by the good balance of generality and computational tractability these models often offer. In addition, the number of useful linkage variables when linking surveys of human populations may be quite limited in practice. There are two aspects of usefulness of linkage variables. First, a linkage variable is useful if it can take sufficiently many values to satisfy the properties studied in Section 3.2. Second, a linkage variable is useful if upon addition to the model it either does not result in between-variables dependence of the comparison outcomes, or this dependence can be taken into account by an identifiable model specification. Linkage variables such as address, surname, first name and date of birth are among those having the largest number of possible values. However, the combination of these variables results in dependencies between comparisons on the address and surname variables, as discussed in Section 2.2.5. It will be shown in this chapter that this model is identifiable. While adding variables such as sex, ethnicity, marital status, relationship, tenure and accommodation type into a linkage model may seem tempting, each of these variables has just a few levels. Furthermore, many of these variables, when used with the address or surname variables, will result in associated comparison outcomes. The identifiability of the corresponding models is not guaranteed. Finally, using more linkage variables will result in some of the comparison patterns $\boldsymbol{\gamma}_p$ having sparse frequency as the number of patterns is 2^K. This sparseness may need additional machinery to deal with, which is beyond the scope of this thesis. Note, that it is sometimes argued that no more than $6 - 10$ linkage variables are needed (Winkler, 2006).

5.1 Linkage models and identifiability

The meaning and justification of a generic mixture-like model

$$\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu}) + (1 - \pi)\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$$

was presented in Section 3.2. In the above model $\mu(\boldsymbol{\gamma}_p; \boldsymbol{\mu})$ and $\nu(\boldsymbol{\gamma}_p; \boldsymbol{\nu})$ are appropriate factorizations of the expected ratios of the matches / non-matches in a pattern p to the number of matches / non-matches for a given model specifications. This is similar to the factorization of the joint probability distribution when some of the variables are independent or conditionally independent. For instance, one may be interested in the model where the between-variables independence of the comparison outcomes among the matches is assumed while allowing for dependence between the comparison outcomes of the

variables v_1 and v_2 in the set of non-matches. Recall that we write this model in a compact form as

$$\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$$

and its parameterization is written as

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4, \\
\pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4, \\
\pi(\boldsymbol{\gamma}_3; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_1(1-\nu_{2|1(1)})\nu_3\nu_4, \\
&\dots \\
\pi(\boldsymbol{\gamma}_{16}; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_4).
\end{aligned}$$

Note that this is the polynomial map, which is the special case of the rational map $\phi : \boldsymbol{\Theta} \to \mathbb{R}^{2^K}$, where $\boldsymbol{\Theta} = \{\pi, \boldsymbol{\mu}, \boldsymbol{\nu}\}$ and $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are collections of parameters, reflecting between-variables dependencies if needed, conditional on the set of matches and set of non-matches. Hence, the results presented in Section 2.5 can be used, irrespective of whether we dealing with regular mixtures of mixture-like models. For some model specifications instead of just polynomial functions the parameterization may actually involve explicit rational functions in the form f/g, where f and g are polynomials.

When using simulated annealing to find $\widehat{\pi}, \mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})$ and $\nu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\nu}})$ such that the χ^2 based on the estimated values $\widehat{\pi}\mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}}) + (1-\widehat{\pi})\nu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\nu}})$ and the given observed data f_p/w is minimized, a generically identifiable model means that we are obtaining unique $\widehat{\pi}, \mu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\mu}})$ and $\nu(\boldsymbol{\gamma}_p; \widehat{\boldsymbol{\nu}})$ that achieve such a minimization. In the case of local identifiability, we either obtain finitely many solution, or we obtain a single solution that satisfies certain knowledge about the model, for instance, that π is small, or that the values of all parameters lie in the interval $(0, 1)$.

In many practical applications, including many record linkage applications, the model of independence conditional on the mixture component is used. In the record linkage case the components are the set of matches and non-matches. This model is generically identifiable up to label switching if the conditions presented in Section 2.5 are satisfied. Recall, however, that when dealing with human populations, such a specification is not always suitable. There may be between-variables dependence of the comparison outcomes in the set of non-matches $\mathcal{U}$ if either entire households are sampled or there is a high likelihood that multiple members of the same household are sampled. When comparing the values of linkage variables for any two individuals in the same household, the address will almost certainly agree while variables like surname, ethnicity, country of birth, address one year ago will have high likelihood of agreeing. On the other hand, if the probability of making a typographical error in a value of one of the linkage variables is associated with the probability of making an error in a value of another variable, there is between-variables dependence in the set of matches $\mathcal{M}$.

The between-variables dependencies in the set of matches may be reduced by the approximate comparison of the values of linkage variables. For example, consider a situation where an interviewer fails to record less familiar, long or hard-to-spell names and surnames correctly, but those recorded values are close to the true ones. Then the events of making a mistake recording the values of surname

and name are associated. Nevertheless, since the values recorded are close to the true ones, and provided that the values of the same variables on the another survey are also spelt close to the correct spelling or correctly, the approximate comparison will result in high similarity scores. Hence, many cases of what would be a simultaneous disagreement on both variables in the case of exact comparison, will often be either a simultaneous agreement of both of the variables, or an agreement on at least one of the variables. The between-variables associations in the set of non-matches is harder to mitigate for a given set of linkage variables as it is related to the way the values of population attributes are structured. Also, if the between-variables associations of comparisons are present in both the set of matches and set of non-matches, different variables will be associated in different sets. For instance, in the set of matches the association between the surname and name variables may be present while the association between the address and surname variables may be present in the set of non-matches. Therefore, different model specification in the set of matches compared to the set of non-matches will be emphasised in the cases considered below. This differs from the examples of establishing identifiability for mixtures or similar models known to us. In the examples available in the literature, identical model specifications are used in all of the model components. For example, in Allman et al. (2015), identifiability of discrete Bayesian networks with hidden variables is discussed and tensor methods are used to check identifiability. While some of these models may seem similar to what is considered below, all of the models in the cited paper have the same model specification on each level of the hidden variable, which is equivalent to using the same model in both the $\mathcal{M}$ and $\mathcal{U}$ sets.

5.2 Checking identifiability for a selection of models

In this section we check the identifiability for a selection of linkage models with four linkage variables. This selection is driven mainly by the potential practical usefulness of the models and in part by academic interest. For every model considered we give a possible set of variables that may lead to this model in real applications. We then check that the number of parameters to estimate is not larger than the number of observables 2^K. If the number of parameters does not exceed the number of observables, then local identifiability is checked using the Wolfram Mathematica software (Wolfram Research, Inc., 2022). If the model is locally identifiable, then generic identifiability is checked using either the tensor or Gröbner basis-based methods. When employing the Gröbner basis-based methods, we either use the Wolfram Mathematica, or Maple (Maple, 2021), or Singular (Decker et al., 2022) or some combination of the software packages. Singular, being specifically designed for polynomial computations, has a wider choice of functionality relevant for Gröbner basis computation and analysis. Often, however, once a basis is computed it is easier to carry out the remaining analysis in Mathematica. There is a sample of code for various identifiability checking approaches and different software packages in Appendix C.

5.2.1 Model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

The first model we are looking at is the model of the conditional between-variables independence of comparisons given the match status. This is the simplest and one of the most frequently considered models (with varying numbers of linkage variables). We already know that this model is identifiable, but it allows a gentle start of the discussion. In real applications, it can be suitable when linkage

variables are, say, surname, first name, date of birth and sex, provided errors in recording the true values of attributes are not correlated or this correlation is very small after applying approximate (fuzzy) comparison functions on the values of variables. It is also reasonable to expect that for non-matching pairs, observing an agreement on one of the variables is not associated with observing an agreement on any other variables. There may in principle be some association between agreement on the surname variable and the name variable, provided individuals with certain cultural backgrounds tend to select from a small pool of names for the first male or female child in a family. However, association is unlikely to be very strong at the population level; see for instance the distribution of baby names (ONS, 2021e). The compact form of this model is

$$\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$$

and the corresponding parameterization is

$$\begin{aligned}
&\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_2\nu_3\nu_4,\\
&\pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_2\nu_3\nu_4,\\
&\dots\\
&\pi(\boldsymbol{\gamma}_{16}; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_2)(1-\nu_3)(1-\nu_4).
\end{aligned}$$

This model has 9 parameters and 16 observables, so we can check for its local identifiability. We work out the Jacobian

$$\begin{aligned}
J(\pi) &= \begin{pmatrix} \frac{\partial\pi_1}{\partial\pi} & \dots & \frac{\partial\pi_1}{\partial\nu_4} \\ \vdots & \ddots & \vdots \\ \frac{\partial\pi_{16}}{\partial\pi} & \dots & \frac{\partial\pi_{16}}{\partial\nu_4} \end{pmatrix}\\
&= \begin{pmatrix} \mu_1\mu_2\mu_3\mu_4 - \nu_1\nu_2\nu_3\nu_4 & \dots & (1-\pi)\nu_1\nu_2\nu_3 \\ \vdots & \ddots & \vdots \\ (1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) & & \\ -(1-\nu_1)(1-\nu_2)(1-\nu_3)(1-\nu_4) & \dots & -(1-\pi)(1-\nu_1)(1-\nu_2)(1-\nu_3) \end{pmatrix}.
\end{aligned}$$

The rank of this Jacobian matrix, computed using the symbolic computation software, is 9. That is, the rank is the same as the number of parameters. Therefore, using the results presented in Section 2.5 the model is locally identifiable. Note the Jacobian is too big to be printed on a page, but it is easy to replicate it using the code provided in Appendix C. We will omit printing even partial results for subsequent examples since the process is exactly the same for all the models when checking local identifiability.

To check that this model is generically identifiable, we can use the established result (19). This is the conditional between-variables independence model given the match status, that is $g = 2$ and we have four linkage variables $K = 4$. We check $K = 4 \geq 2\lceil\log_2 2\rceil + 1 = 3$, which means that the model is generically identifiable up to label switching.

This particular model also allows generic identifiability to be assessed with the Gröbner basis

approach as presented in Section 2.5.6. This approach works for individual parameters. In the context of this thesis the key parameter of interest is π and identifiability of this parameters can be checked. We are dealing with the following ideal

$$I(\phi(\boldsymbol{\Theta})) = \langle \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_2\nu_3\nu_4 - p_1, \ldots, \\ \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_2)(1-\nu_3)(1-\nu_4) - p_{16}\rangle$$

while the ideal corresponding to the augmented map is

$$I(\tilde{\phi}(\boldsymbol{\Theta})) = \langle \pi - q, \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_2\nu_3\nu_4 - p_1, \ldots, \\ \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_2)(1-\nu_3)(1-\nu_4) - p_{16}\rangle.$$

Proposition 16.1.9 in Sullivant (2018) can be applied. We find a reduced Gröbner basis G for the ideal $I(\tilde{\phi}(\boldsymbol{\Theta}))$ with respect to lexicographic or any other elimination order with $\pi \succ \mu_1 \succ \ldots \mu_4 \succ \cdots \succ \nu_1 \succ \ldots \nu_4 \succ q \succ p_1 \succ \cdots \succ p_{16}$. Note, that the relative order of $\pi \succ \mu_1 \succ \ldots \mu_4 \succ \cdots \succ \nu_1 \succ \cdots \succ \nu_4$ is not important, but all these indeterminates have to come before q, as well as the relative order of $p_1 \succ \cdots \succ p_{16}$ is not important after q. Once G is obtained, we either find the elimination ideal $G_q = G \cap \mathbb{R}[q, p_1, \ldots, p_{16}]$ or search for a polynomial of lowest non-zero degree q in the form $g(q, x_1, \ldots, x_n) = \sum_{i=0}^{d} g_i(x_1, \ldots, x_n)q^i \in I(\tilde{\phi}(\boldsymbol{\Theta}))$. In this case, such a polynomial exists, though we do not present it here as it has a very large number of monomials. This polynomial has degree $d = 2$. So there are 2 choices for the parameters. These choices reflect the label switching behaviour of the model. Hence, the model is generically identifiable up to label switching.

Unfortunately, for the majority of record linkage models, the computational complexity is immense and we manage to get results on generic identifiability in real time only for this conditional between-variables independence model. Even this example requires computation of the basis using product order with graded reverse lexicographic orders and then using the Gröbner walk algorithm (Collart et al., 1997) to obtain the basis in lexicographic order; see the related Maple example in Appendix C. However, the situation with rational identifiability appears to be more promising and several examples below allow us to establish rational identifiability.

5.2.2 Model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

The next model is the conditional between-variables independence given the set of matches and association between two variables in the set of non-matches. Clearly, the association can be between any pair without changing the general structure of the model, our presentation uses an association between v_1 and v_2. This is an important model for real applications and it was also analysed in the model justification Section 3.2 of this thesis. It can correspond to the case where there is no dependence in errors for true matches, or this dependence is minimal after use of approximate comparison, as mentioned in Section 2.2.1. Given that entire households are sampled, linkage variables that can lead to such model specifications can be the following: full standardized address, surname, first name and date of birth. As discussed in the linkage experiment Section 2.2.5, we expect the comparisons on the surname variable to agree more often within an agreeing address compared to a non-agreeing address,

which results in lack of independence for comparison outcomes on these two variables in the set on non-matches $\mathcal{U}$.

The model's parameterization is

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4, \\
\pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4, \\
\pi(\boldsymbol{\gamma}_3; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_1(1-\nu_{2|1(1)})\nu_3\nu_4, \\
&\dots \\
\pi(\boldsymbol{\gamma}_6; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)\mu_3\mu_4 + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})\nu_3\nu_4, \\
&\dots \\
\pi(\boldsymbol{\gamma}_{16}; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_4).
\end{aligned}
$$

There are 10 parameters and 16 observables in this model. Checking local identifiability shows that the corresponding Jacobian has rank 10. Since the rank equals the number of parameters, this model is locally identifiable.

We use the tensor methods to assess generic identifiability of this model. Note, that this model is similar to one of those considered in Allman et al. (2015) in the context of discrete Bayesian networks with hidden variables. The difference between the specification of our model of interest and one analysed in the cited paper is that in our case there is between-variables association only in one of the two components of the model (in the set of non-matches), rather than in both components. In fact, it is easier to start with the model where association between linkage variables v_1 and v_2 is present in both $\mathcal{M}$ and $\mathcal{U}$:

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1\mu_{2|1(1)}\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4, \\
\pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_{2|1(0)}\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4, \\
&\dots \\
\pi(\boldsymbol{\gamma}_{16}; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_{2|1(0)})(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_4).
\end{aligned}
$$

We can marginalize variable v_2 in the above model, which yields a model with three linkage variables:

$$
\begin{aligned}
\pi_1^* &= \pi\mu_1\mu_3\mu_4 + (1-\pi)\nu_1\nu_3\nu_4, \\
\pi_2^* &= \pi(1-\mu_1)\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_3\nu_4, \\
&\dots \\
\pi_8^* &= \pi(1-\mu_1)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_3)(1-\nu_4).
\end{aligned}
$$

This model is the equivalent of the model considered in Section 2.5.4 and all 7 parameters are identifiable up to label switching. Now for $\mu_{2|1(1)}, \nu_{2|1(1)}$ and $(1-\mu_{2|1(1)}), (1-\nu_{2|1(1)})$ marginalize over

variables v_3, v_4 to obtain the system of equations

$$x_1 = \pi\mu_1\mu_{2|1(1)} + (1-\pi)\nu_1\nu_{2|1(1)}$$
$$x_2 = \pi\mu_1(1-\mu_{2|1(1)}) + (1-\pi)\nu_1(1-\nu_{2|1(1)}),$$

where π, μ_1, ν_1 are known up to label switching and x_1, x_2 are the sums of observables. Both $\mu_{2|1(1)}$ and $\nu_{2|1(1)}$ can be recovered from the above system up to label switching. In the same way we can obtain the system of equations

$$y_1 = \pi(1-\mu_1)\mu_{2|1(0)} + (1-\pi)(1-\nu_1)\nu_{2|1(0)}$$
$$y_2 = \pi(1-\mu_1)(1-\mu_{2|1(0)}) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(1)}),$$

and recover the parameters $\mu_{2|1(0)}$ and $\nu_{2|1(0)}$ up to label switching.

The above argument can be repeated for the model of interest by replacing $\mu_{2|1(1)}$ and $\mu_{2|1(0)}$ with μ_2 and all the parameters of interest can be recovered. However, this time, there is no label switching. Recall that the label switching means that $\pi\mu_1\mu_2\cdots + (1-\pi)\nu_1\nu_2\cdots = \pi^*\mu_1^*\mu_2^*\cdots + (1-\pi^*)\nu_1^*\nu_2^*\ldots$ for $\pi^* = (1-\pi), \mu_1^* = \nu_1, \mu_2^* = \nu_2$, etc. Now in the model of interest the parameters related to the set of matches are all independent, so that $\mu_{2|1(1)} = \mu_{2|1(0)} = \mu_2$, but $\nu_{2|1(1)}$ and $\nu_{2|1(0)}$ are distinct. Label switching would mean that there is some $\nu^*_{2|1(1)} = \mu_{2|1(1)}$ and $\nu^*_{2|1(0)} = \mu_{2|1(0)}$, but this contradicts the previous statement. Hence, the model parameters are generically identifiable without label switching.

We can also check that parameters are identifiable without label switching by checking rational identifiability. Using the language of computational algebra, the vector of indeterminates is

$$\boldsymbol{\theta} = (\pi, \mu_1, \ldots, \mu_4, \nu_1, \nu_{2|1(1)}, \nu_{2|1(0)}, \ldots, \nu_4)^T$$

and the vector of symbolic parameters is

$$\boldsymbol{t} = (\pi^c, \mu_1^c, \ldots, \mu_4^c, \nu_1^c, \nu_{2|1(1)}^c, \nu_{2|1(0)}^c, \ldots, \nu_4^c)^T.$$

The ideal $I_c \subseteq \mathbb{R}(\pi^c, \mu_1^c, \ldots, \mu_4^c, \nu_1^c, \nu_{2|1(1)}^c, \nu_{2|1(0)}^c, \ldots, \nu_4^c)[\pi, \mu_1, \ldots, \mu_4, \nu_1, \nu_{2|1(1)}, \nu_{2|1(0)}, \ldots, \nu_4]$ is

$$\begin{aligned} I_c = \langle & \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4 - (\pi^c\mu_1^c\mu_2^c\mu_3^c\mu_4^c + (1-\pi^c)\nu_1^c\nu_{2|1(1)}^c\nu_3^c\nu_4^c), \\ & \ldots, \\ & \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_4) \\ & - (\pi^c(1-\mu_1^c)(1-\mu_2^c)(1-\mu_3^c)(1-\mu_4^c) + (1-\pi^c)(1-\nu_1^c)(1-\nu_{2|1(0)}^c)(1-\nu_3^c)(1-\nu_4^c))\rangle. \end{aligned}$$

Our guess is that parameters are identifiable without label switching, so that the 'simplest' functions of the parameters of interest are $f(\pi) = \pi, f(\mu_1) = \mu_1, f(\nu_{2|1(1)}) = \nu_{2|1(1)}$, and so on. To check it we solve (by computing a Gröbner basis) the corresponding ideal membership problems $\pi - \pi^c \in I_c, \mu_1 - \mu_1^c \in I_c, \ldots, \nu_{2|1(1)} - \nu_{2|1(1)}^c \in I_c, \ldots$. Indeed, all such polynomials belong to the ideal I_c. Hence, all individual parameters are rationally identifiable without label switching. Note that $\pi - \pi^c \in I_c$ implies $\pi(1-\pi) - \pi^c(1-\pi^c) \in I_c$. While in the case of label switching the 'simplest' polynomial in π

that belongs to the above ideal would be $\pi(1-\pi) - \pi^c(1-\pi^c) \in I_c$. If this was the case, we would have $\pi - \pi^c \notin I_c$.

5.2.3 Model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{3,4})$

The next model is the one of between-variables independence within the set of matches and association for two non-overlapping pairs of linkage variables in the set of non-matches: v_1 and v_2 are associated as well as v_3 and v_4. Perhaps, this is not a model that would be used very often in practice since not so many meaningful combinations of linkage variables would require such a model specification. A slightly artificial example, in the situation where entire households are sampled, could involve the following set of variables: full standardized address, surname, quinary age, the highest degree attained (or years in education). In this case the agreements on the surname variable would depend on the agreement on the address variable, while the agreement on the highest degree attained would depend to some extent on the agreement on quinary age. This model can be written as

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_{4|3(1)}, \\
\pi(\boldsymbol{\gamma}_2; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_{4|3(1)}, \\
&\dots \\
\pi(\boldsymbol{\gamma}_{16}; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_{4|3(0)}).
\end{aligned}
$$

There are 11 parameters and 16 observables in this case. A check for local identifiability shows that the rank of the corresponding Jacobian equals the number of parameters. Therefore, this model is locally identifiable. Interestingly, the related model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_{1,2}, \gamma_{3,4}) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{3,4})$ with the same sort of between-variables associations in the both sets has 13 parameters, but the rank of the corresponding Jacobian is only 11 and the model is nonidentifiable, this agrees with the results presented in (Allman et al., 2015).

Generic identifiability of this model cannot be checked using the tensor methods. It is impossible to marginalize the given model to obtain a model $\text{Mixt}^2(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$. We resort to checking rational identifiability. Given the ideal

$$
\begin{aligned}
I_c = \langle &\pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_{4|3(1)} - (\pi^c\mu_1^c\mu_2^c\mu_3^c\mu_4^c + (1-\pi^c)\nu_1^c\nu_{2|1(1)}^c\nu_3^c\nu_{4|3(1)}^c), \\
&\dots, \\
&\pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_{4|3(0)}) \\
&- (\pi^c(1-\mu_1^c)(1-\mu_2^c)(1-\mu_3^c)(1-\mu_4^c) + (1-\pi^c)(1-\nu_1^c)(1-\nu_{2|1(0)}^c)(1-\nu_3^c)(1-\nu_{4|3(0)}^c))\rangle
\end{aligned}
$$

in $\mathbb{R}(\pi^c, \mu_1^c, \dots, \mu_4^c, \nu_1^c, \nu_{2|1(1)}^c, \nu_{2|1(0)}^c, \nu_3, \nu_{4|3(1)}^c, \nu_{4|3(0)}^c)[\pi, \mu_1, \dots, \mu_4, \nu_1, \nu_{2|1(1)}, \nu_{2|1(0)}, \nu_3, \nu_{4|3(1)}, \nu_{4|3(0)}]$, we are interested in checking whether polynomials like $\pi - \pi^c \in I_c$ or $\pi(1-\pi) - \pi^c(1-\pi^c) \in I_c$. After finding a Gröbner basis G of I_c, we see that neither of the polynomials of interest are in I_c. In fact, expressing π in terms of other indeterminates gives a complex equation of high degree. So in principle, this model, or at least some parameters of interest, may be identifiable. Since this model is unlikely in practice, we left solution of this equation for future research.

5.2.4 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3},\gamma_4)$

We now consider a model with between-variables independent comparison outcomes in the set of matches and the three-way interaction between variables v_1, v_2 and v_3 in the set of non-matches. Given that households are sampled, this model may be needed if one is dealing with the following linkage variables: full standardized address, surname, country of birth (or ethnicity, or relationship to the head of household) and date of birth. Basically, this model can be considered whenever agreement on address leads to likely agreement on another two linkage variables, while agreement on the fourths variable is independent of all the rest of the agreements. It is important to observe that agreements on v_2 and v_3 given the agreement on address will result in the three-way association since within a given address the population values for the members of this address will tend to be the same for variable v_2 and the same for variable v_3. For example, within an address, all members of the address tend to share the same surname and have the same country of birth. A more detailed explanation is provided in Section 7.2. This model is parameterized as

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_{1(1),2(1),3(1)}\nu_4,\\
\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)\nu_{1(0),2(1),3(1)}\nu_4,\\
&\dots\\
\pi(\boldsymbol{\gamma}_{16};\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4)\\
&\quad + (1-\pi)(1-\nu_{1(1),2(1),3(1)} - \nu_{1(0),2(1),3(1)} - \nu_{1(1),2(0),3(1)} - \nu_{1(1),2(1),3(0)}\\
&\quad - \nu_{1(0),2(0),3(1)} - \nu_{1(0),2(1),3(0)} - \nu_{1(1),2(0),3(0)})(1-\nu_4),
\end{aligned}
$$

where $\nu_{1(1),2(1),3(1)} = \overline{u}_{1(1),2(1),3(1)}/\overline{u}, \dots, \nu_{1(1),2(0),3(0)} = \overline{u}_{1(1),2(0),3(0)}/\overline{u}$.

This model has 13 parameters and 16 observables. The rank of the corresponding Jacobian is 12, which means that the model is not identifiable.

Alternatively, this model can be parameterized as

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_{1|2(1),3(1)}\nu_{2|3(1)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_{1|2(1),3(1)})\nu_{2|3(1)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_3;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_{1|2(0),3(1)}(1-\nu_{2|3(1)})\nu_3\nu_4,\\
&\dots\\
\pi(\boldsymbol{\gamma}_{16};\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4)\\
&\quad + (1-\pi)(1-\nu_{1|2(0),3(0)})(1-\nu_{2|3(0)})(1-\nu_3)(1-\nu_4),
\end{aligned}
$$

Again, model has 13 parameters, but the rank of the corresponding Jacobian is 12, which means that the model is not identifiable.

The fact that this model is not identifiable is important. When entire households are sampled from a population of humans, the majority of variables will result in associated comparisons in the set of non-matches. Therefore, a great deal of care must be taken in specifying linkage models. Either the address variable is included and an identifiable model is specified, such as $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) +$

$(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$, or the combination of variables that lead to nonidentifiable models must be avoided. In both cases, a restriction on the use of the available linkage variables is enforced. That is why it seems unlikely to be able to use many linkage variables in the no-classification linkage and linkage free dual system estimation. Note, that many model specifications found in the literature on record linkage, say Jaro (1989) or Larsen & Rubin (2001), have models that are most likely nonidentifiable as the result of combining multiple associated variables.

5.2.5 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_4)$

The next model for identifiability investigation has independent comparison outcomes between linkage variables in the set of matches, while in the set of non-matches there is association between v_1 and v_2 as well as between v_1 and v_3. There are several practical situations where such a model can be considered. This model can be tried as a simpler alternative for the model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)+(1-\pi)\nu_p(\gamma_{1,2,3},\gamma_4)$. That is, instead of trying to model a three-way dependence, one chooses to take into account the association between the full standardized address and surname as well as the association between the address and country of birth (or ethnicity, relationship to the head of household, etc.). Another situation is when the variables are full standardized address, surname, first name and date of birth of the $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$ model are used, but this time association between the address and surname as well as association between the surname and first name variables are taken into account. To be precise, the corresponding model in this case is $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{2,3},\gamma_4)$, but this model differs from the previous one by the permutation of indices of linkage variables only (that is, the same number and type of associations, but between a different combination of variables) as the model in the title of this section so these are structurally exactly the same models and it is sufficient to establishing identifiability for just one variant of the model. Another structurally the same model is $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,3},\gamma_{2,3},\gamma_4)$, which can be regarded as another simplification of the three-way association model. The parameterization is

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_{3|1(1)}\nu_4,\\
\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_{3|1(0)}\nu_4,\\
&\dots\\
\pi(\boldsymbol{\gamma}_{16};\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_{3|1(0)})(1-\nu_4).
\end{aligned}
$$

There are 11 parameters and the rank of the corresponding Jacobian matrix is also 11. Therefore, the model is locally identifiable.

It is possible to check the rational identifiability of this model in a similar way to several examples examined above. Computing the corresponding Gröbner basis we find that for every parameter of this model the polynomials in the form $\pi - \pi^c$ are in the ideal I_c. Therefore, all parameters are rationally identifiable without label switching.

5.2.6 Model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$

This model can be seen as another simplification of a more complex specification $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$. In terms of the regular probability models a model with all pairwise but no three-way associations is called the homogeneous association model (Agresti, 2002, chap. 8). The mixture-like model considered in this section takes into account only pairwise association between comparison outcomes for three linkage variables in the set of non-matches while the fourth variable is independent of the rest of the variables.

Parameterizing this model is slightly more cumbersome than models considered so far. It is easier to start working with a regular probability model. We can later shift to a mixture-like model since algebraic manipulation with components of this model are equivalent to the regular models. Consider a binary outcome model with three variables $\text{pr}(X_1 = 1, X_2 = 1, X_3 = 1), \text{pr}(X_1 = 0, X_2 = 1, X_3 = 1), \ldots, \text{pr}(X_1 = 0, X_2 = 0, X_3 = 0)$. If the pairwise dependence between all three variables exists, then the following relationship holds (Agresti, 2002, chap. 8):

$$\begin{aligned}&\text{pr}(X_1 = 1, X_2 = 1, X_3 = 1)\text{pr}(X_1 = 0, X_2 = 0, X_3 = 1)\text{pr}(X_1 = 0, X_2 = 1, X_3 = 0)\text{pr}(X_1 = 1, X_2 = 0, X_3 = 0)\\ &= \text{pr}(X_1 = 0, X_2 = 1, X_3 = 1)\text{pr}(X_1 = 1, X_2 = 0, X_3 = 1)\text{pr}(X_1 = 1, X_2 = 1, X_3 = 0)\text{pr}(X_1 = 0, X_2 = 0, X_3 = 0).\end{aligned}$$

This relationship can also be found by computing a Gröbner basis for this homogenous association model expressed in the exponential product form, which can be obtained by taking the exponent of the standard log-linear formulation of the model. Now we can combine the relationship implied by the homogeneous association model and the expression $\text{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \text{pr}(X_1 = x_1 \mid X_2 = x_2, X_3 = x_3)\text{pr}(X_2 = x_2 \mid X_3 = x_3)\text{pr}(X_3 = x_3)$ for the joint probability in order to write one of $\text{pr}(X_1 = x_1 \mid X_2 = x_2, X_3 = x_3)$ as a function of the remaining probabilities and thus reduce the number of parameters from 7 to 6. Denoting $\nu^*_{1|2(1),3(1)} = \text{pr}(\gamma_1 = 1 \mid \mathcal{U}, \gamma_2 = 1, \gamma_3 = 1), \nu^*_{1|2(0),3(1)} = \text{pr}(\gamma_1 = 1 \mid \mathcal{U}, \gamma_2 = 0, \gamma_3 = 1), \nu^*_{1|2(1),3(0)} = \text{pr}(\gamma_1 = 1 \mid \mathcal{U}, \gamma_2 = 1, \gamma_3 = 0)$ and $\nu^*_{1|2(0),3(0)} = \text{pr}(\gamma_1 = 1 \mid \mathcal{U}, \gamma_2 = 0, \gamma_3 = 0)$, it is possible to express, for instance, $\nu^*_{1|2(1),3(1)}$ in terms of the remaining three probabilities. Using the properties of the components of a mixture-like model, the probabilities can be replaced by the ratios of expectations in the resulting expression when dealing with mixture-like record linkage models. Hence, we can then write the corresponding relationship for the expectations of a mixture-like model as

$$\nu_{1|2(1),3(1)} = \frac{\nu_{1|2(1),3(0)}\nu_{1|2(0),3(1)} - \nu_{1|2(1),3(0)}\nu_{1|2(0),3(1)}\nu_{1|2(0),3(0)}}{\nu_{1|2(1),3(0)}\nu_{1|2(0),3(1)} + \nu_{1|2(0),3(0)} - \nu_{1|2(1),3(0)}\nu_{1|2(0),3(0)} - \nu_{1|2(0),3(1)}\nu_{1|2(0),3(0)}}. \tag{59}$$

Therefore, the model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$ can be written

as

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_3\mu_4 + (1-\pi)\nu_{1|2(1),3(1)}\nu_{2|3(1)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_3\mu_4 + (1-\pi)(1-\nu_{1|2(1),3(1)})\nu_{2|3(1)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_3;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_3\mu_4 + (1-\pi)\nu_{1|2(0),3(1)}(1-\nu_{2|3(1)})\nu_3\nu_4,\\
&\dots\\
\pi(\boldsymbol{\gamma}_{16};\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_3)(1-\mu_4)\\
&\quad + (1-\pi)(1-\nu_{1|2(0),3(0)})(1-\nu_{2|3(0)})(1-\nu_3)(1-\nu_4),
\end{aligned}$$

with $\nu_{1|2(1),3(1)}$ replaced by (59).

This model has 12 parameters as the rank of corresponding Jacobian is also 12, hence, the model is locally identifiable.

Checking generic or rational identifiability is hard in this case. The tensor methods cannot be applied here as we cannot marginalize the above model to obtain the model of between-variables conditional independence given $\mathcal{M}$ and $\mathcal{U}$. When checking the rational identifiability with the Gröbner basis method, we did not manage to obtain a basis for this model after hundreds of hours of computations. Note, that the model specification involves rational functions due to the presence of (59) and this should be reflected when computing the Gröbner basis; see Cox et al. (2015, chap. 3).

5.2.7 Model $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$

The final model to consider has association between v_2 and v_3 in the set of matches and association between variables v_1 and v_2 in the set of non-matches. In record linkage and related applications this model specification can be used when linkage variables available are the full standardized address, surname, first name and date of birth. In this case, the association between comparisons on the address and surname variables are expected just as in one of the models presented above. In addition, this time it is assumed that there is association between errors made recording the values of surname and first name and this association either cannot be made nearly ignorable by approximate comparisons or approximate comparisons are not available. Provided there is a high quality address checking mechanism, we regard the dependence between surname and first name as the most likely dependence in the set of matches as these two variables are hardest to record and 'difficult' surnames are often accompanied by 'difficult' first names. This dependence is most likely in an interviewer led survey, where 'unfamiliar' or 'difficult' surnames and names are likely to be misheard or spelt incorrectly. The corresponding parameterization is

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_1;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1\mu_2\mu_{3|2(1)}\mu_4 + (1-\pi)\nu_1\nu_{2|1(1)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_2;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)\mu_2\mu_{3|2(1)}\mu_4 + (1-\pi)(1-\nu_1)\nu_{2|1(0)}\nu_3\nu_4,\\
\pi(\boldsymbol{\gamma}_3;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi\mu_1(1-\mu_2)\mu_{3|2(0)}\mu_4 + (1-\pi)\nu_1(1-\nu_{2|1(1)})\nu_3\nu_4,\\
&\dots\\
&\dots\\
\pi(\boldsymbol{\gamma}_{16};\pi,\boldsymbol{\mu},\boldsymbol{\nu}) &= \pi(1-\mu_1)(1-\mu_2)(1-\mu_{3|2(0)})(1-\mu_4) + (1-\pi)(1-\nu_1)(1-\nu_{2|1(0)})(1-\nu_3)(1-\nu_4).
\end{aligned}$$

There are 11 parameters in this model and the rank of the corresponding Jacobian is 11, which means that the model is locally identifiable.

It is possible to check rational identifiability for this model using the same approach we used for the rest of models considered in this chapter. In this case, Singular software is used. We found that functions such as $\pi - \pi^c$ are in the ideal I_c and hence all parameters are rationally identifiable without label switching. The related Singular example is in Appendix C.

6 Estimating the variance of the linkage free dual system estimator

It was discussed in Section 3.3 that the data generating mechanism in a record linkage problem effectively produces a single realization of a random vector. Therefore, the parameter estimation of a mixture-like model is based on a single observation only. Doing point estimation with a single observation results in a discrepancy between the model, which holds whenever multiple outcomes of the linkage experiment are averaged, and the actual data available in real applications. As a result, estimators of the linkage model and related parameters are not consistent. Nevertheless, even without all the desirable statistical properties, achievable point estimates are still useful and admissible estimates of the linkage model parameters. Given the complexity of the linkage problem, these estimates may well be the best or close to the best results one can obtain. The situation seems more grim when it comes to measuring the uncertainty around these estimates. With a single available observation, the variance estimation seems impossible at first. However, if a special case of blocking, called averaging blocking (see Section 3.4), is achievable, then it is feasible to obtain the approximate variance estimate for the linkage and related parameters using resampling methods.

6.1 General approach

Recall that the averaging blocking splits the population or estimation stratum of interest into non-overlapping groups, $G_1, \ldots, G_\beta, \ldots, G_B$, of equal or nearly equal size τ_β. It is assumed that the values of linkage variables in each G_β are generated by the same or very similar mechanisms. Data sub-samples $S_{1,\beta} \subset S_1$ and $S_{2,\beta} \subset S_2$ are independent enumerations of a block β. The comparisons of the values of linkage variables are made between the records of the sub-sample $S_{1,\beta}$ and the records of the sub-sample $S_{2,\beta}$, but not between any sub-samples $S_{1,\alpha}, S_{2,\beta}, \alpha \neq \beta$. The important assumption behind the averaging blocking is that for each of $j = 1, 2$ every $S_{j,\beta}$ is generated by the same or nearly the same mechanism. Moreover, every $S_{j,\beta}$ is generated by the same mechanism as S_j is generated, only the population size parameters are different. Clearly, $S_{1,\beta}$ and $S_{2,\beta}$ are generally generated by different mechanisms. If such averaging blocking is achievable, then it is possible to estimate the approximate variance of the linkage free estimator $\widetilde{\tau}$, as well as the variance of the other linkage parameters. Our focus here is on the variance estimation of $\widetilde{\tau}$, though.

First, the basic approach for the variance estimation of $\widetilde{\tau}$ for an idealised case where all blocks $G_1, \ldots, G_B$ are of the same size τ_β is presented. Achieving such a fine blocking is hard or even impossible in practice which means there will be an artefactual variability due to varying block sizes. Two ways of addressing this problem in practical applications will be presented. Note, that we are not

discussing variance estimation with complex sampling of either or both of S_1 and S_2 from the finite population. It is assumed in this work, that observations for both S_1 and S_2 are selected using simple random sampling without replacement (though for simplicity we often use models where sampling is with replacement in this thesis). Also, for simplicity, we are not correcting for finite population sampling here. Recall, that in applications like the census, at least one of the samples would be drawn using some stratified multistage sampling design: a selection of output areas would be drawn within a stratum and then a sample of postcodes would be drawn from each selected output area. It is most likely, however, that within a sample stratum every sampled postcode would have the same sampling fraction.

The aim is to obtain an approximate variance estimator $\widehat{\text{Var}}(\widetilde{\tau})$ of the linkage free dual system estimator $\widetilde{\tau} = \frac{1}{\widetilde{\pi}}$ from two independent samples S_1 and S_2, where τ is the size of a population $\mathcal{P}$. In our idealised case we perform the averaging blocking which partitions the population $\mathcal{P}$ into B blocks of nearly equal unknown size τ_β. The linkage free dual system estimates are obtained for each block first, $\widetilde{\tau}_\beta = \frac{1}{\widetilde{\pi}_\beta}$. Consider the following approximation

$$\widetilde{\tau} \approx \sum_{\beta=1}^{B} \widehat{\pi}_\beta^{-1} = \sum_{\beta=1}^{B} \widetilde{\tau}_\beta = B\widetilde{\overline{\tau}}, \tag{60}$$

where

$$\widetilde{\overline{\tau}} = \frac{1}{B} \sum_{\beta=1}^{B} \widetilde{\tau}_\beta.$$

Then we propose to use the following approximation to estimate the variance of $\widetilde{\tau}$

$$\text{Var}(\widetilde{\tau}) \approx B^2 \text{Var}(\widetilde{\overline{\tau}}), \tag{61}$$

which employs the basic fact that for a random variable X and a constant y we have $\text{Var}(yX) = y^2\text{Var}(X)$.

Estimate the variance of $\widetilde{\overline{\tau}}$ using the method of random groups (Wolter, 2007; chap. 2) by treating every block as a random group with common expected population size

$$\widehat{\text{Var}}(\widetilde{\overline{\tau}}) = \frac{1}{B(B-1)} \sum_{\beta=1}^{B} (\widetilde{\tau}_\beta - \widetilde{\overline{\tau}})^2. \tag{62}$$

Finally, obtain the approximate variance estimator of the population size

$$\widehat{\text{Var}}(\widetilde{\tau}) \approx B^2 \widehat{\text{Var}}(\widetilde{\overline{\tau}}). \tag{63}$$

The approximate variance estimator for the 1-to-1 constrained linkage free dual system estimator (53) is defined in exactly the same way. We replace $\widetilde{\tau}$ with $\widetilde{\tau}_c$, $\widetilde{\tau}_\beta$ with $\widetilde{\tau}_{c,\beta}$ and $\widetilde{\overline{\tau}}$ with $\widetilde{\overline{\tau}}_c$ in the above derivations. The linkage free estimate of the size of each block is obtained first. The modified composite linkage free estimator is then used to obtain the size of each block under the 1-to-1 linkage

constraint. The corresponding variance estimator for the population size is $\widehat{\text{Var}}(\widetilde{\tau}_c) \approx B^2\widehat{\text{Var}}(\widetilde{\overline{\tau}}_c)$.

Regarding the distinction that exists in the capture-recapture literature between the variance conditional or unconditional on the observed sample sizes, theoretically, the method presented in this thesis can be used to estimate either the unconditional or conditional variance. This depends on what kind of data generating mechanism the population of interest has: either fixed sample sizes or random sample sizes (but fixed coverage probabilities). As stated above, it is assumed that each block has the same or a very similar generating mechanism to the one used to generate the target population. Then, in theory, for the fixed sample sizes case, each block also has fixed sample sizes. Whereas for the random sample sizes case, each block has a random sample size. The method developed here will always have some artefactual variability when applied in practice which is attributable to some inevitable variability in the population size of each block. This means that the sample sizes will also vary between the blocks, no matter what the true generating mechanism is. The conclusion would be that in practice the proposed variance estimator cannot estimate the conditional variance, even if the sample sizes were fixed at the data collection stage.

6.2 Practical approach: no auxiliary data

As already mentioned, it is not in general possible to achieve the perfect averaging blocking when the linkage exercise is carried out with the goal to estimate the population size (but it may be possible in some linkage exercises where the population size is known). Therefore, in the situations we are interested in, no blocks of approximately equal size τ_β are directly available. Nevertheless, in many well-planned and designed data collection exercises for the population size estimation, there are good quality address frames or / and address listings available. Those are data with the approximate number of addresses for low level geography units such as postcodes and output areas. Knowing how many postcodes are in the population of interest $\mathcal{P}$ and how many addresses are in each of these postcodes, allows postcodes to be collapsed into blocks with the number of addresses being as equal between the blocks as possible. This still will not result in blocks of nearly equal size τ_β for at least two reasons. First, postcodes vary considerably in size. Hence, collapsing postcodes within the population $\mathcal{P}$ into blocks that have equal numbers of addresses is generally unlikely. Second, despite a strong correlation between the number of addresses and the number of individuals within a certain geographical unit, there is still substantial variability of the number of individuals given the number of addresses. Overall, this additional variability will result in (63) overestimating the variance of τ.

Two solutions to the above problem are proposed and presented. Both solutions aim to adjust for differential sizes of the blocks. There may be variations of these solutions, dependent on the situation and data available.

The first solution is rather crude, has certain flaws in the justification and its performance is far from perfect. Nevertheless, it is capable of providing an indicative variance estimate and does not require much effort to be computed. We use the observed sizes of two samples $n_{1,\beta}, n_{2,\beta}$ for a block β to do the adjustment. Of course, there is an obvious flaw in such attempt since the observed size of any of these samples is not only a function of the block's size τ_β, but is also affected by the variability of π_β, which drives the variability we are trying to estimate. The adjustment and variance estimation

are as follow. Let

$$\mu_{\text{adj}} = \frac{1}{B}\sum_{\beta=1}^{B}(n_{1,\beta}n_{2,\beta})^{\frac{1}{2}},$$

be the mean of the square root of the product of the sample sizes within each block. The square root is needed to ensure that μ_{adj} and τ_β are on the same scale. Then every τ_β is adjusted to produce

$$\widetilde{\tau}_{\text{adj},\beta} = \widetilde{\tau}_\beta \frac{\mu_{\text{adj}}}{(n_{1,\beta}n_{2,\beta})^{\frac{1}{2}}}. \tag{64}$$

The method of random groups is used in the way described above, so that

$$\widehat{\text{Var}}(\widetilde{\overline{\tau}}_{\text{adj}}) = \frac{1}{B(B-1)}\sum_{\beta=1}^{B}(\widetilde{\tau}_{\text{adj},\beta} - \widetilde{\overline{\tau}}_{\text{adj}})^2,$$

where

$$\widetilde{\overline{\tau}}_{\text{adj}} = \frac{1}{B}\sum_{\beta=1}^{B}\widetilde{\tau}_{\text{adj},\beta}.$$

The approximate variance estimator is

$$\widehat{\text{Var}}(\widetilde{\tau}) \approx B^2\widehat{\text{Var}}(\widetilde{\overline{\tau}}_{\text{adj}}). \tag{65}$$

The same derivation can be used for the one-to-one constrained linkage free dual system estimator.

6.3 Practical approach: with auxiliary data

The second approach is slightly more elaborated, has a better justification and performs quite well. It relies on the availability of the auxiliary data such as an address frame or address listing, may need some additional processing effort and is more complicated. Let l_β be the number of addresses in the block β as found in the address frame or address listing. Let $l_{1,\beta}$ and $l_{2,\beta}$ be the number of responding addresses in the block in the first and second samples, respectively. Some comments are needed here. Establishing $l_{1,\beta}$ and $l_{2,\beta}$ may involve just counting the numbers of responding addresses in each of the samples irrespective of the counts in the auxiliary data, or it may involve mapping the responding addresses to those contained in the auxiliary data. There may be some discrepancy between those counts and l_β in practical applications. For instance, the observed count may be higher than one found on the address list. Also, responses at the address level may be mapped to the address frame in the case of the first sample, but to the address listing in the case of the second sample. One way or another, there is additional processing required. Nevertheless, the method presented here does not need perfect pre-processing. Rather, it requires high correlation between $n_{1,\beta}$ and $l_{1,\beta}$ as well as between $n_{2,\beta}$ and $l_{2,\beta}$ and reasonably accurate l_β.

