Marginal modelling of capture-recapture data

Elizabeth L. Turner

Department of Mathematics and Statistics,
McGill University, Montréal
Québec, Canada

June, 2007

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-38655-2
Our file Notre référence
ISBN: 978-0-494-38655-2

Abstract

The central theme of this dissertation is the development of a new approach to conceptualize and quantify dependence structures of capture-recapture data for closed populations, with specific emphasis on epidemiological applications. We introduce a measure of source dependence: the Coefficient of Incremental Dependence (CID). Properties of this and the related Coefficient of Source Dependence (CSD) of Vandal, Walker, and Pearson (2005) are presented, in particular their relationships to the conditional independence structures that can be modelled by hierarchical joint log-linear models (HJLLM). From these measures, we develop a new class of marginal log-linear models (MLLM), which we compare and contrast to HJLLMs.

We demonstrate that MLLMs serve to extend the universe of dependence structures of capture-recapture data that can be modelled and easily interpreted. Furthermore, the CIDs and CSDs enable us to meaningfully interpret the parameters of joint log-linear models previously excluded from the analysis of capture-recapture data for reasons of non-interpretability of model parameters.

In order to explore the challenges and features of MLLMs, we show how to produce inference from them under both a maximum likelihood and a Bayesian paradigm. The proposed modelling approach performs well and provides new insight into the fundamental nature of epidemiological capture-recapture data.

Résumé

Le thème central de la présente thèse est le développement d'une nouvelle approche conceptuelle et quantitative envers les structures de dépendance de données de capture-libération obtenues de population fermée, particulièrement en ce qui concerne les applications épidémiologiques. On propose une mesure de dépendance des sources de données : le coefficient de dépendance incrémentielle (CID). On démontre les propriétés de cette mesure et du coefficient de dépendance de source (CSD) de Vandal et al. (2005), en particulier leurs relation avec les structures d'indépendance conditionnelle habituellement modélisées à l'aide de modèles log-linéaires hiérarchiques conjoint (HJLLM). À partir de la forme des deux mesures, on développe une nouvelle classe de modèles log-linéaires marginaux (MLLM), que nous comparons et contrastons aux HJLLM.

On démontre que les MLLMs élargissent l'univers des structures de dépendance de données de capture-libération qui peuvent être modélisées et aisèment interprétées. De plus, les CID et CSD permettent une interprétation des paramètres des modèles log-linéaires non hiérarchiques. Auparavant, ces modèles étaient exclus de l'analyse des données de type capture-libération à cause de l'impossibilité d'interpréter les paramètres du modèle.

Afin d'explorer les problématiques et les caractéristiques des MLLM, on présente des résultats de l'approche de vraisemblance ainsi que de l'approche bayésienne. Pour résumer, l'approche de modélisation proposée offre des résultats satisfaisants et ouvre

de nouvelles perspectives sur la nature même des données de capture-libération dans le domaine de l'épidémiologie.

Acknowledgements

I would first like to thank my advisor, Professor Alain Vandal of the Department of Mathematics and Statistics, for introducing me to the diverse field of capture-recapture and posing research questions which have enabled me to delve into a plethora of related and intersecting branches of statistics. The varied, challenging and fun meetings we have had over the past five years have opened my eyes to many ideas and have helped concretize those learnt in the classroom, with a good balance of theory and practice. I have thoroughly enjoyed being able to discuss a range of other (non-statistical) issues in a challenging fashion. I thank Professor Vandal for his great contributions in developing the foundational work presented in this dissertation, his assistance in formalizing both the notation and the general results contained within, as well as his editorial assistance and assistance with the development of the R code used in Chapter 3.

Several other people have also served as inspiring statistical mentors. I have been fortunate that Dr. Russell Steele has spent much time helping and talking to me, despite him having many commitments to fulfill for his own students. I am indebted to him for the time he has taken over and above that which was expected. He has been incredibly helpful, both statistically and personally. I acknowledge his support, encouragement and wealth of knowledge which greatly aided the development of the Bayesian work including his continual involvement in the development of the C++ code used in Chapter 5 and, indeed, with the general ideas contained in this

dissertation.

I owe much to Professor James Hanley of the Department of Epidemiology and Biostatistics who served as my master's supervisor and has continued to provide much support and guidance, including editorial assistance for this dissertation and introducing me to potential employers. His love of statistics is infectious and is one of the main reasons I began a PhD, following the completion of my MSc work with him. Dr. Nandini Dendukuri, also of the Department of Epidemiology and Biostatistics, in her role as co-supervisor of my MSc work, played a major role in my decision to pursue my studies. Professor David Wolfson, statistician and Chair of the Department of Mathematics and Statistics, has been a guide throughout my studies. He is always available and I cannot overestimate the value of his contributions to my general understanding of statistics, although I have been unfortunate to have never taken a class with such a renowned educator.

Through my work with Professor Tina Wolfson, Director of the Division of Clinical Epidemiology of the McGill University Health Center, in particular on Projet Latitude (jointly with Professor Alain Vandal), I have gained invaluable insight into the workings of real data and I thank her for sharing her enthusiasm and professionalism. I am grateful for the mentorship of Professors Rebecca Fuhrer and Abby Lippman, both of the Department of Epidemiology and Biostatistics, albeit non-statistical! These last three women, together with Dr. Nandini Dndukuri, have been role models, something which is important to me as a young female in academia.

My studies would have not been possible without the financial support of the Canada Memorial Foundation Scholarship, which brought me to Canada, the Commonwealth Scholarship which funded part of my PhD studies and the financial assistance of Professors Alain Vandal, James Hanley and Tina Wolfson. Furthermore, I thank them for their support (financial and otherwise) in my attendance at several conferences, which has served me enormously. Having the opportunity to interact

with a range of statisticians in such a context has been an integral part of my development. I also acknowledge the special Marie Curie awards of the European Union which funded my attendance at the Young Statisticians' Training Camp in Oslo, Norway, as well as a series of Alma Mater awards from McGill, that financed a series of other conferences in Canada and the UK.

Throughout my time in the Department of Mathematics and Statistics I have been lucky to work with very helpful and friendly support staff. I thank Angie, Carla, Carmen, Christina, Jackie, Kelty and Raffaella from the main office, Donna and Greg for their IT support, and Monique in the library. Jennifer and Josette as key team members of the workings of Projet Latitude have been a pleasure to work with.

My friends and colleagues in the Department and around McGill have been there for support, sharing of ideas and many other non-work activities. I thank Benoit, Chris D, Dana, Dave, Ella, Geneviève, Juli, Liang Liang, Marc, Marie-Pierre, Marie-Eve, Nick, Olivier (who kindly provided the Latex layout for my dissertation), Simon (who introduced me to a very special person), Jiguo and all the others who are too numerous to mention. I am very lucky to have spent time sharing an office with Geva, Irina, Nora, Xiaoxing and Ye. I wish to make a special acknowledgement of Geva, who, as my longest standing officemate, has been a tremendous friend these past five years. I couldn't have asked for a better person to spent so much time with. I especially acknowledge her warmth and humor, which has brightened the long hours spent working in our office. I thank her also for spending time reading this dissertation for errors (although I claim responsibility for those that remain!).

Being far from home in a foreign land, I have made '1488' a home away from home with my own surrogate family of Bernard, Cédric, Christine, Jean-Sébastian, Laura and Miriam over the years, with many years (to the present day) spent with Dan

and Mudar, who have been like brothers to me, and more recently Hélène, who too has become part of the family. All of these friends have provided me with a place to live in community with people whose lives are far removed from academia, which has been much needed after long hours working on my thesis! I know I will look back on my time at '1488', as one of great personal discovery. All of the '1488' family have become great friends despite some dispersing to places as far and wide as France, England, Australia and the US.

My dearest friends in Montreal: Cat, Gopi, Isabelle, Karen, Laura (in spirit and since her time at '1488'), Marie-Rose, Megan, Natasha and Sandra have kept me sane, especially during this past year. If I began to describe everything about them which I cherish I would continue for pages! My other friends far and wide: Andu, Anth, Chris, Cloé, Dulfary, Ismäel, Jessica, Joseph, Joyce, Julia, Julien, Kate, Laura R., Michäel, Leonora (who has been like a surrogate grandmother to me), Moa, Oleg and Tansy, have also had a special role to play in my life.

My weekly Monday night rehearsals with *I Medici di McGill* orchestra under the inspiring leadership of Iwan Edwards have provided a perfect, creative and musical counterpoint to the rigors of my daily research. I had put away my violin for several years but my rediscovery of this instrument more than six years ago has opened many doors, no least my appreciation of classical music, an improvement in my musically-challenged ear and the opportunity to play with some outstanding soloists. Most memorable were the performances of the Brahms double violin and cello concerto, as well as the Brahms cello concerto, and the 'post-mortems' delivered by Iwan Edwards after each of the three yearly concerts. Performing in these concerts has been a marvellous opportunity to work towards a great event to share with my friends and colleagues of *I Medici di McGill*, in particular, Aude, Diane, Élise, Julie, Stéphanie, Johanne and Marcus.

Many other activities and places have had quite an effect on me and have con-

tributed to my success these past years (in no particular order): libraries (Bibliothèque Nationale du Québec, and the Birks Reading Room and Nahum Gelber Law Library of McGill); my bookclub (the bookclub where you don't need to read the book!) and the novels read, films watched, discussions had and emotions shared; weekends away in Magog; chalets in the Adirondacks; cross-country skiing and sleeping in refuges; sea-kayaking with the whales at Tadoussac; the roof terrace of '1488' and the unforgettable view; cycling through the Old Port at dusk; walking through Parc Lafontaine in the morning light; the many experiences of living in bilingual Montreal; dancing outside at the many festivals in this wonderful city; and all the other experiences, too numerous to list, which will stay with me.

Throughout all of my experiences related to both my work and my personal life, my family, both close and extended, have been of constant support and encouragement. I thank particularly my mother, father and sister Cath for sharing my experiences on what has been quite a journey of discovery (both research and personal) these past five years. Their love and dedication and superb senses of humour have kept me close to them, despite much of our communication not being in person. I know that their support in my choice to spend a year on exchange in Paris contributed to my decision to move to Montreal for my graduate studies. I especially want to thank my parents for instilling a love of life, travel, dedication to work and, especially for the importance they place on human relationships. And I thank them for helping me fly across the Atlantic when my student income didn't stretch to buying expensive plane tickets!

I wish to acknowledge the role my grandmother has played in my life. Although she died only a month before the completion of this dissertation, I am lucky that she lived such a long life to see over me to this stage of my life. My dissertation is dedicated to her.

Table of Contents

List of Tables

List of Figures

Overview

In this dissertation we present a new modelling approach for capture-recapture data of closed populations. The main motivation for the work is the application of capture-recapture methodology (CRM) to epidemiology; in this context several overlapping sources of data that record individuals with the condition of interest are used. Statistical modelling is used to account for the features of the observed data in order to estimate the unknown total number of individuals N with the condition of interest. CRM is particularly useful for (a) populations that are hard to count perhaps because the individuals, such as intravenous drug users, do not wish to be identified, and (b) for rare conditions, such as amyotrophic lateral sclerosis, an acute disease of the central nervous system, for which alternative enumeration approaches would be prohibitively expensive.

Dependence between sources is the primary feature to take into account when undertaking statistical modelling of capture-recapture data. One of the most common approaches is the use of joint log-linear models in which sources are treated as factors and source dependence is modelled via the inclusion of interaction terms. In particular, hierarchical joint log-linear models are most widely used in order to model conditional dependence structures, maintain interpretability of model parameters and apply theory from graphical models.

The central theme of this dissertation is the development of a new approach to conceptualize and quantify dependence structures of capture-recapture data. In so

doing we present a new framework in which to explore the relationship between N and the dependence structure and the inherent problem of non-identifiability: namely that knowledge of the population size N together with the observed data fully determines the dependence structure of the data. Nonetheless, estimating N when incomplete but overlapping data sources are available, is of relevance to many areas of public health.

We develop a measure of source dependence: the Coefficient of Incremental Dependence (CID), related to the Coefficient of Source Dependence (CSD) of Vandal et al. (2005). For K sources these measures exist for all possible marginal combinations of sources. Properties of the measures are provided, including their relationships to the specific conditional independence structures that can be modelled by hierarchical joint log-linear models.

The form of these measures motivates the development of a new modelling approach in which marginal means are modelled rather than joint (or cell) means as is the case with joint log-linear models, which we name the class of marginal log-linear models (MLLM). Two equivalent parameterizations are presented, the first in terms of the CIDs, the second in terms of the CSDs. Both provide their own useful interpretations. In fact, these measures enable us to meaningfully interpret the parameters of non-hierarchical joint log-linear models. In turn, this enables us to sensibly extend the universe of dependence structures able to be modelled using joint log-linear models which were previously excluded for reasons of non-interpretability of model parameters.

In order to relate the class of MLLMs to structures modelled by the standard analytical approach of hierarchical joint log-linear modelling, we derive the form of the MLLM for joint and conditional independence structures. First it is shown that, for the simple dependence structures of complete independence and mutual dependence, the marginal modelling approach is equivalent to the joint log-linear modelling ap-

proach. However, even for the three-source case, there is no unconstrained marginal model equivalent to the hierarchical joint log-linear model for conditional independence. Consequently, our new approach is more than a mere reparameterization of the standard hierarchial joint log-linear modelling approach. Thus, the universe of dependence structures that can be modelled is extended via this work.

Inference is made using both the likelihood and Bayesian paradigms. In both cases certain constraints must be enforced on the marginal means originating from the multinomial nature of the cell counts. The CIDs are treated as fixed effects in the likelihood approach whilst the Bayesian formulation assumes that the CIDs are random effects. The latter formulation fits into the class of generalized linear mixed models.

Both real and simulated data are analyzed. Simulated data are used to demonstrate that inference is in line with reality. It is shown that our modelling approach out-performs hierarchial joint log-linear models when the true underlying dependence structure is non-hierarchical. The analysis of real data serves to explore the features and challenges of our proposed class of marginal log-linear models.

In short, the proposed marginal modelling approach performs well and provides new insight into the fundamental nature of epidemiological capture-recapture data.

Chapter 1

Introduction and Motivation

Capture-recapture methodology (CRM) is used in epidemiology to estimate the size of a human population by combining several incomplete sources of information. The total population size is estimated by using the information in the overlap of these sources to estimate how many individuals have not been observed.

In order to understand a disease it is important to obtain accurate estimates of its prevalence. Such knowledge helps in the development of strategies to monitor the disease over time, as well as to implement prevention or management programs. CRM can be used for such ends. It has broad application and has been used to estimate the prevalence or incidence of a number of different disorders, such as neurological disorders including multiple sclerosis (Forbes & Swingler, 1999; Corona & Romàn, 2006), amyotrophic lateral sclerosis (Coffman, Horner, Grambow, & Lindquist, 2005; Preux et al., 2000), epilepsy (Debrock, Preux, & Houinato, 2000), Parkinson's disease (Sanchez, Buritica, Pineda, Uribe, & Palacio, 2004) and dementia (?, ?), as well as other conditions including spina bifida (Hook & Regal, 1980), strokes (Tilling, Sterne, & Wolfe, 2001) birth defects (Fienberg, 1972), fetal alcohol syndrome (Egeland, Perham-Hester, & Hook, 1995), diabetes (Fienberg, 1972; Bruno et al., 1994; Ismail, Beeching, Gill, & Bellis, 2000) and HIV (Bartolucci & Forcina, 2006; Abeni,

Brancato, & Perucci, 1994).

Applications in human populations also extend to a broad range of issues in public health and demography, such as in estimating the number of hospitalizations due to influenza (Grijalva et al., 2006), determining the number of intravenous drug users (Hickman et al., 2004; Domingo-Salvany et al., 1998), of street children (Gurgel, da Fonseca, Neyra-Castaeda, Gill, & Cuevas, 2004), estimating the size of a regional lesbian population (Aaron, Chang, Markovic, & LaPorte, 2003), as well as determining fertility and mortality (Aslan, Ozcebe, Bertan, & Karaagaoglu, 2004). It has also found politically relevant application in the realm of human rights. In particular, it has been used to estimate the number of human rights violations including killings in conflicts in Kosovo (Hagan, Schoenfeld, & Palloni, 2006; Ball & Asher, 2002) and East Timor (Silva & Ball, 2005).

It is believed that the first recorded application in human populations was in the 18th century when Laplace used the methodology to estimate the population of France (Hook & Regal, 1995). In terms of applications to human health, the first use appears to have been by Sekar and Deming (International Working Group for & Forecasting, 1995a) who used CRM to estimate birth and death rates in India and to assess the extent of registration of these events.

Applications of CRM are most relevant for populations that are difficult to enumerate. Reasons for such difficulties might be related to the hidden nature of the condition. For instance intravenous drug users are difficult to reach (Hickman et al., 2004; Domingo-Salvany et al., 1998). An alternative reason might be that the condition of interest is a rare condition, such as amyotrophic lateral sclerosis (Preux et al., 2000).

CRM provides an alternative to standard sampling schemes which might be prohibitively expensive, even for common conditions. In using existing data sources, costs can be reduced. Moreover, estimates obtained more appropriately are better

able to correct for under-count, which occurs under sampling schemes with different sampling, or capture, probabilities. Under-count can arise in many instances, and might be quite large for chronic diseases, such as multiple sclerosis (Forbes & Swingler, 1999), as well as infectious diseases such as Severe Acute Respiratory Syndrome (SARS) (Lange & LaPorte, 2003), since such conditions are often treated by a variety of different health professionals and even by private practitioners, (Lange, Chang, & LaPorte, 2004). Such mechanisms pose a variety of challenges to the detection of individuals with the condition of interest.

Under-count is a problem inherent in population censuses. In recent decades much work has been undertaken to use CRM to correct for this under-count. We will discuss such use and the principles associated with it in Section 1.3. Several countries, including the UK (Brown, Diamond, Chambers, Buckner, & Teague, 1999), Canada (Statistics Canada, 2006) and the USA (Freedman, 1991), have employed CRM to correct for the under-count inherent even in national census data (see Fienberg, 1992, for a bibliography). Typically, a two-source approach is adopted with the second source obtained by a post-enumeration survey in which a sample of the population is re-contacted as a follow-up to the census.

There is extensive literature in the application of CRM to ecology. In fact, it is believed to be the application with the most literature starting from the work of Petersen in 1894 (IWGDMF, 1995). The main difference between applications of CRM to counting animals and counting humans is that a sequence of trappings is usually conducted for animal populations. The sequential time effect must be taken into account in statistical modelling. When studying human populations, there is rarely a natural time ordering to the data sources available. Consequently the chronological capture sequence will not be considered in this dissertation and no discussion of models which take this feature into account will be included.

A note on terminology. When used in epidemiology, capture-recapture tech-

niques are also referred to as multi-list or multi-systems methods. We will employ the 'capture-recapture' denomination. 'Source' is used to denote the random event of sampling a given group and 'list' for the observed sample. Both terms will be used in this dissertation.

1.1 Capture-recapture fundamentals

In order to understand the principles of CRM, we present first an example of an epidemiological data set in the form typically employed, that of the incomplete contingency table. Then we state the assumptions made throughout the work presented in this dissertation.

The goal of CRM is to estimate the unknown population size N. This is achieved using information about the appearance of the n_{obs} observed individuals in several overlapping sources. The source membership data of the n_{obs} individuals is typically summarized in what is called an incomplete contingency table (Fienberg, 1972). For instance, Table 1.1 presents an example of a three-source data set from the literature. It summarizes the source membership data of all $n_{obs} = 271$ individuals used to estimate the prevalence of hepatitis in northern Taiwan (Chao, Tsay, Lin, Shau, & Chao, 2001). The three lists, denoted by A, B and C for consistency with the notation used throughout this dissertation, were described as follows (with their corresponding names from Chao et al., 2001):

- A: list of records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan (P-list);
- B: list of hospital records recorded by the National Quarantine Service (Q-list);
- C: list of records collected by epidemiologists (E-list).

	A_{Yes}		A_{No}	
	B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	28	17	18	63
C_{No}	21	69	55	?

Table 1.1: Hepatitis data of Chao et al. (2001)

In order to present the type of notation used in this dissertation, the general incomplete contingency table for three lists is presented in Table 1.2. The table is

	A_{Yes}		A_{No}	
	B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	n_{ABC}	$n_{A\bar{B}C}$	$n_{\bar{A}BC}$	$n_{\bar{A}\bar{B}C}$
C_{No}	$n_{AB\bar{C}}$	$n_{A\bar{B}\bar{C}}$	$n_{\bar{A}B\bar{C}}$	$n_{\bar{A}\bar{B}\bar{C}} = ?$

Table 1.2: Incomplete Contingency Table: Three Source

incomplete since the number of individuals observed in none of the sources n_{unobs} is unknown.

We consider now the information contained in the incomplete contingency table for three sources. For example, the number of individuals observed in all three sources is denoted by n_{ABC}, whilst the number observed in only source A but none of the other two sources is denoted by $n_{A\bar{B}\bar{C}}$. Thus, from Table 1.1, $n_{ABC} = 28$, whilst $n_{A\bar{B}\bar{C}} = 69$. We note that $n_{unobs} = n_{\bar{A}\bar{B}\bar{C}}$ in the three-source case.

An idea which will be used throughout this dissertation is that of obtaining marginal counts from cell (joint, see Remark 1.1) counts. For instance, the number of individuals observed in source A, irrespective of their membership in B and C, is denoted by n_A. It is evident from Table 1.1 that n_A is the sum of the 4 cell counts which appear in the left half of the table. That is, $n_A = 28+21+17+69 = 135$, written symbolically

as $n_A = n_{ABC} + n_{AB\bar{C}} + n_{A\bar{B}C} + n_{A\bar{B}\bar{C}}$. The number of individuals in sources A and B, irrespective of their membership in source C, denoted by n_{AB}, is another example of a marginal count. Again, it is evident from Table 1.1 that n_{AB} is obtained by summing the 2 entries in the first column of data in the table. Therefore, $n_{AB} = 28 + 21 = 49$, given symbolically by $n_{AB} = n_{ABC} + n_{AB\bar{C}}$. Likewise $n_{A\bar{B}}$ is obtained by summing the second column of the table: $n_{A\bar{B}} = n_{A\bar{B}C} + n_{A\bar{B}\bar{C}} = 17 + 69 = 86$. Similarly, n_A can be decomposed as follows: $n_A = n_{AB} + n_{A\bar{B}} = 49 + 86 = 135$.

Remark 1.1 Terminology: joint/cell and marginal.
The terminology *joint* and *cell* will be used interchangeably throughout this dissertation. In so doing we are able to distinguish between the marginal counts and the cell, or joint, counts of the incomplete contingency table. The terminology *joint* is of particular use when referring to modelling techniques, such as those described in Section 1.2.1.

Example 1.1 In the three-source example provided in Table 1.1, $n_{A\bar{B}\bar{C}} = 69$ is a joint count, whilst $n_A = 135$ and $n_{AB} = 49$ are marginal counts. Note that the joint count of the highest order, namely $n_{ABC} = 28$, is also a marginal count. As a consequence, there are as many marginal counts as cell counts (see Remark 1.3).

Remark 1.2 Notation: $\mathbf{n}_{\mathcal{R}}$.
We describe here notation that is applicable to both joint and marginal counts. $\bar{S}$ will be used to denote the complement of any source S. We further let

$$n_{\mathcal{R}} = \left| \bigcap_{S \in \mathcal{R}} S \right|,$$

where $\mathcal{R}$ is a set of sources or source complements.

Example 1.2 Consider the hepatitis data set of Table 1.1. Then, for $\mathcal{R} = \{A, \bar{B}, \bar{C}\}$, $n_{\mathcal{R}} = n_{\{A,\bar{B},\bar{C}\}} = 69$, which we denote by $= n_{A\bar{B}\bar{C}}$ for simplicity. Alternatively, for

$\mathcal{R} = \{A\}$, $n_{\mathcal{R}} = n_{\{A\}} = 135$, which, again for simplicity, we denote by n_A. We note that it is possible to distinguish between the use of $n_{\mathcal{R}}$ for cell and marginal counts by examining the cardinality of $\mathcal{R}$: for the general K-source case, should $|\mathcal{R}| = K$, then $n_{\mathcal{R}}$ refers to a cell count, else to a marginal count should $|\mathcal{R}| < K$ (except in the case of the marginal, equivalently joint, K-way count, which is of order K).

Remark 1.3 Notation: $\mathbf{n}_{\text{cell}}$, $\mathbf{n}_{\text{marg}}$, $\boldsymbol{\mu}$ and $\mathbf{m}$.

For clarity, we will use the three-source case to introduce the notation. The cell and marginal counts can be summarized in vector form as follows

$$\mathbf{n}_{\text{cell}} = \begin{bmatrix} n_{A\bar{B}\bar{C}} \\ n_{\bar{A}B\bar{C}} \\ n_{\bar{A}\bar{B}C} \\ n_{AB\bar{C}} \\ n_{A\bar{B}C} \\ n_{\bar{A}BC} \\ n_{ABC} \end{bmatrix} \quad \text{and} \quad \mathbf{n}_{\text{marg}} = \begin{bmatrix} n_A \\ n_B \\ n_C \\ n_{AB} \\ n_{AC} \\ n_{BC} \\ n_{ABC} \end{bmatrix}.$$

There is a one-to-one linear transformation between the vectors of cell and marginal counts, as ordered as above, summarized by the following relationship

$$\mathbf{n}_{\text{marg}} = \mathbf{A}\mathbf{n}_{\text{cell}},$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{1.1}$$

We notice that the entries in $\mathbf{n}_{\text{marg}}$ are arranged so that marginal counts of the same order are placed together, with single-source counts followed by pairwise counts, then three-way counts. Moreover, the first source subscript A runs faster than B, which runs faster than C. Throughout this dissertation such a convention will be employed. As a result, the ordering of the entries of $\mathbf{n}_{\text{cell}}$ is such that $\mathbf{n}_{\text{marg}} = \mathbf{A}\mathbf{n}_{\text{cell}}$, with $\mathbf{A}$ the upper triangular matrix that corresponds to such an ordering.

Furthermore, when a probability model is assumed for the cell counts (to be described in Section 1.1.3), $\boldsymbol{\mu}$ and $\mathbf{m}$ will be used to denote the cell and marginal means of the corresponding cell and marginal counts given by $\mathbf{n}_{\text{cell}}$ and $\mathbf{n}_{\text{marg}}$, respectively. Thus, the relationship between $\boldsymbol{\mu}$ and $\mathbf{m}$ is given by $\mathbf{m} = \mathbf{A}\boldsymbol{\mu}$, just as $\mathbf{n}_{\text{marg}} = \mathbf{A}\mathbf{n}_{\text{cell}}$.

Remark 1.4 Notation: $P[\mathcal{R}_\cap]$.
As with Remark 1.2, let $\mathcal{R}$ denote a list of sets (sources or source complements). We define $\mathcal{R}_\cap = \cap_{S\in\mathcal{R}} S$. Then

$$P[\mathcal{R}_\cap] = P[\cap_{S\in\mathcal{R}} S].$$

For simplicity we use $P[\mathcal{R}]$ to denote $P[\mathcal{R}_\cap]$.

Example 1.3 Consider again the hepatitis data set of Table 1.1. Then, for $\mathcal{R} = \{A, \bar{B}, \bar{C}\}$,

$$P[\mathcal{R}_\cap] = P\left[A \cap \bar{B} \cap \bar{C}\right].$$

Alternatively, for $\mathcal{R} = \{A\}$

$$P[\mathcal{R}_\cap] = P[A].$$

Assumptions made throughout the dissertation

There are several assumptions made throughout this dissertation, which are in line with those frequently adopted in the literature (see the International Working Group

for Disease Monitoring and Forecasting, IWGDMF, 1995a). The assumptions will be specified here. For assumptions likely to be violated in the epidemiological context, we provide references to work detailing approaches to accommodate such violations. The meaning and implications of the assumptions will become clearer via the discussions in the remainder of the current chapter.

- The population is closed.
- There is no time ordering to the sources.
- All of the population can be observed.
- There is perfect matching (i.e. no tag loss, in the terminology of the ecological capture-recapture literature).
- All cases are true cases, i.e. there are no false positives.

First we consider the assumption that the population under study is a closed population, together with the second assumption of no time ordering. For animal populations there are often situations, generally associated with the sequential time ordering of trappings, that cause the population under study to be open rather than closed. If animals can die or move from the study site then probabilities of capture will be affected and the population is no longer closed. In epidemiology, such difficulties are usually avoided (Chao et al., 2001 discuss differences between animal and human populations). If the study objective is to determine the prevalence of a condition at a given moment, then the population is closed precisely by definition and by the choice of a specific date on which to measure the prevalence. Thus, the population is assumed to be closed by design. For incidence studies, if the period under study is sufficiently short, then births, deaths or emigration will have minimal effect. Again, it is reasonable to assume the population is closed. Thus, for our purposes and throughout this dissertation, we will assume the population is closed. Accordingly,

we will not review the extensive literature on open models with dynamic components (Schwarz & Seber, 1999, provide an extensive overview of capture-recapture methods for animal populations).

Techniques exist to account for violations of the the fourth and fifth assumptions. Seber, Huakau, and Simmons (2000) and Lee, Seber, Holden, and Huakau (2001) discuss list mismatches for the case of two sources and several sources, respectively, whilst Lee (2002) provides a general discussion of violations of the assumption of perfect matching between sources. Should there be false positive diagnoses, Brenner (1994) and Brenner (1996) explore potential consequences, whilst de Greef et al. (2006) present methods to account for such misdiagnoses with analysis of a real data set.

Remark 1.5 Scope of this dissertation.
The methodology presented in this dissertation is intended to be used for three or more sources, as with all modelling approaches described in later chapters. For reasons of clarity, the two-source case will be considered in the next section.

1.1.1 Dependence

We recall that the primary goal of capture-recapture analysis is to estimate N. In order to do so, it is necessary to account for features of the observed data such as relationships between the sources. The primary feature to address is that of dependence of sources.

In order to introduce the notion of dependence, we will discuss the simplest capture-recapture setting for two sources, A and B. Table 1.3 is the incomplete contingency table for the two-source setting. In this case, $n_{unobs} = n_{\bar{A}\bar{B}}$, just as $n_{unobs} = n_{\bar{A}\bar{B}\bar{C}}$ in the three-source example presented above. The well-known Pe-

	A_{Yes}	A_{No}
B_{Yes}	n_{AB}	$n_{A\bar{B}}$
B_{No}	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}} = ?$

Table 1.3: Incomplete Contingency Table: Two Sources

tersen (Petersen, 1896) estimator of N is given by

$$\hat{N} = \frac{n_A n_B}{n_{AB}}, \tag{1.2}$$

which is based on an assumption of independence between sources A and B, using a hypergeometric distribution for the number observed in the overlap between sources (See Alho, 1990, for a description of the hypergeometric model. Section 1.1.3 presents the probability models to be used in this dissertation, i.e. those on the cell counts of the incomplete contingency table). Under the hypergeometric model, there is a nonzero probability that $n_{AB} = 0$, i.e. that no individuals are observed in both sources. In this case the Petersen estimator has infinite bias. Bias corrections methods have been considered by Chapman (1951) for the two-source case and by Evans and Bonett (1994) and Rivest and Lévesque (2001) for the general K-source case.

Remark 1.6 Petersen Estimator: Departures from independence.
The derivation of the Petersen estimator (1.2) relies on the assumption of independence between sources. In the two-source case, the sources are said to be positively dependent when individuals are more likely to appear in source A if they appear in source B, and vice versa. If we erroneously assume independence in such a situation, N is underestimated. Under the independence assumption we would expect that the number of individuals observed in both sources would be a smaller proportion of the total population size than it would be under positive dependence. In the expression of the Petersen estimator, the value of n_{AB} in the positive dependence setting

will be larger than under the assumption of independence. Since it appears in the denominator it would deflate $\hat{N}$ (see page 3129 of Chao et al., 2001 for a similar discussion).

In a similar manner, if we assume independence when in fact the two sources are negatively dependent, then the Petersen estimator will overestimate N since the n_{AB} term of the denominator of (1.2) will be smaller than expected under independence. A related justification is presented in the section entitled "Two-list model" starting on page 1049 of IWGDMF (1995a).

Such relationships prove useful for a preliminary analysis of capture-recapture data. Even if more than two sources are available, it is useful to calculate all Petersen estimators for all pairs of sources (Wittes, Colton, & Sidel, 1974). Whenever a Petersen estimate for two sources is smaller than the number of individuals observed, n_{obs}, there is evidence of possible positive dependence between those two sources. Of course, lack of such relationships does not imply that sources are independent.

It is not possible to test the assumption of independence in the two-source capture recapture setting. As can be seen in Table 1.3, there are three data points available. A test of independence such as the χ^2 test would require knowledge of $n_{\bar{A}\bar{B}}$. Again, precisely because of the nature of CRM, this value is unknown. In fact, this is a feature of the general K-source capture-recapture setting. For K sources there are $2^K - 1$ data points. Testing for K-way independence would require all 2^K data points thus it is not possible to test for the highest level of dependence. Of course, K-way dependence may still be present. In this dissertation, we present new methods to take into account such dependence.

1.1.2 Relationship between heterogeneity and dependence

Since dependence forms the central theme of this dissertation, we now present an example-based discussion describing the equivalent relationship between heterogeneity of capture probabilities and dependence (where homogeneity of capture probabilities is such that for each source, every individual has the same probability of capture).

As an example we consider the use of CRM to determine the number of intravenous drug users in a city. The present discussion, while loosely based on the work of Hickman et al. (2004), IWGDMF (1995a) and Domingo-Salvany et al. (1998), is intended purely as a conceptual example to examine the relationship between the two assumptions.

Example 1.4 Drug Users.

Suppose that, in an attempt to estimate the prevalence of injection drug use in a given time and place, two of the data sources available are drug arrest records and the database of a drug rehabilitation program. It might be expected that drug users who participate in the rehabilitation program are less likely to commit crimes if the program is successful. Consequently, individuals who participate in the program are less likely to be arrested than those who do not participate. Moreover, in order to assume homogeneity of capture probabilities, it would be necessary to assume that all individuals (whether observed or unobserved) are equally likely to appear in the arrest record source and all are equally likely to participate in the drug rehabilitation program. In reality such assumptions are unlikely to hold. Other factors may affect the probability that an individual is arrested, leading to heterogeneity of capture probabilities on the list of arrests. For the drug rehabilitation program, it is likely that certain kinds of individuals are more likely to participate. For instance, pregnant women might be more likely to appear since they should be followed by a medical

professional who is likely to recommend such a program.

The scenario described above, in which those in the drug rehabilitation program are less likely to be arrested than those who are not, would violate the assumption of homogeneity of capture probabilities within a source. Moreover, this situation is likely to induce dependence between the sources of arrest records and the database of the drug rehabilitation program. The direction of this dependence would be related to the proportion of all N individuals who participate in the rehabilitation program. If a large proportion participate, the expected number of individuals who are arrested will decrease, as compared to a situation of homogeneity in capture probabilities. Such a scenario would lead to negative dependence between sources. Alternatively, if all of those arrested are referred to the rehabilitation program, positive dependence between the two sources would arise. In fact, such dependence would not be driven by the characteristics of the individuals according to which capture probabilities vary; rather it would be due to a referral mechanism applied uniformly to all individuals who are arrested. (Note that this distinction is a very fine one. One could also argue that being arrested alters capture probabilities.) It is not difficult to think of many other different ways in which characteristics of the individuals can cause heterogeneity of capture probabilities which in turn induce dependence. In fact, this is a common feature of most capture-recapture data. •

The above discussion leads to the distinction between two types of source dependence:

- Heterogeneity-induced source dependence, and
- Pure source dependence.

As described above, heterogeneity-induced source dependence (also referred to as apparent dependence by IWGDMF, 1995a), can be attributed to characteristics of the individuals whereas the second, pure source dependence (also referred to as local

dependence or list dependence by Chao et al., 2001), cannot be attributed to characteristics of the individual beyond their membership in some sources. It may arise because of how individuals or their conditions are managed.

Stratification to account for heterogeneity

Stratification involves subdividing the sample of observed individuals into strata according to patterns of discrete covariates. Each stratum is analyzed separately and the population estimates for each stratum added together to obtain the overall population estimate (continuous covariates may be discretized, Plante, Rivest, & Tremblay, 1998, and Darroch, Fienberg, Gloneck, & Junker, 1993). In practice, stratification by observed covariate patterns is used to reduce the effects of heterogeneity of capture probabilities. Indeed, when stratum heterogeneity is ignored, results are biased (Kadane, Meyer, & Tukey, 1999). If the observed strata explain all of the dependence, then all individuals in each stratum will have the same probability of capture within any given source; thus the sources will be conditionally independent given the stratum. However, when there are unobserved covariates that modulate capture probabilities, such stratification is unable to fully correct for heterogeneity. Even if stratification is used, residual dependence may occur within strata and thus dependence modelling techniques are still required.

Statistical modelling approaches for more than two sources in general are therefore needed, irrespectively of whether stratification is used. In Sections 1.2 and 1.3 we will describe parametric and nonparametric approaches. In Section 1.2, in which we introduce the standard log-linear modelling approach (which we refer to as joint log-linear modelling for reasons described below), we will describe the dependence structures that can be modelled by the joint log-linear modelling approach, which is essential for the work developed in Chapter 2 and indeed for the very essence of the new methodological work presented in this dissertation.

1.1.3 Likelihoods for incomplete contingency tables

The discussion provided in this section will show that it is natural to work with the multinomial likelihood. For computational convenience, we can use a known relationship between the multinomial and Poisson distributions in order to employ the more computationally (to be discussed in the context of modelling in Section 1.2) convenient Poisson distribution (Sandland & Cormack, 1984).

The multinomial and Poisson likelihoods

Two-source multinomial likelihood

Consider the simple two-source case for sources A and B as introduced in Section 1.1.1, with the incomplete contingency table given by Table 1.3. We will present the development of the multinomial likelihood in this setting. It is natural to consider the N individuals of the population as being assigned to one of four categories, which we denote by $A\bar{B}, \bar{A}B, AB$ and $\bar{A}\bar{B}$ corresponding to the cells of Table 1.3. Then $n_{A\bar{B}}, n_{\bar{A}B}, n_{AB}$ and $n_{\bar{A}\bar{B}}$, individuals are assigned to each of the four categories with probabilities denoted by $p_{A\bar{B}}, p_{\bar{A}B}, p_{AB}$ and $p_{\bar{A}\bar{B}} = 1 - (p_{A\bar{B}} + p_{\bar{A}B} + p_{AB})$. The corresponding likelihood on the unknown parameters N and the vector of cell probabilities, $\mathbf{p}$, is a multinomial likelihood given by

$$L(N, \mathbf{p}; n_{A\bar{B}}, n_{\bar{A}B}, n_{AB}) = \frac{N!}{n_{A\bar{B}}! n_{\bar{A}B}! n_{AB}! n_{\bar{A}\bar{B}}!} p_{A\bar{B}}^{n_{A\bar{B}}} p_{\bar{A}B}^{n_{\bar{A}B}} p_{AB}^{n_{AB}} p_{\bar{A}\bar{B}}^{n_{\bar{A}\bar{B}}}, \qquad (1.3)$$

where

$$p_{A\bar{B}}, p_{\bar{A}B}, p_{AB}, p_{\bar{A}\bar{B}} > 0 \quad \text{and} \quad p_{A\bar{B}} + p_{\bar{A}B} + p_{AB} + p_{\bar{A}\bar{B}} = 1.$$

The data are composed of the 3 observed cell entries of Table 1.3, where $n_{\bar{A}\bar{B}} = N - n_{obs} = N - (n_{A\bar{B}} + n_{\bar{A}B} + n_{AB})$. Examination of likelihood (1.3) shows that there are four unknown parameters to be estimated, namely N and three components of $\mathbf{p}$ (not 4 since the probabilities are constrained to sum to 1) using the three observed cell entries $n_{A\bar{B}}, n_{\bar{A}B}, n_{AB}$. Thus, without further constraints (such as an assumption

of independence as in the case of the Petersen estimator (1.2)), it is not possible to simultaneously estimate all parameters.

K-source multinomial likelihood

A natural extension to the general K-source case, as described by Sanathanan (1972) is given here in line with the notation introduced by Bishop, Fienberg, and Holland (1975). (See Cormack, 1989, Darroch, 1958, Fienberg, 1972 for a similar development; Basu & Ebrahimi, 2001, Casella & George, 1992 and Huggins, 1989, for the equivalent probability model at the level of the individual; Bunge & Fitzpatrick, 1993, for an overview of a range of probability models.)

The number of observed cells of the incomplete contingency table is given by $d = 2^K - 1$. Let $n_1, \ldots, n_d$ denote the observed cell entries of the incomplete contingency table with a single subscript used to denote the cell entry rather than the inclusion/exclusion notation used for the two-source case above in (1.3). The K-source likelihood for N and the $(d+1) \times 1$ vector of cell probabilities $\mathbf{p}$, is given by

$$L(N, \mathbf{p}; \mathbf{n_{incomp}}) = \frac{N!}{(N - n_{obs})! \prod_{i=1}^{d} n_i!} (1 - p^*)^{N - n_{obs}} \prod_{i=1}^{d} p_i^{n_i}, \tag{1.4}$$

where $n_{obs} = \sum_{i=1}^{d} n_i$, $p^* = \sum_{i=1}^{d} p_i \leq 1$ is the probability that an individual is observed, $p^*, p_i > 0$, $i = 1, \ldots, d$ and $\mathbf{n_{incomp}} = (n_1, \ldots, n_d)$ is the vector of observed data of the K-source incomplete contingency table. Such a likelihood (1.4) is invariant to permutations of the cell entries of the incomplete contingency table.

Remark 1.7 We note that there are $d+1$ parameters (i.e. N and the d probabilities of the $(d+1)$-dimensional vector of probabilities $\mathbf{p}$, since $\sum_{i=1}^{d+1} p_i = 1$) to estimate using the d observed cell entries $\mathbf{n_{incomp}}$. In order to address such overparameterization, it is necessary to place constraints on the parameters of the likelihood, such as by specifying a model for the parameters (See section 1.2). In specifying a model on

the capture probabilities $\mathbf{p} = \mathbf{p}(\boldsymbol{\theta})$ with a reduced set of parameters $\boldsymbol{\theta}$, where $|\boldsymbol{\theta}| < d$, it is possible to simultaneously estimate $\mathbf{p} = \mathbf{p}(\boldsymbol{\theta})$ and N (Bishop et al., 1975).

K-source Poisson likelihood

In order to formulate the Poisson likelihood, suppose that the cell entries n_i are independent Poisson random variables with means $\mu_i = Np_i$ (Cormack, 1989). Then the Poisson likelihood for N and the vector of cell probabilities $\mathbf{p}$ is given by

$$L_P(N, \mathbf{p}; \mathbf{n_{incomp}}) = \prod_{i=1}^{d} \exp(-Np_i) \frac{(Np_i)^{n_i}}{n_i!}, \tag{1.5}$$

where n_i and p_i refer to the same cell counts and cell probabilities, respectively, as with the multinomial likelihood above.

Relationship between multinomial and Poisson likelihoods

The maximum likelihood estimates obtained from the multinomial model and the Poisson model conditioned on the population size N are identical (Sandland & Cormack, 1984). The general equivalence of the Poisson conditioned on N and the multinomial is given, for instance, by Christensen (1997):

Result Let $n_1, \ldots, n_d$ be independent with $n_i \sim \text{Poisson}(\mu_i)$ so that $n_1 + \cdots + n_d \sim \text{Poisson}(\mu_1 + \cdots + \mu_d)$. Then $(n_1, \ldots, n_d)|N \sim \text{multinomial}(N, p_1, \ldots, p_d)$, where $N = n_1 + \cdots + n_d$ and $p_i = \frac{\mu_i}{\mu_1 + \cdots + \mu_d}$, $i = 1, \ldots, d$.

However, inference based on the two models is different: the asymptotic variances of the estimators of N under the two models differ (Sandland & Cormack, 1984), with that under the Poisson model larger than under the multinomial model. However, for parameters which do not involve N, Cormack and Jupp (1991) showed that the asymptotic covariances are the same. (See also Baker, 1994.)

Remark 1.8 The parameter of interest: N vs. $\mathbb{E}[N]$.

The fundamental difference between the multinomial and Poisson distributions and

their use in the capture-recapture setting is the parameter of interest: with the multinomial, inference is made on the parameter N whilst for Poisson inference, N is treated as random, so that inference is made on the parameter $\mathbb{E}[N]$ (Farrington, 2002). As noted above, the same point estimates are obtained (Sandland & Cormack, 1984) irrespective of which distribution is adopted and thus irrespective of which parameter is to be estimated. Farrington (2002) pointed out that treating N as a random variable under the Poisson model adds a level of variation which leads to wider confidence intervals, thus providing further support of the asymptotic variance result of Sandland and Cormack (1984), in which the variance is larger under the Poisson model.

Farrington (2002) argues that when interests lies in the underlying prevalence or incidence of disease, rather than the actual number of cases, then the parameter of primary interest is the expectation $\mathbb{E}[N]$ of N over a suitable superpopulation, rather than its realized value. A related difference of the two models is that under the multinomial model $N \geq n_{obs}$, whilst under the Poisson model there is no such constraint on the parameter $\mathbb{E}[N]$ (Farrington, 2002).

Remark 1.9 The likelihood used in this dissertation.
In this dissertation we are most interested in counting the number of individuals rather than determining properties of the underlying process which drives the condition of interest. Therefore, we will adopt the multinomial likelihood for inference purposes. In cases for which there are computational advantages to be gained, the Poisson likelihood we will be adopted (for example, for the frequentist modelling of Chapter 3).

1.2 Literature review: Parametric approaches to modelling CR data

In this next section we provide a literature review of parametric modelling approaches used to reduce the dimensionality of the capture probability vector $\mathbf{p}$ of the equivalent multinomial (1.4) and Poisson (1.5) likelihoods described in the previous section. Since the new marginal modelling approach introduced in this dissertation (see Section 2.4 for the introduction and Chapters 3 and 5 for frequentist and Bayesian approaches to inference) will be compared to the standard modelling approach of joint log-linear modelling (JLLM), we will provide a reasonably comprehensive description of the use of JLLM in the capture-recapture setting. We begin this section with JLLM before describing other models, in particular a range of individual-level models.

1.2.1 Joint log-linear models

Log-linear modelling is one of the most common approaches to modelling epidemiological capture-recapture data (Fienberg, 1972; Cormack, 1989; IWGDMF, 1995a). More specifically, it is generally accepted that such data should be modelled using hierarchical joint log-linear models (HJLLM) (Fienberg, 1972; Cormack, 1989; Madigan, York, & Allard, 1995; Madigan & York, 1997; Stanghellini & van der Heijden, 2004; and implied in the second paragraph of Hook & Regal, 1997, in which the authors enumerate all possible models for different numbers of sources, of which there are 8, 114 and 6893 models for three, four and five sources, respectively). The inclusion of an interaction between a set of sources in such a model entails the inclusion of all lower-order interactions between sources in that set (Bishop et al., 1975). Such HJLLMs model conditional independence structures (Christensen, 1997), which fit within the framework of graphical models (Lauritzen, 1995) (see description below).

Remark 1.10 We note that further evidence for the use of the reduced class of joint log-linear models given by the class of HJLLMs is provided by practices observed in the literature rather than explicit statements. We know of a single reference to a non-HJLLM in the epidemiological capture-recapture literature (Ismail et al., 2000).

Remark 1.11 Terminology: joint log-linear models
We use the prefix *joint* to distinguish between the commonly-called log-linear models, which are models on the *joint* means of the incomplete contingency table (see Remark 1.1) and the new *marginal* log-linear modelling approach introduced in this dissertation.