Let $A_\beta = l_\beta \frac{(n_{1,\beta} n_{2,\beta})^{\frac{1}{2}}}{(l_{1,\beta} l_{2,\beta})^{\frac{1}{2}}}$ and $\bar{A} = \frac{1}{B}\sum_{\beta=1}^{B} A_\beta$. Then every τ_β is adjusted this way

$$\widetilde{\tau}_{\text{aux},\beta} = \widetilde{\tau}_\beta \frac{\bar{A}}{A_\beta}. \tag{66}$$

As before, the method of random groups is used:

$$\widehat{\text{Var}}(\widetilde{\overline{\tau}}_{\text{aux}}) = \frac{1}{B(B-1)} \sum_{\beta=1}^{B} (\widetilde{\tau}_{\text{aux},\beta} - \widetilde{\overline{\tau}}_{\text{aux}})^2,$$

where

$$\widetilde{\overline{\tau}}_{\text{aux}} = \frac{1}{B} \sum_{\beta=1}^{B} \widetilde{\tau}_{\text{aux},\beta}$$

and the approximate variance estimator is

$$\widehat{\text{Var}}(\widetilde{\tau}) \approx B^2 \widehat{\text{Var}}(\widetilde{\overline{\tau}}_{\text{aux}}). \tag{67}$$

Again, the derivation is similar for the one-to-one constrained linkage free dual system estimator.

In Chapter 8 we assess the performance of the above variance estimators through simulations.

6.4 Why not just use bootstrap?

Variance estimation methods presented in this chapter may seem excessive at first glance. Why not do something simpler, say, the bootstrap, instead? The reason why the bootstrap method is not appropriate is related to the sampling or data generating mechanism associated with the record linkage task, as described in Section 2.2.5. It turns out, at least with simple bootstrapping approaches, that it is hard to adequately approximate the original distribution of the record linkage data.

Before showing the inappropriateness of several potential bootstrap approaches to the variance of the linkage free estimator, we make a few important observations. First, recall that we usually model the number of matches either with binomial or hypergeometric distribution. Second, sampling with replacement from $\{x_1, \ldots, x_n\}$, is the same as sampling from a multinomial distribution, with the number of trials being n and the probability of drawing x_i being $1/n$ (Efron & Tibshirani, 1993, chap. 20.2).

The third observation concerns the simplest implementation of the bootstrap method for the variance of the dual system estimator (Buckland & Garthwaite, 1991; Norris & Pollock, 1996). What is resampled with replacement in the case of the non-parametric implementation are the so-called capture histories. That is, the observed matches (in simple dual system estimation, unlike the no-classification methods, matches are treated as observable values), cases in the first sample only and cases in the second sample only. It is easy to see that such an implementation will generally underestimate the true variance. The relationship between sampling with replacement and a multinomial distribution means that we resample from $x_1, \ldots, x_{m+n_{10}+n_{01}}$ capture histories with each x_i taking one of the three possible values. Therefore, given the original sample, the bootstrap samples will be from a multinomial

distribution with the number of trials equal to the number of captures at least in one of the surveys, $m + n_{10} + n_{01}$, and the probabilities of resampling a certain event will be the observed number of that event divided by $m + n_{10} + n_{01}$. It is clear that such an approach does not account for the cases missed in both samples. Therefore it does not always accurately estimate the variance of the dual system estimator. Buckland & Garthwaite (1991) suggest estimating the population size τ first and then sampling from the observed capture histories as well as from the estimated $\widehat{n}_{00}$ histories of being missed completely. The number of observations to resample (or the number of trials in the corresponding multinomial distribution) is $\widehat{\tau}$ in this case. It is a peculiar example, since the bootstrap distribution (multinomial) coincides with the possible original distribution of the events. Note, that in Buckland & Garthwaite (1991) this approach is referred to as the non-parametric bootstrap. The parametric case uses the estimates of τ, π_1, π_2 and, if needed, the observed quantities n_1, n_2 to generate samples from either a multinomial or hypergeometric distribution.

The no-classification methods are quite different from simple dual system estimation due to linkage being an integral part of the estimation. Therefore, one can think that the variance of these methods combines the variance of the simple dual system estimator (which is mainly driven by the variation in the number of matches) and the variance associated with estimating the number of matches. The input data are not $m + n_{10} + n_{01}$ capture histories, but $w = n_1 n_2$ record pairs with the corresponding comparison outcomes. In other words, there are no capture histories available, as there is no classification. Even if we treat the outcomes of 1-to-1 pairings as a kind of capture history, there will be no values of the linkage variables available for the $\widehat{n}_{00}$ estimated cases missed in both surveys. Therefore, there is no straightforward way of creating the comparison patterns for such estimated cases. Below we overview and present arguments against several possible implementations of the bootstrap variance estimation of the linkage free dual system estimator.

One could attempt the non-parametric bootstrap from a set of all ordered pairs $\mathcal{W}$. That is, produce a bootstrap resample by sampling with replacement from the w observed record pairs. However, such a sampling draws each pair independently from the rest of the pairs. It was shown in Section 2.2.5 that there is within-variables associations in the comparison outcomes. Therefore, sampling the original pairs independently would not preserve the underlying associations between record pairs and would not reflect the true variance of the no-classification estimator.

An alternative approach would be to estimate the parameters of the record linkage model and obtaining the no-classification based population size estimate first, then proceed with the parametric bootstrap. Unfortunately, we do not know exactly what the underlying distribution of the record linkage data is. Hence, we cannot implement the parametric bootstrap.

One may try to bootstrap the original survey observations, make the comparison and proceed with the estimation for each bootstrap replicate. Below we discuss two strategies to resample the original survey observations and demonstrate that the resulting variance estimates will be inadequate. The crucial observation here is that the no-classification approaches have the variance component due to variability of M, N_1, N_2 and linkage variability (which is at least partly driven by the distribution of the attributes and errors). Therefore, in order for a bootstrap scheme to work it must approximate well the distribution of matches M in the first place. If the bootstrap distribution of the number of matches is

far from the true distribution of matches, the distribution of the bootstrap no-classification estimates will also be far from the true distribution of estimates. As before, two working distribution for M are the binomial and hypergeometric. Supports for these distributions are $\{0, \ldots, \tau\}$ and $\{\max(0, n_1 + n_2 - \tau), \ldots, \min(n_1, n_2)\}$, respectively. Given that in this thesis we focus on the application of the no-classification methods to estimate a relatively small domain with high coverage in the surveys, the number of matches generally lies between $\max(0, n_1 + n_2 - \tau)$ and $\min(n_1, n_2)$.

The first strategy to resample the original survey observation is the following one. Once the data from two surveys are collected, we use simple random sampling with replacement to draw n_1 records from the first survey and n_2 records form the second survey. These resampled records give rise to w record pairs. Then comparisons of the values of linkage variables are carried out for these pairs and we proceed with the parameter estimation in the same manner we do with the original data. Now observe that the distribution of the number of matches associated with such a resampling scheme has support $\{0, \ldots, n_1 n_2\}$. Here is an example of the extreme case of no matches in a bootstrap sample. Suppose that a record pair (a, b) is a non-match. Then drawing n_1 records $s_{1,a}$ from the first survey and n_2 records $s_{2,b}$ from the second survey will result in 0 matches among $n_1 n_2$ pairs. The extreme case of getting $n_1 n_2$ matches is as above, but with a record pair (a, b) being a match. It is possible to work out the probability mass function of obtaining exactly m matches when using the above bootstrapping scheme. Nevertheless, this probability function is very cumbersome. It has to reflect the fact that exactly m matches can be obtained in multiple ways. For instance, by resampling only a single matching pair that results in m matches; or by resampling any two matching pairs in such a way that m matches are achieved, and so on. In any case, it is clear that this distribution differs substantially both from the binomial and hypergeometric distributions. In order to get an idea how much the variability of matches under such a bootstrap scheme differs from the variability under the two usually used distributions, one can conduct a simple simulation. For instance, let the population size be $\tau = 500$, and the coverage probabilities be $\pi_1 = \pi_2 = 0.9$, so that the expected sample size is 450 for the both surveys. If M follows the hypergeometric distribution with $\tau = 500, n_1 = n_2 = 450$, then its variance equals 4.05. If M follows the binomial distribution with $\tau = 500$ and the probability of success $\pi_1 \pi_2 = 0.81$, then its variance equals 76.95. Suppose now that the collected data are perfectly calibrated to the expected values, so that $n_1 = n_2 = 450, m = 405$ and the bootstrap samples are drawn from these data. The bootstrap should approximate the original distribution reasonably well in order to be useful. The bootstrapped variance of M, however, is 437.80 in this case, which is substantially larger than the possible true variances. It is clear that such a bootstrapping scheme is not approximating well the distribution of the number of matches. Experiments confirmed severe overestimation of the variance of the no-classification estimators under such a bootstrapping scheme.

An alternative approach would be to fix the records of one of the surveys and bootstrap from the remaining one. Without loss of generality, suppose that we fix the sample S_2 and use simple random sampling with replacement to draw n_1 records from S_1. Then we create record pairs and proceed with estimation. Observe, that given the observed records of the first survey each resampled $s_{1,a}$ either has a corresponding match among the records of the second survey or not. Therefore, the number of matches in a bootstrap sample follows the binomial distribution with the number of trials n_1 and

the probability of success $\#\{\text{matches in the original sample}\}/n_1 = \widehat{\pi}_2$. Recall, that if the original model for the number of matches is binomial, then $M \sim Bin(\tau, \pi_1\pi_2)$, not $M \sim Bin(n_1, \widehat{\pi}_2)$ as in the case of the bootstrap. Experiments demonstrated, that in most cases such an approach leads to underestimation of the variance of the no-classification estimators, but the deviation from the true variance is not as severe as when records from the two surveys are resampled. This is what one might expect for good coverage in both surveys, as n_1 is close to τ and $\widehat{\pi}_2$ does not differ substantially from $\pi_1\pi_2$.

The fact that we are dealing with pairwise comparisons in record linkage also means that we cannot use some arbitrary random groups that are independent between the surveys. In this case, we would have an additional source of variation, since there is no guarantee that a population element captured in both S_1 and S_2 will end up in the same random group. Therefore, the averaging blocking based random groups provide the means for a meaningful variance estimation.

7 Simulation study for point estimation

In this chapter a simulation study to assess (a) the validity of the mixture-like conceptualisation of record linkage, as developed in Chapter 3, and (b) the performance of the linkage free and modified linkage free dual system estimators, as described in Chapter 4, is presented. Before the results are summarized and analysed, all the key aims of this simulation study are outlined and the properties of simulated data are introduced alongside a description of the data generating mechanism. The combination of parameters that constitute the simulation scenarios are also given. The population size estimators based on the no-classification linkage approach developed in this thesis are assessed from two perspectives in this study. The first is the practical perspective which considers application of the estimators as they would be applied in most practical situations. That is, given the data, applying the estimator to some estimation strata without averaging blocking or with averaging blocking done similarly to the post-stratification. Such practical application has certain caveats (Section 3.3). Alternatively, the theoretical perspective concerns the application of the estimators in the near perfect averaging blocking set up. This set up leads to a theoretically better justified and data-conforming estimation. However, it is not in general feasible to achieve such a set up. The same data generating mechanism is used in the investigation of point estimates in this chapter, and of variance estimates in Chapter 8, but the variance estimation is assessed on a smaller number of scenarios.

7.1 Aims of the simulation study

The data generating mechanism associated with record linkage (Section 2.2.5), its representation via mixture-like models (Section 3.1), the estimation of the linkage model parameters and deriving the related linkage free dual system estimator using the simulated annealing approach (Sections 3.5 and Section 4.1) and finally reflecting the 1-to-1 matching constraints (Section 4.2) are all too complex to be dealt with analytically. A simulation study, while lacking the ability to establish general properties, allows us to assess the performance of the estimators and check whether certain features predicted in theory are substantiated in empirical results. Therefore several goals are pursued in this study:

1. Check whether the linkage free and modified linkage free dual system estimators have comparable performance to the dual system estimator with the number of links obtained by a usual classification-based approach; what is the performance of the estimators of interest benchmarked against the dual system estimator with perfect linkage;

2. Determine how big, if any, is the reduction in variance and bias when using the modified linkage free estimator compared to the linkage free dual system estimator;

3. Check whether the linkage free dual system estimator constructed to be data-conforming (Section 3.4) functions as anticipated and whether the bias is incurred for certain combinations of the number of unique values of attributes ρ_v as demonstrated using the Taylor series approximations in Section 3.2;

4. Check if the process of replicated values of a certain attribute within the given value of another attribute (like repeating the value for the surname attribute within the household address attribute) in the population results in between-variables dependence of the comparison outcomes as described in the linkage experiment (Section 2.2.5);

5. Check if the probability distribution of the number of comparisons is not multinomial when the data are generated as described in the linkage experiment (Section 2.2.5);

6. Assess whether the averaging blocking strategy outperforms estimation at the aggregate level in practical applications;

7. Check if the identifiability and non-identifiability of models established in Chapter 5 are supported by empirical simulations.

7.2 Simulating data

The data generating mechanism used in this simulated study is an implementation of the linkage experiment described above in Section 2.2.5. Here, we give more details on the kind of data and how they are generated. We also present the actual parameters used in the simulations and scenarios made by combining the individual parameters. All simulation and estimation related code is written in the R language (R Core Team, 2020). Most of the simulation and estimation work was run on the University of Southampton Iridis 5 supercomputer cluster. Some of the simulation and estimation jobs and all the development and testing was carried out on the workstation with 256GB RAM and Intel®Core™ i9-10920X CPU with twelve 3.50GHz dual cores.

There are several stages in the data generation mechanism aiming to mimic as closely as possible the process that brings about the data used in real world record linkage situations. These stages are: (a) generate the target population with a specific nesting of individuals within addresses / households and nesting of households within geographic areas such as postcodes or output areas, also generate the values for attributes associated with each household and individual such as address, surname, etc.; (b) decide which elements in the population responded in the first survey and which responded in the second survey, the decision process is independent for each of the population elements (so that

the decision whether a certain household responds is independent of the decision whether any other household responds, also within a responding household, the decisions on individual responses are made independently for every member of the household) and is independent between two surveys; (c) for each of the surveys, record the population attributes as the values of linkage variables allowing for some of the values to be missing or be recorded incorrectly; (d) pair the records of the two surveys and for every pair carry out a binary comparison of the values of each linkage variable; this process results in a comparison pattern $\boldsymbol{\gamma}(a, b)$ for every record pair (a, b); (e) finally, compute the frequency of the each of 2^K comparison patterns. The data obtained at the last stage (e) are actually used in parameter estimation of the mixture-like model by the simulated annealing approach and the linkage free dual system estimator is based on these estimates. However, the data from stages (b) and (d) are also kept and used in the modified linkage free dual system estimator and in variance estimation. On top of these generated data, there is an output containing the true match status of each pair and the total number of true matches, data tables associated with the set of matches and the set of non-matches, so that the perfect linkage outcome is available for benchmarking. For a simulation scenario, defined by a particular combination of simulation parameters, the simulation process as outlined above runs a specified number of iterations. In this study, the number of iterations is 1000 for all scenarios. In the case of assessing practical application of the linkage free dual system estimator, the simulated data for each iteration are passed to the estimation in turn. Thus, 1000 independent estimates are produced and then analysed. When assessing the possibility to construct a data-conforming estimator, 1000 simulated outcomes from the stage (e) are averaged, thus mimicking the perfect averaging blocking (where each block has the fixed unobservable size τ), and then passed to parameter estimation. The simulated annealing is run with 10 random starts for the same input data to account for the small variability in the outputs of the algorithm for the fixed input data. Then these 10 estimates are averaged to produce only a single estimate which allows assessing whether the estimator converges to the true parameter value or not. All simulations are conducted with $K = 4$ linkage variables. However, the simulation system itself has no limit on the number of linkage variables. Below we provide a more detailed description of the stages and parameter values used in the study.

At the first stage of simulation a population of interest $\mathcal{P}$ is produced using a set of fixed parameters defining this population. The first parameter is the population size τ. Three values of this parameters are used, $\tau \in \{250, 500, 1000\}$. Recall, that in this thesis the dual system estimation of the population total of a relatively small stratum or post-stratum is considered. If the population of interest is large, then it must be further stratified in a meaningful way and estimates for each stratum are obtained in turn. Recall also, that the simple dual system estimators (4) or (5) will have a small bias unless τ is large. We do not explore the scenarios where τ is large enough to make the bias in the dual system estimator near non-existent. This is because the contribution of such bias to the mean square error is small even for the values of τ considered in this thesis. Another reason is that computational burden which is associated with the linkage free estimation grows rapidly because the number of pairs is quadratic in the population size (unless blocking is used, but many of the scenarios of interest do not use any blocking).

Each individual observation representing a person in the population can be nested in a higher level

unit. In this simulation study every person is nested in a household. The average size of households is another parameter defining the population and in all scenarios its value is set to be 2.4. A person can belong to only one household. Also, every household is nested in one and only one artificial geographical unit with the mean size of 60 persons. This unit can be thought as a combination of neighbouring postcodes.

Attributes of the population elements are generated with the help of two parameters. The first parameter is the number ρ_k of unique values the k^{th} attribute can take. The second parameter controls the frequency distribution of the values, with some attributes being uniformly distributed, while others may have certain values to be substantially more frequent than other values of the same attribute. This corresponds to the situation with names or surnames where there are a few very common names or surnames while many other names or surnames are infrequent. The actual values of attributes are just numbers from 0 to $\rho_k - 1$, rather than real names and surnames, with some of the distributional properties similar to the real-world attributes. It is easier and more efficient to work with such numerical attributes. Clearly, if the distribution of values of the k^{th} attribute is uniform, then each value occurs in the population with probability $1/\rho_k$. On the other hand, if the distribution of values is not uniform, then the value $x_{k,i}$, where $x_{k,i} \in \mathbb{Z}$ lies in the interval $[0, \ldots, \rho_k - 1]$, will occur with the probability $\lambda_k \exp(-\lambda_k x_{k,i}) / \sum_{i=1}^{\rho_k} \lambda_k \exp(-\lambda_k x_{k,i})$. Here $\lambda_k \exp(-\lambda_k x_{k,i})$ is the probability density function of the exponential distribution with parameter λ_k and $x_{k,i}$ are the values of the k^{th} attribute. The following values of the parameters ρ_k and λ_k are used to test practical applications of the estimators. The first linkage variable has $\rho_1 = 10000$ and is uniformly distributed (so λ_k is irrelevant in this case). We usually think of the first variable as the full standardized address. The reason to have that many possible values for the address attribute is to mimic the situation where within a small population with the size up to 1000 individuals, the chance of having genuinely the same address is rather small. The uniform distribution of values helps to achieve this. The second and third variables have $\rho_2 = \rho_3 = 500$ and $\lambda_2 = \lambda_3 = 0.01$. We usually think about the second variable as being surname and third one being first name. Clearly, there are a lot more names and surnames in the population. However, the 500 most common names may account for about $75\% - 85\%$ of all names in the population and only around 300 most common surnames may account for 30% of all surnames, while 3500 next most frequent surnames account for another third of all surnames (US Census Bureau, 1990). Given that we work with a population size which does not exceed 1000 individuals, the chosen sizes are reasonable and practical. The values of the fourth variable are uniformly distributed and there are $\rho_4 = 100$ distinct values. This variable can be thought as a sort of 'compressed' or 'binned' date of birth. We note that some of the models considered in Chapter 5 and used in simulations below do not support the presented selection of variables and related numbers of unique values. For instance, we said that v_3 in model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1 - \pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$ can be relationship to the head of household. Clearly, this variable cannot have 500 levels. In order to have the results of simulations comparable between different model specifications, we had to fix the above parameters regardless of such discrepancies.

The final feature related to generation of the population, is the association between the values of certain attributes given another attribute. Such associations in the values of population attributes

lead to dependence between the frequencies of comparison patterns in the set of non-matches $\mathcal{U}$ as obtained at stage (e) of the simulations. There are three settings of such associations in this thesis. The first one is independence between all of the population attributes. That is, every population element has the values of attributes generated in such a way that these values are independent both within the population element and between the values of attributes of the rest of population elements. The resulting stage (e) data generated under such conditions are referred to as the conditional between-variables independence given the set of non-matches. The second setting is the association between the first and second attributes, where the first attribute can be thought of as address and the second as surname. The association is achieved in the following way. Within each address the address value is repeated for all individuals that are nested in this address. The address attribute is thus a 'parent' attribute. Then for the first individual within an address, the value of surname is generated in the way described above. For every remaining individual within the address, the value of the surname is the same as the value of the first individual's with probability 0.8. If the value of surname is different, it is drawn using the same mechanism as the value for the first person within the address. Population attributes generated in this way result in dependence between the frequencies of the first two variables at the final stage of simulation. Recall, that such dependence may be generated for any pair of attributes / linkage variables, the specific choice of the first two attributes is arbitrary. Finally, the third setting is exactly the same as the second one, but the values of the third attribute are generated in the same way as the values of the second attribute. Therefore, there is a multi-way dependence between the frequencies of these three variables. Observe, that while the population attribute generating process does not explicitly produce association between the second and third variables, once the comparisons are made and their frequencies are computed, there will be association in frequencies between the second and third variables. This is because the comparison outcomes are functions of the attribute values of the population elements. Simultaneous agreements on the second and third variables among non-matches within matching households occur only for the households where at least two populations elements have the same values of the second and third attributes. Whereas marginally an agreement among non-matches within matching households for the second (and similarly third) linkage variable occurs not only in the above case, but also when at least two population elements within a household have the same values for the second variable only. Therefore, the product of marginal probabilities is larger than the joint probability of observing simultaneous agreement on the second and third linkage variables.

Once the population is generated, the next step is to produce two data surveys with non-response (undercoverage in the census coverage terminology). There are two stages of non-response in each of the surveys. The first stage is the household response at which any given household on the census address frame or coverage survey address listing either responds or not. The second stage is the within household response at which individuals within responding household either respond or not. The probability of within household response is fixed at 0.98 for both surveys and for all scenarios. The values of the household response are selected in such a way, that the probability of response for a person in a given survey is approximately one of the following: $\pi_1, \pi_2 \in \{0.9, 0.8, 0.7\}$. In order to keep the number of simulation scenarios manageable, the response probabilities for all scenarios satisfy

$\pi_1 = \pi_2$.

The next step is to generate errors in the values of the linkage variables. Those errors are generated using the vectors of the fixed error probabilities $\boldsymbol{\xi}_j$, where $j \in \{1, 2\}$ is the index of the survey. Again, to keep the number of scenarios manageable, $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = \boldsymbol{\xi}$ is used. Three distinct vectors $\boldsymbol{\xi} \in \{(0.003, 0.01, 0.01, 0.01)^T, (0.015, 0.05, 0.05, 0.05)^T, (0.03, 0.1, 0.1, 0.1)^T\}$ of the error probabilities are explored in the studies. The error probabilities in the vector are indexed in the same way as the linkage variables. Observe, that the errors occur in both surveys independently of each other, and the overall probability of error for the k^{th} variable is $\xi_{1,k} + \xi_{2,k} - \xi_{1,k}\xi_{2,k}$. The reason why the first variable has a smaller probability than the rest of the variables is related to the fact that we tend to think about the first variable as the standardized full address. The quality of collection values for this variable is expected to be higher than the quality of the remaining variables for the reasons explained in Section 2.3. There are two types of errors. The first type is a missing or genuinely non-existent value. For instance, a real world equivalent of a genuinely non-existent population value would be if a name 'John' was recorded as 'gnh'. The second type of error is such that while a value of an attribute is recorded incorrectly, the recorded value exists in the population. For instance, a real world equivalent would be if a name 'John' was recorded as 'Johnny'. Given that the attribute's value is erroneously recorded, the probabilities of the error of the first type occurring are: 0 in the first variable (which always has some address from the frame, so is never missing or non-existent in the population), 0.3 in the second and third variables and 0.2 in the fourth variable. The probability of the first type for the surname variable may seem too low as one would expect typographical or phonetic errors to result in surnames that do not genuinely exist in the population. However, assuming that the data undergo thorough data cleaning and validation checks, this process will fix cases with small typos, but will not always fix them correctly relatively to the true attribute's value. For instance, suppose that the actual value of the surname attribute is 'Johnson' and it gets mistyped as 'Jhns'. Then at the data cleaning and validation step the following resolutions are possible: 'Johnson', 'Johnston', 'Johnstone', John', 'Johns', 'Johnsen', 'Johnsey', etc. While all these values exist in the population, only one of them is actually correct. Therefore, it is reasonable to consider a situation where there are fewer cases of completely missing surnames or surnames with failed verification than the cases where verification incorrectly resolved mistyped surname.

The next step is to generate an association between the errors if necessary. There are two settings regarding dependent errors. The first one, is when all errors are independent. The second setting, is an association in errors of the second and third variables, say, surname and name. In this case, the errors are associated for a given element on a given survey, but are not associated between different elements within a household (which can be quite likely in reality, however, it will result in association within a given variable k and the mixture-like model does not make any assumptions of absence of such associations) and are not associated between different surveys. Unlike between-variables dependence in the set of non-matches that are the consequence of the properties of the population attributes, the between-variables dependence in the set of matches is more straightforward. In this case we can directly obtain the joint probability of the errors in two variables such that it does not equal the product of their marginal probabilities. The association is controlled via the odds ratio parameter and

only pairwise dependence is allowed at the moment. In the case that the association between the errors of the second and third variables exists, the corresponding odds ratio parameter is equal to 5.

After surveys are drawn, records from S_1 are paired with records from S_2. For every pair, it is checked for each linkage variable in turn whether a value of the variable on S_1 is the same as the value of the variable on S_2 or not. This process produces the comparison pattern vectors $\boldsymbol{\gamma}(a, b)$. Pairs can be formed either across the entire population or within blocks for the averaging blocking. In the case of pairing across the entire population, the number of pairs in each iteration is $n_1 n_2$ for the corresponding realization n_1 and n_2 of the random variables N_1 and N_2. In the case of averaging blocking, the number of pairs is $\sum_{\beta=1}^{B} n_{1,\beta} n_{2,\beta}$.

Finally, the frequencies of comparison patterns $\boldsymbol{\gamma}_k$ are calculated. If no blocking is used, then the frequencies are calculated across the entire population and the results are ready for the estimation using the simulated annealing approach. In the case of averaging blocking the frequencies are calculated for each block and then averaged for each of the comparison pattern. These are the ultimate simulated data. We stress again, that these final data are not just drawn from some parametric family of probability distributions, but instead constructed hierarchically from 'first principles'. While careful simulating from a certain parametric distributions ensures that the achieved relationships between random variables are as intended (for instance, total between-variables independence, etc.) simulating from 'first principles' leaves some uncertainty about whether the result is exactly as intended. Therefore, we prefer to say that we aim at certain relationships between random variables (say, aim at total between-variables independence).

The scenarios are produced by taking the outer product, denoted $\otimes$, of the individual parameters and we have 189 scenarios when assessing the linkage free population size estimators from a practical perspective. It is easiest to present the combinations of parameters in batches, even though later we use other ways of organising scenarios when presenting our simulation results. This first batch of 108 scenarios is obtained by the following combination:

$$
\begin{aligned}
\{\text{batch } 1\} = {} & \{250, 500, 1000\} \text{ (population size } \tau\text{)} \\
& \otimes \{0.9, 0.8, 0.7\} \text{ (response probabilities } \pi_1, \pi_2\text{)} \\
& \otimes \{(10000, 500, 500, 100)^T\} \text{ (distribution parameter } \boldsymbol{\rho}\text{)} \\
& \otimes \{(0, 0.01, 0.01, 0)^T\} \text{ (distribution parameter } \boldsymbol{\lambda}\text{)} \\
& \otimes \{(0.003, 0.01, 0.01, 0.01)^T, (0.015, 0.05, 0.05, 0.05)^T, (0.03, 0.1, 0.1, 0.1)^T\} \text{ (errors } \boldsymbol{\xi}\text{)} \\
& \otimes \{1\} \text{ (log-odds for errors)} \\
& \otimes \{\text{independence}, \text{association between two attributes}\} \text{ (association between population attributes)} \\
& \otimes \{\text{no blocking, average blocking with } \mathbb{E}(\tau_b) = 100\} \text{ (blocking).}
\end{aligned}
$$

Another bath is formed in the similar way and has 27 scenarios:

$$\begin{aligned}\{\text{batch } 2\} = {} & \{250, 500, 1000\} \text{ (population size } \tau\text{)} \\ & \otimes \{0.9, 0.8, 0.7\} \text{ (response probabilities } \pi_1, \pi_2\text{)} \\ & \otimes \{(10000, 500, 500, 100)^T\} \text{ (distribution parameter } \boldsymbol{\rho}\text{)} \\ & \otimes \{(0, 0.01, 0.01, 0)^T\} \text{ (distribution parameter } \boldsymbol{\lambda}\text{)} \\ & \otimes \{(0.003, 0.01, 0.01, 0.01)^T, (0.015, 0.05, 0.05, 0.05)^T, (0.03, 0.1, 0.1, 0.1)^T\} \text{ (errors } \boldsymbol{\xi}\text{)} \\ & \otimes \{1\} \text{ (log-odds for errors)} \\ & \otimes \{\text{association between three attributes}\} \text{ (association between population attributes)} \\ & \otimes \{\text{no blocking}\} \text{ (blocking).}\end{aligned}$$

The third batch of scenarios contains 54 scenarios:

$$\begin{aligned}\{\text{batch } 3\} = {} & \{250, 500, 1000\} \text{ (population size } \tau\text{)} \\ & \otimes \{0.9, 0.8, 0.7\} \text{ (response probabilities } \pi_1, \pi_2\text{)} \\ & \otimes \{(10000, 500, 500, 100)^T\} \text{ (distribution parameter } \boldsymbol{\rho}\text{)} \\ & \otimes \{(0, 0.01, 0.01, 0)^T\} \text{ (distribution parameter } \boldsymbol{\lambda}\text{)} \\ & \otimes \{(0.003, 0.01, 0.01, 0.01)^T, (0.015, 0.05, 0.05, 0.05)^T, (0.03, 0.1, 0.1, 0.1)^T\} \text{ (errors } \boldsymbol{\xi}\text{)} \\ & \otimes \{\text{log-odds for association between second and third variables is } 5\} \text{ (log-odds for errors)} \\ & \otimes \{\text{independence}, \text{association between two attributes}\} \text{ (association between population attributes)} \\ & \otimes \{\text{no blocking}\} \text{ (blocking).}\end{aligned}$$

7.3 Measures of performance and tuning parameters

In this simulation study the linkage free dual system estimator (44) and its modified version with 1-to-1 constraint (53) are compared to the basic dual system estimator (4) with perfect linkage and to the basic dual system estimator with classification-based linkage as presented in Section 2.2.2. In the case of classification-based linkage, the linkage models and parameter estimation are exactly the same as in the case of linkage free estimation. That is, we are not exploring how the conceptualisation of record linkage with regular mixtures and the use of maximum likelihood estimation is different from the mixture-like conceptualisation with simulated annealing to estimate the model parameters. Nevertheless, we have the means to check the robustness of the classification-based approach when the linkage model is incorrectly specified, and to assess the bias incurred by classification as well as to get an idea about the effort needed to resolve the possible links. Whenever a possible link is 'clerically' resolved, its true match status is used in the simulations below. Of course, in reality clerical linkage is also prone to errors, but in this thesis we impose a perfect outcome.

For all four estimators the following measures of performance are considered: the relative bias,

relative standard error and relative root mean square error. All three measures are relative to and in the same units as the population size and we choose to report them in percentages. For some arbitrary estimator $\widehat{\tau}$, with $\widehat{\tau}_r$ being an estimate at the r^{th} iteration, these measures are calculated as

$$\text{rb}(\widehat{\tau}) = \frac{1000^{-1}\sum_{r=1}^{1000}(\widehat{\tau}_r - \tau)}{\tau} \cdot 100\%; \tag{68}$$

$$\text{rse}(\widehat{\tau}) = \frac{\text{sd}(\widehat{\tau})}{\tau} \cdot 100\%, \tag{69}$$

where

$$\text{sd}(\widehat{\tau}) = \left[1000^{-1}\sum_{r=1}^{1000}\left(\tau_r - \bar{\widehat{\tau}}\right)^2\right]^{0.5} \text{ and } \bar{\widehat{\tau}} = 1000^{-1}\sum_{r=1}^{1000}\widehat{\tau}_r;$$

and

$$\text{rrmse}(\widehat{\tau}) = \frac{\left[1000^{-1}\sum_{r=1}^{1000}(\tau_r - \tau)^2\right]^{0.5}}{\tau} \cdot 100\%. \tag{70}$$

For each of these three measures of performance, the closer the value to zero, the better its performance. The relative root mean square error, which simultaneously accounts for the bias and variance, can be thought of as the best measure of performance. For a given scenario, an estimator with the smallest relative root mean square error, among several estimators under consideration, can be declared as having the best overall performance.

For the classification-based dual system estimator we also report the percentage (relative to the true population size) of unique *records* that constitute record pairs classified as possible links. This figure only gives some general idea of the amount of clerical work needed. For instance, it cannot tell how many record pairs must be clerically checked for each such unlinked record. In real applications, there may be a situation where checking just a handful of pairs is sufficient, but it is also passible to have a situation where a more thorough and labour intensive investigation is needed. On the other hand, there may be cases when a pattern with all disagreements is put into a possible link category and the number of records in this pattern is often large. Obviously, there is virtually no way to clerically resolve such cases and therefore they would not be contributing to clerical resolution effort.

There are several aspects of the classification-based dual system estimator to mention. Unlike the no-classification approach where the outcome depends on the input data and model specification, the classification method also depends on the choice of acceptance and rejection thresholds. Hence, given the data and fixing a model, there are still multiple choices of thresholds, each choice leading to a different relative root mean square error and the clerical effort required to resolve possible links. This means that there is no firm and fair reference to compare the classification and no-classification focused methods. Therefore, one needs to be cautious about making statements in favour or against the no-classification methods when comparing the method involving classification. In the main body of this thesis we use the acceptance 'probability' (or rate) $\pi_{\text{fn}} = 10^{-2}$ for the false negatives and $\pi_{\text{fp}} = 10^{-4}$ for the false positives; see Section 2.2.2 on how the decision is made that a specified threshold is achieved. This choice of thresholds is quite arbitrary, but it results in accepting a majority of patterns where at least three linkage variables agree and rejecting a majority of patterns where

at most one linkage variable agrees across wide range of scenarios. In the Appendix B, we present results with the acceptance rate of 10^{-3} for false negatives and 10^{-6} for false positives, which reduce the error by increasing the amount of clerical work. Another aspect is related to 1-to-1 constraining. Once the constraint enforced through the solution of the assignment problem, the estimated rates of false negatives and false positives corresponding to a chosen threshold are meaningless even if they are correctly estimated. This follows from the fact that the model estimates are for the pairs, and 1-to-1 resolution results in most pairs being discarded in favour of the most likely assignment given the parameter estimates. Overall, the need for thresholds and their arbitrariness is a nuisance when making a comparison with the no-classification methods, and is a weakness of classification-based frameworks in practice.

The simulated annealing approach has several tuning parameters. The values of these parameters when assessing the practical performance of the estimators are following. The cooling rate is 0.995, the target temperature is 10^{-6}, the number of iterations of the Metropolis algorithm at each value of the temperature parameter is 150, the interval from which the values of parameter π are drawn is $\left[7 \cdot 10^{-5}, 3.3 \cdot 10^{-2}\right]$, the interval from which the values of $\mu_k, \mu_{k|j}$ are drawn is $[0.55, 0.99]$ and the interval from which the values of $\nu_k, \nu_{k|j}$ are drawn is $\left[5 \cdot 10^{-8}, 0.55\right]$. The tuning parameters for the verification of the theoretical results are the same, except that the cooling rate is 0.999, which aims to make the differences between the runs with different random starts very small.

7.4 Worked-out examples of no-classification estimators

Before we analyse the results of our simulation study, we present two worked-out examples of application of the linkage free dual system estimator (Section 4.1) and its modified version that accounts for the 1-to-1 constraint (Section 4.2). The data used in these examples are simulated as described in Section 7.2. We generated two input data sets, one per example: the first one aims at between-variables independence in both the set of matches and the set of non-matches (Section 7.4.1), the second aims at between-variables independence in the set of matches and association between the linkage variables v_1 and v_2 in the set of non-matches (Section 7.4.2). When estimating the parameters of a mixture-like model we use the closest mixture-like parameterization to the aimed data structure.

Our simulated population has the size $\tau = 500$ in both examples and the underlying parameters of the distributions of the population attributes are as presented in Section 7.2. The coverage probabilities are 0.9 and the vector of the probabilities of errors is $(0.015, 0.05, 0.05, 0.05)^T$ in both surveys.

Each example is accompanied by two tables. The first table is related to the linkage free dual system estimator. It displays the true frequencies (or numbers) of the record pairs in each of the comparison patterns as well as the corresponding estimates of these frequencies. These frequency estimates are the product of the observed number of record pairs w and the estimated parameters of an appropriate mixture-like model. The parameter estimates are obtained with the simulated annealing algorithm. In the table each row is associated with a particular comparison pattern $\boldsymbol{\gamma}_p$ and the columns are as follows: the first column is the index p of a comparison patter; the second column is the comparison pattern $\boldsymbol{\gamma}_p$ itself; the third column is the observed frequency f_p of a comparison pattern; the fourth column is the true unobserved frequency $f_{p,\mathcal{M}}$ of matching pairs with comparison resulting in a pattern $\boldsymbol{\gamma}_p$; the

fifth column is the true unobserved frequency $f_{p,\mathcal{U}}$ of non-matching pairs with comparison resulting in a pattern $\boldsymbol{\gamma}_p$; the sixth column is the estimated frequency $\widehat{f}_p$ of the corresponding pattern; the seventh column is the estimated number $\widehat{f}_{p,\mathcal{M}}$ of matches and, finally, the last column is the estimated number $\widehat{f}_{p,\mathcal{U}}$ of non-matches. While the linkage free dual system estimator uses only an estimate of π, providing the estimated frequencies alongside the individual parameter estimates gives a broader picture. For instance, simulated annealing searches for the parameter estimates that minimize the distance between f_p and $\widehat{f}_p$, and the table gives a feeling of how close or far apart these two quantities are.

The second table is related to the modified linkage free dual system estimator, but also contains some columns relevant to the classification-based approach. The comparison patterns in both tables are indexed in the same way. The values in the second table are sorted by the 'likelihood' of a pattern to be a link, from the highest to the lowest, based on the pseudo log-odds (50), denoted $\widehat{l}_p$. The first column is the index of a comparison pattern; the second column is the comparison pattern $\boldsymbol{\gamma}_p$ itself; the third column is the frequency $\widetilde{f}_p$ of the corresponding patter after applying the 1-to-1 constraint; the fourth column, $\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$, is the ratio of the contribution of the matches in the p^{th} comparison pattern to the proportion of the p^{th} pattern among all patterns (49); the fifth column is the estimated frequency $\widehat{f}_{p,\mathcal{M}}$ of the links without the 1-to-1 constraint; the sixth column is the modified linkage free estimate $\widetilde{m}_{c,p}$ of the number of links in the corresponding pattern (51). The remaining columns are related to the classification-based approach. The seventh column is the estimated 'probability' $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$ of observing a certain comparison pattern given a record pair is a match; the eighth column is the estimated 'probability' $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$ of observing a certain comparison pattern given a record pair is a non-match; the ninth column is the value of pseudo log-odds ratio $\widehat{l}_p$. Note, that this log-odds is also used in the modified estimator and is based on the $\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$, rather than on $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$ and $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$ as in Section 2.2.2. However, it does not affect the ordering in the examples considered. Finally, the last column, $\text{d}(\boldsymbol{\gamma}_p)$, is the classification decision, where 'l' stands for 'link', 'u' stands for 'unresolved' (possible) link and 'n' stands for 'non-link'. Recall, that in this thesis the classification-based approach classifies cases after constraining for the 1-to-1 matches. Note, that the number of decimal places vary between the columns since the values in some of columns tend to be small.

7.4.1 Between-variables independence in both sets of matches and non-matches

In this example the simulated data aim at between-variables independence in the both sets of matches and non-matches. The observed frequencies are displayed in the third column of Table 8. In this case $n_1 = 453$, $n_2 = 452$ and the number of observed record pairs is $w = 204756$. The closest possible identifiable mixture-like parameterization of the simulated data is the model of between-variables independence of the comparisons in the set of matches and non-matches, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1 - \pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

The simulated annealing algorithm searches for the parameter estimates that minimizes the value of the chi-squared statistic (39). In this case the achieved value of the chi-squared statistic is 4.95 and the corresponding parameter estimate are (with varying number of decimal places) as follows: $\widehat{\pi} = 0.001986$, $\widehat{\mu}_1 = 0.956$, $\widehat{\mu}_2 = 0.903$, $\widehat{\mu}_3 = 0.927$, $\widehat{\mu}_4 = 0.917$, $\widehat{\nu}_1 = 0.000124$, $\widehat{\nu}_2 = 0.004630$, $\widehat{\nu}_3 = 0.004890$, $\widehat{\nu}_4 = 0.009942$. The linkage free dual system estimate in this case is $\widetilde{\tau} = 1/\widehat{\pi} = 0.001986^{-1} =$

503.52. These estimates can be used to obtain the estimated frequencies $\widehat{f}_{p,\mathcal{M}}$ and $\widehat{f}_{p,\mathcal{U}}$ by plugging the estimates into the corresponding parameterization of a p^{th} pattern to get $\widehat{\pi}_p$ and then computing the related frequency $\widehat{f}_p = w\widehat{\pi}_p$.

Table 8: Between-variables independence in both the set of matches and the set of non-matches: observed data, true and estimated frequencies of comparison patterns in the set of matches and the set of non-matches

p	$\boldsymbol{\gamma}_p$	f_p	$f_{p,\mathcal{M}}$	$f_{p,\mathcal{U}}$	$\widehat{f}_p$	$\widehat{f}_{p,\mathcal{M}}$	$\widehat{f}_{p,\mathcal{U}}$
1	0000	200366	0	200366	200371.18	0.01	200371.17
2	0001	2015	0	2015	2012.26	0.12	2012.14
3	0010	986	0	986	984.77	0.13	984.64
4	0011	9	1	8	11.37	1.48	9.89
5	0100	931	0	931	932.06	0.10	931.97
6	0101	10	1	9	10.45	1.09	9.36
7	0110	7	3	4	5.82	1.24	4.58
8	0111	14	14	0	13.77	13.73	0.05
9	1000	25	0	25	25.12	0.23	24.89
10	1001	2	2	0	2.80	2.55	0.25
11	1010	6	6	0	3.02	2.90	0.12
12	1011	32	32	0	32.12	32.11	0.00
13	1100	1	1	0	2.25	2.14	0.12
14	1101	28	28	0	23.67	23.67	0.00
15	1110	28	28	0	26.94	26.94	0.00
16	1111	296	296	0	298.41	298.41	0.00

Table 9: Between-variables independence in both sets of matches and non-matches: outputs for the modified linkage free dual system estimator and classification-based linkage

p	$\boldsymbol{\gamma}_p$	$\widetilde{f}_p$	$\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$	$\widehat{f}_{p,\mathcal{M}}$	$\widetilde{m}_{c,p}$	$\widehat{\mathrm{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$	$\widehat{\mathrm{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$	$\widehat{l}_p$	$\mathrm{d}(\boldsymbol{\gamma}_p)$
16	1111	296	1.0000	298.41	296.00	0.73349	0.00000	35.54	l
15	1110	28	1.0000	26.94	28.00	0.06621	0.00000	21.53	l
12	1011	32	1.0000	32.11	32.00	0.07894	0.00000	20.34	l
14	1101	28	1.0000	23.67	28.00	0.05818	0.00000	19.84	l
8	0111	14	0.9967	13.73	14.00	0.03374	0.00000	11.40	l
11	1010	6	0.9595	2.90	5.87	0.00713	0.00000	6.33	l
13	1100	1	0.9486	2.14	1.06	0.00525	0.00000	5.83	l
10	1001	2	0.9106	2.55	2.05	0.00626	0.00000	4.64	l
7	0110	3	0.2130	1.24	1.61	0.00305	0.00002	-2.61	u
4	0011	1	0.1300	1.48	1.42	0.00363	0.00005	-3.80	u
6	0101	1	0.1042	1.09	1.08	0.00268	0.00005	-4.30	n
9	1000	0	0.0092	0.23	0.23	0.00057	0.00012	-9.37	n
3	0010	7	0.0001	0.13	0.13	0.00033	0.00482	-17.81	n
5	0100	6	0.0001	0.10	0.10	0.00024	0.00456	-18.31	n
2	0001	4	0.0001	0.12	0.12	0.00029	0.00985	-19.50	n
1	0000	23	0.0000	0.01	0.01	0.00003	0.98053	-33.51	n

The above parameter estimates can be used to produce the modified linkage free dual system estimate of the population size that account for the 1-to-1 match constraint. First, the linkage free dual system estimates are used to compute $\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$ and the pseudo log-odds ratios $\widehat{l}_p$, which are (49) and (50) in Section 4.2, respectively. These log-odds ratios are fed to the assignment algorithm that produces the frequencies $\widetilde{f}_p$ which reflect the 1-to-1 constraint. The next step is for the composite estimator (51) to use $\widetilde{f}_p$, $\widehat{f}_{p,\mathcal{M}}$ and $\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$ to estimate the number of matches in each of the

comparison pattern. By summing these estimates, we obtain the estimated number of matches under the 1-to-1 match constraint. We then use this estimate with (53) to produce the modified linkage free dual system estimate of the population size, which is $\widetilde{\tau}_c = 497.52$.

The classification-based approach uses the estimates $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$ and $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$ to classify the records, or more precisely, patterns as links, non-links and possible links. Our chosen acceptance thresholds are 0.01 for false positives and 0.0001 for false negatives. We start summing the values of $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$, as presented in Table 9, from top to bottom. As long as the cumulative sum up to and including a given pattern is smaller or equal to the threshold, we accept the pattern as a link. Then, we sum the values of $\widehat{\text{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$ from the bottom to top. As long as this cumulative sum at a given pattern is larger or equal to the threshold, we declare the pattern as a non-link. Everything in between is unresolved and needs clerical resolution. Recall, that the thresholds are computed with respect to the total number of record pairs, whereas after the accounting for 1-to-1, the majority of pairs are dropped from the process. Since we fixed the thresholds in advance for all scenarios (rather than doing it once the parameters are estimated), there may be situations where no pattern is declared as a potential link while some patterns are simultaneously declared as links and non-links. In this case, such overlapping classifications are clerically revised.

7.4.2 Between-variables independence in the set of matches, association between v_1 and v_2 in the set of non-matches

In this example the simulated data aim at between-variables independence of the comparison outcomes in the set of matches and association in comparison outcomes of the first and second linkage variables in the set of non-matches. The estimation model used is the closest identifiable mixture-like parameterization to the simulated data, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$.

Table 10: Between-variables independence in the set of matches, association between the first and second variable in the set of non-matches: observed data, true and estimated frequencies of comparison patterns in the set of matches and the set of non-matches

p	$\boldsymbol{\gamma}_p$	f_p	$f_{p,\mathcal{M}}$	$f_{p,\mathcal{U}}$	$\widehat{f}_p$	$\widehat{f}_{p,\mathcal{M}}$	$\widehat{f}_{p,\mathcal{U}}$
1	0000	199663	0	199663	199674.19	0.03	199674.16
2	0001	1980	0	1980	1979.91	0.29	1979.62
3	0010	978	0	978	975.34	0.27	975.07
4	0011	12	2	10	12.00	2.33	9.67
5	0100	1092	0	1092	1087.39	0.37	1087.02
6	0101	12	3	9	14.00	3.22	10.78
7	0110	5	2	3	8.30	2.99	5.31
8	0111	29	29	0	26.21	26.15	0.05
9	1000	381	0	381	380.17	0.34	379.83
10	1001	11	4	7	6.75	2.99	3.77
11	1010	2	1	1	4.62	2.77	1.85
12	1011	27	27	0	24.25	24.23	0.02
13	1100	527	4	523	527.71	3.83	523.88
14	1101	38	29	9	38.70	33.51	5.19
15	1110	41	39	2	33.60	31.05	2.56
16	1111	267	267	0	271.86	271.83	0.03

The observed frequencies of the comparison patterns are displayed in Table 10. Note the difference

from the case of between-variables independence presented in Table 8. Namely, the frequencies of the patterns with agreements on both the address and surname variables (first two variables) are substantially larger. The size of S_1 is $n_1 = 441$ and the size of S_2 is $n_2 = 465$, the number of observed record pairs is 205065. The value of the chi-squared statistic achieved in this case is 9.60. The corresponding parameter estimates are $\widehat{\pi} = 0.001981$, $\widehat{\mu}_1 = 0.912$, $\widehat{\mu}_2 = 0.918$, $\widehat{\mu}_3 = 0.890$, $\widehat{\mu}_4 = 0.897$. We report the joint parameter estimates for the associated comparisons: $\widehat{\nu}_{1,1} = 0.000618$, $\widehat{\nu}_{1,0} = 0.324592$, $\widehat{\nu}_{0,1} = 0.001281$. Finally, $\widehat{\nu}_3 = 0.004860$, $\widehat{\nu}_4 = 0.009816$. The linkage free dual system estimate in this case is $\widetilde{\tau} = 1/\widehat{\pi} = 0.001981^{-1} = 504.80$

The derivation of the modified linkage free dual system estimate is the same as in the previous example in Section 7.4.1. The modified estimate is 502.33. The corresponding data are displayed in Table 11. Note that both examples considered have the estimates close to the true population value. However, in general, they can be quite spread out around the parameter τ.

Table 11: Observed data, true and estimated frequencies of comparison patterns in the sets of matches and non-matches: outputs for the modified linkage free dual system estimator and classification-based linkage

p	$\boldsymbol{\gamma}_p$	$\widetilde{f}_p$	$\widehat{r}(\mathcal{M} \mid \boldsymbol{\gamma}_p)$	$\widehat{f}_{p,\mathcal{M}}$	$\widetilde{m}_{c,p}$	$\widehat{\mathrm{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{M})$	$\widehat{\mathrm{pr}}(\boldsymbol{\gamma}_p \mid \mathcal{U})$	$\widehat{l}_p$	$\mathrm{d}(\boldsymbol{\gamma}_p)$
16	1111	267	0.9999	271.83	267.00	0.66923	0.00000	18.56	l
12	1011	27	0.9992	24.23	27.00	0.05966	0.00000	14.37	l
8	0111	29	0.9980	26.15	28.99	0.06439	0.00000	12.42	l
15	1110	39	0.9239	31.05	38.39	0.07643	0.00001	4.99	l
14	1101	28	0.8658	33.51	28.74	0.08249	0.00003	3.73	l
11	1010	1	0.5987	2.77	1.71	0.00681	0.00001	0.80	l
10	1001	4	0.4423	2.99	3.43	0.00735	0.00002	-0.46	l
7	0110	2	0.3601	2.99	2.63	0.00735	0.00003	-1.15	l
6	0101	3	0.2302	3.22	3.17	0.00794	0.00005	-2.41	u
4	0011	2	0.1943	2.33	2.27	0.00574	0.00005	-2.84	u
13	1100	5	0.0073	3.83	3.84	0.00942	0.00256	-9.84	n
9	1000	0	0.0009	0.34	0.34	0.00084	0.00186	-14.03	n
5	0100	4	0.0003	0.37	0.37	0.00091	0.00531	-15.98	n
3	0010	8	0.0003	0.27	0.27	0.00066	0.00476	-16.41	n
2	0001	9	0.0001	0.29	0.29	0.00071	0.00967	-17.68	n
1	0000	13	0.0000	0.03	0.03	0.00008	0.97564	-31.24	n

7.5 Simulations assessing practical performance

In this section we demonstrate the results of the simulation study assessing practical application of the linkage free and modified linkage free dual system estimators. Each subsection below corresponds to a specific type of a simulation model. For instance, a model aiming at the between-variables independence of comparison outcomes. Recall, we prefer saying 'aiming at' some between-variables independence / dependence relationships of the comparison outcomes since the data are predominantly generated in hierarchical way using the 'first principles' approach rather than employing some parametric distribution that guarantees such relationships. The parametric approach is only well-justified and used to generate between-variables associations in the set of matches. When discussing the outputs of each simulation model type, we follow this order. First, for a given type of the simulated data, an estimation model that has the closest parameterization to the aimed data structure is discussed. This is

followed by presenting either a less complex or a more complex estimation model, which depends on the simulated outcome. Some results are moved to Appendix B.1.

There are several cases where we compare the estimates of τ obtained without blocking against those obtained with the averaging blocking (in a way that is achievable in practice); however, the majority of the outputs are for applications without blocking (at the population level). Tables with the results without blocking have the following structure. The first three columns contain these simulation parameters: the population size τ; the response probability for two surveys $\pi_j = \pi_1 = \pi_2$; the probability of the error recording the values of linkage variables. To save space, here the value $\xi = 0.01$ corresponds to the vector of errors $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = (0.003, 0.01, 0.01, 0.01)^T$, $\xi = 0.05$ to $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = (0.015, 0.05, 0.05, 0.05)^T$, and $\xi = 0.10$ to $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = (0.03, 0.1, 0.1, 0.1)^T$. The fourth column, labelled $\mathrm{C_{FS}}$, contains the percentage of clerically resolved *records* relative to the population size in the regular classification-based approach with the acceptance rate of 10^{-2} for the false negatives and 10^{-4} for the false positives. The remaining columns are divided into three blocks for the relative bias, relative standard error and relative root mean square error. All three statistics are reported in percentages for each of the estimators: the dual system estimator with the regular classification-based linkage, $\widehat{\tau}_{\mathrm{FS}}$; the linkage free dual system estimator, $\widetilde{\tau}$; the modified linkage free dual system estimator, $\widetilde{\tau}_c$; the dual system estimator with perfect linkage, $\widehat{\tau}$. Tables comparing the estimation without blocking to the estimation with the averaging blocking contain the simulation parameters as well as the relative bias, relative standard error and relative root mean square error in percentages for two estimators: the linkage free dual system estimator without blocking, $\widetilde{\tau}$; and the linkage free dual system estimator with the averaging blocking, $\widetilde{\tau}_G$.

7.5.1 Between-variables independence in both sets of matches and non-matches

We start with the analysis of the estimates obtained with the simulated data aiming at between-variables independence in both the sets of matches and non-matches. The first estimation model we consider is the closest possible mixture-like parameterization of the simulated data, that is the model of between-variables independence of the comparisons in the set of matches and non-matches, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1 - \pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$. This model is identifiable (see Section 5.2.1). The results are presented in Table 12.

Certain characteristic trends observed in this between-variables independence case repeat in the majority of simulation and estimation scenarios explored in this thesis, provided that estimation models are identifiable. In particular, the relative bias and the relative standard error of the linkage free, modified linkage free and classification-based dual system estimators increase when either one of the following occurs (or both occur simultaneously): when the probabilities of making errors recording the values of population attributes increase, or when the coverage probabilities of the surveys decrease. Obviously, the relative root mean square error also increases in these situations. Hence, the worst performance is expected in the scenarios with the highest errors and lowest coverage, which seems quite intuitive. The relative bias and relative standard error decrease, and so does the relative root mean square error as a result, when the population size increases. Recall, however, that following the discussion in Section 3.3, we know that the linkage free dual system estimator is not anticipated to

behave well as the population size tends to infinity. The dual system estimator with perfect linkage, of course, does not vary with the errors in linkage variables (except for some random fluctuations in the simulated outcomes) and depends only on the population size and coverage probabilities.

When the probabilities of making errors recording the values of population attributes are small, then the linkage free, modified linkage free and the classification-based dual system estimators have very similar performance in terms of the relative root mean square error to performance of the dual system estimator with perfect linkage. As the probabilities of errors increase, the gap in performance between the dual system estimator with perfect linkage and the rest of the estimators become noticeable.

Table 12: Simulated data: between-variables independence in sets of both matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.08	0.06	-0.17	0.02	0.01	1.12	1.23	1.14	1.11	1.12	1.24	1.14	1.11
	0.9	0.05	0.67	1.04	2.33	0.35	0.09	1.36	2.13	1.34	1.09	1.71	3.16	1.38	1.10
		0.10	1.92	1.62	3.90	0.58	-0.02	1.57	3.25	1.69	1.12	2.25	5.07	1.79	1.12
		0.01	1.73	0.04	-0.25	0.01	0.01	2.63	2.68	2.64	2.63	2.63	2.69	2.64	2.63
250	0.8	0.05	0.52	1.01	2.09	0.26	0.10	2.85	3.18	2.77	2.68	3.02	3.81	2.78	2.68
		0.10	1.95	1.93	4.80	0.84	0.20	3.07	4.67	3.12	2.73	3.63	6.70	3.23	2.74
		0.01	1.25	0.25	-0.29	0.19	0.20	4.73	4.72	4.74	4.72	4.73	4.73	4.74	4.72
	0.7	0.05	0.48	1.06	2.26	0.39	0.21	4.64	5.02	4.63	4.54	4.76	5.50	4.64	4.54
		0.10	2.11	1.95	5.88	0.91	0.16	4.95	6.51	4.97	4.62	5.32	8.77	5.05	4.63
		0.01	1.97	0.04	-0.07	0.00	0.00	0.77	0.82	0.78	0.76	0.77	0.82	0.78	0.76
	0.9	0.05	0.43	1.06	1.35	0.18	0.00	1.00	1.49	1.06	0.78	1.46	2.01	1.07	0.78
		0.10	2.58	1.51	1.78	0.35	0.03	1.00	2.04	1.30	0.74	1.81	2.71	1.34	0.74
		0.01	1.64	0.22	0.09	0.19	0.18	1.87	1.92	1.88	1.87	1.89	1.93	1.89	1.88
500	0.8	0.05	0.41	1.06	1.70	0.27	0.07	2.02	2.44	2.06	1.90	2.29	2.97	2.07	1.90
		0.10	3.12	1.54	2.34	0.46	0.06	2.03	2.99	2.25	1.82	2.55	3.80	2.30	1.82
		0.01	1.34	0.13	-0.05	0.08	0.09	3.21	3.24	3.22	3.21	3.22	3.24	3.22	3.21
	0.7	0.05	0.39	1.01	1.76	0.22	0.12	3.46	3.70	3.46	3.36	3.60	4.10	3.47	3.36
		0.10	3.25	1.61	3.32	0.66	0.24	3.39	4.44	3.56	3.25	3.75	5.54	3.62	3.25
		0.01	1.88	0.04	0.00	0.01	0.01	0.57	0.61	0.59	0.57	0.58	0.61	0.59	0.57
	0.9	0.05	0.33	1.04	0.62	0.07	0.00	0.69	1.03	0.78	0.54	1.25	1.20	0.78	0.54
		0.10	3.57	1.34	0.77	0.17	0.01	0.73	1.47	1.05	0.54	1.53	1.66	1.06	0.54
		0.01	1.65	0.08	-0.01	0.03	0.04	1.37	1.39	1.37	1.37	1.37	1.39	1.37	1.37
1000	0.8	0.05	0.34	0.93	0.80	0.04	-0.01	1.38	1.64	1.42	1.32	1.66	1.82	1.42	1.32
		0.10	4.66	1.20	1.10	0.15	0.04	1.46	1.99	1.66	1.38	1.89	2.27	1.67	1.38
		0.01	1.46	0.22	0.14	0.17	0.18	2.41	2.42	2.42	2.41	2.42	2.42	2.42	2.41
	0.7	0.05	0.33	0.69	0.91	-0.08	-0.05	2.36	2.63	2.42	2.32	2.46	2.78	2.42	2.33
		0.10	5.25	0.98	1.50	0.10	0.17	2.36	2.99	2.58	2.32	2.55	3.35	2.59	2.33

While the classification-based approach is not our estimator of interest, it is worth examining its performance. Recall, that this performance is somehow arbitrary due to the chosen acceptance and rejection thresholds. The relative root mean square error of the classification-based dual system estimator is up to 2.8 times of the dual system estimator with perfect linkage. The error is around 1.5 times that of the perfect linkage based estimator across the majority of scenarios. The main source of the increased error in the classification-based method is bias, but there is also a slight increase in the standard error. For the chosen thresholds, the bias is positive. This indicates that the false negative errors dominate the false positives. In other words, there are more matches in the comparison patterns

classified as non-links than there are non-matches in the patterns classified as links. With the selected thresholds and all caveats mentioned in Section 7.3, the amount of clerical resolution is reasonably low, which is largely due to the 1-to-1 constraint leaving a small number of records to resolve. Observe, that for the fixed population size and coverage probabilities the amount of clerical work does not gradually increase as the error recording the values of population attributes increases for the chosen thresholds (it does increase for different choices of thresholds; see Appendix B). This may seem counter-intuitive at first glance, but can be easily explained.

When the errors in recording the values of population attributes are small, the majority of matches are concentrated in the patterns with agreements on all or nearly all linkage variables, while most of the non-matches are concentrated in the patterns with no agreements or agreements on a few variables only. There are several aspects contributing to the observed behaviour. Consider the sequence of patterns ordered by the ratio of 'likelihoods'. Since there are only 2^K comparison patterns, it is difficult to get a fine classification in the sense that both the desired thresholds are achieved and no substantial changes in the results occur if one more comparison pattern is classified as a link or non-link instead of being classified as a possible link. Say, for a small error, the following comparison patterns $(1,1,1,1)^T, (1,1,1,0)^T, (1,1,0,1)^T$ are accepted as links because the specified threshold for the false positives is achieved, while $(0,1,1,1)^T$ is classified as a possible link, despite containing very few non-matches, but containing reasonably many matches after application of the 1-to-1 linkage constraint. On the other end of the sequence of ordered patterns, the following comparison patterns $(0,0,0,0)^T, (1,0,0,0)^T, (0,1,0,0)^T, (0,0,1,0)^T, (0,0,0,1)^T, (1,1,0,0)^T, (0,1,1,0)^T$ are classified as non-links. Because the error probability is low and the 1-to-1 constraint is used, the pattern $(0,1,1,1)^T$ may make a substantial contribution to the number of records classified as possible links among all the patterns with such classification. When the error gradually increases, more matches start to appear in the patterns with agreements on some but not all linkage variables, while there is little to no effect on the true non-matches. The increasing error affects how much matches and non-matches are mixed in each pattern as well as how many patterns end up classified as possible links. The latter is again not a smooth process, and some substantial shifts can appear in the number of records classified as possible links. Thus, at some point the discussed pattern $(0,1,1,1)^T$ starts getting classified as a link. Therefore, the number of records for clerical resolution drops. When the error keeps increasing, more matches and non matches start appearing in the patterns like $(1,1,0,0)^T, (0,1,1,0)^T$ that gradually start to be classified as possible links and the amount of clerical work increases. This is what we observe for our combination of increment in the errors and acceptance / rejection thresholds. In short, for a sufficiently large increase in the errors, we will definitely see the increase in clerical effort. However, it is possible to find such increments in the errors, for which this effort will not increase or may even decrease.

Back to our key estimator of interest, we can see that the linkage free dual system estimator demonstrates a good performance given it is a fully automated approach. Unsurprisingly, its performance is inferior to the performance of other estimators considered in this thesis. Its relative root mean square error can be up to 3.6 times that of the dual system estimator with perfect linkage in the hardest of scenarios. However, the error rarely exceeds 1.8 times what can be achieved in the perfect case

and 1.3 times the error of the classification-based approach with the chosen thresholds. This seems a reasonable trade-off between the need to resolve some records clerically and the loss of accuracy. Both the bias and variance contribute to the increase in the relative root mean square error for this no-classification approach. The increased variance of the linkage free dual system estimator compared to the classification approach is largely explainable by the absence of the 1-to-1 constrain in the basic no-classification approach. As discussed in Section 3.3, the linkage free dual system estimator is not consistent and thus most likely biased. While the empirical results confirm that the estimator is biased, we do not have the means to explain the bias incurring mechanism in practical applications except the factors related to mixture-like representation of the record linkage data (Section 3.2).

Table 13: Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
250	0.9	0.01	1.80	0.07	-0.13	0.05	0.01	1.12	1.23	1.15	1.11	1.13	1.23	1.15	1.11
		0.05	0.65	1.23	2.49	0.52	0.09	1.47	2.14	1.38	1.09	1.92	3.28	1.48	1.10
		0.10	2.73	1.67	4.41	0.95	-0.02	1.62	3.39	1.84	1.12	2.33	5.56	2.08	1.12
	0.8	0.01	1.42	0.06	-0.21	0.04	0.01	2.63	2.68	2.65	2.63	2.63	2.69	2.65	2.63
		0.05	0.51	1.26	2.26	0.48	0.10	2.89	3.19	2.80	2.68	3.15	3.91	2.84	2.68
		0.10	2.53	2.06	5.35	1.30	0.20	3.13	4.84	3.27	2.73	3.74	7.21	3.52	2.74
	0.7	0.01	0.98	0.26	-0.24	0.22	0.20	4.73	4.72	4.74	4.72	4.74	4.72	4.75	4.72
		0.05	0.44	1.45	2.46	0.71	0.21	4.74	5.05	4.70	4.54	4.96	5.62	4.75	4.54
		0.10	2.57	2.19	6.49	1.50	0.16	5.05	6.58	5.06	4.62	5.50	9.24	5.27	4.63
500	0.9	0.01	1.93	0.04	-0.04	0.02	0.00	0.77	0.82	0.78	0.76	0.77	0.82	0.78	0.76
		0.05	0.44	1.08	1.49	0.30	0.00	1.00	1.51	1.07	0.78	1.47	2.12	1.12	0.78
		0.10	2.82	1.53	2.20	0.59	0.03	1.00	2.16	1.37	0.74	1.83	3.08	1.49	0.74
	0.8	0.01	1.58	0.22	0.12	0.22	0.18	1.87	1.93	1.89	1.87	1.89	1.93	1.90	1.88
		0.05	0.42	1.10	1.84	0.41	0.07	2.03	2.48	2.09	1.90	2.31	3.09	2.13	1.90
		0.10	3.36	1.58	2.79	0.74	0.06	2.06	3.08	2.30	1.82	2.60	4.15	2.42	1.82
	0.7	0.01	1.31	0.13	-0.01	0.11	0.09	3.21	3.24	3.22	3.21	3.22	3.24	3.23	3.21
		0.05	0.39	1.07	1.91	0.38	0.12	3.46	3.73	3.48	3.36	3.62	4.19	3.50	3.36
		0.10	3.35	1.69	3.81	1.01	0.24	3.44	4.55	3.67	3.25	3.83	5.94	3.80	3.25
1000	0.9	0.01	1.89	0.04	0.01	0.03	0.01	0.57	0.61	0.59	0.57	0.58	0.61	0.59	0.57
		0.05	0.34	1.04	0.72	0.13	0.00	0.70	1.04	0.79	0.54	1.25	1.26	0.80	0.54
		0.10	3.60	1.36	1.08	0.31	0.01	0.73	1.54	1.07	0.54	1.54	1.88	1.12	0.54
	0.8	0.01	1.66	0.08	0.01	0.05	0.04	1.37	1.39	1.37	1.37	1.37	1.39	1.37	1.37
		0.05	0.35	0.93	0.92	0.14	-0.01	1.38	1.65	1.44	1.32	1.66	1.89	1.44	1.32
		0.10	4.58	1.23	1.43	0.32	0.04	1.46	2.06	1.68	1.38	1.91	2.51	1.71	1.38
	0.7	0.01	1.47	0.22	0.16	0.20	0.18	2.41	2.42	2.41	2.41	2.42	2.43	2.42	2.41
		0.05	0.34	0.69	1.04	0.03	-0.05	2.37	2.64	2.44	2.32	2.47	2.84	2.44	2.33
		0.10	5.16	1.03	1.87	0.31	0.17	2.36	3.13	2.65	2.32	2.58	3.64	2.67	2.33

The modified linkage free dual system estimator demonstrates a remarkable performance given it is a fully automated approach. Recall that this estimator is the no-classification method combining the linkage free dual system estimator and the fact (or assumption) that the matches obey the 1-to-1 constraint. The modified version outperforms substantially the basic version of the linkage free dual system estimator both in terms of the bias and variance across all scenarios. Actually, the modified version of the no-classification method has better performance than the classification-based approach (with chosen thresholds) across the majority of scenarios. This outperformance of the classification-

based method by the modified no-classification estimator is due to the absence of classification and accurate estimation of the parameters when the model is correctly specified. That is, the error incurred by accepting all the cases belonging to a particular comparison pattern after applying the 1-to-1 constraint exceeds the estimation error of the modified no-classification approach. The performance of the modified linkage free dual system estimator is not too far away from the performance of the perfect dual system estimator in the case of between-variables independence of comparisons in both the set of matches and the set of non-matches. The error of this no-classification estimator is on average 1.17 the estimator with perfect linkage. In summary, the modified linkage free dual system estimator for this type of relationship between the variables is the best choice among all the estimators available in practice.

Table 14: Single block vs averaged blocking: between-variables independence in both sets of matches and non-matches

			RB		RSE		RRMSE	
τ	π_j	ξ	$\widetilde{\tau}$	$\widetilde{\tau}_G$	$\widetilde{\tau}$	$\widetilde{\tau}_G$	$\widetilde{\tau}$	$\widetilde{\tau}_G$
		0.01	-0.17	-0.25	1.23	1.18	1.24	1.20
	0.9	0.05	2.33	2.11	2.13	1.91	3.16	2.85
		0.10	3.90	4.08	3.25	2.98	5.07	5.05
		0.01	-0.25	-0.16	2.68	2.49	2.69	2.50
250	0.8	0.05	2.09	1.96	3.18	3.30	3.81	3.84
		0.10	4.80	4.89	4.67	4.37	6.70	6.56
		0.01	-0.29	-0.45	4.72	4.57	4.73	4.59
	0.7	0.05	2.26	2.23	5.02	5.02	5.50	5.49
		0.10	5.88	5.81	6.51	6.27	8.77	8.55
		0.01	-0.07	-0.01	0.82	0.85	0.82	0.85
	0.9	0.05	1.35	1.64	1.49	1.54	2.01	2.25
		0.10	1.78	1.90	2.04	1.83	2.71	2.63
		0.01	0.09	-0.09	1.92	1.94	1.93	1.94
500	0.8	0.05	1.70	1.97	2.44	2.35	2.97	3.06
		0.10	2.34	2.48	2.99	2.79	3.80	3.73
		0.01	-0.05	0.21	3.24	3.19	3.24	3.19
	0.7	0.05	1.76	2.10	3.70	3.80	4.10	4.34
		0.10	3.32	3.36	4.44	4.29	5.54	5.45
		0.01	0.00	-0.01	0.61	0.61	0.61	0.61
	0.9	0.05	0.62	0.84	1.03	0.98	1.20	1.29
		0.10	0.77	0.94	1.47	1.11	1.66	1.45
		0.01	-0.01	0.06	1.39	1.34	1.39	1.34
1000	0.8	0.05	0.80	1.22	1.64	1.67	1.82	2.07
		0.10	1.10	1.18	1.99	1.76	2.27	2.11
		0.01	0.14	0.12	2.42	2.26	2.42	2.26
	0.7	0.05	0.91	1.46	2.63	2.60	2.78	2.98
		0.10	1.50	1.59	2.99	2.77	3.35	3.19

Table 13 contains the outputs obtained by fitting the model of between-variables independence in the set of matches and association between two variables in the set of non-matches, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$, to the data aiming at between-variables independence in both the matches and non-matches. When an identifiable, but more complex than needed, estimation model is specified, the increase in the relative root mean square error of the no-classification approaches is driven not only by increased variance, but also by increased bias. Increased variance in this case is

what is expected in general statistical model-based estimation applications. A possible explanation for increased bias is related to the fact that we are dealing with the unobservable data in the tables of matches and non-matches. So that fitting a model to random fluctuations incurs the bias in the estimate of the parameter π, rather than just increasing the variance. Meanwhile, the errors for the classification based approach are very similar to the correct model case.

We complete the discussion of the between-variables independence model by comparing the results obtained without blocking, $\widetilde{\tau}$, to those with the practically achievable averaging blocking, $\widetilde{\tau}_G$. Results are displayed in Table 14. Such averaging blocking requires the information from the address or sampling frame in order to construct the blocks of approximately equal size. As we can see, there is no clear-cut evidence that would allow us to say that one approach performs better than the other. Neither the bias nor variance is considerably lower or higher consistently for the averaging blocking approach comparing to the estimation at the population level. Tentatively, one can say that the averaging blocking tends to produce slightly better results judging by the relative root mean square error. Certainly, the linkage free dual system estimator with the averaging blocking has better statistical justification than the estimator applied at the population level. However, it is questionable if the observed gains are worth additional effort creating the blocks of approximately the same size in the case of independence.

7.5.2 Between-variables independence in the set of matches, association between v_1 and v_2 in the set of non-matches

The next set of results corresponds to the data simulation model that aims at the between-variables independence of comparison outcomes in the set of matches and association in comparison outcomes of the first and second linkage variables in the set of non-matches. We start with the estimates produced by the model that has the closest identifiable mixture-like parameterization to the simulated data, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$.

The results are tabulated in Table 15. All the main trends are similar to those observed in the case of between-variables independence in Section 7.5.1. However, once the between-variables association of the comparison outcomes is present in the set of non-matches, the relative root mean square error increases in both no-classification estimators as well as in the classification-based estimator. The approach with classification can have the error up to 3 times of the perfect dual system estimator's in this case and is roughly 1.5 times the perfect case in the majority of scenarios. The amount of clerical resolution needed is also larger than in the case of between-variables independence.

As for the linkage free dual system estimator and its modified version, the increase in the relative root mean square error is mainly explained by the increased variance. This is not surprising and is similar to what one usually observes in regular statistical models. Unlike the case of between-variables independence, here the relationships in the data are more complicated so that the simulation outcomes are more variable while the estimation model is also more complex. Hence, the increased variability. What looks like striking behaviour at first glance is the reduction in the relative bias compared to the case of independence. However, it is unlikely to be a real bias reduction. It is most likely that this between-variables dependence model is slightly negatively biased; see the corresponding results

for a data-conforming theoretical version in Table 27. So that the tendency of the no-classification estimators in practical applications to have positive bias is counterbalanced by the negative bias in this between-variables dependence model. We will come back to other aspects related to the bias in a paragraph below. The basic linkage free dual system estimator has again the highest error among the all estimators considered. It demonstrates reasonable performance given it is fully automated, does not use 1-to-1 constraining and has to deal with the between-variables association of the comparison outcomes. Its modified version, however, demonstrates a remarkable performance once again. The modified linkage free dual system estimator either slightly outperforms or slightly falls behind the classification-based approach. The relative root mean square error of the modified no-classification approach is between 0.67 and 1.27 times the classification-based one. The error of the modified no-classification estimator is between 1.02 and 2.7 times the dual system estimator with perfect linkage.

Table 15: Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.71	0.51	-0.26	0.18	0.03	1.28	1.96	1.30	1.05	1.38	1.97	1.32	1.05
	0.9	0.05	2.18	1.76	2.27	0.56	0.02	1.57	2.86	1.67	1.06	2.36	3.65	1.76	1.06
		0.10	8.77	1.91	3.60	0.97	0.01	1.81	4.33	2.51	1.12	2.63	5.63	2.69	1.12
		0.01	1.38	0.70	-0.21	0.34	0.15	2.85	3.24	2.88	2.78	2.94	3.25	2.90	2.78
250	0.8	0.05	1.69	1.80	2.27	0.52	0.02	3.08	3.99	3.07	2.73	3.57	4.59	3.11	2.73
		0.10	7.52	2.12	4.39	1.15	0.13	3.26	5.62	3.83	2.70	3.88	7.13	4.00	2.70
		0.01	1.19	0.84	-0.25	0.52	0.32	4.77	5.11	4.87	4.76	4.84	5.12	4.90	4.77
	0.7	0.05	1.43	1.90	2.09	0.53	0.14	5.05	5.63	5.01	4.75	5.40	6.01	5.04	4.75
		0.10	6.18	2.42	5.26	1.33	0.25	5.41	7.50	5.69	4.95	5.92	9.16	5.85	4.95
		0.01	1.90	0.19	0.10	0.23	0.05	0.75	1.37	0.95	0.73	0.77	1.37	0.98	0.73
	0.9	0.05	0.89	1.57	1.10	0.29	0.02	1.08	1.97	1.25	0.76	1.91	2.26	1.28	0.76
		0.10	6.59	1.53	1.68	0.57	-0.03	1.08	2.87	1.85	0.75	1.87	3.33	1.93	0.75
		0.01	1.59	0.22	0.10	0.27	0.08	1.82	2.20	1.92	1.82	1.84	2.20	1.94	1.82
500	0.8	0.05	0.77	1.65	1.44	0.39	0.10	2.11	2.90	2.20	1.89	2.68	3.23	2.23	1.89
		0.10	6.08	1.55	1.98	0.51	-0.01	2.07	3.71	2.68	1.83	2.58	4.21	2.73	1.83
		0.01	1.27	0.25	0.04	0.32	0.09	3.30	3.64	3.39	3.27	3.31	3.64	3.41	3.27
	0.7	0.05	0.64	1.68	1.72	0.42	0.18	3.57	4.27	3.65	3.40	3.94	4.61	3.68	3.40
		0.10	5.57	1.76	2.55	0.61	0.07	3.74	5.16	4.14	3.45	4.13	5.76	4.19	3.45
		0.01	1.76	0.16	0.08	0.14	0.04	0.56	1.01	0.70	0.54	0.58	1.01	0.72	0.54
	0.9	0.05	0.64	1.43	0.60	0.21	-0.01	0.81	1.40	0.95	0.54	1.64	1.53	0.98	0.54
		0.10	5.35	1.38	0.78	0.30	0.01	0.72	1.96	1.37	0.52	1.56	2.10	1.41	0.52
		0.01	1.52	0.09	0.07	0.12	-0.03	1.32	1.60	1.39	1.31	1.32	1.60	1.40	1.31
1000	0.8	0.05	0.60	1.28	0.60	0.13	-0.03	1.42	2.01	1.53	1.27	1.91	2.10	1.54	1.27
		0.10	5.93	1.17	0.91	0.23	0.01	1.47	2.58	2.01	1.30	1.87	2.74	2.02	1.30
		0.01	1.26	0.06	0.00	0.09	-0.06	2.23	2.50	2.32	2.23	2.23	2.50	2.32	2.23
	0.7	0.05	0.57	1.11	0.87	0.15	0.02	2.38	2.87	2.49	2.27	2.63	3.00	2.50	2.27
		0.10	5.99	0.91	1.12	0.06	0.04	2.32	3.30	2.74	2.29	2.49	3.49	2.74	2.29

Table 16 contains the outputs for a simpler parameterization of the mixture-like estimation model, namely, the model of between-variables independence of the comparison outcomes in both the set of matches and the set of non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$. The results in this table demonstrate several points. First, comparing to Table 12, suggests that the sim-

ulation approach indeed produces data different from the between-variables independence data (since the independence model results in substantial bias). Second, comparing with Table 15, we see that these associations are between the first and second variables (since the mixture-model parameterized for this association produce high-quality outputs). Third, comparing with Table 15 again, we see that the model specification with the between-variables associations works very well and that the model is identifiable as predicted by our theoretical exploration. Fourth, that failure to specify the correct model when relying on the pure estimation-based approach without classification may lead to very poor results.

Table 16: Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.66	0.11	-3.01	-19.19	0.03	1.08	1.87	7.22	1.05	1.09	3.54	20.50	1.05
	0.9	0.05	0.53	1.41	-1.45	-19.29	0.02	1.46	2.82	4.19	1.06	2.03	3.17	19.74	1.06
		0.10	4.84	1.38	-1.86	-8.55	0.01	1.52	4.09	3.16	1.12	2.06	4.49	9.12	1.12
		0.01	1.28	0.26	-3.07	-18.96	0.15	2.80	3.20	7.78	2.78	2.82	4.43	20.50	2.78
250	0.8	0.05	0.43	1.35	-1.62	-19.17	0.02	2.95	4.08	4.92	2.73	3.24	4.39	19.80	2.73
		0.10	4.59	1.56	-2.54	-8.20	0.13	3.06	6.00	4.16	2.70	3.43	6.52	9.20	2.70
		0.01	0.78	0.41	-3.06	-18.34	0.32	4.77	5.03	9.32	4.76	4.79	5.89	20.57	4.77
	0.7	0.05	0.41	1.42	-1.89	-18.65	0.14	4.94	5.65	6.44	4.75	5.13	5.95	19.73	4.75
		0.10	4.07	1.78	-2.13	-8.28	0.25	5.29	8.01	6.27	4.95	5.58	8.29	10.39	4.95
		0.01	1.94	0.14	-2.72	-20.34	0.05	0.75	1.34	5.14	0.73	0.76	3.03	20.97	0.73
	0.9	0.05	0.34	1.16	-2.68	-19.33	0.02	1.01	2.07	2.87	0.76	1.53	3.39	19.54	0.76
		0.10	3.89	1.20	-4.25	-8.28	-0.03	1.01	2.80	2.07	0.75	1.56	5.09	8.54	0.75
		0.01	1.60	0.16	-2.73	-20.14	0.08	1.82	2.18	6.07	1.82	1.83	3.50	21.03	1.82
500	0.8	0.05	0.30	1.21	-2.42	-19.25	0.10	2.06	2.98	3.62	1.89	2.39	3.84	19.59	1.89
		0.10	4.28	1.15	-5.14	-8.12	-0.01	2.01	4.14	2.79	1.83	2.32	6.60	8.58	1.83
		0.01	1.24	0.18	-2.76	-19.65	0.09	3.28	3.59	7.13	3.27	3.29	4.53	20.90	3.27
	0.7	0.05	0.31	1.22	-2.13	-19.12	0.18	3.48	4.31	4.56	3.40	3.69	4.81	19.65	3.40
		0.10	4.44	1.24	-4.69	-8.02	0.07	3.61	5.68	4.07	3.45	3.82	7.36	8.99	3.45
		0.01	1.93	0.10	-2.63	-20.39	0.04	0.55	0.99	3.78	0.54	0.56	2.81	20.74	0.54
	0.9	0.05	0.24	0.97	-3.23	-19.57	-0.01	0.69	1.51	2.11	0.54	1.19	3.57	19.68	0.54
		0.10	3.97	1.09	-5.79	-8.25	0.01	0.70	1.90	1.40	0.52	1.30	6.10	8.37	0.52
		0.01	1.69	0.03	-2.64	-20.32	-0.03	1.31	1.59	4.40	1.31	1.31	3.08	20.79	1.31
1000	0.8	0.05	0.23	0.85	-3.25	-19.60	-0.03	1.34	2.10	2.50	1.27	1.59	3.87	19.76	1.27
		0.10	4.94	0.94	-6.44	-8.31	0.01	1.42	3.03	1.88	1.30	1.71	7.12	8.52	1.30
		0.01	1.46	-0.01	-2.71	-20.31	-0.06	2.23	2.47	5.18	2.23	2.23	3.67	20.96	2.23
	0.7	0.05	0.25	0.67	-2.95	-19.57	0.02	2.31	2.91	3.20	2.27	2.40	4.15	19.83	2.27
		0.10	5.91	0.68	-6.31	-8.43	0.04	2.32	3.71	2.73	2.29	2.41	7.32	8.86	2.29

When the estimation model does not account for the between-variables associations of the comparison outcomes in the set of non-matches, the no-classification estimators are negatively biased. Note an irregular trend in the bias of the linkage free dual system estimator as the coverage and error probabilities vary. Most likely, this is due to the positive bias in estimation interfering with the negative bias of model misspecification. Another very interesting outcome of ignoring the between-variables associations is increased variance of the estimates produced by a simpler model. This looks like yet another irregularity of the mixture / mixture-like model in comparison with the regular models (see some

known irregularities mentioned in Section 2.2.3). It is also interesting to see how the modified linkage free dual system estimator, which demonstrates a very good performance under the correctly specified model, performs worse than any other estimators when the model does not take the associations into account. The difference in estimates between the linkage free estimator and its modified version are often so large, that maybe such a difference can be used as an indicator of a model misspecification.

Table 17: Single block vs averaged blocking: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches

			RB		RSE		RRMSE	
τ	π_j	ξ	$\widetilde{\tau}$	$\widetilde{\tau}_G$	$\widetilde{\tau}$	$\widetilde{\tau}_G$	$\widetilde{\tau}$	$\widetilde{\tau}_G$
		0.01	-0.26	-1.05	1.96	1.96	1.97	2.23
	0.9	0.05	2.27	1.54	2.86	2.88	3.65	3.27
		0.10	3.60	3.40	4.33	4.07	5.63	5.30
		0.01	-0.21	-1.19	3.24	3.34	3.25	3.55
250	0.8	0.05	2.27	1.92	3.99	4.17	4.59	4.59
		0.10	4.39	3.73	5.62	5.51	7.13	6.65
		0.01	-0.25	-1.23	5.11	5.27	5.12	5.41
	0.7	0.05	2.09	1.86	5.63	5.63	6.01	5.93
		0.10	5.26	4.56	7.50	7.19	9.16	8.51
		0.01	0.10	-0.85	1.37	1.38	1.37	1.62
	0.9	0.05	1.10	0.95	1.97	2.15	2.26	2.35
		0.10	1.68	1.48	2.87	2.62	3.33	3.01
		0.01	0.10	-0.91	2.20	2.32	2.20	2.49
500	0.8	0.05	1.44	1.28	2.90	2.91	3.23	3.18
		0.10	1.98	1.91	3.71	3.45	4.21	3.95
		0.01	0.04	-0.93	3.64	3.57	3.64	3.69
	0.7	0.05	1.72	1.44	4.27	4.08	4.61	4.32
		0.10	2.55	2.81	5.16	5.04	5.76	5.77
		0.01	0.08	-0.86	1.01	0.98	1.01	1.30
	0.9	0.05	0.60	0.33	1.40	1.47	1.53	1.51
		0.10	0.78	0.58	1.96	1.73	2.10	1.82
		0.01	0.07	-0.75	1.60	1.63	1.60	1.79
1000	0.8	0.05	0.60	0.55	2.01	2.10	2.10	2.17
		0.10	0.91	1.00	2.58	2.26	2.74	2.48
		0.01	0.00	-0.78	2.50	2.50	2.50	2.62
	0.7	0.05	0.87	1.04	2.87	2.92	3.00	3.10
		0.10	1.12	1.31	3.30	3.33	3.49	3.58

One interesting behaviour related to the classification-based approach is worth mentioning here. For the chosen thresholds, the simpler incorrect model (Table 16) has better performance than the correct model (Table 15) in terms of the overall error as well as the amount of clerical work. This is another counter-intuitive result which can be easily explained. Regarding the amount of clerical work, the correct model yields reliable parameter estimates and therefore the cut-off points for accepting / rejecting the comparison patterns as links and non-links do correspond well to the chosen nominal thresholds (with respect to the record pairs). These result in fewer patterns being accepted as links than in the case of the between-variables independence model. The between-variables independence model results in an incorrect but more optimistic acceptance and rejection decision. As a result, there is less clerical review under the incorrect model. The main source of the reduction of the error in the classification-based approach with the between-variables independence model is the bias. Recall,

that the overall bias in the classification-based approaches is a linear combination of the number of matches classified as non-links (false negatives) and the number of non-matches classified as links (false positives). Across all scenarios, the bias is positive, meaning that there are more false negatives than false positives. As observed above, the incorrect model classifies several patterns as links while the correct model classifies the same patterns as possible links. Therefore, the classification with the incorrect model has a few more false positive errors than the classification with the correct model. These additional errors reduce the absolute value of the linear combination of the errors which manifests in the simulation work in the reduced bias. In other words, this is just a case of uncontrolled cancellation of errors. We will see a similar behaviour in some other scenarios, but not in all.

Finally, comparing the estimates obtained with the averaging blocking to those obtained estimating directly at the population level, outputs displayed in Table 17, we see again that there is no strong evidence in favour of one or another approach. There are several cases where the averaging blocking leads to slightly better results. However, there is no clear consistency in such performance. More thorough comparison of these two approaches in practical situations is left for future research.

7.5.3 Association between v_2 and v_3 in the set of matches, between-variables independence in the set of non-matches

The next set of results we analyse are those for the simulation model aiming at the association in the comparison outcomes of the second and third variables in the set of matches, and between-variables independence in the set of non-matches. In other words, this is a situation where the errors of recording two population attributes are correlated, rather than the values of population attributes being correlated. As always, we start with the closest parameterization of the mixture-like model, which is the identifiable model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$. The relevant results are displayed in Table 18.

Yet again, the main trends are similar to those presented in Section 7.5.1 for the case of the between-variables independence model. In fact, the quality of the estimates in the event of associations between comparison outcomes of two variables in the set of matches is broadly similar to the quality achieved under between-variables independence. However, for the no-classification methods the relative root mean square error is smaller in the case of the associations under discussion when compared with the error in the case of between-variables independence. This 'improvement' in performance is largely attributable to the decreased bias. As we can see in Table 27 the equivalent model under perfect averaging blocking has a small negative bias across the majority of scenarios. Therefore, we think that this is not a genuine improvement, but the effect of blending the positive bias in the linkage free dual system estimation in practical applications and the negative bias in the between-variables dependence model of interest. We have already seen a similar behaviour in Section 7.5.2.

The amount of clerical effort generally exceeds the amount of reviews needed when no between-variables associations are present. At the same time, the percentage of clerically reviewed records is smaller when comparing the current model with the model aiming at the between-variables associations in the set of non-matches. The caveat in the last statement is that the strengths of the between-variables associations of the outcomes in the set of matches and the strengths of the between-variables

associations of the outcomes in the set of non-matches are not easy to compare due to the different mechanisms by which these associations are achieved.

As in all the results seen so far, the modified version of the linkage free dual system estimator shows a good performance and outperforms the classification-based approach with the chosen thresholds. The modified no-classification estimator beats the classification approach both in terms of the bias and variance. Note that the estimates in the case of the between-variables associations in the set of non-matches presented in Table 15 are more variable than the estimates presented in this section. Again, despite the results, we cannot claim that this is a general property, because of the different ways the associations are generated. However, given that the set of matches is usually very small compared to the set on non-matches, it is likely that associations between the linkage variables in the set of non-matches are main contributors to the variability.