Fienberg (1972), Cormack (1989) and Chao et al. (2001) propose three equivalent parameterizations of HJLLMs. As stated by Cormack (1989), the parameterization of Fienberg (1972) uses main effects and interactions averaged over all levels of other factors thus necessitating a series of parameter constraints, whilst that of Cormack is such that a main effect of a particular list contrasts the number of individuals not seen in the list but seen in every other, with those seen in all samples. In such a case the intercept-term corresponds to the logarithm of the expected number of individuals observed in all sources. We adopt the parameterization of Chao et al. (2001), in which the intercept-term corresponds to $\log \mathbb{E}[n_{unobs}]$, so that $\hat{N} = n_{obs} + \hat{n}_{unobs}$, with $\hat{n}_{unobs}$ obtained from the fitted intercept-term. The main effect terms correspond to deviations from $\log \mathbb{E}[n_{unobs}]$. Since the primary goal of a capture-recapture analysis is to estimate N, which is done via estimation of n_{unobs} in the case of HJLLMs, it is useful that one of the model parameters corresponds to n_{unobs}. Moreover, unlike the parameterization of Fienberg (1972), no intricate constraints are required.

Example 1.5 Three-source HJLLMs
For the three-source case, the model for independence and the most general saturated

model are given below, parameterized according to Chao et al. (2001).

HJLLM of Independence			Saturated HJLLM		
$\log \mathbb{E}[n_{A\bar{B}\bar{C}}]$	$=$	$\alpha + \alpha_A$	$\log \mathbb{E}[n_{A\bar{B}\bar{C}}]$	$=$	$\alpha + \alpha_A$
$\log \mathbb{E}[n_{\bar{A}B\bar{C}}]$	$=$	$\alpha + \alpha_B$	$\log \mathbb{E}[n_{\bar{A}B\bar{C}}]$	$=$	$\alpha + \alpha_B$
$\log \mathbb{E}[n_{\bar{A}\bar{B}C}]$	$=$	$\alpha + \alpha_C$	$\log \mathbb{E}[n_{\bar{A}\bar{B}C}]$	$=$	$\alpha + \alpha_C$
$\log \mathbb{E}[n_{AB\bar{C}}]$	$=$	$\alpha + \alpha_A + \alpha_B$	$\log \mathbb{E}[n_{AB\bar{C}}]$	$=$	$\alpha + \alpha_A + \alpha_B + \alpha_{AB}$
$\log \mathbb{E}[n_{A\bar{B}C}]$	$=$	$\alpha + \alpha_A + \alpha_C$	$\log \mathbb{E}[n_{A\bar{B}C}]$	$=$	$\alpha + \alpha_A + \alpha_C + \alpha_{AC}$
$\log \mathbb{E}[n_{\bar{A}BC}]$	$=$	$\alpha + \alpha_B + \alpha_C$	$\log \mathbb{E}[n_{\bar{A}BC}]$	$=$	$\alpha + \alpha_B + \alpha_C + \alpha_{BC}$
$\log \mathbb{E}[n_{ABC}]$	$=$	$\alpha + \alpha_A + \alpha_B + \alpha_C$	$\log \mathbb{E}[n_{ABC}]$	$=$	$\alpha + \alpha_A + \alpha_B + \alpha_C + \alpha_{AB} + \alpha_{AC} + \alpha_{BC}$

Remark 1.12 Further to the stated equivalence of the Poisson conditioned on N and the multinomial as given in Section 1.1.3, Lang (1996b) provides a discussion of the similarities and differences of inference under both likelihood models for general log-linear models, whilst Cormack (1989) does so in the capture-recapture setting.

HJLLMs, conditional independence structures and graphical models

HJLLMs model conditional independence structures (Christensen, 1996). As stated by Madigan and York (1997), such dependence structures can be represented by undirected, chordal graphs, which are termed decomposable graphical models (Whittaker, 1990 and Lauritzen, 1995). Dependence is modelled by the inclusion of interaction terms, like those of the saturated model of Example 1.5.

Using the three-source case as an illustration, we introduce the hierarchical models notation of Christensen (1997) to be used in Chapter 2.

Example 1.6 Consider the three-source capture-recapture setting. We use $\perp\!\!\!\perp$ to denote independence.

- Mutual independence. In this case $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$ so that $\alpha_{AB} = \alpha_{AC} = \alpha_{BC} = 0$. Such a model is represented by $[A][B][C]$ (Christensen, 1997) and given in Example 1.5.

- Joint independence. For example, $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$ but A and B are dependent. All terms of the saturated model of Example 1.5 are included except for the interactions corresponding to independence. That is $\alpha_{AC} = 0$ and $\alpha_{BC} = 0$. Such a model is represented by $[AB][C]$ (Christensen, 1997).

- Conditional independence. Without loss of generality, we suppose that B and C are independent, conditionally on A, that is $B \perp\!\!\!\perp C|A$. Then $\alpha_{BC} = 0$ in the expression of the saturated model of Example 1.5 but all other terms are included. Such a model is represented by $[AB][AC]$ (Christensen, 1997).

Remark 1.13 HJLMM descriptor notation.
Let $\mathcal{R}$ and $\mathcal{T}$ denote arbitrary sets of sources. Then we use $[\mathcal{R}, \mathcal{T}]$ to denote the HJLLM specification that includes all hierarchical terms of $\mathcal{R} \cup \mathcal{T}$. For example, for $\mathcal{R} = \{A, B\}$ and $\mathcal{T} = \{C\}$ with $\mathcal{R} \cup \mathcal{T} = \{A, B, C\}$, a HJLLM with the descriptors $[\mathcal{R}, \mathcal{T}]$ contains terms $\alpha, \alpha_A, \alpha_B, \alpha_C$, $\alpha_{AB}, \alpha_{AC}, \alpha_{BC}$ and α_{ABC}.

Remark 1.14 All of the theory applicable to the general complete contingency table case, including results for parameter estimation via maximum likelihood, is applicable to the incomplete contingency table (Bishop et al., 1975). As described earlier, the sole difference is that a model corresponding to K-way dependence for the K-source example is not estimable. Any HJLLM that includes the highest order term would be fully determined and all cells would be estimated perfectly equal to their observed value.

Remark 1.15 Irrespective of which interaction terms are included in the model, the single-source marginal counts are sufficient statistics for the model parameters (Bishop

et al., 1975). Under maximum likelihood estimation, estimators of the parameters are such that the marginal counts are fitted exactly equal to the observed counts. That is, $\hat{m}_A = n_A$, $\hat{m}_B = n_B$ and $\hat{m}_C = n_C$ in the three-source case. An approximate relationship of this sort will be observed for the models introduced in this dissertation, with the frequentist approach of Chapter 3 and the Bayesian approach of Chapter 5.

Whenever the dependence structure is more complex than that of independence, sufficient statistics correspond to the highest order terms in the model (Bishop et al., 1975). For the joint independence case, for example $[AB][C]$, MLEs of the model parameters are such that $\hat{m}_{AB} = n_{AB}$. For conditional independence, for example $[AB][AC]$, a similar relationship is observed: $\hat{m}_{AB} = n_{AB}$ and $\hat{m}_{AC} = n_{AC}$.

Remark 1.16 Accounting for heterogeneity-induced source dependence. To account for heterogeneity-induced source dependence, IWGDMF (1995a) propose a method which adds a homogeneous term within each level of interaction. That is, for example, the same term is used to represent all pairwise dependence rather than including a different interaction term for each pair of sources. Such an approach is useful when covariates thought to explain the dependence are not available (Stanghellini & van der Heijden, 2004).

Several alternatives have been proposed to account for unobserved heterogeneity which are essentially variations on log-linear models with latent variables. Stanghellini and van der Heijden (2004) present such a model to account for observed and unobserved heterogeneity. Since the modelling approach in this thesis will not distinguish between the two forms of dependence we will not describe such models for heterogeneity here; rather we will discuss them in the next section.

Precision estimation and confidence intervals

Asymptotic normality of the estimator $\hat{N}$ can be used to form $(1-\alpha)\%$ Wald confidence intervals of the form $\hat{N} \pm z_{\alpha/2}\widehat{\text{se}}(\hat{N})$, where $\widehat{\text{se}}(\hat{N})$ is obtained by the δ-method (Seber, 1982), using the estimated intercept-term of the HJLLM, denoted below by α. The δ-method is used to obtain the estimated variance under the assumption of an underlying Poisson likelihood

$$\widehat{\text{Var}}_P\left[\hat{N}\right] = \widehat{\text{Var}}\left[\hat{n}_{unobs}\right] \simeq \exp(2\hat{\alpha})\widehat{\text{Var}}(\widehat{\alpha})$$

so that

$$\widehat{\text{se}}_P(\widehat{N}) = \widehat{\text{se}}(\widehat{n}_{unobs}) \simeq \exp(\hat{\alpha})\widehat{\text{se}}(\widehat{\alpha}).$$

When the multinomial likelihood is assumed, the estimated variance is given by (Rivest & Lévesque, 2001)

$$\widehat{\text{Var}}_m\left[\hat{N}\right] = \exp(\hat{\alpha}) + \widehat{\text{Var}}_P\left[\hat{N}\right].$$

However, the distribution of $\hat{N}$ is typically skewed (IWGDMF, 1995a), implying that the asymptotic normality approach may give misleading results. Transformations of $\hat{N}$ may alleviate such concerns (for example the log-transformation of Borchers, Buckland, & Zucchini, 2002).

An alternative approach is that of profile likelihood (Cormack, 1992 and Regal & Hook, 1984). As noted by IWGDMF (1995a), the advantage of such an approach is that it is like working with the best possible transformation of N. Buckland (1984) developed bootstrap methods for the capture-recapture setting, as advocated by several other authors (Buckland & Garthwaite, 1991; Norris & Pollock, 1996).

We will use Wald intervals assuming an underlying Poisson likelihood in Chapter 3 for computational convenience, despite the concerns raised above related to the use of Wald confidence intervals. We consider such usage reasonable since the purpose of the analysis in that chapter will be one of model comparison.

Model selection

When working within the inferential framework of maximum likelihood it is natural to select amongst nested models using likelihood ratio tests (IWGDMFI, 1995a). As further noted by IWGDMF (1995a), information criteria can be used to select amongst nonnested models. Burnham, White, and Anderson (1995) discuss the use of several information criteria for capture-recapture data including the AIC (Sakamoto, Ishiguro, & Kitigawa, 1986) and BIC (Draper, 1995) given by

$$AIC = -2\log(L(\hat{N}, \hat{\alpha})) + 2q \tag{1.6}$$

$$BIC = -2\log(L(\hat{N}, \hat{\alpha})) + \log(n)q, \tag{1.7}$$

where q is the number of model parameters and $L(\hat{N}, \hat{\alpha})$ the likelihood evaluated at the MLEs of N and the model parameters. There is disagreement as to what n should be (IWGDMF, 1995a), with some suggesting that it should be fixed at n_{obs} (Hook & Regal, 1997). The BIC penalizes large models more heavily than the AIC, thus the BIC tends to select more parsimonious models than the AIC. Simulations by Hook and Regal (1997) suggest that the AIC is preferable in the capture-recapture setting; thus it is the model selection criterion we adopt for model selection in Chapter 3. In practice, the AIC is calculated for all models in order to select the model with the lowest value as the best model.

Model averaging

In basing inference on a single model, selected according to a criterion such as AIC, no measure of model mis-specification is incorporated into confidence intervals. Thus, inference is conditional on the correct model having been chosen (IWGDMF, 1995). Buckland, Burnham, and Augustin (1997) propose a model weighting approach to account for model uncertainty, whilst Madigan and York (1997) use a Bayesian approach. In this dissertation we will not perform model averaging; rather we will

undertake model selection in the frequentist analysis of Chapter 3.

1.2.2 Individual-level models

As described in Section 1.1.2, heterogeneity of capture probabilities (i.e different probabilities of inclusion in a list for different individuals) is termed *observable* when it can be explained by observed covariates, such as age or sex. Stratification by the covariate believed to be associated with heterogeneous capture probabilities is one approach (Bishop et al., 1975). An alternative parametric modelling approach is to consider the class of models that accommodate heterogeneity by modelling individual-level list-inclusion (or capture) probabilities. Unlike HJLLMs, which model source-level inclusion probabilities, such models stratify the population at the finest level, that of the individual.

Alho (1990) and Huggins (1989) independently developed comparable logistic regression approaches that accommodate both categorical and continuous covariates. This approach assumes that the lists operate independently at the individual level (Alho, 1990), with the work extended by Zwane and van der Heijden (2005) to allow for possible dependence between lists.

When covariates explaining heterogeneity are not available, several approaches have been proposed. The latent class approach advocated by Pledger (2000) assumes that individuals cluster into several latent classes such that all individuals within a class have the same probability of inclusion on a given list (Agresti, 1994; Coull & Agresti, 1999; Fienberg, Johnson, & Junker, 1999). The model of Coull and Agresti (1999) and Fienberg et al. (1999) is the Rasch model (Rasch, 1969), first introduced in the context of educational testing. Both Coull and Agresti (1999) and Fienberg et al. (1999) demonstrate that the Rasch model is equivalent to a log-linear model of quasi-symmetry (Bishop et al., 1975).

As stated by Bartolucci and Forcina (2006), a basic assumption of such latent

class models is that lists operate independently within homogeneous subjects, so that marginal association is due entirely to unobserved heterogeneity. Stanghellini and van der Heijden (2004) present an approach to allow for marginal association between lists via bivariate interactions between lists conditional on the latent, for a model with categorical covariates, whilst Bartolucci and Forcina (2001) adopt an alternative approach which allows separate modelling of the univariate marginals and the bivariate associations (as stated in Bartolucci & Forcina, 2006) in the presence of categorical covariates. They further extend their work on the modelling of observable and unobservable heterogeneity (Bartolucci & Forcina, 2006) to accommodate continuous covariates whilst also allowing for conditional dependence amongst lists. The authors point out that these models are only slightly more flexible than those of Bartolucci and Forcina (2001) when only discrete covariates are available.

Remark 1.17 We note that the inferential framework adopted in all references detailed in the present section is that of maximum likelihood estimation, except for the Rasch model of Fienberg et al. (1999), in which a Bayesian approach was adopted. In Section 1.4 we present additional Bayesian work.

1.3 Literature review: Nonparametric approaches to modelling CR data

In this section we consider work from the literature related to census under-count as itself related to sample coverage. The ideas presented here are required for Chapter 2. In particular, we define the Coefficient of Covariation (CCV) related to the measures of source dependence introduced in Chapter 2.

In order to assess the extent of under-count, a post-enumeration survey (PES) is undertaken on a sample of households to generate a two source capture-recapture data

set (Chao & Tsay, 1998). We note that "correlation bias" is a term commonly used to denote what is often referred to as dependence in the general capture-recapture literature. As stated in the introduction to this chapter, several countries including Canada, the UK and the USA, now routinely collect PES data to correct for under-count. See Redfern (2004) for a description of practical issues of under-count in the UK census and Mitchell, Dorling, Martin, and Simpson (2002) concerning the 1.2 million people missed from the 1991 UK census. Robinson, West, and Adlakha (2002) presents an assessment of the under-count in Census 2000 in the USA, whilst Anderson and Fienberg (2002) discuss the controversy surrounding adjustment of under-count in censuses.

Sample coverage

First we introduce the foundational work of Good (1953). He presents a method of sample coverage which has subsequently been applied to species estimation (see Bunge & Fitzpatrick, 1993, for a review) and to the correction of under-count of census estimates (Chao et al., 2001). Good describes the method as follows. Consider drawing a random sample from a population of various species. Then r/N is not a good estimate of the population frequency p of a particular species, when r, the number of times that particular species is observed, is small. Good provides methods for estimating p with very few assumptions on the underlying population. The estimate of p is expressed in terms of n_r $(r = 1, 2, 3, \dots)$, where n_r is the number of distinct species that are observed r times in the sample. An estimate of the proportion of the species occurring in the sample can be obtained directly.

Chao, Lee, and Jeng (1992) describe how the relationship between sample coverage and population size has been used to obtain estimates of N from capture-recapture data for animal populations. However, as noted by Chao and Tsay (1998), the animal population methods of Chao et al. (1992) cannot be applied directly to census

under-count, partly because there are typically only two or three samples in a census under-count setting (as opposed to considerably more, usually, in animal abundance estimation). Additionally, there is no time-ordering to the lists (as stated in the assumptions in Section 1.1).

Therefore, Chao and Tsay (1998) develop nonparametric methods for census under-count correction. They introduce the CCV and its relationship to sample coverage. We proceed with the description here based directly on that of Chao and Tsay (1998), with the notation and terminology adapted to that used in this dissertation.

For two sources A and B, the sample coverage of the two sources is defined as (Chao & Tsay, 1998)

$$C = \frac{1}{2}\left[\frac{\sum_i \mathbb{E}\left[X_{iB}|X_{iA}\right]I[X_{iA}>0]}{\sum_i \mathbb{E}\left[X_{iB}|X_{iA}\right]} + \frac{\sum_i \mathbb{E}\left[X_{iA}|X_{iB}\right]I[X_{iB}>0]}{\sum_i \mathbb{E}\left[X_{iA}|X_{iB}\right]}\right], \tag{1.8}$$

where

$$X_{iS} = I[\text{the } i\text{th individual is listed in source S}],\ S = A, B,$$

and $I[.]$ is the usual indicator function. When dependence is present, Chao and Tsay (1998) note that it is difficult to count the number unobserved directly but that the sample coverage can be well estimated. An estimator is given by

$$\hat{C} = \frac{1}{2}\left(\frac{n_{AB}}{n_A} + \frac{n_{AB}}{n_B}\right). \tag{1.9}$$

When no dependence is present, Chao and Tsay (1998) demonstrate that

$$\hat{N} = \frac{n_A + n_B}{2\hat{C}},$$

where $\hat{C} = n_{AB}(1/n_A + 1/n_B)/2$, which leads to the Petersen (1.2) estimator $\hat{N} = n_A n_B / n_{AB}$.

Coefficient of covariation

Consider the K-source setting to estimate the true (unknown) population size N. Let the data for each of the N individuals be included in an $N \times K$ matrix $\mathbf{X} = (X_{ij})$

where

$$X_{ij} = I[\text{the } i\text{th individual is listed in source } S_j],$$

and $I[.]$ is the usual indicator function. Assume that individuals act independently and define the average inclusion probability for source S_j as $\mu_{S_j} = (1/N)\sum_{i=1}^{N} E(X_{ij})$. The definition of the coefficient of covariation (CCV) of sources S_j and S_k is given by

$$\omega_{S_jS_k} = \frac{1}{N}\sum_{i=1}^{N} E[(X_{ij} - \mu_{S_j})(X_{ik} - \mu_{S_k})]/(\mu_{S_j}\mu_{S_k}). \tag{1.10}$$

The CCV measures the degree of dependence between sources and is equal to 0 in the case of independent sources. It is defined for the general K-source case, for sources $S_1, \ldots, S_K$, as follows:

$$\omega_{S_1,\ldots,S_K} = \frac{1}{N}\sum_{i=1}^{N} E[(X_{iS_1} - \mu_{S_1})(X_{iS_2} - \mu_{S_2})\ldots(X_{iS_K} - \mu_{S_K})]/(\mu_{S_1}\ldots\mu_{S_K}).$$

When capture probabilities are homogeneous (i.e. do not differ by individuals), Chao and Tsay (1998) show that (1.10) simplifies to

$$\omega_{S_jS_k} = \frac{p_{S_jS_k}}{p_{S_j}p_{S_k}} - 1, \tag{1.11}$$

where $p_{S_jS_j}$ is the probability of appearing in sources S_j and S_k, likewise p_{S_j} and p_{S_k} are the marginal probabilities of appearing in sources S_j and S_k, respectively. For three sources

$$\omega_{S_jS_kS_m} = \frac{p_{S_jS_kS_m}}{p_{S_j}p_{S_k}p_{S_m}} - (\omega_{S_jS_k} + \omega_{S_jS_m} + \omega_{S_kS_m}) - 1. \tag{1.12}$$

Furthermore, they state that analogous expressions for higher order CCVs can be derived. In Chapter 2 we will state an explicit relationship between the CID for two and three sources and the CCV for two and three sources given by (1.11) and (1.12), respectively.

Estimation of N

In the two-source case, Chao and Tsay (1998) derive an expression relating the population size N to the CCV. For clarity, let the two sources be denoted by A and B. Then the relationship between N and ω_{AB} is given by

$$N = \frac{\mathbb{E}[n_A]\mathbb{E}[n_B]}{\mathbb{E}[n_{AB}]}(1+\omega_{AB}).$$

As further noted by the authors, ω_{AB} must be estimated. However, there are insufficient degrees of freedom to test whether ω_{AB} is different from 0. Thus, if there is dependence, such an approach cannot be used in practice (unless a Bayesian approach is used and a prior distribution placed on ω_{AB}, an approach not adopted by Chao & Tsay, 1998). Such relationships become useful for three or more sources (see Chao & Tsay, 1998 for a discussion), as is the case with the new modelling approach presented in this dissertation.

1.4 Literature review: Bayesian approaches to modelling CR data

Much of the Bayesian capture-recapture literature is based on two classes of models described in Section 1.2, namely HJLLMs and the Rasch latent class model, and variations thereon. Here we provide an overview of the features of the different approaches presented in the literature.

Both Madigan and York (1997) and King and Brooks (2001a) describe Bayesian approaches for HJLLMs. For the former, a three-source data set was analyzed by fitting all 7 HJLLMs with inference based on an average over all models according to the posterior probability of each model. King and Brooks (2001a) combine the work of Madigan and York (1997) and Dellaportas and Forster (1999) by using reversible jump MCMC (Green, 1995) to move between all models in the class of HJLLMs.

Casella and George (1992) use Gibbs sampling to calculate Bayes estimates for a simplified version of the Rasch model, in which capture probabilities are assumed to be homogeneous. Fienberg et al. (1999) extend that work to the full Rasch model.

As stated by Fienberg et al. (1999), prior to the use of MCMC methods (see below for a description) for capture-recapture data, several authors developed Bayesian approaches with minimal computational challenges. Roberts (1967) dealt with the two-source setting with homogeneous capture probabilities. Castledine (1981) extended the approach to multiple captures, again with homogeneous capture probabilities to obtain the marginal posterior of N under a specific prior specification. In the same setting, Smith (1991) used empirical Bayes approaches to obtain the posterior of N. Garthwaite, Yu, and Hope (1995) examined the sensitivity of the posterior of N to different priors on N and showed that the number of captures in each sample (equivalently list) typically provides little information about N. Recent work by Wang, He, and Sun (2007) discusses the difficulties of obtaining noninformative priors for a Bayesian capture-recapture model.

1.4.1 Overview of the Bayesian paradigm and MCMC methods

In this section we provide a brief overview of the Bayesian paradigm and associated computational techniques. A comprehensive description of Bayesian data analysis can be found in Gelman, Carlin, Stern, and Rubin (2004) and Gilks, Richardson, and Spiegelhalter (1996). The description given here is based on a similar section of the author's work in Turner (2002), with notation similar to that used by Gelman et al. (2004). The interested reader can consult the text for a more complete description.

Bayes' theorem

We consider Bayes' theorem in its simplest form for a single scalar parameter. For

ease of exposition, we will assume that all random variables are continuous. The theory applies equally to discrete random variables. Let θ be the parameter to be estimated and y (which depends on θ) the data observed. Then

$$\pi(\theta|y) = \frac{f(\theta, y)}{f(y)} = \frac{\pi(\theta)f(y|\theta)}{f(y)} = \frac{\pi(\theta)L(\theta|y)}{f(y)} \tag{1.13}$$

where $\pi(\theta|y)$ is known as the posterior density of θ, $\pi(\theta)$ the prior density of θ, $f(y|\theta)$ is the conditional density function of y, $L(y|\theta)$ is the likelihood of θ given the observed y and $f(y) = \int \pi(\theta)f(y|\theta)\, d\theta$ is the integrated likelihood. Since $f(y)$ does not depend on θ it can be considered a constant for fixed y. Hence, (1.13) is equivalent to

$$\pi(\theta|y) \propto \pi(\theta)L(\theta|y) \tag{1.14}$$

These expressions form the basis of Bayesian inference. Of course, the same relationship exists for multivariate data, $\mathbf{Y}$, and a multi-dimensional parameter, $\boldsymbol{\theta}$. Notice that the posterior density of θ is composed of information from the data y, in the form of the likelihood, and an *a priori* distribution for θ. It is a compromise between information from these two sources Consequently, if there is very little known about θ *a priori*, a noninformative prior distribution should be accorded to θ. (We note that the terms vague, flat and diffuse are used somewhat similarly, depending on the context. See Gelman et al., 2004.) Such a noninformative prior on θ will play a minor role in forming the posterior distribution of θ via (1.13). Consequently, the posterior distribution will be largely determined by the observed data.

Conjugacy

A family of prior distributions is known as a conjugate family for data/samples from a particular distribution (i.e. with a particular likelihood) if the posterior distribution is in the same family as the prior. We consider a simple example as illustration. Let $\mathbf{X}$ be a random sample of size n from $Bernoulli(\theta)$. Then $Y = \sum_{i=1}^{n} X_i \sim Bin(n, \theta)$. Suppose we observe $Y = y, (y = 0, 1, \ldots, n)$. Furthermore, suppose that

$\theta \sim Beta(\alpha, \beta)$, where α and β are assumed to be known. Then, by (1.13)

$$\begin{aligned}\pi(\theta|y) &= \frac{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int_0^1 \pi(\theta)f(y|\theta)\,d\theta} \\ &= \frac{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\frac{B(y+\alpha,n-y+\beta)}{B(\alpha,\beta)}} \\ &= \frac{\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{B(y+\alpha,n-y+\beta)} \qquad (1.15)\end{aligned}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Therefore, $\theta|y \sim Beta(y+\alpha, n-y+\beta)$. In this case the posterior distribution of θ is a Beta distribution just as the prior was, indicating that the Beta family is the conjugate family for data/samples from the Binomial distribution (i.e. for the binomial likelihood). Moreover, this example demonstrates one of the important properties of conjugacy stated next.

Remark 1.18 Adopting a conjugate prior, whether it be uni-dimensional, as in the previous example, or multi-dimensional, often aids in computation and in simulation procedures that are based on the Bayesian paradigm. It is a property that will be adopted in Chapters 4 and 5.

Hierarchical prior structure

When the parameters of the prior distribution of $\pi(\theta)$, termed the hyperparameters, are unknown, it is necessary to place a distribution on them. In such a way it is possible to generate the posterior distribution of θ. Let ϕ denote the hyperparameters of the prior distribution of θ denoted by $\pi(\theta|\phi)$, where the distribution of ϕ is denoted by $f(\phi)$. Then the joint prior distribution is $\pi(\theta,\phi) = \pi(\theta|\phi)f(\phi)$. The likelihood is now $L(\theta,\phi|y) = f(y|\theta,\phi)$. But the sampling distribution of y is assumed to be independent of the hyperparameters, ϕ. Therefore, $L(\theta,\phi|y) = f(y|\theta,\phi) = f(y|\theta) =$

$L(\theta|y)$. Hence, using (1.14), the joint posterior distribution is

$$\begin{aligned}\pi(\theta,\phi|y) &\propto f(\phi)\pi(\theta|\phi)f(y|\theta,\phi)\\ &= f(\phi)\pi(\theta|\phi)f(y|\theta)\\ &= \pi(\theta,\phi)f(y|\theta)\\ &= \pi(\theta,\phi)L(\theta|y) \qquad (1.16)\end{aligned}$$

A clear example of this can be found in Chapter 5 of Gelman et al. (2004). We recall that θ is the parameter (or vector of parameters) of interest. A joint posterior distribution is generated using the hierarchical model (1.16). The conditional posterior density of θ, given the hyperparameter(s), ϕ, can then be determined analytically. Moreover, the marginal distribution of ϕ can be determined by integrating out θ.

MCMC computational methods

There are many challenges to estimation of the components of a Bayesian model. Thus, simulation techniques are often used. The difficulties may arise because the integration necessary to determine the densities of (1.16) is too complex or because the densities may not even exist in a closed form that can be calculated analytically. Here we will provide a brief introduction to some standard Bayesian computational methods, in particular Markov chain Monte Carlo methods.

The methods adopted to determine the posterior distribution of the parameter(s) of interest (i.e. θ or both θ and ϕ in the section above) are based on the simulation of a random walk which converges to an equilibrium distribution corresponding to the posterior distribution, $\pi(\theta|y)$, with the properties of irreducibility and ergodicity. Detailed balance ensures that there is time irreversibility (see Chapter 1 of Gilks et al., 1996). [1] The random walk is generated by successively sampling values from

[1]Note that we will adopt the notation $\pi(\theta|y)$ throughout. It will be assumed that θ could be a

the marginal posterior conditional distributions (e.g. the marginal posterior conditional distribution of θ_1, $f(\theta_1|\theta_2,\ldots,\theta_m,y)$). These distributions can be determined by examining the joint distribution $f(\theta,y)=\pi(\theta)L(\theta|y)$ in the expression of the hierarchical model (1.16). The simulated values form a chain which is Markov, since each draw depends only on the previous one and is independent of all those that came before. Depending on the nature of the target distribution (i.e. that of θ), various methods exist to generate such a chain. Here we will discuss two of them which will be used in Chapters 4 and 5. They are the Gibbs Sampler and the Metropolis algorithm. We note that the Gibbs Sampler is a special case of the Metropolis algorithm, which in turn is a special case of the Metropolis-Hastings algorithm.

The Gibbs sampler

Casella and George (1992) provide a thorough and relatively simple introduction to the Gibbs Sampler, as do Gelman et al. (2004). We suppose that our parameter of interest, θ, is an m-dimensional vector (recall that this could include the hyperparameter(s) ϕ). Consider the model in (1.14) or equivalently in (1.16). We are interested in determining the joint posterior distribution, $\pi(\theta|y)$. In order to do so we will sample from the full conditional distributions, which are of the form $\pi(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_m,y)$, $i=1,\ldots,m$. These distributions can be determined by examining the joint distribution $f(\theta,y)$ should the prior/likeihood combination by conjugate. An initial value of θ, $(\theta_1^{(0)},\ldots,\theta_m^{(0)})$ is chosen (see Gelman et al., 2004 for a discussion). The Gibbs sampler algorithm then samples successively from the full conditional distributions according to the following iteration scheme:

(i) sample $\theta_1^{(i+1)}$ from $\pi(\theta_1|\theta_2^{(i)},\ldots,\theta_m^{(i)},y)$

(ii) sample $\theta_2^{(i+1)}$ from $\pi(\theta_2|\theta_1^{(i+1)},\theta_3^{(i)},\ldots,\theta_m^{(i)},y)$

vector and could include the hyperparameters of the hierarchical model (1.16). Therefore, in this discussion $\pi(\theta|y)$ will also be used to represent $\pi(\theta,\phi|y)$.

.....

(iii) sample $\theta_m^{(i+1)}$ from $\pi(\theta_m|\theta_1^{(i+1)},\ldots,\theta_{m-1}^{(i+1)},y)$.

At each run through the loop a vector $\theta^{(i)}$, $i=1,2,\ldots$ is generated. The sequence of vectors, $\theta^{(1)},\theta^{(2)},\ldots$, form a Markov chain since each realization of the vector, $\theta^{(i)}$, depends only on the value of the vector in the previous iteration, $\theta^{(i-1)}$. Gelman et al. (2004) demonstrate that the chain converges towards the joint posterior distribution that we are interested in, $\pi(\theta|y)$. Both Casella and George (1992) and Gelman et al. (2004) discuss how to determine whether the chain has converged. Casella and George (1992) discuss how the generated sequence of vectors, $\theta^{(1)},\theta^{(2)},\ldots$, is used to determine properties of the posterior distribution $\pi(\theta|y)$ by using the fact that $\theta^{(1)},\theta^{(2)},\ldots$, is a sample from $\pi(\theta|y)$, or that after some run-in period, r ($r\geq 1$) iterations say, the sequence $\theta^{(r+1)},\theta^{(r+2)},\ldots$, is considered a sample from $\pi(\theta|y)$. For instance, the mean and variance of $\pi(\theta|y)$ can be determined, just as can any percentiles of the distribution. We are, of course, ultimately interested in determining the unconditional marginal distributions, $f(\theta_i)$, $i=1,\ldots,m$. Casella and George (1992) indicate that the quantities $\pi(\theta_i|\theta_1^{(j)},\ldots,\theta_{i-1}^{(j)},\theta_{i+1}^{(j)},\ldots,\theta_m^{(j)},y)$, $i=1,\ldots,m, j=1,2,\ldots$, calculated using the simulated values $\theta_1^{(j)},\ldots,\theta_{i-1}^{(j)},\theta_{i+1}^{(j)},\ldots,\theta_m^{(j)}$ at the jth iteration contain more information about $f(\theta_i)$ than the simulated values, $\theta_i^{(0)},\theta_i^{(1)},\theta_i^{(2)},\ldots$, themselves. For instance, the mean of $f(\theta_i)$ can be estimated by $(1/n)\sum_{j=0}^{n-1}\theta_i^{(j)}$. In the limit this expression tends to the true mean of $f(\theta_i)$. However, a better estimate uses the following: $(1/n)\sum_{j=0}^{n-1}\mathbb{E}\left[\theta_i|\theta_1^{(j)},\ldots,\theta_{i-1}^{(j)},\theta_{i+1}^{(j)},\ldots,\theta_m^{(j)},y\right]$ if these conditional expectations can be evaluated.

The Metropolis algorithm

The Gibbs Sampler is a special case of the Metropolis Algorithm. Recall that at each stage of sampling in the Gibbs Sampler there was no decision taken as to whether or not to accept the updated vector-value of the parameter. The Metropolis Algorithm

includes, at each iteration, a decision process as to whether or not to accept the updated parameter value. Both Tanner (1996) and Gelman et al. (2004) indicate how this decision is made. Suppose that the vector $\theta^{(*)}$ is the proposed updated vector at the ith iteration. If $\theta^{(*)}$ increases the posterior density then it is accepted. If it decreases the posterior density then it is only kept (*accepted*) with a certain probability.

We will provide an outline of this scheme which uses the same notation as Gelman et al. (2004). We will include all of the main ideas but it will be necessary to consult the text for a full description. As with the Gibbs sampler we choose a starting value of θ from a *starting distribution* and name this $\theta^{(0)}$. Then proceed with the following:

- Sample a *candidate point* $\theta^{(*)}$ from a symmetric *jumping distribution* at time t, $J(\theta^{(*)}|\theta^{(t-1)})$.
- Calculate the ratio of the conditional density at the proposed point and at the previous point in the chain

$$r = \frac{\pi(\theta^{(*)}|y)}{\pi(\theta^{(t-1)}|y)}. \tag{1.17}$$

- Set

$$\theta^{(t)} = \begin{cases} \theta^{(*)}, & \text{with probability min(r,1)} \\ \theta^{(t-1)}, & \text{otherwise.} \end{cases}$$

This algorithm generates a Markov chain. In order to generate the posterior density of θ, the Markov chain is used in the same way as that generated by the Gibbs sampler.

The Metropolis Algorithm involves a so-called *rejection/acceptance* component. However, the Gibbs Sampler will always accept the proposed parameter vector $\theta^{(*)}$, even if it decreases the posterior density. Both methods will accept $\theta^{(*)}$ with probability 1 if it increases the posterior density. It is important to note that the power of these methods, as indicated by Gelman et al. (2004), is not the Markov property but

the fact that the distributions are essentially improved at each iteration. Forming a Markov chain is desirable for ease of proving the convergence of the chain and other such properties.

Remark 1.19 The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm, for which the jumping rules J_t are not necessarily symmetric, as is the case with the Metropolis algorithm presented here. Since we will use only symmetric proposal distributions in Chapters 4 and 5, we will not discuss the more general Metropolis-Hastings algorithm here.

Convergence and Run-In Period. It is natural to expect that the choice of starting values will influence the number of iterations required to achieve convergence of the Markov Chain. Moreover, the question of how to determine whether convergence has been achieved is, of course, one of great importance. An issue linked to this question is how great a 'run-in' period to allow. All of these issues, particularly that of convergence, have been examined extensively by many authors. Both Gelman et al. (2004) and Brooks (1998) address all of these questions in some detail.

It is recommended to run not only one sequence of iterations but several simultaneously, each of which begins with different starting values. Moreover, the use of starting values from overdispersed distributions is advocated (Gelman et al., 2004). One can then compare within sequence variation to between sequence variation. Initially, between sequence variation will be greater than within sequence variation. As the chain approaches the equilibrium distribution the two variations will become increasingly similar. This fact has been exploited in developing a statistic that is used widely to determine convergence. It is described in some detail by Gelman et al. (2004) and by Brooks and Gelman (1998). The statistic is known as the Gelman-Rubin convergence statistic, as indicated in Gelman et al. (2004). We will not describe the statistic in detail here; suffice to say that it is a statistic calculated for each parameter

that is being estimated by the algorithm. In order to conclude that convergence has been achieved, the value of the statistic should be approximately equal to 1 for each parameter being estimated.

1.5 Literature review: General parametric modelling techniques

In this section we describe modelling techniques that enable us to place the work presented in this dissertation relative to developments broader than those only in the capture-recapture setting.

Generalized linear mixed models

McCulloch and Searle (2001) introduce the class of generalized linear mixed models (GLMM), building on the work of McCullagh and Nelder (1999), who present the class of generalized linear models (GLM). We let $\mathbf{y}$ be a vector of response variables and g a link function (not necessarily linear) which operates on each element of $\mathbf{y}$. The GLM is defined by

$$g(\mathbb{E}[\mathbf{y}]) = \mathbf{X}\boldsymbol{\beta} \tag{1.18}$$

extended to the definition of the GLMM, given by

$$g(\mathbb{E}[\mathbf{y}|\mathbf{u}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \text{ where } \mathbf{u} \sim f_U(\mathbf{u}), \tag{1.19}$$

where, in each case, $\mathbf{y}$ is taken from an exponential family and $\mathbf{X}$ is the design matrix for the fixed effects $\boldsymbol{\beta}$. In the case of the GLMM, $\mathbf{Z}$ is the design matrix for the random effects $\mathbf{u}$, which are assumed to be distributed according to the distribution $f_U(\mathbf{u})$. It is usually assumed that the response vector $\mathbf{y}$ consists of conditionally

independent elements, each with a distribution from the exponential family so that

$$y_i|\mathbf{u} \sim \text{ indep } f_{Y_i|\mathbf{u}}(y_u|\mathbf{u})$$
$$f_{Y_i|\mathbf{u}}(y_u|\mathbf{u}) = \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i, \tau)\}.$$

See McCulloch and Searle (2001) for further details.

We note that by adding random effects to the GLM, it is possible to incorporate correlation and to undertake broader inference since a greater range of models fit into such a class (McCulloch & Searle, 2001). For example, the analysis of longitudinal data necessitates the incorporation of correlation between observations on the same individuals (Diggle, Heagerty, Liang, & Zeger, 2002). The model we introduce in Chapter 4 is a further generalization of the GLMM, in which the link function g takes linear combinations of the entries of $\mathbb{E}[\mathbf{y}|\mathbf{u}]$, equivalently, the model is of the form $\mathbf{A}(\mathbb{E}[\mathbf{y}|\mathbf{u}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, where $\mathbf{A}$ is an invertible matrix. Zhao, Staudenmayer, Coull, and Wand (2006) discuss a general framework for Bayesian GLMMs, in which it is understood that the link function g operates element-wise on the vector $\mathbb{E}[\mathbf{y}|\mathbf{u}]$. The model introduced in Chapter 4 of this dissertation is a Bayesian model of a similar form. However, there are additional challenges faced in our setting in which the link function operates on linear combinations of the entries of $\mathbb{E}[\mathbf{y}|\mathbf{u}]$, rather than the simpler element-wise operation of g in (1.19).

Marginal models for categorical data

The new modelling approach introduced in this dissertation is a model on marginal means of a contingency table (in particular the incomplete contingency table of the capture-recapture setting). In such a case both models (1.18) and (1.19) assume that the elements of $\mathbf{y}$ are distributed as independent Poisson random variables or that $\mathbf{y}$ has a multinomial distribution with N as the population size. Here we present a description of general marginal models for categorical data. Such models typically

fit into the class of GLMMs as described above, with g not necessarily a function on each element of $\mathbf{y}$, but one which might take linear combinations of those elements.

Haber (1985b) describes maximum likelihood methods for log-linear models for categorical data (thought of as arranged in a contingency table), with the most general model given by

$$\mathbf{C} \log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{1.20}$$

where $\mathbf{C}$ is a matrix which serves to form linear combinations of the entries of $\boldsymbol{\mu}$, the cell means of the contingency table, corresponding to the vector of cell means $\boldsymbol{\mu}_{cell}$ introduced in Section 1.1. Haber and Brown (1986) further extend the method to the case where expected frequencies are subject to linear constraints.

Lang and Agresti (1994) build on this work to develop a more general model form to simultaneously model the joint and marginal distribution of multivariate categorical responses. The model takes the form

$$\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \ \text{ident}(\boldsymbol{\mu}) = \mathbf{0} \tag{1.21}$$

where $\mathbf{C} = \mathbf{C}_J \bigoplus \mathbf{C}_M$, $\mathbf{B}' = (\mathbf{A}'_J, \mathbf{A}'_M)$, $\mathbf{X} = \mathbf{X}_J \bigoplus \mathbf{X}_M$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_J, \boldsymbol{\beta}'_M)$, and $\text{ident}(\boldsymbol{\mu}) = \mathbf{0}$ denotes the multinomial identifiability constraints, and J refers to the model on the joint means and M to that on the marginal means. We note that (1.20) is a special case of (1.21), with $\boldsymbol{\alpha}$ composed only of parameters related to the joint cell means of the contingency table and not the marginal means.

The marginal model we introduce in Chapter 2, which we treat via maximum likelihood estimation in Chapter 3, is of the form given by (1.21), in which $\boldsymbol{\alpha}$ is composed only of entries related to the marginal distribution and not the joint distribution, as is the case with (1.20). The matrix $\mathbf{A}$ is also composed of entries relating to the marginal distribution and not the joint distribution. It is an upper triangular matrix given by (1.1) for the three-source case. Details of variance estimation of the model parameters of Chapter 3 are given in Appendix F, in particular for the model of that

chapter, derived according to the results of Lang and Agresti (1994).

Lang, McDonald, and Smith (1999) present a class of association-marginal models for multivariate categorical data of a form similar to (1.21), where the model is given by

$$\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

We note that this form is more general than the specifications of (1.21). The authors describe a maximum likelihood approach to model fitting. Molenberghs and Lesaffre (1999) describe a method in which the joint distribution of $\boldsymbol{\mu}$ is expressed in terms of the marginal mean functions and pairwise and higher order association measures.

A general theoretical framework is presented by Bergsma and Rudas (2002) who introduce a general definition of marginal log-linear parameters. They describe conditions under which the model parameters are smooth and variation independent, and conditions under which large-sample theory applies.

Marginal models for capture-recapture data

As described above in Section 1.2.2, we have identified two references in which the authors develop a marginal model specific to the capture-recapture setting (Bartolucci & Forcina, 2002 and Bartolucci & Forcina, 2006). In both instances the marginal modelling does not extend to the modelling of complete marginal association. The marginal modelling approach introduced in this dissertation considers the complete marginal distribution of all orders including the highest K-way for the K-source setting.

Remark 1.20 We note that we have found no reference to a Bayesian approach to marginal modelling and in particular no such reference to Bayesian marginal modelling of capture-recapture data. Thus, the work presented in Chapters 4 and 5 is a new approach to combining both components.

1.6 Outline of dissertation

This dissertation presents a marginal log-linear modelling approach able to model arbitrary dependence structures, which are a larger class than those modelled by hierarchical log-linear models. In so doing we provide a new approach to conceptualizing dependence in capture-recapture data.

Chapter 2

A measure of dependence for capture-recapture data named the Coefficient of Incremental Dependence (CID) is introduced, related to the Coefficient of Source Dependence (CSD) of Vandal et al. (2005). Both measures are defined for all possible $2^K - 1$ combinations of sources for the K-source case. Properties of both measures are derived. The CID is related to the CSD in that it measures changes in dependence due to moving from margins of lower dimension to those of higher dimension. On the other hand the CSDs measure dependence on a more absolute scale.

These measures form the basis for a new class of marginal log-linear models (MLLM). Unlike hierarchical joint log-linear models (HJLLM), MLLMs in their most general form are able to accommodate dependence structures that are non-hierarchial and not necessarily of a conditional independence form. Moreover, MLLMs may provide an indication of whether K-way dependence is present for which it is not possible to test using JLL models, even nonhierarchical ones. We derive the class of models and examine its relationship to the conditional independence structures modelled by HJLLMs. We see that for the HJLLM of independence and joint dependence there is an equivalent MLLM, whilst the MLLM equivalent to the HJLLM of conditional independence is a constrained model.

Chapter 3

For the new marginal log-linear model formulated in terms of the CIDs, we present a maximum likelihood approach to parameter estimation. The goal of this chapter is to further understand the relationship between MLLMs and HJLLMs. As such, a robust method will not be presented. Rather, we will present specific examples of data sets, both real and simulated in order to demonstrate that the MLLM performs well. In some cases, in particular for a nonhierarchical dependence structure, we observe that the MLLM out-performs the best HJLLM.

Chapter 4

An alternative approach to the parametric models of Chapter 3 is presented in this chapter. The dependence structure of the incomplete contingency table is modelled using random effects. A general form of the model is presented together with the development of a general Bayesian framework. Such a model fits into the class of generalized linear mixed models described in Section 1.5. This approach, in working with a model on the marginal means, is new in the field of capture-recapture, except for a single related model of the form given by Bartolucci and Forcina (2001) and Bartolucci and Forcina (2006). An MCMC scheme for parameter estimation subject to constraints is discussed.

Chapter 5

In this chapter we present a specific form of the random effects model introduced in Chapter 4 parameterized in terms of the CIDs (as introduced in Chapter 2). The CIDs are treated as random effects, which differs to Chapter 3, in which the CIDs were treated as fixed effects. We describe the specific details of the MCMC scheme introduced in Chapter 4 and present results from an analysis of the real data set

analyzed via maximum likelihood in Chapter 3. We explore the sensitivity of posterior inference to the prior specification on N and the random effects variance.

Chapter 6

In this chapter we present an overview of the dissertation and an indication of the original work developed therein.

Chapter 2

Two Measures of Source Dependence

2.1 Introduction and overview of source dependence

Statistical modelling of capture-recapture data must account for possible dependence between sources. As discussed in Section 1.1.1, there is rarely full independence between sources. For this reason, the concept of source dependence, and approaches to modelling it in order to estimate the true unknown population size N, is the central theme of this dissertation.

In this chapter we present a new approach to understanding source dependence. This is done by the introduction of a new measure of dependence named the Coefficient of Incremental Dependence (CID) derived from the Coefficient of Source Dependence (CSD) of Vandal et al. (2005). Both measures are defined for all possi-

ble marginal combinations of sources. That is, for the general K-source case there are K single-source margins, $\binom{K}{2}$ two-source margins, all the way up to the single K-way marginal combination of sources. These measures form the basis of a new marginal modelling approach, which will also be introduced in this chapter.