Table 18: Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.25	0.00	-0.50	-0.08	-0.06	1.14	1.22	1.16	1.13	1.14	1.32	1.17	1.13
	0.9	0.05	0.49	0.85	1.60	0.03	0.03	1.31	2.06	1.30	1.08	1.56	2.61	1.31	1.08
		0.10	3.79	1.98	3.66	0.83	0.06	1.65	3.17	1.98	1.09	2.57	4.84	2.15	1.09
		0.01	1.74	0.12	-0.50	0.01	0.05	2.67	2.69	2.67	2.65	2.67	2.74	2.67	2.65
250	0.8	0.05	0.45	0.61	0.67	-0.09	0.06	2.75	3.16	2.75	2.69	2.81	3.23	2.75	2.69
		0.10	2.13	1.64	3.62	0.44	-0.13	3.09	4.29	3.24	2.77	3.50	5.61	3.27	2.77
		0.01	0.95	0.22	-0.75	0.11	0.18	4.41	4.33	4.42	4.4	4.41	4.40	4.42	4.41
	0.7	0.05	0.38	0.76	0.67	0.15	0.29	4.57	4.76	4.54	4.47	4.64	4.80	4.54	4.48
		0.10	0.93	1.77	3.88	0.46	0.10	4.93	5.82	4.89	4.65	5.24	6.99	4.91	4.65
		0.01	1.47	0.04	-0.40	-0.01	0.02	0.75	0.85	0.76	0.75	0.75	0.94	0.76	0.75
	0.9	0.05	0.37	0.66	0.56	-0.22	-0.03	0.90	1.35	0.96	0.75	1.11	1.46	0.99	0.75
		0.10	2.99	1.40	1.12	-0.01	-0.03	1.03	2.08	1.40	0.77	1.74	2.36	1.40	0.77
		0.01	1.27	0.20	-0.31	0.13	0.17	1.84	1.87	1.84	1.83	1.85	1.89	1.85	1.84
500	0.8	0.05	0.40	0.64	0.75	-0.15	0.04	1.96	2.24	1.98	1.89	2.06	2.36	1.99	1.89
		0.10	4.12	1.42	1.70	0.11	0.00	2.05	3.02	2.30	1.85	2.49	3.46	2.30	1.85
		0.01	1.14	0.06	-0.48	-0.04	0.03	3.23	3.24	3.22	3.22	3.23	3.27	3.22	3.22
	0.7	0.05	0.38	0.74	1.00	-0.06	0.17	3.41	3.60	3.40	3.38	3.49	3.74	3.40	3.38
		0.10	4.61	1.37	2.42	0.21	0.05	3.35	4.24	3.56	3.24	3.62	4.88	3.57	3.24
		0.01	1.81	0.03	-0.27	-0.05	0.00	0.54	0.63	0.55	0.53	0.54	0.68	0.55	0.53
	0.9	0.05	0.13	1.24	0.53	0.08	0.01	0.72	0.97	0.79	0.54	1.44	1.11	0.80	0.54
		0.10	4.17	1.56	0.79	0.29	0.02	0.82	1.51	1.18	0.53	1.76	1.70	1.21	0.53
		0.01	1.40	0.07	-0.26	-0.01	0.04	1.28	1.33	1.29	1.28	1.29	1.36	1.29	1.28
1000	0.8	0.05	0.14	1.07	0.65	0.06	-0.05	1.38	1.63	1.45	1.30	1.75	1.76	1.45	1.30
		0.10	5.90	1.29	0.75	-0.01	0.02	1.45	2.03	1.72	1.29	1.94	2.16	1.72	1.29
		0.01	1.28	0.00	-0.39	-0.11	-0.03	2.36	2.36	2.36	2.36	2.36	2.39	2.36	2.36
	0.7	0.05	0.33	0.56	0.45	-0.27	0.05	2.28	2.45	2.31	2.26	2.35	2.49	2.32	2.26
		0.10	7.52	1.03	0.97	-0.18	0.08	2.38	2.82	2.54	2.26	2.59	2.98	2.54	2.26

Applying a simpler between-variables independence model to the data with the associations under discussion, we observe several curious features in the results; see Table 19. The classification-based approach requires less clerical intervention with this simpler but incorrect model than it requires with the correct model. This is similar to what we have seen in the case of the associations between the

comparison outcomes of two variables in the set on non-matches. However, this time the relative bias increases when the incorrect model is used. It again demonstrates some unpredictability in performance of the classification-based approach. The results for the linkage free dual system estimator and its modified version, when comparing with the results in Table 18, show that the correct model specification is important in order to obtain accurate estimates. Nevertheless, there is not such a dramatic loss of accuracy when a model is misspecified in comparison with the situation when between-variables associations are present in the set of non-matches. Specifically, the modified linkage free dual system estimator still performs very well and outperforms the classification-based method.

Table 19: Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.56	0.00	-0.20	-0.07	-0.06	1.14	1.20	1.16	1.13	1.14	1.21	1.17	1.13
	0.9	0.05	0.77	0.86	2.65	0.21	0.03	1.31	2.30	1.30	1.08	1.57	3.51	1.31	1.08
		0.10	2.68	2.65	7.28	2.15	0.06	1.74	3.71	1.94	1.09	3.17	8.18	2.90	1.09
		0.01	2.25	0.12	-0.18	0.03	0.05	2.67	2.69	2.67	2.65	2.67	2.70	2.67	2.65
250	0.8	0.05	0.65	0.62	1.42	0.04	0.06	2.74	3.25	2.75	2.69	2.81	3.55	2.75	2.69
		0.10	1.36	2.39	7.42	1.57	-0.13	3.22	4.91	3.29	2.77	4.01	8.90	3.65	2.77
		0.01	1.39	0.22	-0.35	0.13	0.18	4.41	4.35	4.42	4.4	4.41	4.37	4.42	4.41
	0.7	0.05	0.63	0.77	1.49	0.28	0.29	4.58	4.89	4.54	4.47	4.64	5.11	4.55	4.48
		0.10	0.61	2.39	7.71	1.39	0.10	4.99	6.46	4.92	4.65	5.53	10.06	5.11	4.65
		0.01	1.64	0.04	-0.15	0.00	0.02	0.75	0.81	0.76	0.75	0.75	0.82	0.76	0.75
	0.9	0.05	0.65	0.67	1.72	-0.02	-0.03	0.90	1.58	0.96	0.75	1.12	2.34	0.96	0.75
		0.10	1.83	1.96	4.44	1.11	-0.03	1.20	2.55	1.38	0.77	2.30	5.12	1.77	0.77
		0.01	1.45	0.20	-0.03	0.14	0.17	1.84	1.86	1.84	1.83	1.85	1.86	1.85	1.84
500	0.8	0.05	0.70	0.66	1.79	0.03	0.04	1.96	2.38	1.99	1.89	2.07	2.98	1.99	1.89
		0.10	2.03	2.06	5.27	1.33	0.00	2.20	3.50	2.30	1.85	3.02	6.33	2.65	1.85
		0.01	1.30	0.06	-0.19	-0.02	0.03	3.23	3.23	3.22	3.22	3.23	3.24	3.22	3.22
	0.7	0.05	0.63	0.79	2.07	0.13	0.17	3.41	3.70	3.39	3.38	3.50	4.24	3.40	3.38
		0.10	2.12	2.11	5.88	1.39	0.05	3.48	4.71	3.58	3.24	4.07	7.53	3.84	3.24
		0.01	1.89	0.03	-0.10	-0.04	0.00	0.54	0.57	0.55	0.53	0.54	0.58	0.55	0.53
	0.9	0.05	0.32	1.28	1.63	0.34	0.01	0.72	1.09	0.78	0.54	1.47	1.97	0.85	0.54
		0.10	2.43	2.11	4.09	1.47	0.02	0.90	1.77	1.14	0.53	2.29	4.46	1.86	0.53
		0.01	1.50	0.07	-0.09	0.00	0.04	1.28	1.30	1.29	1.28	1.29	1.30	1.29	1.28
1000	0.8	0.05	0.29	1.16	1.70	0.30	-0.05	1.39	1.75	1.45	1.30	1.81	2.44	1.48	1.30
		0.10	2.55	1.92	4.13	1.20	0.02	1.54	2.39	1.71	1.29	2.46	4.77	2.09	1.29
		0.01	1.41	0.00	-0.18	-0.09	-0.03	2.36	2.35	2.36	2.36	2.36	2.36	2.36	2.36
	0.7	0.05	0.51	0.66	1.55	-0.04	0.05	2.29	2.56	2.32	2.26	2.38	2.99	2.32	2.26
		0.10	3.02	1.63	4.34	1.05	0.08	2.38	3.17	2.55	2.26	2.88	5.37	2.76	2.26

7.5.4 Association between v_2 and v_3 in the set of matches, association between v_1 and v_2 in the set of non-matches

The input data used to produce the estimates in this section were simulated aiming at the association between the comparison outcomes of the variables v_2 and v_3 in the set of matches and association between the comparison outcomes of the variables v_1 and v_2 in the set of non-matches. As always, we consider the closest mixture-like parameterization first. This identifiable parameterization

is $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$.

All the general trends seen before are present in the results of this section, displayed in Table 20. One exception to the trends is that the percentage of clerical resolutions in the classification-based approach increases as the error recording the population attributes increases for the population size $\tau = 250$. The previously observed non-monotone trend persists for other values of the population size parameter.

Table 20: Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.30	0.69	-0.48	0.04	0.05	1.36	1.95	1.33	1.07	1.53	2.01	1.33	1.07
	0.9	0.05	2.91	1.61	1.38	0.30	0.01	1.60	2.87	1.74	1.13	2.27	3.19	1.77	1.13
		0.10	11.25	1.86	2.78	0.78	-0.03	1.67	4.03	2.65	1.11	2.50	4.89	2.76	1.11
		0.01	1.09	0.81	-0.53	0.17	0.16	2.79	3.24	2.83	2.71	2.90	3.28	2.84	2.72
250	0.8	0.05	1.51	1.28	0.63	-0.09	-0.07	2.95	3.84	2.92	2.63	3.22	3.89	2.92	2.63
		0.10	7.28	2.05	2.86	0.72	0.03	3.03	4.76	3.48	2.61	3.66	5.55	3.56	2.61
		0.01	0.90	0.88	-0.87	0.15	0.18	4.77	5.05	4.84	4.79	4.85	5.12	4.85	4.79
	0.7	0.05	1.08	1.69	0.82	0.31	0.44	4.96	5.52	4.95	4.74	5.24	5.58	4.96	4.76
		0.10	4.62	2.05	2.57	0.25	0.06	5.02	6.68	5.35	4.69	5.42	7.16	5.36	4.69
		0.01	1.72	0.16	-0.22	0.09	0.07	0.75	1.41	0.95	0.73	0.76	1.42	0.95	0.73
	0.9	0.05	0.53	1.27	0.48	-0.02	0.00	0.98	1.99	1.20	0.77	1.61	2.04	1.20	0.77
		0.10	6.26	1.47	1.00	0.13	-0.04	1.07	2.67	1.84	0.76	1.82	2.85	1.84	0.76
		0.01	1.39	0.14	-0.34	0.02	0.03	1.84	2.19	1.92	1.83	1.84	2.22	1.92	1.83
500	0.8	0.05	0.44	1.39	0.63	0.12	0.05	2.10	2.75	2.18	1.95	2.52	2.82	2.19	1.95
		0.10	6.48	1.63	1.39	0.26	0.03	2.10	3.63	2.73	1.83	2.66	3.89	2.74	1.83
		0.01	1.12	0.35	-0.28	0.24	0.24	3.15	3.45	3.25	3.14	3.17	3.46	3.26	3.15
	0.7	0.05	0.48	1.59	1.08	0.35	0.18	3.47	3.95	3.50	3.35	3.81	4.09	3.52	3.35
		0.10	6.43	1.42	1.49	0.00	-0.13	3.49	4.75	3.91	3.37	3.77	4.98	3.91	3.37
		0.01	1.60	0.09	-0.18	-0.04	0.01	0.56	0.98	0.70	0.55	0.57	1.00	0.70	0.55
	0.9	0.05	0.84	1.36	0.53	0.28	0.03	0.78	1.47	0.99	0.53	1.56	1.56	1.03	0.53
		0.10	5.69	1.53	0.61	0.29	0.01	0.86	2.04	1.51	0.57	1.75	2.13	1.53	0.57
		0.01	1.39	0.12	-0.20	-0.01	0.04	1.30	1.60	1.37	1.30	1.31	1.61	1.37	1.30
1000	0.8	0.05	0.83	1.27	0.61	0.29	-0.03	1.47	2.00	1.61	1.30	1.94	2.09	1.63	1.30
		0.10	6.75	1.36	0.53	0.01	0.03	1.47	2.45	1.97	1.34	2.00	2.50	1.97	1.34
		0.01	1.20	0.12	-0.18	0.01	0.03	2.35	2.56	2.40	2.34	2.35	2.57	2.40	2.34
	0.7	0.05	0.50	0.84	0.18	-0.20	0.06	2.21	2.75	2.37	2.21	2.37	2.75	2.38	2.21
		0.10	7.60	1.08	0.86	0.04	0.02	2.38	3.35	2.86	2.26	2.61	3.46	2.86	2.26

The linkage free dual system estimators with this model specification account well for the associations between the comparisons of linkage variables. The modified linkage free estimator yet again shows very good performance outperforming the classification-based approach. Having said that the no-classification approach outperforms the classification-based, we have to remain a little bit critical about such a claim. If we compare the relative root mean square error of the current simulation and estimation model case with the error for the case of between-variables associations of the comparison outcomes in the set of non-matches only (Table 15), we notice that the former is smaller. This again can be explained by the fact that the models that account for associations have a tendency to incur some negative bias. Theoretical results in Table 27 confirm that the model with between-variables as-

sociations in both the set of matches and non-matches has the largest negative bias among the models explored.

Table 21: Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.62	0.51	-0.16	0.17	0.05	1.27	1.89	1.33	1.07	1.37	1.90	1.34	1.07
	0.9	0.05	2.48	1.92	2.65	0.85	0.01	1.73	3.01	1.83	1.13	2.59	4.01	2.02	1.13
		0.10	10.71	2.86	6.28	2.41	-0.03	2.02	4.66	2.77	1.11	3.50	7.82	3.67	1.11
		0.01	1.39	0.66	-0.19	0.29	0.16	2.77	3.22	2.83	2.71	2.85	3.23	2.85	2.72
250	0.8	0.05	1.26	1.56	1.53	0.27	-0.07	2.99	3.88	2.96	2.63	3.38	4.17	2.97	2.63
		0.10	6.63	3.30	6.45	2.25	0.03	3.25	5.48	3.66	2.61	4.63	8.46	4.30	2.61
		0.01	1.16	0.71	-0.49	0.28	0.18	4.76	5.03	4.85	4.79	4.81	5.06	4.86	4.79
	0.7	0.05	0.93	2.05	1.81	0.73	0.44	5.06	5.60	5.01	4.74	5.46	5.89	5.06	4.76
		0.10	4.11	3.29	6.40	1.70	0.06	5.26	7.36	5.55	4.69	6.20	9.75	5.80	4.69
		0.01	1.87	0.15	0.04	0.20	0.07	0.75	1.35	0.94	0.73	0.76	1.35	0.96	0.73
	0.9	0.05	0.44	1.43	1.66	0.55	0.00	0.97	2.06	1.27	0.77	1.73	2.64	1.39	0.77
		0.10	5.00	2.16	4.16	1.49	-0.04	1.38	3.07	1.89	0.76	2.56	5.17	2.41	0.76
		0.01	1.53	0.14	-0.09	0.12	0.03	1.84	2.16	1.91	1.83	1.84	2.17	1.92	1.83
500	0.8	0.05	0.36	1.53	1.66	0.58	0.05	2.09	2.80	2.20	1.95	2.59	3.25	2.28	1.95
		0.10	4.48	2.37	4.70	1.69	0.03	2.29	4.02	2.77	1.83	3.29	6.18	3.24	1.83
		0.01	1.22	0.35	-0.01	0.35	0.24	3.15	3.46	3.25	3.14	3.17	3.46	3.27	3.15
	0.7	0.05	0.43	1.70	2.28	0.87	0.18	3.47	4.05	3.56	3.35	3.86	4.65	3.66	3.35
		0.10	3.89	2.28	4.82	1.45	-0.13	3.73	5.22	4.02	3.37	4.37	7.10	4.27	3.37
		0.01	1.67	0.09	0.01	0.05	0.01	0.56	0.96	0.69	0.55	0.57	0.96	0.69	0.55
	0.9	0.05	0.59	1.69	1.64	0.85	0.03	0.85	1.54	1.04	0.53	1.89	2.25	1.34	0.53
		0.10	3.90	2.08	3.61	1.58	0.01	0.93	2.27	1.53	0.57	2.28	4.26	2.20	0.57
		0.01	1.47	0.12	0.00	0.08	0.04	1.30	1.57	1.36	1.30	1.31	1.57	1.36	1.30
1000	0.8	0.05	0.61	1.60	1.70	0.85	-0.03	1.53	2.05	1.63	1.30	2.21	2.66	1.84	1.30
		0.10	3.66	1.93	3.68	1.35	0.03	1.52	2.73	2.01	1.34	2.46	4.58	2.42	1.34
		0.01	1.25	0.11	0.05	0.11	0.03	2.35	2.56	2.41	2.34	2.35	2.56	2.41	2.34
	0.7	0.05	0.37	1.03	1.29	0.34	0.06	2.24	2.81	2.42	2.21	2.46	3.09	2.44	2.21
		0.10	3.70	1.71	4.01	1.39	0.02	2.46	3.67	2.91	2.26	3.00	5.44	3.22	2.26

We now briefly look at the results for three simpler and identifiable but incorrect models: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$, $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, and $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

The outputs for the model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$ are displayed in Table 21. This estimation model accounts only for the between-variables association in the set of non-matches. The increase in the relative root means square error is observed when comparing this simpler model with the model accounting for the between-variables associations in both the set of matches and non-matches. This increase is mainly inflicted by the increased bias, but there is some contribution from the variability as well. The quality of the modified linkage free dual system estimator seems largely acceptable across all of the scenarios. The quality of the basic linkage free estimator is mainly acceptable for the scenarios with higher coverage and smaller errors.

Table 22: Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.23	0.15	-3.14	-17.85	0.05	1.10	1.82	7.13	1.07	1.11	3.63	19.22	1.07
	0.9	0.05	0.51	1.05	-2.15	-19.72	0.01	1.46	2.79	4.06	1.13	1.80	3.52	20.13	1.13
		0.10	7.01	1.62	-2.02	-9.21	-0.03	1.59	4.05	3.46	1.11	2.27	4.53	9.84	1.11
		0.01	0.90	0.28	-3.32	-18.61	0.16	2.73	3.21	7.84	2.71	2.75	4.62	20.19	2.72
250	0.8	0.05	0.35	0.91	-2.88	-21.72	-0.07	2.82	3.82	4.70	2.63	2.96	4.79	22.22	2.63
		0.10	3.90	1.71	-2.98	-10.09	0.03	2.92	5.15	4.63	2.61	3.38	5.95	11.10	2.61
		0.01	0.60	0.28	-3.66	-19.42	0.18	4.79	4.99	9.01	4.79	4.80	6.19	21.41	4.79
	0.7	0.05	0.30	1.23	-2.63	-21.62	0.44	4.78	5.49	6.01	4.74	4.94	6.09	22.44	4.76
		0.10	1.96	1.55	-3.51	-10.80	0.06	4.88	7.04	6.33	4.69	5.12	7.87	12.52	4.69
		0.01	1.42	0.13	-3.01	-20.15	0.07	0.74	1.38	5.43	0.73	0.75	3.31	20.87	0.73
	0.9	0.05	0.29	0.76	-3.09	-20.90	0.00	0.92	2.02	2.90	0.77	1.19	3.69	21.10	0.77
		0.10	4.21	1.20	-4.22	-9.67	-0.04	1.04	2.62	2.31	0.76	1.58	4.96	9.95	0.76
		0.01	1.21	0.10	-3.14	-20.65	0.03	1.84	2.18	5.85	1.83	1.84	3.82	21.46	1.83
500	0.8	0.05	0.29	0.77	-3.06	-20.94	0.05	2.03	2.82	3.56	1.95	2.17	4.16	21.24	1.95
		0.10	5.23	1.24	-4.57	-9.60	0.03	2.02	3.94	3.11	1.83	2.37	6.03	10.09	1.83
		0.01	1.04	0.31	-3.05	-19.93	0.24	3.14	3.43	7.08	3.14	3.16	4.60	21.15	3.15
	0.7	0.05	0.30	0.86	-2.62	-20.50	0.18	3.36	3.94	4.50	3.35	3.47	4.73	20.98	3.35
		0.10	5.73	1.04	-4.32	-9.89	-0.13	3.46	4.98	4.23	3.37	3.61	6.59	10.76	3.37
		0.01	1.78	0.06	-2.87	-20.51	0.01	0.56	0.98	3.84	0.55	0.56	3.03	20.87	0.55
	0.9	0.05	0.11	1.17	-3.05	-20.00	0.03	0.71	1.53	2.07	0.53	1.37	3.42	20.11	0.53
		0.10	4.83	1.34	-5.20	-9.28	0.01	0.77	2.01	1.67	0.57	1.55	5.57	9.43	0.57
		0.01	1.38	0.10	-2.90	-20.97	0.04	1.30	1.58	4.20	1.30	1.30	3.31	21.39	1.30
1000	0.8	0.05	0.13	1.03	-3.03	-19.72	-0.03	1.45	2.05	2.58	1.30	1.78	3.66	19.89	1.30
		0.10	6.61	1.19	-5.69	-9.55	0.03	1.44	2.72	2.08	1.34	1.87	6.31	9.78	1.34
		0.01	1.29	0.08	-2.85	-20.36	0.03	2.35	2.53	5.26	2.34	2.35	3.82	21.03	2.34
	0.7	0.05	0.25	0.45	-3.43	-21.09	0.06	2.25	2.77	3.18	2.21	2.30	4.41	21.33	2.21
		0.10	8.15	0.92	-5.39	-9.75	0.02	2.36	3.63	2.82	2.26	2.54	6.50	10.15	2.26

The situation is less favourable for the no-classification approaches when only the between-variables association in the set of matches is accounted for. The corresponding results are presented in Table 22. The results are similar to those in Table 19 and the linkage free estimators are doing poorly when the estimation model is misspecified for the set of non-matches. Note that among all four models considered in this section, the classification-based approach achieves the best performance for $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$. This is another example of somehow unpredictable behaviour of the classification approach.

Finally, the results for the estimation model with between-variables independence in both the set of matches and the set of non-matches are tabulated in Table 23. These results are very similar to those discussed in the previous paragraph. The linkage free estimators suffer from model misspecification. The classification-based approach demonstrates an impressive level of robustness, but it is difficult to determine its exact behaviour for any given combination of the input data and model specification. Also, it is obvious that the classification-based approach cannot be robust across all the choices of the thresholds and it is not clear how one would choose such a set of thresholds that guarantee robustness in practice.

Table 23: Simulated data: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.58	2.74	-2.89	-17.79	0.05	1.10	1.80	7.18	1.07	1.10	3.41	19.18	1.07
	0.9	0.05	0.67	3.78	-1.20	-19.30	0.01	1.51	2.96	4.07	1.13	1.94	3.20	19.73	1.13
		0.10	6.69	4.35	0.22	-7.72	-0.03	1.74	4.81	3.27	1.11	2.75	4.82	8.38	1.11
		0.01	1.19	6.80	-3.03	-18.49	0.16	2.72	3.19	7.88	2.71	2.74	4.40	20.09	2.72
250	0.8	0.05	0.44	7.12	-2.12	-21.55	-0.07	2.85	3.87	4.68	2.63	3.02	4.41	22.06	2.63
		0.10	3.42	7.75	-0.35	-8.67	0.03	3.10	6.00	4.50	2.61	3.86	6.01	9.77	2.61
		0.01	0.76	12.00	-3.33	-19.30	0.18	4.80	4.98	9.04	4.79	4.81	5.99	21.31	4.79
	0.7	0.05	0.41	12.09	-1.78	-21.42	0.44	4.84	5.58	6.02	4.74	5.03	5.85	22.25	4.76
		0.10	1.72	12.62	-0.64	-9.42	0.06	5.05	7.91	6.19	4.69	5.46	7.93	11.27	4.69
		0.01	1.68	3.70	-2.80	-20.03	0.07	0.74	1.33	5.45	0.73	0.75	3.10	20.76	0.73
	0.9	0.05	0.41	4.69	-2.17	-20.45	0.00	0.94	2.12	2.91	0.77	1.24	3.03	20.66	0.77
		0.10	3.22	5.37	-2.24	-8.34	-0.04	1.07	3.19	2.17	0.76	1.88	3.89	8.62	0.76
		0.01	1.44	9.17	-2.92	-20.52	0.03	1.83	2.15	5.88	1.83	1.84	3.63	21.34	1.83
500	0.8	0.05	0.40	10.13	-2.24	-20.60	0.05	2.03	2.88	3.55	1.95	2.18	3.65	20.91	1.95
		0.10	3.41	10.48	-2.16	-8.21	0.03	2.10	4.41	3.01	1.83	2.67	4.91	8.74	1.83
		0.01	1.21	15.71	-2.82	-19.80	0.24	3.14	3.44	7.11	3.14	3.16	4.45	21.04	3.15
	0.7	0.05	0.42	16.89	-1.64	-20.12	0.18	3.38	4.04	4.52	3.35	3.50	4.36	20.62	3.35
		0.10	3.49	17.77	-1.83	-8.51	-0.13	3.55	5.52	4.21	3.37	3.85	5.82	9.49	3.37
		0.01	1.93	5.59	-2.71	-20.39	0.01	0.56	0.96	3.85	0.55	0.56	2.88	20.75	0.55
	0.9	0.05	0.21	7.00	-2.19	-19.53	0.03	0.70	1.62	2.05	0.53	1.42	2.72	19.64	0.53
		0.10	3.29	8.16	-3.40	-8.02	0.01	0.82	2.48	1.55	0.57	1.85	4.21	8.17	0.57
		0.01	1.52	13.01	-2.73	-20.85	0.04	1.30	1.56	4.22	1.30	1.30	3.14	21.27	1.30
1000	0.8	0.05	0.19	14.30	-2.17	-19.27	-0.03	1.43	2.11	2.57	1.30	1.84	3.02	19.44	1.30
		0.10	3.82	14.61	-3.43	-8.19	0.03	1.46	3.06	2.01	1.34	2.12	4.60	8.43	1.34
		0.01	1.42	23.44	-2.67	-20.24	0.03	2.34	2.53	5.27	2.34	2.35	3.68	20.91	2.34
	0.7	0.05	0.36	22.52	-2.54	-20.67	0.06	2.25	2.84	3.20	2.21	2.32	3.81	20.92	2.21
		0.10	4.33	23.83	-3.15	-8.41	0.02	2.38	4.02	2.76	2.26	2.71	5.11	8.85	2.26

7.5.5 Between-variables independence in the set of matches, three-way association between v_1, v_2 and v_3 in the set of non-matches

Until now all the examples considered involved data with associations for which an identifiable parametric proxy of the mixture-like model existed. Such identifiable models demonstrated the feasibility of the no-classification methods for population size estimation. In this section we will look at an example of data that have between-variables associations of the comparison outcomes without a corresponding identifiable parameterization of the mixture-like model. The simulation model in this case aims at between-variables independence of the comparisons in the set of matches and a 3-way association in the comparison outcomes between variables v_1, v_2 and v_3 in the set of non-matches. The closest parameterization is $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3},\gamma_4)$, which is non-identifiable (see Section 5.2.4). A simpler candidate is the model with pairwise associations between three variables in the set of non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_{1,3},\gamma_{2,3},\gamma_4)$. It is not known whether this model is identifiable or not.

Table 24: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.23	0.28	-51.47	-6.80	0.00	1.29	2.79	4.82	1.12	1.32	51.54	8.33	1.12
	0.9	0.05	9.40	1.24	-44.18	-17.85	-0.02	1.47	3.54	4.43	1.09	1.92	44.32	18.40	1.09
		0.10	18.25	1.65	-33.21	-19.81	0.01	1.46	6.60	4.23	1.02	2.20	33.86	20.25	1.02
		0.01	1.61	0.37	-51.17	-6.58	0.05	2.67	3.28	5.59	2.60	2.69	51.27	8.64	2.60
250	0.8	0.05	7.28	1.29	-44.12	-17.40	0.03	2.94	4.24	5.16	2.69	3.21	44.32	18.15	2.69
		0.10	14.78	2.06	-32.93	-19.05	0.26	2.99	7.86	5.53	2.59	3.63	33.85	19.83	2.61
		0.01	1.08	0.43	-51.19	-6.21	0.16	4.67	4.22	6.60	4.61	4.69	51.36	9.07	4.61
	0.7	0.05	5.62	1.37	-44.06	-16.79	0.01	4.77	4.98	6.73	4.57	4.96	44.34	18.08	4.57
		0.10	11.78	1.73	-32.47	-18.27	-0.12	5.12	9.74	7.06	4.86	5.41	33.90	19.59	4.86
		0.01	1.87	0.22	-51.48	-7.60	0.00	0.83	1.87	3.60	0.77	0.86	51.52	8.41	0.77
	0.9	0.05	8.85	1.34	-44.28	-19.01	0.01	1.12	2.39	3.06	0.79	1.75	44.34	19.26	0.79
		0.10	18.75	1.56	-34.50	-21.07	0.03	1.05	3.37	2.47	0.77	1.88	34.66	21.22	0.77
		0.01	1.55	0.28	-51.44	-7.72	0.04	1.92	2.24	4.36	1.90	1.95	51.49	8.87	1.90
500	0.8	0.05	6.82	1.44	-44.39	-18.53	0.01	2.09	2.88	3.82	1.86	2.54	44.48	18.92	1.86
		0.10	15.69	1.70	-34.14	-20.68	0.10	2.04	4.33	3.13	1.88	2.65	34.42	20.92	1.89
		0.01	1.06	0.45	-51.27	-7.55	0.23	3.25	2.99	5.36	3.21	3.28	51.36	9.26	3.22
	0.7	0.05	5.32	1.79	-44.03	-18.20	0.34	3.40	3.64	4.79	3.25	3.84	44.18	18.82	3.26
		0.10	12.94	1.62	-33.68	-20.33	0.04	3.57	5.94	4.61	3.36	3.92	34.20	20.85	3.36
		0.01	0.92	0.19	-51.03	-9.78	0.02	0.56	1.37	2.91	0.52	0.59	51.05	10.21	0.52
	0.9	0.05	8.61	1.40	-44.26	-20.13	0.00	0.89	1.54	2.12	0.54	1.66	44.28	20.24	0.54
		0.10	19.56	1.40	-34.72	-21.69	-0.01	0.74	1.98	1.58	0.56	1.58	34.78	21.75	0.56
		0.01	0.70	0.17	-51.15	-9.55	0.02	1.30	1.62	3.31	1.28	1.32	51.18	10.11	1.28
1000	0.8	0.05	6.67	1.46	-44.23	-20.02	0.03	1.52	1.90	2.60	1.29	2.11	44.27	20.19	1.29
		0.10	17.14	1.43	-34.68	-21.55	-0.01	1.48	2.40	2.06	1.35	2.06	34.77	21.65	1.35
		0.01	0.59	0.23	-51.14	-9.40	0.10	2.22	2.06	4.08	2.21	2.23	51.18	10.24	2.21
	0.7	0.05	5.15	1.61	-44.01	-19.70	0.18	2.55	2.55	3.25	2.39	3.01	44.09	19.96	2.39
		0.10	14.81	1.41	-34.48	-21.39	-0.04	2.30	3.86	3.14	2.21	2.69	34.69	21.62	2.21

The results produced by the non-identifiable model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$ are displayed in Table 24. It is clear that the no-classification approaches perform very poorly in this case. It is what was expected with a non-identifiable model. The relative bias is extremely large across all the scenarios. In fact, it is often so large, that the population size estimate is smaller than the number of observed individuals in the surveys used in the estimation. Hence, if somebody risks using this non-identifiable model in practice, it may be easy to see from the outputs that it does not work well. It is worth noting that the variance of the linkage free estimators is very low given that the model is not identifiable. This is somehow unexpected, as one would anticipate the simulated annealing to produce estimates of the population size τ that are far apart. However, it is possible that there are infinitely many solutions concentrated around a certain region. This may be a topic for future investigations.

The classification-based approach performs well despite the model being non-identifiable, but now requites substantial clerical contribution. Note the regular pattern in the percentage of clerically resolved records as the error ξ increases.

The results for a simpler model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$

are presented in Table 25. These results are very similar to those obtained with the more complex non-identifiable model. It is reasonable to assume that this simpler model is also non-identifiable.

Several other simpler but identifiable models produce outputs similar in nature to those seen in the above sections when an incorrect identifiable model is used instead of the closest parameterization of the mixture-like model, these outputs are placed in Appendix B.1.

Table 25: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
250	0.9	0.01	1.74	0.33	-52.44	-8.94	0.00	1.35	2.73	5.01	1.12	1.39	52.51	10.24	1.12
		0.05	8.25	0.53	-46.91	-16.46	-0.02	1.63	3.08	4.72	1.09	1.71	47.01	17.12	1.09
		0.10	15.35	0.18	-38.63	-18.92	0.01	1.78	3.53	3.30	1.02	1.79	38.79	19.21	1.02
	0.8	0.01	1.26	0.37	-52.15	-8.57	0.05	2.70	3.20	5.71	2.60	2.73	52.24	10.30	2.60
		0.05	6.29	0.63	-46.91	-15.83	0.03	3.00	3.71	5.50	2.69	3.07	47.06	16.76	2.69
		0.10	12.59	0.62	-38.38	-18.11	0.26	3.17	4.44	4.33	2.59	3.23	38.64	18.62	2.61
	0.7	0.01	0.86	0.45	-52.15	-8.10	0.16	4.66	4.15	6.69	4.61	4.68	52.31	10.51	4.61
		0.05	4.87	0.60	-46.89	-15.31	0.01	4.94	4.35	7.00	4.57	4.97	47.09	16.84	4.57
		0.10	10.19	0.28	-38.40	-17.67	-0.12	5.29	5.65	6.03	4.86	5.29	38.82	18.67	4.86
500	0.9	0.01	1.40	0.26	-52.44	-11.79	0.00	0.90	1.84	3.84	0.77	0.93	52.47	12.40	0.77
		0.05	8.10	0.57	-47.04	-18.85	0.01	1.16	2.27	3.33	0.79	1.30	47.09	19.14	0.79
		0.10	16.06	0.51	-39.25	-20.75	0.03	1.20	2.41	2.35	0.77	1.31	39.32	20.88	0.77
	0.8	0.01	1.18	0.36	-52.39	-11.83	0.04	1.95	2.19	4.36	1.90	1.99	52.43	12.61	1.90
		0.05	6.26	0.74	-47.14	-18.34	0.01	2.12	2.65	4.08	1.86	2.24	47.21	18.79	1.86
		0.10	13.56	0.65	-39.07	-20.37	0.10	2.23	2.92	2.95	1.88	2.32	39.18	20.58	1.89
	0.7	0.01	0.83	0.50	-52.21	-11.52	0.23	3.27	2.96	5.36	3.21	3.31	52.30	12.70	3.22
		0.05	4.95	1.17	-46.80	-17.94	0.34	3.46	3.34	5.03	3.25	3.65	46.92	18.63	3.26
		0.10	11.41	0.63	-38.86	-20.21	0.04	3.61	3.76	4.01	3.36	3.67	39.04	20.60	3.36
1000	0.9	0.01	0.65	0.21	-52.00	-16.12	0.02	0.58	1.34	2.78	0.52	0.62	52.01	16.36	0.52
		0.05	8.54	0.63	-47.09	-22.16	0.00	0.83	1.45	2.17	0.54	1.04	47.11	22.27	0.54
		0.10	17.12	0.55	-39.43	-23.11	-0.01	0.84	1.65	1.65	0.56	1.01	39.47	23.17	0.56
	0.8	0.01	0.49	0.20	-52.10	-15.77	0.02	1.31	1.58	3.10	1.28	1.33	52.13	16.08	1.28
		0.05	6.60	0.66	-47.06	-21.98	0.03	1.48	1.80	2.69	1.29	1.62	47.10	22.15	1.29
		0.10	15.29	0.55	-39.39	-22.94	-0.01	1.53	1.99	2.09	1.35	1.62	39.44	23.03	1.35
	0.7	0.01	0.45	0.27	-52.08	-15.45	0.10	2.23	2.01	3.89	2.21	2.25	52.12	15.93	2.21
		0.05	5.12	0.88	-46.83	-21.54	0.18	2.50	2.41	3.36	2.39	2.65	46.89	21.80	2.39
		0.10	13.49	0.56	-39.40	-22.88	-0.04	2.35	2.54	2.67	2.21	2.41	39.48	23.03	2.21

The results produced using the estimation model of between-variables independence in both the set of matches and non-matches are in Table 26. As in all the examples seen in this thesis, the simple between-variables independence model with the classification-based approach and 1-to-1 constraint (and the chosen thresholds) demonstrates a good performance no matter how complicated the between-variables associations in the underlying data are. Interestingly, the modified linkage free dual system estimator is not as extremely biased as it is in other cases of applying the between-variables independence model in the presence of between-variables associations of the comparison outcomes.

The results in this chapter demonstrate empirically the importance of the identifiability of models in the no-classification estimation of the population size τ. At the same time, the classification approach with clerical resolutions does not appear affected much by the use of non-identifiable models.

Table 26: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.16	0.00	-54.37	-4.18	0.00	1.14	2.35	2.23	1.12	1.14	54.43	4.74	1.12
	0.9	0.05	0.77	0.27	-53.36	-4.86	-0.02	1.21	1.98	1.86	1.09	1.24	53.40	5.20	1.09
		0.10	3.32	0.96	-49.58	-7.72	0.01	1.37	2.25	2.17	1.02	1.67	49.63	8.02	1.02
		0.01	0.16	0.06	-54.09	-4.24	0.05	2.62	2.91	3.38	2.60	2.62	54.17	5.42	2.60
250	0.8	0.05	0.62	0.28	-53.22	-4.79	0.03	2.77	2.53	3.01	2.69	2.78	53.28	5.66	2.69
		0.10	3.19	1.24	-49.25	-7.29	0.26	2.83	2.89	3.33	2.59	3.09	49.34	8.01	2.61
		0.01	0.13	0.12	-54.01	-4.09	0.16	4.62	3.86	4.99	4.61	4.62	54.15	6.45	4.61
	0.7	0.05	0.60	0.25	-53.07	-4.80	0.01	4.68	3.33	4.74	4.57	4.69	53.17	6.75	4.57
		0.10	3.12	0.77	-48.97	-7.53	-0.12	5.00	4.06	5.19	4.86	5.06	49.14	9.15	4.86
		0.01	0.19	-0.03	-54.38	-4.42	0.00	0.78	1.55	1.69	0.77	0.78	54.41	4.73	0.77
	0.9	0.05	0.63	0.22	-53.76	-5.87	0.01	0.85	1.44	1.42	0.79	0.88	53.77	6.04	0.79
		0.10	3.27	0.97	-50.42	-8.91	0.03	0.99	1.55	1.75	0.77	1.39	50.44	9.08	0.77
		0.01	0.21	0.03	-54.38	-4.59	0.04	1.91	1.95	2.45	1.90	1.91	54.42	5.20	1.90
500	0.8	0.05	0.56	0.17	-53.68	-5.77	0.01	1.89	1.76	2.27	1.86	1.90	53.71	6.20	1.86
		0.10	3.85	1.02	-50.23	-8.87	0.10	1.97	1.90	2.50	1.88	2.22	50.27	9.21	1.89
		0.01	0.20	0.21	-54.27	-4.48	0.23	3.22	2.59	3.58	3.21	3.23	54.33	5.73	3.22
	0.7	0.05	0.56	0.40	-53.44	-5.60	0.34	3.29	2.42	3.41	3.25	3.31	53.49	6.56	3.26
		0.10	4.23	0.78	-50.10	-8.99	0.04	3.47	2.57	3.71	3.36	3.55	50.16	9.72	3.36
		0.01	0.21	0.00	-54.36	-5.20	0.02	0.53	1.12	1.24	0.52	0.53	54.37	5.34	0.52
	0.9	0.05	0.63	0.15	-54.07	-7.21	0.00	0.59	0.99	1.22	0.54	0.61	54.08	7.31	0.54
		0.10	3.83	0.89	-50.87	-10.57	-0.01	0.71	1.06	1.45	0.56	1.14	50.89	10.66	0.56
		0.01	0.27	-0.01	-54.41	-5.21	0.02	1.29	1.32	1.75	1.28	1.29	54.42	5.50	1.28
1000	0.8	0.05	0.64	0.03	-54.03	-7.25	0.03	1.31	1.22	1.73	1.29	1.31	54.04	7.46	1.29
		0.10	5.00	0.76	-50.82	-10.56	-0.01	1.45	1.35	1.96	1.35	1.63	50.84	10.74	1.35
		0.01	0.32	0.07	-54.38	-5.24	0.10	2.20	1.74	2.58	2.21	2.21	54.41	5.84	2.21
	0.7	0.05	0.60	-0.07	-53.80	-7.16	0.18	2.35	1.69	2.54	2.39	2.35	53.83	7.60	2.39
		0.10	6.10	0.43	-50.78	-10.74	-0.04	2.23	1.80	2.65	2.21	2.27	50.81	11.07	2.21

7.5.6 Conclusions

This simulation study for the practical applications of the linkage free dual system estimator and its modified version has several important points to summarize.

The linkage free dual system estimators can work very well, especially the modified version that utilizes the 1-to-1 linkage constraint. The performance of the modified linkage free estimator is as good or even better than the standard classification-based approach with clerical resolution, bearing in mind that the performance of the latter depends on the choice of acceptance and rejection thresholds.

The successful performance of the linkage free estimation largely depends on the estimation model taking into account the between-variables associations of the comparison outcomes, if those associations are present in the data. Model misspecification often leads to unacceptable estimates. Specifying a more complex model than needed is less of an issue than specifying a model that misses associations.

Identifiability of an estimation model is paramount in obtaining reliable parameter estimates. Estimates produced by a non-identifiable model may be nowhere near the true value of the parameter of interest.

While we are not focussed on the classification-based approach in this thesis and were only using it

for comparison purposes, the simulation results demonstrate its remarkable robustness. On the other hand, this approach lacks regularity in its behaviour and its performance depends on the choice of the thresholds, making it difficult to allocate the clerical resources and anticipate the extent of the error.

Given the performance of the modified linkage free dual system estimator is superior to the performance of the simple no-classification estimator in most scenarios, we would recommend to use it in the majority of real life applications. Simple linkage free estimates are faster to obtain and they constitute the basis for the modified estimates. Hence, the simple version may be used when fast preliminary estimates are needed.

7.6 Simulations verifying theoretical results

In this section we empirically verify the Taylor series approximation-based justification of the suitability of the mixture-like models for the no-classification record linkage and related population size estimation (Section 3.2). We also empirically assess the conjecture that it is possible to construct a data-conforming estimator with the help of the averaging blocking (Section 3.4). The underlying simulation approach is as described in Section 7.2. Unlike in the simulations assessing the practical applications of the linkage free dual system estimation, here we are only interested in the simulated data that have corresponding identifiable parameterizations. We are also only interested in the mixture-like model parameterizations closest to the simulated data.

There are three batches of simulations. The first one uses the same parameters $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ characterizing the distribution of the population attributes as in the simulations above (Section 7.2). The second batch only differs from the first by having the binary fourth linkage variable, v_4. In other words, the parameter $\boldsymbol{\rho}$ takes the value $(10000, 500, 500, 2)^T$ in this case. The third batch differs from the first by using a population where attributes have many distinct uniformly distributed values. That is, the parameter $\boldsymbol{\rho}$ takes the value $(10000, 1000000, 1000000, 1000000)^T$ and the parameter $\boldsymbol{\lambda}$ takes the value $(0, 0, 0, 0)^T$ in this case. Note, that it is unlikely to have that many distinct uniformly distributed attributes in a real situation (unless the attributes are telephone number, national insurance number, etc.), but such an extreme selection allows the anticipated behaviour to be checked more easily.

Across all three batches we generate data that aim at four types of the associations between the outputs of comparison outcomes (with the closest mixture-like model parameterization in brackets):

- between-variables independence in outcomes in both the set of matches and non-matches ($\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$);
- between-variables independence in the set of matches and dependence between v_1 and v_2 in the set of non-matches ($\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$);
- dependence between v_2 and v_3 in the set of matches and between-variables independence in the set of non-matches ($\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$);
- dependence between v_2 and v_3 in the set of matches and dependence between v_1 and v_2 in the set of non-matches ($\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$).

The results are presented in tables where the first three columns are the simulation parameters τ, π_j, ξ followed by four columns, each containing the relative bias of the four models under consideration. There are 1000 iterations of the simulations. Simulated data for each iteration are regarded as an output for a single block. Therefore, the blocks are perfectly same-sized for each parameter τ. The simulated data are averaged across all iterations achieving the perfect average blocking and the averaged data are passed to the estimation. We use ten random starts for the simulated annealing and the estimates are averaged across these ten trials. This is to minimize the effect of variation in the simulated annealing output for a fixed data set. Only the linkage free dual system estimator is used.