Both the CID and CSD are defined as ratios of population-level probabilities of source membership. Of course, the true values of the CIDs and CSDs are unknown. This is a direct consequence of the nature of capture-recapture data: the population size, N, is unknown and is to be estimated and therefore the true underlying dependence structure is also unknown. Correctly modelling the true underlying dependence structure provides the means to estimate N. To this end, two parameterizations of a new marginal modelling approach will be presented, the first based on the CIDs and the second on the CSDs. For the purposes of this introductory discussion, the two parameterizations are essentially interchangeable. Whenever we discuss them we will refer to CIDs and write CSD in parentheses.

Further motivation for this new marginal modelling approach is related to what we view as the current restrictive practices of modelling dependence for capture-recapture data. As described in Section 1.2, hierarchical joint log-linear models (HJLLMs) are believed to be the most widely adopted class of models used to analyze epidemiological capture-recapture data. They are restrictive in their inability to model non-hierarchical dependence structures. Rather, HJLLMs model conditional independence structures, which are hierarchical in nature. In this chapter we will demonstrate that non-hierarchical dependence structures may arise in practice, by providing a simple example. Nonetheless, the use of HJLLMs is advocated by many authors (Bishop et al., 1975; see description in Section 1.2.1 for more advocates), in part because of the interpretability of model parameters (see Section 1.2.1) and their relationship to graphical models.

Few authors choose to fit non-hierarchical joint log-linear models (non-HJLLMs) in

addition to HJLLMs and then select the best model amongst all joint log-linear models (JLLMs). Indeed, we know of only one case in the literature, that of Ismail et al. (2000). The major drawback of non-HJLLMs, and the one raised by most advocates of HJLLMs, is lack of interpretability of model parameters. But, as described above, in excluding non-HJLLMs there are dependence structures which cannot be modelled well. Thus, there is a need for interpretable, alternative models of dependence in order to well estimate N when the true underlying dependence structure is not well modelled by hierarchical models. The CIDs (CSDs) provide a way in which to interpret model parameters of non-HJLLM. Furthermore, the marginal modelling presented in this thesis (whose model parameters are interpretable in terms of the CIDs (CSDs)), provide a universe of dependence structures complementary to those that can be modelled by JLLMs.

This chapter is organized as follows. In Section 2.2 the definition of the CSD is presented. Properties of the CSD are provided, including explicit results stating the specific form of the CSD for the conditional dependence structures that can be modelled by the standard approach of HJLLMs. (Examples are presented later in Section 2.4.2 for the three and four-source cases). In so doing the restrictive nature of HJLLMs is demonstrated. Further properties of the CSDs are presented including statements concerning bounds and the rate of change of CSDs with changes in N. Next an explicit relationship of the CSD to the Coefficient of Covariation (CCV) measure of Chao and Tsay (1998) is presented. This measure is closely related to that of our CID which is defined in Section 2.3. Properties of the CID are presented in Section 2.3.3. Specifically, we state and prove the general K-source one-to-one relationship between the CIDs and CSDs.

In Section 2.4, we introduce the marginal log-linear model (MLLM), using early parts of this chapter as motivation. The CID and CSD parameterizations are presented along with specific relationships to the conditional dependence structures mod-

elled by HJLLMs. We do so by simultaneously deriving the explicit form of the CIDs and the CSDs (from Section 2.2) for such dependence structures. It is shown that marginal models with some CIDs (CSDs) fixed at 0 are equivalent to simple dependence structures modelled by HJLLMs (independence and joint independence) but that the marginal model equivalent to the HJLLM for conditional independence is a constrained MLLM (in practice Lagrange multipliers could be used to enforce the constraints). Thus, there is no unconstrained MLLM equivalent to the HJLLM for conditional independence. Examples are presented for the three and four-source cases.

Remark 2.1 Note that in this and in subsequent chapters, we will not distinguish between heterogeneity-induced source dependence (described in Section 1.1.1 and more specifically in Section 1.1.2) and pure source dependence. The reasons are two-fold: first, there is more than enough to say with respect to dependence in general in the new framework presented in this dissertation and, second, covariate information is not always readily available to enable more sophisticated modelling which might account for some dependence via the inclusion of covariate information. Such work will be undertaken in the future.

2.2 Coefficients of source dependence

In this section we present the definition of the Coefficient of Source Dependence (CSD) as first introduced in Vandal et al. (2005) and further described in Melocco (2002).

2.2.1 Definition

Consider a source denoted by S. Let $P[S]$ denote the probability that a randomly chosen individual in the population appears in source S; such a probability is a property of the source rather than of the individuals in the population under study

and represents the average probability of inclusion in source S, averaged over all individuals in the population. A natural measure of dependence for a set of sources denoted by $\mathcal{Q}$ is one which measures departures from independence, given by:

Definition 2.1 $c_{\mathcal{Q}} = \dfrac{P\left[\cap_{S\in\mathcal{Q}} S\right]}{\prod_{S\in\mathcal{Q}} P\left[S\right]}$,

which represents the ratio of the joint probability of membership in all sources divided by the joint probability under independence. The Coefficient of Source Dependence (CSD) for the set of sources in $\mathcal{Q}$, denoted by $C_{\mathcal{Q}}$, is defined as the natural logarithm of $c_{\mathcal{Q}}$.

Definition 2.2 $C_{\mathcal{Q}} = \log\left(c_{\mathcal{Q}}\right) = \log\left(\dfrac{P\left[\cap_{S\in\mathcal{Q}} S\right]}{\prod_{S\in\mathcal{Q}} P\left[S\right]}\right)$.

Remark 2.2 Consider $\mathcal{Q}$ composed of K sources and given by $\mathcal{Q} = \{S_1, \ldots, S_K\}$. Then, $C_{\mathcal{Q}} = C_{\{S_1,\ldots,S_K\}}$. To lighten notation, let $C_{S_1,\ldots,S_K} = C_{\{S_1,\ldots,S_K\}}$.

Remark 2.3 From definition 2.2, the single-source CSD corresponding to source S is identically zero. In this case $\mathcal{Q} = \{S\}$, and $C_{\mathcal{Q}} = C_{\{S\}}$ which we denote by C_S (see the previous remark). Then $C_S = 0$, which is referred to as the single-source CSD corresponding to source S. Further, we define the intersection of sources in the empty set to be the whole population and the empty product to be equal to 1, so that $C_{\varnothing} = 0$.

The motivation for the development of the CSD measure is twofold. First, as will be seen in Section 2.4, the definition of the CSD (and that of the related CID) naturally leads to a marginal modelling approach for capture-recapture data (hence the motivation to define the CSD according to Definition 2.2 rather than Definition 2.1). Secondly, it is a useful tool in its own right, complete with interesting properties and a useful reformulation, which we present in Section 2.3.

The CSD measures the strength and direction of dependence between any marginal combination of sources, of which there are $2^K - 1$ for K sources, ranging from the K single-source CSDs, which we term the one-way CSDs, to the single K-source CSD, which we term the K-way CSD.

Example 2.1 Consider the two-source case with sources denoted by A and B. There are 2 one-way CSDs identically equal to 0 (see Remark 2.3) and a single two-way CSD given by

$$C_{AB} = \log\left(\frac{P[A \cap B]}{P[A]P[B]}\right).$$

Examination of the form of C_{AB} shows that a value of 0 corresponds to independence, whilst $C_{AB} > 0$ corresponds to positive dependence ($P[A|B] > P[A]$ and $P[B|A] > P[B]$) and $C_{AB} < 0$ negative dependence ($P[A|B] < P[A]$ and $P[B|A] < P[B]$).

Example 2.2 Now consider the three-source case, with sources A, B and C. There are $2^3 - 1 = 7$ CSDs in total. The 3 one-way CSDs are identically equal to 0 (see Remark 2.3), whilst the 3 two-way CSDs are given by:

$$C_{AB} = \log\left(\frac{P[A \cap B]}{P[A]P[B]}\right), C_{AC} = \log\left(\frac{P[A \cap C]}{P[A]P[C]}\right), C_{BC} = \log\left(\frac{P[B \cap C]}{P[B]P[C]}\right),$$

and the single three-way CSD given by

$$C_{ABC} = \log\left(\frac{P[A \cap B \cap C]}{P[A]P[B]P[C]}\right).$$

We note that the interpretation of the CSD sign as indicating the direction of independence still holds for the three-way CSD. For instance if $C_{ABC} > 0$, then $P[A \cap B|C] > P[A]P[B]$, and so on.

2.2.2 Properties

In this section we consider the behaviour of CSDs in several respects: we consider the CSDs implied by HJLLMs, some inequalities for CSDs, their relationship to N, and

their relationship to the Coefficient of Covariation of Chao et al. (2001).

HJLLMs and CSDs

In order to provide an interpretive background to CSDs in terms of known dependence structures, we first examine the behaviour of the CSDs for all dependence structures that can be represented by a hierarchical joint log-linear model. Recall that, for such models, all possible interactions of lower order which can be formed from each of the higher order interactions must be included in the model. The available dependence structures are mutual independence, joint independence, conditional independence and mutual dependence (as described in Chapter 1). Before examining the form of these dependence structures, consider a general result to be used for the specific conditional independence structures for the three-source and four-source cases presented in Section 2.4.2 below where we simultaneously present the corresponding results for the CIDs and the marginal models (to be introduced in Section 2.4).

Theorem 2.3 *Let $\mathcal{R}$, $\mathcal{T}$ and $\mathcal{S}$ denote arbitrary sets of sources such that $\mathcal{R} \cap \mathcal{S} = \varnothing$. We recall the HJLLM conventions presented in Section 1.2.1. Then if the groups $[\mathcal{R}, \mathcal{T}]$ and $[\mathcal{S}, \mathcal{T}]$ appear in the HJLLM specification with no $[\mathcal{A}, \mathcal{B}]$ specification where $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{S}$ (so that sources in $\mathcal{R}$ and sources in $\mathcal{S}$ are conditionally independent given sources in $\mathcal{T}$)*

$$C_{\mathcal{R} \cup \mathcal{S} \cup \mathcal{T}} = C_{\mathcal{R} \cup \mathcal{T}} + C_{\mathcal{S} \cup \mathcal{T}} - C_{\mathcal{T}}.$$

The proof of this Theorem appears in Appendix D on page 222.

Remark 2.4 In the simplest of cases, $\mathcal{R} = \{\mathcal{B}\}$, $\mathcal{S} = \{\mathcal{C}\}$ and $\mathcal{T} = \{\mathcal{A}\}$. In this instance, if $[AB]$ and $[AC]$ belong in an HJLLM description (entailing conditional independence of B and C given A), the relationship $C_{ABC} = C_{AB} + C_{AC} - C_A$ must hold by Theorem 2.3. In other words the 3-way component of dependence is decomposed in such a manner.

Remark 2.5 The general results presented in this section will enable us to explore specific characteristics of the CSDs for the three-source and four-source case, to be presented in Section 2.4.2. Theorem 2.3 states that a HJLLM of conditional independence entails specific linear equality constraints on the CSDs.

Bounds on CSDs

We now explore relationships between the CSDs. Specifically, we consider bounds on the CSDs, which, as will be demonstrated, are a direct consequence of the order relationships amongst the underlying marginal source probabilities.

There is an inherent ordering on marginal probabilities. If $\mathcal{Q}$ is a set of sources and $\mathcal{Q}^* \subset \mathcal{Q}$ is any subset, then $P\left[\cap_{S\in\mathcal{Q}^*} S\right] \leq P\left[\cap_{S\in\mathcal{Q}} S\right]$. Consider, for instance, the probability $P[A \cap B]$ that a randomly selected individual is captured by both sources A and B. This probability is bounded above by the individual probabilities of membership in each of sources A and B, denoted by $P[A]$ and $P[B]$, respectively. That is

$$P_{AB} \leq \min\{P_A, P_B\},$$

where P_A, P_B and P_{AB} denote $P[A]$, $P[B]$ and $P[A \cap B]$, respectively.

These ideas form the basis of the following Proposition, which places upper and lower bounds on a CSD relating to an arbitrary number of sources.

Proposition 2.1 *Let $\mathcal{R}$ denote a set of sources, with corresponding CSD $C_\mathcal{R}$. We define $P_\mathcal{R} = P\left[\cap_{S\in\mathcal{R}} S\right]$. Then*

$$\log\left(\frac{P_\mathcal{R}}{\overline{P}_\mathcal{R}^{|\mathcal{R}|}}\right) \leq C_\mathcal{R} \leq \log\left(\frac{P_\mathcal{R}}{\prod_{S\in\mathcal{R}} P_S}\right),$$

where

$$\begin{aligned}\overline{P}_{\mathcal{R}} &= \max_{S \in \mathcal{R}} Pr[S] \\ \underline{P}_{\mathcal{R}} &= \min_{\substack{\mathcal{Q} \in \mathbb{P}(\mathcal{R}) \\ |\mathcal{Q}| = |\mathcal{R}| - 1}} P_{\mathcal{Q}},\end{aligned}$$

and where $\mathbb{P}(\mathcal{R})$ *is the power set of* $\mathcal{R}$.

Example 2.3 We consider the workings of Proposition 2.1 in the two-source case. Without loss of generality, assume that $P_A \geq P_B$. Then

$$P_{AB} \leq P_B.$$

Therefore

$$\frac{P_{AB}}{P_A P_B} \leq \frac{P_B}{P_A P_B} = \frac{1}{P_A} \leq \frac{1}{P_B},$$

and the corresponding CSD, C_{AB}, is bounded in the following manner

$$\log\left(\frac{P_{AB}}{P_A^2}\right) \leq C_{AB} = \log\left(\frac{P_{AB}}{P_A P_B}\right) \leq -\log(P_A). \tag{2.1}$$

Example 2.4 For three sources, a more extensive series of bounds exists. Consider sources A, B and C. Similarly to the two-source case above, without loss of generality, assume that $P_A \geq P_B \geq P_C$. Then

$$\begin{aligned}P_{AB} &\leq \min\{P_A, P_B\} = P_B \\ P_{AC} &\leq \min\{P_A, P_C\} = P_C \\ P_{BC} &\leq \min\{P_B, P_C\} = P_C \\ P_{ABC} &\leq \min\{P_{AB}, P_{AC}, P_{BC}\} \leq P_C,\end{aligned}$$

which leads to upper bounds on the three pairwise CSDs given by

$$\begin{aligned}C_{AB} &\leq -\log(P_A) \\ C_{AC} &\leq -\log(P_C) \\ C_{BC} &\leq -\log(P_C),\end{aligned}$$

and an upper bound on the threeway CSD given by

$$C_{ABC} = \log\left(\frac{P_{ABC}}{P_A P_B P_C}\right) \leq \log\left(\frac{P_C}{P_A P_B P_C}\right) \leq \log\left(\frac{1}{P_A P_B}\right) = -\log(P_A P_B),$$

with a tighter upper bound given by

$$C_{ABC} = \log\left(\frac{P_{ABC}}{P_A P_B P_C}\right) \leq \log\left(\frac{P_2}{P_A P_B P_C}\right),$$

where $P_2 = \min\{P_{AB}, P_{AC}, P_{BC}\}$.

Lower bounds on the pairwise CSDs are given as follows

$$\begin{aligned} C_{AB} &\geq \log\left(\frac{P_{AB}}{P_A^2}\right) \\ C_{AC} &\geq \log\left(\frac{P_{AC}}{P_A^2}\right) \\ C_{BC} &\geq \log\left(\frac{P_{BC}}{P_B^2}\right). \end{aligned} \tag{2.2}$$

Remark 2.6 Figure 2.1 shows the region of feasible values for C_{AB} for the case where $P_A > P_B > P_C$ with fixed marginal probabilities $P_{AB} = 0.2, P_{AC} = 0.1$ and $P_{BC} = 0.099$. In this case, as shown in (2.1), the upper bound is controlled by P_A, whilst the lower bound is controlled by both P_A and P_{AB}, as shown by (2.2).

For any capture-recapture data set, the set of feasible values of all of the CSDs is thus constrained by the magnitude and ordering of the marginal source probabilities. This relationship demonstrates that, although the CSDs range over $\mathbb{R}$, there exist constraints on the set of all CSDs for a given data set that cannot be overlooked.

Relationship to the Coefficient of Covariation

The CSD is related to the Coefficient of Covariation (CCV) introduced by Chao and Tsay (1998), which was discussed in Section 1.3. For the two-source case, the relationship is given by the following proposition.

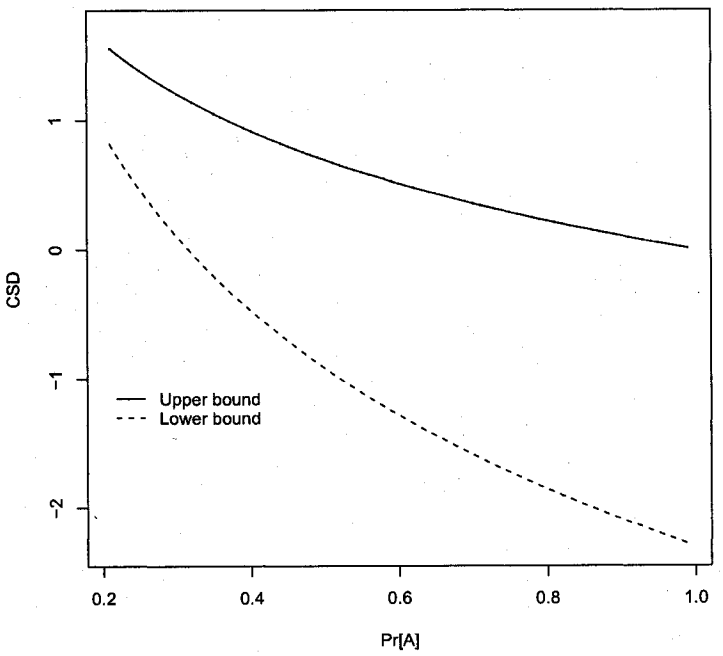


Figure 2.1: Upper and lower bounds for C_{AB} for $P_A > P_B > P_C$ with $P_{AB} = 0.2$, $P_{AC} = 0.1$ and $P_{BC} = 0.099$.

Proposition 2.2 *Consider two sources, A and B. Let ω_{AB} denote the two-source CCV defined by* (1.10)*. Then*

$$\omega_{AB} = \exp(C_{AB}) - 1 = c_{AB} - 1. \tag{2.3}$$

Proof. Consider the definition of the CCV for sources A and B, as given by (1.10). Recall that X_{iA} and X_{iB} are indicator functions that individual i belongs to sources A and B, respectively, and μ_A and μ_B denote average inclusion probabilities for sources A and B, respectively. Now, since these random variables, X_{iA} and X_{iB}, are indicator functions, the expectation of the product is simply $\mathbb{E}[X_{iA}X_{iB}] = Pr[X_{iA} = 1, X_{iB} = 1] = Pr[A \cap B|I = i]$, in terms of the notation developed here in this dissertation.

That is, the expectation of the product of the random variables X_{iA} and X_{iB} is equal to the probability that the ith individual is in both sources A and B. Thus, from the definition of ω_{AB} we obtain

$$\begin{aligned}
\omega_{AB} &= \frac{1}{N}\frac{\sum_{i=1}^{N}\mathbb{E}\left[(X_{iA}-\mu_A)(X_{iB}-\mu_B)\right]}{\mu_A\mu_B} \\
&= \frac{1}{N}\frac{\sum_{i=1}^{N}\mathbb{E}\left[X_{iA}X_{iB}-\mu_A X_{iB}-\mu_B X_{iA}+\mu_A\mu_B\right]}{\mu_A\mu_B} \\
&= \frac{1}{N}\frac{\sum_{i=1}^{N}\mathbb{E}\left[X_{iA}X_{iB}\right]}{\mu_A\mu_B} - \frac{(\mu_A N\mu_B+\mu_B N\mu_A-N\mu_A\mu_B)}{N\mu_A\mu_B} \\
&= \frac{1}{N}\frac{\sum_{i=1}^{N}\mathbb{E}\left[X_{iA}X_{iB}\right]}{\mu_A\mu_B} - 1, \qquad (2.4)
\end{aligned}$$

Under random sampling of individuals in the population of interest, $Pr[I=i] = 1/N$. We can therefore equate the average probability μ_A of inclusion in source A and the probability that a randomly selected individual is observed in source A.

$$\begin{aligned}
\mu_A &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[X_{iA}\right] \\
&= \sum_{i=1}^{N} Pr[A|I=i]Pr[I=i] \\
&= \sum_{i=1}^{N} Pr[A, I=i] \\
&= Pr[A]. \qquad (2.5)
\end{aligned}$$

Likewise $\mu_B = Pr[B]$.

Consider the numerator of (2.4). Using the Law of Total Probability, it can be re-expressed as

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[X_{iA}X_{iB}\right] &= \sum_{i=1}^{N} Pr[A\cap B|I=i]Pr[I=i] \\
&= \sum_{i=1}^{N} Pr[A\cap B, I=i] \\
&= Pr[A\cap B]. \qquad (2.6)
\end{aligned}$$

Substituting from (2.5) and (2.6) into (2.4) leads to

$$\begin{aligned}\omega_{AB} &= \frac{1}{N}\frac{\sum_{i=1}^{N} Pr[A \cap B | I = i]}{\mu_A \mu_B} - 1 \\ &= \frac{Pr[A \cap B]}{Pr[A]Pr[B]} - 1 \\ &= c_{AB} - 1,\end{aligned}$$

as required. □

Remark 2.7 This relationship demonstrates that the two-way CCV is a translation of the exponential of the corresponding CSD. It is immediately obvious that such a relationship holds in the case of independence: when sources A and B are independent, the CSD is known to be given by $C_{AB} = 0$ and $c_{AB} = \exp(C_{AB}) = 1$, just as the CCV is given by $\omega_{AB} = 0$.

Remark 2.8 An approximation can be derived as follows. Taking the first-order Taylor expansion of the natural logarithm of c_{AB}, i.e. of C_{AB}, which is valid for c_{AB} close to 1 (i.e. close to independence) yields

$$\omega_{AB} = c_{AB} - 1 \simeq \log c_{AB} = C_{AB}. \tag{2.7}$$

Proposition 2.3 *Consider three sources, A, B and C. Let ω_{ABC} denote the three-way CCV. Then*

$$\omega_{ABC} = (c_{ABC} - 1) - (c_{AB} - 1) - (c_{BC} - 1) - (c_{AC} - 1). \tag{2.8}$$

If the CSDs are close to zero, then

$$\omega_{ABC} \simeq C_{ABC} - (C_{AB} + C_{BC} + C_{AC}) + (C_A + C_B + C_C) \tag{2.9}$$

Here we present an outline of the proof using the same assumptions as for the relationship for two sources derived above. The full derivation is presented in Appendix D on page 224.

Proof.

$$\begin{aligned}
\omega_{ABC} &= (c_{ABC} - 1) - (c_{AB} - 1) - (c_{BC} - 1) - (c_{AC} - 1) \\
&\simeq \log c_{ABC} - \log c_{AB} - \log c_{BC} - \log c_{AC}, \text{ from the first order Taylor expansion for log} \\
&= \log\left(\frac{c_{ABC}}{c_{AB}c_{BC}c_{AC}}\right) \\
&= \log\left(\frac{c_{ABC}c_Ac_Bc_C}{c_{AB}c_{BC}c_{AC}}\right), \text{ since by definition } c_A = c_B = c_C = 1 \\
&= C_{ABC} - (C_{AB} + C_{BC} + C_{AC}) + (C_A + C_B + C_C), \text{ by definition of the CSDs}
\end{aligned}$$

□

Remark 2.9 Thus, the CCV for three sources, ω_{ABC}, is approximately equal to a linear combination of all three-way, two-way and single-source CSDs. Moreover, the linear combination has an inclusion/exclusion form. Although it may appear unimportant to include the single-source CSDs, since they are identically equal to 0, it is formally useful to do so for the development of an alternative measure of dependence, the Coefficient of Incremental Dependence, which will be introduced in the next section.

Remark 2.10 For four sources the following relationship can be derived using the same assumptions as for the three source case.

$$\begin{aligned}
\omega_{ABCD} &= (c_{ABCD} - 1) + (c_{AB} - 1) + (c_{BC} - 1) + (c_{BD} - 1) + (c_{AC} - 1) \\
&\quad + (c_{AD} - 1) + (c_{CD} - 1) - (c_{ABC} - 1) - (c_{ABD} - 1) - (c_{ACD} - 1) - (c_{BCD} - 1) \\
&\simeq \log\left(\frac{c_{ABCD}c_{AB}c_{AC}c_{AD}c_{BC}c_{BD}c_{CD}}{c_{ABC}c_{ABD}c_{ACD}c_{BCD}}\right), \text{ if the CSDs are close to zero} \\
&= \log\left(\frac{c_{ABCD}c_{AB}c_{AC}c_{AD}c_{BC}c_{BD}c_{CD}}{c_{ABC}c_{ABD}c_{ACD}c_{BCD}c_Ac_Bc_Cc_D}\right) \\
&= C_{ABCD} - (C_{ABC} + C_{ABD} + C_{ACD} + C_{BCD}) \\
&\quad + (C_{AB} + C_{AC} + C_{AD} + C_{BC} + C_{BD} + C_{CD}) - (C_A + C_B + C_C + C_D)
\end{aligned}$$

Again, as with the three-source case, the four-way CCV ω_{ABCD}, is approximately equal to an inclusion-exclusion linear combination of the four-way CSD and all three-way, two-way and single-source CSDs. Such relationships play an important role in motivating the reformulation of the CSDs presented in the next section.

Rate of Change of CSDs with change in N

We wish to examine the form of the CSDs with a change in the assumed value of N. To do so, we consider completing the incomplete contingency table of n_{obs} individuals with various additional numbers of individuals corresponding to different values of N. See Table 2.1 for a prototypical three-source incomplete contingency table. In order to present Example 2.5 below, a three-source example of the form of the CSDs for a change in the assumed value of N, we first present some necessary theory.

Consider the data collected from a capture-recapture study for a set of K sources, denoted by $\dot{\mathcal{Q}}$. There are $2^K - 1$ data points corresponding to the source membership data aggregated over all n_{obs} observed individuals. For each completed table, it is possible to obtain the exact value of the $2^K - 1$ CSDs denoted by $C_{\mathcal{R}}$, where $\mathcal{R} \subset \mathcal{Q}$, for $|\mathcal{R}| = 1, \ldots, K$. When N is known, the marginal probabilities of belonging to a specific combination of sources can be expressed in terms of marginal means and N.

Proposition 2.4 *Let $\mathcal{Q}$ be a set of sources and let N be known. Then $P[S] = m_S/N$ for a source $S \in \mathcal{Q}$ and $P[\mathcal{R}] = m_{\mathcal{R}}/N$, for $\mathcal{R} \subset \mathcal{Q}$.*

Remark 2.11 Notation.
We use $P[\mathcal{R}]$ to denote $P[\mathcal{R}_\cap]$ and $m_{\mathcal{R}}$ to denote $m_{\mathcal{R}_\cap}$, where $\mathcal{R}_\cap = \cap_{\mathcal{S} \in \mathcal{R}} \mathcal{S}$ for $\mathcal{R}$ a set of sources.

The relationships characterized by Proposition 2.4 enables the CSDs to be reformulated in the following manner. Substituting for such relationships into Defini-

tion 2.1 yields the following

$$c_{\mathcal{R}} = \frac{Pr[\mathcal{R}]}{\prod_{S \in \mathcal{R}} Pr[S]} = N^{|\mathcal{R}|-1} \frac{m_{\mathcal{R}}}{\prod_{S \in \mathcal{R}} m_S},$$

which is a ratio of marginal means scaled by some power of N. This, in turn, leads to an alternative expression for the CSD in terms of marginal means and N rather the marginal probabilities of the original definition of the CSD given by Definition 2.2.

Proposition 2.5 *Let $\mathcal{Q}$ be a set of sources and let N be known. Then*

$$C_{\mathcal{R}} = (|\mathcal{R}| - 1) \log N + \log m_{\mathcal{R}} - \sum_{S \in \mathcal{R}} \log m_S, \text{ for } \mathcal{R} \subset \mathcal{Q}. \tag{2.10}$$

Remark 2.12 Now, when N is known, the expected marginal counts are simply equal to the observed counts, since the complete table accounts for the entire population. That is, $m_{\mathcal{R}} = n_{\mathcal{R}}$, for $\mathcal{R} \subset \mathcal{Q}$.

Using (2.10) for the complete table, the relationship between N and the CSDs is given by the following proposition.

Proposition 2.6 *Let $\mathcal{Q}$ be a set of sources and let N be known. Then a natural estimator $\widehat{C}_{\mathcal{R}}$ of $C_{\mathcal{R}}$ is*

$$\widehat{C}_{\mathcal{R}} = (|\mathcal{R}| - 1) \log N + \log n_{\mathcal{R}} - \sum_{S \in \mathcal{R}} \log n_S, \text{ for } \mathcal{R} \subset \mathcal{Q}.$$

Remark 2.13 It is evident from Proposition 2.6 that all estimated k-way CSDs, $k = 1, \ldots, |\mathcal{Q}|$, increase with N at the same rate since only the $\log N$ term changes with N in a nonlinear manner (as a function of N). Pairwise CSDs, for example $\widehat{C}_{AB} = \log N + \log n_{AB} - (\log n_A + \log n_B)$, have a positive rate of change $1/N$ which decreases as N increases.

Example 2.5 With this example we will explore the properties of the CSDs for N known, using the theory outlined in Propositions 2.4- 2.6. Consider a three-source

simulated data example for which source membership for each of 1000 individuals is assigned according to a multinomial distribution with fixed probabilities for each of the 2^3 cells of the complete contingency table. As usual in the capture-recapture setting, the cell corresponding to the number of individuals observed in none of the three sources would be unobserved. An example of simulated data is presented in Table 2.1 (a description of the data generation mechanism is presented in Appendix C). Of the 1000 individuals of the simulated population, 754 were observed. For this

	B_{Yes}		B_{No}	
	C_{Yes}	C_{No}	C_{Yes}	C_{No}
A_{Yes}	117	96	64	72
A_{No}	109	134	162	$n_{unobs} = ?$

Table 2.1: Observed sample simulated from population size 1000

capture-recapture data set we wish to examine the form of the CSDs with a change in the assumed value of N, i.e with departures from the true value of $N = 1000$. That is, it is assumed that the data set of the incomplete contingency table remains the same but that N changes. The values of the estimated CSDs corresponding to different values of N will be obtained using Proposition 2.6. In this way it will be possible to observe the effect of a changing N on the dependence structure present in such data, as well as to explore the relationships between the CSDs themselves, specifically between the pairwise CSDs and the three-way relative to the pairwise.

The relationship of Proposition 2.6 is confirmed by Figure 2.2. We see equal change for all k-way CSDs with increasing N (Remark 2.13). It is clear that the change is equal for all pairwise CSDs: the three lines are parallel. For the single three-way CSD, C_{ABC}, the change is governed by $2 \log N$ rather than by $\log N$, as N increases. Again, such a difference is clearly apparent in Figure 2.2, in which the line

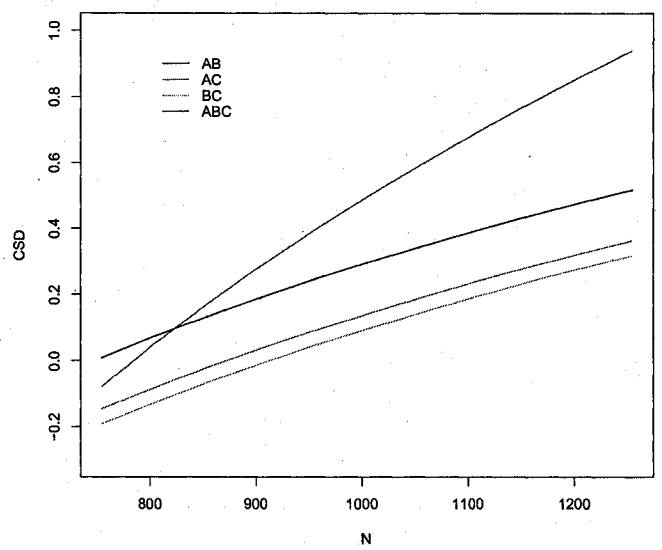


Figure 2.2: Estimated CSDs for data of Table 2.1 for different values of N

corresponding to C_{ABC} diverges from the three pairwise CSDs.

2.3 Coefficients of incremental dependence

In this section we introduce a second measure, the Coefficient of Incremental Dependence (CID) (see Definition 2.4) which is designed to decompose dependence into its K-way components.

2.3.1 Motivation

The CSDs measure dependence in a manner which can be thought of as absolute. As an illustration, consider the three-way CSD, C_{ABC}. It can be thought of as operating on an absolute scale since it does not quantify how much additional dependence

can be explained by all three sources compared to that measured by the 3 pairwise CSDs, C_{AB}, C_{AC} and C_{BC}. Rather, it quantifies all dependence at the level of the three sources irrespective of that present at all of the two-way levels. An alternative approach would be to develop a measure which quantifies the additional three-way dependence not already accounted for in the two-way dependence structure.

In this section we present an example, Example 2.6, for the three-source case. This example provides more specific motivation for a reformulation of the CSDs.

Example 2.6 Consider a simple three-source example for which the only dependence present occurs between sources A and B, i.e. an underlying model of joint independence denoted by $[AB][C]$. Suppose that $C_{AB} = \log(1.2)$. By assumption, the other two pairwise CSDs are given by $C_{AC} = C_{BC} = 0$. In this case, as was shown above in Section 2.2.2, $C_{ABC} = C_{AB} = \log(1.2)$. But given that, by assumption, there is no three-way dependence, a measure of three-way dependence that would take on a value of zero would be attractive in this case.

Consider the following expression:

$$C_{ABC} - (C_{AB} + C_{AC} + C_{BC}) = \log(1.2) - (\log(1.2) + 0 + 0) = 0.$$

Being equal to 0, this expression provides a sense of the magnitude of how much additional dependence is accounted for jointly by all three sources relative to that which can be accounted for by all pairs. It would appear to be a natural measure to adopt for such an example.

Remark 2.14 The second measure of dependence which we propose, namely the Coefficient of Incremental Dependence (see Definition 2.4), is of a form related to the property described in Example 2.6. In the next section the CIDs, which are designed to decompose dependence into its several K-way components, will be defined and their properties explored.

2.3.2 Definition

Definition 2.4 *Let $\mathcal{Q}$ be any set of sources. Then the Coefficient of Incremental Dependence (CID) $\gamma_{\mathcal{Q}}$ for the sources contained in $\gamma_{\mathcal{Q}}$ is given by*

$$\gamma_{\mathcal{Q}} = \sum_{j=1}^{n} (-1)^{j_n} \sum_{\substack{\mathcal{R} \subset \mathcal{Q} \\ |\mathcal{R}| = j}} C_{\mathcal{R}}, \tag{2.11}$$

where (here and hereafter) the notation $\sum_{\substack{\mathcal{R} \subset \mathcal{Q} \\ |\mathcal{R}| = j}} C_{\mathcal{R}}$ *indicates that the sum is taken over all subsets $\mathcal{R}$ of $\mathcal{Q}$ with cardinality j, and $j_n = j$ if n is even and $j + 1$ if n is odd. We also define $\gamma_{\varnothing} = 0$.*

Remark 2.15 For completeness, we note that the single-source CIDs are set equal to the single-source CSDs, which are defined to be 0.

Example 2.7 The differences between the form of the CID (see Definition 2.4) when the cardinality of $\mathcal{R}$ is even compared to when it is odd, will be illustrated using the three-source case. Using Definition 2.4 for three sources, A, B and C, the CIDs are defined as follows

$$\begin{aligned}
\gamma_A &= C_A = 0 \\
\gamma_B &= C_B = 0 \\
\gamma_C &= C_C = 0 \\
\gamma_{AB} &= C_{AB} \\
\gamma_{AC} &= C_{AC} \\
\gamma_{BC} &= C_{BC} \\
\gamma_{ABC} &= C_{ABC} - (C_{AB} + C_{AC} + C_{BC})
\end{aligned} \tag{2.12}$$

Remark 2.16 Non-single source CIDs are related to non-single source CSDs by a linear transformation represented by $\boldsymbol{\gamma} = \mathbf{GC}$ in matrix/vector terms, where $\mathbf{C}$ and $\boldsymbol{\gamma}$ denote the $(2^K - K - 1) \times 1$ vector of non-single source CSDs and CIDs, respectively, and $\mathbf{G}$ represents the linear transformation given by (2.11).

Example 2.8 In the three-source case, we can take

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{BC} \\ \gamma_{ABC} \end{bmatrix}, \mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} C_{AB} \\ C_{AC} \\ C_{BC} \\ C_{ABC} \end{bmatrix} \tag{2.13}$$

Remark 2.17 Notice that the approximate relationships observed between the CCV and CSD, given by (2.7), and (2.9) for two and three sources, respectively, are precisely equal to the definition of the CIDs, as exemplified here in the three-source case (2.12). Thus, there is an approximate equality between the CID and CCV for a given set of sources.

Remark 2.18 Consider the definition of the three-way term, γ_{ABC} in (2.12). It is defined as the difference between the three-way CSD, C_{ABC}, and the sum of the three pairwise CSDs, C_{AB}, C_{AC} and C_{BC}. Such a difference represents the additional three-way dependence not explained jointly by the pairwise dependence, which is measured by the sum of the pairwise CSDs.

2.3.3 Properties

In this section we present properties of the CIDs and their relationship to the CSDs. The main result of this section, Theorem 2.5, presents the general form of the CSDs in terms of the CIDs by deriving the inverse of the relationship provided by Definition 2.4.

Theorem 2.5 *Let $\mathcal{Q}$ be any set of sources and denote by $C_{\mathcal{Q}}$ the CSD associated with this set of sources. Then*

$$C_{\mathcal{Q}} = \sum_{\mathcal{S} \subset \mathcal{Q}} \gamma_{\mathcal{S}}. \tag{2.14}$$

The proof of this Theorem appears in Appendix D on page 225.

Remark 2.19 Theorem 2.5 serves to make explicit the manner in which the CIDs decompose dependence. It formalizes the inverse relationship between the CIDs and the CSDs. That is, it provides the inverse transformation of the CSDs in terms of the CIDs. The original transformation, with the CIDs expressed in terms of the CSDs, is that used as the definition of the CIDs, given by Definition 2.4. The specific form of this relationship for the three-source case was described above with the complete set of CIDs given by (2.12). The corresponding inverse relationship, with the system of CSDs expressed in terms of the CIDs, is given by (2.15) in the following example. It is evident that there is a one-to-one relationship between the system of CSDs and CIDs for three sources as is the case for the general K-source case. The specific form of the $\mathbf{G}$ matrix for the four-source case is provided in Appendix A.

Remark 2.20 Non-single source CSDs are related to non-single source CIDs by the inverse of the linear transformation given by Remark 2.16. Thus, the inverse transformation is represented by $\mathbf{C} = \mathbf{G}^{-1}\boldsymbol{\gamma}$ in matrix/vector terms, where $\mathbf{G}^{-1}$ represents the transformation for non-single sources given, for example, by (2.13) for three sources.

Example 2.9 For three sources, the inverse transformation which expresses the single-source CSDs in terms of the CIDs is given by

$$C_A = \gamma_A = 0$$

$$C_B = \gamma_B = 0$$

$$C_C = \gamma_C = 0,$$

whilst that for the non-single source CSDs is given by $\mathbf{C} = \mathbf{G}^{-1}\boldsymbol{\gamma}$, where

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{2.15}$$

with $\boldsymbol{\gamma}$ and $\mathbf{C}$ are given in Example 2.8.

Example 2.10 (Example 2.5 continued.)
We consider now the rate of change of the CIDs for changes in N. Figure 2.3 shows the CIDs changing with N in the way that Figure 2.2 showed the CSDs changing with N. The same data set is used. It is evident that the three-way CID changes at the same rate as each of the pairwise CIDs rather than at a different rate as was the case with the CSDs. It can be shown that the same relationship is observed for the general K-source case. Thus, working with the Coefficients of Incremental Dependence rather than the Coefficients of Source Dependence enables us to measure dependence between any number of sources on the same scale rather than on different scales. Therefore, the CIDs are comparable in terms of magnitude irrespective of the number of sources, whereas the corresponding CSDs are not.

Theorem 2.5 provides an expression for a k-way CSD in terms of all CIDs of equal or lower order. It leads to the following result relating marginal source probabilities and the CIDs.

Corollary 2.1 *Let* $\pi_{\mathcal{Q}} = \log P\left[\cap_{S\in\mathcal{Q}} S\right]$. *Then*

$$\pi_{\mathcal{Q}} = \sum_{S\in\mathcal{Q}} \pi_{\{S\}} + \sum_{\mathcal{R}\subset\mathcal{Q}} \gamma_{\mathcal{R}} \tag{2.16}$$

Proof. The proof is immediate from $C_{\mathcal{Q}} = \pi_{\mathcal{Q}} - \sum_{S\in\mathcal{Q}} \pi_{\{S\}}$ and Theorem 2.5. □

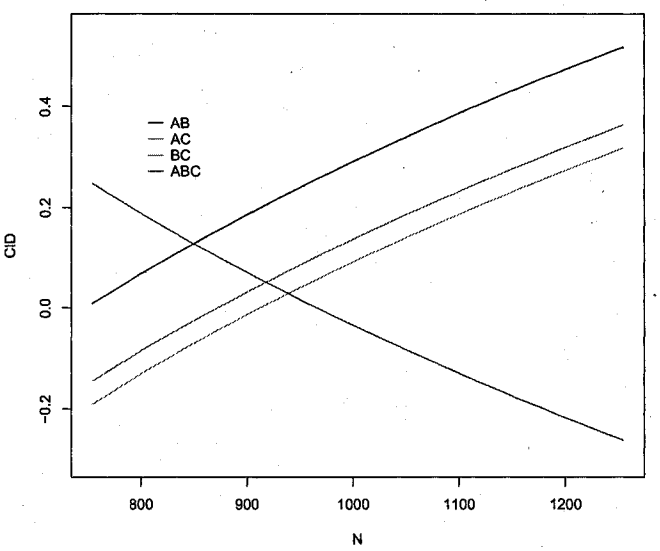


Figure 2.3: Estimated CIDs corresponding to estimated CSDs of Figure 2.2

Remark 2.21 Thus, Corollary 2.1 provides an explicit decomposition of the log-marginal source probability $\pi_{\mathcal{Q}}$ corresponding to the set of sources $\mathcal{Q}$. Such a decomposition is formed by the constituent single-source marginal probabilities (the $\pi_{\{S\}}$ for $S \in \mathcal{Q}$) of $\mathcal{Q}$ and all CIDs of order $|\mathcal{Q}|$ and lower. In the next section, we will see that using Corollary 2.1 to decompose $\pi_{\mathcal{Q}}$, the log-marginal probability for a set of sources $\mathcal{Q}$, naturally leads to the form of a marginal model. We present the form of the model and describe its properties.

Corollary 2.1 leads, in turn, to Theorem 2.6.

Theorem 2.6 *Let $\mathcal{Q}$ be a set of sources and $\pi_{\mathcal{Q}} = \log P[\cap_{S\in\mathcal{Q}} S]$. Then for $|\mathcal{Q}| = n \geq 2$ and j_n defined as per Definition 2.4*

$$\gamma_{\mathcal{Q}} = \sum_{j=1}^{n} (-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q} \\ |\mathcal{R}|=j}} \pi_{\mathcal{R}}. \tag{2.17}$$

The proof of this Theorem appears in Appendix D on page 229.

Remark 2.22 Theorem 2.6 expresses the CID for the set of sources $\mathcal{Q}$ as an inclusion/exclusion form on marginal source probabilities. The equivalent form in terms of the CSDs, as per Definition 2.4, is exactly the same but with $C_{\mathcal{R}}$ replacing $\pi_{\mathcal{R}}$: any CID can be equivalently expressed in terms of all CSDs of equal or lower order or all marginal probabilities of equal or lower order.

2.4 Marginal log-linear models using source dependence measures

Throughout the discussion of measures of dependence it has been emphasized that one of the goals of such theory is to develop a new modelling approach. In this section we will describe the development of the model from Corollary 2.1.

First, let $\mathcal{Q}$ denote a set of sources. Elementary results from sampling theory for population size N yield

$$P[\mathcal{Q}] = \frac{m_{\mathcal{Q}}}{N},$$

where $P[\mathcal{Q}]$ and $m_{\mathcal{Q}}$ are used to denote $P[\mathcal{Q}_{\cap}]$ and $m_{\cap_{S\in\mathcal{Q}}}$, respectively, which correspond to the probability of, and expected number of individuals in, the intersection of the sources contained in $\mathcal{Q}$, respectively. Equivalently

$$\pi_{\mathcal{Q}} = \log\left(\frac{m_{\mathcal{Q}}}{N}\right), \tag{2.18}$$

for known N.

We have the following proposition.

Proposition 2.7 *Let $\mathcal{Q}$ be a set of sources and let $m_{\mathcal{Q}}$ be defined as above. Then*

$$\log m_{\mathcal{Q}} = -(|\mathcal{Q}| - 1)\log N + \sum_{S\in\mathcal{Q}} \log m_S + \sum_{\mathcal{R}\subset\mathcal{Q}} \gamma_{\mathcal{R}}.$$

Proof. Combining (2.18) with Corollary 2.1 yields

$$\begin{aligned} \pi_{\mathcal{Q}} = \log\left(\tfrac{m_{\mathcal{Q}}}{N}\right) &= \textstyle\sum_{S\in\mathcal{Q}} \pi_{\{S\}} + \sum_{\mathcal{R}\subset\mathcal{Q}} \gamma_{\mathcal{R}} \\ &= \textstyle\sum_{S\in\mathcal{Q}} \log\left(\tfrac{m_S}{N}\right) + \sum_{\mathcal{R}\subset\mathcal{Q}} \gamma_{\mathcal{R}}, \end{aligned}$$

whence the result. □

Remark 2.23 Thus, Proposition 2.7 provides the general form of the marginal mean $m_{\mathcal{Q}}$ in terms of the single-source marginal means and all CIDs of equal or lower order.

The full development of the marginal model is obtained by parameterizing the single-source marginal means as $\beta_S = \log m_S$ and $\beta_0 = \log N$. This leads us to define the marginal log-linear model in the following manner:

Definition 2.7 *Let $\mathcal{Q}$ be a set of sources and $K = |\mathcal{Q}|$. Then the following system of $2^K - 1$ equations constitutes the marginal log-linear model (MLLM) for the K sources of $\mathcal{Q}$.*

$$\log m_S = \beta_S, \quad \textit{for } S \in \mathcal{Q};$$
$$\log m_{\mathcal{R}} = -(|\mathcal{R}| - 1)\beta_0 + \sum_{S \in \mathcal{Q}} \beta_S + \sum_{\mathcal{T} \subset \mathcal{R}} \gamma_{\mathcal{T}}, \quad \textit{for } \mathcal{R} \subset \mathcal{Q}$$

It is a model on marginal means rather than on the joint cell means that are modelled by the joint log-linear modelling approach (see Section 1.2.1).

An equivalent parameterization in terms of the CSDs is obtained using Theorem 2.5, which yields the following proposition.

Proposition 2.8 *Let $\mathcal{Q}$ be a set of sources and let $d = 2^{|\mathcal{Q}|} - 1$. Then the following system of d equations is equivalent to the marginal log-linear model for $\mathcal{Q}$ as given by Definition 2.7.*

$$\log m_S = \beta_S, \quad \textit{for } S \in \mathcal{Q}$$
$$\log m_{\mathcal{R}} = -(|\mathcal{R}| - 1)\beta_0 + \sum_{S \in \mathcal{Q}} \beta_S + C_{\mathcal{R}}, \quad \textit{for } \mathcal{R} \subset \mathcal{Q}$$

Example 2.11 Consider the three-source case for source A, B and C. Using Definition 2.7, with $d = 2^3 - 1 = 7$, the full specification of the three-source marginal log-linear model is given by:

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + \gamma_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C + \gamma_{AC} \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C + \gamma_{BC} \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \gamma_{ABC}. \qquad (2.19)
\end{aligned}$$

The equivalent CSD parameterization, obtained from Proposition 2.8 is

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + C_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C + C_{AC} \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C + C_{BC} \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + C_{ABC}.
\end{aligned}$$

Definition 2.8 *Let $\mathcal{Q}$ be a set of sources and let $d = 2^{|\mathcal{Q}|} - 1$. Then the matrix form of the system of d equations for the $|\mathcal{Q}|$-source capture-recapture setting is given by:*

$$\log \mathbf{m} = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \tag{2.20}$$

where $\mathbf{m}$, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are $d \times 1$ vectors of marginal means, cell means, and CIDs, respectively, $\boldsymbol{\beta}$ is the $(|\mathcal{Q}| + 1) \times 1$ vector of all such parameters. $\mathbf{A}$ is the $d \times d$ upper triangular matrix which transforms cell means into marginal means, $\mathbf{X}$ is the $d \times (|\mathcal{Q} + 1|)$ design matrix for $\boldsymbol{\beta}$ and $\mathbf{Z}$ the design matrix for $\boldsymbol{\gamma}$.

Remark 2.24 Model form (2.20) fits into the class of models described by Lang and Agresti (1994) which simultaneously model the joint and marginal distributions of multivariate categorical responses. We note that their setting applies to complete contingency tables rather than the capture-recapture setting.

Example 2.12 (Example 2.11 continued.) Consider the three-source setting for sources A, B and C. Then $d = 2^3 - 1 = 7$. The matrix form of the marginal log-linear

model is given by Definition 2.8, with

$$\mathbf{m} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix} ; \boldsymbol{\mu} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} ; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_A \\ \gamma_B \\ \gamma_C \\ \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{BC} \\ \gamma_{ABC} \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ; \mathbf{X} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 \\ -2 & 1 & 1 & 1 \end{bmatrix} ; \mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Proposition 2.9 *The matrix form equivalent to that of Definition 2.8 in terms of the CSDs, for the general form of the $|\mathcal{Q}|$-source marginal model is given by:*

$$\log \mathbf{m} = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{C},$$

where all vectors and matrices are given by Definition 2.8, with $\mathbf{C}$ *the* $d \times 1$ *vector of CSDs. Note that the design matrix on* $\mathbf{C}$ *is the identity matrix, so is not written explicitly.*

Proof. The proof is immediate from the expression of the CSDs in terms of the CIDs, given by Theorem 2.5. □

Example 2.13 (Example 2.11 continued.) The matrix form in terms of the CSDs is given by Proposition 2.9, with all matrices as above and

$$\mathbf{C} = \begin{bmatrix} C_A \\ C_B \\ C_C \\ C_{AB} \\ C_{AC} \\ C_{BC} \\ C_{ABC} \end{bmatrix}.$$

2.4.1 Constrained parameter space of marginal log-linear model

Modelling marginal means presents some additional challenges as compared to modelling cell means. Since all cell means must be non-negative, i.e. $\boldsymbol{\mu} \geq \mathbf{0}$, only marginal means which correspond to non-negative cell means are feasible. That is, $\mathbf{m}$ is such that $\mathbf{A}^{-1}\mathbf{m} \geq \mathbf{0}$. In order to ensure non-negativity of the cell means, we examine the relationship between marginal and cell means via the form of the model.

The constraints placed on the marginal means via the marginal log-linear model are given by the following proposition.

Proposition 2.10 *Let $\mathcal{Q}$ be a set of sources. Then the marginal log-linear model given by Definition 2.7, and its equivalent reparameterization given by Proposition 2.8, is defined subject to the following system of constraints*

$$\begin{aligned} &\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{m} \geq \mathbf{0} \\ \Longrightarrow\ &\boldsymbol{\mu} = \mathbf{A}^{-1}\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) \geq \mathbf{0} \end{aligned}$$

Example 2.14 (Example 2.11 continued.) Consider again the three-source case. The relationship between the 7 cell means and 7 marginal means is given by

$$\begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix}$$

$$= \begin{bmatrix} m_A - m_{AB} - m_{AC} + m_{ABC} \\ m_B - m_{AB} - m_{BC} + m_{ABC} \\ m_C - m_{AC} - m_{BC} + m_{ABC} \\ m_{AB} - m_{ABC} \\ m_{AC} - m_{ABC} \\ m_{BC} - m_{ABC} \\ m_{ABC} \end{bmatrix} \tag{2.21}$$

Thus, ensuring non-negativity of the cell means corresponds to the following order relationship on the marginal means:

$$\begin{aligned} m_A - m_{AC} \geq m_{AB} - m_{ABC} \geq 0 \\ m_B - m_{AB} \geq m_{BC} - m_{ABC} \geq 0 \\ m_C - m_{BC} \geq m_{AC} - m_{ABC} \geq 0 \\ m_{ABC} \geq 0. \end{aligned}$$

Such an ordering is intuitive. For example, the marginal mean for sources A and C must be larger than that for sources A, B and C. This relationship is given by $m_{AB} - m_{ABC} \geq 0$.

2.4.2 Relationship to hierarchical joint log-linear models

In this section we will relate the marginal models to hierarchical joint log-linear models (HJLLM) in order to understand their similarities and differences and their place relative to standard analytical approaches for capture-recapture data. First we present a result (Theorem 2.9) that characterises the form of the CIDs corresponding to the conditional independence structures modelled by hierarchical joint log-linear models. We then examine the behaviour of the marginal model for all dependence structures that can be represented by a hierarchical joint log-linear model. The available dependence structures are mutual independence, joint independence, conditional independence and mutual dependence (as described in Chapter 1, Section 1.2.1). We consider the three and four-source cases as examples to present the specific form of the marginal model for each of the dependence structures. In each case we will obtain the form of the CSDs (using Theorem 2.3) and the CIDs (using Theorem 2.9 below) which characterize the MLLMs corresponding to the HJLLMs of each dependence structure. In the three-source case, we will state explicitly the form of the corresponding marginal model.

Remark 2.25 Notation.

We introduce the following notation: if $\mathcal{R}$, $\mathcal{S}$, $\mathcal{T}$ ... are disjoint sets of sources, we write $\mathbb{P}(\mathcal{R})$ for the power set of $\mathcal{R}$ and let $\mathbb{P}^*(\mathcal{R}) = \mathbb{P}(\mathcal{R})\backslash\{\varnothing\}$, and $\mathbb{P}^*(\mathcal{R}, \mathcal{S}, \mathcal{T}, \dots) = \mathbb{P}^*(\mathcal{R}) \oplus \mathbb{P}^*(\mathcal{S}) \oplus \mathbb{P}^*(\mathcal{T}) \oplus \dots$, where if $\mathbb{A}$ and $\mathbb{B}$ are classes of sets, $\mathbb{A} \oplus \mathbb{B} = \{S = A \cup B : A \in \mathbb{A}, B \in \mathbb{B}\}$. (We hold to the convention that $\mathbb{A} \oplus \mathbb{B} \oplus \varnothing = \mathbb{A} \oplus \mathbb{B}$.) Thus $\mathbb{P}^*(\mathcal{R}, \mathcal{S}, \mathcal{T})$ consists of lists of sources in which at least one source comes from each of $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{T}$.

Finally we let

$$\begin{aligned}\gamma[\mathcal{R}] &= \sum_{\mathcal{Q}\in\mathbb{P}^*(\mathcal{R})}\gamma_{\mathcal{Q}},\\ \gamma[\mathcal{R},\mathcal{S}] &= \sum_{\mathcal{Q}\in\mathbb{P}^*(\mathcal{R},\mathcal{S})}\gamma_{\mathcal{Q}},\end{aligned}$$

and so on.

Example 2.15 In the simplest form, $\mathcal{R} = \{A\}$ and $\mathcal{S} = \{B\}$. Then $\gamma[\mathcal{R}] = \gamma_A$, $\gamma[\mathcal{S}] = \gamma_B$ and $\gamma[\mathcal{R},\mathcal{S}] = \gamma_{AB}$.

Similarly to Theorem 2.3, we can characterize conditional independence hierarchical joint log-linear models using CIDs as follows.

Theorem 2.9 *Suppose that an HJLLM description contains the pair of descriptors* $[\mathcal{R}^*,\mathcal{T}]$ *and* $[\mathcal{S}^*,\mathcal{T}]$, $\mathcal{R}^*\cap\mathcal{S}^* = \varnothing$, *and no descriptor of the form* $[\mathcal{R},\mathcal{S}]$ *such that* $\mathcal{R}\subset\mathcal{R}^*$ *and* $\mathcal{S}\subset\mathcal{S}^*$, *so that sources in* $\mathcal{R}^*$ *and sources in* $\mathcal{S}^*$ *are conditionally independent given the sources in* $\mathcal{T}$.

Then for any proper subsets $\mathcal{R}\subset\mathcal{R}^*$ *and* $\mathcal{S}\subset\mathcal{S}^*$ *we have*

$$\gamma[\mathcal{R},\mathcal{S},\mathcal{T}] = -\gamma[\mathcal{R},\mathcal{S}]. \tag{2.22}$$

Proof. First, we note that we can rewrite Theorem 2.5 as

$$C_{\mathcal{R},\mathcal{T}} = \gamma[\mathcal{R}] + \gamma[\mathcal{T}] + \gamma[\mathcal{R},\mathcal{T}] \tag{2.23}$$

and similarly

$$C_{\mathcal{R},\mathcal{S},\mathcal{T}} = \gamma[\mathcal{R}] + \gamma[\mathcal{S}] + \gamma[\mathcal{T}] + \gamma[\mathcal{R},\mathcal{S}] + \gamma[\mathcal{R},\mathcal{T}] + \gamma[\mathcal{S},\mathcal{T}] + \gamma[\mathcal{R},\mathcal{S},\mathcal{T}]. \tag{2.24}$$

Now by Theorem 2.3 and under the presence of $[\mathcal{R}^*,\mathcal{T}]$ and $[\mathcal{S}^*,\mathcal{T}]$ in the description of the HJLLM, we have

$$\begin{aligned}C_{\mathcal{R},\mathcal{S},\mathcal{T}} &= C_{\mathcal{R},\mathcal{T}} + C_{\mathcal{S},\mathcal{T}} - C_{\mathcal{T}}\\ &= \gamma[\mathcal{R}] + \gamma[\mathcal{T}] + \gamma[\mathcal{R},\mathcal{T}] + \gamma[\mathcal{S}] + \gamma[\mathcal{T}] + \gamma[\mathcal{S},\mathcal{T}] - \gamma[\mathcal{T}]\end{aligned} \tag{2.25}$$

by (2.23). Equating (2.24) and (2.25) yields the result. □

Remark 2.26 In the simplest of cases, $\mathcal{R} = \{A\}$, $\mathcal{S} = \{B\}$ and $\mathcal{T} = \{C\}$. In this instance, if $[AC][BC]$ is the HJLLM description of the model (entailing conditional independence of A and B given C), the relationship $\gamma_{ABC} = -\gamma_{AB}$ must hold by (2.22). In other words the 3-way component of dependence exactly cancels out the 2-way component of dependence borne by the conditionally independent sources.

Corollary 2.2 *Let an HJLLM description be given by* $[\mathcal{T}_1^*][\mathcal{T}_2^*]\ldots[\mathcal{T}_k^*]$, *where* $\mathcal{T}_i^* \cap \mathcal{T}_j^* = \varnothing$ *for* $i \neq j$. *Let* $\mathcal{T}_i \subset \mathcal{T}_i^*$ *and* $\mathcal{T}_j \subset \mathcal{T}_j^*$. *Then* $\gamma[\mathcal{T}_i, \mathcal{T}_j] = 0$ *for* $i \neq j$.

Proof. Let $\mathcal{T} = \varnothing$. Then $\gamma[\mathcal{T}_i, \mathcal{T}_j, \mathcal{T}] = \gamma[\mathcal{T}_i, \mathcal{T}_j]$. But the descriptors $[\mathcal{T}_i^*, \mathcal{T}] = [\mathcal{T}_i^*]$ and $[\mathcal{T}_j^*, \mathcal{T}] = [\mathcal{T}_j^*]$ satisfy the conditions of Theorem 2.9 above. Therefore $\gamma[\mathcal{T}_i, \mathcal{T}_j, \mathcal{T}] = -\gamma[\mathcal{T}_i, \mathcal{T}_j]$ as well, which can only occur if $\gamma[\mathcal{T}_i, \mathcal{T}_j] = 0$. □

Remark 2.27 The above corollary states that the CIDs corresponding to sets of sources that exhibit independence will be zero. Dependence is not increased by jointly considering mutually independent sets of sources together.

Example 2.16 Three-source hierarchical dependence structures.
We now consider the three-source capture-recapture setting for sources A, B and C. We recall that there are 3 pairwise CIDs (equivalently CSDs) and a single three-way CID (equivalently CSD), whilst the 3 single-source CIDs (equivalently CSDs) are all identically equal to 0 The non-single source CSDs are not known, precisely because only a portion of the population is observed in a capture-recapture study.

We will show the following, where, for brevity in this list, 'joint model' refers to 'hierarchical joint log-linear model',

- For joint dependence structures (including independence), the marginal and joint model are equivalent with certain CSDs set equal to 0. Such a marginal model is thus a re-parameterization of the joint model.

- A conditional dependence structure corresponds to a marginal model with constraints placed on the CIDs. In such a case, the marginal model is not a reparameterization of the joint model.

In the four cases presented next, purporting to mutual independence, joint independence, conditional independence and mutual dependence, all models will be expressed in terms of N and probabilities (joint or marginal) rather than the corresponding parameterizations in terms of log-means (joint and marginal). This is in line with the descriptions outlined in Bishop et al. (1975) for complete contingency tables and enables us to gain compactness. Although N is unknown in the capture-recapture setting, all of the underlying theory for the form of hierarchical joint log-linear models for capture-recapture data is the same as that for complete contingency tables, again as stated in Bishop et al. (1975). For instance, the model for independence is expressed in the same manner for complete and incomplete contingency tables, with the only difference being the number of data points available from which to estimate model parameters.

As already noted in Chapter 1, there are some models which cannot be fitted to incomplete contingency tables. As always, it is possible to fit an identifiable model with at most as many parameters as data points. In the case of the incomplete contingency table for K-source capture-recapture data, the number of parameters cannot exceed $2^K - 1$, the number of available data points.

Here we present the four cases. All vectors and matrices are as defined in Example 2.12 and provided in Appendix A

Mutual Independence. Mutual independence implies that all subsets of the three sources exhibit independence. Such a dependence structure is represented as $[A][B][C]$ in the notation introduced in Christensen (1997) and Bishop et al. (1975) as described in Section 1.2.1. By Corollary 2.2, all CIDs, equivalently all CSDs, are identically zero. That is $\gamma_{AB} = \gamma_{AC} = \gamma_{BC} = \gamma_{ABC} = 0$.

The corresponding HJLLM for mutual independence (Bishop et al., 1975), expressed in terms of the six marginal probabilities $p_A = Pr[A]$, $p_B = Pr[B]$ and $p_C = Pr[C]$ and $p_{\bar{A}} = 1 - p_A$, $p_{\bar{B}} = 1 - p_B$ and $p_{\bar{C}} = 1 - p_C$, is given by the following model on the cell means $\boldsymbol{\mu}$:

$$\begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} = \begin{bmatrix} Np_Ap_{\bar{B}}p_{\bar{C}} \\ Np_{\bar{A}}p_Bp_{\bar{C}} \\ Np_{\bar{A}}p_{\bar{B}}p_C \\ Np_Ap_Bp_{\bar{C}} \\ Np_Ap_{\bar{B}}p_C \\ Np_{\bar{A}}p_Bp_C \\ Np_Ap_Bp_C \end{bmatrix}.$$

The equivalent expression in terms of the marginal means $\mathbf{m}$ is obtained by pre-multiplying $\boldsymbol{\mu}$ by the transformation matrix $\mathbf{A}$, so that $\mathbf{m} = \mathbf{A}\boldsymbol{\mu}$, to yield

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Np_Ap_{\bar{B}}p_{\bar{C}} \\ Np_{\bar{A}}p_Bp_{\bar{C}} \\ Np_{\bar{A}}p_{\bar{B}}p_C \\ Np_Ap_Bp_{\bar{C}} \\ Np_Ap_{\bar{B}}p_C \\ Np_{\bar{A}}p_Bp_C \\ Np_Ap_Bp_C \end{bmatrix} = \begin{bmatrix} N(p_Ap_{\bar{B}}p_{\bar{C}} + p_Ap_Bp_{\bar{C}} + p_Ap_{\bar{B}}p_C + p_Ap_Bp_C) \\ N(p_{\bar{A}}p_Bp_{\bar{C}} + p_Ap_Bp_{\bar{C}} + p_{\bar{A}}p_Bp_C + p_Ap_Bp_C) \\ N(p_{\bar{A}}p_{\bar{B}}p_C + p_Ap_{\bar{B}}p_C + p_{\bar{A}}p_Bp_C + p_Ap_Bp_C) \\ N(p_Ap_Bp_{\bar{C}} + p_Ap_Bp_C) \\ N(p_Ap_{\bar{B}}p_C + p_Ap_Bp_C) \\ N(p_{\bar{A}}p_Bp_C + p_Ap_Bp_C) \\ Np_Ap_Bp_C \end{bmatrix}$$

$$= \begin{bmatrix} N(p_Ap_{\bar{C}} + p_Ap_C) \\ N(p_Bp_{\bar{C}} + p_Bp_C) \\ N(p_{\bar{B}}p_C + p_Bp_C) \\ N(p_Ap_B(p_{\bar{C}} + p_C))) \\ N(p_Ap_C(p_{\bar{B}} + p_B)) \\ N(p_Bp_C(p_{\bar{A}} + p_A)) \\ Np_Ap_Bp_C \end{bmatrix} = \begin{bmatrix} Np_A \\ Np_B \\ Np_C \\ Np_Ap_B \\ Np_Ap_C \\ Np_Bp_C \\ Np_Ap_Bp_C \end{bmatrix}.$$

The final expression is precisely that of the marginal model for independence given by Definition 2.7 with all CIDs set to 0. Thus, the joint log-linear model and marginal log-linear model forms are equivalent for the case of independence and are simply two different parameterizations of the same model. Note that N is indeed unknown in the case of capture-recapture data but the form of the model is unaffected by the fact that one data point of the contingency table is missing.

Joint Independence. Without loss of generality, suppose that sources A and B are dependent but that there is no other dependence exhibited between the three sources. Such a structure is given by $[AB][C]$ (Christensen, 1997). By Corollary 2.2, $\gamma_{AC} = \gamma_{BC} = 0$. Equivalently, $C_{AC} = C_{BC} = 0$, whereas $\gamma_{AB} = C_{AB} \neq 0$ since sources A and B are dependent, by assumption. Furthermore, $\gamma_{ABC} = 0$, by Corollary 2.2. Thus $C_{ABC} = C_{AB}$, as $C_{ABC} = \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \gamma_{ABC}$ by Theorem 2.5. Thus 3-way dependence, is completely explained by 2-way dependence.

The corresponding HJLLM for joint independence (Bishop et al., 1975), expressed in terms of the five marginal probabilities $p_{AB} = P[A \cap B]$, $p_{\bar{A}B} = P\left[\bar{A} \cap B\right]$, $p_{A\bar{B}} = P\left[A \cap \bar{B}\right]$, $p_C = P[C]$ and $p_{\bar{C}} = 1 - p_C$, is given by the following model on the cell means $\boldsymbol{\mu}$:

$$\begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} = \begin{bmatrix} Np_{A\bar{B}}p_{\bar{C}} \\ Np_{\bar{A}B}p_{\bar{C}} \\ Np_{\bar{A}\bar{B}}p_C \\ Np_{AB}p_{\bar{C}} \\ Np_{A\bar{B}}p_C \\ Np_{\bar{A}B}p_C \\ Np_{AB}p_C \end{bmatrix}.$$

The equivalent expression in terms of the marginal means $\mathbf{m}$ is obtained using $\mathbf{m} =$

$\mathbf{A}\boldsymbol{\mu}$, to yield

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Np_{A\bar{B}}p_{\bar{C}} \\ Np_{\bar{A}B}p_{\bar{C}} \\ Np_{\bar{A}\bar{B}}p_{C} \\ Np_{AB}p_{\bar{C}} \\ Np_{A\bar{B}}p_{C} \\ Np_{\bar{A}B}p_{C} \\ Np_{AB}p_{C} \end{bmatrix} = \begin{bmatrix} N(p_{A\bar{B}}p_{\bar{C}} + p_{AB}p_{\bar{C}} + p_{A\bar{B}}p_{C} + p_{AB}p_{C}) \\ N(p_{\bar{A}B}p_{\bar{C}} + p_{AB}p_{\bar{C}} + p_{\bar{A}B}p_{C} + p_{AB}p_{C}) \\ N(p_{\bar{A}\bar{B}}p_{C} + p_{A\bar{B}}p_{C} + p_{\bar{A}B}p_{C} + p_{AB}p_{C}) \\ N(p_{AB}p_{\bar{C}} + p_{AB}p_{C}) \\ N(p_{A\bar{B}}p_{C} + p_{AB}p_{C}) \\ N(p_{\bar{A}B}p_{C} + p_{AB}p_{C}) \\ Np_{AB}p_{C} \end{bmatrix}$$

$$= \begin{bmatrix} Np_{A}(p_{\bar{C}} + p_{C}) \\ Np_{B}(p_{\bar{C}} + p_{C}) \\ Np_{C}(p_{\bar{B}} + p_{B}) \\ Np_{AB}(p_{\bar{C}} + p_{C}) \\ Np_{A}p_{C} \\ Np_{B}p_{C} \\ Np_{AB}p_{C} \end{bmatrix} = \begin{bmatrix} Np_{A} \\ Np_{B} \\ Np_{C} \\ Np_{AB} \\ Np_{A}p_{C} \\ Np_{B}p_{C} \\ Np_{AB}p_{C} \end{bmatrix}.$$

The final expression is precisely that of the marginal model for joint independence given by Definition 2.7 with all CIDs set to 0 except for γ_{BC}, whose form is not constrained to be set to 0 since there is dependence between sources A and B, as described above. Thus, the joint log-linear model and marginal log-linear model forms are equivalent for the case of joint independence and are simply two different parameterizations of the same model. Again, note that N is indeed unknown in the case of capture-recapture data but the form of the model is unaffected by the fact that one data point of the contingency table is missing.

Conditional Independence Without loss of generality, suppose that sources B and C are independent, conditionally on membership in source A. Such a structure is given by $[AB][AC]$ (Christensen, 1997) and implies that A and B are pairwise dependent,

as are A and C. In this case $\gamma_{AB} \neq 0$ and $\gamma_{AC} \neq 0$ (equivalently $C_{AB} \neq 0$ and $C_{AC} \neq 0$). Moreover, since conditional independence of B and C given A does not imply marginal independence of B and C, it follows that $\gamma_{BC} \neq 0$ (equivalently $C_{BC} \neq 0$). We show in Appendix D, page 231, that C_{BC} is completely determined by the CSDs C_{AB} and C_{AC} and the marginal probability $Pr[A]$ (equivalently γ_{BC} is completely determined by the CIDs γ_{AB} and γ_{AC} and the marginal probability $Pr[A]$), since

$$c_{BC} = \frac{Pr[A]}{1 - Pr[A]} \left(c_{AB}c_{AC} - c_{AB} - c_{AC} + \frac{1}{Pr[A]} \right).$$

For the three-way CID, we get

$$\gamma_{ABC} = -\gamma_{AB} \tag{2.26}$$

from Theorem 2.9, just as Theorem 2.3 yields the equivalent relationship for the three-way CSD:

$$C_{ABC} = C_{AB} + C_{AC} - C_A, \tag{2.27}$$

stated equivalently as $C_{ABC} = C_{AB} + C_{AC}$, from Remark 2.3.

Therefore, a MLLM for conditional independence of $B \perp\!\!\!\perp C|A$ would require that the constraint given by (2.26) (equivalently (2.27)) be enforced. Consequently, unlike the cases of mutual independence and joint independence, the marginal model equivalent to the HJLLM of conditional independence is a constrained MLLM. Thus, the MLLM is not a direct reparameterization of the HJLLM with some CIDs set to zero.

We note that that the HJLLM for conditional independence is expressed as (Bishop et al., 1975)

$$\begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} = \begin{bmatrix} N\frac{p_{A\bar{B}}p_{A\bar{C}}}{p_A} \\ N\frac{p_{\bar{A}B}p_{\bar{A}\bar{C}}}{p_{\bar{A}}} \\ N\frac{p_{\bar{A}\bar{B}}p_{\bar{A}C}}{p_{\bar{A}}} \\ N\frac{p_{AB}p_{A\bar{C}}}{p_A} \\ N\frac{p_{A\bar{B}}p_{AC}}{p_A} \\ N\frac{p_{\bar{A}B}p_{\bar{A}C}}{p_{\bar{A}}} \\ N\frac{p_{AB}p_{AC}}{p_A} \end{bmatrix},$$

whose equivalent marginal form is expressed as (details not provided)

$$\begin{bmatrix} Np_A \\ Np_B \\ Np_C \\ Np_{AB} \\ Np_{AC} \\ N(\frac{p_{\bar{A}B}p_{\bar{A}C}}{p_{\bar{A}}} + \frac{p_{AB}p_{AC}}{p_A}) \\ N\frac{p_{AB}p_{AC}}{p_A} \end{bmatrix},$$

using $\mathbf{m} = \mathbf{A}\boldsymbol{\mu}$.

Mutual Dependence In this case, represented by $[ABC]$ there is three-way dependence and none of the CSDs, neither the pairwise nor the three-way, are necessarily equal to 0. Such a dependence structure can be represented by an HJLLM but, as discussed in Section 1.1.1, cannot be uniquely estimated since there are insufficient degrees of freedom.

Remark 2.28 The previous example serves to make explicit the relationships between HJLLMs and MLLMs. For simple dependence structures, the MLLMs equivalent to hierarchical joint log-linear models are obtained by constraining certain CIDs

(equivalently CSDs) to be equal to 0. Such a relationship is intuitive. Given that the CIDs (CSDs) measure departures from independence it seems natural that setting a CSD equal to 0 in the marginal model is similar to omitting the corresponding interaction term of the hierarchical joint log-linear model. However, HJLLMS of conditional independence correspond to specific linear constraints on the CIDs (CSDs).

Example 2.17 Four-source hierarchical dependence structures.
We now examine the dependence structures that can be represented by hierarchical joint log-linear models among four sources, A, B, C and D. Although the simpler dependence structures are exactly those presented for three sources, we will describe each fully for completeness. We will present results in terms of the CSDs only. The corresponding results for the CIDs are obtained by Theorem 2.9, Corollary 2.2 or Definition 2.4, whilst the corresponding marginal model forms can be obtained using the same approach as in the previous example. We note that the three-way CSDs are expressed in terms of the three-way CIDs as follows (Theorem 2.5):

$$\begin{aligned} C_{ABC} &= \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \gamma_{ABC} \\ C_{ABD} &= \gamma_{AB} + \gamma_{AD} + \gamma_{BD} + \gamma_{ABD} \\ C_{ACD} &= \gamma_{AC} + \gamma_{AD} + \gamma_{CD} + \gamma_{ACD} \\ C_{BCD} &= \gamma_{BC} + \gamma_{BD} + \gamma_{CD} + \gamma_{BCD}, \end{aligned}$$

with the four-way CSD given by

$$\begin{aligned} C_{ABCD} = \gamma_{AB} + \gamma_{AC} + \gamma_{AD} + \gamma_{BC} + \gamma_{BD} + \gamma_{CD} \\ + \gamma_{ABC} + \gamma_{BCD} + \gamma_{ACD} + \gamma_{BCD} + \gamma_{ABCD}. \end{aligned} \tag{2.28}$$

Mutual Independence. Mutual independence, represented by $[A][B][C][D]$, implies that all subsets of the four sources exhibit independence. Thus, all of the 6 pairwise,

4 three-way and the single four-way CSDs are exactly equal to 0.

Joint Independence.

Case 1 Without loss of generality, suppose that sources A and B are dependent but no other dependence is exhibited. Such a structure is denoted by $[AB][C][D]$. Then $C_{AB} \neq 0$, whereas $C_{AC} = C_{AD} = C_{BC} = C_{BD} = C_{CD} = 0$. As seen for the three-source case above

$$C_{ABC} = C_{AB} \quad \text{and} \quad C_{ABD} = C_{AB}.$$

Further,

$$C_{ABCD} = C_{AB}.$$

The remaining three-way CSDs are equal to 0,

$$C_{ACD} = C_{BCD} = 0.$$

Case 2 Another case of joint independence structure occurs with structure $[AB][CD]$. That is, assume without loss of generality that sources A and B are jointly independent of sources C and C. Then, using similar reasoning to the previous case,

$$C_{ABC} = C_{ABD} = C_{AB}.$$

Likewise,

$$C_{ACD} = C_{BCD} = C_{CD},$$

and

$$C_{ABCD} = C_{AB} + C_{CD}.$$

Conditional Independence. Without loss of generality, suppose that C and D are independent, conditionally on membership in sources A and B. Such a structure is denoted by $[ABC][ABD]$. As with the example of conditional independence for the

three-source case above, conditional independence of C and D given A and B, does not imply that C and D are marginally independent. Thus, it is not possible to say that $C_{CD} = 0$. In this case none of the other pairwise CSDs are equal to 0 since, by the principle of hierarchy for such a model, all other pairs must be dependent. Moreover, the 2 three-way CSDs, C_{ABC} and C_{ABD}, are not equal to 0 since, by assumption, each of these triples exhibits dependence.

For the other 2 three-way CSDs, C_{ACD} and C_{BCD}, it is not possible to obtain a direct relationship to any of the known CSDs. This can be seen by expanding from the definition of c_{ACD}, (2.1), in the following manner

$$c_{ACD} = \frac{Pr[A \cap C \cap D]}{Pr[A]Pr[C]Pr[D]} = \frac{Pr[C \cap D|A]Pr[A]}{Pr[A]Pr[C]Pr[D]} = \frac{Pr[C \cap D|A]}{Pr[C]Pr[D]}.$$

C and D are not necessarily conditionally independent given A, and the conditional distributions of $C|A$ and $D|A$ are arbitrary. The same reasoning applies to C_{BCD}.

For the four-way CSD, Theorem 2.3 states that

$$C_{ABCD} = C_{ABC} + C_{ABD} - C_{AB}. \tag{2.29}$$

As with the three-source case of conditional independence above (i.e.[AB][AC] for sources A, B and C), Theorem 2.3 is used to decompose the highest order CSD. With this four-source example, we note that conditioning is made on two sources rather than the single source in the three-source example and the joint effect of both of the two conditioning sources must be accounted for.

If sources A and B are assumed independent, equivalently $c_{AB} = 1$ and $C_{AB} = 0$, then we obtain

$$C_{ABCD} = C_{ABC} + C_{ABD},$$

from Theorem 2.3. That is, the four-way CSD is equal to the sum of the 2 three-way CSDs, C_{ABC} and C_{ABD}, when there is also pairwise independence of A and B.

Mutual Dependence. In this case there is four-way dependence and none of the CSDs are equal to 0. Again, as in the three-source case, such a set-up can be modelled by a hierarchical log-linear model at the cost of over-parameterization.

Remark 2.29 As with the three-source case given in Example 2.16, the previous example serves to make explicit the relationships between HJLLMs and MLLMs. For simple dependence structures, the MLLMs equivalent to hierarchical joint log-linear models are obtained by constraining certain CIDs (equivalently CSDs) to be equal to 0. However, HJLLMs of conditional independence correspond to specific linear constraints on the CIDs (CSDs) and thus the MLLM equivalent to the HJLLM is a constrained model and not a mere reparameterization of the HJLLM as is the case for simple dependence structures.

Remark 2.30 Nonhierarchical dependence. With the discussion of hierarchical dependence structures of the current section, we close by considering nonhierarchical dependence structures. On page 232 of Appendix D we provide an example of three events that are pairwise independent but jointly dependent, thus demonstrating that nonhierarchical dependence can occur. In the capture-recapture setting, a possible scenario like that of the example would arise for three lists should the probability that an individual appear on a list depends on his/her membership on both of the 2 other lists, whilst knowing only that an individual appears on one such list provides no information concerning his/her appearance on each of the other lists separately.

It is known that HJLLMs, designed to accommodate hierarchical dependence structures, do not necessarily well model nonhierarchical dependence. We suggest that the class of MLLMs offer greater flexibility in modelling such structures. A data analysis of data generated according to a nonhierarchical dependence structure will be presented in the next chapter.

Summary

We will provide evidence as to the flexibility of the marginal model in the following chapters, in which we undertake parameter estimation and inference. A frequentist approach will be presented in Chapter 3. In Chapter 4 a Bayesian development for a general model form will be presented, followed by a specific model parameterization in Chapter 5.

Chapter 3

Frequentist Marginal Log-Linear Models

3.1 Introduction

In this chapter we consider the general form of the marginal log-linear model introduced in Chapter 2 (see Definition 2.7). We describe a family of marginal models which consists of all those derived from the general form by fixing different combinations of CIDs (or CSDs) to zero. We follow with a description of parameter estimation via maximum likelihood, which includes reference to the challenges posed by fitting a model on marginal means using a likelihood on joint cell means.

Data analysis is performed for 3 four-source data sets: the diabetes data set of Bruno et al. (1994) and two simulated data sets. Of these latter, the first is generated according to a conditional independence model and the second according to a nonhierarchical dependence structure. The data analysis consists of fitting all possible joint log-linear models, both hierarchical and nonhierarchical, as well as all possible

marginal log-linear models, and to compare the models using simple model selection criterion. The relationship to joint log-linear models will be presented and the manner in which marginal log-linear models account for dependence structures similarly and differently to hierarchical joint log-linear models will be described. These examples will serve to provide concrete empirical evidence for the properties of the CIDs and marginal log-linear models described in Chapter 2.

The work presented in this chapter is not intended to be the definitive frequentist formulation. Rather the goal of this work is twofold: first, to demonstrate that it is possible to obtain reasonable maximum likelihood estimates using the CID model formulation; and, second, to begin to explore dependence structures which are not well modelled by hierarchical log-linear models and demonstrate that our model thus parameterized is able to out-perform both the best-performing hierarchical log-linear model and nonhierarchical log-linear model (see Section 3.5.4).

With these goals in mind, this chapter presents ideas to aid exploration and understanding whilst still offering a possible route to analysing capture-recapture data. It should be noted that in demonstrating that the marginal model performs better than the best hierarchical model for certain types of capture-recapture data, we provide weight to the argument in favour of using non-hierarchical joint log-linear models in capture-recapture as well as more general settings. Before presenting the results of the analysis of such data, we begin with a discussion of the model and the inferential procedure to be used, that is maximum likelihood estimation.

3.2 The marginal log-linear model

The most general form of the marginal log-linear model introduced in Chapter 2 is given by Definition 2.7 for a set of sources $\mathcal{Q}$. We include the definition here again for completeness.

Definition 3.1 *Let $\mathcal{Q}$ be a set of sources and let $K = |\mathcal{Q}|$. Then the following system of $2^K - 1$ equations constitutes the marginal log-linear model for the K sources of $\mathcal{Q}$.*

$$\begin{aligned} &\log m_S = \beta_S, \ \ \textit{for } S \in \mathcal{Q}; \\ &\log m_{\mathcal{R}} = -(|\mathcal{R}| - 1)\log N + \sum_{S \in \mathcal{Q}} \log m_S + \sum_{\mathcal{T} \subset \mathcal{R}} \gamma_{\mathcal{T}}, \ \ \textit{for } \mathcal{R} \subset \mathcal{Q}. \end{aligned} \tag{3.1}$$

Equivalently, the formulation can be re-expressed in terms of the CSDs using Proposition 2.8, again presented here for completeness.

Proposition 3.1 *Let $\mathcal{Q}$ be a set of sources and let $K = |\mathcal{Q}|$. Then the following system of $2^K - 1$ equations is equivalent to the marginal log-linear model for $\mathcal{Q}$ as given by Definition 3.1.*

$$\begin{aligned} &\log m_S = \beta_S, \ \ \textit{for } S \in \mathcal{Q} \\ &\log m_{\mathcal{R}} = -(|\mathcal{R}| - 1)\log N + \sum_{S \in \mathcal{Q}} \log m_S + C_{\mathcal{R}}, \ \ \textit{for } \mathcal{R} \subset \mathcal{Q} \end{aligned} \tag{3.2}$$

Example 3.1 Consider the three-source capture-recapture setting, with sources A, B and C. Then the following system of $2^3 - 1 = 7$ linear equations constitutes the three-source marginal model parameterized in terms of the CIDs:

$$\begin{aligned} \log m_A &= \beta_A \\ \log m_B &= \beta_B \\ \log m_C &= \beta_C \\ \log m_{AB} &= -\beta_0 + \beta_A + \beta_B + \gamma_{AB} \\ \log m_{AC} &= -\beta_0 + \beta_A + \beta_C + \gamma_{AC} \\ \log m_{BC} &= -\beta_0 + \beta_B + \beta_C + \gamma_{BC} \\ \log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \gamma_{ABC}, \end{aligned} \tag{3.3}$$

whilst the equivalent formulation in terms of the CSDs is given by:

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + C_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C + C_{AC} \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C + C_{BC} \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + C_{ABC},
\end{aligned} \tag{3.4}$$

•

3.3 Family of marginal models and parameter reduction

Both the CIDs and CSDs measure dependence. As described in the previous chapter, the manner in which they achieve this differs. In the case of the CSDs (see Definition 2.2), dependence is measured as a departure from marginal independence. For CIDs (see Definition 2.4), it is measured as the incremental dependence injected by a subset of sources into a set of sources. For instance, a three-way CID measures the additional dependence in the three corresponding sources not accounted for by all of the 3 pairs of sources marginally.

Setting CIDs (CSDs) equal to zero in the CID parameterization (3.1) (the CSD parameterization (3.4)), alters the form of the dependence structure being modelled. In so doing, the number of parameters to be estimated is reduced. Setting CIDs (CSDs) equal to zero is equivalent to omitting the corresponding term in the model and can be thought of in a similar manner to that of omitting interaction terms in

a joint log-linear model. As described in the previous chapter, marginal models are able to model dependence differently to hierarchical joint log-linear models; first, since they are not constrained to adhere to the principle of hierarchy (Bishop et al., 1975), and, secondly, since the form of the model necessarily implies that the modelling is done so differently.

In working with the CID parameterization rather than the CSD parameterization, given for the three-source case by (3.3) and (3.4), respectively, we gain modelling flexibility. Whenever a CSD is fixed at 0, the equation for the corresponding marginal mean is the same as the model for independence. However, such a situation does not arise under the CID MLLM formulation wherein, for models of dimension greater or equal to 3, it is possible to set some CIDs equal to zero whilst retaining other non-zero CID terms. In so doing, it is possible to retain some terms that model dependence rather than necessarily removing all dependence terms in the case of the CSD formulation. We illustrate this point with the three-source case.

Example 3.2 Consider the three-source case, for sources A, B and C. We place the following constraint on the CSDs in (3.4): $C_{AC} = C_{BC} = C_{ABC} = 0$. Then the model is given by

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + C_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C, \qquad (3.5)
\end{aligned}$$

which is the model for mutual dependence of A and B. However, if $\gamma_{AC} = \gamma_{BC} =$

$\gamma_{ABC} = 0$ in the CID parameterization (3.3), the corresponding model is not that of mutual dependence of A and B, rather it is given by

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + \gamma_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + \gamma_{AB},
\end{aligned}$$

In this case, there is a term that measures dependence in the expression for m_{ABC}, the three-way marginal (equivalently joint) mean. In both cases the number of parameters to be estimated is reduced by 3 but the nature of the modelled dependence is different. The CID formulation equivalent to (3.5) would impose the constraint $\gamma_{ABC} = -\gamma_{AB}$ (as per the discussion in Section 2.3). Further, we note that such a model is not that of conditional independence of $B \perp\!\!\!\perp A|C$, which would have non-zero CIDs γ_{AC} and γ_{BC}, as per the setting presented in Example 2.16.

The previous example serves to demonstrate that a model in which CSDs are set to zero is less plausible than a model setting CIDs equal to zero, since every combination of lists is given its own offset term, as it were, in the form of a CSD. We therefore focus our discussion on the CID parameterization, which we employ for data analysis in Sections 3.5.2 - 3.5.4.

Remark 3.1 The general form of the CID formulation of the marginal model (and also the CSD formulation) is over-parameterized. For the general K-source capture-recapture setting, there are 2^K parameters to estimate, corresponding to $(K+1)$ β terms (i.e. $\beta_0, \beta_A, \ldots, \beta_K$) and $(2^K - 1) - K$ CIDs (CSDs). However, there are

only $2^K - 1$ data points, corresponding to the $2^K - 1$ cell entries of the incomplete contingency table (given by Table 1.2, for the three-source case), available to estimate these parameters. Thus, the system is under-determined with insufficient data to simultaneously estimate all model parameters.

Parameter reduction must be undertaken in order to achieve model identifiability. Even with a parameter reduction of 1, the number of parameters to estimate would equal the number of data points to be used for this estimation. That is, the model would be saturated. In this situation, as with any statistical model which seeks to fit a model to a data set with as many parameters as data points, the maximum likelihood estimation procedure used would yield model parameter estimates which fit to the data perfectly. In the capture-recapture setting, this translates into estimated model parameters which yield estimated cell means that are equal to the observed cell counts.

Unsaturated models can be obtained by setting more than a single CID (CSD) equal to zero. In fact, a family of marginal models exists in which we consider all possible combinations of CIDs (CSDs) that are fixed at zero and the marginal models that correspond to these combinations of zero-valued CIDs. The following proposition specifies the number of possible models in this family, for which all main effect terms corresponding to the single-source marginal means are included. This is a specific feature of capture-recapture data: since the goal is to use all of the K data sources available in the general case, main effect terms should be included for all of these K sources. This principle will be adhered to henceforth.

Proposition 3.2 *Let $\mathcal{Q}$ denote a set of sources and $K = |\mathcal{Q}|$. Let s equal the number of non-single source CIDs (CSDs) of $\mathcal{Q}$. Then there are s possible CID (CSD) parameters of model* (3.1) *(model* (3.2)*) to be estimated, where*

$$s = (2^K - 1) - K.$$

Furthermore

1. *The number of marginal models with all possible combinations of CIDs (CSDs) fixed at zero is given by* 2^s.

2. *There are* $2^s - 1$ *models that are not over-parameterized.*

3. *Of these, there are* $2^s - (s+1)$ *unsaturated models.*

Proof. The proof is immediate from simple combinatorial arguments.

1. The number of possible models with all combinations of s CIDs (CSDs) fixed at zero, ranging from 0 CIDs (CSDs) to all s CIDs (CSDs), is given by
$$\binom{s}{0} + \binom{s}{1} + \binom{s}{2} + \cdots + \binom{s}{s-1} + \binom{s}{s} = 2^s.$$

2. Since there are $2^K - 1$ data points, the number of model parameters cannot exceed this value. The only model with more than $2^K - 1$ parameters is that with none of the s CIDs (CSDs) fixed at zero (i.e. the model that includes all CIDs (CSDs)). Thus, there are $2^s - 1$ models that are not over-parameterized.

3. Unsaturated models contain fewer parameters than the number of data points with which to estimate the parameters. Of the $2^s - 1$ models that are not over-parameterized, we remove the $\binom{s}{1} = s$ models with one of the s CIDs (CSDs) fixed at zero to yield $2^s - (1+s)$.

□

Example 3.3 For the three-source case there are $s = (2^3 - 1) - 3 = 4$ CIDs. Thus, using Proposition 3.1, 16 unique models exist. Of those 16 models, 1 of them is over-parameterized (i.e. the model parameters cannot be estimated uniquely). Of the 15 models that are not over-parameterized, 4 of them are saturated (i.e. there are as many data points as parameters to estimate). Thus, there are 11 unsaturated models to be considered.

Example 3.4 For the four-source case there are $s = (2^4 - 1) - 4 = 11$ CIDs and consequently many more possible marginal models than in the three-source case. Using Proposition 3.1, 2048 unique models exist. Of those 2048 models, 1 of them is over-parameterized (i.e. the model parameters cannot be estimated uniquely). Of the 2047 models that are not over-parameterized, 11 are saturated (i.e. there are as many data points as parameters to estimate). Thus, there are 2036 unsaturated models to be considered.

Remark 3.2 Note that in the case of HJLMMs there is a single saturated model for the general K-source case, whereas in considering non-hierarchical models, as we do here with the marginal models, there are $\binom{K}{1} = K$ such models, which correspond to the number of ways in which one of the CIDs can be fixed at zero.

Before being able to select amongst a set of models, it is necessary to specify the family amongst which we must select. Sensibly, it is only possible to consider the family that consists of all marginal models which are not over-parameterized. Thus, for the four-source data sets analyzed in this chapter, we consider the family of 2047 marginal models as described in the previous example, Example 3.4. In order to perform data analysis using this family of non-over parameterized marginal models, we must select the parameter estimation procedure to be used. In the next section we describe the maximum likelihood approach we choose to use, as well as the Fisher Scoring algorithm we employ to obtain the corresponding parameter estimates.

3.4 Maximum likelihood estimation

In general, maximum likelihood estimates can be obtained by using Fisher Scoring when there are no constraints on the parameter space. In working with a model on the marginal means $\mathbf{m}$, it is necessary to ensure that the corresponding cell means

$\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{m}$, are positive (for a description of the matrix $\mathbf{A}$, which transforms cell means into marginal means, see the matrix form below, as first introduced in (2.20)). Such constraints on the $2^K - 1$ cell means place corresponding non-linear constraints on the model parameters and define the feasible parameter space. In practice, again since the goal of this chapter is not to provide a definitive approach to maximum likelihood estimation for the best model formulation, it is possible to use Fisher Scoring starting the algorithm from a value known to be reasonably well inside the feasible parameter space. Indeed, when we used estimates obtained by fitting the joint log-linear model for independence in R (*R Programming Language*, 2004) (by using the inbuilt `glm` function with log link and a Poisson likelihood) as our starting values, the Fisher Scoring algorithm moved out of the feasible parameter space in very few instances. In situations where it did, an adjustment was made to move the parameter estimates to the nearest feasible point in the direction of maximum change as used in the Fisher Scoring algorithm. This was not difficult to implement and satisfied the exploratory goal of this chapter.

Implementation of the Fisher Scoring algorithm requires both the score vector and information matrix of the likelihood together with the model as specified by (3.1). Given the relationship between the multinomial and Poisson likelihoods outlined in Section 1.1.3 and described in Sandland and Cormack (1984), combined with the fact that the score and information matrix of the Poisson likelihood are much easier to deal with than those of the multinomial likelihood, we choose to work with the Poisson likelihood in this chapter.