Table 27: Population attributes as in the main simulations, Section 7.5

			Relative bias			
τ	π_j	ξ	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$
		0.01	0.00	-0.08	-0.06	-0.04
	0.9	0.05	0.14	0.07	-0.25	-0.14
		0.10	0.00	0.00	0.27	0.00
		0.01	-0.04	0.08	-0.03	-0.01
250	0.8	0.05	-0.02	0.03	-0.17	-0.56
		0.10	0.07	0.01	-0.07	-0.12
		0.01	-0.03	0.08	-0.07	-0.19
	0.7	0.05	0.03	-0.18	0.25	-0.13
		0.10	0.00	-0.04	-0.35	-0.71
		0.01	-0.01	0.05	0.01	0.01
	0.9	0.05	0.02	-0.08	-0.38	-0.30
		0.10	-0.02	0.10	-0.38	-0.26
		0.01	0.15	0.04	0.12	-0.08
500	0.8	0.05	0.09	0.00	-0.31	-0.24
		0.10	0.03	-0.13	-0.21	-0.29
		0.01	-0.02	-0.02	-0.10	-0.01
	0.7	0.05	0.00	0.00	-0.22	-0.02
		0.10	0.26	-0.16	-0.12	-0.65
		0.01	0.00	0.01	-0.04	-0.05
	0.9	0.05	-0.02	0.09	0.04	0.17
		0.10	-0.04	0.06	0.14	0.10
		0.01	0.00	0.01	0.00	-0.05
1000	0.8	0.05	-0.03	-0.01	0.08	0.16
		0.10	0.04	-0.03	-0.07	-0.16
		0.01	0.13	-0.11	-0.12	-0.02
	0.7	0.05	-0.13	-0.01	-0.26	-0.34
		0.10	0.13	-0.08	-0.14	-0.08

We refer to Sections 3.2 and 3.4 for a detailed discussion of the mixture-like conceptualization of the record linkage and no-classification population size estimation as well as the idea of the average blocking. We repeat here that our analysis carried out in the relevant sections suggested that the mixture-like representation should yield a good approximation overall. Yet, we identified several cases, where this approximation may not be as accurate. Specifically, our analysis shows that as a model becomes more complex, the approximation becomes less accurate. This loss of accuracy manifests itself in the increased bias in the estimates of the linkage model and the related parameters, such as the linkage free dual system estimate of the population size parameter. The accuracy also drops if one or

more linkage variables has very few levels or, on the contrary, linkage variables can take excessively many distinct values.

Table 27 displays the results obtained using the distributional parameters of the population attributes identical to those in the simulations assessing the practical applications of the no-classification estimators. The between-variables independence model produces results that are very close to the true values of the parameters. It is hard to determine if there is any true under- or overestimation, as there is no clear pattern in the way results vary. Increasing the number of iterations may be helpful in this case, but we tried to keep the theoretical simulations in line with the practical, because the increase in iterations is computationally expensive. It is also quite likely that the simulation parameters such as population size and coverage probabilities also contribute to the results, but we did not attempt to analyse this contribution in Sections 3.2. In addition, we chose the distributional parameters of the population attributes to make a realistic population, rather than to find a set of parameters that would lead to unbiased estimators. It is likely, that there are no such parameters at all, and the estimator is always at least slightly biased.

Table 28: One binary population attribute

			Relative bias			
τ	π_j	ξ	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_{2,3},\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_{2,3},\gamma_3,\gamma_4)$
		0.01	-0.01	-1.17	-0.03	-0.16
	0.9	0.05	0.11	0.48	-0.34	0.76
		0.10	-0.01	-0.21	0.73	1.43
		0.01	-0.06	-1.36	-0.12	0.01
250	0.8	0.05	-0.10	0.88	-0.24	-0.65
		0.10	0.12	-0.37	0.26	0.43
		0.01	-0.13	-2.04	-0.09	-1.30
	0.7	0.05	0.10	0.50	-0.05	0.03
		0.10	-0.01	0.27	-0.24	-0.83
		0.01	-0.01	-0.04	0.02	0.16
	0.9	0.05	0.04	0.67	-0.37	-0.39
		0.10	-0.24	0.17	-0.44	-0.03
		0.01	0.10	-0.74	-0.14	0.37
500	0.8	0.05	0.02	0.06	-0.23	-0.01
		0.10	-0.11	0.34	0.03	0.37
		0.01	0.03	-0.86	-0.04	0.24
	0.7	0.05	-0.06	0.58	-0.26	-0.35
		0.10	0.11	0.07	0.04	0.32
		0.01	0.02	-0.31	-0.12	-0.37
	0.9	0.05	-0.01	0.46	0.17	-0.03
		0.10	-0.04	0.12	0.41	0.62
		0.01	0.04	-0.10	-0.06	0.26
1000	0.8	0.05	-0.05	-0.06	0.04	0.62
		0.10	-0.04	0.11	-0.10	-0.27
		0.01	-0.21	0.10	-0.09	-0.42
	0.7	0.05	0.05	0.39	-0.21	-0.62
		0.10	-0.03	0.33	-0.11	-0.65

As the model becomes more complex, the accuracy drops and mainly negative bias is incurred. For the model with a single association in the set of non-matches, there is no clear tendency towards

negative bias. However, for the model with a single association in the set of matches and the model with associations in both sets, this tendency is more prominent. We observe that not all scenarios lead to negative bias, which again suggests that other simulation parameters may influence the results.

Table 28 presents the results of the simulations with the population where the fourth attribute, corresponding to the variable v_4, is binary. There is a slight but not substantial difference for the between-variables independence model between this batch of the simulations and the one considered above. However, as anticipated by our theoretical analysis in Section 3.2 once a binary attribute is present and an estimation model becomes more complex, the approximation becomes less accurate. This is clearly visible for all three models with the between-variables associations. The interesting feature of these results is that there is no clear pattern in the bias. As already suggested, it is likely that there is a complex interplay between the parameters $\boldsymbol{\rho}, \tau, \pi_j$ and $\boldsymbol{\xi}$ that has not been analysed.

Table 29: Population attributes with excessively many uniformly distributed values

			Relative bias			
τ	π_j	ξ	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$	$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$
		0.01	0.00	-0.59	0.09	-0.56
	0.9	0.05	0.07	-0.75	0.14	-0.65
		0.10	0.28	-1.21	0.47	-0.06
		0.01	0.01	-0.63	0.07	-0.60
250	0.8	0.05	0.08	-0.85	0.09	-0.77
		0.10	0.19	-1.42	0.55	-1.01
		0.01	-0.18	-0.33	0.08	-0.61
	0.7	0.05	0.29	-0.87	0.44	-0.84
		0.10	0.09	-1.10	0.52	-1.73
		0.01	0.02	-0.36	0.07	-0.32
	0.9	0.05	0.28	-0.65	0.21	-0.59
		0.10	0.17	-1.05	0.26	-0.96
		0.01	-0.01	-0.23	0.02	-0.33
500	0.8	0.05	0.21	-0.71	0.14	-0.45
		0.10	0.33	-1.08	0.32	-1.07
		0.01	0.18	-0.38	0.02	-0.21
	0.7	0.05	0.20	-0.91	0.12	-0.46
		0.10	0.24	-1.07	0.59	-0.67
		0.01	-0.03	-0.18	0.06	-0.13
	0.9	0.05	0.28	-0.52	0.49	-0.41
		0.10	0.38	-0.81	0.35	-0.43
		0.01	0.02	-0.18	0.07	-0.25
1000	0.8	0.05	0.20	-0.45	0.48	-0.35
		0.10	0.30	-0.84	0.51	-0.76
		0.01	0.08	-0.13	0.10	-0.13
	0.7	0.05	0.34	-0.36	0.12	-0.63
		0.10	0.52	-0.73	0.54	-0.53

Finally, we look at the behaviour of the well-constructed linkage free dual system estimator in the case when there are many distinct uniformly distributed values of the population attributes, the results are displayed in Table 29. As predicted, the accuracy of the mixture-like model based estimation decreases. Positive bias is most frequently incurred in the cases of between-variables independence and the between-variables association of the comparison outcomes in the set of matches. Meanwhile,

between-variables association in the set of non-matches and between-variables association in both sets result in negative bias.

7.6.1 Conclusions

The simulation results support our analysis presented in Sections 3.2 and 3.4. It is possible to construct the data-conforming no-classification dual system estimator by the means of averaging blocking. While the simulations show that this estimator is (as expected) biased, conceptually this estimator corresponds better to the mixture-like conceptualization than the estimator without averaging blocking and therefore in theory has a superior performance (compare with the results presented in Section 7.5).

We also confirmed that the accuracy of the mixture-like representation of record linkage data depends on the complexity of an underlying model, in terms of the number of associated comparison patterns, and the number of unique values the population attributes can take. The more complex is the model specification, the less accurate is the mixture-like approximation. Also, the accuracy of the approximation drops when the population attributes have very few unique values or, on the contrary, when the attributes tend to be unique for every element of the population.

This study suggests that there are other factors contributing to the accuracy of the mixture-like parameterization, but which our analysis did not account for. In future, a larger number of simulation iterations would be helpful to discern the trends more clearly. Also, a more detailed investigation of how the entire ensemble of the parameters work together and affect the approximation may be carried out.

7.7 Comparing the data generated according to the linkage experiment against a parametric approach

A two component mixture of the probability mass functions of binary random variables (8) implies that for a fixed number of draws the observed frequency of the comparison patterns follows a multinomial distribution with cell probability $\mathrm{pr}(\boldsymbol{\gamma}_p)$. However, our discussions on the linkage experiment (Section 2.2.5) and justification of the mixture-like model (Section 3.2) demonstrate that the standard mixture is not an appropriate model for record linkage and related problems. In this section, we present some empirical results providing an idea how large the discrepancy between the record linkage data generated from 'first principles' mimicking the linkage experiment (including within-variables dependence) and the parametrically generated data resulting in within-variables independent record pairs can be.

We should bear in mind, that there may be more than one way of demonstrating the discrepancies between the record linkage data generated from the 'first principles' and the parametric approach. While the standard mixture model relevant to this thesis generates multinomial data, any multinomial distribution requires a fixed number of trials. The number of trials corresponds to the number of record pairs in the context of record linkage. However, we generally have varying numbers of record pairs due to the variability of the survey sizes while keeping the coverage probabilities fixed. We are not interested in the conditional distribution given the fixed sizes of the two surveys whenever our main interest is the population size estimation. Hence, there is a room for various alternative approaches that generate independent observations for a non-fixed number of trials.

We approach the task in the following way. Once the data are generated from the 'first principles', we can compute the standard deviation of the frequency of each of the comparison patterns across 1000 simulation iterations. We can also compute the probability of a record pair being in each pattern, or a cell in the standard categorical data language, using the same data. We can imagine that there exists a probability distribution, possibly quite complex, that uses the above vector of cell probabilities as its fixed parameter.

Table 30: Standard deviations of the simulated data cells vs parametric approach with within-variables independence: $\tau = 500, \pi_1 = \pi_2 = 0.9$

		$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$		$\pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$		$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$		$\pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4)$ $+(1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$	
ξ	$\boldsymbol{\gamma}$	sd(S)	sd(M)	sd(S)	sd(M)	sd(S)	sd(M)	sd(S)	sd(M)
0.01	0000	6671.90	6671.09	7033.22	7045.09	7120.83	7118.19	7126.79	7145.57
	0001	88.20	67.19	94.05	70.95	94.17	71.69	94.30	71.87
	0010	75.85	33.44	76.52	35.29	76.65	35.78	75.99	36.08
	0011	4.39	0.34	4.35	0.36	4.15	0.37	4.33	0.36
	0100	76.27	33.62	143.80	36.02	75.83	35.79	144.02	36.22
	0101	4.33	0.35	4.34	0.37	4.15	0.37	4.49	0.36
	0110	3.10	0.17	2.99	0.18	2.98	0.18	3.09	0.19
	0111	5.35	0.25	5.37	0.27	4.59	0.20	4.64	0.20
	1000	6.12	0.67	45.38	9.95	6.25	0.71	46.69	10.15
	1001	0.71	0.01	2.35	0.10	0.61	0.01	2.49	0.11
	1010	0.57	0.01	1.67	0.06	0.49	0.01	1.78	0.06
	1011	2.65	0.25	2.71	0.27	2.75	0.26	2.69	0.26
	1100	0.57	0.01	60.84	22.68	0.49	0.01	62.77	23.33
	1101	2.83	0.26	4.51	0.50	2.67	0.26	4.48	0.49
	1110	2.77	0.25	3.80	0.39	2.62	0.25	3.86	0.38
	1111	14.63	12.60	15.25	13.35	14.98	13.56	15.00	13.66
0.05	0000	6889.83	6887.89	6960.17	6976.37	6987.70	6986.72	7065.39	7072.94
	0001	86.94	68.13	91.32	68.95	89.77	69.16	92.62	70.07
	0010	71.36	32.67	71.72	33.08	69.57	33.15	71.24	33.38
	0011	4.30	0.43	4.22	0.43	4.02	0.38	4.14	0.38
	0100	71.84	32.54	142.73	35.57	73.01	33.09	140.04	35.59
	0101	4.50	0.43	4.60	0.45	4.18	0.38	4.44	0.41
	0110	3.32	0.26	3.39	0.28	3.13	0.23	3.16	0.23
	0111	9.19	0.98	9.02	0.97	8.43	0.78	8.26	0.80
	1000	6.49	0.70	47.05	10.62	6.23	0.72	47.81	11.05
	1001	1.87	0.12	2.78	0.22	2.75	0.27	3.49	0.38
	1010	1.76	0.11	2.37	0.16	1.82	0.11	2.21	0.15
	1011	5.50	1.01	5.48	1.04	5.10	0.90	5.20	0.92
	1100	1.87	0.11	57.77	18.76	1.70	0.10	55.58	19.74
	1101	5.24	1.01	6.59	1.21	4.92	0.90	5.95	1.11
	1110	5.13	1.00	5.74	1.12	5.60	1.06	5.76	1.16
	1111	15.27	9.38	15.16	9.56	15.26	9.96	15.27	10.13
0.10	0000	7276.55	7277.48	7383.38	7385.21	6781.68	6782.68	7017.72	7027.00
	0001	88.78	70.60	91.91	71.68	89.17	65.91	92.96	68.40
	0010	71.69	32.08	65.96	32.39	66.95	29.96	68.71	30.90
	0011	5.02	0.63	4.93	0.65	4.82	0.53	4.58	0.55
	0100	68.54	31.99	127.03	37.07	68.19	29.89	129.63	35.00
	0101	4.95	0.64	5.34	0.69	4.61	0.53	5.22	0.58
	0110	4.12	0.46	4.38	0.49	4.11	0.43	4.38	0.47
	0111	9.31	1.38	9.30	1.42	9.25	1.33	9.22	1.39
	1000	6.54	0.80	47.83	11.79	6.99	0.80	46.51	11.24
	1001	3.25	0.37	4.03	0.47	4.32	0.63	4.80	0.76
	1010	3.20	0.36	3.43	0.42	2.70	0.26	3.10	0.31
	1011	6.54	1.53	6.41	1.55	5.97	1.14	5.79	1.17
	1100	3.27	0.37	52.58	16.03	2.76	0.27	48.06	15.42
	1101	6.57	1.53	7.15	1.70	5.79	1.12	6.23	1.34
	1110	6.66	1.52	6.67	1.61	6.42	1.47	6.81	1.61
	1111	13.40	6.53	13.76	6.68	13.95	6.50	13.98	6.75

There are other parameters of this distribution, but we do not know them. Then, given the number of the record pairs for each iteration of the first principles approach, we use this imagined distribution to generate exactly the same data we as generated from the first principles. Then the standard deviation computed above is also the standard deviation for this distribution. Note, that this distribution preserves the associations between individual observations, if such exist. Now we take the same vector of the cell probabilities as the parameter of the multinomial distribution. For each iteration, we use the number of record pairs used in the 'first principles' approach as the number of trials and generate the data that follow the multinomial distribution. This time, however, we know that for each iteration individual data points are independent (in terms of within-variables independence). We then work out the standard deviation of the frequencies of each comparison pattern obtained in this way. If these two distributions are not the same, then comparing the two sets of standard deviations should result in substantial differences between the 'first principles' based data (or imagined distribution) and the multinomial data with a varying number of trials.

Results are displayed in Table 30. We limit presentation to the case where the population size is $\tau = 500$ individuals and coverage probabilities are $\pi_1 = \pi_2 = 0.9$, but vary the errors, which are in the first column of the table. The second column contains the comparison patterns $\boldsymbol{\gamma}$. The rest of the table is split into four blocks, one for each type of identifiable model, as in the case of the verification of the theoretical results in Section 7.6. For each model, two values of the standard deviations are reported. The first one sd(S) is for the data simulated from the 'first principles' while sd(M) is for the parametric approach.

The results demonstrate that, apart from the comparison pattern with all disagreements, (0000), there is a substantial difference between the data generated from the first principles, which should be reasonably close to what one expects to encounter when dealing with record linkage tasks, and the data generated by the parametric approach outlined above. When the probabilities of making errors recording the population attributes are low, the standard deviation for the pattern with all agreements, (1111), also has a good degree of similarity. However, as these probabilities increases, the discrepancy increases.

These results show the value in simulating the data from 'first principles', rather than relying on purely parametric simulations. Also, these results support the use of the parameter estimation method that does not assume independence between record pairs (also known as the within-variables independence) as well as relying on the averaging blocking in variance estimation instead of an attempt to bootstrap the observed pairs.

8 Simulation study for variance estimation

Finally, we present the result of the simulation study assessing the performance of the variance estimators for the linkage free dual system estimator and its modified version (see Chapter 6). Recall, that two methods were developed for practical applications. The basic one with no auxiliary data (Section 6.2) and the second one which seeks to enhance the performance using the auxiliary data such as the address frame or address listing (Section 6.3).

The simulation approach and parameters are as described in Section 7.2. However, we are only considering a subset of scenarios that produce the data aiming at independence and at the between-variables associations with the identifiable parameterizations of the mixture-like model. We consider only those estimation models that have the closest parameterizations to the generated data: between-variables independence in outcomes in both the set of matches and non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$; between-variables independence in the set of matches and dependence between v_1 and v_2 in the set of non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$; dependence between v_2 and v_3 in the set of matches and between-variables independence in the set of non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$; dependence between v_2 and v_3 in the set of matches and dependence between v_1 and v_2 in the set of non-matches, $\pi(\boldsymbol{\gamma}_p;\pi,\boldsymbol{\mu},\boldsymbol{\nu}) = \pi\mu_p(\gamma_1,\gamma_{2,3},\gamma_4) + (1-\pi)\nu_p(\gamma_{1,2},\gamma_3,\gamma_4)$.

Table 31: Variance estimation: between-variables independence in both the set of matches and the set of non-matches

			Emp		Est				Est aux			
τ	π_j	ξ	$\text{Var}(\widetilde{\tau})$	$\text{Var}(\widetilde{\tau}_c)$	$\widehat{\text{Var}}(\widetilde{\tau})$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}))$	$\widehat{\text{Var}}(\widetilde{\tau}_c)$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}_c))$	$\widehat{\text{Var}}(\widetilde{\tau})$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}))$	$\widehat{\text{Var}}(\widetilde{\tau}_c)$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}_c))$
		0.01	9	8	48	71	49	76	21	32	20	32
	0.9	0.05	28	11	59	96	47	77	33	49	20	30
		0.10	66	18	114	171	52	78	81	135	28	42
		0.01	45	44	142	213	154	235	74	112	76	120
250	0.8	0.05	63	48	163	247	139	212	95	147	70	109
		0.10	136	61	222	345	140	227	143	214	78	125
		0.01	139	140	274	422	311	568	171	256	184	302
	0.7	0.05	157	134	357	610	330	571	224	348	196	314
		0.10	265	154	439	687	284	437	282	423	172	245
		0.01	17	15	101	96	102	101	45	39	42	39
	0.9	0.05	56	28	124	114	99	94	70	61	43	40
		0.10	104	42	223	216	113	114	151	139	55	51
		0.01	93	89	281	262	301	285	145	142	146	149
500	0.8	0.05	149	106	325	320	304	305	184	174	153	146
		0.10	224	127	491	476	295	295	311	279	160	150
		0.01	262	260	631	694	714	843	378	385	409	447
	0.7	0.05	343	299	718	787	702	785	452	466	412	459
		0.10	492	317	928	991	627	710	614	580	379	360
		0.01	38	34	200	124	203	128	87	54	83	52
	0.9	0.05	106	61	251	158	200	132	142	91	88	53
		0.10	217	110	444	271	213	136	304	188	108	62
		0.01	192	189	569	354	611	387	287	168	292	169
1000	0.8	0.05	268	202	680	456	620	423	366	236	296	192
		0.10	395	275	966	646	602	405	606	385	322	204
		0.01	586	584	1285	984	1427	1123	736	488	784	539
	0.7	0.05	690	585	1455	1176	1382	1132	896	623	788	560
		0.10	896	667	1869	1330	1276	893	1205	801	760	469

As in all the simulations seen so far, we rely on 1000 iterations. For the data simulated at each iteration, the corresponding variance estimate is obtained and those estimates are then summarized. We are interested in the mean of the variance estimates and the standard deviation of the distribution of these estimates.

The results in tables are organized in the following way. The first three columns are the simulation parameters τ, π_j and ξ as before. The remaining columns are organized into three blocks. The first block, titled 'Emp', contains the empirical variance of the linkage free dual system estimator, $\text{Var}(\widetilde{\tau})$, and the empirical variance of the modified linkage free estimator, $\text{Var}(\widetilde{\tau}_c)$. By empirical variance we mean the variance of the estimates obtained by a specified estimator in the simulation study in Section 7.5. We sometimes refer to these values as the true variance. The second block, titled 'Est' contains the results for the simple variance estimator with no auxiliary data. The columns $\overline{\widehat{\text{Var}}}(\widetilde{\tau})$ and $\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}))$ are the mean and standard deviation for the distribution of the variance estimates of the linkage free dual system estimator. The columns $\overline{\widehat{\text{Var}}}(\widetilde{\tau}_c)$ and $\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}_c))$ are the mean and standard deviation for the distribution of the variance estimates of the modified linkage free dual system estimator. The third bock, titled 'Est aux' is similar to the second block, but contains the results for the variance estimation with the auxiliary data.

Table 32: Variance estimation: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches

			Emp		Est				Est aux			
τ	π_j	ξ	$\text{Var}(\widetilde{\tau})$	$\text{Var}(\widetilde{\tau}_c)$	$\overline{\widehat{\text{Var}}}(\widetilde{\tau})$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}))$	$\overline{\widehat{\text{Var}}}(\widetilde{\tau}_c)$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}_c))$	$\overline{\widehat{\text{Var}}}(\widetilde{\tau})$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}))$	$\overline{\widehat{\text{Var}}}(\widetilde{\tau}_c)$	$\text{sd}(\widehat{\text{Var}}(\widetilde{\tau}_c))$
		0.01	24	11	70	107	52	83	41	59	23	34
	0.9	0.05	51	17	96	147	57	93	67	95	31	44
		0.10	117	39	181	273	71	108	140	206	47	72
		0.01	66	52	176	272	160	249	102	151	79	115
250	0.8	0.05	100	59	195	304	141	213	130	201	81	124
		0.10	197	92	310	500	158	263	214	338	105	166
		0.01	163	148	339	506	340	507	228	327	205	305
	0.7	0.05	198	157	384	624	312	541	260	403	200	324
		0.10	352	203	602	972	335	527	421	674	230	344
		0.01	47	23	133	125	102	104	79	68	44	40
	0.9	0.05	97	39	182	161	110	97	127	112	57	52
		0.10	206	85	323	287	132	115	258	233	92	84
		0.01	121	92	342	317	317	301	201	184	160	146
500	0.8	0.05	210	121	400	362	302	288	255	227	167	154
		0.10	345	180	606	555	314	289	431	388	209	194
		0.01	331	288	693	840	688	920	447	452	404	445
	0.7	0.05	457	334	772	796	625	653	525	490	400	370
		0.10	665	429	1103	1174	600	615	777	770	426	413
		0.01	101	49	271	160	207	133	161	92	89	53
	0.9	0.05	197	91	364	215	224	149	251	155	120	83
		0.10	382	189	664	415	280	185	527	306	197	137
		0.01	256	194	686	430	631	415	383	231	299	193
1000	0.8	0.05	404	235	837	561	634	438	513	302	332	201
		0.10	667	403	1254	823	658	432	864	517	427	273
		0.01	625	537	1459	1033	1433	1080	917	598	807	532
	0.7	0.05	825	622	1633	1172	1349	1017	1063	669	826	534
		0.10	1091	749	2202	1604	1201	884	1555	1035	843	542

Table 31 contains the results for the simulation / estimation model of the between-variables independence of the comparison outputs in both the set of matches and non-matches. This table contains all the characteristic trends that will be visible in all of the simulation / estimation models. We can see that for the simple approach without the auxiliary data, the mean of the variance estimates indicates a

substantial overestimation of the variance for both the linkage free and the modified estimators. This is attributable to all the additional variability in the variance estimation process that is not present in the no-classification point estimation (see again the discussion in Section 6.2). On the positive side, there is a good correlation between the true variance and the variance estimation obtained without the auxiliary data. The standard deviation of the variance estimates is large. This is due to some extreme results (a few very low and very high estimates) being present in the distribution.

The approach with the auxiliary data, despite resulting in overestimation of the variance, demonstrates a fair performance for both estimators. On average, the variance estimates are considerably closer to the true variance. The standard deviations of the variance estimates are still large even when the auxiliary data are used. The reason is the same as in the case without the auxiliary data: a few extreme estimates.

Table 33: Variance estimation: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches

			Emp		Est				Est aux			
τ	π_j	ξ	$\mathrm{Var}(\widetilde{\tau})$	$\mathrm{Var}(\widetilde{\tau}_c)$	$\widehat{\overline{\mathrm{Var}}}(\widetilde{\tau})$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}))$	$\widehat{\overline{\mathrm{Var}}}(\widetilde{\tau}_c)$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}_c))$	$\widehat{\overline{\mathrm{Var}}}(\widetilde{\tau})$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}))$	$\widehat{\overline{\mathrm{Var}}}(\widetilde{\tau}_c)$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}_c))$
		0.01	9	8	47	71	48	73	21	29	19	27
	0.9	0.05	27	11	62	91	49	74	34	51	21	31
		0.10	63	25	120	185	63	92	77	117	31	47
		0.01	45	44	127	187	139	211	73	106	75	111
250	0.8	0.05	63	47	148	214	139	200	82	115	67	97
		0.10	115	66	224	349	153	245	130	194	79	121
		0.01	117	122	288	494	334	586	180	301	202	356
	0.7	0.05	141	129	317	506	314	499	190	291	182	295
		0.10	212	149	364	637	299	588	237	367	173	335
		0.01	18	14	101	99	104	110	47	44	44	44
	0.9	0.05	46	23	117	115	102	99	61	55	42	39
		0.10	108	49	193	174	110	112	129	118	55	53
		0.01	87	85	271	284	295	317	137	121	140	129
500	0.8	0.05	125	98	327	343	317	328	175	181	155	150
		0.10	228	132	422	410	294	296	259	235	159	153
		0.01	262	259	577	603	656	720	346	332	377	378
	0.7	0.05	325	288	653	700	632	680	418	396	378	363
		0.10	449	317	886	994	635	697	567	631	385	379
		0.01	40	30	196	130	202	137	89	61	83	58
	0.9	0.05	95	63	252	160	202	126	140	87	90	54
		0.10	227	139	418	288	233	156	279	191	124	76
		0.01	178	166	572	417	630	475	290	191	302	210
1000	0.8	0.05	266	209	667	476	600	458	374	241	304	216
		0.10	411	295	901	640	611	445	553	354	334	210
		0.01	556	556	1217	843	1402	1043	712	484	780	559
	0.7	0.05	602	532	1313	944	1296	931	796	532	743	516
		0.10	793	643	1769	1262	1224	844	1144	740	740	466

The results for the models with various between-variables associations of the comparison outcomes, displayed in Table 32, Table 33 and Table 34 are similar to those observed for the between-variables independence model. The approach with no auxiliary data provides only indicative estimates, in the sense that if the estimate is low, then one can expect that the true variance is also low. The variance estimator with the auxiliary data enhanced approach performs well, but suffers from sporadic extreme

variance estimates.

Overall, we can claim that the variance estimation approach with the auxiliary data appears to be promising, given that there are not many attempts to estimate the variance of the linkage parameters in the frequentist setting and without unrealistic distributional assumptions. There is an issue of extreme variance estimates that needs further research.

Table 34: Variance estimation: association between the second and third variable in the set of matches, association between the first and second variable in the set of non-matches

			Emp		Est				Est aux			
τ	π_j	ξ	$\mathrm{Var}(\widetilde{\tau})$	$\mathrm{Var}(\widetilde{\tau}_c)$	$\widehat{\mathrm{Var}}(\widetilde{\tau})$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}))$	$\widehat{\mathrm{Var}}(\widetilde{\tau}_c)$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}_c))$	$\widehat{\mathrm{Var}}(\widetilde{\tau})$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}))$	$\widehat{\mathrm{Var}}(\widetilde{\tau}_c)$	$\mathrm{sd}(\widehat{\mathrm{Var}}(\widetilde{\tau}_c))$
		0.01	24	11	68	103	52	80	40	62	24	35
	0.9	0.05	52	19	96	145	56	88	67	106	31	52
		0.10	101	44	156	231	69	118	122	174	49	77
		0.01	66	50	151	235	140	229	97	149	78	121
250	0.8	0.05	92	53	182	259	139	206	113	167	76	119
		0.10	142	76	273	432	151	223	189	270	101	149
		0.01	159	147	336	535	334	560	208	320	187	296
	0.7	0.05	191	153	361	578	303	468	229	355	180	274
		0.10	279	179	473	991	284	514	342	666	215	362
		0.01	49	22	134	123	103	98	79	71	44	42
	0.9	0.05	99	36	175	151	111	107	115	99	56	51
		0.10	178	84	276	246	133	119	209	196	84	80
		0.01	120	92	328	320	308	317	194	171	153	140
500	0.8	0.05	190	119	360	325	273	245	235	198	159	140
		0.10	329	186	570	579	303	299	397	388	201	191
		0.01	298	264	676	682	669	730	448	442	385	387
	0.7	0.05	389	306	773	773	633	663	524	466	406	392
		0.10	564	383	1036	1112	535	562	696	676	363	328
		0.01	97	49	279	180	210	145	170	104	92	58
	0.9	0.05	215	98	366	218	222	147	251	146	117	78
		0.10	414	227	594	348	275	172	462	266	191	123
		0.01	257	188	690	519	647	506	409	264	324	227
1000	0.8	0.05	401	258	805	554	596	435	511	292	328	194
		0.10	599	387	1120	741	601	394	775	496	393	250
		0.01	655	578	1349	1007	1375	1093	884	548	807	542
	0.7	0.05	754	564	1533	1041	1283	924	990	603	785	512
		0.10	1122	816	2184	1670	1169	865	1470	941	797	503

9 Summary, conclusions and future work

We finish this thesis with a short summary of the achieved results, concluding notes and several open questions as well as suggested areas where improvements can be made.

The conceptual closeness of the automated record linkage problem and dual system estimation was explored in order to derive a population size estimator that seamlessly integrates record linkage and dual system estimation within a single framework. The main contribution of this thesis to the existing corpus of methods and knowledge in the fields of record linkage and dual system estimation is the development of the no-classification dual system estimator. This estimator is derived from a purely estimation-based record linkage that does not classify records into links and non-links. The

proposed approach is fully automated so that there is no need for a clerical resolution of record pairs that neither have a strong evidence of being links nor a strong evidence of being non-links. This is the main practical advantage of the developed method over the majority of existing approaches. We also developed a modified version of the linkage free dual system estimator and this modified approach demonstrates a remarkable performance. We made several auxiliary developments needed to enable no-classification dual system estimation in this thesis. These developments include modelling of record linkage data, explaining the nature of associations between comparison and identifiability issues. While this thesis was being written, a research project focusing on the closeness of record linkage and capture-recapture had appeared (Tancredi et al., 2020) and contributed to the topic from the Bayesian perspective. All the developments in this thesis are within the frequentist paradigm and have a benefit of relative conceptual and practical simplicity.

In order to appropriately conceptualize the task of record linkage (Fellegi & Sunter, 1969), the notion of a mixture-like model was introduced in this thesis. Unlike regular mixture-models, the mixture-like model more adequately represents the complex nature of the record linkage problem without making or implying strong distributional assumptions, as in the cases of Jaro (1989); Winkler & Thibaudeau (1991); Larsen & Rubin (2001); Larsen (2005). The mixture-like models allow the assessment of how accurately a model can approximate the outcomes of the linkage experiment (see Section 2.2.5) and how the accuracy of approximation is related to the parameters that are not explicitly taken into account by the linkage model, such as parameters defining the distribution of the values of the population attributes. The mixture-like approach also demonstrates certain limitations related to the parameter estimation of record linkage models and functions of these parameters, such as the linkage free dual system estimator. For the actual parameter estimation we proposed to use simulated annealing, which is a well-established Markov chain Monte Carlo approach. This approach is attractive since it does not impose unrealistic distributional assumptions and is relatively easy to implement and adapt for different models.

A special case of blocking, called the averaging blocking, was proposed in the thesis. Unlike the usual blocking strategies that aim to reduce the number of pairs in a linkage process (Herzog et al., 2007, chap. 12; Christen, 2012, chap. 4), averaging blocking allows to construct, at least in principle, a linkage free dual system estimator as well as estimators of the parameters of record linkage models that accord with the data generating mechanisms in repeated record linkage tasks. In addition, averaging blocking enables a framework for variance estimation.

Another contribution of this thesis is the development of two variance estimation methods for the linkage free dual system estimator and its modified version. These methods assume that the data for each block in the averaging blocking are produced by the identical data generating mechanism, but unlike the existing methods (Chipperfield & Chambers, 2015) do not make any distributional assumptions about the mechanism itself. The proposed methods are also suitable for variance estimation of the linkage model parameters.

From the outset of the theory of record linkage (Fellegi & Sunter, 1969), there were concerns about the comparison outcomes of the linkage variables not being independent and several approaches to deal with between-variables associations were proposed by Winkler (1993); Armstrong & Mayda

(1993); Thibaudeau (1993); Belin & Rubin (1995); Larsen & Rubin (2001). In this thesis we explicitly show how between-variables associations of the comparison outputs emerge in the set of matches and non-matches. Particular combinations of the population attributes that lead to such associations were discussed. We also presented the corresponding parameterizations of the mixture-like models that account for these associations.

Complex models with between-variables associations of the comparison outcomes may or may not be identifiable. Identifiability of record linkage models was largely overlooked in the past. In this thesis we reviewed and applied a variety of methods from the field of algebraic statistics to check the identifiability of several models that are expected to be frequently used in practical applications of the proposed no-classification approaches to population size estimation.

Finally, a realistic simulation study, in which data were predominantly simulated from 'first principles' rather than using parametric distributions, was conducted. In this study the performance of the linkage free dual system estimator and its modified version was assessed against the usual classification-based approach and the dual system estimator with perfect linkage. As a by-product, certain features of the classification-based approach were also assessed.

The results of the simulation study demonstrate that all the individual pieces of the research listed above fit together. The most important outcome of this thesis is a demonstration that the linkage free dual system estimation is feasible. More specifically, the fully automated modified no-classification estimation displays a promising performance. This performance is close to and often better than the performance of the classical approach with classification, subject to the choice of thresholds for the classification-based method. The key benefit of the no-classification methods developed in this thesis is full automation that makes clerical resolutions redundant. Also, the no-classification methods, assuming the correct estimation model is specified, produce the objectively best outputs given the input data. The 'objectively best' here is assessed through the minimization of the modified chi-squared statistic. On the contrary, the classification-based approaches have a certain level of subjective choice involved in specifying the acceptance and rejection thresholds. The ability to automatically complete a record linkage task by only maximizing some objective function makes the methods developed in this thesis akin to recent developments in the field of automated linkage (Lee et al., 2022).

While the work presented in this thesis is self-sufficient in the sense that it answers the key questions enabling no-classification dual system estimation to be used, there are many questions of a theoretical and practical importance that are left for future research. We list several questions that, in our opinion, are worth attention.

Dual system estimation, whether it is employed with the usual classification-based linkage or using the no-classification approaches, only adjusts for undercoverage, or non-response, in two surveys. It is often, however, the case that one of the surveys is subject to overcoverage and the corresponding adjustment of the population size estimate is needed. Therefore, extending the no-classification approaches to deal with overcoverage is one of the main strands of the future research. Overcoverage estimation usually involves linking records outside small geographical units. Therefore, the notion of the averaging blocking may need revision in this case.

In Section 4.2 we derived the modified linkage free dual system estimator (58) for the domains such

as age-sex group by the coverage survey strata. However, in the simulation study we only assessed the no-classification approaches applied at the population level. Assessment of the linkage free domain estimator is an important area of research, since in practice we are often interested in such domain estimates.

Another domain related research question concerns the fact that dual system estimation may be biased if the data used in estimation are pooled over domains with variable responses; see Section 2.1.2. Given that dual system estimation and record linkage are closely related, further research could attempt to answer the following question: may pooling the data over the domains with variable responses result in heterogeneity bias in the estimates of the record linkage model parameters or not?

In Section 4.2 we also showed how a weight (54) reflecting the contribution of each individual record in a given survey to the total number of the estimated links can be derived. At the moment, these weights are always positive, but can be greater than one. Ideally, we would like to have weights lying strictly within the interval $[0, 1]$. Forcing weights into this interval can also be considered as an additional constraint, that can further improve the modified linkage free dual system estimator.

When justifying the use of the mixture-like models for record linkage and the dual system estimation, we considered how certain population parameters affect the mixture-like approximation. The simulation work indicates that there may be more parameters influencing the accuracy of the approximation, and a more complicated relationship between the parameters may exist than that discovered so far. Also, the justification part is not fully formalized at the current stage of research and obtaining a rigorous demonstration is an outstanding task.

We did not manage to check identifiability for all the models considered in this thesis. The work on identifiability can be continued, especially as new developments in the field of computational algebra become available.

So far, we relied on a heuristic approach to model selection. This approach is based on the knowledge of the underlying population structure and attributes used in record linkage and population size estimation. We would like to have more formal procedures, or at least some indicators of whether a certain model fits the data well or not. While working on the simulation study, we observed that if a model is well specified or is more complex than needed, then the chi-squared statistic is very small and the estimated frequencies for each pattern are very close to the observed frequencies, whereas in the cases of model misspecification, where insufficiently complex model was used, the chi-squared statistic was very large and a substantial discrepancy between the observed and estimated frequencies was present.

The precision of the variance estimators developed in this thesis suffered from extreme estimates. Finding a way to avoid such extreme cases would make the variance estimates obtained by the approach with the auxiliary data substantially more accurate.

Simulation work can be improved so that more complex between-variables associations of the comparison outcomes become available. Some additional functionality, such as correlation between the errors for the members of the same household, would make simulations even more realistic.

The no-classification population size estimation as developed in this thesis focuses on the combination of census and the census coverage survey or similar high quality population surveys. Not all

population size estimation tasks can be supported by such high quality data. Therefore, it would be interesting to investigate, whether the no-classification population size estimation can be adapted to work with other types of data.

Finally, assessing the linkage free dual system estimators on real data would allow to identify any features that are potentially missing in the current framework and any further developments needed to make the linkage free dual system estimation fully functional in practice.

References

Agresti A. (2002). *Categorical data analysis*. Second edition. John Wiley & Sons.

Alho J. (1990). Logistic regression in capture-recapture models. *Biometrics*, *46*(3), 623–635. https://doi.org/10.2307/2532083

Allman E. S., Cervantes H. B., Hosten S., Kubjas K., Lemke D., Rhodes J., & Zwiernik P. (2019). Maximum likelihood estimation of the latent class model through model boundary decomposition. *Journal of Algebraic Statistics*, *10*(1), 51–84. https://doi.org/10.18409/jas.v10i1.75

Allman E. S., Matias C., & Rhodes J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*(6A), 3099–3132. https://doi.org/10.1214/09-aos689

Allman E. S., Rhodes J. A., Stanghellini E., & Valtorta M. (2015). Parameter identifiability of discrete Bayesian networks with hidden variables. *Journal of Causal Inference*, *3*(2), 189–205. https://doi.org/10.1515/jci-2014-0021

Anderson M. J., & Fienberg S. E. (1999). *Who counts? The politics of census-taking in contemporary America*. Russel Sage Foundation.

Armstrong J. B., & Mayda J. E. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, *19*(2), 137–147.

Ball P., Betts W., Scheuren F. J., Dudukovich J., & Asher J. (2002). *Killings and Refugee Flow in Kosovo, March – June 1999*. American Association for the Advancement of Science.

Becker T., & Weispfenning V. (1993). *Gröbner basis: a computational approach to commutative algebra*. Springer.

Belin T. B., & Rubin D. B. (1995). A method for calibrating false matches in record linkage. *Journal of the American Statistical Association*, *90*(430), 694–707. https://doi.org/10.1080/01621459.1995.10476563

Bertsekas D. (1998). *Network optimization: continuous and discrete models*. Athena Scientific.

Biemer P. P. (1988). Modeling matching error and its effect on estimates of census coverage error. *Survey Methodology 14*(1), 117–134.

Binette O., & Steorts R. C. (2022). (Almost) all of entity resolution. *Science Advances*, *8*(12). https://doi.org/10.1126/sciadv.abi8021

Bishop Y. M. M., Fienberg S. E., & Holland P. W. (1975). *Discrete multivariate analysis*. The MIT Press.

Böhning D. (2005). *Computer-assisted analysis of mixtures and applications*. Chapman & Hall /CRC.

Böhning D., van der Heijden P. G. M. & Bunge J. (Eds.) (2018). *Capture-recapture methods for the social and medical sciences*. Chapman & Hall /CRC.

Brown J. J. (2000). Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration. PhD thesis. University of Southampton.

Brown J. J., Abbott O., & Diamond I. D. (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 883–902. https://doi.org/10.1111/j.1467-985X.2006.00431.x

Brown J. J., Abbott O., & Smith P. A. (2011). Design of the 2001 and 2011 Census coverage surveys for England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(4), 881–906. https://doi.org/10.1111/j.1467-985X.2011.00697.x

Brown J. J., Diamond I. D., Chambers R. L., Buckner L. J., & Teague A. D. (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 247–267. https://doi.org/10.1111/1467-985x.00133

Brown J. J., Sexton C., Abbott O., & Smith P. A. (2019). The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS*, *35*(3), 481–499.

Buckland S. T., & Garthwaite P.H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, *47*(1), 255–268. https://doi.org/10.2307/2532510

Burke D., & Račinskij V. (2020). The 2021 Census coverage survey: sample allocation strategy. *Census external assurance panel report*, *EAP127*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP127-CCS-2021-allocation-strategy.docx

Cabinet Office (2008). Helping to shape tomorrow – the 2011 Census of population and housing in England and Wales. *CM 7513*.

Cabinet Office (2018). Help shape our future – the 2021 Census of population and housing in England and Wales. *CM 9745*.

Castaldo A. (2018). 2021 Census coverage survey design strategy. *Census external assurance panel report*, *EAP103*. Available from https://uksa.statisticsauthority.gov.uk /wp-content/uploads/2020/07/EAP103-Census-Coverage-Survey-Design-Strategy.docx

Chipperfield J. O., & Chambers R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics*, *31*(3), 397–414.

Christen P. (2012). *Data Matching*. Springer.

Cochran W. G. (1977). *Sampling techniques*. Third edition. John Wiley & Sons.

Collart S., Kalkbrener M., & Mall D.(1997). Converting Bases with the Gröbner Walk. *Journal of Symbolic Computation*, *24*(3-4). https://doi.org/10.1006/jsco.1996.0145

Cox D., Little J., & O'Shea D. (2004). *Using algebraic geometry*. Second edition. Springer.

Cox D., Little J., & O'Shea D. (2015). *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Fourth edition. Springer.

Coumans A. M., Cruyff M., van der Heijden P. G. M., Wolf J., & Schmeets H. (2017). Estimating homelessness in the Netherlands using a capture-recapture approach. *Social Indicators Research 130*, 189–212. https://doi.org/10.1007/s11205-015-1171-7

de Wolf P.-P., van der Laan J., & Zult D. (2019). Connecting correction methods for linkage error in capture-recapture. *Journal of Official Statistics*, *35*(3), 577–597. https://doi.org/10.2478/jos-2019-0024

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1) 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Di Consiglio L., & Tuoto T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, *31*(3), 415–429. https://doi.org/10.1515/jos-2015-0025

Ding Y., & Fienberg S. E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology 20*(2), 149–158.

Dini E. (2018). Hard to count index for the 2021 Census . *Census external assurance panel report, EAP102*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP102-Hard-to-Count-index-for-the-2021-Census.docx

Decker, W., Greuel, G.-M., Pfister, G., & Schönemann, H. (2022). SINGULAR 4-3-0 — A computer algebra system for polynomial computations. https://www.singular.uni-kl.de.

Drton M., Sturmfels B., & Sullivant S. (2009). *Lectures on Algebraic Statistics*. Birkhäuser.