It is useful to re-express model (3.1) in matrix form in order to obtain easily readable expressions for the score and information matrix. For the general case of K sources, let $d = 2^K - 1$ be the number of entries in the incomplete contingency table and let $s = d - K$ be the number of possible non-single source CIDs to be estimated (Recall that some can be fixed at zero, equivalently omitted from the model, to reduce

the model dimension). Let $\mathbf{A}$ be the $d \times d$ matrix that transforms the cell means into marginal means, $\boldsymbol{\mu}$ the $d \times 1$ vector of cell means, $\mathbf{m}$ the $d \times 1$ vector of marginal means, $\mathbf{X}$ the $d \times (K+1)$ design matrix for the $(K+1) \times 1$ vector $\boldsymbol{\beta}$ and $\mathbf{Z}$ the $d \times p$ design matrix for the $p \times 1$ vector of CIDs to be estimated, which is denoted by $\boldsymbol{\gamma}$, where $p \leq s-1$. For the K sources $S_1, \ldots, S_K$,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{S_1} \\ \vdots \\ \beta_{S_K} \end{bmatrix}. \tag{3.6}$$

As described in Chapter 2, the matrix form of the marginal log-linear model (3.1), is given by

$$\log \mathbf{m} = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}.$$

For convenience, this model can be expressed in the following equivalent form as

$$\log \mathbf{m} = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{Y}\boldsymbol{\delta}, \tag{3.7}$$

where

$$\boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} \mathbf{X} : \mathbf{Z} \end{bmatrix}.$$

We use q to denote the dimension of $\boldsymbol{\delta}$ with $q = (K+1)+p$, where p is the number of non-zero CIDs in the model. Then we require $q \leq d$ or equivalently $p \leq d-(K+1) = (2^K-1)-(K+1) = s-1$, where we recall that s is the number of non-single source CIDs. The matrix $\mathbf{Y}$ is the $d \times q$ design matrix corresponding to the q-dimensional vector $\boldsymbol{\delta}$.

Remark 3.3 As noted in Remark 2.24, we recognize that model (3.7) is a member of the class of models described by Lang and Agresti (1994) to simultaneously model

the joint and marginal distributions of multivariate categorical responses. Again we note that the setting of Lang and Agresti (1994) applies to complete contingency tables rather than the capture-recapture setting.

Example 3.5 In particular, for three sources, $K = 3$, $d = 2^K - 1 = 7$ and $s = 4$. The most general model (i.e. with no CIDs fixed at zero) is over-parameterized and is given by

$$\log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{Y}\boldsymbol{\delta},$$

where

$$\mathbf{m} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix}; \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix} \mathbf{Y} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ -2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \boldsymbol{\delta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \\ \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{BC} \\ \gamma_{ABC} \end{bmatrix}.$$

Remark 3.4 Reduced models are obtained by removing appropriate columns from the design matrix $\mathbf{Y}$ and the corresponding CID parameters in the parameter vector

$\boldsymbol{\delta}$. As mentioned above, we require $q \leq d$ in order for the model parameters to be estimable. The general model given by (3.7) is over-parameterized whenever $q > d$.

Example 3.6 For three sources, the model that contains only the three-way CID is given by

$$\log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{Y}\boldsymbol{\delta},$$

where

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ -2 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \boldsymbol{\delta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \\ \gamma_{ABC} \end{bmatrix}.$$

Then $q = 5$ for $d = 7$ data points.

Remark 3.5 All of the theory derived below applies to a general form of $\mathbf{Y}$ and $\boldsymbol{\delta}$, for $q \leq d$, where q is the dimension of the parameter vector $\boldsymbol{\delta}$.

Likelihood

The Poisson log-likelihood of $\boldsymbol{\delta}$ can be expressed as

$$l(\delta) = \mathbf{n}' \log[\boldsymbol{\mu}(\boldsymbol{\delta})] - \mathbf{e}'\boldsymbol{\mu}(\boldsymbol{\delta}) = \mathbf{n}' \log[\mathbf{A}^{-1} \exp(\mathbf{Y}\boldsymbol{\delta})] - \mathbf{e}'\boldsymbol{\mu}(\boldsymbol{\delta}), \tag{3.8}$$

where $\mathbf{n}$ is the $d \times 1$ vector of observed cell counts and $\mathbf{e}$ is a $d \times 1$ vector of 1s. The corresponding $q \times 1$, $q \leq d$, score vector is given by

$$U(\boldsymbol{\delta}) = \mathbf{Y}'\text{diag}(\exp(\mathbf{Y}\boldsymbol{\delta}))(\mathbf{A}^{-1})' \left[\mathbf{n} \circ \boldsymbol{\mu}(\boldsymbol{\delta})^{-I} - \mathbf{e}\right]$$

using the conventions described in Appendix B (with exponentiation of a matrix applied element-wise, as introduced by Gentleman & Vandal, 2001) whilst the negative $q \times q$ information matrix is given by

$$-I(\boldsymbol{\delta}) = \mathbf{Y}'\mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\left(\mathbf{D}_{(\mathbf{A}^{-1})'(\mathbf{n}\circ\boldsymbol{\mu}^{-I}-\mathbf{e})} - (\mathbf{A}^{-1})'\mathbf{D}_{\mathbf{n}\circ\boldsymbol{\mu}^{-2I}}\mathbf{A}^{-1}\mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right)\mathbf{Y}, \qquad (3.9)$$

where, for ease of notation, dependence of $\boldsymbol{\mu}$ on $\boldsymbol{\delta}$ has been suppressed and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\delta})$. The derivations of the score vector and information matrix are reproduced from Vandal et al. (2005) in Appendix F.

The Newton-Raphson algorithm uses the following scheme to update the $\boldsymbol{\delta}$ parameters at the sth step from a value of $\boldsymbol{\delta}_s$:

$$\boldsymbol{\delta}_{s+1} = \boldsymbol{\delta}_s + I(\boldsymbol{\delta}_s)^{-1}U(\boldsymbol{\delta}_s),$$

starting from an initial value denoted by $\boldsymbol{\delta}_0$. Fisher scoring, a well-known likelihood maximization method, replaces the observed information matrix $I(\boldsymbol{\delta})$ with its expectation $\mathbb{E}[I(\boldsymbol{\delta})]$, thereby stabilizing the algorithm. Expectation is taken with respect to the distribution of the vector of cell counts $\mathbf{n}$, which is assumed to be a vector of independent Poisson random variables with expectation $\boldsymbol{\mu}$, as per (3.8). Taking expectations entry-wise in (3.9), we see that each matrix entry is a linear function of the cell counts of the $d \times p$ matrix, $I(\boldsymbol{\delta})$. Thus, we have a linear function on each entry from $\mathbf{n}$ and the expectation operator can be applied directly to each entry of $\mathbf{n}$. Specifically we can apply $\mathbb{E}[\mathbf{n}] = \boldsymbol{\mu}$ to (3.9) to obtain

$$-\mathbb{E}[I(\boldsymbol{\delta}_s)] = -\mathbf{Y}'\mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}(\mathbf{A}^{-1})'\mathbf{D}_{\boldsymbol{\mu}^{-I}}\mathbf{A}^{-1}\mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\mathbf{Y}. \qquad (3.10)$$

The Fisher scoring iteration is given by

$$\boldsymbol{\delta}_{s+1} = \boldsymbol{\delta}_s + \mathbb{E}[I(\boldsymbol{\delta}_s)]^{-1}U(\boldsymbol{\delta}_s),$$

As alluded to earlier in this section, there are additional challenges posed by fitting a model on marginal means using a likelihood on the cell means. Positivity

of the cell means must be ensured as Fisher Scoring updates $\boldsymbol{\delta}$. Ensuring positivity of $\boldsymbol{\mu}$ corresponds to an ordering of the marginal probabilities of $\mathbf{m}$ as outlined in the previous chapter in Section 2.4.1. These constraints are non-linear in model parameters, $\boldsymbol{\delta}$, which implies that maximum likelihood estimation must be performed subject to the d non-linear constraints given by

$$\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{m} \geq \mathbf{0},$$

where the inequality applies element-wise.

As mentioned above, in working with simulated data it is possible to start from reasonable starting values which would not cause the Fisher scoring algorithm to pass through infeasible values of $\boldsymbol{\mu}$. Thus, such an approach will be adopted without recourse to complex optimization algorithms which enforce such constraints. In fact, in most cases, starting the algorithm from the parameter estimates obtained from fitting the joint log-linear model for independence (using the `glm` function with log link and the Poisson family in the *R Programming Language*, 2004) is sufficient to prevent the algorithm from leaving the feasible parameter space. Furthermore, in assuming a Poisson likelihood and performing inference on $\mathbb{E}[N]$, it is not necessary to ensure that the parameter of interest, $\mathbb{E}[N]$ is larger than the number of observed individuals, as would be the case if the multinomial likelihood had been assumed and inference were to be performed on the parameter N. Since the results presented here are proof of concept, and the computational techniques involved not the focus of our discussion, it is not a serious drawback that the optimization technique used will not apply to all data and all starting values.

3.5 Data analysis

In this section we will analyze one real and two simulated four-source data sets. The analysis consists in fitting all possible 2047 MLLMs to each data set by implementa-

tion of the Fisher Scoring algorithm derived in the previous section, as well as fitting all possible 2047 JLLMs, including those which are nonhierarchical. In each case, we will use the AIC model selection criterion (Sakamoto et al., 1986) to select the best model amongst all all models of each family. The best MLLM will be compared to the best HJLLM. In so doing, we will demonstrate that the family of marginal log-linear models is complementary to the family of hierarchical joint log-linear models and, in some cases, are selected over a HJLLM. Moreover, the analyses serve another purpose: they provide weight to our suggestion that the universe of models to be considered using the standard joint log-linear modelling approach should be extended to include nonhierarchical models.

Remark 3.6 Note on the choice of AIC. The work of Hook and Regal (1997) supports the use of AIC in the capture-recapture setting. (See Section 1.2.1 for further details.) They found that the performance of other information criteria for model selection were found to be roughly equivalent. The use of AIC simplifies the discussion and is sufficient to demonstrate the viability of MLLMs as alternatives to JLLMs.

Remark 3.7 Note on terminology. The suffix *LLM* of JLLM and MLLM will be used interchangeably for log-linear models and log-linear modelling, where the context will dictate the meaning.

3.5.1 Modelling approach

We will consider the four-source capture-recapture setting. As described earlier in this chapter, the different marginal log-linear models to be considered are obtained by fixing all possible combinations of the 11 non-single source CIDs equal to 0. From Proposition 3.1, and as shown in Example 3.4, there are 2047 non-overparameterized models, of which 2036 are unsaturated. Likewise, using the same reasoning related

to the 11 possible interaction terms, there are 2047 non-overparameterized joint log-linear models, of which 2036 are unsaturated. Of these, 113 models are hierarchical.

The Fisher Scoring algorithm (described in Section 3.4) will be used to obtain maximum likelihood estimates for the marginal log-linear models. Code was written in R (*R Programming Language*, 2004) to implement the algorithm (see Appendix G for the code). The inbuilt `glm` function (with log link and the Poisson family) in R was used to run all joint log-linear models. We computed AIC values without using the inbuilt function in R for the JLLMs. In such a way the same code was used for all models and we avoided any potential differences in definition for the AIC. Estimates of precision for $\hat{N}$ were obtained using the asymptotic standard errors described in Section 1.2.1 for the JLLMs assuming an underlying Poisson likelihood (which we note will give tighter confidence intervals that those obtained assuming a multinomial likelihood - see Section 1.2.1) and using the approach of Lang and Agresti (1994) for the MLLM, also described in Chapter 1. Details are provided in Appendix F.2.

Estimating CIDs

For the best model in each family, the CIDs will estimated and contrasted using two different approaches: the model-based approach and the non-parametric approach, which we describe here.

Model-based approach

For MLLMs, the CIDs are parameters to be estimated. Thus, their value is obtained directly from the estimated model. In the case of JLLMs, the fitted model parameters are used to obtain the corresponding fitted cell probabilities, from which the CIDs are calculated.

Non-parametric approach

This approach is the same for both MLLMs and JLLMs. The estimated $\hat{N}$ is used to complete the table of observed cell counts to obtain an estimated cell mean for the

unobserved cell. The corresponding table of cell probabilities for the complete table is used to calculate the CIDs.

Simulation of data sets

The simulated data sets to be analysed are described in Sections 3.5.3 and 3.5.4 below. In both cases, an underlying multinomial likelihood is assumed. Given the equivalence of point estimates obtained using a Poisson model to those obtained using the multinomial model (Sandland & Cormack, 1984), we will fit the models using the more computationally straightforward Poisson likelihood, as mentioned in Section 3.4.

3.5.2 Real data: Diabetes data set

In this section we consider a real four-source data set that has been analysed several times in the literature (Bruno et al., 1994; IWGDMF, 1995a; Fienberg et al., 1999; Bartolucci & Forcina, 2001). The data set is provided in Table 3.1. Bruno et al. (1994) sought to enumerate all individuals with diabetes in a northern region of Italy on October 1, 1988. The four sources, denoted by A, B, C and D for consistency with the notation used throughout this dissertation, were described in IWGDMF (1995a) as follows:

- A: list of all patients with a previous diagnosis of insulin-dependent diabetes mellitus or non-insulin dependent diabetes mellitus via a diabetic clinic and/or family physicians;
- B: list of all patients discharged with a primary or secondary diagnosis of diabetes in all public and private hospitals in the region;
- C: computerized database list of insulin and oral hypoglycemic prescriptions for 1988;

- D: list of all residents of the region who requested a reimbursement for insulin and reagent strips.

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	58	46	14	8
	D_{No}	157	650	20	182
C_{No}	D_{Yes}	18	12	7	10
	D_{No}	104	709	74	?

Table 3.1: Diabetes data set of Bruno et al. (1994).

Of the 2069 cases, most were observed in source A. There were 1754, 452, 1135 and 173, observed in each of sources A, B, C and D, respectively. Thus, very few individuals were observed in each of sources B and D. This is not surprising given that these sources correspond to hospital discharges and reagent syringes, which we would expect to capture fewer individuals than the diabetic clinics and prescriptions. Further details of the data set can be found in the original article (Bruno et al., 1994).

In line with the approach adopted by IWGDMF (1995a) and described in Section 1.1.1, we present the results from each of the pairwise Petersen estimates as a first step in our analysis We notice that three of these estimates fall below the observed number of individuals, 2069. As stated in IWGDMF (1995a), and suggested in Wittes et al. (1974), the very low value (relative to the rest) of $\hat{N}$ for sources B and D suggests the need to consider positive dependence amongst sources. Thus, we do not anticipate that a model of independence be selected as the best model.

We now proceed to present results from all 2047 MLLMs and 2047 JLLMs. First, Figure 3.1 presents a plot of all AIC values against the number of model parameters. Secondly, Figure 3.2 presents the corresponding values of $\hat{N}$ for each of the models,

Sources	$\hat{N}$
A,B	2351
A,C	2185
A,D	2262
B,C	2057
B,D	803
C,D	1555

Table 3.2: Pairwise Petersen estimates of Diabetes data set of Bruno et al. (1994).

again compared to the number of parameters in the model. In both cases, we use colours and symbols to distinguish between the MLLMs and the JLLMs, and to further distinguish between those that are hierarchical and nonhierarchical in the latter. Note that there were three marginal models[1], all of which contained the four-way CID γ_{ABCD}, for which there were computational difficulties using the Fisher Scoring algorithm. We do not include these three models in the discussion below.

Figure 3.1 shows the AIC value for models related to the number of model parameters. We see that the minimum AIC of all 2047 JLLMs and that of all 2047 MLLMs are very close. In both cases the models contain 12 parameters. Table 3.3 summarizes these results and presents the form of each of these models. The best JLLM slightly outperforms the best MLLM with an AIC of 24.91 compared to 25.07 for the best MLLM.

[1]

$$A+B+C+D+AB+AC+AD+BC+ABC+ABD+ACD+ABCD$$

$$A+B+C+D+AB+AC+AD+BC+CD+ABC+ACD+BCD+ABCD$$

$$A+B+C+D+AB+AC+AD+BC+CD+ABC+ABD+ACD+ABCD$$

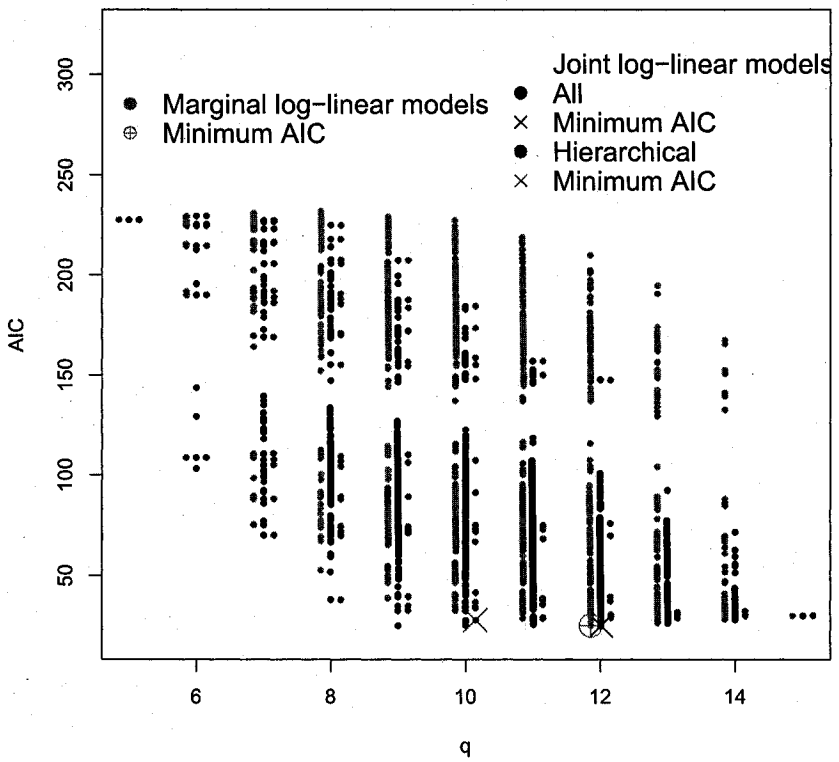


Figure 3.1: AIC vs. number of model parameters for all four-source JLLMs and MLLMs for diabetes data of Table 3.1

We notice that both the best-performing JLLM and best-performing MLLM (as selected by the AIC criterion) are nonhierarchical. In both instances, nonhierarchical models are preferred over the best-performing HJLLM, which is more parsimonious with 10 parameters but with a larger AIC value of 27.62. The best overall model is the nonhierarchical JLLM which provides an estimate of $\hat{N} = 3092$, followed by the MLLM with an estimate which is considerably smaller at $\hat{N} = 2345$. Should we constrain ourselves to the standard analysis of choosing amongst the best-performing

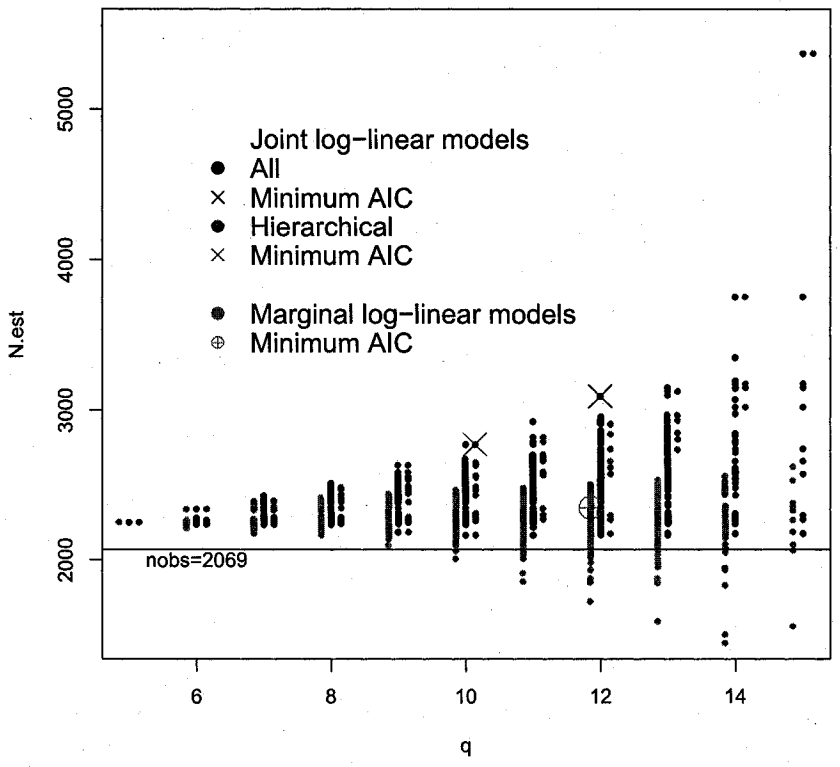


Figure 3.2: $\hat{N}_{MLE}$ vs. number of model parameters for all four-source JLLMs and MLLMs for diabetes data of Table 3.1

HJLLM, we would estimate an intermediate value of $\hat{N} = 2771$.

As a result of the differences in $\hat{N}$ according to which of the three families of models we select amongst, there is quite a difference in the point estimate of the proportion of the population observed in the data set. For the best-performing JLLM, 67% (2069/3092) of the diabetes population is assumed to have been observed, for the best-performing HJLLM, 75% (2069/2771), with a much higher proportion at 88% (2069/2345) estimated to have been observed for the best-performing MLLM.

Model		Model parameters													95%
Family	AIC	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD	q	$\hat{N}$	C.I.
JLLM	24.91	1	1	1	1	1	1	0	0	0	0	1	12	3092	(2573, 3611)
HJLLM	27.62	1	1	0	1	1	1	0	0	0	0	0	10	2771	(2491, 3050)
MLLM	25.07	0	1	0	1	1	1	1	0	0	1	1	12	2345	(2295, 2396)

Table 3.3: Best-performing models by model family for diabetes data

Remark 3.8 Both the best performing JLLM and MLLM contain the same number of parameters (see Table 3.3). Although the model form does not contain the same types of dependence terms, we observe that in both cases a term related to four-way dependence is included: in the case of the JLLM this is a four-way interaction term (note that the model is not saturated since it has fewer than 15 parameters), whilst in the case of the MLLM this term is the 4-way CID, denoted by γ_{ABCD}. Further, we note that the 4-way interaction in the JLLM models the differential behavior of the corresponding cell with respect to other cells, while the 4-way term in the MLLM models the extra dependence induced by considering 4 sources, as distinct from 3- and 2-way dependence.

Table 3.4 provides additional information concerning the fit of each of the best performing models in each of the three families. It contains the estimated CIDs (using both the model-based and nonparametric techniques described on page 112) for each of the three best-performing models, as well as the estimated marginal means. We choose to examine the estimated marginal means rather than the estimated cell means since, as will be described below in terms of sufficiency, they exhibit more meaningful properties than the cell means. We notice that the CIDs for each of the best performing models of each of the three families are different since the models relate to different dependence structures. Nonetheless, the CIDs serve to provide in-

Margin	Estimated CID						Observed	Estimated marginal mean		
	JLLM		HJLLM		MLLM					
	MB	NP	MB	NP	MB	NP	count	JLLM	HJLLM	MLLM
A	0	0	0	0	0	0	1754	1754	1754	1754
B	0	0	0	0	0	0	452	452	452	452.04
C	0	0	0	0	0	0	1135	1135	1135	1136.03
D	0	0	0	0	0	0	173	173	173	173
AB	0.273	0.273	0.164	0.164	0	-0.003	337	337	337	338.04
AC	0.347	0.347	0.238	0.238	0.069	0.071	911	911	911	910.25
AD	0.312	0.312	0.174	0.202	0	0.035	134	134	130.27	129.37
BC	0.406	0.406	0.296	0.296	0.134	0.130	249	249	249	250.45
BD	1.344	1.344	1.235	1.235	1.068	1.068	97	97	97	97.02
CD	0.685	0.685	0.576	0.576	0.402	0.409	126	126	126	125.30
ABC	-0.199	-0.200	-0.102	-0.091	0.069	0.076	215	215.22	212.65	215.06
ABD	-0.257	-0.262	-0.105	-0.152	0	0.014	76	76.36	77.47	72.55
ACD	-0.272	-0.284	-0.144	-0.174	0	-0.007	104	105.17	104.23	100.40
BCD	-0.412	-0.387	-0.240	-0.277	-0.114	-0.111	72	70.25	74.77	71.70
ABCD	0.172	0.165	0.075	0.055	-0.107	-0.112	58	58	63.85	55.35

Table 3.4: CIDs and marginal means for diabetes data from best-performing JLLM, HJLLM and MLLM, where MB and NP indicate model-based and nonparametric estimation, respectively.

terpretation to model parameters of nonhierarchical JLLMs, which are often excluded from the analysis of capture-recapture data. For the MLLM, the nonparametrically estimated CIDs corresponding to the CID-terms omitted from the model are not zero as with the model-based approach. This is to be expected given that the fitted model is not used in the estimation procedure aside from using the value of $\hat{N}$ from the fitted model. Further features observed in that table occur for all data sets. We delay discussion of these common features until Section 3.5.5, which follows the analysis of the two simulated data sets presented in Sections 3.5.3 and 3.5.4.

Figure 3.2 shows the point estimates of $\hat{N}$ relative to the number of model parameters q for all 2047 JLLMs and all 2047 MLLMs. The interest of such a plot is that some models (independence, saturated, some 2-way interaction models) can be immediately identified on the graph as having the same fit for JLLMs and MLLMs. (As with the description of features of Table 3.4, several of the features observed in Figure 3.2 occur for all data sets. Again, we delay discussion of these common features until Section 3.5.5.)

A feature of Figure 3.2, particular to the diabetes data set, is that the estimates of $\hat{N}$ obtained from MLLMs tend to be smaller than those obtained by using JLLMs. Moreover, we observe that there are point estimates of $\hat{N}$ smaller than $n_{obs} = 2069$ for some of the MLLMs but for none of the JLLMs. In choosing to adopt the Poisson model for computational reasons, it is possible that the corresponding point estimate $\hat{N}$ be smaller than n_{obs}, since the parameter of interest is $\mathbb{E}[N]$ rather than N. Nonetheless, in the case of the best performing MLLM, $\hat{N} = 2345 > n_{obs}$, and the lower limit of the asymptotic 95% confidence interval, given by $(2295, 2396)$, exceeds n_{obs}.

3.5.3 Simulated data: Conditional independence

The data analysis undertaken in the previous section concerns a real data set. Consequently we do not have a notion of the true N and it is not possible to determine how well the model performs in terms of estimating that true value. Using simulated data provides a means to verify the performance of modelling techniques. In this section and the next, we analyze two different simulated capture-recapture data sets with a true underlying population size of $N = 1000$ in each case. Two different dependence structures will be used: first, conditional independence and, second, a nonhierarchical dependence structure. In so doing, we will be able to assess how well the model selected by the AIC criterion performs for the specific data set under study.

Remark 3.9 We note that in working with a single simulated data set in each case, there is sampling variability which will not be accounted for explicitly by determining frequentist coverage properties of the model. This is not the goal of the current chapter. We draw the reader's attention to the fact that it is possible that, by chance, the simulated data set is somewhat different to the true generating structure. Nevertheless, the analyses presented in this section and the next section, serve the useful purpose of exploring the use of MLLMs. We will hold to requiring coverage of the true value of $N = 1000$ by the 95% confidence intervals. Further, such intervals may be unreasonable tight because of the method used (the Wald, asymptotic approach assuming an underlying Poisson likelihood, using $\widehat{\text{Var}}_P\left[\hat{N}\right]$ of Section 1.2.1).

The four-source simulated data set considered in this section is generated according to a known conditional dependence structure given by $[AB][AC][D]$. The results from this analysis serve to confirm the result given by Theorem 2.9 for HJLLMs, concerning the value of the CIDs for a model of conditional independence.

The data are shown in Table 3.5 and the details of the generating mechanism can be found in Appendix E. We now proceed to present results from all 2047 MLLMs

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	16	24	5	10
	D_{No}	61	92	34	47
C_{No}	D_{Yes}	50	72	16	26
	D_{No}	165	224	50	?

Table 3.5: Data generated according to the conditional independence structure given by $[AB][AC][D]$ (see Appendix E).

and 2047 JLLMs. First, Figure 3.3 presents a plot of all AIC values against the number of model parameters. Secondly, Figure 3.4 presents the corresponding values of $\hat{N}$ for each of the models, again compared to the number of parameters in the model. In both cases, we use colours and symbols to distinguish between the MLLMs and the JLLMs, and to further distinguish between those that are hierarchical and nonhierarchical in the latter.

Figure 3.3 shows the AIC value for models related to the number of model parameters q. We see that the minimum AIC of all 2047 JLLMs and that of all 2047 MLLMs are very close. In both cases the models contain 7 parameters. In fact, it is clear that the best-performing JLLM is in fact a HJLLM and, thus, it is also the best-performing HJLLM. Table 3.6 confirms this relationship, as it summarizes these results and presents the form of each of the best-performing models.

The best-performing MLLM slightly outperforms the best-performing JLLM with an AIC of 16.13 compared to 16.24 for the best HJLLM. We notice that both the best-performing MLLM and JLLM contain the same CID and interaction terms in the case of the MLLM and HJLLM, respectively. The best overall model is the MLLM which provides an estimate of $\hat{N} = 964$, compared to the estimate of $\hat{N} =$

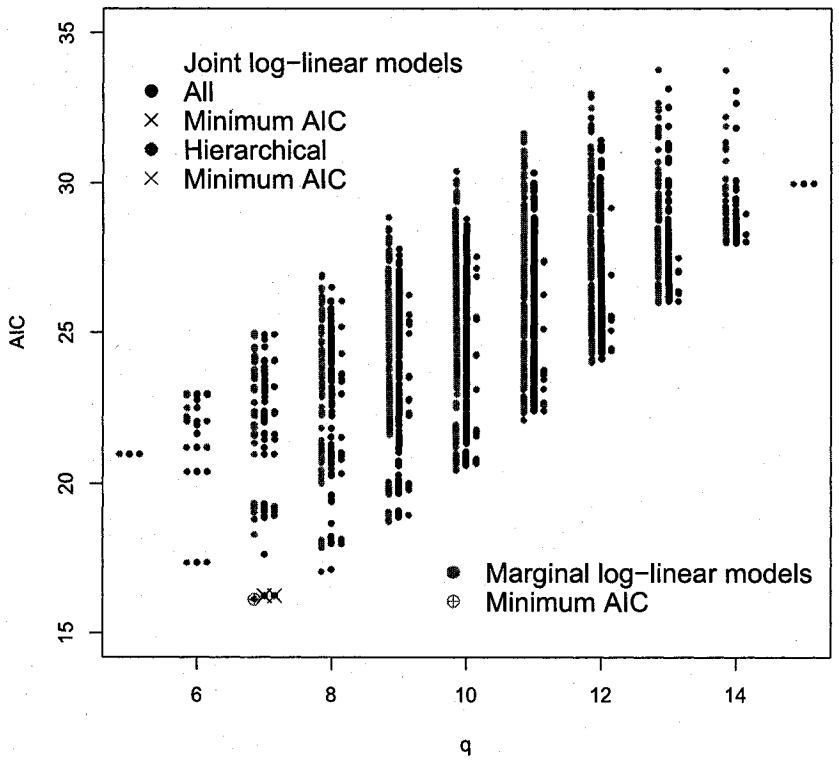


Figure 3.3: AIC vs. number of model parameters for all four-source JLLMs and MLLMs for data of Table 3.5 generated according to $[AB][AC][D]$

966 for the best-performing JLMM. Thus, with point estimates of 964 and 966, the estimated proportion of the population observed in the capture-recapture data set is 92% (obtained using 892/965 by taking the average of the two very close point estimates). We observe that none of the 95% confidence intervals contain the true value $N = 1000$ (see Remark 3.9). Thus, our MLLM performs no worse than the best-performing JLLM.

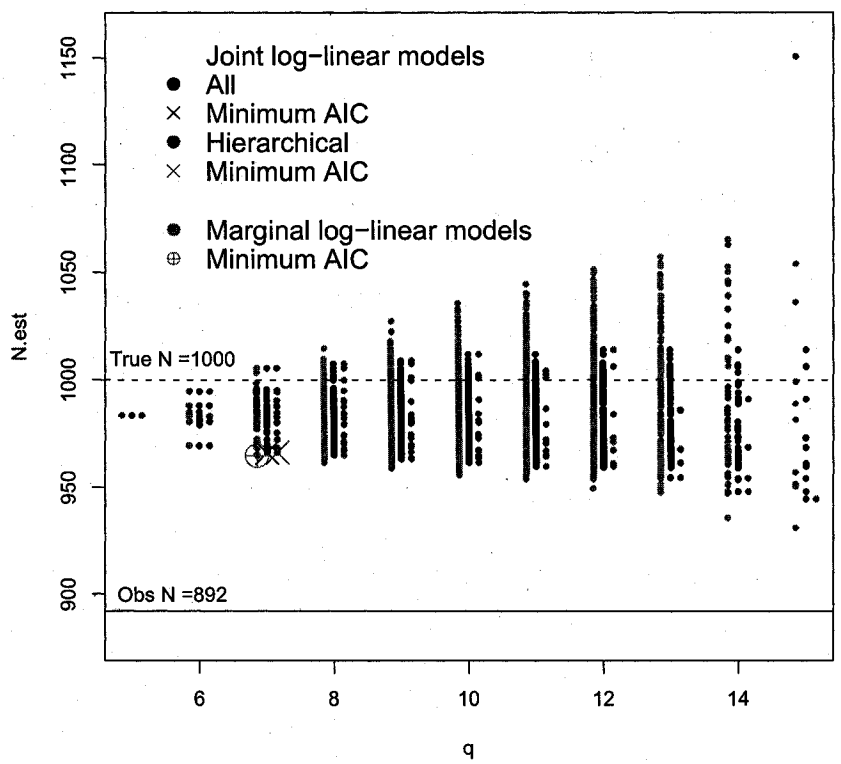


Figure 3.4: $\hat{N}_{MLE}$ vs. number of model parameters for all four-source JLLMs and MLLMs for data of Table 3.5 generated according to $[AB][AC][D]$

Table 3.7 provides additional information concerning the fit of each of the best-performing models in each of the JLLM and MLLM families. It contains the estimated CIDs (using both the model-based and nonparametric techniques described on page 112) and estimated marginal means for each of the best-performing models. As with the diabetes data in Section 3.5.2, we choose to examine the estimated marginal means rather than the estimated cell means. A feature of this table serves to confirm results shown in Chapter 2.

Model Category	AIC	Model parameters											q	$\hat{N}$	95% C.I.
		AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD			
JLLM	16.24	0	1	0	0	0	1	0	0	0	0	0	7	966	(948, 984)
HJLLM	16.24	0	1	0	0	0	1	0	0	0	0	0	7	966	(948, 984)
MLLM	16.13	0	1	0	0	0	1	0	0	0	0	0	7	964	(943, 987)

Table 3.6: Best-performing models by model family for data of Table 3.5 generated according to $[AB][AC][D]$.

The best fitted JLLM is the model of conditional independence of A and D given C, denoted by $[AC][CD][B]$, which serves to confirm the result contained in Theorem 2.9. The theorem translates into the following for the HJLLM model given by $[AC][CD][B]$: $\gamma_{ACD} = -\gamma_{AD}$. We see that this relationship holds, with $\hat{\gamma}_{AD} = 0.006$ and $\hat{\gamma}_{ACD} = -0.006$, using the model-based estimates of the CIDs. Moreover, we observe that all CIDs that should be fitted at zero are estimated as zero via the model-based estimates. We note that the same relationship does not hold for the nonparametric estimates of the CIDs, precisely because they are not obtained from the fitted model.

Remark 3.10 We note that neither the best-performing JLLM nor best-performing MLLM correspond to the correct underlying model given by $[AB][AC][D]$. Rather they are both models of a related conditional independence structure given by $[AC][CD][B]$. (Again, see Remark 3.9.)

Figure 3.4 shows the point estimates of $\hat{N}$ relative to the number of model parameters q for all 2047 JLLMs and all 2047 MLLMs. A feature of Figure 3.4 that is particular to the simulated data set under study in this section is that the estimates of N, denoted by $\hat{N}$, obtained from MLLMs are more variable than those obtained by using JLLMs. Unlike the results from the diabetes data set, we observe that none

	CID					Marginal means		
		Estimated						
		JLLM		MLLM		Observed	Estimated	
Margin	True	MB	NP	MB	NP	count	JLLM	MLLM
A	0	0	0	0	0	704	704	704
B	0	0	0	0	0	397	397	397
C	0	0	0	0	0	289	289	289.03
D	0	0	0	0	0	219	219	219
AB	0.069	0	0.009	0	0.008	292	289.27	289.67
AC	-0.049	-0.087	-0.087	-0.089	-0.089	193	193	192.91
AD	0	0.006	0.015	0	0.014	162	160.48	159.79
BC	-0.00797	0	-0.023	0	-0.025	116	118.75	118.93
BD	0	0	-0.034	0	-0.035	87	89.98	90.11
CD	0	-0.175	-0.175	-0.178	-0.176	55	55	54.88
ABC	0.00797	0	-0.015	0	-0.014	77	79.30	79.38
ABD	0	0	0.016	0	0.017	66	65.94	65.75
ACD	0	-0.006	0.070	0	0.072	40	36.73	36.63
BCD	0	0	-0.016	0	-0.015	21	22.60	22.58
ABCD	0	0	0.037	0	0.035	16	15.09	15.07

Table 3.7: CIDs and marginal means for conditional independence data from best performing JLLM and MLLM, where MB and NP indicate model-based and non-parametric estimation, respectively.

of the point estimates of $\hat{N}$ is smaller than $n_{obs} = 892$. Moreover, the width of the asymptotic 95% confidence intervals for both the best JLLM and best MLLM, are reasonably tight with neither interval containing the true value of $N = 1000$ (See Remark 3.9).

The data analysis undertaken in the current section demonstrates that a MLLM model can outperform the standard data analytical approach of HJLLM, even for a data set generated according to a known hierarchical dependence structure. Such an example supports the necessity expand the universe of models to be considered to include nonhierarchical models. In all cases, the CIDs serve to provide meaning to the model parameters, which is a criticism levelled against the use of nonhierarchical models. Moreover, we see that our MLLM can perform at least as well as JLLMs.

3.5.4 Simulated data: Nonhierarchical dependence

In this section we consider a four-source simulated data set generated according to a known nonhierarchical dependence structure in which the only dependence present is in two of the three-way margins, specifically in the two sets of sources given by $\{A, B, C\}$ and $\{A, B, D\}$. Note that, unlike a hierarchical structure, there is no pairwise dependence. Such a nonhierarchical dependence structure cannot be expressed in the notation of hierarchical models; in particular, it is not represented by $[ABC][ABD]$. Such a nonhierarchical structure is not well modelled by hierarchical dependence structures.

The data are shown in Table 3.8 and the details of the generating mechanism can be found in Appendix E. We now proceed to present results from all 2047 MLLMs and 2047 JLLMs. First, Figure 3.5 presents a plot of all AIC values against the number of model parameters q. Secondly, Figure 3.6 presents the corresponding values of $\hat{N}$ for each of the models, again compared to the number of parameters in the model. In both cases, we use colours and symbols to distinguish between the MLLMs and the JLLMs,

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	21	9	6	26
	D_{No}	61	59	60	70
C_{No}	D_{Yes}	35	27	17	60
	D_{No}	82	195	128	?

Table 3.8: Data generated according to the nonhierarchical dependence structure given Appendix E.

and to further distinguish between those that are hierarchical and nonhierarchical in the latter.

Figure 3.5 shows the AIC value for models related to the number of parameters q in the model. We see that, unlike in the previous two sections for the diabetes data and the simulated data of conditional independence, the minimum AIC of all 2047 JLLMs and that of all 2047 MLLMs are considerably different. For the MLLMs, it is given by 19.46 for 12 parameters and for the JLLMs by 22.86 for 11 parameters. Neither the best-performing MLLM nor the best-performing JLLM is hierarchical. In fact, the best-performing HJLLM does considerably worse than both of these models, with an AIC of 24.83 and 12 parameters. Table 3.9 summarizes these results and presents the form of each of these models.

The point estimates $\hat{N}$ are close to each other, given by 1011, 1017 and 1032 for the JLLMs, HJLLMs and MLLMs, respectively. We see that the asymptotic 95% confidence intervals contain the true value $N = 1000$ for the JLLMs and HJLMMs but not in the case of the MLLM, as we would wish. It should be noted that these intervals, derived from asymptotic arguments, are likely to be tighter than those obtained by an alternative non-asymptotic method and we note that the lower bound

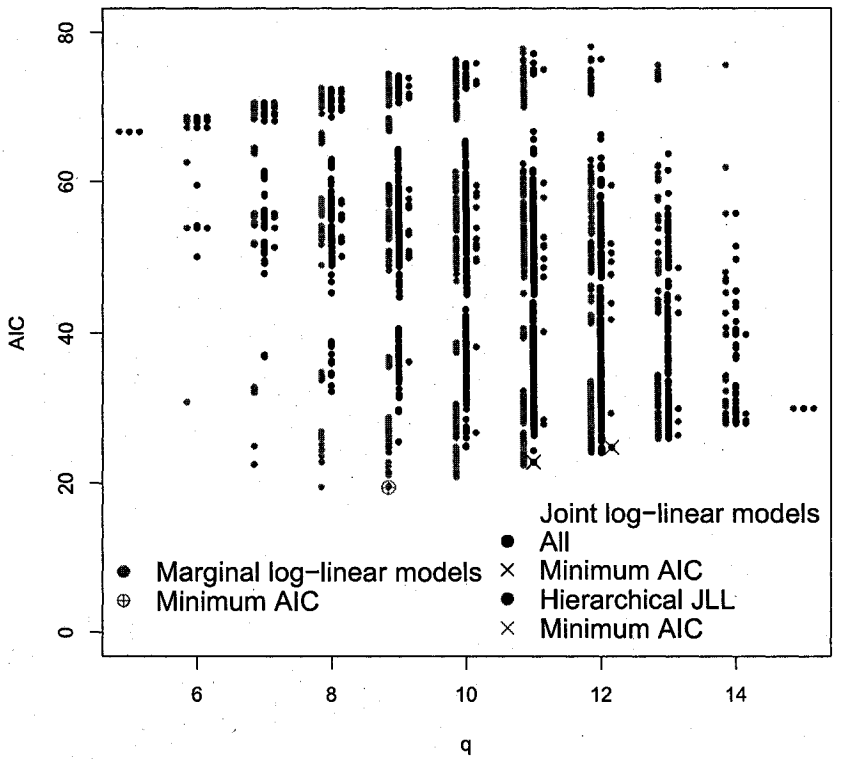


Figure 3.5: AIC vs. number of model parameters for all four-source JLLMs and MLLMs for data in Table 3.8, generated according to a nonhierarchical dependence scheme (see Appendix E).

of 1003 is close to the truth of $N = 1000$. Moreover, we observe that the interval for the MLLM is tighter than those for both the best-performing JLLM and HJLLM, although it does not contain the true value of $N = 1000$.

Remark 3.11 The best-performing JLLM and MLLM contain different numbers of parameters (see Table 3.9) and very different terms (which are interaction terms in

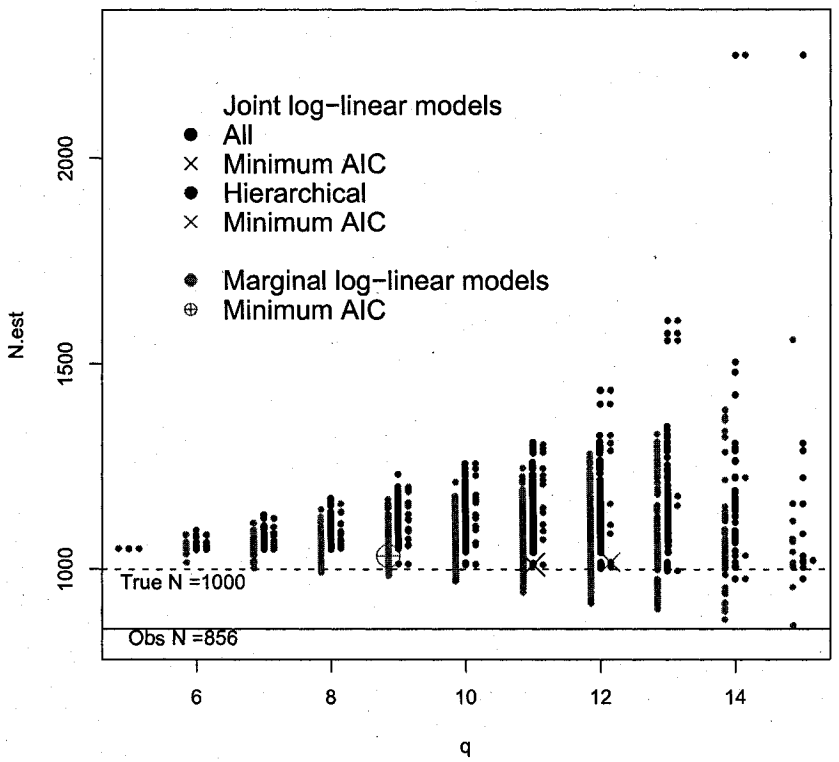


Figure 3.6: $\hat{N}_{MLE}$ vs. number of model parameters for all four-source JLLMs and MLLMs for data of Table 3.8 generated according to a nonhierarchical dependence scheme (see Appendix E).

the case of JLLMs and CID terms in the case of MLLMs). The best-performing model overall, the MLLM, is close to the true underlying generating dependence structure for which the only nonzero CIDs are γ_{ABC} and γ_{ABD} (see Table 3.10). It contains the CID terms γ_{ABC} and γ_{ABD}, as we would wish, and also γ_{BC} and γ_{BCD}.

Figure 3.6 shows the point estimates of $\hat{N}$ relative to the number of model para-

Model Category	AIC	Model parameters											q	$\hat{N}$	95% C.I.
		AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD			
JLLM	22.86	1	1	1	0	1	0	1	1	0	0	0	11	1011	(958, 1064)
HJLLM	24.83	1	1	1	1	1	0	1	1	0	0	0	12	1017	(934, 1100)
MLLM	19.46	0	0	0	1	0	0	1	1	0	1	0	12	1032	(1003,1063)

Table 3.9: Best-performing models by model family for nonhierarchical dependence data of Table 3.8

meters q for all 2047 JLLMs and all 2047 MLLMs. A feature of Figure 3.6 that is particular to the nonhierarchical data set under study is that the estimates $\hat{N}$ of N obtained from MLLMs spread out similarly to those from the JLLMs but attain lower values than for the JLLMs. Nonetheless, none of the point estimates $\hat{N}$ are smaller than $n_{obs} = 856$. The features of Table 3.10, which provides additional information concerning the fit of each of the best models, will be discussed in the next section.

The results from this section serve to provide evidence that the fit of a nonhierarchical model, whether it be a marginal or a joint model, can improve that of a HJLLM for nonhierarchical data.

3.5.5 Observations common to all analyses

In this section we describe the common features of the analysis of the three data sets in Sections 3.5.2- 3.5.4. For convenience we will use the terms "Figures-AIC" to denote Figures 3.1, 3.3 and 3.5, and "Figures-$\hat{N}$" to denote Figures 3.2, 3.4 and 3.6. "Tables-BestModels" will identify Tables 3.3, 3.6 and 3.9, whilst "Tables-Estimates" will be used to denote Tables 3.4, 3.7 and 3.10, which present the estimated CIDs and marginal means for each of the best-performing models for the three data sets.

Remark 3.12 For all three tables of Tables-Estimates, we choose not to round the

Table 3.10: CIDs and marginal means for nonhierarchical dependence data of Table E.2 from best performing JLLM, HJLLM and MLLM, where MB and NP indicate model-based and nonparametric estimation, respectively.

	CID							Marginal means			
		Estimated									
		JLLM		HJLLM		MLLM		Observed	Estimated		
Margin	True	MB	NP	MB	NP	MB	NP	count	JLLM	HJLLM	MLLM
A	0	0	0	0	0	0	0	489	489	489	488.99
B	0	0	0	0	0	0	0	410	410	410	411.41
C	0	0	0	0	0	0	0	312	312	312	311.31
D	0	0	0	0	0	0	0	201	201	201	199.59
AB	0	0.004	0.004	0.001	0.010	0	0.025	199	199	199	194.75
AC	0	-0.005	-0.005	0.0004	0.0004	0	0.015	150	150	150	147.37
AD	0	-0.054	-0.054	-0.049	-0.049	0	-0.034	92	92	92	94.48
BC	0	0.153	0.158	0.163	0.163	0.169	0.178	148	147.36	148	146.88
BD	0	-0.030	-0.030	-0.025	-0.025	0	-0.010	79	79	79	79.49
CD	0	0.052	0.0004	0.049	0.006	0	0.021	62	65.28	64.71	60.15
ABC	0.1515	0.142	0.137	0.132	0.132	0.115	0.117	82	82	82	78.00
ABD	0.4055	0.433	0.433	0.427	0.427	0.400	0.412	56	56	56	56.14
ACD	0	0.058	0.061	0.062	0.055	0	0.040	30	31.52	31.52	28.47
BCD	0	0.010	-0.055	0.011	-0.061	-0.124	-0.076	27	30.20	30.27	25.06
ABCD	0	-0.120	-0.100	-0.121	-0.095	0	-0.079	21	23.08	23.08	19.86

estimated marginal means to the nearest integer in order to differentiate between those estimated perfectly and those not. Thus, whenever an estimated marginal mean is an integer it is estimated so perfectly.

The features in common to the analysis of each of the three data sets are given as follows:

- The AIC and $\hat{N}$ estimates coincide for the JLLM of independence and the MLLM for independence, since these are equivalent (see the models for $q = 5$ in Figures-AIC and Figures-$\hat{N}$).

- For the range of saturated models, whether JLLM or MLLM, the model fit is the same but $\hat{N}$ varies. More specifically, consider the following. All models with 15 parameters are saturated: there is a single saturated HJLLM, but 11 saturated JLLMs and 11 saturated MLLMs. Since all such models fit the data perfectly they have the same fit, as observed by a single point value at $q = 15$ in Figures-AIC. However, there are multiple values of $\hat{N}$ corresponding to the range of the 11 saturated JLLMs and the range of the 11 saturated MLLM visible as a range of distinct values at $q = 15$ in Figures-$\hat{N}$. For the HJLLMs of each of the three plots of Figures-$\hat{N}$, there is a single $\hat{N}$ value corresponding to the single saturated HJLLM.

- For the plots of Figures-$\hat{N}$, we observe that, as the number of model parameters q increases, there is a funnelling form to the $\hat{N}$ estimates. As a consequence, the corresponding estimates of N are more variable and, thus, more spread out.

- Tables-Estimates provide empirical support to results concerning the fit of sufficient statistics of JLLMs as proved in Bishop et al. (1975). For the best-performing JLLMs for each of the three data sets, both hierarchical and non-hierarchical, the single-source marginal means are estimated perfectly equal to

the observed marginal means. In fact, such a property holds for all JLLMs, as described in Bishop et al. (1975). They prove that parameters corresponding to all sufficient statistics, which include the single-source marginal means, are estimated perfectly by JLLMs. Such a property also explains some of the other perfect fits observed for the JLLMs.

- Tables-Estimates also show that whenever an interaction term is included in a JLLM, the corresponding marginal mean is fitted perfectly equal to the observed marginal mean, since the marginal counts corresponding to interactions included in the model are sufficient statistics (again, see Bishop et al., 1975). Thus, for each of the selected JLLMs and HJLLMs shown in Tables-Estimates, the marginal mean corresponding to any of the interaction terms included in the model, as well as the single-source marginal means are estimated perfectly. For example, consider the best-performing JLLM for the data set of conditional independence in Section 3.5.3. The chosen model contains two interaction terms corresponding to AC and CD (see Table 3.6). From Table 3.7 we see that the corresponding marginal means, m_{AC} and m_{CD}, are estimated equal to the observed marginal counts of $n_{AC} = 193$ and $n_{CD} = 55$. Moreover, the single-source marginal means are also estimated perfectly as $\hat{m}_A = n_A = 704$, $\hat{m}_B = n_B = 397$, $\hat{m}_C = n_C = 289$ and $\hat{m}_B = n_B = 219$.

- An equivalent relationship is not observed with the MLLMs, although further investigations would be required to determine whether this is a numerical artefact. First, the single-source marginal means are not consistently estimated equal to the observed single-source marginal counts. Nonetheless they are very close, within less than 1 individual in all cases. Second, whenever a CID-term is included in the model, the equivalent marginal mean is not necessarily estimated perfectly equal to the corresponding observed marginal count. For

example, consider again the conditional independence data of Section 3.5.3. The best-performing MLLM contains only 2 non-zero CIDs, given by γ_{AC} and γ_{CD}, whilst the corresponding estimated marginal means $\hat{m}_{AC} = 192.91$ and $\hat{m}_{CD} = 54.88$ are not estimated perfectly equal to the corresponding observed marginal counts of $n_{AC} = 193$ and $n_{CD} = 55$.

- For the estimated CIDs, several features are observed (see Tables-Estimates). First, we notice that the non-parametric (NP) CID estimates are close to the model-based (MB) estimates. This is to be expected since the MB estimates use the fitted cell means in place of the observed cell counts, as is the case with the NP estimates. When the model fits the data well, the fitted cell means are close to the observed cell counts, which they aim to fit. In both cases, the same value of $\hat{N}$ is used, i.e. that obtained from the fitted model.

- For the MB-based CIDs from the MLLM, those CIDs corresponding to CID-terms omitted from the model are exactly 0, as they are set to 0 by design.

3.6 Summary

The nature of the goals of the current chapter was exploratory. These goals were twofold: first, to demonstrate that it is possible to obtain reasonable maximum likelihood estimates under the CID model formulation; and, second, to begin to explore dependence structures which are not well modelled by hierarchical log-linear models and demonstrate that our model thus parameterized is able to out-perform both the best-performing hierarchical log-linear model and nonhierarchical log-linear model. The work presented in this chapter has enabled us to achieve these goals.

We have demonstrated that a frequentist maximum likelihood approach to the analysis of MLLMs can work. For data generated according to a nonhierarchical de-

pendence structure, a MLLM outperforms even the best-performing nonhierarchical JLLM (see Section 3.5.4). Moreover, as seen in Section 3.5.3, MLLMs can even outperform (in terms of model fit as measured by AIC) the best-performing HJLLM for data generated according to a hierarchical dependence structure. For the real diabetes data set of Section 3.5.2, the best-performing model overall is a nonhierarchical JLLM and the MLLM still outperforms the best-performing HJLLM (again as judged by AIC). The three analyses serve to provide weight to the suggestion that the universe of models to be considered in the analysis of epidemiological capture-recapture should be expanded to included nonhierarchical and marginal models.

The three data analyses have also served to confirm Theorem 2.9 and Corollary 2.2, as well as the general results related to JLLMs and sufficiency as per Bishop et al. (1975). They have enabled us to explore the nature of the CID measures introduced in Chapter 2 and to suggest that these measures can be used to provide interpretability to the model fit of nonhierarchical models, even nonhierarchial JLLMs.

In short, we see that there is scope for using nonhierarchical modelling in the analysis of capture-recapture data and the family of marginal log-linear models provide a complementary and universally interpretable class of models to that of joint log-linear models.

Chapter 4

Bayesian Marginal Log-Linear Models with Random Effects

4.1 Modelling dependence with random effects

An alternative approach to the parametric models seen in Chapter 3 is to model the dependence structure of the incomplete contingency table using random effects (see Coull & Agresti, 2003). Such an approach is not only attractive in its ability to reduce the number of parameters to estimate but also, as we shall see in Chapter 5, because of the potential form of the model and its interpretation.

In this chapter we introduce the general form of such a model. In Section 4.2 we produce its parametric form. In Section 4.3 we motivate the use of the Bayesian paradigm for parameter estimation. In Section 4.4, we present the full Bayesian formulation consisting of the specification of the likelihood and prior distributions. A description of the Markov chain Monte Carlo (MCMC) scheme which can be employed to obtain a sample from the joint posterior distribution is provided in Section 4.5. In

the following chapter we will implement one such random effects model in which the CIDs are treated as random effects.

4.2 The random effects model

The mixed effects marginal model formulation for the general K-source capture-recapture setting is given by

$$\begin{aligned} \log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi}, \\ \boldsymbol{\xi} &\sim N(\mathbf{0}, \Sigma), \end{aligned} \tag{4.1}$$

where $\mathbf{A}$ is the $d \times d$ matrix which transforms the $d \times 1$ vector of cell means $\boldsymbol{\mu}$, into the $d \times 1$ vector of marginal means $\mathbf{m}$, for $d = 2^K - 1$, the number of observed cell entries in the incomplete K-way contingency table. The $(K+1) \times 1$ vector of fixed effects is denoted by $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}^\star]' = [\beta_0, \beta_{S_1}, \ldots, \beta_{S_K}]'$, for sources $S_1, \ldots, S_K$, with corresponding $d \times (K+1)$ design matrix, $\mathbf{X}$. The $r \times 1$ vector of random effects is denoted by $\boldsymbol{\xi}$ with corresponding $d \times r$ design matrix, $\mathbf{Z}$. Σ denotes the $r \times r$ covariance matrix of the assumed distributional form of $\boldsymbol{\xi}$. Let $\boldsymbol{\theta}$ denote the $(K+1+r) \times 1$ concatenation of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ given by

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\xi} \end{bmatrix}.$$

In this chapter we will specify neither the vector of random effects $\boldsymbol{\xi}$, nor the corresponding design matrix $\mathbf{Z}$, aside from making assumptions on the dimension of $\mathbf{Z}$, which will have d rows. One specification for $\mathbf{Z}$ and $\boldsymbol{\xi}$ is presented in Chapter 5.

Remark 4.1 Note that although $\boldsymbol{\theta}$ may represent the same model components as the $\boldsymbol{\delta}$ vector of Chapter 3, with $\boldsymbol{\xi}$ equal to the vector of CIDs denoted by $\boldsymbol{\gamma}$, we choose to keep the notation different to reiterate that the current chapter introduces

the general form of the mixed effects model. In such a way a distinction is made between the two treatments of $\boldsymbol{\theta}$: as fixed effects in Chapter 3, where $\boldsymbol{\theta} = \boldsymbol{\delta}$, and as random effects in the current and subsequent chapters.

Model (4.1) parallels that given by Coull and Agresti (2003), in which the generalized log-linear model of the form

$$\mathbf{C}\log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi}, \tag{4.2}$$

is presented, where $\mathbf{C}$ is a matrix of appropriate dimension. Such a model is more general than that expressed in (4.1), and is equivalent when $\mathbf{C}$ is the identity matrix. However, the applications considered by Coull and Agresti (2003) do not relate to the capture-recapture setting; rather they relate to smoothing of sparse data for large contingency tables. The mixed effects model given by (4.1) has a more specific form in the capture-recapture setting presented in this dissertation. Specifically, the form of $\mathbf{X}$ and $\boldsymbol{\beta}$ correspond to the form of the marginal model as it was first introduced in Chapter 2.

Remark 4.2 Relationship to Generalized Linear Mixed Models.
The class of models denoted by (4.2), and consequently that denoted by (4.1), are related to the class of generalized linear mixed models as presented by McCulloch and Searle (2001). (See Section 1.5 of this dissertation.) We note that the broad description of McCulloch and Searle (2001) suggests that they work with the case when $\mathbf{C} = \mathbf{A} = \mathbf{I}$ and the link function is of a general form, not necessarily the log link of (4.2) and (4.1).

Example 4.1 Consider the three-source capture-recapture setting for sources A, B and C. The data are conveniently summarized in the three-source incomplete contingency table, given by Table 4.1, as first introduced in Chapter 1.

	A_{Yes}		A_{No}	
	B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	n_{ABC}	$n_{A\bar{B}C}$	$n_{\bar{A}BC}$	$n_{\bar{A}B\bar{C}}$
C_{No}	$n_{AB\bar{C}}$	$n_{A\bar{B}\bar{C}}$	$n_{\bar{A}B\bar{C}}$	$n_{\bar{A}\bar{B}\bar{C}} = ?$

Table 4.1: Incomplete Contingency Table: Three Source

The components of model (4.1) are given as follows (except for the random effects vector $\boldsymbol{\xi}$ and corresponding design matrix $\mathbf{Z}$, whose form remains unspecified in the current chapter), where we use the subscript '3' for clarity:

$$\mathbf{m_3} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix}; \mathbf{A_3} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \boldsymbol{\mu_3} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix},$$

and

$$\mathbf{X_3} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 \\ -2 & 1 & 1 & 1 \end{bmatrix}; \boldsymbol{\beta_3} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix} = \begin{bmatrix} \log N \\ \log m_A \\ \log m_B \\ \log m_C \end{bmatrix},$$

where N is the population size to be estimated. Note that the corresponding matrices for the four-source setting are found in Appendix A.

Remark 4.3 As described in Chapter 3, it is convenient to introduce notation to distinguish between the parameter vector, $\boldsymbol{\beta}$, which consists of $\log N$ together with all single source log-marginal means, and that which does not include $\log N$, denoted by $\boldsymbol{\beta}^*$. In particular, for three sources $\boldsymbol{\beta_3}$ denotes the full set of β parameters, whilst $\boldsymbol{\beta_3}^*$ denotes the reduced set which does not include β_0. Thus,

$$\boldsymbol{\beta_3} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}, \text{ and, } \boldsymbol{\beta_3}^* = \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}. \tag{4.3}$$

4.3 Parameter estimation

In developing model (4.2), Coull and Agresti (2003) provide an outline of various schemes to undertake maximum likelihood estimation of parameters. Unless the model parameters can be isolated in the likelihood function, such an approach is challenging and usually requires, depending on the specific model form, special numerical tools to integrate out the random effects. One such model form is that given by (4.1) for the capture-recapture setting, where isolating the model parameters of the likelihood function is impossible. Its likelihood (see Section 1.1.3) is the multinomial likelihood

$$L(N, \boldsymbol{\mu}(\boldsymbol{\theta}); \mathbf{n}_{\text{incomp}}) = \frac{N!}{(N - n_{obs})! \prod_{i=1}^{d} n_i!} \left(1 - \sum_{i=1}^{d} \frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{N - n_{obs}} \prod_{i=1}^{d} \left(\frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{n_i}, \tag{4.4}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi})$ is given by model (4.1) and the indexing is assumed to be over all d cells of the incomplete contingency table, which, for three sources, is shown in Table 4.1. We use $\mathbf{n}_{\text{incomp}}$ to represent the data of the incomplete contingency table. Therefore, in the three-source case, $\mathbf{n}_{\text{incomp}}$ represents the 7 observed cell entries of Table 4.1.

There are specific challenges associated with maximum likelihood (ML) estimation for model (4.1) under likelihood (4.4). First, as indicated in Chapter 3, one such challenge is associated with working with a model on marginal means but a likelihood on cell means. Model (4.1) is a linear model of the model parameters $\boldsymbol{\theta}$, for the marginal means $\mathbf{m}$, but a nonlinear function of $\boldsymbol{\theta}$ for the cell means $\boldsymbol{\mu}$, since the equivalent model on the cell means is given by

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \mathbf{A}^{-1} \exp\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi}\right).$$

Unlike joint log-linear models on cell means (see Section 1.2.1) for which the canonical log link enforces non-negativity of the cell means, it is necessary to enforce such constraints explicitly for model (4.1). That is

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi}) \geq \mathbf{0}. \tag{4.5}$$

Furthermore, μ_{unobs}, the cell mean corresponding to the unobserved cell must be non-negative. That is

$$\mu_{unobs} \geq 0. \tag{4.6}$$

In the multinomial likelihood given by (4.4) it is necessary to ensure that the cell means, including μ_{unobs}, sum to N, or equivalently

$$\mu_{unobs} + \sum_{i=1}^{d} \mu_i = N. \tag{4.7}$$

Thus, ML estimation involves maximization of a nonlinear function of $\boldsymbol{\theta}$ given by likelihood (4.4), subject to nonlinear inequality constraints on $\boldsymbol{\theta}$ given by (4.5)- (4.7). A further constraint imposed by the likelihood is that $N \geq n_{obs}$.

From a computational viewpoint, Bayesian MCMC offers a flexible and convenient approach to deal with the challenges of enforcing constraints when dealing with a function which is is non-linear in the model parameters. Moreover, from an inferential viewpoint, it is meaningful to treat both $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ as random variables and perform

inference which depends on updating prior distributions according to the information contained in the observed data, as quantified by the likelihood. Thus, we are able to make meaningful probability statements about all parameters not only the random effects, as would be the case with a traditional maximum likelihood approach. In such a way, the so-called random effects are not so different from the so-called fixed effects of model (4.1). In this chapter and the following chapter we adopt a Bayesian approach for parameter estimation.

The Bayesian paradigm provides a natural framework in which to accommodate random effects, to incorporate constraints and deal with the challenges posed by the nature of the likelihood. Although the known relationship between the multinomial and Poisson likelihoods (see Section 1.1.3) allows the use of the latter, we choose to work with the multinomial likelihood in which N is considered the quantity of interest. Since MCMC is able to account for the specific challenges of parameter estimation associated with our model and likelihood, it is not necessary to use the equivalence, whereas with frequentist joint log-linear modelling and marginal log-linear modelling there are considerable computational advantages to adopting the Poisson likelihood over the multinomial (see Chapter 3).

A marginal model with random effects is a new idea in the field of capture-recapture. In fact, the use of random effects for the generalized log-linear model has not been discussed extensively in the literature. A known single reference is Coull and Agresti (2003), who did not discuss the use of a Bayesian approach. Only two applications of marginal modelling to capture-recapture data have been found (Bartolucci & Forcina, 2001 and Bartolucci & Forcina, 2006) and, specifically, no application of marginal models with random effects.

4.4 Bayesian model formulation

As with all Bayesian models, a likelihood and prior structure must be specified, together with the corresponding model on the parameters (see Section 1.4.1). Here we describe all components of the Bayesian model. The multinomial likelihood is repeated here for completeness.

Likelihood

$$L(N, \boldsymbol{\mu}(\boldsymbol{\theta}); \mathbf{n}_{\text{incomp}}) = \frac{N!}{(N - n_{obs})! \prod_{i=1}^{d} n_i!} \left(1 - \sum_{i=1}^{d} \frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{N - n_{obs}} \prod_{i=1}^{d} \left(\frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{n_i},$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi})$ is given by the model.

Model

$$\begin{aligned} \log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi} \\ \implies \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi}) &= \mathbf{A}^{-1} \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi}) \end{aligned}$$

$$\begin{aligned} \text{subject to } \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi}) &\geq \mathbf{0}, \\ \mu_{unobs} &\geq 0 \\ N = \mu_{unobs} + \sum_{i=1}^{d} \mu_i & \\ N &\geq n_{obs} \end{aligned} \tag{4.8}$$

Hierarchical prior structure

$$
\begin{aligned}
\boldsymbol{\xi}|\Sigma &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &\sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_0^{-1}, \mathbf{M}) \\
\beta_0 &\sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}) \\
\beta_i &\sim_{ind} N(\mu_{\beta_i}, \sigma^2_{\beta_i}), \text{ for } i = 1, \ldots K,
\end{aligned}
$$

where $\boldsymbol{\Sigma}_0^{-1}$, $\mathbf{M}$, $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_i}$, for $i = 1, \ldots K$, are fixed. We note that the prior structure is expressed in terms of the model components, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. The model components $\boldsymbol{\beta}$ are chosen rather than their corresponding N and marginal means $\mathbf{m}$ so that the range is given by $\mathbb{R}$ rather than $\mathbb{Z}^+$, as would be the case for a prior on N.

In practice, as will be seen in the following chapter, we assume *a priori* that the random effects are independent. In such a case the prior variance matrix $\boldsymbol{\Sigma}$ has all off-diagonal entries set to zero. Thus, if all random effects are assumed to have common variance, it is possible to simplify the prior form for the random effects as follows

$$
\begin{aligned}
\xi_i|\sigma^2_\xi &\sim_{iid} N(0, \sigma^2_\xi), \text{ for all random effects}, i = 1, \ldots, r \\
\sigma^2_\xi &\sim \text{scaled inv} - \chi^2(\nu_0, s_0^2) \\
\beta_0 &\sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}) \\
\beta_i &\sim_{ind} N(\mu_{\beta_i}, \sigma^2_{\beta_i}), \text{ for } i = 1, \ldots K,
\end{aligned} \tag{4.9}
$$

where ν_0 and s_0^2 are fixed and the parameterization of the scaled inverse-χ^2 distribution is described in Section 4.5.1, with density plots of the scaled inverse-χ^2 distribution given in Figure 5.8 (as per the scaled inverse-χ^2 formulation of Gelman et al., 2004).

At this stage, the hierarchical prior structure is not extended to include an additional level of prior information on the hyper-parameters themselves; rather, they

are assumed fixed. Motivation for such an extension would arise should we observe sensitivity of the posterior distributions to the values at which the highest level hyper-parameters are fixed. In practice they are fixed at levels such that the priors are noninformative. For example, large values of $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_i}$ are used in order that the prior distributions on the components of $\boldsymbol{\beta}$ remain noninformative.

Remark 4.4 Implied priors.

In working with a log reparameterization of N and $\mathbf{m}$, we must be attentive to the nature of the prior structure. For example, for the model parameter $\beta_0 = \log N$ a normal prior on β_0 corresponds to a log-normal prior on N. We note that although N is integer-valued, it is treated as a continuous random variable in the Bayesian parameterization in terms of β_0, just as with the frequentist joint log-linear models in which n_{unobs} is parameterized as $\log n_{unobs}$ and treated as a continuous parameter. Such a log-normal prior on N is a continuous, right-skewed distribution that places most mass in the left tail of the distribution. The prior density of N is given by

$$f_N(N; \mu_{\beta_0}, \sigma^2_{\beta_0}) = \frac{\exp(-(\log N - \mu_{\beta_0})^2/2\sigma^2_{\beta_0})}{N\sigma_{\beta_0}\sqrt{2\pi}} \tag{4.10}$$

with corresponding (implied) prior mean, variance and skewness given by

$$\begin{aligned} \mathbb{E}[N] &= \exp\left(\mu_{\beta_0} + \frac{\sigma^2_{\beta_0}}{2}\right) \\ \text{Var}[N] &= (e^{\sigma^2_{\beta_0}} - 1)e^{2\mu_{\beta_0} + \sigma^2_{\beta_0}} \\ \text{Skew}[N] &= (e^{\sigma^2_{\beta_0}} + 2)\sqrt{e^{\sigma^2_{\beta_0}} - 1}. \end{aligned} \tag{4.11}$$

Thus, it is apparent from the form of $\mathbb{E}[N]$ that placing a vague prior on β_0 by increasing variance $\sigma^2_{\beta_0}$ will have the effect of artificially increasing the prior mean on N. Such considerations indicate that summary measures such as posterior means must be taken for the native parameters, not the reparametrized version.

For joint log-linear models in which the model is on the cell means rather than the marginal means of our model, King and Brooks (2001b) have considered the prior structure implied on reparameterizations of the model parameters.

Full posterior

By Bayes' Theorem, the full joint posterior of all model parameters, which we denote by $p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2 | \mathbf{n}_{\text{incomp}})$, is given by

$$p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2 | \mathbf{n}_{\text{incomp}}) = \frac{L(N, \boldsymbol{\mu}(\boldsymbol{\theta}); \mathbf{n}_{\text{incomp}}) p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2)}{p(\mathbf{n}_{\text{incomp}})}, \tag{4.12}$$

where $p(\mathbf{n}_{\text{incomp}})$ is referred to in the literature as the integrated (data) likelihood. The joint prior of all parameters $p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2)$ is obtained using the hierarchical prior structure of (4.9) as follows

$$p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2) = p(\beta_0) \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{r} p(\xi_i | \sigma_\xi^2) \times p(\sigma_\xi^2),$$

with dependence on the fixed hyper-prior parameters of (4.9) suppressed.
From (4.12), the following proportional relationship holds

$$\begin{aligned} & p(\beta_0, \boldsymbol{\beta}^\star, \boldsymbol{\xi}, \sigma_\xi^2 | \mathbf{n}_{\text{incomp}}) \propto \\ & \frac{N!}{(N - n_{obs})! \prod_{i=1}^{d} n_i!} \left(1 - \sum_{i=1}^{d} \frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{N - n_{obs}} \prod_{i=1}^{d} \left(\frac{\mu_i(\boldsymbol{\theta})}{N}\right)^{n_i} \\ & \times \exp\left(\frac{-1}{2\sigma_{\beta_0}^2}(\beta_0 - \mu_{\beta_0})^2\right) \times \prod_{all \beta_k} \exp\left(\frac{-1}{2\sigma_{\boldsymbol{\beta}_k}^2}(\beta_k - \mu_{\beta_k})^2\right) \\ & \times \prod_{all \xi_j} \frac{1}{\sigma_\xi} \exp\left(\frac{-1}{2\sigma_\xi^2} \xi_j^2\right) \times \frac{1}{(\sigma_\xi^2)^{(\nu_0/2+1)}} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma_\xi^2}\right), \end{aligned} \tag{4.13}$$

where $\mathbf{n}_{\text{incomp}}$ denotes the data contained in the incomplete contingency table and $N = \exp(\beta_0)$.

4.5 MCMC specifications

In this section we describe in detail an MCMC scheme to sample from the joint posterior distribution given by (4.13). Examination of the joint posterior (4.13) shows that it is only possible to obtain a full conditional in closed form for the random effects variance σ_ξ^2. Thus, all other parameters must be updated using an alternative principle to that of a Gibbs update (see Section 1.4). An appropriate approach is to use a Metropolis-Hastings proposal for each of the other parameters.[1] The specific details of the full MCMC scheme will first be described here in words and then described symbolically. This scheme will be used in the next chapter to implement a full Bayesian data analysis of a real data set.

The scheme proceeds in the following order (details are provided below):

- Sample an integer value of N conditional on the current value of N via a Metropolis-Hastings step. From the updated value of N, obtain the corresponding value of $\beta_0 = \log N$
- Sample the components of $\boldsymbol{\beta}^\star$ via a Metropolis-Hastings step conditional on the current value of $\boldsymbol{\beta}^\star$.
- For the random effects, the scheme will depend on the specific form of $\mathbf{Z}$ and $\boldsymbol{\xi}$. Details of one such case will be provided in the next chapter.
- Perform a Gibbs update on the random effects variance, σ_ξ^2. This is possible since the scaled inverse-χ^2 prior is conjugate for the normal "data" of the random effects, in the sense described below.

All Metropolis-Hastings proposals are Gaussian proposals, except for that of N, which is a symmetric discrete uniform. In so doing, the symmetry of the Gaussian distrib-

[1]The proposals are all symmetric, thus the proposals are all like the Metropolis algorithm described in Section 1.4.1. We use the more general Metropolis-Hastings terminology throughout.

ution ensures that the forward and backward probabilities are equal and thus cancel in the transition probability calculations.

Remark 4.5 Constraints.
In order to enforce the model constraints given by $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\xi}) \geq \mathbf{0}$ and $\mu_{unobs} \geq 0$ of (4.8), the value of $\boldsymbol{\mu}(\boldsymbol{\theta})$ is obtained after each Metropolis-Hastings proposal. Should $\boldsymbol{\mu}(\boldsymbol{\theta})$ violate the constraints, and thus lie outside of the feasible parameter space, the proposed parameter values are rejected. Likewise, when a value of N is proposed, it is verified that the proposed value is within the feasible parameter space defined by $N \geq n_{obs}$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) \geq \mathbf{0}$ and $N = \mu_{unobs} + \sum_{i=1}^{d} \mu_i$.

4.5.1 Details of the MCMC simulation scheme

Consider the chain at iteration $(t-1)$ with parameters denoted by $N^{(t-1)}$ (equivalently $\beta_0^{(t-1)}$), $\boldsymbol{\beta}^{\star(t-1)}, \boldsymbol{\xi}^{(t-1)}$ and $\sigma_{\xi}^{2,(t-1)}$. For each of the components of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\xi})$, Metropolis-Hastings updates will be used as described above. The details of the update are provided here.

Details: N (equivalently β_0)

Draw N^{prop} from Discrete $U(N^{t-1} - N_{\text{width}}, N^{t-1} + N_{\text{width}})$, where N_{width} is chosen so that acceptance rates are in the range 15% - 30% (for justification, see the theoretical work described on page 55, Chapter 3 of Gilks et al., 1996). Calculate the Metropolis-Hastings ratio, r_N (see Chapter 1 for details, in particular the description of the Metropolis algorithm in Section 1.4.1), and update N as follows

$$N^{(t)} = \begin{cases} N^{\text{prop}}, & r_N \geq u \\ N^{(t-1)}, & r_N < u \end{cases}$$

where $u \sim U(0,1)$.

Details: $\boldsymbol{\beta}^{\star}$

Draw $\boldsymbol{\beta}^{\star(\text{prop})}$ from $MVN(\boldsymbol{\beta}^{\star(t-1)}, \Sigma)$, where Σ is a variance-covariance matrix controlled by the user, again so that acceptance rates are in the range 15% -30%. Calculate the Metropolis-Hastings ratio, $r_{\boldsymbol{\beta}^{\star}}$ (again, see Chapter 1 for details), and update $\boldsymbol{\beta}^{\star}$ as follows

$$\boldsymbol{\beta}^{\star(t)} = \begin{cases} \boldsymbol{\beta}^{\star(\text{prop})}, & r_{\boldsymbol{\beta}^{\star}} \geq u \\ \boldsymbol{\beta}^{\star(t-1)}, & r_{\boldsymbol{\beta}^{\star}} < u \end{cases} \tag{4.14}$$

where $u \sim U(0,1)$.

Details: $\boldsymbol{\xi}$

One such case is presented in the following chapter.

Details: σ_{ξ}^2

From the form of the joint posterior distribution (4.13), it is clear that the full conditional of σ_{ξ}^2 exists in closed form. It is the standard case of normal data with known mean and unknown variance, with a scaled inverse-χ^2 prior on the variance term. The random effects themselves, $\boldsymbol{\xi}$, are our 'data'. It is straight-forward to show that the scaled inverse-χ^2 prior is conjugate for normal data. The general scheme is given here.

$$\begin{aligned} \sigma_{\xi}^2 &\sim \text{scaled inv} - \chi^2(\nu_0, s_0^2) \\ \xi_i &\sim_{iid} N(\mu, \sigma_{\xi}^2), \ \ for\ i = 1, \ldots, r \end{aligned}$$

which corresponds to

$$\begin{aligned} p(\sigma_{\xi}^2) &\propto (\sigma_{\xi}^2)^{-(\frac{\nu_0}{2}+1)} \exp\left(\frac{-\nu_0 s_0^2}{2\sigma_{\xi}^2}\right) \\ p(\boldsymbol{\xi}|\sigma_{\xi}^2) &\propto (\sigma_{\xi}^2)^{-\frac{n}{2}} \exp\left(\frac{-n\nu}{2\sigma_{\xi}^2}\right), \end{aligned}$$

where $\nu = \sum \frac{(\xi_i - \mu)^2}{r}$, r is the dimension of the random effects vector $\boldsymbol{\xi}$, μ the assumed common mean of the random effects.

In order to understand the nature of such a scaled inv-χ^2 prior, consider the value of its mode, expectation and variance. Suppose θ, some random variable, is distributed as $\theta \sim \text{scaled inv} - \chi^2(\nu, s^2)$. Then

$$\begin{aligned} \text{mode}(\theta) &= \frac{\nu}{\nu+2} s^2 \\ \mathbb{E}[\theta] &= \frac{\nu}{\nu-2} s^2, \qquad \text{for } \nu > 2 \\ \text{Var}[\theta] &= \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4, \qquad \text{for } \nu > 4. \end{aligned} \tag{4.15}$$

Thus, for small s_0^2, the prior mode of σ_ξ^2 will be close to 0. It is known that, for the standard random effects model described above, a prior mode close to zero can cause unexpected results in terms of posterior inference. We thus seek to ensure that the mode is not too close to 0.

For this prior and data we obtain a conjugate posterior given by

$$p(\sigma_\xi^2 | \boldsymbol{\xi}) \propto (\sigma_\xi^2)^{-(\frac{\nu_0 + r}{2} + 1)} \exp\left(-\frac{r\nu + \nu_0 s_0^2}{2\sigma_\xi^2} \right),$$

which is proportional to

$$\sigma_\xi^2 \sim \text{scaled inv} - \chi^2(\nu_0 + r, \tfrac{\nu_0 s_0^2 + r\nu}{\nu_0 + r}).$$

Since the full conditional distribution exists in closed form, a Gibbs update can be made for σ_ξ^2 after the joint Metropolis-Hastings update is performed for all the other parameters, i.e. for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\xi})$.

4.6 Bayesian Model Developments

In the following chapter we present a random effects formulation of the form (4.1) in which the CIDs are treated as random effects. The general principles and MCMC

simulation scheme developed in this chapter will be adopted and modified appropriately, according to the specifics of the model. We will perform a Bayesian analysis of the four-source diabetes data set of Bruno et al. (1994), as analyzed in Chapter 3.

Chapter 5

Bayesian Random Effects Modelling: CID Formulation

5.1 Introduction

In this chapter we present a specific form of the general Bayesian random effects model introduced in Chapter 4 in which we assume the CIDs are random. The model form is the Bayesian version of the marginal log-linear model introduced in Chapter 3 with random rather than fixed CIDs, as was the case in Chapter 3. In so doing, we achieve parameter reduction, which is desirable given the limited number of data points available for capture-recapture data.

In Section 5.2 we describe the specific form of the model relative to the general form (4.1). In Section 5.3 we describe specifics of the MCMC simulation scheme to lead to Section 5.4, in which we present results of a full Bayesian analysis for the four-source diabetes data set of Bruno et al. (1994), analyzed in Chapter 3. The analysis consists in demonstrating properties of the Bayesian model of independence, followed

by an analysis via a reduced model in which only two CIDs are included in the model (we describe below the motivation for the choice of the specific model). We conclude with an analysis via the fully parameterized model in which all CIDs are included in the model. We discuss the problem of nonidentifiability of model parameters in the case of the full model, and compare to the analysis of the reduced model for which there are no parameter identifiability issues. In Section 5.5 we conclude with a discussion, in particular related to the relationship between the random effects variance and N, which is explored via the data analysis of this chapter.

5.2 CID-based random effects models

Definition 5.1 *Let $\mathcal{Q}$ be a set of sources and $K = |\mathcal{Q}|$. Then the following system of $2^K - 1$ equations constitutes the CID random effects marginal log-linear model for the K sources of $\mathcal{Q}$.*

$$\log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},$$
$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}),$$

where $\boldsymbol{\gamma}$ is the $r \times 1$ vector of all non-single source CIDs, (see Definition 2.11), with the $d \times r$ design matrix $\mathbf{Z}$ given by the following concatenation:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{0} \\ \mathbf{G^{-1}} \end{bmatrix},$$

where $\mathbf{G^{-1}}$ is the $(d - K) \times r$ matrix described in Remark 2.20 and $\mathbf{0}$ is the $K \times r$ matrix of zeros. The dimension of $\boldsymbol{\gamma}$ is given by $r = d - K$, for $d = 2^K - 1$. All other matrices and vectors are defined as per the discussion immediately following (4.1).

Example 5.1 Consider the three-source capture-recapture setting for sources A, B

and C. Then the three-source CID random effects marginal model is given as follows

$$\log(\mathbf{m}) = \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},$$
$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}),$$

where, in particular

$$\gamma_{AB} \sim N(0, \sigma_\gamma^2)$$
$$\gamma_{AC} \sim N(0, \sigma_\gamma^2)$$
$$\gamma_{BC} \sim N(0, \sigma_\gamma^2)$$
$$\gamma_{ABC} \sim N(0, \sigma_\gamma^2),$$

with

$$\mathbf{m} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix}; \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \boldsymbol{\mu} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix}.$$

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 \\ -2 & 1 & 1 & 1 \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}; \mathbf{Z} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{BC} \\ \gamma_{ABC} \end{bmatrix}.$$

Compare this form to (3.3), which is the most general form of the fixed effects three-source marginal log-linear model. We note that the sole difference with the form in Chapter 3 is that the CIDs are treated as fixed rather than random effects. As stated in Chapter 3, the model given by (3.3) is over-parameterized. In that case, parameter reduction is achieved by setting combinations of CIDs equal to zero, thus omitting those terms from the model. Even for the most general form of the random effects model for three sources given above, the model is no longer over-parameterized. There are $2^3 - 1 = 7$ cell entries available in the three-source contingency table (see Table 1.2) with which to estimate the 5 model parameters, namely β_0, β_A, β_B, β_C and σ^2_γ. Nonetheless, from an inferential point of view, estimating 5 parameters with only 7 data points is a demanding setting. The Bayesian paradigm provides a means to incorporate information via prior distributions to lighten the burden of parameter estimation, so to speak, on the 7 cell entries of the three-source incomplete contingency table.

Remark 5.1 The full expression of all matrices for the three and four-source cases are presented in Appendix A.

5.3 Bayesian CID-based random effects models

The likelihood and prior structure are as those described in general terms in Section 4.4. Likewise, the MCMC scheme used to generate samples from the joint posterior distribution of the model parameters is the same as the general form described in Chapter 4.

Details of the MCMC scheme

Here we describe the specific MCMC update performed for the random effects vector $\boldsymbol{\gamma}$ of CIDs, with corresponding design matrix $\mathbf{Z}$ described above. For convenience,

subdivide the vector of random effects in the following manner

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_{\text{pair}} \\ \boldsymbol{\gamma}_{3-\text{way}} \\ \vdots \\ \boldsymbol{\gamma}_{\text{K}-\text{way}} \end{bmatrix}. \tag{5.1}$$

Consider the MCMC chain at iteration $(t-1)$ for random effects vector $\boldsymbol{\gamma}^{(t-1)}$. Metropolis-Hastings updates are made in blocks for the elements of $\boldsymbol{\gamma}^{(t-1)}$, with blocks corresponding to the number of sources, as defined by (5.1). We proceed as follows:

- Pairwise random effects $\boldsymbol{\gamma}_{\text{pair}}$
  Draw $\boldsymbol{\gamma}_{\text{pair}}^{\text{prop}}$ from $N(\boldsymbol{\gamma}_{\text{pair}}^{(t-1)}, \sigma_{\text{pair}}^2 \mathbf{I})$. Calculate the Metropolis-Hastings ratio r_{pair} and update $\boldsymbol{\gamma}_{\text{pair}}$ as follows

$$\boldsymbol{\gamma}_{\text{pair}}^{(t)} = \begin{cases} \boldsymbol{\gamma}_{\text{pair}}^{\text{prop}}, & r_{\text{pair}} \geq u \\ \boldsymbol{\gamma}_{\text{pair}}^{(t-1)}, & r_{\text{pair}} < u \end{cases}$$

  where $u \sim U(0,1)$.

- Three-way random effects $\boldsymbol{\gamma}_{3\text{way}}$
  Draw $\boldsymbol{\gamma}_{3\text{way}}^{\text{prop}}$ from $N(\boldsymbol{\gamma}_{3\text{way}}^{(t-1)}, \sigma_{3\text{way}}^2 \mathbf{I})$. Calculate the Metropolis-Hastings ratio $r_{3\text{way}}$ and update $\boldsymbol{\gamma}_{3\text{way}}$ as follows

$$\boldsymbol{\gamma}_{3\text{way}}^{(t)} = \begin{cases} \boldsymbol{\gamma}_{3\text{way}}^{\text{prop}}, & r_{3\text{way}} \geq u \\ \boldsymbol{\gamma}_{3\text{way}}^{(t-1)}, & r_{3\text{way}} < u \end{cases}$$

  where $u \sim U(0,1)$.

- Proceed in a similar manner in block form up to the single K-way random effect.

Remark 5.2 In each case, the proposal variances (i.e. σ_{pair}^2, $\sigma_{3\text{way}}^2$, etc.) are adjusted by the user so that acceptance rates for each block are in the range 15% -30% (see

page 55, Chapter 3 of Gilks et al., 1996 for some justification). Model constraints are enforced as per Remark 4.5. Further, we note that for the CID random effects formulation of the current chapter, Σ, the proposal variance-covariance matrix of $\boldsymbol{\beta}^{\star}$ is a diagonal matrix with zero off-diagonal entries.

5.4 Data analysis

In this section we will revisit the diabetes data set of Bruno et al. (1994), as first introduced in Chapter 3 and analyzed using frequentist marginal log-linear models in Section 3.5.2. The data are presented again for reference in Table 5.1. Recall that for this particular data set very few of the $n_{obs} = 2069$ observed individuals were observed in sources B and D with marginal single-source observed counts given by $n_A = 1754$, $n_B = 452$, $n_C = 1135$ and $n_D = 173$.

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	58	46	14	8
	D_{No}	157	650	20	182
C_{No}	D_{Yes}	18	12	7	10
	D_{No}	104	709	74	?

Table 5.1: Diabetes data set of Bruno et al. (1994).

We will analyze the diabetes data set using the Bayesian random effects marginal model for the four-source capture-recapture setting. The analysis consists in fitting two models: first we analyze the data set via a specific reduced model in which only 2 of the 11 unknown CIDs are included in the model, and secondly we analyze the data set via the full Bayesian model which includes all random effects (i.e. none of them

are fixed equal to zero). We demonstrate that the parameter identifiability issues associated with fitting the full model are not encountered when fitting the reduced model. We note that the specific form of the reduced model selected for this chapter is that containing the 2 pairwise CIDs given by γ_{BD} and γ_{CD}. These two pairwise CIDs were selected based on preliminary analysis of the diabetes data set via both calculation of the 6 Petersen estimates as given in Table 3.2.

In order to demonstrate that the Bayesian formulation of the marginal model performs sensibly for a known case, we will first fit the independence model, followed by analysis for a single hierarchical prior structure for the reduced model. We will then present results from a range of prior distributions for the reduced model in order to determine the sensitivity of the posterior distribution to different prior parameters, in particular to the prior distributions on the main parameters of interest, namely β_0 and σ^2_γ. The same analysis will then be repeated for the full model; that is we first present results from the analysis using a single hierarchical prior structure followed by results from a sensitivity analysis.

5.4.1 Modelling approach

In all analyses described in the following sections, we generate two chains using the MCMC scheme described in Section 4.5. The two chains, each of length 10,020,000, start from two sets of dispersed starting values. In so doing, we wish to reduce dependence of the posterior on the starting values (see Section 1.4.1). Each chain is thinned by 200 to yield a sample of 51,000 draws from the joint posterior distribution given by (4.12). We then remove the first 1000 iterations of each sample (i.e. burn-in) to yield a sample of size 50,000 from the joint posterior distribution for each chain. Although thinning by a factor of 200 reduces dependence on starting values, we nonetheless remove the first 1000 iterations of the thinned chain for good measure. We combine the two samples each of size 50,000, corresponding to the two MCMC

runs, to obtain a single sample of size 100,000 from the full joint posterior distribution of all parameters.

Remark 5.3 Preliminary analyses for both the reduced and full models demonstrated that there was high posterior correlation amongst model parameters, in particular between β_0 and the components of γ. As a result, a large thinning factor of 200 was used to reduce posterior correlation amongst model parameters. Although, it is desirable to minimize posterior correlations between model parameters in order to obtain representative samples from the joint posterior distribution, with modern computational power it is not a major drawback if reparameterizations of the model do not yield such features. If the chain is run for long enough, a reasonable sample will be obtained.

We choose posterior samples of size 50,000 for each chain in line with the approach adopted by Fienberg et al. (1999), who also faced similar issues of high posterior correlation when working with their Bayesian formulation of the Rasch model. Tables 5.2 and 5.3 presents such posterior correlations for a single thinned (by 200) MCMC chain for the main parameters of interest β_0 and σ^2_γ relative to $\boldsymbol{\gamma}$ for two different models. First, Table 5.2 presents such correlations (including those with the elements of $\boldsymbol{\beta}^\star$) for the reduced model which contains only two pairwise CIDs, namely γ_{BD} and γ_{CD}; we note again that all other CIDs are fixed at zero rather than being estimated like γ_{BD} and γ_{CD}. We described above the motivation for the choice of such a model. Secondly, Table 5.3 presents such correlations for the full model in which all CIDs are included in the model with none fixed at zero. (Note that we have not removed the lower-half of each of the two tables to remove the repeated symmetric correlations.)

The posterior correlations are reasonable for the reduced model (see Table 5.2). The largest correlation of 0.241 is between β_B and β_D, whilst the correlation between the two random effects, γ_{BD} and γ_{CD}, and each of β_0 and σ^2_γ is at most 0.134, which

Table 5.2: Posterior correlations between β_0, σ^2_γ, γ_{BD}, γ_{CD} and $\boldsymbol{\beta}^\star$ for the reduced model containing only two CIDs

	β_0	σ^2_γ	γ_{BD}	γ_{CD}	β_A	β_B	β_C	β_D
β_0	1.00	0.004	0.096	0.134	0.003	0.004	0.012	-0.017
σ^2_γ	0.004	1.000	0.016	-0.007	0.003	0.001	-0.006	-0.004
γ_{BD}	0.096	0.016	1.000	-0.065	-0.002	-0.231	-0.001	-0.177
γ_{CD}	0.134	-0.007	-0.065	1.000	-0.003	-0.007	-0.153	-0.112
β_A	0.003	0.003	-0.002	-0.003	1.000	-0.008	-0.003	-0.009
β_B	0.004	0.001	-0.231	-0.007	-0.008	1.000	0.003	0.241
β_C	0.011	-0.006	-0.001	-0.153	-0.003	0.003	1.000	0.107
β_D	-0.017	-0.004	-0.177	-0.112	-0.009	0.241	0.107	1.000

again is perfectly reasonable. The main effect parameter β_A is almost uncorrelated with the random effects γ_{BD} and γ_{CD}, whereas the main effects β_B and β_D corresponding to sources B and D are correlated with the corresponding random effect γ_{BD}, whilst β_C and β_D are correlated with the corresponding random effect γ_{CD}.

The posterior correlations for the full model (see Table 5.3) are not as reasonable. In fact, we observe very high correlation (close to 1) between the components of the random effects vector $\boldsymbol{\gamma}$ and β_0, and consequently between the random effects themselves, for the full model . The correlation between the components of $\boldsymbol{\gamma}$ and σ^2_γ are lower and perfectly equal at a reasonably low value of 0.34 (modulo the sign) to the posterior correlation between β_0 and σ^2_γ. Such high posterior correlations arise because of nonidentifiability of model parameters (see discussion below). Despite such high correlations, the analysis via the full model serves to demonstrate certain features of a non-identifiable Bayesian model. We note in addition that the corresponding posterior correlations between the components of $\boldsymbol{\beta}^*$, and of each component with β_0 and σ^2_γ, are at most 0.24 (table not presented).

The trace plots presented in Figures 5.1 and 5.2 for the reduced and full model, respectively, are for a single set of starting values for a chain thinned by a factor of 200. The chains correspond to the single analyses to be presented in Sections 5.4.3 and 5.4.4 below. Both figures further confirm the correlation patterns observed in Tables 5.2 and 5.3. Figure 5.1 demonstrates that there is good mixing of the chain over the parameter space for each of β_0, σ_γ and the two random effects γ_{BD} and γ_{CD}, whereas we observe serial correlation in Figure 5.2 in each of the corresponding plots of β_0, γ_{BD} and γ_{CD} and for the other 9 random effects (plots not presented here) of the full model.