Dunn H. L. (1946). Record linkage. *American Journal of Public Health*, *36*(12), 1412–1416. https://doi.org/10.2105/AJPH.36.12.1412

Efron B., & Tibshirani R. J. (1993). *An introduction to the Bootstrap*. Chapman and Hall.

Everitt B. S., & Hand D. J. (1981). *Finite mixture distributions*. Chapman and Hall.

Fellegi I. P., & Sunter A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Fienberg S. E. (1972). The analysis of incomplete multi-way contingency tables. *Biometrics*, *28*(1). 177–202. https://doi.org/10.2307/2528967

Fienberg S. E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, *18*(1), 143–154.

Fienberg S. E., Hersh P., Rinaldo A., & Zhou Y. (2009). Maximum likelihood estimation in latent class models for contingency table data. In *Algebraic and geometric methods in statistics* (Eds. Gibilisco P., Riccomagno E., Rogantin M. P., & Wynn H. P.), 27–62. Cambridge University Press.

Fortini M., Liseo B., Nuccitelli A., & Scanu M. (2001). On Bayesian record linkage. *Researh in Official Statistics*, *4*(1), 185–198.

Frühwirth-Schnatter S., Celeux G., & Robert C. P. (Eds.) (2019). *Handbook of mixture analysis*. Chapman & Hall/CRC.

García-Puente L. D., Spielvogel S., & Sullivant S. (2010). Identifying causal effects with computer algebra. In *Proceedings of the 26th conference of uncertainty in artificial intelligence* (Eds. Grünwald P., & Spirtes P.). AUAI Press.

Geiger D., Heckerman D., King H., & Meek C. (2001). Stratified exponential families: graphical models and model selection. *The Annals of Statistics*, *29*(2), 505–529. https://doi.org/10.1214/aos/1009210550

Geiger H., & Werner A. (1924). Die Zahl der von Radium ausgesandten α-Teilchen. *Zeitschrift für Physik*, *21*, 187–203. https://doi.org/10.1007/BF01328262

Godfrey K. R., & DiStefano J. J. (1987). Identifiability of model parameters. In *Identifiability of parametric models* (Ed. Walker E.), 1–20, Pergamon Press.

Grimmet G. R., & Stirzaker D. R. (2001). *Probability and random processes*. Third edition. Oxford University Press.

Gyllenberg M., Koski T., Reilink E., & Verlaan M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, *31*(02), 542–548. https://doi.org/10.1017/s0021900200045034

Haberman S.J. (1974). Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *The Annals of Statistics*, *2*(5), 911–924. https://doi.org/10.1214/aos/1176342813

Hartshorne R. (1977). *Algebraic geometry*. Springer.

Herzog T. N., Scheuren F. J., & Winkler W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer.

Hogan H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, *46*(4), 261–269. https://doi.org/10.2307/2685308.

Hogan H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, *88*(423), 1047–1060. https://doi.org/10.1080/01621459.1993.10476374

Hosten S., Khetan A., & Sturmfels B. (2005). Solving the likelihood equations. *Foundations of Computational Mathematics*, *5*, 389–407. https://doi.org/10.1007/s10208-004-0156-8

Jaro M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Johndrow J. E., Lum K., & Dunson D. B. (2018). Theoretical limits of microclustering for record linkage. *Biometrika*, *105*(2), 431–446. https://doi.org/10.1093/biomet/asy003

Kirkpatrick S., Gelatt C. D., & Vecchi M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Krampe A., & Kuhnt S. (2010). Model selection for contingency tables with algebraic statistics. In *Algebraic and geometric methods in statistics* (Eds. Gibilisco P., Riccomagno E., Rogantin M. P., & Wynn H. P.), 83–98. Cambridge University Press.

Kruskal J. B. (1976). More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, *41*,(3), 281 – 293. https://doi.org/10.1007/bf02293554

Kruskal J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, *18*(2), 95–138. https://doi.org/10.1016/0024-3795(77)90069-6

Laplace P.-S. (1783). Sur les naissances, les mariages et les morts. In *Oeuvres complètes*, vol. 11, 1895, pp. 35–46. Gauthier-Villars.

Larsen M. D. (2005). Advances in record linkage theory: hierarchical Bayesian record linkage theory. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 3277–3283.

Larsen M. D., & Rubin. D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, *96*(453), 32–41. https://doi.org/10.1198/016214501750332956

Lee D., Zhang L. C., & Kim L. K. (2022). Maximum entropy classification for record linkage. *Survey Methodology*, *48*(1), 1–23.

Lehmann E. L., & Casella G. (1998). *Theory of point estimation*. Second edition. Springer.

Lincoln F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. *U.S. Department of Agriculture*, *118*, 1–4.

Liu J. S. (2004). *Monte Carlo strategies in scientific computing*. Springer.

Maple 2021. Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario.

Marks E. S., Mauldin W.P., & Nisselson H. (1953). The post-enumeration survey of the 1950 census: a case history in survey design. *Journal of the American Statistical Association*, *48*(262), 220–243. https://doi.org/10.1080/01621459.1953.10483469

McLachlan G. J., & Basford K. E. (1988). *Mixture models: inference and applications to clustering*. Marcel Dekker Inc.

McLachlan G. J. & Krishnan T. (2008). *The EM algorithm and extensions*. Second edition. John Wiley & Sons.

McLachlan G. J., & Peel D. (2000). *Finite mixture models*. John Wiley & Sons.

McCrea R. S., & Morgan B. J. T. (2015). *Analysis of capture-recapture data*. Chapman & Hall / CRC.

Meshkat N., Eisenberg J., & DiStefano J.J. (2009). An algorithm for finding globally identifiable parameter combinations of nonlinear ODE models using Gröbner bases. *Mathematical Biosciences*, *222*(2), 61–72. https://doi.org/10.1016/j.mbs.2009.08.010

Newcombe H. B., Kennedy J. M., Axford S. J., & James A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954–959. https://doi.org/10.1126/science.130.3381.954

Newcombe H. B., & Kennedy J. M. (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, *5*(11), 563–566. https://doi.org/10.1145/368996.369026

Norris J. L., & Pollock K. H. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, *3*, 235–244. https://doi.org/10.1007/bf00453012

Office for National Statistics (2010). Building the address register for the 2011 Census. *Technical report*. Available from https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/designforthe2011census/compilingtheaddressregister/address-register-overview_tcm77-189342.pdf

Office for National Statistics (2012). 2011 Census: Census response, return and coverage rates. Available from http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release–quality-assurance-and-methodology-papers/census-response-rates.xls

Office for National Statistics (2015). 2011 Census. *General report for England and Wales*.

Office for National Statistics (2018a). 2021 Census editing and imputation strategy. *Census external assurance panel report*, *EAP110*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP110-2021-Census-Editing-and-Imputation-Strategy.docx

Office for National Statistics (2018b). Methodology report on coverage matching for the 2021 Census. *Census external assurance panel report*, *EAP107*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP107-Methodology-report-on-coverage-matching-for-the-2021-Census.docx

Office for National Statistics (2018c). Census to census matching strategy 2021. *Census external assurance panel report*, *EAP121*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP121-Census-to-census-matching-strategy-2021.docx

Office for National Statistics (2021a). Design for Census 2021. *Technical report*. Available from https://www.ons.gov.uk/census/planningforcensus2021/censusdesign/designforcensus2021#design-and-build

Office for National Statistics (2021b). Search-as-you-type and address look-up functionality for Census 2021. *Technical report*. Available from https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/searchaasyoutypeandaddresslookupfunctionalityforcensus2021

Office for National Statistics (2021c). Census 2021 paper questionnaires. *Sample questionnaires*. Available from https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/census2021paperquestionnaires

Office for National Statistics (2021d). Census 2021 geographies. Available from https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies

Office for National Statistics (2021e). All data related to baby names in England and Wales: 2021. Available from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2021/relateddata

Office for National Statistics (2023). Evaluation of addressing quality: Census 2021. *Technical report*. Available from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/evaluationofaddressingqualitycensus2021

Pawitan Y. (2013). *In all likelihood*. Oxford University Press.

Petersen C. G. J. (1896). The yearly immigration of young plaice into the Limford from the German sea. *Report of The Danish Biological Station*, *6*, 5–84.

Pollock K. H. (1976). Building models of capture-recapture experiments. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *25*(4), 253–259. https://doi.org/10.2307/2988083

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Račinskij V., & Hammond C. (2018). Over-coverage estimation strategy for the 2021 Census of England and Wales. *Census external assurance panel report*, *EAP112*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP112-Over-coverage-estimation-strategy-for-the-2021-Census-of-England-and-Wales.docx

Račinskij V. (2018). Coverage estimation strategy for the 2021 Census of England and Wales. *Census external assurance panel report*, *EAP105*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP105-Coverage-Estimation-Strategy-for-the-2021-Census-of-England-and-Wales.docx

Račinskij V. (2020). Dealing with informative sampling in the coverage estimation of the 2021 Census of England and Wales. *Census external assurance panel report*, *EAP128*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP128-Infomative-sampling-in-coverage-estimation-of-Census-2021.pdf

Račinskij V. (2022). Adjusting for the dependence bias in the coverage estimation of the 2021 Census of England and Wales. *Census external assurance panel report*, *EAP160*. Available from https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2022/08/EAP160-Adjusting-for-the-dependence-bias-in-the-Census-2021-coverage-estimation.pdf

Rao P. B. L. S. (1992). *Identifiability in stochastic models: characterization of probability distributions*. John Wiley & Sons.

Sadinle M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, *112*(518), 600–612. https://doi.org/10.1080/01621459.2016.1148612

Seber G. A. F. (1982). *The estimation of animal abundance and related parameters*. Second edition. Charles Griffin & Company Ltd.

Sekar C. C., & Deming W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, *44*(245), 101–115. https://doi.org/10.1080/01621459.1949.10483294

Schnabel Z. E. (1938). The estimation of total fish population of a lake. *The American Mathematical Monthly*, *45*(6), 348–352. https://doi.org/10.2307/2304025

Silverman B. W. (2020). Model fitting in multiple systems analysis for the quantification of modern slavery: classical and Bayesian approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 691-736. https://doi.org/10.1111/rssa.12505

Sullivant S. (2018). *Algebraic statistics*. American Mathematical Society.

Tancredi A., & Liseo B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, *5*(2B), 1553–1585. https://doi.org/10.1214/10-aoas447

Tancredi A., Steorts R., & Liseo B. (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian analysis*, *15*(2), 633–682. https://doi.org/10.1214/19-ba1146

Teicher H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, *32*(1), 244–248. https://doi.org/10.1214/aoms/1177705155

Teicher H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *34*(4), 1265–1269. https://doi.org/10.1214/aoms/1177703862

Teicher H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, *38*(4), 1300–1302. https://doi.org/10.1214/aoms/1177698805

Thibaudeau Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology*, *19*(1), 31–38.

Titterington D. M., Smith A. F. M., & Makov U. E. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

Tuoto T. (2016). New proposal for linkage error estimation. *Statistical Journal of the IAOS*, *32*(3), 413–420. https://doi.org/10.3233/sji-160995

US Census Bureau (1990). Frequently occurring surnames from Census 1990. Names Files. Available from https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html

US Census Bureau (2008). 2010 Census Coverage Measurement Estimation Methodology. *DSSD 2010 Census coverage measurement memorandum series #2010-E-18*. US Bureau of the Census. Available from https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/ccm-workshop/201-e-18.pdf

Vermunt J. K. & Magidson J. (2004). Latent class analysis. In *The SAGE encyclopedia of social science research methods* (Eds. Lewis-Beck M., Bryman A., & Liao T. F.), vol. 1, 550–553. SAGE Publications, Inc.

Watanabe S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge University Press.

Winkler W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 274–279.

Winkler W. E. (2002). Methods for record linkage and Bayesian networks. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3743–3748.

Winkler W. E. (2006). Overview of record linkage and current research directions. *Technical report RR2006/02*, US Bureau of the Census.

Winkler W. E., & Thibaudeau Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census. *Technical report RR91-09*. US Bureau of the Census.

Wolter K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, *81*(394), 338–346. https://doi.org/10.2307/2289222

Wolfram Research, Inc. (2022). Mathematica, Version 13.2, Champaign, IL.

Yakowitz S.J., & Spragins J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*(1), 209–214. https://doi.org/10.1214/aoms/1177698520

Zehna P. W. (1966). Invariance of maximum likelihood estimators. *The Annals of Mathematical Statistics*, *37*(3),744–744. https://doi.org/10.1214/aoms/1177699475

Zwiernik P. (2016). *Semialgebraic statistics and latent tree models*. Chapman & Hall/ CRC.

Appendices

A Primer on Gröbner basis

This primer is based on Cox et al. (2015, chaps. 1 – 3) and Sullivant (2018, chap. 3). It is a very short summary of the referenced sources. Whenever applicable, we give the reference to definitions and theorems.

Definition 1 *A **field** is a set $\mathbb{K}$ with two binary operators "+" and "·" defined on $\mathbb{K}$ with the properties listed below.*
Associativity: $(a+b)+c = a+(b+c),\ (a\cdot b)\cdot c = a\cdot(b\cdot c)$ *for all* $a,b,c \in \mathbb{K}$.
Commutativity: $a+b = b+a,\ a\cdot b = b\cdot a$ *for all* $a,b \in \mathbb{K}$.
Distributivity: $a\cdot(b+c) = a\cdot b + a\cdot c$ *for all* $a,b,c \in \mathbb{K}$.
Identities: there are $0,1 \in \mathbb{K}$ *such that* $a+0 = a\cdot 1 = a$ *for all* $a \in \mathbb{K}$.
Additive inverses: given $a \in \mathbb{K}$ *there is* $b \in \mathbb{K}$ *such that* $a+b = 0$.
Multiplicative inverses: $a \in \mathbb{K},\ a \neq 0$ *there is* $c \in \mathbb{K}$ *such that* $a\cdot c = 1$.

Most frequently we are dealing with one of the following fields. The rational numbers $\mathbb{Q}$ which are useful to perform computations. The real numbers $\mathbb{R}$ which are key in dealing with probabilities since probabilities are real. The real numbers are also useful for plotting purposes. Finally, the complex numbers $\mathbb{C}$ which are crucial for proving theorems. We use $\mathbb{K}$ to denote an arbitrary field and $\mathbb{K}(x_1,\ldots,x_r)$ to denote an arbitrary field in indeterminates $x_1,\ldots,x_r$. Often, either when r is not specified or for brevity, we write simply $\mathbf{x}$ instead of listing indeterminates $x_1,\ldots,x_r$. Note that when discussing algebra and geometry we use the word 'indeterminates' instead of 'variables' to avoid confusion with random variables, but also because polynomials in a strict sense do not involve variables.

Definition 2 *A **commutative ring** is a set S with two binary operators "+" and "·" defined on S with the properties listed below.*
Associativity: $(a+b)+c = a+(b+c),\ (a\cdot b)\cdot c = a\cdot(b\cdot c)$ *for all* $a,b,c \in S$.
Commutativity: $a+b = b+a,\ a\cdot b = b\cdot a$ *for all* $a,b \in S$.
Distributivity: $a\cdot(b+c) = a\cdot b + a\cdot c$ *for all* $a,b,c \in S$.
Identities: there are $0,1 \in S$ *such that* $a+0 = a\cdot 1 = a$ *for all* $a \in S$.
Additive inverses: given $a \in S$ *there is* $b \in S$ *such that* $a+b = 0$.

Ring has no multiplicative inverses. Examples of a commutative ring are fields, integers $\mathbb{Z}$ and sets of polynomials in indeterminates $x_1,\ldots,x_r$ with coefficients in a field $\mathbb{K}$, written $\mathbb{K}[x_1,\ldots,x_r]$, that satisfy the above definition.

Definition 3 (Definition 1, Cox et al. (2015, chap. 1.2)**)** *A **monomial** in $x_1,\ldots x_r$ is the product in form*

$$x_1^{\alpha_1}x_2^{\alpha_2}\ldots x_r^{\alpha_r},$$

where $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_r)$ are non-negative integers.

Sometimes, we can simplify the notation by letting $\alpha = (\alpha_1, \ldots, \alpha_r)$ be a n-tuple of non-negative integers. Then we write

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_r^{\alpha_r}.$$

Monomials can be combined in a certain way to produce the entity which is in the core of our discussion.

Definition 4 (Definition 2, Cox et al. (2015, chap. 1.2)) *A **polynomial** f in indeterminates $x_1, \ldots, x_r$ with coefficients in a field $\mathbb{K}$ is a finite linear combination of monomials*

$$f = \sum_\alpha a_\alpha x^\alpha, \ a_\alpha \in \mathbb{K},$$

where the sum is over a finite number of n-tuples $\alpha = (\alpha_1, \ldots, \alpha_r)$. The set of all polynomials in $x_1, \ldots, x_r$ with coefficients in $\mathbb{K}$ is denoted $\mathbb{K}[x_1, \ldots, x_r] = \mathbb{K}[\mathbf{x}]$.

The set $\mathbb{K}[x_1, \ldots, x_r] = \mathbb{K}[\mathbf{x}]$ is a ring since the sum of two polynomials is a polynomial and the product of two polynomials is again a polynomial; associativity, commutativity and distributivity properties also hold for polynomials. In addition, there are identities and additive inverses. However, there is no multiplicative inverse, say, if $f = x$ then $1/x$ is not a polynomial.

Definition 5 (Definition 3, Cox et al. (2015, chap.i̇1.2)) *Let $f = \sum_\alpha a_\alpha x^\alpha$ be a polynomial in $\mathbb{K}[x_1, \ldots, x_r]$.*

1. *We call a_α the **coefficient** of the monomial x^α.*
2. *If $a_\alpha \neq 0$, then we call $a_\alpha x^\alpha$ a **term** of f.*
3. *The **total degree** of $f \neq 0$, denoted $deg(f)$, is the maximum $|\alpha|$ such that the coefficient a_α is non-zero. The total degree of the zero polynomial is undefined.*

For instance, the polynomial $f = 5x^3y^4z + 3y^2z^3 + (3/5)xz$ has three terms and the total degree eight.

What makes possible to link algebra and geometry is the ability to consider a polynomial $f = \sum_\alpha a_\alpha x^\alpha \in \mathbb{K}[x_1, \ldots, x_r]$ as a function

$$f : \mathbb{K}^r \to \mathbb{K}$$

by replacing every x_i in f by a_i from an r-tuple $(a_1, \ldots, a_r) \in \mathbb{K}^r$.

Definition 6 (Definition 1, Cox et al. (2015, chap. 1.2)) *Let $\mathbb{K}$ be a field, and let $f_1 \ldots, f_s$ be polynomials in $\mathbb{K}[x_1, \ldots, x_r]$. The we call the set*

$$\mathbf{V} = \{(a_1, \ldots, a_r) \in \mathbb{K}^r \mid f_i(a_1, \ldots, a_r) = 0 \text{ for all } \ 1 \leq i \leq s\}$$

*the **variety** defined by $f_1, \ldots, f_s$.*

In other words, a variety $\mathbf{V}(f_1, \ldots, f_s) \subseteq \mathbb{K}^r$ is the set of all solutions of the system of equations $f_1(x_1, \ldots, x_r) = \cdots = f_s(x_1, \ldots, x_r) = 0$. For example, graphs of polynomial functions are variates:

the graph of $y = f(x)$ is $\mathbf{V}(y - f(x))$. Another example of varieties is a set of solutions of a system of linear equations. Such variety is a linear variety. Here, we are mainly interested in non-linear systems, but as we will see Gröbner basis generalizes certain ideas from linear algebra.

Definition 7 (Definition 1, Cox et al. (2015, chap. 1.3)) *Let $\mathbb{K}$ be a field. A **rational function** in $t_1, \ldots, t_m$ with coefficients in $\mathbb{K}$ is a quotient f/g of two polynomials $f, g \in \mathbb{K}[t_1, \ldots, t_m]$, where g is not the zero polynomial (a polynomial with all coefficients being 0). The set of all such rational functions is denoted $\mathbb{K}(t_1, \ldots, t_m)$.*

Note, that $\mathbb{K}(t_1, \ldots, t_m)$ is a field.

A variety can be represented either parametrically or implicitly. Let $V = \mathbf{V}(f_1, \ldots, f_s) \subseteq \mathbb{K}^r$ be a variety. A **rational parametric representation** of V consists of rational functions $g_1, \ldots, g_r \in \mathbb{K}(t_1, \ldots, t_m)$ such that the points given by

$$\begin{aligned} x_1 &= g_1(t_1, \ldots, t_m), \\ &\vdots \\ x_r &= g_r(t_1, \ldots, t_m) \end{aligned}$$

lie in V. In addition, we requite that V is the smallest (it is beyond the scope of this primer to explain in what sense a variety is smallest) variety that contains the above points. It is often the case that $g_1, \ldots, g_r$ in the above expression are polynomial rather than rational functions. We call such representation of V a **polynomial parametric representation**.

The original defining equations $f_1 = \cdots = f_s = 0$ of V are called an **implicit representation** of V.

In general, not every variety has a parametric representation. However, given a parametric representation of a variety, it is possible to find a corresponding implicit representation or the defining equations. This fact will be crucial for our discussion of identifiability.

Definition 8 (Definition 1, Cox et al. (2015, chap. 1.4)) *A subset $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ is an **ideal** if it satisfies:*

1. $0 \in I$.
2. *If $f, g \in I$, then $f + g \in I$.*
3. *If $f \in I$ and $h \in \mathbb{K}[x_1, \ldots, x_r]$ then $hf \in I$.*

An ideal is the basic algebraic object featuring in our discussion.

For a collection of polynomials $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_r]$, define the set of polynomials $\langle f_1, \ldots, f_s \rangle$ as

$$\langle f_1, \ldots, f_s \rangle = \left\{ \sum_{i=1}^{s} h_i f_i \mid h_1, \ldots, h_s \in \mathbb{K}[x_1, \ldots, x_r] \right\}.$$

It is easy to check that $\langle f_1, \ldots, f_s \rangle$ is an ideal of $\mathbb{K}[x_1, \ldots, x_r]$ and we call $\langle f_1, \ldots, f_s \rangle$ the **ideal generated by** $f_1, \ldots, f_s$. (These are Definition 2 and Lemma 3 in Cox et al. (2015, chap. 1.4)).

The ideal $\langle f_1, \ldots, f_s \rangle$ can be seen as an infinite set of all possible polynomial equations that can be obtained from a given collection $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_r]$ of polynomials satisfying

$$\begin{aligned} f_1 &= 0, \\ &\vdots \\ f_s &= 0 \end{aligned}$$

in the following way: multiply every f_i by $h_i \in \mathbb{K}[x_1, \ldots, x_r]$ and add products together. The resulting polynomial $h_1 f_1 + \cdots + h_s f_s = 0$ is the consequence of the original system and $h_1 f_1 + \cdots + h_s f_s \in \langle f_1, \ldots, f_s \rangle$.

An ideal I is **finitely generated** if there exist $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_r]$ such that $I = \langle f_1, \ldots, f_s \rangle$. In such case we say that $f_1, \ldots, f_s$ is a **basis** of I. In fact, the following important result, known as the Hilbert Basis Theorem, states that every ideal of $\mathbb{K}[x_1, \ldots, x_r]$ is finitely generated.

Theorem 1 (Hilbert Basis Theorem [Theorem 4, Cox et al. (2015, chap. 2.5)**)** *] Every ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ has a finite generating set. That is, $I = \langle f_1, \ldots, f_s \rangle$ for some $f_1, \ldots, f_s \in I$.*

It is important to note that a given ideal may have many different bases. However, some of the bases are more useful than other, as we will see.

There are analogical concepts to ideal and bases in linear algebra. An ideal is similar to a subspace that is closed both under addition and multiplication. However, in the case of subspace, multiplication is by scalars, while in the case of ideals multiplication is by polynomials. Also, the ideal generated by polynomials $f_1, \ldots, f_s$ is similar to the span of finite number of vectors $v_1, \ldots, v_s$. Polynomial coefficients are used for the ideals while field coefficients are used in the span.

An important fact is that a variety depends only on the ideal generated by its defining equations. In other words, if $f_1, \ldots, f_s$ and $g_1, \ldots, g_t$ are bases of the same ideal in $\mathbb{K}[x_1, \ldots, x_r]$, so that $\langle f_1, \ldots, f_s \rangle = \langle g_1, \ldots, g_t \rangle$, then $\mathbf{V}(f_1, \ldots, f_s) = \mathbf{V}(g_1, \ldots, g_t)$. In practice, changing a basis of an ideal by more 'useful' one makes easier to determine the variety. In fact, varieties are determined by ideals, not equations.

So far, we have been discussing situation where a common zero set $\mathbf{V}(f_1, \ldots, f_s)$ (a geometric object) was produces for a given collection of polynomials $f_1, \ldots, f_s$ (an algebraic object). However, the opposite construction can be also regarded: given a zero set, construct the set of polynomials that vanish on it.

Definition 9 (Definition 3.2.1, Sullivant (2018, chap. 3)**)** *Let $V \subseteq \mathbb{K}^r$ be a variety. Then we set*

$$\mathbf{I}(V) = \{ f \in \mathbb{K}[x_1, \ldots, x_r] \mid f(a_1, \ldots, a_r) = 0 \text{ for all } (a_1, \ldots, a_r) \in V \}.$$

An important fact is that $\mathbf{I}(V) \subseteq \mathbb{K}[x_1, \ldots, x_r]$ is an ideal. We call $\mathbf{I}(V)$ the **ideal of** V, or the **vanishing ideal of** V, or the **defining ideal of** V. If $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_r]$, then $\langle f_1, \ldots, f_s \rangle \subseteq \mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$, but equality need not occur.

Before introducing Gröbner basis, we will briefly discuss the division algorithm for polynomials in $\mathbb{K}[x]$ and in $\mathbb{K}[x_1, \ldots, x_r]$ as well as concepts related to division of polynomials. This discussion will motivate the need of and facilitate understanding of Gröbner basis.

For division of polynomials in one indeterminate, consider a polynomial $f \in \mathbb{K}[x]$. To perform division of polynomials, we need to define the leading term of a polynomial in one indeterminate.

Definition 10 (Definition 1, Cox et al. (2015, chap. 1.5)) *Given a non-zero polynomial $f \in \mathbb{K}[x]$, let*

$$f = c_0x^m + c_1x^{m-1} + \cdots + c_m,$$

*where $c_i \in \mathbb{K}$ and $c_0 \neq 0$, so that $m = deg(f)$. Then we say than c_0x^m is the **leading term** of f. We write $LT(f) = c_0x^m$.*

For instance, if $f = 5x^4 + 2x^2 - 8x + 3$, then $\mathrm{LT}(f) = 5x^4$. An important observation here is that if f and g are non-zero polynomials, then

$$deg(f) \leq deg(g) \iff \mathrm{LT}(f) \text{ divides } \mathrm{LT}(g).$$

If $g \in \mathbb{K}$ is a non-zero polynomial, then every $f \in \mathbb{K}$ can be written as

$$f = qg + r,$$

where $q, r \in \mathbb{K}$, and either $r = 0$ or $deg(f) \leq deg(g)$. Furthermore, q and r are unique, and the **division algorithm** (Proposition 2, Cox et al. (2015, chap. 1.5)) below produces q and r.

input: g, f
output: q (quotient), r (remainder)
$q := 0$; $r := f$
while $r \neq 0$ **and** $\mathrm{LT}(g)$ divides $\mathrm{LT}(r)$ **do**
    $q := q + \mathrm{LT}(r)/\mathrm{LT}(g)$
    $r := r - (\mathrm{LT}(r)/\mathrm{LT}(g))g$
return q, r

For example, let $f = 5x^4 + 2x^2 - 8x + 3$ and $g = 3x^2 + 6x + 1$, then

$$f = qg + r = \left(\frac{5}{3}x^2 + \frac{10}{3}x + \frac{61}{9}\right)\left(3x^2 + 6x + 1\right) - \left(\frac{136}{3}x + \frac{34}{9}\right)$$

.

Before we move to the division algorithm in $\mathbb{K}[x_1, \ldots, x_r]$, it is worth making a few remarks about ideals in $\mathbb{K}$ as this helps to understand better the case of several indeterminates. We say that a polynomial h is a **greatest common divisor** of polynomials $f_1, \ldots, f_r \in \mathbb{K}$ if h divides $f_1, \ldots, f_r$ and if p is another polynomial which divides $f_1, \ldots, f_r$ then p divides h. We write, $h = \gcd(f_1, \ldots, f_r)$. The greatest common divisor can be algorithmically computed by repeatedly applying the Euclidean Algorithm, which computes the greatest common divisor of two polynomials. What is important, is that for polynomials in one indeterminate, the $\gcd(f_1, \ldots, f_r)$ is the generator of $\langle f_1, \ldots, f_r \rangle$.

For instance, let $f = x^8 - 1$ and $g = x^6 - 1$ and consider the ideal $\langle f, g \rangle \subseteq \mathbb{K}[x]$. Then $h = \gcd(f, g) = x^2 - 1$ and h is the generator of $\langle f, g \rangle \subseteq \mathbb{K}[x]$. Put it another way, $\langle f, g \rangle = \langle x^2 - 1 \rangle$. Hence, both f and g (and infinitely many other polynomials in the ideal), are the consequences of h.

Using the above results, it is possible to address algorithmically the **ideal membership problem** for the the polynomials in one indeterminate. This problem can be stated as follows: given an ideal $\langle f_1, \ldots, f_r \rangle \subseteq \mathbb{K}[x]$ and a polynomial $f \in \mathbb{K}[x]$, determine whether f lies in $\langle f_1, \ldots, f_r \rangle$. The solution has two steps. First, find the generator $h = \gcd(f_1, \ldots, f_r)$. Then, use the division algorithm to write $f = qh + r$, where $\deg(r) < \deg(h)$. Then f is in the ideal if and only if $r = 0$.

For instance, using the ideal from the previous example $I = \langle x^8 - 1, x^6 - 1 \rangle \subseteq \mathbb{K}[x]$. The generator of this ideal is $\langle x^2 - 1 \rangle$. Then $x^5 + 4x^3 + 2x^2 - 5x - 2 \in \langle x^2 - 1 \rangle$, since $x^5 + 4x^3 + 2x^2 - 5x - 2 = (x^3 + 5x + 2)(x^2 - 1)$. While $x^3 + 2x \notin \langle x^2 - 1 \rangle$, since $x^3 + 2x = x(x^2 - 1) + 3x$.

Looking at the division algorithm in one indeterminate, we see that the leading terms are playing the key role here. We first divide the leading term of our dividend by the leading term of the divisor and then we subtract the product of the resulting quotient and the divisor from the dividend to cancel out the leading term of the original dividend. The process is repeated until the degree of remainder becomes smaller than the degree of the divisor. A notion of **ordering of terms** is crucial for determining a leading term on each iteration of the division algorithm. In the case of a single indeterminate, term ordering is simply based on the degree of monomials:

$$\cdots \succ x^{m+1} \succ x^m \succ \cdots \succ x^2 \succ x \succ 1.$$

Symbols $\succ$ (succeeds) and $\prec$ (precedes) are denoting term orders, quite often standard symbols $>$ and $<$ are used instead, respectively.

In the case of several indeterminates the task of determining order of terms is more complicated. If all monomials in a polynomial are of degree 1 and there are r indeterminates, we can order indeterminates in $r!$ ways. Say,

$$x_1 \succ x_2 \succ \cdots \succ x_r$$

or

$$x_r \succ x_{r-1} \succ \cdots \succ x_1.$$

In fact, in the row-reduction algorithm on matrices we employ one of those $r!$ orderings, quite often this ordering is arbitrary. When monomials of a polynomials have higher degrees and there are r indeterminates, we need more work to arrange the terms of a polynomial in an unambiguously descending manner. In order to compare every pair of monomials and establish relative positions our ordering must be **linear** or **total** ordering: for every pair of monomials x^α and x^β exactly one the following three statements

$$x^\alpha \succ x^\beta, x^\alpha = x^\beta, x^\alpha \prec x^\beta$$

should be true. A total order also must be **transitive**, so that $x^\alpha \succ x^\beta$ and $x^\beta \succ x^\gamma$ always imply $x^\alpha \succ x^\gamma$. The following definition reflects the above considerations.

Definition 11 (Definition 1, Cox et al. (2015, chap. 2.2)) *A **monomial ordering** $\succ$ on $\mathbb{K}[x_1, \ldots, x_r]$ is a relation $\succ$ on $\mathbb{Z}^r_{\geq 0}$ (equivalently, a relation on the set of monomials $x^\alpha, \alpha \in \mathbb{Z}^r_{\geq 0}$), satisfying:*

1. *$\succ$ is a total ordering on $\mathbb{K}[x_1, \ldots, x_r]$.*

2. *If* $\alpha \succ \beta$ *and* $\gamma \in \mathbb{Z}^r_{\geq 0}$*, then* $\alpha + \gamma \succ \beta + \gamma$*.*

3. *If* $A \subseteq \mathbb{Z}^r_{\geq 0}$ *is non-empty, then there is* $\alpha \in A$ *such that* $\beta \succ \alpha$ *for every* $\beta \neq \alpha$ *in* A*.*

Clearly, there are many possible term orders. We are particularly interested in two term orders: the lexicographic term order and the graded reverse lexicographic term order, also referred to as reverse degree lexicographic order.

Definition 12 (Lexicographic monomial order, Definition 3, Cox et al. (2015, chap. 2.2)**)** *Let* $\alpha = (\alpha_1, \ldots, \alpha_r)$ *and* $\beta = (\beta_1, \ldots, \beta_r)$ *be in* $\mathbb{Z}^r_{\geq 0}$*. We say that* $\alpha \succ_{lex} \beta$ *if the leftmost non-zero entry of the vector difference* $\alpha - \beta \in \mathbb{Z}^r_{\geq 0}$ *is positive. We write* $x^\alpha \succ_{lex} x^\beta$ *if* $\alpha \succ_{lex} \beta$*.*

Note that in the lexicographic monomial ordering we assume that indeterminates are in some fixed alphabetical order, say, $x \succ y \succ z$, or $y \succ x \succ z$, or order based on indices, $x_1 \succ_{lex} x_2 \succ_{lex} \cdots \succ_{lex} x_r$.

As an example of lexicographic order with $x \succ y \succ z$ consider $xy^3 \succ_{lex} y^5z^4$ (or equivalently $\alpha = (1, 3, 0) \succ_{lex} \beta = (0, 5, 4)$) since $\alpha - \beta = (1, -2, -4)$.

Definition 13 (Graded reverse lexicographic order, Definition 6, Cox et al. (2015, chap. 2.2)**)** *Let* $\alpha, \beta \in \mathbb{Z}^r_{\geq 0}$*. We say* $\alpha \succ_{grevlex} \beta$ *if*

$$|\alpha| = \sum_{i=1}^{r} \alpha_i \succ |\beta| = \sum_{i=1}^{r} \beta_i, \text{ or } |\alpha| = |\beta| \text{ and the rightmost non-zero entry of } \alpha - \beta \text{ is negative.}$$

Grevlex orders by total degree and if degrees of two monomials are equal, it breaks ties by looking at the rightmost (or smallest) indeterminate in a monomial and favours the smallest power. Hence, the notion of reverse lexicographic does not mean reversing the order of variables $x \succ y \succ \ldots w$ used the lexicographic order. The lexicographic order looks for the leftmost (largest) indeterminate and favours the largest power. For instance, let $x \succ y \succ z$. Then $x^6yz^2 \succ_{lex} x^5yz^3$ since we are looking at the largest indeterminate x having the largest power. On this occasion, we also have $x^6yz^2 \succ_{grevlex} x^5yz^3$. This is because the total degree of two monomials is the same, 9, but the time the smallest indeterminate z having the smallest power. So essentially, it is a double reversing.

In many cases graded reverse lexicographic order is computationally very efficient while lexicographic order allows obtaining important algebraic and geometric results. Often computations are carried out with graded reverse lexicographic order and then transformed to lexicographic order using various methods, like Gröbner walk (Collart et al., 1997).

Definition 14 (Definition 7, Cox et al. (2015, chap. 2.2)**)** *Let* $f = \sum_\alpha a_\alpha x^\alpha$ *be a non-zero polynomial in* $\mathbb{K}[x_1, \ldots, x_r]$ *and let* $\succ$ *be a monomial order.*

1. *The* ***multidegree*** *of* f *is* $\text{multideg}(f) = \max(\alpha \in \mathbb{Z}^r_{\geq 0} \mid a_\alpha \neq 0)$*, where the maximum is taken with respect to* $\succ$*.*

2. *The* ***leading coefficient*** *of* f *is* $\text{LC}(f) = a_{\text{multideg}(f)} \in \mathbb{K}$*.*

3. *The* ***leading monomial*** *of* f *is* $\text{LM}(f) = x^{\text{multideg}(f)}$*.*

4. *The* ***leading term*** *of* f *is* $\mathrm{LT}(f) = \mathrm{LC}(f) \cdot \mathrm{LM}(f)$.

For example, let $f = 7x^3y^2z - 2x^4y + 8y^6z^3$ and $\succ$ is lex order with $x \succ y \succ z$. Then

$$\begin{aligned} \mathrm{multideg}(f) &= (4, 1, 0), \\ \mathrm{LC}(f) &= -2, \\ \mathrm{LM}(f) &= x^4y, \\ \mathrm{LT}(f) &= -2x^4y. \end{aligned}$$

Note that $\mathrm{multideg}(fg) = \mathrm{multideg}(f) + \mathrm{multideg}(g)$.

We are now equipped for discussion of the **division algorithm** in $\mathbb{K}[x_1, \ldots, x_r]$. Let $\succ$ be a monomial order on $\mathbb{Z}^r_{\geq 0}$ and let $F = (f_1, \ldots, f_s)$ be an ordered s-tuple of polynomials in $\mathbb{K}[x_1, \ldots, x_r]$. Then every $f \in \mathbb{K}[x_1, \ldots, x_r]$ can be written as

$$f = q_1f_1 + \cdots + q_sf_s + r,$$

where $q_i, r \in \mathbb{K}[x_1, \ldots, x_r]$, and either $r = 0$ or r is a linear combination with the coefficients in $\mathbb{K}$, of monomials, none of which is divisible by any of $\mathrm{LT}(f_1), \ldots, \mathrm{LT}(f_s)$. We call r a **remainder** of f on division by F. If $q_if_i \neq 0$ then

$$\mathrm{multideg}(f) \geq \mathrm{multideg}(q_if_i).$$

The division algorithm, presented in Theorem 3, Cox et al. (2015, chap. 2.3), produces q_i and r.

input: $f_1, \ldots, f_s, f$
output: $q_1, \ldots, q_s, r$
$q_1 := 0; \ldots; q_s := 0; r := 0$
$p := f$
while $p \neq 0$ **do**
    $i := 1$
    $d := 0$
    while $i \leq s$ **and** $d = 0$ **do**
        if $\mathrm{LT}(f_i)$ divides $\mathrm{LT}(p)$ **then**
            $q_i := q_i + \mathrm{LT}(p)/\mathrm{LT}(f_i)$
            $p := p - \mathrm{LT}(p)/\mathrm{LT}(f_i)f_i$
            $d := 1$
        else
            $i := i + 1$
    if $d = 0$ **then**
        $r := r + \mathrm{LT}(p)$
        $p := p - \mathrm{LT}(p)$
return $q_1, \ldots, q_s, r$

To illustrate the division algorithm and some unwanted behaviour associated with it consider an example from Cox et al. (2015). Suppose we divide $f = x^2y + xy^2 + y^2$ by $f_1 = y^2 - 1$, $f_2 = xy - 1$

using lex order with $x \succ y$. We get $x^2y + xy^2 + y^2 = (x+1)(y^2-1) + x(xy-1) + 2x + 1$, so that the reminder is $2x + 1$. Now divide the same f by the same polynomials but permute them, that is f_2, f_1, using the same lex order with $x \succ y$. We get $x^2y + xy^2 + y^2 = (x+y)(xy-1) + 1 \cdot (y^2-1) + x + y + 1$. The remainder is $x + y + 1$ this time around.

The reminder in the division algorithm is not in general uniquely determined. This is an obstacle in solving the ideal membership problem for $f \in \mathbb{K}[x_1, \ldots, x_r]$. It can be shown, that $r = 0$ is only sufficient condition for ideal membership, so that even $f \in \langle f_1, \ldots, f_r \rangle$ it is possible to obtain a non-zero reminder on division by $(f_1, \ldots, f_r)$. The idea of Gröbner basis allows to overcome the issue of non-uniqueness of a division in $\mathbb{K}[x_1, \ldots, x_r]$.

Definition 15 (Ideal of leading terms, Definition 1, Cox et al. (2015, chap. 2.5)**)** *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ be an ideal other than $\{0\}$ and fix a monomial ordering on $\mathbb{K}[x_1, \ldots, x_r]$. Then*

1. *The set of leading terms of non-zero elements of I is defined as*

$$\mathrm{LT}(I) = \{cx^\alpha \mid \text{ there exists } f \in I \backslash \{0\} \text{ with } \mathrm{LT}(f) = cx^\alpha\}$$

2. *We denote by $\langle \mathrm{LT}(I) \rangle$ the ideal generated by the elements of $\mathrm{LT}(I)$.*

In other words, $\langle \mathrm{LT}(I) \rangle$ is an infinite set of all possible consequences produced by monomials in an infinite set $\mathrm{LT}(I)$. Note that $I = \langle f_1, \ldots, f_s \rangle$ then $\langle \mathrm{LT}(f_1), \ldots, \mathrm{LT}(f_s) \rangle$ and $\langle \mathrm{LT}(I) \rangle$ may be different ideals. It is true that $\langle \mathrm{LT}(f_1), \ldots, \mathrm{LT}(f_s) \rangle \subseteq \langle \mathrm{LT}(I) \rangle$, but $\langle \mathrm{LT}(I) \rangle$ can be strictly larger.

We are now in a position to introduce Gröbner basis, a subset of a given ideal, that has remarkable properties and even more remarkable applications.

Definition 16 (Definition 5, Cox et al. (2015, chap. 2.5)**)** *Fix a monomial order on the polynomial ring $\mathbb{K}[x_1, \ldots, x_r]$. A finite subset $G = \{g_1, \ldots, g_t\}$ of an ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ different from $\{0\}$ is said to be a* ***Gröbner basis*** *or* ***standard basis*** *if*

$$\langle \mathrm{LT}(g_1), \ldots, \mathrm{LT}(g_t) \rangle = \langle \mathrm{LT}(I) \rangle.$$

The Gröbner basis of the zero ideal is the empty set $\emptyset$.

Informally, we can say that a set $\{g_1, \ldots, g_t\} \subseteq I$ is a Gröbner basis of I if and only if the leading term of any element of I is divisible by one of the $\mathrm{LT}(g_i)$. There exist several equivalent definitions of Gröbner basis (Becker & Weispfenning, 1993), but most of them require additional knowledge to what was introduced in this summary. However, we will see one more definition when discussing the properties of Gröbner basis.

Recall, that the Hilbert Basis Theorem guarantees that every ideal is finitely generated. Furthermore, it can be shown that for a fixed monomial order every ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ has a Gröbner basis and that any Gröbner basis for an ideal I is a basis of I. Note, that a Gröbner basis for an ideal I is not unique. However, we will later introduce a Gröbner basis which is unique for an ideal I.

An important consequence of the Hilbert Basis Theorem is a geometric one and allows us to consider a variety as being defined by an ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$.

Definition 17 Definition 8, Cox et al. (2015, chap. 2.5) *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ be an ideal. Let $\mathbf{V}(I)$ be the following set*

$$\mathbf{V}(I) = \{(a_1, \ldots, a_r) \in \mathbb{K}^r \mid f(a_1, \ldots, a_r) = 0 \text{ for all } f \in I\}.$$

In fact, $\mathbf{V}(I)$ is a variety and if $I = \langle f_1, \ldots, f_s \rangle$, then $\mathbf{V}(I) = \mathbf{V}(f_1, \ldots, f_s)$. In other words, varieties are determined by ideals, or $\mathbf{V}(f_1, \ldots, f_s) = \mathbf{V}(g_1, \ldots, g_t)$ whenever $\langle f_1, \ldots, f_s \rangle = \langle g_1, \ldots, g_t \rangle$. For real life problems is means that given a variety $\mathbf{V}(I)$, we can find the 'right' generating set for an I that allows a better understanding of the variety $\mathbf{V}(I)$.

We now discuss the properties of Gröbner bases. Let $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ be an ideal, $G = \{g_1, \ldots, g_t\}$ be a Gröbner basis for I. Also, let $f \in \mathbb{K}[x_1, \ldots, x_r]$. Then there is a unique $r \in \mathbb{K}[x_1, \ldots, x_r]$ with the following properties:

1. No term on r is divisible by any of $\mathrm{LT}(g_1), \ldots, \mathrm{LT}(g_t)$.
2. There is $g \in I$ such that $f = g + r$.

In particular, r is the remainder on division of f by G no matter how the elements of G are listed when using the division algorithm in $\mathbb{K}[x_1, \ldots, x_r]$. Also, $f \in I$ if and only if the remainder on division of f by G is zero. The latter can be regarded as another definition of a Gröbner basis.