In all cases, C++ code was used to implement the MCMC algorithm. The code can be found in Appendix H. Jumping proposal parameters for the Metropolis-Hastings steps were adjusted in order to achieve acceptance rates in the range of 15% - 30%,

Table 5.3: Posterior correlations between β_0, σ^2_γ and $\boldsymbol{\gamma}$ for the full Bayesian model

	β_0	σ^2_γ	$\boldsymbol{\gamma}$										
β_0	1.00	0.34	0.98	0.99	0.96	0.96	0.90	0.94	-0.99	-0.96	-0.98	-0.94	0.97
σ^2_γ	0.34	1.00	0.34	0.34	0.33	0.34	0.34	0.34	-0.34	-0.34	-0.34	-0.34	0.34
$\boldsymbol{\gamma}$	0.98	0.34	1.00	0.98	0.95	0.96	0.89	0.93	-0.98	-0.96	-0.96	-0.94	0.97
	0.99	0.34	0.98	1.00	0.96	0.96	0.89	0.94	-0.98	-0.96	-0.98	-0.94	0.97
	0.96	0.33	0.95	0.96	1.00	0.92	0.86	0.92	-0.94	-0.94	-0.96	-0.91	0.94
	0.96	0.34	0.96	0.96	0.92	1.00	0.88	0.93	-0.96	-0.94	-0.94	-0.95	0.95
	0.90	0.34	0.89	0.89	0.86	0.88	1.00	0.85	-0.89	-0.87	-0.89	-0.87	0.89
	0.94	0.34	0.93	0.94	0.92	0.93	0.85	1.00	-0.93	-0.92	-0.94	-0.92	0.93
	-0.99	-0.34	-0.98	-0.98	-0.94	-0.96	-0.89	-0.93	1.00	0.96	0.97	0.94	-0.98
	-0.96	-0.34	-0.96	-0.96	-0.94	-0.94	-0.87	-0.92	0.96	1.00	0.95	0.93	-0.97
	-0.98	-0.34	-0.96	-0.98	-0.96	-0.94	-0.89	-0.94	0.97	0.95	1.00	0.93	-0.97
	-0.94	-0.34	-0.94	-0.94	-0.91	-0.95	-0.87	-0.92	0.94	0.93	0.93	1.00	-0.94
	0.97	0.34	0.97	0.97	0.94	0.95	0.89	0.93	-0.98	-0.97	-0.97	-0.94	1.00

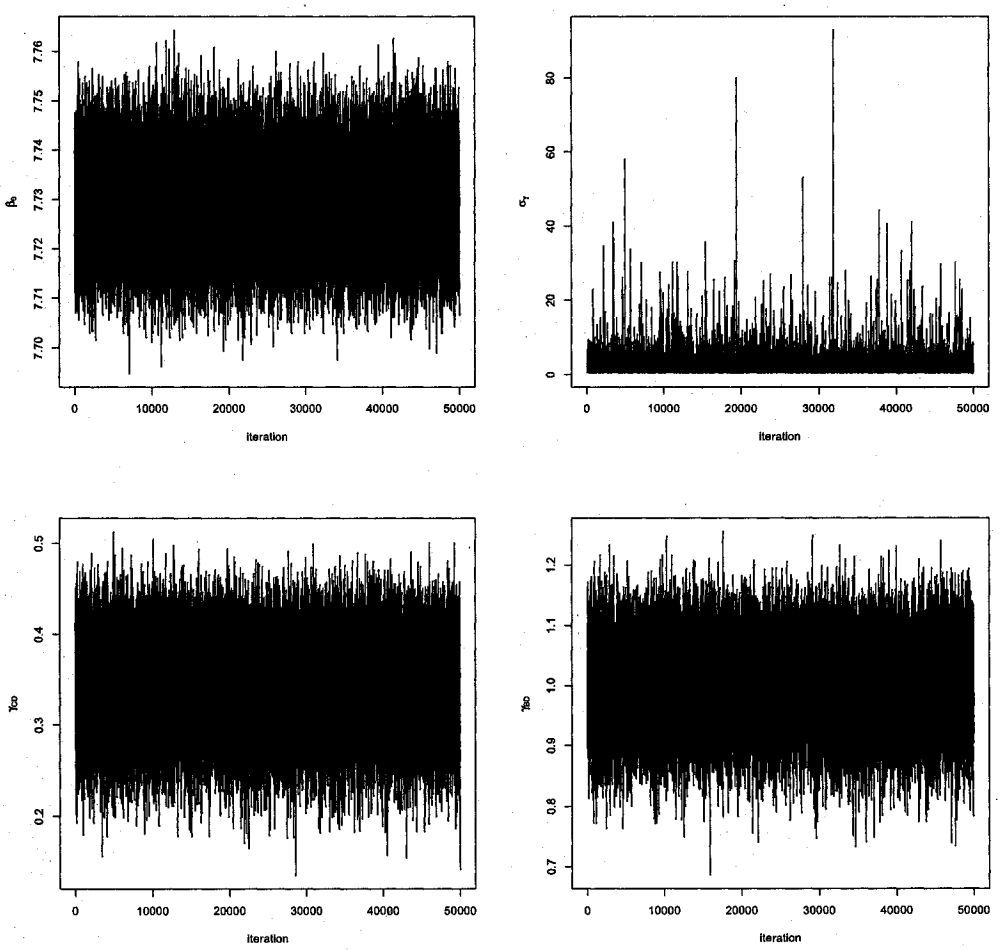


Figure 5.1: Trace plots of β_0, σ_γ, γ_{BD} and γ_{CD} for the reduced model

as described above.

The same vague prior distributions were assumed for the components of the $\beta^\star$ vector for each chain. Preliminary analyses demonstrated minimal sensitivity of the joint posterior of all model parameters to the prior distribution on $\beta^\star$. In particular these preliminary analyses demonstrated that there was minimal sensitivity of the

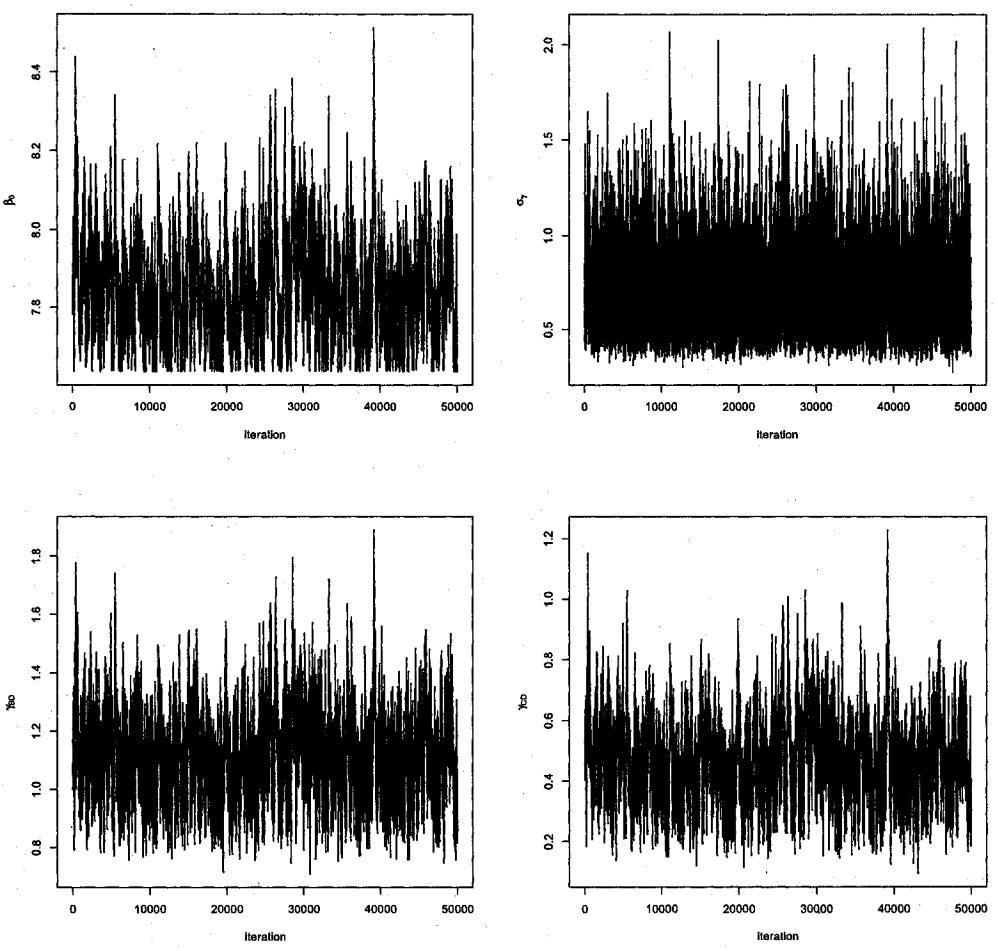


Figure 5.2: Trace plots of β_0, σ_γ, γ_{BD} and γ_{CD} for the full model

posterior distributions of the elements of $\boldsymbol{\beta}^\star$ to their own prior distributions (note that we state this relationship without including a series of sensitivity plots). In fact, they are uniformly close to centered on the corresponding log-marginal counts, given by $\log n_A$, $\log n_B$, $\log n_C$ and $\log n_D$ for the four-source case. Thus, we set the same prior distribution of each element of $\boldsymbol{\beta}^\star$ with prior parameters given by $\mu_{\beta^\star} = 5.15$

and $\sigma^2_{\beta^\star} = 1.74$ for each of the four components of $\beta^\star$, namely for each of β_A, β_B, β_C and β_D. Note that these prior parameters are somewhat arbitrary and correspond to a low marginal mean count of $\exp(5.15) = 172$ for each of the four components.

For the first analysis presented in each of sections 5.4.2, 5.4.3 and 5.4.4, we fix the prior parameters of β_0 at $\mu_{\beta_0} = 8.06$ and $\sigma_{\beta_0} = 0.28$, which roughly corresponds to a prior 90% range on the corresponding N parameter of 2000-5000 (see following discussion).

Remark 5.4 Prior parameters for β_0.

In proposing prior parameters for the $N(\mu_{\beta_0}, \sigma^2_{\beta_0})$ prior on β_0, it is useful to consider the corresponding prior range on $N = \exp \beta_0$. Let N_u and N_l be the upper and lower levels of the nominal 90% prior range we wish to place on the parameter N. In order to obtain the corresponding prior parameters on β_0, i.e. μ_{β_0}, $\sigma^2_{\beta_0}$, we equate N_u and N_l with the exponential of the corresponding 90% bounds of the $N(\mu_{\beta_0}, \sigma^2_{\beta_0})$ prior distribution. Then

$$N_u = \exp(\mu_{\beta_0} + 1.64\sigma_{\beta_0})$$
$$N_l = \exp(\mu_{\beta_0} - 1.64\sigma_{\beta_0}),$$

which leads to

$$\mu_{\beta_0} = (\log N_u + \log N_l)/2$$
$$\sigma_{\beta_0} = (\log N_u - \log N_l)/(2 \times 1.64).$$

Thus, the prior range of 2000-5000 on N corresponds to $\mu_{\beta_0} = (\log 5000 + \log 2000)/2 = 8.06$ and $\sigma_{\beta_0} = (\log 5000 - \log 2000)/(2 \times 1.64) = 0.28$.

5.4.2 Case of Independence: $\sigma^2_\gamma = 0$

In this short section we will demonstrate that the Bayesian random effects model performs as expected for the model of independence. That is, we show that the mar-

ginal posterior distribution for $\beta_0 = \log N$ is centered on $\log \hat{N}_{MLE}$, the log-maximum likelihood estimate of N obtained using the frequentist formulation of the marginal model, in which the components of the γ are treated as fixed effects (see Chapter 3). In so doing, we confirm that the Bayesian modelling approach presented in the current chapter performs as it should in a known setting, thus providing evidence that the modelling approach is reasonable.

It is known by design that the random effects, γ, measure departures from independence. Thus, under independence, the components of γ must be equal to their assumed means, i.e. equal to 0. Such a case of full independence corresponds to a random effects variance, σ^2_γ, of 0. Consequently, simulation from the full posterior distribution (4.13) via MCMC methods is made easier. Since the random effects, γ, are known exactly, it is not necessary to update them in the MCMC scheme described in Section 5.3. The only updating in the MCMC scheme is of N and $\boldsymbol{\beta}^*$ with $\sigma^2_\gamma = 0$ and all components of γ equal to zero.

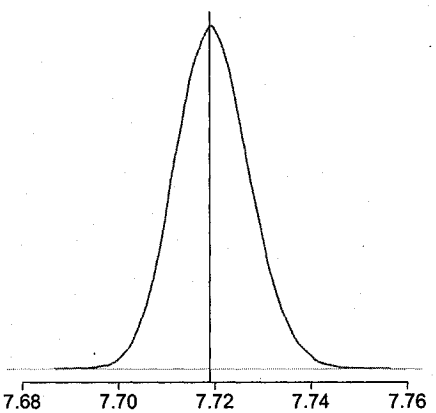


Figure 5.3: Posterior of β_0 for $N(8.06, 0.28)$ prior on β_0 under model of independence.

Figure 5.3 presents the marginal posterior distribution of β_0. Note that the vertical axis is not specified since it is not important for our purposes. We use this plot to serve the purpose of the current section: to confirm that the Bayesian model for independence performs as it should. In Chapter 2, we demonstrated that the marginal log-linear model for independence (i.e. with all components of $\boldsymbol{\gamma}$ fixed at zero in the fixed effects model of Chapter 3) is equivalent to the joint log-linear model for independence. The (unrounded) point estimate for N obtained from the frequentist joint log-linear model for independence[1] is given by $\hat{N} = 2250.601$, with corresponding value $\hat{\beta}_0 = \log \hat{N} = \log(2250.601) = 7.719$. The vertical line of Figure 5.3 is at this value of $\beta_0 = 7.719$. We observe that the symmetric-looking marginal posterior distribution of β_0 is close to perfectly centered on the value of 7.719, thus confirming that the Bayesian marginal model for independence performs as it should.

We note that all the important features of the model and MCMC chain are present in the analysis using the model of full dependence presented in the next section. Thus we include no further details of the analysis using the model of independence.

5.4.3 Case of dependence: $\sigma_\gamma^2 > 0$; a specific form of the reduced model

In this section we consider a model of dependence for which σ_γ^2 is not fixed at zero. The specific model, which henceforth will be referred to as the reduced model versus the full model to be considered in the following section, is that which contains only two CIDs treated as random effects, namely γ_{BD} and γ_{CD}. All other CIDs are fixed at zero, equivalently they are not included in the model, hence the term 'reduced' model.

The 2 pairwise CIDs given by γ_{BD} and γ_{CD} were selected according to preliminary

[1]Obtained using the inbuilt `glm` function in the *R Programming Language* (2004), with Poisson family and log link.

analyses observed via both calculation of the 6 Petersen estimates as given in Table 3.2 and via the full Bayesian analysis to be presented below (in particular as evinced by Figure 5.12 in which we observe posterior distributions of γ_{BD} and γ_{CD} furthest removed from 0). Such analyses suggest that most of the dependence exhibited in the diabetes data set is explained in the two marginal pairs given by $\{B, D\}$ and $\{C, D\}$. The purpose of the current section is to demonstrate that it is possible to obtain identifiable parameter estimates from a model of dependence more complex than the simplest model, that of independence. We note that the goal of the current chapter is not to perform model selection and obtain the best model according to some model selection criterion, rather to demonstrate that a bayesian treatment of an identifiable model provides sensible estimates.

We note that the model containing only γ_{BD} and γ_{CD} is not the model of conditional independence of B and C given D, since by Theorem 2.9 such a model would require $\gamma_{BCD} = -\gamma_{BC}$. Thus, the model considered in the current section is a non-hierarchical model. The model considered here is identifiable (in terms of model parameters) unlike the full model presented in the next section.

Features of analysis for a single prior specification

We first undertake the full Bayesian analysis for a single hierarchical prior structure. The specific prior structure on the parameters of interest, β_0 and σ^2_γ, is given by Table 5.4. Hyperprior values of $\mu_{\beta_0} = 8.06$ and $\sigma^2_{\beta_0} = 0.28$ correspond to a prior 90%

Parameter	β_0		σ^2_γ	
Hyperprior parameter	μ_{β_0}	$\sigma^2_{\beta_0}$	ν_0	s^2_0
Value	8.06	0.28	5	0.5

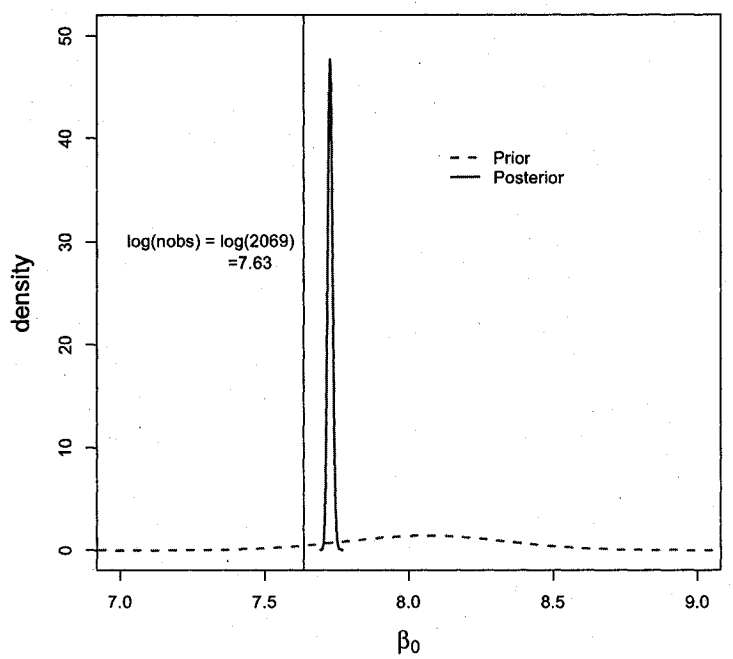
Table 5.4: Hyperprior parameters of β_0 and σ^2_γ for Bayesian analysis

range on N from 2000 to 5000 (as described above in Section 5.4.1). Figure 5.8 shows the form of the scaled-inverse χ^2 prior on σ_γ^2. We will discuss further details of this prior below when we examine a range of prior distributions. For the moment, we select a single prior distribution for σ_γ^2 that is loosely informative and enables mass to be placed near 0 (i.e. near to independence) since, *a priori*, we do not wish to place great mass away from the case of independence since we prefer to be noninformative in our selection of prior distribution. One such prior is defined by $\nu_0 = 5$ and $s_0^2 = 0.5$ (see Figure 5.4 for the spread of this distribution) and is the one selected for the present analysis (see Table 5.4).

Inference on N, β_0 and σ_γ^2

Figure 5.4 shows the marginal posterior distributions of each of β_0 and σ_γ^2. The posteriors represent the effect of updating (via the likelihood, given by (4.4)) the prior distributions of each of β_0 and σ_γ^2 using the observed capture-recapture data of the incomplete contingency table given by Table 5.1. The upper panel of Figure 5.4 shows that the posterior of β_0 is bounded below by $\log n_{obs} = \log 2069 = 7.63$ as per (4.8). The posterior distributions of β_0 and σ_γ^2 are shifted slightly to the left and right, respecitvely with great reduction in variability in the case of β_0, with an increase in variability in the case of σ_γ^2. As expected by conjugacy, the posterior of σ_γ^2 is a scaled inverse-χ^2 distribution.

Table 5.5 provides posterior summaries of β_0 and σ_γ^2, as well as the corresponding posterior summaries of N as obtained from $N = \exp(\beta_0)$ and of σ_γ. The upper and lower bounds of the symmetric 90% and 95% credible intervals are provided, together with the minimum and maximum values in order to examine the range of posterior values of each parameter. It should be noted that, although the quantiles of both N and σ_γ are obtained by direct transformation of those on β_0 and σ_γ^2, respectively, the posterior mean of N and σ_γ cannot be obtained by direct transformation of the



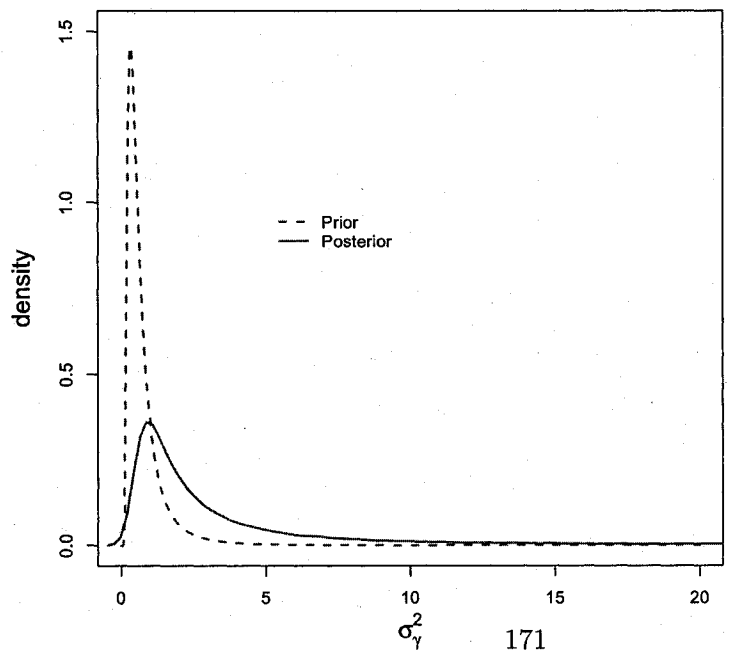


Figure 5.4: Posteriors of β_0 and σ_γ^2 for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2.

	Quantile							
Parameter	0 %	2.5%	5%	50%	95%	97.5%	100%	Mean
β_0	7.695	7.713	7.716	7.729	7.743	7.746	7.768	7.729
N	2197	2236	2243	2273	2305	2312	2363	2273
σ_γ	0.391	0.658	0.723	1.389	4.168	5.543	149.61	1.824
σ_γ^2	0.153	0.433	0.523	1.929	17.372	30.719	22384	6.846

Table 5.5: Posterior summaries for β_0, N and σ_γ for Bayesian analysis

corresponding posterior means of β_0 and σ_γ^2, respectively, since it is not possible to reverse the order of a nonlinear function (in this case the exponential operator and square root operator) and the expectation operator. Thus, the posterior expectations of N and σ_γ were obtained by taking the expectation of the suitably transformed values of each of the 100,000 generated points in the simulated chain.

In working with a log-linear model, the variance of exponentiated parameters operates multiplicatively on those exponentiated parameters. For example, the random effects variance σ_γ^2 represents the variability of the components of $\boldsymbol{\gamma}$, which themselves measure departures of the log-marginal means from independence. Thus, σ_γ operates multiplicatively via $\exp(\sigma_\gamma)$ on the corresponding marginal means. Therefore, a value of $\sigma_\gamma^2 = 1$ for the random effects corresponds to a multiplicative factor of $\exp(\sigma_\gamma) = 2.71$ on the marginal means.

Remark 5.5 In Remark 4.4, we stated that summary measures of the natural parameters in the hierarchical prior specification were preferable to summary measures of their transformation, since the implied priors on the latter may be sensitive to the hyper-parameters of the original parameters. For instance, the parameter $\beta_0 = \log N$ is used in the hierarchical prior structure rather than N. By the reasoning of Remark 4.4, we prefer to examine summary measures of β_0 rather than N. Nonetheless,

since the primary goal of capture-recapture methodology is to perform inference on N, we include the appropriate summaries for N in Table 5.7.

Further details of inference on N

The posterior mean and median of N coincide at a value of 2273 (see Table 5.5) because of the tight, symmetric nature of the posterior distribution. The 95% credible interval for N is given by $(2236, 2312)$. The point estimate and 95% asymptotic confidence interval of the best (in terms of AIC) frequentist marginal log-linear model (MLLM) of Section 3.5.2 are given by $\hat{N}_{MLE} = 2345$ and $(2295, 2396)$, respectively. Thus, we observe similarity in terms of location and width of confidence bounds between the results from the Bayesian analysis of the specific reduced model considered here in this section and the best-performing frequentist marginal log-linear model. We note however that Wald asymptotic confidence intervals were used in Chapter 3 (as per the methods of Appendix F.2), which are likely to be unreasonably tight.

Inference on $\boldsymbol{\beta}^{\star}$

Figure 5.5 shows the nearly symmetric posterior distributions of the components of $\boldsymbol{\beta}^{\star}$ which correspond to the logarithms of the single-source marginal means, m_A, m_B, m_C and m_D. In each case the solid vertical line corresponds to the log-observed marginal counts, given by $\log n_A = \log 1754 = 7.46$, $\log n_B = \log 452 = 6.11$, $\log n_C = \log 1135 = 7.03$ and $\log n_D = \log 173 = 5.15$. We observe that each of the four posterior distributions is centered on the value that we seek to fit. This is in line with results observed for frequentist joint log-linear models, as in Chapter 3 where we observed that the MLEs for these four quantities were exactly equal to the observed quantities (see discussion in Section 3.5.5). Note that the same noninformative prior used for each component of $\boldsymbol{\beta}^{\star}$ cannot be seen on the four plots of Figure 5.5, since the distribution is so vague that the prior density values are too small to be seen on the range of the posterior of each component of $\boldsymbol{\beta}^{\star}$. The patterns observed in each of the four panels of Figure 5.5 provide further evidence that the Bayesian model per-

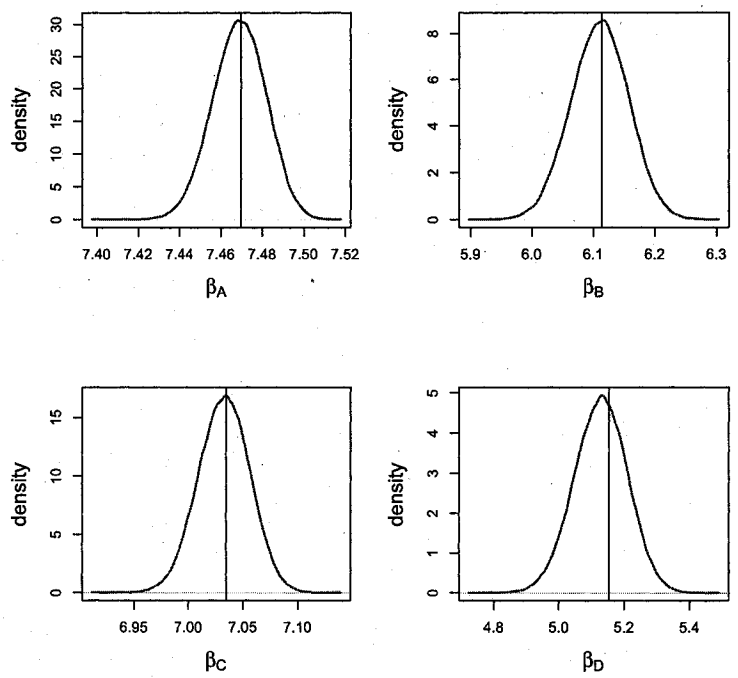


Figure 5.5: Posterior of $\boldsymbol{\beta}^{\star}$ for the reduced model with $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_{γ}. The vertical lines correspond to the log-observed marginal counts.

forms as desired since it is to be expected that the log-single source marginal means are estimated close to the corresponding log-observed counts. Heuristically, almost all the information gained concerning these parameters originates from the observed data.

Inference on the CID random effects $\boldsymbol{\gamma}$

Figure 5.6 shows the posterior distributions of the two non-zero CIDs of the reduced model, denoted by $\boldsymbol{\gamma}$. Recall that the postulated distribution of each of the two

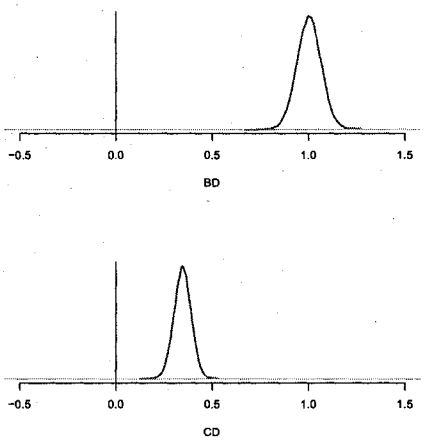


Figure 5.6: Posterior of non-zero CIDs for the reduced model with $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2.

components of $\boldsymbol{\gamma}$ is given by $\gamma_i|\sigma_\gamma^2 \sim_{iid} N(0, \sigma_\gamma^2)$ in the hierarchical prior structure. In both cases the distributions are close to symmetric, and are located above and away from zero thus demonstrating that the two marginal pairs given by $\{B, D\}$ and $\{C, D\}$ do explain some of the dependence present in the diabetes data set.

Model fit evaluated by fit of $\boldsymbol{\mu}$ to observed data $\mathbf{n_{incomp}}$

Figure 5.7 demonstrates the fit of the model. The 15 panels show the posterior distributions of the 15 cell means of $\boldsymbol{\mu}$. The vertical lines correspond to the observed cell count in each case. For clarity, we choose to omit the vertical axis corresponding to the density. The purpose of the plot is to assess the model fit. We see that in some cases the center of the posterior distribution of each cell mean is very close to the observed cell count (in the case of $\mu_{\bar{A}BCD}$ and $\mu_{AB\bar{C}D}$), whilst in others (e.g. $\mu_{ABC\bar{D}}$) the observed data are not fitted perfectly. Thus the model fit is not perfect but reasonable.

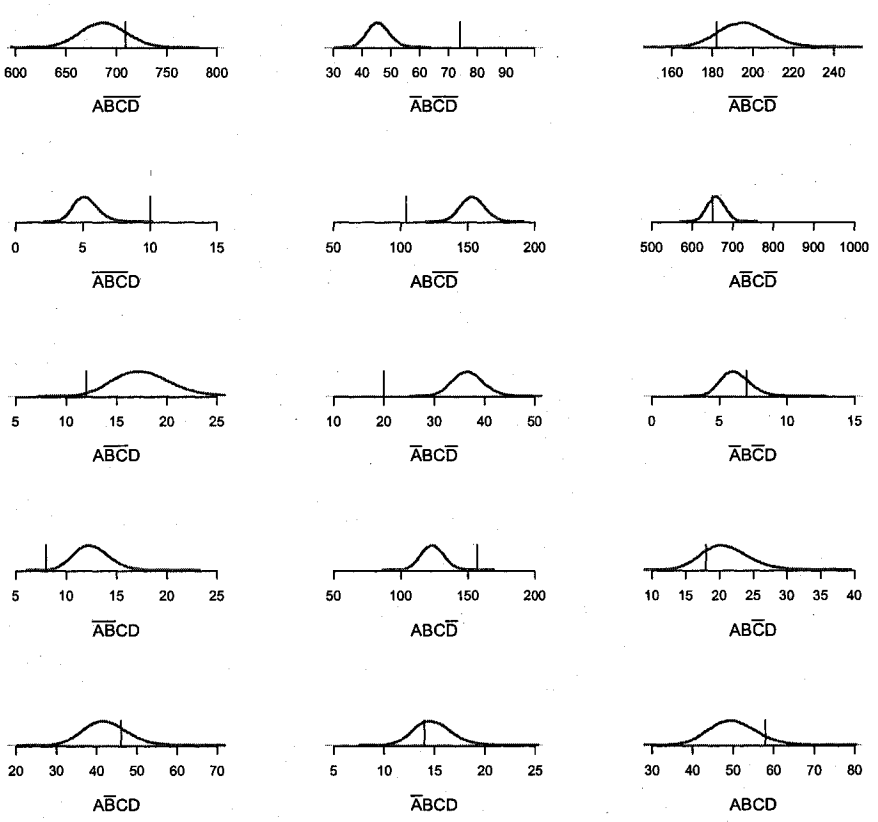


Figure 5.7: Posteriors of 15 cell means for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2.

Sensitivity of analysis to a range of prior specifications

We have seen that the reduced marginal log-linear model with the two CIDs γ_{BD} and γ_{CD} treated as random effects provides reasonable results with inference performed within the Bayesian framework. Thus, we wish to determine the sensitivity of posterior inference to the prior distributions placed on the main parameters of interest, namely on β_0 and σ^2_γ. Next we provide results from a series of analyses with different prior structures.

We will consider a range of different priors on each of β_0 and of σ^2_γ. More specifically, we will proceed according to the following two steps. First we will work with the same $N(8.06, 0.28)$ prior on β_0 of the previous section and examine the effect on the posterior of both β_0 and σ^2_γ for a range of priors on σ^2_γ. Second, we will do the reverse and work with the same scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_γ as in the previous section and examine the effect on the posterior of both β_0 and σ^2_γ for a range of priors on β_0.

Remark 5.6 Before proceeding with the first step of the sensitivity analysis of the current section, we will present the reasoning used to choose which priors we place on σ^2_γ. Figure 5.8 shows density plots of three different priors with degrees of freedom 2,5 and 10, each at eight different levels of the scale parameter (0.5,1,1.5,3,4,5,10,15). For a fixed level ν_0 of the prior degrees of freedom, it is clear that the distribution becomes increasingly skewed to the right as the scale parameter s^2_0 increases. This can be seen by looking at any one of the three panels in Figure 5.8. For fixed s^2_0, the distribution becomes more concentrated towards 0 as ν_0 increases, as can be seen by looking at one of the colored lines down across each of the three panels of Figure 5.8. Since a change in prior degrees of freedom ν_0 has only a small effect in terms of change in spread at fixed prior scale parameter s^2_0, we choose to work with a fixed level of ν_0 to investigate the effect of a change in prior on the full joint posterior distribution.

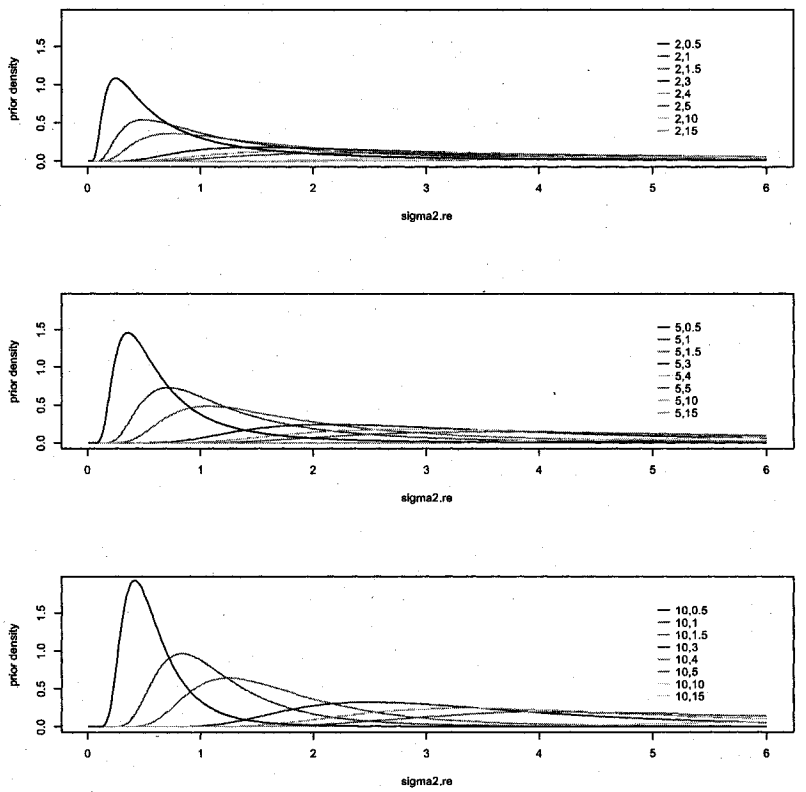


Figure 5.8: Prior distributions for σ^2_γ at three different prior degrees of freedom: 2,5 and 10, and for each at scale parameter: 0.5,1,1.5,3,4,5,10,15

We select $\nu_0 = 5$ since it is the minimum value of ν_0 for which the variance of the distribution is defined.

Since preliminary analyses at fixed levels of ν_0 demonstrated that there was minimal sensitivity to the prior distributions with changes in s_0^2, we also selected a single prior with $\nu_0 = 2$ and a single prior with $\nu_0 = 10$. The corresponding scale parameter s_0^2 was selected so that the mode of the distribution was equal to that for the inverse-$\chi^2(5, 0.5)$, i.e. with a mode of 0.357 (see page 151).

Our first step is to explore sensitivity of the posterior distributions to a range of priors on σ_γ^2 for the $N(8.06, 0.28)$ prior on β_0. As just described we fix the prior degrees of freedom ν_0 at 5 and let the scale parameter s_0^2 adopt the four different values of 0.1, 0.5, 1 and 1.5, since they place a reasonable amount of mass close to zero and we wish to be noninformative about departures from independence *a priori*. We also take a single case each with $\nu_0 = 2$ and $\nu_0 = 10$.

Sensitivity of β_0 to prior on σ_γ^2

There is no sensitivity in the posterior of β_0 to the prior on σ_γ^2 (plots not provided since in all cases the posterior of β_0 is that given in Figure 5.4).

Sensitivity of σ_γ^2 to prior on σ_γ^2

Figure 5.9 shows that there is some sensitivity of the posterior of σ_γ^2 to the prior on σ_γ^2 at a fixed prior on β_0. For $\nu_0 = 5$, we observe that as s_0^2 increases, the posterior of σ_γ^2 shifts to the right (see the first four panels of Figure 5.9). The final two panels of Figure 5.9 further demonstrate sensitivity to the prior value of ν_0 and s_0^2.

Now we move to the second step of the sensitivity analysis as described in the introduction to the current section. We wish to explore the sensitivity of the posteriors of both β_0 and σ_γ^2 to changes in the prior on β_0 for fixed prior on σ_γ^2. We choose three priors on β_0 given by $N(8.06, 0.28)$ (as used for the full analysis in the previous section), $N(8.06, 0.702)$ and $N(7.71, 0.49)$, which correspond to 90% prior ranges of 2000-5000, 1000-10,000 and 1000-5000, respectively.

Sensitivity of β_0 to prior on β_0

As with the case of sensitivity to the prior on σ_γ^2 above, there is no sensitivity in the posterior of β_0 to the prior on β_0. Again plots are not provided since in all cases the

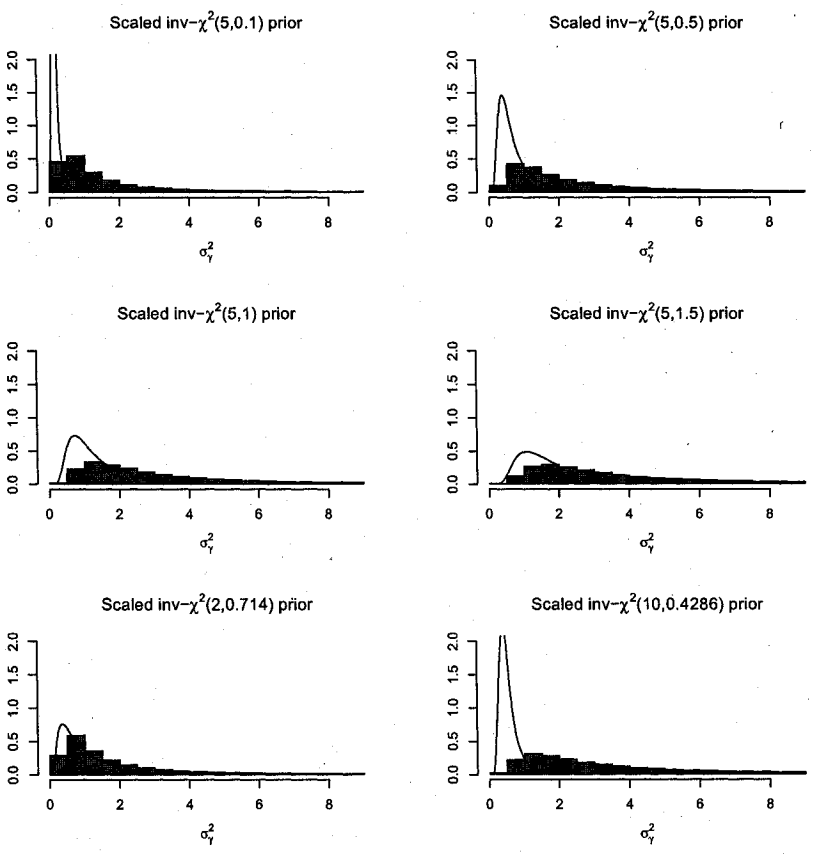


Figure 5.9: Posterior histograms of σ_γ^2 for various priors on σ_γ^2 and fixed prior on β_0.

posterior of β_0 is that given in Figure 5.4.

Sensitivity of σ_γ^2 to prior on β_0

There is no sensitivity in the posterior on σ_γ^2 to the prior on β_0 (again plots not included) unlike the observed sensitivity to the prior on σ_γ^2 as observed in Figure 5.9.

In summary, the discussion above and Figure 5.9 demonstrate that the greatest

posterior sensitivity is in σ_γ^2 with a change in its own prior. The posterior of β_0 is sensitive to neither changes in its own prior nor changes in the prior on σ_γ^2, just as there is no sensitivity in the posterior of σ_γ^2 to changes in the prior on β_0. Thus, using the range of prior distributions selected for this sensitivity analysis, changes in the prior on σ_γ^2 have a greater effect than changes in the prior on β_0.

5.4.4 Case of dependence: $\sigma_\gamma^2 > 0$; full model

In this section we consider the more flexible model of dependence, for which σ_γ^2 is not fixed at zero and none of the CIDs (treated as random effects) are fixed at zero. In such a case, the full analysis consists in mixing over the distribution of σ_γ^2 rather than working at a single slice of that distribution, namely at $\sigma_\gamma^2 = 0$, as was the case with the model of independence in Section 5.4.2. Consequently, the components of the random effects vector $\boldsymbol{\gamma}$ are to be estimated rather than being set to zero and more specifically, all elements are to be estimated as opposed to only γ_{BD} and γ_{CD} as was the case in the previous section.

The manner in which we proceed in this section parallels that for the reduced model in the previous section. We first present results for a single prior structure. We then examine the sensitivity of the joint posterior of all model parameters to the prior specification used by examining a range of prior specifications. More specifically, we adjust the prior parameters and hyperparameters relating to the main parameters of interest, namely β_0 and σ_γ^2, of the hierarchical prior structure given by (4.9).

Before presenting the results, we first address the important issue of nonidentifiability of model parameters for the full Bayesian model in order that all results will be evaluated in this light. The high posterior correlations observed in Table 5.3, despite thinning by a factor of 200, suggest that there is a problem with mixing and that it is likely that the posterior parameter space has not been explored completely. Such problems may arise from nonidentifiability of model parameters. Before presenting

the analysis of the full model we will describe the nature of such nonidentifiability by presenting the case of the full three-source model given first by Example 2.11, with the system of equations repeated here for completeness.

$$\begin{aligned}
\log m_A &= \beta_A \\
\log m_B &= \beta_B \\
\log m_C &= \beta_C \\
\log m_{AB} &= -\beta_0 + \beta_A + \beta_B + \gamma_{AB} \\
\log m_{AC} &= -\beta_0 + \beta_A + \beta_C + \gamma_{AC} \\
\log m_{BC} &= -\beta_0 + \beta_B + \beta_C + \gamma_{BC} \\
\log m_{ABC} &= -2\beta_0 + \beta_A + \beta_B + \beta_C + \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \gamma_{ABC}. \qquad (5.2)
\end{aligned}$$

Suppose we fix all 7 marginal means (and by design the corresponding β_A, β_B and β_C terms) at known values. We will demonstrate that there are a countable number of solutions to the system of equations when β_0 is fixed at a known value.

First, the pairwise CIDs are obtained from (5.2) and are given by

$$\begin{aligned}
\gamma_{AB} &= \log m_{AB} + \beta_0 - (\beta_A + \beta_B) \\
\gamma_{AC} &= \log m_{AC} + \beta_0 - (\beta_A + \beta_C) \\
\gamma_{BC} &= \log m_{BC} + \beta_0 - (\beta_B + \beta_C), \qquad (5.3)
\end{aligned}$$

which, in turn, yields the following expression for γ_{ABC}

$$\begin{aligned}
\gamma_{ABC} &= \log m_{ABC} + 2\beta_0 - (\beta_A + \beta_B + \beta_C) - (\gamma_{AB} + \gamma_{AC} + \gamma_{BC}) \\
&= \log m_{ABC} - (\log m_{AB} + \log m_{AC} + \log m_{BC}) - \beta_0 + (\beta_A + \beta_B + \beta_C). \qquad (5.4)
\end{aligned}$$

Now suppose that, for the same 7 fixed marginal means, β_0 is shifted by a constant term c. Then we have an alternative system of solutions for $\gamma_{AB}, \gamma_{AC}, \gamma_{BC}$ and γ_{ABC}

given by

$$\begin{aligned}
\gamma_{AB} &= \log m_{AB} + (\beta_0 + c) - (\beta_A + \beta_B) \\
\gamma_{AC} &= \log m_{AC} + (\beta_0 + c) - (\beta_A + \beta_C) \\
\gamma_{BC} &= \log m_{BC} + (\beta_0 + c) - (\beta_B + \beta_C),
\end{aligned} \tag{5.5}$$

which, in turn, yields the following expression for γ_{ABC}

$$\begin{aligned}
\gamma_{ABC} &= \log m_{ABC} - (\log m_{AB} + \log m_{AC} + \log m_{BC}) + \\
&\quad 2(\beta_0 + c) - (\beta_A + \beta_B + \beta_C) - (\gamma_{AB} + \gamma_{AC} + \gamma_{BC}) \\
&= \log m_{ABC} - (\log m_{AB} + \log m_{AC} + \log m_{BC}) - (\beta_0 + c) + (\beta_A + \beta_B + \beta_C).
\end{aligned} \tag{5.6}$$

Thus, there are a countable number of solutions for β_0, $\gamma_{AB}, \gamma_{AC}, \gamma_{BC}$ and γ_{ABC} for fixed marginal means since we can add a constant to β_0 and obtain the same solution to the system of equations given by (5.2) by adding the same constant to each of the three CID terms given by γ_{AB}, γ_{AC} and γ_{BC} and subtracting that term from γ_{ABC}. (We note that there are a countable rather than uncountable number since N is an integer.) Therefore, nonidentifiability arises (as it does for the full model for four sources but not stated explicitly here for brevity's sake).

There are problems as evinced by the high posterior correlations of Table 5.3 and Figure 5.2. Despite such problems we include the analysis via the full model to demonstrate its properties and highlight the challenges of working with such a model. It should be noted that the interpretation of the results must be considered dubious in the light of the issue of nonidentifiability. Better model fit is achieved than for the reduced model precisely because there are so many model parameters which are able to provide a close-to-perfect fit. We further hypothesize that reparameterization of the model could alleviate some of the problems. Such work will be undertaken in the future.

Features of analysis for a single prior specification

As with the reduced model in the previous section, we first undertake the full Bayesian analysis for a single hierarchical prior structure. The specific prior structure on the parameters of interest, β_0 and σ_γ^2, is the same as for the reduced model and is given here again by Table 5.6.