Hence, we can see that a Gröbner basis overcomes an issue of a possibly non-unique remainder in the division algorithm in $\mathbb{K}[x_1, \ldots, x_r]$. Therefore, if we can find a Gröbner basis for a given ideal, we can solve the ideal membership problem. Luckily, given an ideal I a Gröbner basis for I can be constructed in a finite numbers of steps. The Buchberger's algorithm was the first algorithm designed to construct a Gröbner basis and still often used in many symbolic algebra packages. More recent and often faster than the Buchberger's algorithms are Faugère's F_4, F_5 and various related variations. A discussion of algorithms that construct a Gröbner basis is beyond the scope of this primer. We note here that the row reduction algorithm in matrix algebra is essentially a special case of the Buchberger's algorithm.

What is important in practical application is the complexity issues, that is a possibility that a Gröbner basis might not be computed in real time despite the fact that every ideal has a Gröbner basis and that Gröbner basis can be constructed in a finite number of steps. The success of constructing the Gröbner basis may depend on the choice of monomial ordering with grevlex ordering of indeterminates often being quite efficient. Yet, as we will see, solving certain problems may require lexicographic monomial ordering. While conversion from one ordering to another is always possible in theory, this is also not always doable in real time.

An ideal I may have infinitely many Gröbner bases. It is possible, however, to construct the Gröbner basis which is in a sense better than others.

Definition 18 (Definition 4, Cox et al. (2015, chap. 2.7)) *A **reduced Gröbner basis** for a polynomial ideal I is a Gröbner basis G for I such that*

1. $\mathrm{LC}(p) = 1$*, for all polynomials* $p \in G$.

2. For all $p \in G$, no monomial of p lies in $\langle LT(G \setminus \{p\})\rangle$.

For an ideal $I \neq \{0\}$ and a given monomial ordering, I has a reduced Gröbner basis, and the reduced Gröbner basis is unique. The uniqueness property of a Gröbner basis allows checking whether two sets of polynomials $\{f_1, \ldots, f_s\}$ and $\{g_1, \ldots, g_t\}$ generate the same ideal. To do that, fix a monomial order and compute a reduced Gröbner basis for $\langle f_1, \ldots, f_s\rangle$ and $\langle g_1, \ldots, g_t\rangle$. Two ideals are equal if and only if the Gröbner bases are the same.

We have already seen the ideal membership problem for polynomials in one indeterminate. Combining Gröbner bases with the division algorithm in several indeterminates allows to solve the ideal membership problem in general. Suppose $I = \langle f_1, \ldots, f_s\rangle$ and we interested in whether f lies in I. To check it, we compute a Gröbner basis $G = \{g_1, \ldots, g_t\}$ for I and use the division algorithm to find the remainder on division of f by G. Then $f \in I$ if and only if the remainder is zero.

A very important result for our discussion of identifiability is the Elimination theorem. The Elimination theorem provides the conceptual tools for solving the implicitization problem algorithmically.

Definition 19 Definition 1, Cox et al. (2015, chap. 3.1) *Given $I = \langle f_1, \ldots, f_r\rangle \subseteq \mathbb{K}[x_1, \ldots, x_r]$, the l-th **elimination ideal** I_l is the ideal of $\mathbb{K}[x_{l+1}, \ldots, x_r]$ defined by*

$$I_l = I \cap \mathbb{K}[x_{l+1}, \ldots, x_r].$$

I_l is an ideal of $\mathbb{K}[x_{l+1}, \ldots, x_r]$. It consists of all consequences of $f_1 = \cdots = f_s = 0$ which eliminate the indeterminates $x_1, \ldots, x_l$. Eliminating $x_1, \ldots, x_l$ means finding non-zero polynomials in the l-th elimination ideal. With certain term orderings, it is possible to find the l-th elimination ideal. Any term ordering that allows solving this problem is called an **elimination order**. We are not discussing elimination orders in general, it is sufficient for our discussion to know that lex order is an elimination order.

Theorem 2 (The Elimination Theorem, Theorem 2, Cox et al. (2015, chap. 3.1)**)** *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_r]$ be an ideal and let G be a Gröbner basis of I with respect to lex order where $x_1 \succ x_2 \succ \cdots \succ x_r$. Then, for every $0 \leq l \leq r$, the set*

$$G_l = G \cap \mathbb{K}[x_{l+1}, \ldots, x_r]$$

is a Gröbner basis of the l-th elimination ideal I_l.

B Extra simulation results

B.1 Additional scenarios for main results

Table B1: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	1.73	0.67	-43.66	-22.72	0.00	1.48	5.61	2.65	1.12	1.62	44.02	22.87	1.12
	0.9	0.05	8.33	0.87	-40.02	-23.54	-0.02	1.48	4.96	3.00	1.09	1.72	40.33	23.73	1.09
		0.10	16.90	1.17	-34.36	-21.20	0.01	1.44	3.89	2.75	1.02	1.85	34.58	21.38	1.02
		0.01	1.36	0.86	-43.39	-22.22	0.05	2.85	6.67	3.58	2.60	2.98	43.90	22.51	2.60
250	0.8	0.05	6.67	0.97	-40.10	-23.41	0.03	2.92	5.25	3.72	2.69	3.08	40.44	23.71	2.69
		0.10	13.91	1.51	-33.98	-20.79	0.26	2.89	4.83	3.75	2.59	3.26	34.32	21.13	2.61
		0.01	1.12	1.01	-43.66	-21.95	0.16	4.87	7.37	5.35	4.61	4.97	44.27	22.59	4.61
	0.7	0.05	5.40	1.12	-39.89	-23.13	0.01	4.83	6.67	5.23	4.57	4.95	40.44	23.71	4.57
		0.10	11.34	1.19	-33.77	-20.64	-0.12	5.06	6.31	5.55	4.86	5.20	34.36	21.37	4.86
		0.01	1.84	0.58	-42.97	-23.29	0.00	0.93	3.94	1.75	0.77	1.09	43.15	23.36	0.77
	0.9	0.05	8.37	0.73	-40.16	-24.28	0.01	0.97	3.10	1.93	0.79	1.21	40.28	24.36	0.79
		0.10	16.08	1.10	-35.18	-22.20	0.03	1.03	2.49	1.86	0.77	1.51	35.27	22.28	0.77
		0.01	1.46	0.65	-43.13	-23.23	0.04	2.01	4.50	2.42	1.90	2.11	43.37	23.36	1.90
500	0.8	0.05	6.63	0.87	-40.11	-24.10	0.01	2.01	3.95	2.60	1.86	2.20	40.30	24.23	1.86
		0.10	14.11	1.16	-35.09	-21.99	0.10	2.01	2.82	2.35	1.88	2.32	35.20	22.11	1.89
		0.01	1.12	0.86	-43.16	-22.89	0.23	3.31	5.41	3.62	3.21	3.42	43.50	23.17	3.22
	0.7	0.05	5.29	1.24	-39.94	-23.82	0.34	3.34	4.55	3.65	3.25	3.57	40.19	24.10	3.26
		0.10	12.29	1.01	-34.85	-21.95	0.04	3.53	3.89	3.48	3.36	3.67	35.06	22.23	3.36
		0.01	1.77	0.49	-41.95	-23.61	0.02	0.59	3.12	1.31	0.52	0.77	42.07	23.65	0.52
	0.9	0.05	8.34	0.57	-40.42	-24.99	0.00	0.67	2.02	1.34	0.54	0.88	40.47	25.03	0.54
		0.10	17.91	0.95	-35.91	-23.30	-0.01	0.72	1.71	1.31	0.56	1.19	35.95	23.33	0.56
		0.01	1.42	0.46	-42.19	-23.56	0.02	1.34	3.45	1.70	1.28	1.41	42.33	23.62	1.28
1000	0.8	0.05	6.74	0.59	-40.30	-24.94	0.03	1.37	2.42	1.73	1.29	1.49	40.38	25.00	1.29
		0.10	16.30	0.90	-35.85	-23.24	-0.01	1.49	2.06	1.72	1.35	1.74	35.91	23.30	1.35
		0.01	1.11	0.52	-42.34	-23.50	0.10	2.24	3.95	2.39	2.21	2.30	42.53	23.62	2.21
	0.7	0.05	5.41	0.73	-40.18	-24.67	0.18	2.41	3.05	2.46	2.39	2.52	40.29	24.79	2.39
		0.10	14.92	0.79	-35.80	-23.35	-0.04	2.28	2.62	2.37	2.21	2.41	35.90	23.47	2.21

Table B2: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.95	0.14	-53.07	-4.92	0.00	1.23	2.48	2.96	1.12	1.24	53.13	5.74	1.12
	0.9	0.05	7.21	0.48	-49.21	-10.40	-0.02	1.52	2.45	3.67	1.09	1.60	49.27	11.03	1.09
		0.10	15.70	0.30	-42.25	-13.53	0.01	1.74	3.15	3.49	1.02	1.76	42.37	13.98	1.02
		0.01	0.73	0.22	-52.80	-4.80	0.05	2.65	3.02	3.82	2.60	2.66	52.88	6.14	2.60
250	0.8	0.05	5.45	0.52	-49.09	-10.21	0.03	2.98	3.10	4.42	2.69	3.03	49.19	11.13	2.69
		0.10	12.40	0.67	-41.94	-12.94	0.26	3.13	3.91	4.46	2.59	3.20	42.12	13.69	2.61
		0.01	0.51	0.27	-52.76	-4.58	0.16	4.66	3.95	5.46	4.61	4.67	52.91	7.13	4.61
	0.7	0.05	3.79	0.61	-49.02	-9.97	0.01	4.85	3.76	6.04	4.57	4.89	49.16	11.66	4.57
		0.10	9.55	0.38	-41.80	-12.77	-0.12	5.26	5.03	6.25	4.86	5.27	42.10	14.21	4.86
		0.01	0.34	0.05	-53.27	-4.46	0.00	0.81	1.59	1.73	0.77	0.81	53.29	4.78	0.77
	0.9	0.05	2.18	0.72	-49.99	-9.71	0.01	1.02	1.70	2.15	0.79	1.25	50.02	9.94	0.79
		0.10	8.79	1.27	-43.96	-12.57	0.03	1.09	1.94	2.36	0.77	1.68	44.01	12.79	0.77
		0.01	0.31	0.11	-53.26	-4.52	0.04	1.91	2.00	2.50	1.90	1.92	53.30	5.17	1.90
500	0.8	0.05	1.67	0.82	-50.02	-9.44	0.01	2.04	2.02	2.87	1.86	2.20	50.06	9.86	1.86
		0.10	7.87	1.33	-43.73	-12.50	0.10	2.11	2.34	3.04	1.88	2.49	43.79	12.86	1.89
		0.01	0.25	0.28	-53.13	-4.45	0.23	3.21	2.66	3.70	3.21	3.23	53.20	5.79	3.22
	0.7	0.05	1.50	1.13	-49.71	-9.30	0.34	3.34	2.73	3.94	3.25	3.52	49.78	10.10	3.26
		0.10	6.87	1.24	-43.60	-12.34	0.04	3.58	3.11	4.24	3.36	3.79	43.71	13.05	3.36
		0.01	0.18	0.08	-53.27	-4.75	0.02	0.55	1.14	1.10	0.52	0.55	53.28	4.87	0.52
	0.9	0.05	2.01	0.58	-50.55	-9.60	0.00	0.73	1.08	1.42	0.54	0.93	50.56	9.71	0.54
		0.10	7.58	1.19	-44.76	-12.28	-0.01	0.75	1.29	1.69	0.56	1.41	44.78	12.40	0.56
		0.01	0.20	0.07	-53.34	-4.72	0.02	1.30	1.36	1.63	1.28	1.30	53.36	4.99	1.28
1000	0.8	0.05	1.68	0.63	-50.50	-9.63	0.03	1.42	1.34	1.93	1.29	1.56	50.52	9.82	1.29
		0.10	7.34	1.16	-44.73	-12.19	-0.01	1.48	1.58	2.20	1.35	1.88	44.75	12.39	1.35
		0.01	0.19	0.13	-53.32	-4.71	0.10	2.21	1.78	2.48	2.21	2.22	53.35	5.32	2.21
	0.7	0.05	1.40	0.83	-50.28	-9.44	0.18	2.48	1.86	2.71	2.39	2.62	50.31	9.82	2.39
		0.10	7.05	0.96	-44.70	-12.24	-0.04	2.31	2.09	2.92	2.21	2.50	44.75	12.59	2.21

Table B3: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.57	0.02	-53.40	-4.02	0.00	1.16	2.41	2.03	1.12	1.16	53.45	4.50	1.12
	0.9	0.05	1.37	0.57	-51.10	-7.12	-0.02	1.33	2.10	2.12	1.09	1.45	51.14	7.43	1.09
		0.10	5.83	1.04	-46.26	-9.56	0.01	1.40	2.45	2.58	1.02	1.75	46.33	9.90	1.02
		0.01	0.46	0.12	-53.12	-3.91	0.05	2.63	2.98	3.09	2.60	2.63	53.21	4.98	2.60
250	0.8	0.05	1.05	0.60	-50.93	-7.13	0.03	2.87	2.66	3.16	2.69	2.93	51.00	7.80	2.69
		0.10	4.94	1.35	-45.88	-9.23	0.26	2.83	3.13	3.58	2.59	3.13	45.99	9.90	2.61
		0.01	0.33	0.18	-53.06	-3.74	0.16	4.65	3.92	4.87	4.61	4.65	53.20	6.14	4.61
	0.7	0.05	0.88	0.65	-50.78	-7.08	0.01	4.73	3.53	4.75	4.57	4.77	50.90	8.53	4.57
		0.10	4.38	0.95	-45.63	-9.29	-0.12	5.04	4.34	5.31	4.86	5.13	45.83	10.70	4.86
		0.01	0.40	0.02	-53.44	-4.26	0.00	0.80	1.57	1.49	0.77	0.80	53.47	4.51	0.77
	0.9	0.05	1.16	0.45	-51.46	-8.10	0.01	0.94	1.51	1.65	0.79	1.04	51.48	8.27	0.79
		0.10	5.20	1.00	-47.08	-10.80	0.03	1.00	1.68	1.94	0.77	1.41	47.11	10.97	0.77
		0.01	0.35	0.07	-53.44	-4.31	0.04	1.91	2.00	2.30	1.90	1.92	53.48	4.88	1.90
500	0.8	0.05	0.88	0.50	-51.40	-7.96	0.01	1.97	1.87	2.38	1.86	2.03	51.43	8.31	1.86
		0.10	5.08	1.07	-46.87	-10.70	0.10	2.00	2.04	2.63	1.88	2.27	46.91	11.02	1.89
		0.01	0.31	0.22	-53.33	-4.21	0.23	3.21	2.64	3.52	3.21	3.22	53.39	5.49	3.22
	0.7	0.05	0.74	0.76	-51.13	-7.77	0.34	3.30	2.56	3.45	3.25	3.38	51.19	8.50	3.26
		0.10	4.94	0.93	-46.74	-10.70	0.04	3.48	2.76	3.82	3.36	3.60	46.82	11.36	3.36
		0.01	0.19	0.07	-53.43	-4.84	0.02	0.55	1.15	1.14	0.52	0.55	53.44	4.97	0.52
	0.9	0.05	1.13	0.31	-51.76	-9.29	0.00	0.68	1.03	1.28	0.54	0.75	51.77	9.37	0.54
		0.10	5.60	0.89	-47.49	-12.27	-0.01	0.73	1.15	1.50	0.56	1.15	47.50	12.36	0.56
		0.01	0.22	0.05	-53.48	-4.83	0.02	1.29	1.36	1.66	1.28	1.29	53.50	5.10	1.28
1000	0.8	0.05	0.95	0.21	-51.71	-9.32	0.03	1.36	1.29	1.77	1.29	1.38	51.73	9.49	1.29
		0.10	6.25	0.75	-47.44	-12.21	-0.01	1.46	1.47	2.00	1.35	1.64	47.46	12.37	1.35
		0.01	0.23	0.11	-53.45	-4.81	0.10	2.21	1.78	2.51	2.21	2.21	53.48	5.43	2.21
	0.7	0.05	0.79	0.18	-51.49	-9.18	0.18	2.40	1.78	2.54	2.39	2.40	51.52	9.52	2.39
		0.10	6.64	0.44	-47.40	-12.33	-0.04	2.25	1.95	2.70	2.21	2.29	47.44	12.63	2.21

B.2 Results with very strict acceptance thresholds for the classical approach

Table B4: Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.11	0.11	-0.17	0.02	0.01	1.14	1.23	1.14	1.11	1.15	1.24	1.14	1.11
	0.9	0.05	2.48	0.30	2.33	0.35	0.09	1.16	2.13	1.34	1.09	1.20	3.16	1.38	1.10
		0.10	8.79	0.45	3.90	0.58	-0.02	1.72	3.25	1.69	1.12	1.77	5.07	1.79	1.12
		0.01	0.11	0.10	-0.25	0.01	0.01	2.64	2.68	2.64	2.63	2.64	2.69	2.64	2.63
250	0.8	0.05	2.98	0.32	2.09	0.26	0.10	2.73	3.18	2.77	2.68	2.75	3.81	2.78	2.68
		0.10	10.26	0.64	4.80	0.84	0.20	2.81	4.67	3.12	2.73	2.88	6.70	3.23	2.74
		0.01	0.07	0.29	-0.29	0.19	0.20	4.74	4.72	4.74	4.72	4.75	4.73	4.74	4.72
	0.7	0.05	3.45	0.45	2.26	0.39	0.21	4.59	5.02	4.63	4.54	4.62	5.50	4.64	4.54
		0.10	11.69	0.68	5.88	0.91	0.16	4.95	6.51	4.97	4.62	5.00	8.77	5.05	4.63
		0.01	0.09	0.09	-0.07	0.00	0.00	0.78	0.82	0.78	0.76	0.78	0.82	0.78	0.76
	0.9	0.05	3.22	0.18	1.35	0.18	0.00	0.81	1.49	1.06	0.78	0.83	2.01	1.07	0.78
		0.10	10.24	0.39	1.78	0.35	0.03	0.83	2.04	1.30	0.74	0.92	2.71	1.34	0.74
		0.01	0.08	0.28	0.09	0.19	0.18	1.88	1.92	1.88	1.87	1.90	1.93	1.89	1.88
500	0.8	0.05	4.33	0.27	1.70	0.27	0.07	1.94	2.44	2.06	1.90	1.96	2.97	2.07	1.90
		0.10	13.63	0.43	2.34	0.46	0.06	1.90	2.99	2.25	1.82	1.95	3.80	2.30	1.82
		0.01	0.08	0.18	-0.05	0.08	0.09	3.22	3.24	3.22	3.21	3.23	3.24	3.22	3.21
	0.7	0.05	5.18	0.31	1.76	0.22	0.12	3.40	3.70	3.46	3.36	3.42	4.10	3.47	3.36
		0.10	15.75	0.64	3.32	0.66	0.24	3.30	4.44	3.56	3.25	3.36	5.54	3.62	3.25
		0.01	0.05	0.11	0.00	0.01	0.01	0.58	0.61	0.59	0.57	0.59	0.61	0.59	0.57
	0.9	0.05	4.23	0.16	0.62	0.07	0.00	0.56	1.03	0.78	0.54	0.58	1.20	0.78	0.54
		0.10	12.41	0.30	0.77	0.17	0.01	0.60	1.47	1.05	0.54	0.67	1.66	1.06	0.54
		0.01	0.06	0.13	-0.01	0.03	0.04	1.37	1.39	1.37	1.37	1.38	1.39	1.37	1.37
1000	0.8	0.05	6.58	0.15	0.80	0.04	-0.01	1.33	1.64	1.42	1.32	1.34	1.82	1.42	1.32
		0.10	17.09	0.35	1.10	0.15	0.04	1.43	1.99	1.66	1.38	1.48	2.27	1.67	1.38
		0.01	0.07	0.27	0.14	0.17	0.18	2.42	2.42	2.42	2.41	2.43	2.42	2.42	2.41
	0.7	0.05	8.53	0.10	0.91	-0.08	-0.05	2.32	2.63	2.42	2.32	2.32	2.78	2.42	2.33
		0.10	20.30	0.48	1.50	0.10	0.17	2.37	2.99	2.58	2.32	2.42	3.35	2.59	2.33

Table B5: Simulated data: between-variables independence in both sets of matches and non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.08	0.14	-0.13	0.05	0.01	1.14	1.23	1.15	1.11	1.15	1.23	1.15	1.11
	0.9	0.05	2.86	0.31	2.49	0.52	0.09	1.17	2.14	1.38	1.09	1.21	3.28	1.48	1.10
		0.10	9.94	0.44	4.41	0.95	-0.02	1.26	3.39	1.84	1.12	1.33	5.56	2.08	1.12
		0.01	0.07	0.14	-0.21	0.04	0.01	2.65	2.68	2.65	2.63	2.65	2.69	2.65	2.63
250	0.8	0.05	3.27	0.33	2.26	0.48	0.10	2.73	3.19	2.80	2.68	2.75	3.91	2.84	2.68
		0.10	11.21	0.66	5.35	1.30	0.20	2.81	4.84	3.27	2.73	2.89	7.21	3.52	2.74
		0.01	0.05	0.31	-0.24	0.22	0.20	4.74	4.72	4.74	4.72	4.75	4.72	4.75	4.72
	0.7	0.05	3.71	0.46	2.46	0.71	0.21	4.60	5.05	4.70	4.54	4.62	5.62	4.75	4.54
		0.10	12.42	0.66	6.49	1.50	0.16	4.69	6.58	5.06	4.62	4.74	9.24	5.27	4.63
		0.01	0.09	0.10	-0.04	0.02	0.00	0.78	0.82	0.78	0.76	0.78	0.82	0.78	0.76
	0.9	0.05	3.44	0.19	1.49	0.30	0.00	0.81	1.51	1.07	0.78	0.83	2.12	1.12	0.78
		0.10	11.17	0.41	2.20	0.59	0.03	0.84	2.16	1.37	0.74	0.93	3.08	1.49	0.74
		0.01	0.07	0.29	0.12	0.22	0.18	1.88	1.93	1.89	1.87	1.90	1.93	1.90	1.88
500	0.8	0.05	4.48	0.27	1.84	0.41	0.07	1.94	2.48	2.09	1.90	1.96	3.09	2.13	1.90
		0.10	14.28	0.45	2.79	0.74	0.06	1.89	3.08	2.30	1.82	1.95	4.15	2.42	1.82
		0.01	0.08	0.19	-0.01	0.11	0.09	3.22	3.24	3.22	3.21	3.23	3.24	3.23	3.21
	0.7	0.05	5.31	0.32	1.91	0.38	0.12	3.40	3.73	3.48	3.36	3.42	4.19	3.50	3.36
		0.10	16.20	0.67	3.81	1.01	0.24	3.30	4.55	3.67	3.25	3.37	5.94	3.80	3.25
		0.01	0.05	0.10	0.01	0.03	0.01	0.58	0.61	0.59	0.57	0.59	0.61	0.59	0.57
	0.9	0.05	4.23	0.16	0.72	0.13	0.00	0.56	1.04	0.79	0.54	0.59	1.26	0.80	0.54
		0.10	12.62	0.33	1.08	0.31	0.01	0.61	1.54	1.07	0.54	0.69	1.88	1.12	0.54
		0.01	0.06	0.13	0.01	0.05	0.04	1.37	1.39	1.37	1.37	1.38	1.39	1.37	1.37
1000	0.8	0.05	6.55	0.15	0.92	0.14	-0.01	1.33	1.65	1.44	1.32	1.33	1.89	1.44	1.32
		0.10	17.19	0.37	1.43	0.32	0.04	1.43	2.06	1.68	1.38	1.48	2.51	1.71	1.38
		0.01	0.07	0.27	0.16	0.20	0.18	2.42	2.42	2.41	2.41	2.43	2.43	2.42	2.41
	0.7	0.05	8.42	0.10	1.04	0.03	-0.05	2.32	2.64	2.44	2.32	2.33	2.84	2.44	2.33
		0.10	20.45	0.50	1.87	0.31	0.17	2.37	3.13	2.65	2.32	2.43	3.64	2.67	2.33

Table B6: Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.92	0.43	-0.26	0.18	0.03	1.20	1.96	1.30	1.05	1.28	1.97	1.32	1.05
	0.9	0.05	15.87	0.28	2.27	0.56	0.02	1.17	2.86	1.67	1.06	1.20	3.65	1.76	1.06
		0.10	31.28	0.42	3.60	0.97	0.01	1.25	4.33	2.51	1.12	1.32	5.63	2.69	1.12
		0.01	2.28	0.59	-0.21	0.34	0.15	2.84	3.24	2.88	2.78	2.90	3.25	2.90	2.78
250	0.8	0.05	13.35	0.33	2.27	0.52	0.02	2.81	3.99	3.07	2.73	2.83	4.59	3.11	2.73
		0.10	28.01	0.57	4.39	1.15	0.13	2.81	5.62	3.83	2.70	2.87	7.13	4.00	2.70
		0.01	1.80	0.81	-0.25	0.52	0.32	4.86	5.11	4.87	4.76	4.93	5.12	4.90	4.77
	0.7	0.05	11.43	0.46	2.09	0.53	0.14	4.81	5.63	5.01	4.75	4.83	6.01	5.04	4.75
		0.10	24.78	0.69	5.26	1.33	0.25	5.02	7.50	5.69	4.95	5.07	9.16	5.85	4.95
		0.01	2.99	0.36	0.10	0.23	0.05	0.90	1.37	0.95	0.73	0.97	1.37	0.98	0.73
	0.9	0.05	16.35	0.20	1.10	0.29	0.02	0.79	1.97	1.25	0.76	0.81	2.26	1.28	0.76
		0.10	26.91	0.33	1.68	0.57	-0.03	0.84	2.87	1.85	0.75	0.90	3.33	1.93	0.75
		0.01	2.38	0.44	0.10	0.27	0.08	1.90	2.20	1.92	1.82	1.95	2.20	1.94	1.82
500	0.8	0.05	14.53	0.29	1.44	0.39	0.10	1.90	2.90	2.20	1.89	1.92	3.23	2.23	1.89
		0.10	26.23	0.37	1.98	0.51	-0.01	1.91	3.71	2.68	1.83	1.94	4.21	2.73	1.83
		0.01	1.85	0.52	0.04	0.32	0.09	3.40	3.64	3.39	3.27	3.44	3.64	3.41	3.27
	0.7	0.05	12.60	0.40	1.72	0.42	0.18	3.43	4.27	3.65	3.40	3.45	4.61	3.68	3.40
		0.10	25.30	0.50	2.55	0.61	0.07	3.52	5.16	4.14	3.45	3.55	5.76	4.19	3.45
		0.01	1.63	0.28	0.08	0.14	0.04	0.70	1.01	0.70	0.54	0.76	1.01	0.72	0.54
	0.9	0.05	11.54	0.17	0.60	0.21	-0.01	0.56	1.40	0.95	0.54	0.59	1.53	0.98	0.54
		0.10	26.40	0.33	0.78	0.30	0.01	0.60	1.96	1.37	0.52	0.69	2.10	1.41	0.52
		0.01	1.36	0.23	0.07	0.12	-0.03	1.38	1.60	1.39	1.31	1.40	1.60	1.40	1.31
1000	0.8	0.05	12.02	0.16	0.60	0.13	-0.03	1.28	2.01	1.53	1.27	1.29	2.10	1.54	1.27
		0.10	27.82	0.35	0.91	0.23	0.01	1.36	2.58	2.01	1.30	1.41	2.74	2.02	1.30
		0.01	1.10	0.26	0.00	0.09	-0.06	2.34	2.50	2.32	2.23	2.35	2.50	2.32	2.23
	0.7	0.05	12.06	0.20	0.87	0.15	0.02	2.28	2.87	2.49	2.27	2.29	3.00	2.50	2.27
		0.10	27.90	0.41	1.12	0.06	0.04	2.33	3.30	2.74	2.29	2.36	3.49	2.74	2.29

Table B7: Simulated data: between-variables independence in the set of matches, association between the first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.13	0.16	-3.01	-19.19	0.03	1.08	1.87	7.22	1.05	1.10	3.54	20.50	1.05
	0.9	0.05	4.44	0.26	-1.45	-19.29	0.02	1.16	2.82	4.19	1.06	1.19	3.17	19.74	1.06
		0.10	14.99	0.34	-1.86	-8.55	0.01	1.23	4.09	3.16	1.12	1.27	4.49	9.12	1.12
		0.01	0.12	0.30	-3.07	-18.96	0.15	2.82	3.20	7.78	2.78	2.83	4.43	20.50	2.78
250	0.8	0.05	4.36	0.28	-1.62	-19.17	0.02	2.78	4.08	4.92	2.73	2.79	4.39	19.80	2.73
		0.10	15.50	0.50	-2.54	-8.20	0.13	2.80	6.00	4.16	2.70	2.84	6.52	9.20	2.70
		0.01	0.09	0.46	-3.06	-18.34	0.32	4.78	5.03	9.32	4.76	4.80	5.89	20.57	4.77
	0.7	0.05	4.56	0.41	-1.89	-18.65	0.14	4.81	5.65	6.44	4.75	4.83	5.95	19.73	4.75
		0.10	15.56	0.64	-2.13	-8.28	0.25	5.02	8.01	6.27	4.95	5.06	8.29	10.39	4.95
		0.01	0.08	0.12	-2.72	-20.34	0.05	0.76	1.34	5.14	0.73	0.77	3.03	20.97	0.73
	0.9	0.05	4.39	0.12	-2.68	-19.33	0.02	0.80	2.07	2.87	0.76	0.81	3.39	19.54	0.76
		0.10	14.84	0.04	-4.25	-8.28	-0.03	0.81	2.80	2.07	0.75	0.81	5.09	8.54	0.75
		0.01	0.09	0.15	-2.73	-20.14	0.08	1.82	2.18	6.07	1.82	1.82	3.50	21.03	1.82
500	0.8	0.05	5.58	0.20	-2.42	-19.25	0.10	1.90	2.98	3.62	1.89	1.91	3.84	19.59	1.89
		0.10	17.63	0.06	-5.14	-8.12	-0.01	1.88	4.14	2.79	1.83	1.88	6.60	8.58	1.83
		0.01	0.09	0.18	-2.76	-19.65	0.09	3.29	3.59	7.13	3.27	3.29	4.53	20.90	3.27
	0.7	0.05	6.04	0.30	-2.13	-19.12	0.18	3.42	4.31	4.56	3.40	3.43	4.81	19.65	3.40
		0.10	19.50	0.21	-4.69	-8.02	0.07	3.47	5.68	4.07	3.45	3.48	7.36	8.99	3.45
		0.01	0.06	0.08	-2.63	-20.39	0.04	0.55	0.99	3.78	0.54	0.56	2.81	20.74	0.54
	0.9	0.05	4.76	0.04	-3.23	-19.57	-0.01	0.56	1.51	2.11	0.54	0.56	3.57	19.68	0.54
		0.10	15.05	0.05	-5.79	-8.25	0.01	0.55	1.90	1.40	0.52	0.56	6.10	8.37	0.52
		0.01	0.07	0.00	-2.64	-20.32	-0.03	1.32	1.59	4.40	1.31	1.32	3.08	20.79	1.31
1000	0.8	0.05	7.00	0.03	-3.25	-19.60	-0.03	1.28	2.10	2.50	1.27	1.28	3.87	19.76	1.27
		0.10	19.84	0.06	-6.44	-8.31	0.01	1.32	3.03	1.88	1.30	1.32	7.12	8.52	1.30
		0.01	0.09	-0.03	-2.71	-20.31	-0.06	2.24	2.47	5.18	2.23	2.24	3.67	20.96	2.23
	0.7	0.05	8.74	0.08	-2.95	-19.57	0.02	2.29	2.91	3.20	2.27	2.29	4.15	19.83	2.27
		0.10	23.17	0.09	-6.31	-8.43	0.04	2.30	3.71	2.73	2.29	2.30	7.32	8.86	2.29

Table B8: Simulated data: between-variables independence in the set of matches, association between the second and third variable in the set of matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.04	0.05	-0.50	-0.08	-0.06	1.15	1.22	1.16	1.13	1.16	1.32	1.17	1.13
	0.9	0.05	2.27	0.26	1.60	0.03	0.03	1.27	2.06	1.30	1.08	1.29	2.61	1.31	1.08
		0.10	12.03	0.44	3.66	0.83	0.06	1.20	3.17	1.98	1.09	1.28	4.84	2.15	1.09
		0.01	0.06	0.16	-0.50	0.01	0.05	2.66	2.69	2.67	2.65	2.67	2.74	2.67	2.65
250	0.8	0.05	2.24	0.20	0.67	-0.09	0.06	2.72	3.16	2.75	2.69	2.72	3.23	2.75	2.69
		0.10	10.64	0.21	3.62	0.44	-0.13	2.99	4.29	3.24	2.77	3.00	5.61	3.27	2.77
		0.01	0.05	0.26	-0.75	0.11	0.18	4.42	4.33	4.42	4.4	4.42	4.40	4.42	4.41
	0.7	0.05	2.19	0.42	0.67	0.15	0.29	4.52	4.76	4.54	4.47	4.54	4.80	4.54	4.48
		0.10	10.32	0.39	3.88	0.46	0.10	4.70	5.82	4.89	4.65	4.71	6.99	4.91	4.65
		0.01	0.03	0.08	-0.40	-0.01	0.02	0.76	0.85	0.76	0.75	0.76	0.94	0.76	0.75
	0.9	0.05	2.34	0.13	0.56	-0.22	-0.03	0.80	1.35	0.96	0.75	0.81	1.46	0.99	0.75
		0.10	10.77	0.17	1.12	-0.01	-0.03	0.82	2.08	1.40	0.77	0.84	2.36	1.40	0.77
		0.01	0.04	0.23	-0.31	0.13	0.17	1.85	1.87	1.84	1.83	1.86	1.89	1.85	1.84
500	0.8	0.05	3.25	0.20	0.75	-0.15	0.04	1.91	2.24	1.98	1.89	1.92	2.36	1.99	1.89
		0.10	15.45	0.23	1.70	0.11	0.00	1.88	3.02	2.30	1.85	1.90	3.46	2.30	1.85
		0.01	0.04	0.08	-0.48	-0.04	0.03	3.22	3.24	3.22	3.22	3.23	3.27	3.22	3.22
	0.7	0.05	4.26	0.36	1.00	-0.06	0.17	3.39	3.60	3.40	3.38	3.41	3.74	3.40	3.38
		0.10	18.70	0.31	2.42	0.21	0.05	3.28	4.24	3.56	3.24	3.29	4.88	3.57	3.24
		0.01	0.03	0.08	-0.27	-0.05	0.00	0.54	0.63	0.55	0.53	0.55	0.68	0.55	0.53
	0.9	0.05	6.17	0.18	0.53	0.08	0.01	0.57	0.97	0.79	0.54	0.60	1.11	0.80	0.54
		0.10	13.31	0.25	0.79	0.29	0.02	0.55	1.51	1.18	0.53	0.61	1.70	1.21	0.53
		0.01	0.04	0.11	-0.26	-0.01	0.04	1.29	1.33	1.29	1.28	1.29	1.36	1.29	1.28
1000	0.8	0.05	10.30	0.14	0.65	0.06	-0.05	1.30	1.63	1.45	1.30	1.31	1.76	1.45	1.30
		0.10	18.86	0.23	0.75	-0.01	0.02	1.32	2.03	1.72	1.29	1.34	2.16	1.72	1.29
		0.01	0.07	0.03	-0.39	-0.11	-0.03	2.36	2.36	2.36	2.36	2.36	2.39	2.36	2.36
	0.7	0.05	8.05	0.18	0.45	-0.27	0.05	2.28	2.45	2.31	2.26	2.28	2.49	2.32	2.26
		0.10	22.75	0.32	0.97	-0.18	0.08	2.27	2.82	2.54	2.26	2.30	2.98	2.54	2.26

Table B9: Simulated data: association between the second and third variable in the set of matches, between-variables independence in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.12	0.06	-0.20	-0.07	-0.06	1.16	1.20	1.16	1.13	1.16	1.21	1.17	1.13
	0.9	0.05	1.84	0.3	2.65	0.21	0.03	1.15	2.30	1.30	1.08	1.19	3.51	1.31	1.08
		0.10	10.94	0.99	7.28	2.15	0.06	1.38	3.71	1.94	1.09	1.69	8.18	2.90	1.09
		0.01	0.09	0.17	-0.18	0.03	0.05	2.67	2.69	2.67	2.65	2.67	2.70	2.67	2.65
250	0.8	0.05	1.62	0.27	1.42	0.04	0.06	2.73	3.25	2.75	2.69	2.74	3.55	2.75	2.69
		0.10	8.42	0.71	7.42	1.57	-0.13	3	4.91	3.29	2.77	3.08	8.90	3.65	2.77
		0.01	0.06	0.27	-0.35	0.13	0.18	4.41	4.35	4.42	4.4	4.42	4.37	4.42	4.41
	0.7	0.05	1.52	0.52	1.49	0.28	0.29	4.54	4.89	4.54	4.47	4.57	5.11	4.55	4.48
		0.10	7.29	0.79	7.71	1.39	0.10	4.74	6.46	4.92	4.65	4.81	10.06	5.11	4.65
		0.01	0.09	0.09	-0.15	0.00	0.02	0.76	0.81	0.76	0.75	0.76	0.82	0.76	0.75
	0.9	0.05	1.67	0.2	1.72	-0.02	-0.03	0.81	1.58	0.96	0.75	0.83	2.34	0.96	0.75
		0.10	8.81	0.81	4.44	1.11	-0.03	0.95	2.55	1.38	0.77	1.25	5.12	1.77	0.77
		0.01	0.06	0.24	-0.03	0.14	0.17	1.85	1.86	1.84	1.83	1.86	1.86	1.85	1.84
500	0.8	0.05	1.98	0.27	1.79	0.03	0.04	1.92	2.38	1.99	1.89	1.94	2.98	1.99	1.89
		0.10	12.03	0.82	5.27	1.33	0.00	1.98	3.50	2.30	1.85	2.15	6.33	2.65	1.85
		0.01	0.08	0.09	-0.19	-0.02	0.03	3.23	3.23	3.22	3.22	3.23	3.24	3.22	3.22
	0.7	0.05	2.64	0.43	2.07	0.13	0.17	3.39	3.70	3.39	3.38	3.42	4.24	3.40	3.38
		0.10	14.04	0.99	5.88	1.39	0.05	3.35	4.71	3.58	3.24	3.49	7.53	3.84	3.24
		0.01	0.03	0.09	-0.10	-0.04	0.00	0.54	0.57	0.55	0.53	0.55	0.58	0.55	0.53
	0.9	0.05	4.53	0.29	1.63	0.34	0.01	0.58	1.09	0.78	0.54	0.64	1.97	0.85	0.54
		0.10	10.64	1.08	4.09	1.47	0.02	0.77	1.77	1.14	0.53	1.33	4.46	1.86	0.53
		0.01	0.04	0.11	-0.09	0.00	0.04	1.29	1.30	1.29	1.28	1.3	1.30	1.29	1.28
1000	0.8	0.05	7.16	0.24	1.70	0.30	-0.05	1.32	1.75	1.45	1.30	1.34	2.44	1.48	1.30
		0.10	14.41	1.07	4.13	1.20	0.02	1.43	2.39	1.71	1.29	1.79	4.77	2.09	1.29
		0.01	0.06	0.03	-0.18	-0.09	-0.03	2.36	2.35	2.36	2.36	2.36	2.36	2.36	2.36
	0.7	0.05	5.04	0.25	1.55	-0.04	0.05	2.28	2.56	2.32	2.26	2.30	2.99	2.32	2.26
		0.10	17.84	1.09	4.34	1.05	0.08	2.38	3.17	2.55	2.26	2.62	5.37	2.76	2.26

Table B10: Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.59	0.42	-0.48	0.04	0.05	1.22	1.95	1.33	1.07	1.29	2.01	1.33	1.07
	0.9	0.05	16.70	0.27	1.38	0.3	0.01	1.20	2.87	1.74	1.13	1.23	3.19	1.77	1.13
		0.10	32.96	0.35	2.78	0.78	-0.03	1.22	4.03	2.65	1.11	1.27	4.89	2.76	1.11
		0.01	2.36	0.53	-0.53	0.17	0.16	2.78	3.24	2.83	2.71	2.83	3.28	2.84	2.72
250	0.8	0.05	11.33	0.16	0.63	-0.09	-0.07	2.69	3.84	2.92	2.63	2.69	3.89	2.92	2.63
		0.10	28.02	0.36	2.86	0.72	0.03	2.69	4.76	3.48	2.61	2.72	5.55	3.56	2.61
		0.01	2.06	0.53	-0.87	0.15	0.18	4.84	5.05	4.84	4.79	4.87	5.12	4.85	4.79
	0.7	0.05	8.81	0.66	0.82	0.31	0.44	4.78	5.52	4.95	4.74	4.83	5.58	4.96	4.76
		0.10	24.37	0.36	2.57	0.25	0.06	4.71	6.68	5.35	4.69	4.73	7.16	5.36	4.69
		0.01	2.66	0.51	-0.22	0.09	0.07	1.00	1.41	0.95	0.73	1.12	1.42	0.95	0.73
	0.9	0.05	16.02	0.17	0.48	-0.02	0.00	0.81	1.99	1.20	0.77	0.82	2.04	1.20	0.77
		0.10	27.32	0.17	1.00	0.13	-0.04	0.82	2.67	1.84	0.76	0.84	2.85	1.84	0.76
		0.01	2.35	0.41	-0.34	0.02	0.03	1.92	2.19	1.92	1.83	1.97	2.22	1.92	1.83
500	0.8	0.05	13.76	0.26	0.63	0.12	0.05	1.99	2.75	2.18	1.95	2.01	2.82	2.19	1.95
		0.10	28.12	0.27	1.39	0.26	0.03	1.87	3.63	2.73	1.83	1.89	3.89	2.74	1.83
		0.01	1.75	0.64	-0.28	0.24	0.24	3.20	3.45	3.25	3.14	3.27	3.46	3.26	3.15
	0.7	0.05	12.65	0.41	1.08	0.35	0.18	3.40	3.95	3.50	3.35	3.42	4.09	3.52	3.35
		0.10	27.93	0.15	1.49	0.00	-0.13	3.40	4.75	3.91	3.37	3.40	4.98	3.91	3.37
		0.01	1.51	0.21	-0.18	-0.04	0.01	0.68	0.98	0.70	0.55	0.71	1.00	0.70	0.55
	0.9	0.05	13.54	0.23	0.53	0.28	0.03	0.56	1.47	0.99	0.53	0.60	1.56	1.03	0.53
		0.10	26.68	0.27	0.61	0.29	0.01	0.60	2.04	1.51	0.57	0.66	2.13	1.53	0.57
		0.01	1.31	0.35	-0.20	-0.01	0.04	1.43	1.60	1.37	1.30	1.47	1.61	1.37	1.30
1000	0.8	0.05	15.94	0.21	0.61	0.29	-0.03	1.33	2.00	1.61	1.30	1.34	2.09	1.63	1.30
		0.10	29.29	0.27	0.53	0.01	0.03	1.36	2.45	1.97	1.34	1.38	2.50	1.97	1.34
		0.01	1.14	0.33	-0.18	0.01	0.03	2.44	2.56	2.40	2.34	2.46	2.57	2.40	2.34
	0.7	0.05	12.34	0.24	0.18	-0.20	0.06	2.22	2.75	2.37	2.21	2.23	2.75	2.38	2.21
		0.10	30.62	0.30	0.86	0.04	0.02	2.29	3.35	2.86	2.26	2.31	3.46	2.86	2.26

Table B11: Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.58	0.44	-0.16	0.17	0.05	1.23	1.89	1.33	1.07	1.30	1.90	1.34	1.07
	0.9	0.05	16.24	0.36	2.65	0.85	0.01	1.24	3.01	1.83	1.13	1.29	4.01	2.02	1.13
		0.10	31.99	0.84	6.28	2.41	-0.03	1.44	4.66	2.77	1.11	1.66	7.82	3.67	1.11
		0.01	2.35	0.55	-0.19	0.29	0.16	2.80	3.22	2.83	2.71	2.85	3.23	2.85	2.72
250	0.8	0.05	10.79	0.21	1.53	0.27	-0.07	2.69	3.88	2.96	2.63	2.70	4.17	2.97	2.63
		0.10	25.62	0.82	6.45	2.25	0.03	2.77	5.48	3.66	2.61	2.88	8.46	4.30	2.61
		0.01	2.05	0.53	-0.49	0.28	0.18	4.85	5.03	4.85	4.79	4.88	5.06	4.86	4.79
	0.7	0.05	8.16	0.75	1.81	0.73	0.44	4.80	5.60	5.01	4.74	4.86	5.89	5.06	4.76
		0.10	20.86	0.73	6.40	1.70	0.06	4.74	7.36	5.55	4.69	4.80	9.75	5.80	4.69
		0.01	2.68	0.48	0.04	0.20	0.07	0.97	1.35	0.94	0.73	1.08	1.35	0.96	0.73
	0.9	0.05	15.25	0.26	1.66	0.55	0.00	0.83	2.06	1.27	0.77	0.87	2.64	1.39	0.77
		0.10	25.07	0.82	4.16	1.49	-0.04	0.97	3.07	1.89	0.76	1.27	5.17	2.41	0.76
		0.01	2.37	0.36	-0.09	0.12	0.03	1.92	2.16	1.91	1.83	1.96	2.17	1.92	1.83
500	0.8	0.05	12.52	0.34	1.66	0.58	0.05	2.00	2.80	2.20	1.95	2.03	3.25	2.28	1.95
		0.10	24.21	0.87	4.70	1.69	0.03	1.97	4.02	2.77	1.83	2.15	6.18	3.24	1.83
		0.01	1.74	0.62	-0.01	0.35	0.24	3.23	3.46	3.25	3.14	3.29	3.46	3.27	3.15
	0.7	0.05	10.73	0.53	2.28	0.87	0.18	3.40	4.05	3.56	3.35	3.44	4.65	3.66	3.35
		0.10	22.95	0.73	4.82	1.45	-0.13	3.45	5.22	4.02	3.37	3.53	7.10	4.27	3.37
		0.01	1.51	0.19	0.01	0.05	0.01	0.67	0.96	0.69	0.55	0.70	0.96	0.69	0.55
	0.9	0.05	11.57	0.33	1.64	0.85	0.03	0.56	1.54	1.04	0.53	0.65	2.25	1.34	0.53
		0.10	23.60	1.07	3.61	1.58	0.01	0.79	2.27	1.53	0.57	1.33	4.26	2.20	0.57
		0.01	1.33	0.28	0.00	0.08	0.04	1.37	1.57	1.36	1.30	1.40	1.57	1.36	1.30
1000	0.8	0.05	12.36	0.31	1.70	0.85	-0.03	1.33	2.05	1.63	1.30	1.37	2.66	1.84	1.30
		0.10	24.26	1.07	3.68	1.35	0.03	1.45	2.73	2.01	1.34	1.80	4.58	2.42	1.34
		0.01	1.12	0.28	0.05	0.11	0.03	2.43	2.56	2.41	2.34	2.44	2.56	2.41	2.34
	0.7	0.05	9.03	0.31	1.29	0.34	0.06	2.22	2.81	2.42	2.21	2.24	3.09	2.44	2.21
		0.10	24.70	1.09	4.01	1.39	0.02	2.41	3.67	2.91	2.26	2.65	5.44	3.22	2.26

Table B12: Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.12	0.15	-3.14	-17.85	0.05	1.11	1.82	7.13	1.07	1.12	3.63	19.22	1.07
	0.9	0.05	5.30	0.27	-2.15	-19.72	0.01	1.19	2.79	4.06	1.13	1.22	3.52	20.13	1.13
		0.10	18.47	0.37	-2.02	-9.21	-0.03	1.23	4.05	3.46	1.11	1.29	4.53	9.84	1.11
		0.01	0.13	0.29	-3.32	-18.61	0.16	2.72	3.21	7.84	2.71	2.74	4.62	20.19	2.72
250	0.8	0.05	3.47	0.19	-2.88	-21.72	-0.07	2.70	3.82	4.70	2.63	2.70	4.79	22.22	2.63
		0.10	15.35	0.43	-2.98	-10.09	0.03	2.70	5.15	4.63	2.61	2.73	5.95	11.10	2.61
		0.01	0.11	0.28	-3.66	-19.42	0.18	4.79	4.99	9.01	4.79	4.80	6.19	21.41	4.79
	0.7	0.05	3.01	0.73	-2.63	-21.62	0.44	4.77	5.49	6.01	4.74	4.83	6.09	22.44	4.76
		0.10	13.93	0.42	-3.51	-10.80	0.06	4.72	7.04	6.33	4.69	4.74	7.87	12.52	4.69
		0.01	0.08	0.13	-3.01	-20.15	0.07	0.75	1.38	5.43	0.73	0.76	3.31	20.87	0.73
	0.9	0.05	4.75	0.11	-3.09	-20.90	0.00	0.80	2.02	2.90	0.77	0.81	3.69	21.10	0.77
		0.10	16.08	0.00	-4.22	-9.67	-0.04	0.84	2.62	2.31	0.76	0.84	4.96	9.95	0.76
		0.01	0.08	0.09	-3.14	-20.65	0.03	1.84	2.18	5.85	1.83	1.84	3.82	21.46	1.83
500	0.8	0.05	5.07	0.18	-3.06	-20.94	0.05	1.97	2.82	3.56	1.95	1.98	4.16	21.24	1.95
		0.10	19.24	0.11	-4.57	-9.60	0.03	1.89	3.94	3.11	1.83	1.89	6.03	10.09	1.83
		0.01	0.09	0.30	-3.05	-19.93	0.24	3.15	3.43	7.08	3.14	3.17	4.60	21.15	3.15
	0.7	0.05	6.26	0.33	-2.62	-20.50	0.18	3.36	3.94	4.50	3.35	3.37	4.73	20.98	3.35
		0.10	21.14	-0.01	-4.32	-9.89	-0.13	3.39	4.98	4.23	3.37	3.39	6.59	10.76	3.37
		0.01	0.07	0.05	-2.87	-20.51	0.01	0.56	0.98	3.84	0.55	0.56	3.03	20.87	0.55
	0.9	0.05	7.06	0.16	-3.05	-20.00	0.03	0.55	1.53	2.07	0.53	0.58	3.42	20.11	0.53
		0.10	17.01	0.06	-5.20	-9.28	0.01	0.61	2.01	1.67	0.57	0.61	5.57	9.43	0.57
		0.01	0.07	0.08	-2.90	-20.97	0.04	1.31	1.58	4.20	1.30	1.31	3.31	21.39	1.30
1000	0.8	0.05	10.69	0.13	-3.03	-19.72	-0.03	1.32	2.05	2.58	1.30	1.33	3.66	19.89	1.30
		0.10	21.61	0.07	-5.69	-9.55	0.03	1.36	2.72	2.08	1.34	1.36	6.31	9.78	1.34
		0.01	0.11	0.06	-2.85	-20.36	0.03	2.35	2.53	5.26	2.34	2.35	3.82	21.03	2.34
	0.7	0.05	9.47	0.14	-3.43	-21.09	0.06	2.21	2.77	3.18	2.21	2.21	4.41	21.33	2.21
		0.10	24.67	0.10	-5.39	-9.75	0.02	2.28	3.63	2.82	2.26	2.28	6.50	10.15	2.26

Table B13: Simulated data: association between the second and third variable in the set of matches, association between first and second variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.12	0.15	-2.89	-17.79	0.05	1.11	1.80	7.18	1.07	1.12	3.41	19.18	1.07
	0.9	0.05	4.92	0.30	-1.20	-19.30	0.01	1.20	2.96	4.07	1.13	1.24	3.20	19.73	1.13
		0.10	18.14	0.57	0.22	-7.72	-0.03	1.34	4.81	3.27	1.11	1.45	4.82	8.38	1.11
		0.01	0.13	0.29	-3.03	-18.49	0.16	2.72	3.19	7.88	2.71	2.73	4.40	20.09	2.72
250	0.8	0.05	3.01	0.17	-2.12	-21.55	-0.07	2.68	3.87	4.68	2.63	2.69	4.41	22.06	2.63
		0.10	13.89	0.72	-0.35	-8.67	0.03	2.76	6.00	4.50	2.61	2.85	6.01	9.77	2.61
		0.01	0.10	0.29	-3.33	-19.30	0.18	4.80	4.98	9.04	4.79	4.81	5.99	21.31	4.79
	0.7	0.05	2.46	0.70	-1.78	-21.42	0.44	4.75	5.58	6.02	4.74	4.80	5.85	22.25	4.76
		0.10	11.40	0.72	-0.64	-9.42	0.06	4.76	7.91	6.19	4.69	4.81	7.93	11.27	4.69
		0.01	0.05	0.09	-2.80	-20.03	0.07	0.75	1.33	5.45	0.73	0.76	3.10	20.76	0.73
	0.9	0.05	3.87	0.12	-2.17	-20.45	0.00	0.80	2.12	2.91	0.77	0.81	3.03	20.66	0.77
		0.10	15.08	0.24	-2.24	-8.34	-0.04	0.94	3.19	2.17	0.76	0.97	3.89	8.62	0.76
		0.01	0.06	0.07	-2.92	-20.52	0.03	1.84	2.15	5.88	1.83	1.84	3.63	21.34	1.83
500	0.8	0.05	3.79	0.18	-2.24	-20.60	0.05	1.98	2.88	3.55	1.95	1.98	3.65	20.91	1.95
		0.10	17.49	0.34	-2.16	-8.21	0.03	1.94	4.41	3.01	1.83	1.97	4.91	8.74	1.83
		0.01	0.07	0.28	-2.82	-19.80	0.24	3.15	3.44	7.11	3.14	3.16	4.45	21.04	3.15
	0.7	0.05	4.34	0.33	-1.64	-20.12	0.18	3.35	4.04	4.52	3.35	3.36	4.36	20.62	3.35
		0.10	18.52	0.24	-1.83	-8.51	-0.13	3.42	5.52	4.21	3.37	3.43	5.82	9.49	3.37
		0.01	0.04	0.03	-2.71	-20.39	0.01	0.56	0.96	3.85	0.55	0.56	2.88	20.75	0.55
	0.9	0.05	5.84	0.19	-2.19	-19.53	0.03	0.56	1.62	2.05	0.53	0.60	2.72	19.64	0.53
		0.10	16.28	0.18	-3.40	-8.02	0.01	0.69	2.48	1.55	0.57	0.71	4.21	8.17	0.57
		0.01	0.05	0.05	-2.73	-20.85	0.04	1.30	1.56	4.22	1.30	1.31	3.14	21.27	1.30
1000	0.8	0.05	8.22	0.16	-2.17	-19.27	-0.03	1.33	2.11	2.57	1.30	1.34	3.02	19.44	1.30
		0.10	20.25	0.24	-3.43	-8.19	0.03	1.41	3.06	2.01	1.34	1.43	4.60	8.43	1.34
		0.01	0.08	0.04	-2.67	-20.24	0.03	2.35	2.53	5.27	2.34	2.35	3.68	20.91	2.34
	0.7	0.05	6.29	0.15	-2.54	-20.67	0.06	2.22	2.84	3.20	2.21	2.22	3.81	20.92	2.21
		0.10	23.08	0.27	-3.15	-8.41	0.02	2.38	4.02	2.76	2.26	2.39	5.11	8.85	2.26

Table B14: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	5.57	0.21	-51.47	-6.80	0.00	1.22	2.79	4.82	1.12	1.23	51.54	8.33	1.12
	0.9	0.05	28.70	0.28	-44.18	-17.85	-0.02	1.21	3.54	4.43	1.09	1.24	44.32	18.40	1.09
		0.10	55.12	0.28	-33.21	-19.81	0.01	1.11	6.60	4.23	1.02	1.14	33.86	20.25	1.02
		0.01	4.44	0.30	-51.17	-6.58	0.05	2.65	3.28	5.59	2.60	2.67	51.27	8.64	2.60
250	0.8	0.05	22.95	0.34	-44.12	-17.40	0.03	2.77	4.24	5.16	2.69	2.79	44.32	18.15	2.69
		0.10	46.93	0.56	-32.93	-19.05	0.26	2.69	7.86	5.53	2.59	2.75	33.85	19.83	2.61
		0.01	3.60	0.42	-51.19	-6.21	0.16	4.64	4.22	6.60	4.61	4.66	51.36	9.07	4.61
	0.7	0.05	19.82	0.32	-44.06	-16.79	0.01	4.61	4.98	6.73	4.57	4.62	44.34	18.08	4.57
		0.10	39.17	0.26	-32.47	-18.27	-0.12	4.91	9.74	7.06	4.86	4.91	33.90	19.59	4.86
		0.01	4.70	0.28	-51.48	-7.60	0.00	0.83	1.87	3.60	0.77	0.87	51.52	8.41	0.77
	0.9	0.05	21.46	0.34	-44.28	-19.01	0.01	0.86	2.39	3.06	0.79	0.92	44.34	19.26	0.79
		0.10	43.18	0.27	-34.50	-21.07	0.03	0.83	3.37	2.47	0.77	0.87	34.66	21.22	0.77
		0.01	3.67	0.31	-51.44	-7.72	0.04	1.92	2.24	4.36	1.90	1.95	51.49	8.87	1.90
500	0.8	0.05	18.15	0.33	-44.39	-18.53	0.01	1.91	2.88	3.82	1.86	1.94	44.48	18.92	1.86
		0.10	40.12	0.37	-34.14	-20.68	0.10	1.93	4.33	3.13	1.88	1.96	34.42	20.92	1.89
		0.01	2.78	0.47	-51.27	-7.55	0.23	3.23	2.99	5.36	3.21	3.27	51.36	9.26	3.22
	0.7	0.05	15.55	0.71	-44.03	-18.20	0.34	3.29	3.64	4.79	3.25	3.36	44.18	18.82	3.26
		0.10	36.52	0.33	-33.68	-20.33	0.04	3.42	5.94	4.61	3.36	3.44	34.20	20.85	3.36
		0.01	4.82	0.28	-51.03	-9.78	0.02	0.56	1.37	2.91	0.52	0.63	51.05	10.21	0.52
	0.9	0.05	15.25	0.34	-44.26	-20.13	0.00	0.58	1.54	2.12	0.54	0.67	44.28	20.24	0.54
		0.10	34.56	0.23	-34.72	-21.69	-0.01	0.62	1.98	1.58	0.56	0.66	34.78	21.75	0.56
		0.01	3.80	0.24	-51.15	-9.55	0.02	1.31	1.62	3.31	1.28	1.33	51.18	10.11	1.28
1000	0.8	0.05	14.28	0.37	-44.23	-20.02	0.03	1.32	1.90	2.60	1.29	1.37	44.27	20.19	1.29
		0.10	34.84	0.25	-34.68	-21.55	-0.01	1.39	2.40	2.06	1.35	1.41	34.77	21.65	1.35
		0.01	2.90	0.30	-51.14	-9.40	0.10	2.22	2.06	4.08	2.21	2.24	51.18	10.24	2.21
	0.7	0.05	13.30	0.53	-44.01	-19.70	0.18	2.42	2.55	3.25	2.39	2.48	44.09	19.96	2.39
		0.10	34.00	0.26	-34.48	-21.39	-0.04	2.24	3.86	3.14	2.21	2.25	34.69	21.62	2.21

Table B15: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_{2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	4.29	0.26	-52.44	-8.94	0.00	1.23	2.73	5.01	1.12	1.25	52.51	10.24	1.12
	0.9	0.05	23.19	0.43	-46.91	-16.46	-0.02	1.31	3.08	4.72	1.09	1.38	47.01	17.12	1.09
		0.10	47.34	0.60	-38.63	-18.92	0.01	1.42	3.53	3.30	1.02	1.54	38.79	19.21	1.02
		0.01	3.32	0.38	-52.15	-8.57	0.05	2.66	3.20	5.71	2.60	2.68	52.24	10.30	2.60
250	0.8	0.05	18.86	0.48	-46.91	-15.83	0.03	2.81	3.71	5.50	2.69	2.85	47.06	16.76	2.69
		0.10	40.69	0.86	-38.38	-18.11	0.26	2.80	4.44	4.33	2.59	2.93	38.64	18.62	2.61
		0.01	2.76	0.47	-52.15	-8.10	0.16	4.65	4.15	6.69	4.61	4.67	52.31	10.51	4.61
	0.7	0.05	16.64	0.47	-46.89	-15.31	0.01	4.69	4.35	7.00	4.57	4.71	47.09	16.84	4.57
		0.10	34.26	0.54	-38.40	-17.67	-0.12	5.08	5.65	6.03	4.86	5.11	38.82	18.67	4.86
		0.01	3.29	0.30	-52.44	-11.79	0.00	0.86	1.84	3.84	0.77	0.92	52.47	12.40	0.77
	0.9	0.05	16.53	0.42	-47.04	-18.85	0.01	0.87	2.27	3.33	0.79	0.96	47.09	19.14	0.79
		0.10	35.44	0.39	-39.25	-20.75	0.03	0.91	2.41	2.35	0.77	0.99	39.32	20.88	0.77
		0.01	2.60	0.36	-52.39	-11.83	0.04	1.94	2.19	4.36	1.90	1.98	52.43	12.61	1.90
500	0.8	0.05	14.52	0.44	-47.14	-18.34	0.01	1.92	2.65	4.08	1.86	1.97	47.21	18.79	1.86
		0.10	34.06	0.53	-39.07	-20.37	0.10	1.95	2.92	2.95	1.88	2.02	39.18	20.58	1.89
		0.01	2.00	0.51	-52.21	-11.52	0.23	3.25	2.96	5.36	3.21	3.29	52.30	12.70	3.22
	0.7	0.05	13.18	0.80	-46.80	-17.94	0.34	3.32	3.34	5.03	3.25	3.42	46.92	18.63	3.26
		0.10	32.31	0.51	-38.86	-20.21	0.04	3.45	3.76	4.01	3.36	3.48	39.04	20.60	3.36
		0.01	3.41	0.33	-52.00	-16.12	0.02	0.56	1.34	2.78	0.52	0.65	52.01	16.36	0.52
	0.9	0.05	11.47	0.42	-47.09	-22.16	0.00	0.60	1.45	2.17	0.54	0.73	47.11	22.27	0.54
		0.10	27.75	0.30	-39.43	-23.11	-0.01	0.63	1.65	1.65	0.56	0.70	39.47	23.17	0.56
		0.01	2.71	0.28	-52.10	-15.77	0.02	1.31	1.58	3.10	1.28	1.34	52.13	16.08	1.28
1000	0.8	0.05	11.59	0.42	-47.06	-21.98	0.03	1.32	1.80	2.69	1.29	1.38	47.10	22.15	1.29
		0.10	29.66	0.31	-39.39	-22.94	-0.01	1.41	1.99	2.09	1.35	1.44	39.44	23.03	1.35
		0.01	2.13	0.33	-52.08	-15.45	0.10	2.23	2.01	3.89	2.21	2.25	52.12	15.93	2.21
	0.7	0.05	11.87	0.56	-46.83	-21.54	0.18	2.41	2.41	3.36	2.39	2.48	46.89	21.80	2.39
		0.10	30.50	0.31	-39.40	-22.88	-0.04	2.23	2.54	2.67	2.21	2.25	39.48	23.03	2.21

Table B16: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{1,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	77.21	0.07	-43.66	-22.72	0.00	1.15	5.61	2.65	1.12	1.15	44.02	22.87	1.12
	0.9	0.05	74.36	0.13	-40.02	-23.54	-0.02	1.14	4.96	3.00	1.09	1.14	40.33	23.73	1.09
		0.10	72.25	0.18	-34.36	-21.20	0.01	1.09	3.89	2.75	1.02	1.10	34.58	21.38	1.02
		0.01	59.42	0.15	-43.39	-22.22	0.05	2.63	6.67	3.58	2.60	2.63	43.90	22.51	2.60
250	0.8	0.05	58.08	0.18	-40.10	-23.41	0.03	2.73	5.25	3.72	2.69	2.73	40.44	23.71	2.69
		0.10	60.40	0.45	-33.98	-20.79	0.26	2.64	4.83	3.75	2.59	2.68	34.32	21.13	2.61
		0.01	44.25	0.26	-43.66	-21.95	0.16	4.66	7.37	5.35	4.61	4.66	44.27	22.59	4.61
	0.7	0.05	45.61	0.20	-39.89	-23.13	0.01	4.63	6.67	5.23	4.57	4.63	40.44	23.71	4.57
		0.10	49.46	0.11	-33.77	-20.64	-0.12	4.90	6.31	5.55	4.86	4.90	34.36	21.37	4.86
		0.01	17.83	0.39	-42.97	-23.29	0.00	0.88	3.94	1.75	0.77	0.96	43.15	23.36	0.77
	0.9	0.05	26.80	0.37	-40.16	-24.28	0.01	0.86	3.10	1.93	0.79	0.93	40.28	24.36	0.79
		0.10	42.59	0.26	-35.18	-22.20	0.03	0.83	2.49	1.86	0.77	0.87	35.27	22.28	0.77
		0.01	14.51	0.45	-43.13	-23.23	0.04	1.96	4.50	2.42	1.90	2.02	43.37	23.36	1.90
500	0.8	0.05	23.84	0.40	-40.11	-24.10	0.01	1.93	3.95	2.60	1.86	1.97	40.30	24.23	1.86
		0.10	40.31	0.37	-35.09	-21.99	0.10	1.92	2.82	2.35	1.88	1.96	35.20	22.11	1.89
		0.01	11.02	0.63	-43.16	-22.89	0.23	3.28	5.41	3.62	3.21	3.34	43.50	23.17	3.22
	0.7	0.05	21.04	0.76	-39.94	-23.82	0.34	3.31	4.55	3.65	3.25	3.40	40.19	24.10	3.26
		0.10	37.53	0.33	-34.85	-21.95	0.04	3.42	3.89	3.48	3.36	3.43	35.06	22.23	3.36
		0.01	3.42	0.41	-41.95	-23.61	0.02	0.57	3.12	1.31	0.52	0.71	42.07	23.65	0.52
	0.9	0.05	17.41	0.41	-40.42	-24.99	0.00	0.60	2.02	1.34	0.54	0.72	40.47	25.03	0.54
		0.10	33.60	0.25	-35.91	-23.30	-0.01	0.61	1.71	1.31	0.56	0.66	35.95	23.33	0.56
		0.01	2.75	0.37	-42.19	-23.56	0.02	1.33	3.45	1.70	1.28	1.38	42.33	23.62	1.28
1000	0.8	0.05	16.92	0.46	-40.30	-24.94	0.03	1.32	2.42	1.73	1.29	1.39	40.38	25.00	1.29
		0.10	34.55	0.26	-35.85	-23.24	-0.01	1.39	2.06	1.72	1.35	1.42	35.91	23.30	1.35
		0.01	2.21	0.41	-42.34	-23.50	0.10	2.23	3.95	2.39	2.21	2.26	42.53	23.62	2.21
	0.7	0.05	16.28	0.62	-40.18	-24.67	0.18	2.42	3.05	2.46	2.39	2.49	40.29	24.79	2.39
		0.10	34.27	0.27	-35.80	-23.35	-0.04	2.23	2.62	2.37	2.21	2.25	35.90	23.47	2.21

Table B17: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_{2,3}, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	3.02	0.14	-53.07	-4.92	0.00	1.21	2.48	2.96	1.12	1.22	53.13	5.74	1.12
	0.9	0.05	17.06	-0.13	-49.21	-10.40	-0.02	1.41	2.45	3.67	1.09	1.41	49.27	11.03	1.09
		0.10	35.17	-0.72	-42.25	-13.53	0.01	1.68	3.15	3.49	1.02	1.83	42.37	13.98	1.02
		0.01	2.30	0.24	-52.80	-4.80	0.05	2.64	3.02	3.82	2.60	2.65	52.88	6.14	2.60
250	0.8	0.05	14.33	-0.09	-49.09	-10.21	0.03	2.85	3.10	4.42	2.69	2.85	49.19	11.13	2.69
		0.10	31.33	-0.32	-41.94	-12.94	0.26	2.96	3.91	4.46	2.59	2.98	42.12	13.69	2.61
		0.01	1.73	0.25	-52.76	-4.58	0.16	4.65	3.95	5.46	4.61	4.66	52.91	7.13	4.61
	0.7	0.05	11.82	-0.09	-49.02	-9.97	0.01	4.69	3.76	6.04	4.57	4.69	49.16	11.66	4.57
		0.10	27.56	-0.55	-41.80	-12.77	-0.12	4.99	5.03	6.25	4.86	5.02	42.10	14.21	4.86
		0.01	2.96	0.08	-53.27	-4.46	0.00	0.82	1.59	1.73	0.77	0.82	53.29	4.78	0.77
	0.9	0.05	15.66	-0.09	-49.99	-9.71	0.01	0.95	1.70	2.15	0.79	0.96	50.02	9.94	0.79
		0.10	31.05	-0.84	-43.96	-12.57	0.03	1.06	1.94	2.36	0.77	1.35	44.01	12.79	0.77
		0.01	2.44	0.15	-53.26	-4.52	0.04	1.93	2.00	2.50	1.90	1.93	53.30	5.17	1.90
500	0.8	0.05	13.61	-0.06	-50.02	-9.44	0.01	1.92	2.02	2.87	1.86	1.92	50.06	9.86	1.86
		0.10	30.49	-0.83	-43.73	-12.50	0.10	2.12	2.34	3.04	1.88	2.27	43.79	12.86	1.89
		0.01	1.78	0.31	-53.13	-4.45	0.23	3.23	2.66	3.70	3.21	3.25	53.20	5.79	3.22
	0.7	0.05	12.26	0.29	-49.71	-9.30	0.34	3.33	2.73	3.94	3.25	3.34	49.78	10.10	3.26
		0.10	29.23	-0.84	-43.60	-12.34	0.04	3.46	3.11	4.24	3.36	3.56	43.71	13.05	3.36
		0.01	0.90	0.04	-53.27	-4.75	0.02	0.54	1.14	1.10	0.52	0.54	53.28	4.87	0.52
	0.9	0.05	13.22	-0.02	-50.55	-9.60	0.00	0.63	1.08	1.42	0.54	0.63	50.56	9.71	0.54
		0.10	31.39	-0.80	-44.76	-12.28	-0.01	0.74	1.29	1.69	0.56	1.09	44.78	12.40	0.56
		0.01	0.71	0.03	-53.34	-4.72	0.02	1.29	1.36	1.63	1.28	1.29	53.36	4.99	1.28
1000	0.8	0.05	12.80	-0.02	-50.50	-9.63	0.03	1.34	1.34	1.93	1.29	1.34	50.52	9.82	1.29
		0.10	32.23	-0.81	-44.73	-12.19	-0.01	1.45	1.58	2.20	1.35	1.66	44.75	12.39	1.35
		0.01	0.69	0.11	-53.32	-4.71	0.10	2.21	1.78	2.48	2.21	2.21	53.35	5.32	2.21
	0.7	0.05	12.50	0.14	-50.28	-9.44	0.18	2.39	1.86	2.71	2.39	2.40	50.31	9.82	2.39
		0.10	32.11	-0.78	-44.70	-12.24	-0.04	2.26	2.09	2.92	2.21	2.39	44.75	12.59	2.21

Table B18: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	2.35	0.08	-53.40	-4.02	0.00	1.18	2.41	2.03	1.12	1.19	53.45	4.50	1.12
	0.9	0.05	11.41	0.01	-51.10	-7.12	-0.02	1.19	2.10	2.12	1.09	1.19	51.14	7.43	1.09
		0.10	24.27	-0.18	-46.26	-9.56	0.01	1.17	2.45	2.58	1.02	1.18	46.33	9.90	1.02
		0.01	1.87	0.19	-53.12	-3.91	0.05	2.63	2.98	3.09	2.60	2.63	53.21	4.98	2.60
250	0.8	0.05	10.02	0.07	-50.93	-7.13	0.03	2.72	2.66	3.16	2.69	2.72	51.00	7.80	2.69
		0.10	23.03	0.10	-45.88	-9.23	0.26	2.66	3.13	3.58	2.59	2.66	45.99	9.90	2.61
		0.01	1.45	0.27	-53.06	-3.74	0.16	4.65	3.92	4.87	4.61	4.66	53.20	6.14	4.61
	0.7	0.05	8.91	0.06	-50.78	-7.08	0.01	4.61	3.53	4.75	4.57	4.61	50.90	8.53	4.57
		0.10	21.81	-0.24	-45.63	-9.29	-0.12	4.86	4.34	5.31	4.86	4.86	45.83	10.70	4.86
		0.01	1.62	-0.05	-53.44	-4.26	0.00	0.79	1.57	1.49	0.77	0.79	53.47	4.51	0.77
	0.9	0.05	9.80	-0.13	-51.46	-8.10	0.01	0.86	1.51	1.65	0.79	0.87	51.48	8.27	0.79
		0.10	22.80	-0.44	-47.08	-10.80	0.03	0.91	1.68	1.94	0.77	1.01	47.11	10.97	0.77
		0.01	1.32	0.02	-53.44	-4.31	0.04	1.91	2.00	2.30	1.90	1.91	53.48	4.88	1.90
500	0.8	0.05	9.25	-0.11	-51.40	-7.96	0.01	1.90	1.87	2.38	1.86	1.91	51.43	8.31	1.86
		0.10	23.85	-0.35	-46.87	-10.70	0.10	1.94	2.04	2.63	1.88	1.97	46.91	11.02	1.89
		0.01	1.04	0.19	-53.33	-4.21	0.23	3.22	2.64	3.52	3.21	3.22	53.39	5.49	3.22
	0.7	0.05	9.19	0.24	-51.13	-7.77	0.34	3.29	2.56	3.45	3.25	3.30	51.19	8.50	3.26
		0.10	24.30	-0.38	-46.74	-10.70	0.04	3.42	2.76	3.82	3.36	3.44	46.82	11.36	3.36
		0.01	0.21	0.04	-53.43	-4.84	0.02	0.54	1.15	1.14	0.52	0.54	53.44	4.97	0.52
	0.9	0.05	4.91	0.10	-51.76	-9.29	0.00	0.59	1.03	1.28	0.54	0.60	51.77	9.37	0.54
		0.10	16.52	-0.04	-47.49	-12.27	-0.01	0.62	1.15	1.50	0.56	0.62	47.50	12.36	0.56
		0.01	0.26	0.03	-53.48	-4.83	0.02	1.29	1.36	1.66	1.28	1.29	53.50	5.10	1.28
1000	0.8	0.05	6.65	0.12	-51.71	-9.32	0.03	1.30	1.29	1.77	1.29	1.31	51.73	9.49	1.29
		0.10	20.76	-0.02	-47.44	-12.21	-0.01	1.38	1.47	2.00	1.35	1.38	47.46	12.37	1.35
		0.01	0.31	0.10	-53.45	-4.81	0.10	2.21	1.78	2.51	2.21	2.22	53.48	5.43	2.21
	0.7	0.05	8.17	0.29	-51.49	-9.18	0.18	2.39	1.78	2.54	2.39	2.41	51.52	9.52	2.39
		0.10	23.79	-0.04	-47.40	-12.33	-0.04	2.24	1.95	2.70	2.21	2.24	47.44	12.63	2.21

Table B19: Simulated data: between-variables independence in the set of matches, association between the first, second and third variable in the set of non-matches. Estimation model: $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$

				RB				RSE				RRMSE			
τ	π_j	ξ	C_{FS}	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$	$\widehat{\tau}_{FS}$	$\widetilde{\tau}$	$\widetilde{\tau}_c$	$\widehat{\tau}$
		0.01	0.15	-0.02	-54.37	-4.18	0.00	1.15	2.35	2.23	1.12	1.15	54.43	4.74	1.12
	0.9	0.05	3.66	0.00	-53.36	-4.86	-0.02	1.14	1.98	1.86	1.09	1.14	53.40	5.20	1.09
		0.10	12.71	-0.06	-49.58	-7.72	0.01	1.12	2.25	2.17	1.02	1.12	49.63	8.02	1.02
		0.01	0.19	0.04	-54.09	-4.24	0.05	2.61	2.91	3.38	2.60	2.61	54.17	5.42	2.60
250	0.8	0.05	4.04	0.05	-53.22	-4.79	0.03	2.70	2.53	3.01	2.69	2.70	53.28	5.66	2.69
		0.10	13.82	0.18	-49.25	-7.29	0.26	2.63	2.89	3.33	2.59	2.64	49.34	8.01	2.61
		0.01	0.21	0.12	-54.01	-4.09	0.16	4.63	3.86	4.99	4.61	4.63	54.15	6.45	4.61
	0.7	0.05	4.65	0.03	-53.07	-4.80	0.01	4.61	3.33	4.74	4.57	4.61	53.17	6.75	4.57
		0.10	14.81	-0.16	-48.97	-7.53	-0.12	4.88	4.06	5.19	4.86	4.89	49.14	9.15	4.86
		0.01	0.13	-0.05	-54.38	-4.42	0.00	0.78	1.55	1.69	0.77	0.78	54.41	4.73	0.77
	0.9	0.05	3.83	0.02	-53.76	-5.87	0.01	0.82	1.44	1.42	0.79	0.82	53.77	6.04	0.79
		0.10	12.63	-0.06	-50.42	-8.91	0.03	0.84	1.55	1.75	0.77	0.84	50.44	9.08	0.77
		0.01	0.18	0.00	-54.38	-4.59	0.04	1.90	1.95	2.45	1.90	1.90	54.42	5.20	1.90
500	0.8	0.05	5.03	0.00	-53.68	-5.77	0.01	1.89	1.76	2.27	1.86	1.89	53.71	6.20	1.86
		0.10	15.89	0.03	-50.23	-8.87	0.10	1.91	1.90	2.50	1.88	1.91	50.27	9.21	1.89
		0.01	0.26	0.18	-54.27	-4.48	0.23	3.22	2.59	3.58	3.21	3.22	54.33	5.73	3.22
	0.7	0.05	6.55	0.35	-53.44	-5.60	0.34	3.27	2.42	3.41	3.25	3.29	53.49	6.56	3.26
		0.10	18.30	-0.04	-50.10	-8.99	0.04	3.39	2.57	3.71	3.36	3.39	50.16	9.72	3.36
		0.01	0.17	-0.03	-54.36	-5.20	0.02	0.53	1.12	1.24	0.52	0.53	54.37	5.34	0.52
	0.9	0.05	4.29	0.03	-54.07	-7.21	0.00	0.57	0.99	1.22	0.54	0.57	54.08	7.31	0.54
		0.10	12.95	0.04	-50.87	-10.57	-0.01	0.60	1.06	1.45	0.56	0.60	50.89	10.66	0.56
		0.01	0.26	-0.03	-54.41	-5.21	0.02	1.29	1.32	1.75	1.28	1.29	54.42	5.50	1.28
1000	0.8	0.05	7.07	0.06	-54.03	-7.25	0.03	1.29	1.22	1.73	1.29	1.29	54.04	7.46	1.29
		0.10	17.87	0.05	-50.82	-10.56	-0.01	1.38	1.35	1.96	1.35	1.38	50.84	10.74	1.35
		0.01	0.35	0.04	-54.38	-5.24	0.10	2.20	1.74	2.58	2.21	2.21	54.41	5.84	2.21
	0.7	0.05	9.31	0.21	-53.80	-7.16	0.18	2.38	1.69	2.54	2.39	2.39	53.83	7.60	2.39
		0.10	21.57	0.00	-50.78	-10.74	-0.04	2.22	1.80	2.65	2.21	2.22	50.81	11.07	2.21

C Sample code to check identifiability

Wolfram Mathematica code to check local identifiability of the model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$:

```
(* ideal *)
id =
{
(p*mx*my*mz*mw + (1 - p)*ux*uy*uz*uw) - p1,
(p*(1 - mx)*my*mz*mw + (1 - p)*(1 - ux)*uy*uz*uw) - p2,
(p*mx*(1 - my)*mz*mw + (1 - p)*ux*(1 - uy)*uz*uw) - p3,
(p*(1 - mx)*(1 - my)*mz*mw + (1 - p)*(1 - ux)*(1 - uy)*uz*uw) - p4,
(p*mx*my*(1 - mz)*mw + (1 - p)*ux*uy*(1 - uz)*uw) - p5,
(p*(1 - mx)*my*(1 - mz)*mw + (1 - p)*(1 - ux)*uy*(1 - uz)*uw) - p6,
(p*mx*(1 - my)*(1 - mz)*mw + (1 - p)*ux*(1 - uy)*(1 - uz)*uw) - p7,
(p*(1 - mx)*(1 - my)*(1 - mz)*mw + (1 - p)*(1 - ux)*(1 - uy)*(1 - uz)*uw) - p8,
```

```
(p*mx*my*mz*(1 - mw) + (1 - p)*ux*uy*uz*(1 - uw)) - p9,
(p*(1 - mx)*my*mz*(1 - mw) + (1 - p)*(1 - ux)*uy*uz*(1 - uw)) - p10,
(p*mx*(1 - my)*mz*(1 - mw) + (1 - p)*ux*(1 - uy)*uz*(1 - uw)) - p11,
(p*(1 - mx)*(1 - my)*mz*(1 - mw) + (1 - p)*(1 - ux)*(1 - uy)*uz*(1 - uw)) - p12,
(p*mx*my*(1 - mz)*(1 - mw) + (1 - p)*ux*uy*(1 - uz)*(1 - uw)) - p13,
(p*(1 - mx)*my*(1 - mz)*(1 - mw) + (1 - p)*(1 - ux)*uy*(1 - uz)*(1 - uw)) - p14,
(p*mx*(1 - my)*(1 - mz)*(1 - mw) + (1 - p)*ux*(1 - uy)*(1 - uz)*(1 - uw)) - p15,
(p*(1 - mx)*(1 - my)*(1 - mz)*(1 - mw) + (1 - p)*(1 - ux)*(1 - uy)*(1 - uz)*(1 - uw)) - p16
}

(* compute Jacobian *)
jac = D[id, {#}] & /@ {mx, my, mz, mw, ux, uy, uz, uw, p}

(* rank of the Jacobian *)
MatrixRank[Transpose[jac]]
```

Maple code to check generic identifiability of the parameter π in the model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4) + (1-\pi)\nu_p(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ using Gröbner basis approach:

```
# number of cpus to use
kernelopts(numcpus = 8);

# Groebner package
with(Groebner);

# define ideal, t is used to get rid of trivial output p1 + ... + p16 - 1
id :=
[t*p - q,
t*(p*mx*my*mz*mw + (1 - p)*ux*uy*uz*uw) - p1,
t*(p*(1 - mx)*my*mz*mw + (1 - p)*(1 - ux)*uy*uz*uw) - p2,
t*(p*mx*(1 - my)*mz*mw + (1 - p)*ux*(1 - uy)*uz*uw) - p3,
t*(p*(1 - mx)*(1 - my)*mz*mw + (1 - p)*(1 - ux)*(1 - uy)*uz*uw) - p4,
t*(p*mx*my*(1 - mz)*mw + (1 - p)*ux*uy*(1 - uz)*uw) - p5,
t*(p*(1 - mx)*my*(1 - mz)*mw + (1 - p)*(1 - ux)*uy*(1 - uz)*uw) - p6,
t*(p*mx*(1 - my)*(1 - mz)*mw + (1 - p)*ux*(1 - uy)*(1 - uz)*uw) - p7,
t*(p*(1 - mx)*(1 - my)*(1 - mz)*mw + (1 - p)*(1 - ux)*(1 - uy)*(1 - uz)*uw) - p8,
t*(p*mx*my*mz*(1 - mw) + (1 - p)*ux*uy*uz*(1 - uw)) - p9,
t*(p*(1 - mx)*my*mz*(1 - mw) + (1 - p)*(1 - ux)*uy*uz*(1 - uw)) - p10,
t*(p*mx*(1 - my)*mz*(1 - mw) + (1 - p)*ux*(1 - uy)*uz*(1 - uw)) - p11,
t*(p*(1 - mx)*(1 - my)*mz*(1 - mw) + (1 - p)*(1 - ux)*(1 - uy)*uz*(1 - uw)) - p12,
t*(p*mx*my*(1 - mz)*(1 - mw) + (1 - p)*ux*uy*(1 - uz)*(1 - uw)) - p13,
```

```
t*(p*(1 - mx)*my*(1 - mz)*(1 - mw) + (1 - p)*(1 - ux)*uy*(1 - uz)*(1 - uw)) - p14,
t*(p*mx*(1 - my)*(1 - mz)*(1 - mw) + (1 - p)*ux*(1 - uy)*(1 - uz)*(1 - uw)) - p15,
t*(p*(1 - mx)*(1 - my)*(1 - mz)*(1 - mw)
+ (1 - p)*(1 - ux)*(1 - uy)*(1 - uz)*(1 - uw)) - p16];

# suggest orders
vord1 := SuggestVariableOrder(id, [p, mx, my, mz, mw, ux, uy, uz, uw, t]);
vord2 := SuggestVariableOrder(id, [q,p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15,p16]);

# compute using the product order with tdeg (graded reverse lexicographic)
gb := Basis(id, prod(tdeg(vord1), tdeg(vord2))):;

# convert to lex using the Groebner walk
gb2 := Walk(gb, prod(tdeg(vord1),tdeg(vord2)),
plex(mw, mz, my, mx, uw, uz, uy, ux, p, t,
p16, p15, p14, p12, p8, p13, p11, p10, p7, p6, p4, p9, p5, p3, p2, p1, q)):;
```

Singular code to check rational identifiability of the model $\pi(\boldsymbol{\gamma}_p; \pi, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi\mu_p(\gamma_1, \gamma_{2,3}, \gamma_4) + (1 - \pi)\nu_p(\gamma_{1,2}, \gamma_3, \gamma_4)$

```
/* set-up */
option(prot);
short = 0;
ring r = (0, uy1c, uy0c, mz1c, mz0c, mxc, myc, mwc, uxc, uzc, uwc, pc),
(uy1, uy0, mz1, mz0, mx, my, mw, ux, uz, uw, p), lp;

/* ideal */
ideal I = p*mx*my*mz1*mw + (1 - p)*ux*uy1*uz*uw -
(pc*mxc*myc*mz1c*mwc + (1 - pc)*uxc*uy1c*uzc*uwc),
p*(1 - mx)*my*mz1*mw + (1 - p)*(1 - ux)*uy0*uz*uw -
(pc*(1 - mxc)*myc*mz1c*mwc + (1 - pc)*(1 - uxc)*uy0c*uzc*uwc),
p*mx*(1 - my)*mz0*mw + (1 - p)*ux*(1 - uy1)*uz*uw -
(pc*mxc*(1 - myc)*mz0c*mwc + (1 - pc)*uxc*(1 - uy1c)*uzc*uwc),
p*(1 - mx)*(1 - my)*mz0*mw + (1 - p)*(1 - ux)*(1 - uy0)*uz*uw -
(pc*(1 - mxc)*(1 - myc)*mz0c*mwc + (1 - pc)*(1 - uxc)*(1 - uy0c)*uzc*uwc),
p*mx*my*(1 - mz1)*mw + (1 - p)*ux*uy1*(1 - uz)*uw -
(pc*mxc*myc*(1 - mz1c)*mwc + (1 - pc)*uxc*uy1c*(1 - uzc)*uwc),
p*(1 - mx)*my*(1 - mz1)*mw + (1 - p)*(1 - ux)*uy0*(1 - uz)*uw -
(pc*(1 - mxc)*myc*(1 - mz1c)*mwc + (1 - pc)*(1 - uxc)*uy0c*(1 - uzc)*uwc),
p*mx*(1 - my)*(1 - mz0)*mw + (1 - p)*ux*(1 - uy1)*(1 - uz)*uw -
(pc*mxc*(1 - myc)*(1 - mz0c)*mwc + (1 - pc)*uxc*(1 - uy1c)*(1 - uzc)*uwc),
```

```
p*(1 - mx)*(1 - my)*(1 - mz0)*mw + (1 - p)*(1 - ux)*(1 - uy0)*(1 - uz)*uw -
(pc*(1 - mxc)*(1 - myc)*(1 - mz0c)*mwc + (1 - pc)*(1 - uxc)*(1 - uy0c)*(1 - uzc)*uwc),
p*mx*my*mz1*(1 - mw) + (1 - p)*ux*uy1*uz*(1 - uw) -
(pc*mxc*myc*mz1c*(1 - mwc) + (1 - pc)*uxc*uy1c*uzc*(1 - uwc)),
p*(1 - mx)*my*mz1*(1 - mw) + (1 - p)*(1 - ux)*uy0*uz*(1 - uw) -
(pc*(1 - mxc)*myc*mz1c*(1 - mwc) + (1 - pc)*(1 - uxc)*uy0c*uzc*(1 - uwc)),
p*mx*(1 - my)*mz0*(1 - mw) + (1 - p)*ux*(1 - uy1)*uz*(1 - uw) -
(pc*mxc*(1 - myc)*mz0c*(1 - mwc) + (1 - pc)*uxc*(1 - uy1c)*uzc*(1 - uwc)),
p*(1 - mx)*(1 - my)*mz0*(1 - mw) + (1 - p)*(1 - ux)*(1 - uy0)*uz*(1 - uw) -
(pc*(1 - mxc)*(1 - myc)*mz0c*(1 - mwc) + (1 - pc)*(1 - uxc)*(1 - uy0c)*uzc*(1 - uwc)),
p*mx*my*(1 - mz1)*(1 - mw) + (1 - p)*ux*uy1*(1 - uz)*(1 - uw) -
(pc*mxc*myc*(1 - mz1c)*(1 - mwc) + (1 - pc)*uxc*uy1c*(1 - uzc)*(1 - uwc)),
p*(1 - mx)*my*(1 - mz1)*(1 - mw) + (1 - p)*(1 - ux)*uy0*(1 - uz)*(1 - uw) -
(pc*(1 - mxc)*myc*(1 - mz1c)*(1 - mwc) + (1 - pc)*(1 - uxc)*uy0c*(1 - uzc)*(1 - uwc)),
p*mx*(1 - my)*(1 - mz0)*(1 - mw) + (1 - p)*ux*(1 - uy1)*(1 - uz)*(1 - uw) -
(pc*mxc*(1 - myc)*(1 - mz0c)*(1 - mwc) + (1 - pc)*uxc*(1 - uy1c)*(1 - uzc)*(1 - uwc)),
p*(1 - mx)*(1 - my)*(1 - mz0)*(1 - mw) + (1 - p)*(1 - ux)*(1 - uy0)*(1 - uz)*(1 - uw) -
(pc*(1 - mxc)*(1 - myc)*(1 - mz0c)*(1 - mwc) +
(1 - pc)*(1 - uxc)*(1 - uy0c)*(1 - uzc)*(1 - uwc));

/* compute Groebner basis */
ideal G = slimgb(I);

/* check if f = p - pc is contained in I */
/* if contained, normal form (NF) is 0 */
NF(p - pc, std(I));

/* check the rest in the same way... */
```