Parameter	β_0		σ_γ^2	
Hyperprior parameter	μ_{β_0}	$\sigma_{\beta_0}^2$	ν_0	s_0^2
Value	8.06	0.28	5	0.5

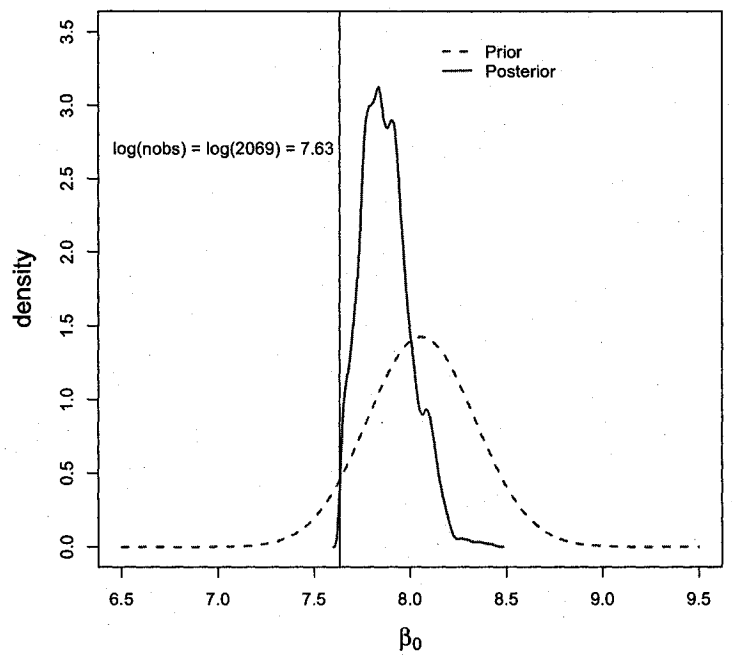
Table 5.6: Hyperprior parameters of β_0 and σ_γ^2 for full Bayesian analysis

Inference on N, β_0 and σ_γ^2

Figure 5.10 shows the marginal posterior distributions of each of β_0 and σ_γ^2. The posteriors represent the effect of updating (via the likelihood, given by (4.4)) the prior distributions of each of β_0 and σ_γ^2 using the observed capture-recapture data of the incomplete contingency table given by Table 5.1. The upper panel of Figure 5.10 shows that the posterior of β_0 is bounded below by $\log n_{obs} = \log 2069 = 7.63$ as per (4.8).[2] The posterior distributions of β_0 and σ_γ^2 are shifted slightly to the left in both cases with a reduction in variability. As expected by conjugacy, the posterior of σ_γ^2 is a scaled inverse-χ^2 distribution.

Table 5.7 provides posterior summaries of β_0 and σ_γ^2, as well as the corresponding posterior summaries of N as obtained from $N = \exp(\beta_0)$ and of σ_γ. The upper and lower bounds of the symmetric 90% and 95% credible intervals are provided, together

[2]The plot was produced using the inbuilt `density` function of The *R Programming Language* (2004). Thus, the appearance of the posterior of β_0 below its lower bound of $\log n_{obs}$ is an artefact of the method used.



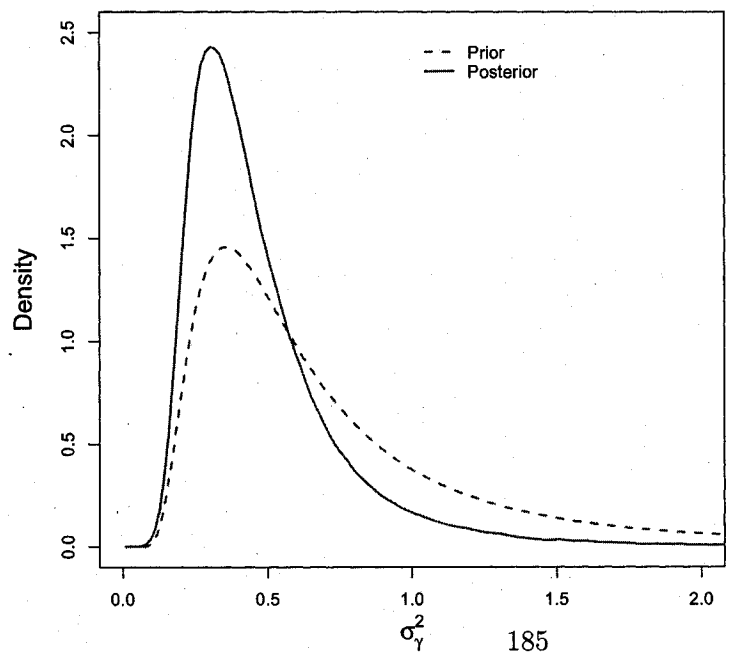


Figure 5.10: Posteriors of β_0 and σ_γ^2 for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2.

	Quantile							
Parameter	0 %	2.5%	5%	50%	95%	97.5%	100%	Mean
β_0	7.635	7.663	7.685	7.864	8.114	8.153	8.448	7.877
N	2069	2128	2175	2602	3341	3475	4668	2658
σ_γ	0.302	0.427	0.452	0.636	0.971	1.067	1.998	0.664
σ_γ^2	0.091	0.183	0.205	0.404	0.942	1.138	3.994	0.468

Table 5.7: Posterior summaries for β_0, N and σ_γ for Bayesian analysis

with the minimum and maximum values in order to examine the range of posterior values of each parameter.

Further details of inference on N

The posterior mean and median of N are given by 2658 and 2602, respectively. With a posterior mean larger than the posterior median, the posterior distribution is right-skewed as confirmed by the left panel of Figure 5.10. Such skewness is in part due to the normal prior on β_0, which in turn implies a right-skewed log-normal prior on N. The 95% credible interval for N (see Table 5.7) is given by $(2128, 3475)$ for a posterior mean and median of 2658 and 2602, respectively. The point estimate and 95% asymptotic confidence interval of the best (in terms of AIC) frequentist marginal log-linear model (MLLM) of Section 3.5.2 are given by $\hat{N}_{MLE} = 2345$ and $(2295, 2396)$, respectively. Thus, we observe a nontrivial difference, in terms of both location and the width of confidence bounds, between $\hat{N}_{MLE}$ from the best frequentist MLLM and the posterior mean and median of the full Bayesian analysis, as we observe such a difference between the results of the present full Bayesian and that from the reduced model in the previous section. We note however that Wald asymptotic confidence intervals were used in Chapter 3 (as per the methods of Appendix F.2), which are likely to be unreasonably tight.

It should be noted that the corresponding point estimates and confidence intervals from the best joint log-linear model (JLLM) and hierarchical JLLM are closer in location to the center of the posterior distribution of N obtained from the present Bayesian analysis. The point estimates and 95% confidence intervals are given by $\hat{N}_{MLE,JLLM} = 3092$ and (2573,3611) for the best-performing JLLM, which outperforms the best HJLLM with $\hat{N}_{MLE,HJLLM} = 2771$ and 95% confidence interval given by (2491,3050). The greater width of the 95% credible interval over the 95% confidence intervals of the best frequentist models is in part due to the fact that model selection has not been performed in the Bayesian analysis of the current chapter. Rather we have chosen to work with the fully specified model with none of the CIDs fixed at zero, unlike the frequentists models of Chapter 3, in which model selection was performed amongst all possible 2047 MLLMs.

Inference on $\boldsymbol{\beta}^{\star}$

Figure 5.11 shows the nearly symmetric posterior distributions of the components of $\boldsymbol{\beta}^{\star}$ which correspond to the logarithms of the single-source marginal means, m_A, m_B, m_C and m_D (as did Figure 5.5 for the reduced model). In each case the solid vertical line corresponds to the log-observed marginal counts, given by $\log n_A = \log 1754 = 7.46$, $\log n_B = \log 452 = 6.11$, $\log n_C = \log 1135 = 7.03$ and $\log n_D = \log 173 = 5.15$. We observe that each of the four posterior distributions is centered on the value that we seek to fit. This is in line with the results for the reduced model (see Figure 5.5) and with the results observed for frequentist joint log-linear models, as in Chapter 3 where we observed that the MLEs for these four quantities were exactly equal to the observed quantities (see discussion in Section 3.5.5). As with Figure 5.5 we note that the same noninformative prior used for each component of $\boldsymbol{\beta}^{\star}$ cannot be seen on the four plots of Figure 5.11, since the distribution is so vague that the prior density values are too small to be seen on the range of the posterior of each component of $\boldsymbol{\beta}^{\star}$. The patterns observed in each of the four panels of Figure 5.11 provide further

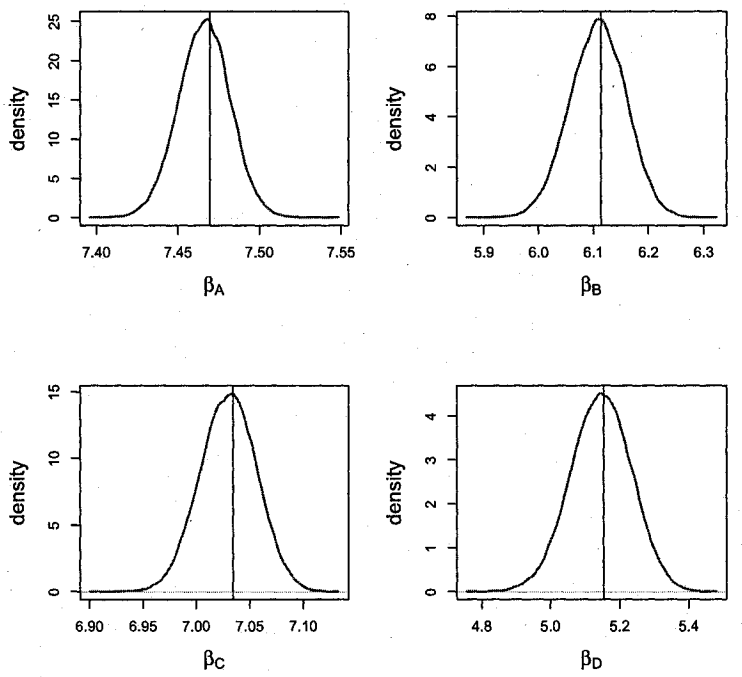


Figure 5.11: Posterior of $\boldsymbol{\beta}^{\star}$ for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_{γ}. The vertical lines correspond to the log-observed marginal counts.

evidence that the Bayesian model performs as desired since it is to be expected that the log-single source marginal means are estimated close to the corresponding log-observed counts. Heuristically, almost all the information gained concerning these parameters originates from the observed data.

Inference on the CID random effects $\boldsymbol{\gamma}$

Figure 5.12 shows the posterior distributions of all 11 CIDs, denoted by $\boldsymbol{\gamma}$. Recall that the postulated distribution of each of the components of $\boldsymbol{\gamma}$ is given by $\gamma_i|\sigma^2_{\gamma} \sim_{iid} N(0, \sigma^2_{\gamma})$ in the hierarchical prior structure. The six panels of the left

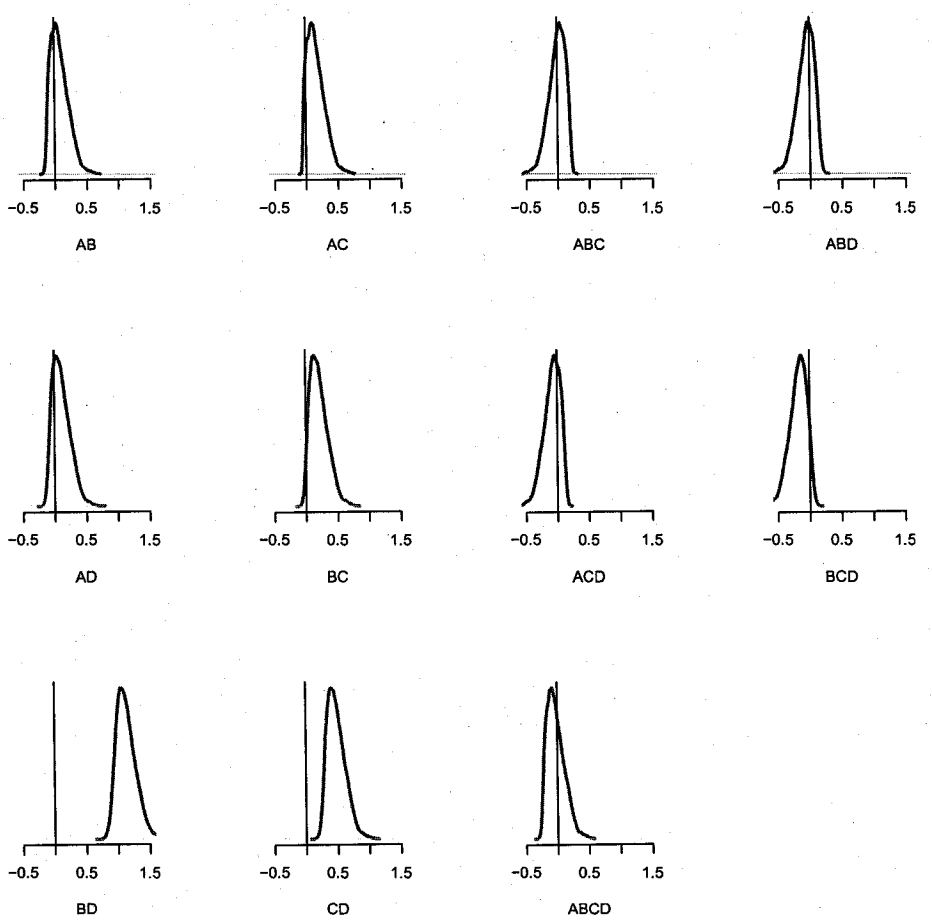


Figure 5.12: Posterior of CIDs for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2.

side of Figure 5.12 show the posteriors for each of the six pairwise CIDs, denoted by $\gamma_{AB}, \gamma_{AC}, \gamma_{AD}, \gamma_{BC}, \gamma_{BD}, \gamma_{CD}$. In all cases, the distributions are close to symmetric, with some slight right-skewness observed in the slightly longer right tails of those distributions. We observe that the posterior distributions for γ_{BD} and γ_{CD} lie the

furthest from zero. The four upper panels on the right side of Figure 5.12 show the posteriors for each of the 4 three-way CIDs denoted by $\gamma_{ABC}, \gamma_{ABD}, \gamma_{ACD}, \gamma_{BCD}$, with the lowest panel showing that for the single four-way CID γ_{ABCD}. In this case, we observe that the posteriors of the three-way CIDs are slightly left-skewed and quite close to 0 in all cases, with the posterior of the sole four-way CID approximately symmetric. Thus, there tends to be slight right-skewness in the pairwise CIDs, with slight left-skewness in the three-way CIDs.

The location of the posteriors of the CIDs are loosely in line with the results of Chapter 3. The frequentist marginal log-linear model results of Section 3.5.2 show that the model selected by the AIC contains each of the 7 CID terms given by $\gamma_{AC}, \gamma_{BC}, \gamma_{BD}, \gamma_{CD}, \gamma_{ABC}, \gamma_{BCD}, \gamma_{ABCD}$ suggesting that there is a range of dependence present in the diabetes data of Table 5.1. Such a feature is confirmed by the posteriors of the components of $\boldsymbol{\gamma}$ being centered around values away from zero and in some cases (i.e. γ_{BD} and γ_{CD}) quite considerably removed from zero. The Bayesian analysis of this section serves to provide an indication as to which margins explain most of the dependence. Figure 5.12 shows that the $\{B, D\}$ and $\{C, D\}$ margins are believed to explain much of the dependence thus providing support to the choice of the reduced model used in the previous section.

Questioning which of the CIDs explain most dependence is equivalent to questioning which random effects are most different from zero. Such questioning relates to the fundamental issue of random effects modelling: is the primary interest in the variance parameter controlling the distribution of the random effects or in the random effects themselves? In the capture-recapture setting it can be useful to understand the dependence structure itself, as summarized by the CIDs (i.e. the components of the random effects vector $\boldsymbol{\gamma}$), but the primary goal is to obtain reliable estimates of N. Modelling the dependence structure can considered as a means to the end of estimation of N. As a consequence, σ^2_γ arguably holds greater interest than the

CIDs themselves when these are treated as random effects. Nonetheless, we reiterate that there are problems of nonidentifiability with the full Bayesian model which must addressed in the future.

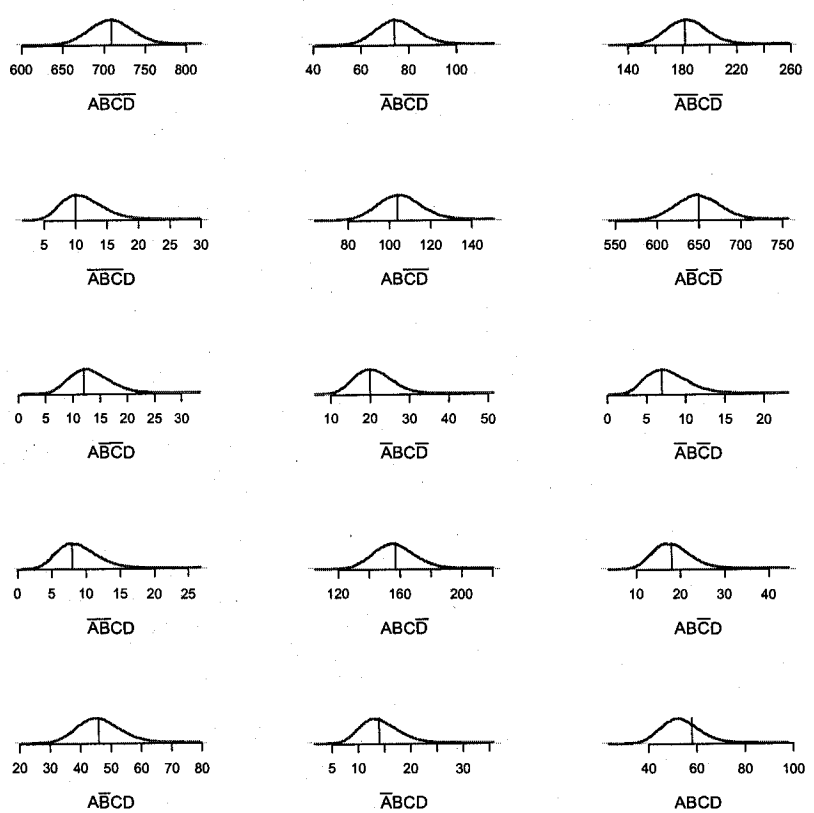


Figure 5.13: Posteriors of 15 cell means for $N(8.06, 0.28)$ prior on β_0 and scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_γ.

Model fit evaluated by fit of $\boldsymbol{\mu}$ to observed data $\mathbf{n_{incomp}}$

Figure 5.13 demonstrates the fit of the model. The 15 panels show the posterior distributions of the 15 cell means of $\boldsymbol{\mu}$. The vertical lines correspond to the observed cell count in each case. For clarity, we choose to omit the vertical axis corresponding to the density. The purpose of the plot is to demonstrate that the center of the posterior distribution of each cell mean is very close to the observed cell count thus demonstrating that the model fits well. Such a result is to be expected given the nature of the nonidentifiability of model parameters. We so many model parameters it is possible for the model to fit almost perfectly to the observed data unlike the fit of the reduced model observed in Figure 5.7.

Sensitivity of analysis to a range of prior specifications

We have seen that the marginal log-linear model with all CIDs treated as random effects and inference performed within the Bayesian framework is able to fit the data close to perfectly (see Figure 5.13). In light of the challenges of model nonidentifiability, we will present some results to highlight the sensitivity of posterior inference to the prior distributions placed on the main parameters of interest, namely on β_0 and σ^2_γ, as was done for the reduced model in the previous section.

We will consider a range of different priors on each of β_0 and of σ^2_γ. More specifically, as for the reduced model, we will proceed according to the following two steps. First we will work with the same $N(8.06, 0.28)$ prior on β_0 of the previous section and examine the effect on the posterior of both β_0 and σ^2_γ for a range of priors on σ^2_γ. Second, we will do the reverse and work with the same scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_γ as in the previous section and examine the effect on the posterior of both β_0 and σ^2_γ for a range of priors on β_0.

Our first step is to explore sensitivity of the posterior distributions to a range of priors on σ^2_γ for the $N(8.06, 0.28)$ prior on β_0. As just described we fix the prior

degrees of freedom ν_0 at 5 and let the scale parameter s_0^2 adopt the four different values of 0.1, 0.5, 1 and 1.5, since they place a reasonable amount of mass close to zero and we wish to be noninformative about departures from independence *a priori*.

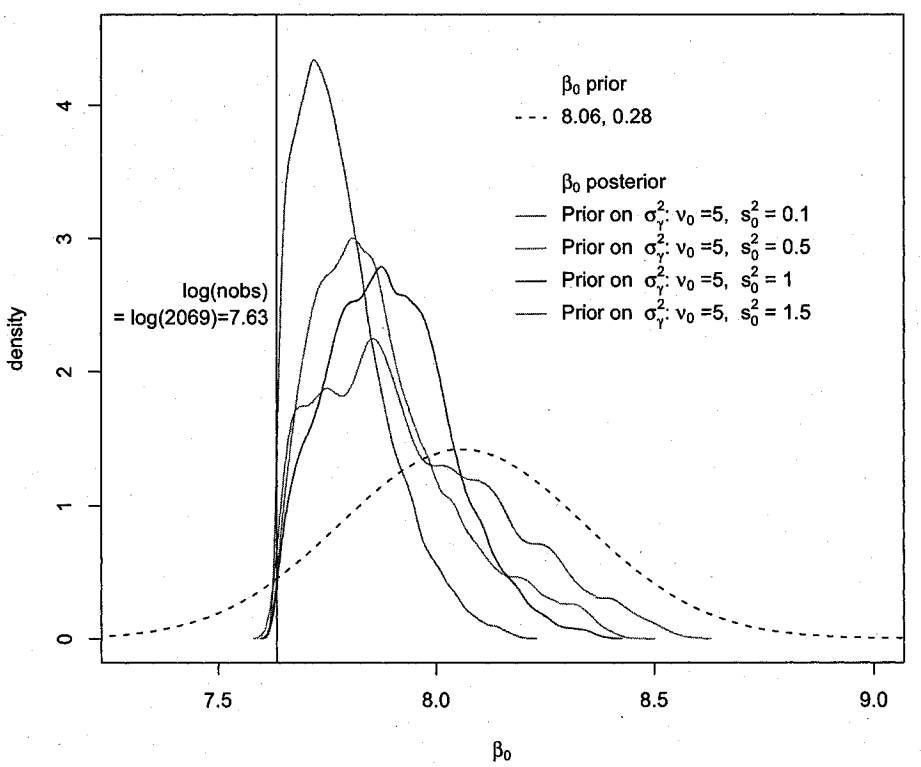


Figure 5.14: Posterior of β_0 for various priors on σ_γ^2 and fixed prior on β_0.

Sensitivity of β_0 to prior on σ_γ^2

Figure 5.14 shows the four posterior distributions on β_0 corresponding to the four different priors on σ_γ^2 for the same $N(8.06, 0.28)$ prior on β_0. In all four cases the pos-

terior distributions are less variable than the prior and are shifted to the left. All four distributions are bounded below by $\log n_{obs} = 7.63$. [3] In all four cases the posterior distributions are close in terms of location. The observed suggestions of bimodality of the posterior of β_0, in particular for a prior on σ_γ^2 with $\nu_0 = 1$ and $\nu_0 = 1.5$, is a result of running two chains and the mixing of the chains. Nonetheless, the results of Figure 5.14 serve to demonstrate that there is some sensitivity in the posterior of β_0 to the prior on σ_γ^2.

Sensitivity of σ_γ^2 to prior on σ_γ^2

Figure 5.15 shows three of the posterior distributions on σ_γ^2 corresponding to the four different priors on σ_γ^2 for the same $N(8.06, 0.28)$ prior on β_0. Only three of the four densities are shown for clarity. The pattern for the fourth (for the case with $\nu_0 = 1.5$) was consistent with the other three cases. In each case we observe that the posterior distribution is shifted from the prior. We notice that the direction of this shift differs: in the case of the prior with $\nu_0 = 0.1$, the posterior is shifted to the right relative to the prior, whereas for $\nu_0 = 0.5$ and $\nu_0 = 1$, it is shifted in the opposite direction. The plots of Figure 5.14 demonstrate that there is sensitivity of the posterior of σ_γ^2 to changes in the prior on σ_γ^2.

Now we move to the second step of the sensitivity analysis as described in the introduction to the current section. We wish to explore the sensitivity of the posteriors of both β_0 and σ_γ^2 to changes in the prior on β_0 for fixed prior on σ_γ^2. We choose three priors on β_0 given by $N(8.06, 0.28)$ (as used for the full analysis in the previous section), $N(8.06, 0.702)$ and $N(7.71, 0.49)$, which correspond to 90% prior ranges of

[3] Note that, as with the upper panel of Figure 5.10, the inbuilt `density` function of the *R Programming Language* is used to obtain the plot. Thus the observed part of the posterior densities of β_0 below 7.63 is an artefact of the inbuilt function.

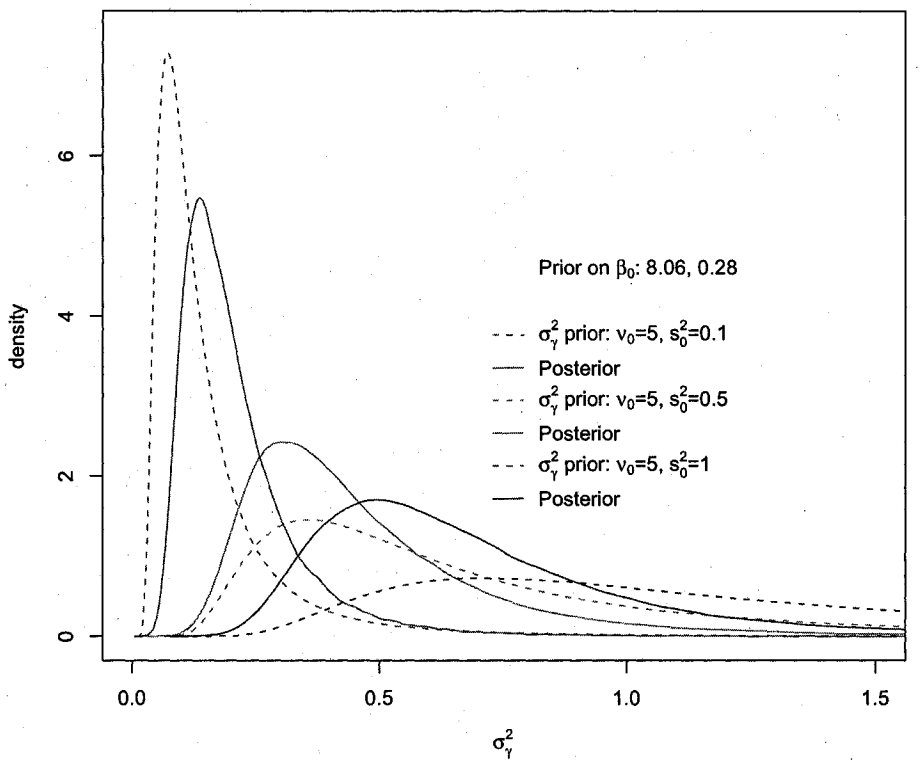


Figure 5.15: Posterior of σ^2_γ for various priors on σ^2_γ and fixed prior on β_0.

2000-5000, 1000-10,000 and 1000-5000, respectively.

Sensitivity of β_0 to prior on β_0

Figure 5.16 shows the three posterior distributions of β_0 corresponding to the three different priors on β_0 for the same scaled-inverse $\chi^2(5, 0.5)$ prior on σ^2_γ. In all three cases the posterior distributions are less variable than the three priors and tend to be shifted to the left of their corresponding prior. These patterns are similar to those

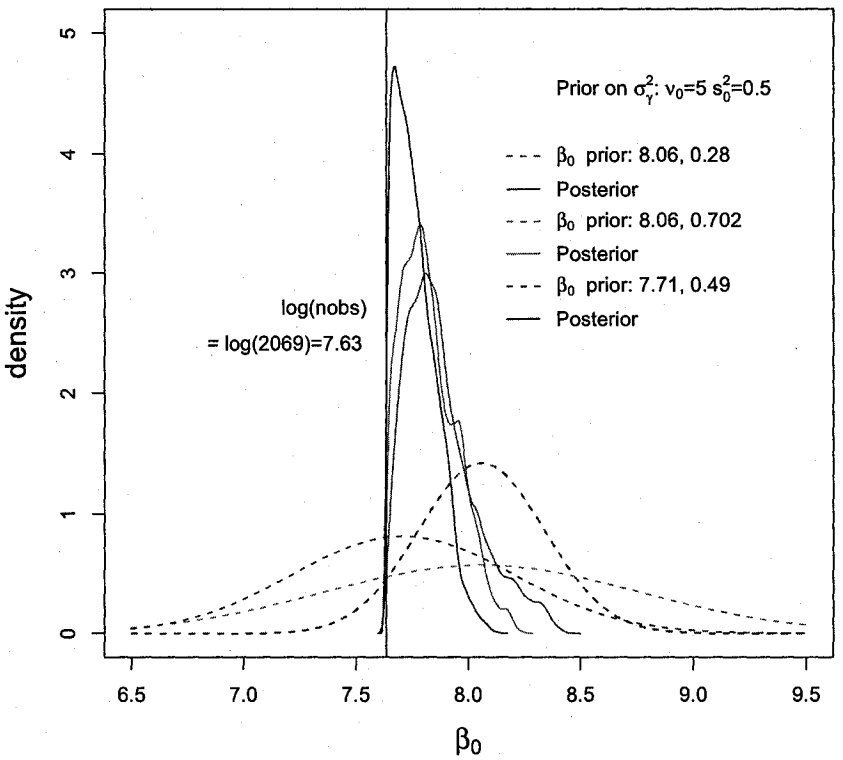


Figure 5.16: Posterior of β_0 for various priors on β_0 and fixed prior on σ_γ^2.

observed in Figure 5.14 above. As in that figure, all three distributions are bounded below by $\log n_{obs} = 7.63$. [4] The four posterior distributions are close to each other in terms of location. There is some observed bimodality of the posterior of β_0. Despite the imperfect mixing, the results of Figure 5.16 serve to demonstrate that there is

[4]Note that again, as with the upper panel of Figures 5.10 and 5.14, the inbuilt 'density' function of the *R Programming Language* is used to obtain the plot. Thus the observed part of the posterior densities of β_0 below 7.63 is an artefact of the inbuilt function.

some sensitivity in the posterior of β_0 to its own prior distribution.

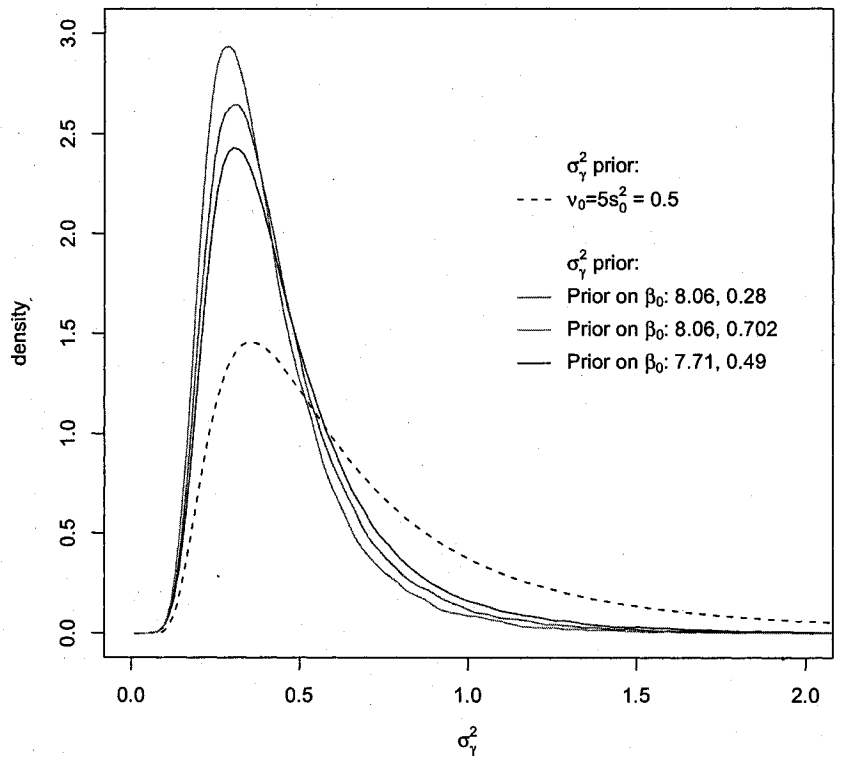


Figure 5.17: Posterior of σ_γ^2 for various priors on β_0 and fixed prior on σ_γ^2.

Sensitivity of σ_γ to prior on β_0

Figure 5.17 shows the posterior distributions of σ_γ^2 corresponding to the three different priors on β_0 for the same scaled-inverse $\chi^2(5, 0.5)$ prior on σ_γ^2. In each case we observe that the posterior distribution is barely shifted from the prior. Therefore, we observe that the posterior of σ_γ^2 is much less sensitive to a change in prior on β_0 than

it is to a change in its own prior.

In summary, Figures 5.14-5.17 demonstrate that the greatest posterior sensitivity is in σ_γ^2 with a change in its own prior. The posterior of β_0 is somewhat sensitive to changes in both its own prior and the prior on σ_γ^2, whereas there is minimal sensitivity of the posterior of σ_γ^2 to changes in the prior on β_0. Thus, using the range of prior distributions selected for this sensitivity analysis, changes in the prior on σ_γ^2 have a greater effect than changes in the prior on β_0. Again we evaluate such results in light of the challenges of parameter nonidentifiability. Unlike the somewhat restrictive reduced model for which there was minimal sensitivity to the prior specification, we see greater sensitivity here for the full model. We believe that such a highly determined model is able to fit the data close to perfectly, thus there is sensitivity to the prior specification since the parameters are able to be adjusted to obtain the close-to-perfect fit.

5.5 Summary and discussion

In this chapter we have performed a Bayesian analysis of the complete marginal log-linear model using two different models: the first with two CIDs in the model and the second with all CIDs included in the model. In both cases the CIDS were treated as random effects. Unlike the frequentist approach of Chapter 3, in which the CIDs were treated as fixed effects, we have not performed model selection. Moreover, in treating the CIDs as random rather than fixed effects we achieved parameter reduction although there are issues related to the nonidentifiability of model parameters of the full model which must be addressed.

By performing inference via the Bayesian paradigm we were able to incorporate prior information to neatly accommodate the random effects. Furthermore, we have

been able to explore more explicitly the relationship between N and the dependence structure measured by the CIDs γ, whose distributions are controlled by σ_γ^2. Overall, our approach is a seemingly promising methodology, but further thought must be given to the nonidentifiability.

Chapter 6

Conclusion

6.1 Overview

In this dissertation we have presented a new class of marginal log-linear models (MLLM) for population size estimation using capture-recapture data, with an emphasis on epidemiological applications for closed populations. We presented a new framework in which to quantify source dependence with a new measure of dependence, the Coefficient of Incremental Dependence (CID), which in turn led to new modelling approaches in which dependence is modelled via the inclusion of source dependence measure terms. Two alternative approaches to inference were presented: the likelihood approach in which the measures were treated as fixed effects and a Bayesian approach in which they were treated as random. Real and simulated data analyses were performed in both cases. Through these analyses we demonstrated that we obtain plausible results using our class of marginal models, which were compared and contrasted to those obtained using the standard modelling approach of hierarchical joint log-linear modelling. Further, we demonstrated the manner in which MLLMs extend the universe of models for capture-recapture, thus enabling us to obtain reasonable population size estimates for dependence structures not necessarily

well-modelled by existing methods. In short, the proposed marginal modelling approach performs well and provides new insight into the fundamental nature of the use of epidemiological capture-recapture data.

6.2 Contributions

In this section we will guide the reader chapter by chapter through the original material introduced in this dissertation in order to highlight the specific contributions of the work to the literature.

Chapter 2

The cornerstone of the work presented in this dissertation has been the new manner in which we quantify and, indeed, conceptualize source dependence. To this end, we introduced a new measure of dependence: the Coefficient of Incremental Dependence (CID)(see Definition 2.4), that we relate to the Coefficient of Source Dependence (CSD) which was first proposed by Vandal et al. (2005), and subsequently developed by Melocco (2002). In both instances, these measures exist for all possible combinations of the K available sources, with the CIDs defined in terms of the CSDs. We characterized the (non-trivial) inverse linear relationship between the CSDs and CIDs, that is, with the CSDs expressed as a linear combination of all CIDs of equal and lower order (Theorem 2.5). In such a way, we demonstrated the manner in which the CIDs decompose source dependence. Further, we derived the properties relating these measures to (1) the conditional independence structures modelled by hierarchical joint log-linear models (HJLLM) (Theorems 2.3 and 2.9), (2) simple dependence structures modelled by HJLLMs (Corollary 2.2) (3) the bounds resulting from the nested form of the measures (of higher order marginal means nested within lower order marginal means, see Proposition 2.1), and, (4) the Coefficient of Covariation of

Chao and Tsay (1998) (Propositions 2.2 and 2.3 and Remark 2.10).

The form of the measures motivated the development of a new class of marginal models, named the marginal log-linear models (MLLM), in which marginal means are modelled rather than joint (or cell) means, as is the case with joint log-linear models. Two parameterizations were presented, the first in terms of the CIDs (Definition 2.7), the second in terms of the CSDs (Proposition 2.8). Both provide their own useful interpretations, with the CID parameterization preferred for reasons of modelling flexibility and hence used for the data analyses of Chapters 3 and 5.

The universe of dependence structures that can be modelled is extended by the use of our new marginal modelling approach. First it is shown that, for the simple dependence structures of complete independence and mutual dependence, the marginal modelling approach is equivalent to the joint log-linear modelling approach. However, even for the three-source case, there is no unconstrained marginal model equivalent to the HJLLM for conditional independence (see Example 2.16). Consequently, our new approach is more than a mere reparameterization of the standard hierarchial joint log-linear modelling approach.

Chapter 3

We compared and contrasted the marginal log-linear modelling approach to the standard modelling approach of HJLLMs. We proposed the family of MLLMs formed by setting all possible combinations of CIDs (CSDs) equal to zero. The likelihood approach was used as the inferential framework to fit the CID parameterization of the MLLM (with the CIDs treated as fixed effects).

A known equivalence between the multinomial and Poisson likelihoods enabled us to exploit the computational simplicity of the latter for the purposes of model fitting. We developed (and coded in the *R Programming Language*, 2004) a Fisher Scoring algorithm to obtain maximum likelihood estimates for MLLMs (see code in

Appendix G). The procedure was shown to work satisfactorily on real and simulated data sets.

As per the stated objectives of Chapter 3, we showed, via the analyses of these data sets, that our new MLLM (1) gives plausible results for real data and known dependence structures of simulated data, and (2) can better accommodate nonhierarchical dependence than the standard modelling approach of HJLLMs. Additionally, we demonstrated that the CIDs (CSDs) enable us to interpret the model parameters of nonhierarchical joint log-linear models. As a consequence, such a feature enables us to sensibly extend the universe of dependence structures able to be modelled using joint log-linear models which were previously excluded for reasons of non-interpretability of model parameters.

Chapter 4

We presented an alternative manner in which to parameterize the MLLM by working with a mixed effects model formulation, in which the random effects are used to model dependence (see Definition 4.1). Such a model formulation, which to our knowledge is new in the field of capture-recapture methodology, is related to the work on the generalized log-linear model with random effects of Coull and Agresti (2003), the larger class of generalized linear mixed models and the limited literature on general marginal modelling techniques.

We presented motivation for the adoption of the Bayesian inferential framework to accommodate the random effects and the model constraints, like those enforced in Chapter 3. A full Bayesian specification was presented for the general form of the model, together with a description of the MCMC scheme to be adopted for parameter estimation.

Chapter 5

Adopting the general Bayesian mixed effects model form of Chapter 4, we proposed one specific model formulation in which the random effects were set equal to the CIDs. Details of the MCMC simulation scheme for the specific form of the model were described (and coded in the C++ language, see Appendix H) and a full Bayesian data analysis of the real data set analyzed in Chapter 3 was presented. Such an analysis further demonstrated features of the MLLM and, indeed of the fundamental nature of capture-recapture data. In the first respect, we saw that the center of the posterior distributions of the fixed effect terms were close to those of the frequentist formulation of Chapter 3.

Minimal sensitivity of the posterior distributions of both N (via the β_0 parameter) and the random effects variance to the priors on each of these parameters was observed for the reduced model, with more sensitivity for the full model, likely as a result of the nonidentifiability of model parameters. In particular, we observed greater sensitivity of each posterior to the prior on the random effects variance for the full model. The methodology presented in this chapter is promising. We anticipate that future work related to the challenges of nonidentifiability will present interesting results and will further illuminate the nature of the marginal model, as well as clarifying the appropriateness of the model parameterization for inference performed via the Bayesian paradigm.

Summary

Estimating N when incomplete but overlapping data sources are available, is of relevance to many areas of public health. However, there are a range of challenges faced when analyzing capture-recapture data. Through the framework presented in this dissertation, we have sought to advance the field by developing a framework in which to conceptualize and quantify source dependence, whilst presenting a new

marginal modelling approach to fully exploit what can be learnt from the incomplete capture-recapture data.

6.3 Future work

We view the work presented in this dissertation as positioned in both the field of capture-recapture methodology and the broader field of general categorical data analysis. In both cases, the primary extensions of our work relate to the use of our measures of source dependence and the marginal modelling approach itself. For the latter, one specific goal is to determine ways in which to incorporate the covariate data often available for epidemiological capture-recapture studies into the modelling approach. Such a extension offers the possibility of accounting for heterogeneity-induced source dependence.

Future work we propose, in relation to the CSDs and CIDs, centers around their relationships to other known measures of association for contingency tables, including those for complete contingency tables, such as the odds ratio. We would seek to formalize the general K-source relationship between the CCV of Chao and Tsay (1998) and the CIDs (CSDs) as demonstrated for the 2, 3 and 4-source case in Chapter 2. We propose to further develop the expression of the CSD (and thus, the CID) in terms of available covariate information aggregated over each margin, as explored in Melocco (2002). Such a measure could be incorporated into an alternative marginal model, with covariate information, which should reduce residual variability in the marginal modelling approach. Moreover, the measures could be extended to continuous variables in addition to the bivariate ones measured here in the capture-recapture setting.

Noting the demonstrated relationship between our CIDs (and CSDs) and the CCV, we propose to explore how the CCV might form the basis of a marginal modelling

approach at the level of the individual, rather than at the level of the source as with our approach. We note further that Chang, LaPorte, Aaron, and Songer (1999) use a nonparametric approach to the estimation of N; a modelling approach based on the CCV would be a new contribution to the literature.

Through the work of this dissertation, we conjecture that every unconstrained joint log-linear model corresponds to a constrained marginal log-linear model and vice versa. For the frequentist work, an algorithm to fully accommodate the model constraints should be developed to construct a software package implementing the frequentist methods presented in Chapter 3. In so doing we would hope to extend and explore the modelling framework for complete contingency tables.

There are a range of issues related to the Bayesian formulations of Chapters 4 and 5. There were computational challenges posed by imperfect mixing, as demonstrated by the high posterior correlations of the random effects and N for the full model. Such difficulties can be overcome in part by obtaining long MCMC chains thinned by a large factor. Alternatively, model reparameterization offers a means to reduce such posterior correlations. We plan on exploiting the ideas of hierarchical centered parameterizations of Gelfand, Sahu, and Carlin (1995) and Gelfand, Sahu, and Carlin (1996), for normal linear mixed models and generalized linear models, respectively, and the ideas of Papaspiliopoulos, Roberts, and Sköld (2003) related to nonhierarchical centered parameterizations.

With our experience of the Bayesian fitting of our model, we anticipate that an uncentered parameterization would be preferred, as we believe that the high posterior correlation is driven by the nonidentifiability of N and the random effects variance. Thus, we would need to tease apart the components of the model expressed directly in terms of N. This offers a clue as to how to obtain a reasonable reparameterization. Further extending the hyperprior structure is another alternative.

The Bayesian model fitting techniques and corresponding MCMC simulation scheme

should be examined for applications to the broader field of generalized log-linear models, as presented in Coull and Agresti (2003). The authors presented a frequentist, not a Bayesian, approach to model fitting, dealing with examples of sparse contingency table data (not, we note, the capture-recapture setting).

We note that the model used in the data analysis of Chapter 5 served the purposes of exploration and was via (a) the reduced model containing only two random effects (CIDs), and, (b) the full model containing all random effects (CIDs). That is, in the latter, none of them is set to zero, contrary to the family of models examined in the frequentist analysis of Chapter 3. Further, we note that although there were more parameters than data points with which to estimate them, by treating the CIDs as random effects, parameter reduction occurs naturally. In the future we hope to explore how the Bayesian paradigm, via the inclusion of informative prior information (rather than the vague priors used in the analyses of Chapter 5), enables us to gain more from the available data but there remains work to be done in order to further explore the issue of nonidentifiability. Some assessment of the number of effective parameters in our model, along the lines of the discussion in Spiegelhalter, Best, Carlin, and van der Linde (2002), could be useful.

In order to reduce the variability in our posterior estimates attributed to the presence of unnecessary CIDs, we propose to use reversible jump MCMC (Green, 1995) to move around the space of the family of models with some CIDs set to zero in order obtain posterior model summaries which are averaged over all models, with greater weight given to preferred models.

A related issue, and one which we have begun to explore with preliminary analyses, is that of non-identifiability of N and the dependence structure present in the capture-recapture data: namely that knowledge of the population size N together with the observed data fully determines the dependence structure of the data. We propose to undertake sensitivity analyses by running the MCMC scheme of Chapter 5 at fixed

levels of the random effects variance σ_γ^2, rather than mixing over the distribution of σ_γ^2, as was the case with the Bayesian analysis of Chapter 5. Preliminary results indicate that there is a strong relationship between N and σ_γ^2, with N increasing as σ_γ^2 increases.

The marginal modelling approach should be explored for its application to complete contingency tables for known N. Furthermore, we seek to develop a model selection criterion for joint log-linear models based on the CIDs (CSDs). We propose to compare the model-based CIDs (CSDs) to those obtained from the saturated model to form the basis of an information criterion, along the lines of the AIC, for example.

There is scope to use the CIDs (CSDs) to measure the information available from pilot studies for capture-recapture studies to obtain information concerning the source dependence structure of the available data. In so doing, the goal would be to obtain informative prior information with which to undertake the Bayesian analysis of the full study, thereby minimizing the problems posed by the nonidentifiability of N and the dependence structure.

Such ideas are related to those of optimal design for capture-recapture study, including issues of optimizing costs. Through pilot studies our ideas may be able to be incorporate into a design framework to develop optimal designs for capture-recapture studies in epidemiology.

Overall, we see that there is scope for a wealth of extensions to a diverse range of applications. The work in this dissertation offers the potential to open up a broad field of research.

Appendix A

Matrices and vectors used in dissertation

A.1 Three sources

For three sources, A, B and C , the incomplete contingency table is given by Table A.1

	A_{Yes}		A_{No}	
	B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	n_{ABC}	$n_{A\bar{B}C}$	$n_{\bar{A}BC}$	$n_{\bar{A}B\bar{C}}$
C_{No}	$n_{AB\bar{C}}$	$n_{A\bar{B}\bar{C}}$	$n_{\bar{A}B\bar{C}}$	$n_{\bar{A}\bar{B}\bar{C}} = ?$

Table A.1: Incomplete Contingency Table: Three Source

$$\mathbf{m} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_{AB} \\ m_{AC} \\ m_{BC} \\ m_{ABC} \end{bmatrix} ; \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ; \boldsymbol{\mu} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}} \\ \mu_{\bar{A}B\bar{C}} \\ \mu_{\bar{A}\bar{B}C} \\ \mu_{AB\bar{C}} \\ \mu_{A\bar{B}C} \\ \mu_{\bar{A}BC} \\ \mu_{ABC} \end{bmatrix},$$

and

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 \\ -2 & 1 & 1 & 1 \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix} = \begin{bmatrix} \log N \\ \log m_A \\ \log m_B \\ \log m_C \end{bmatrix} ; \boldsymbol{\beta}^* = \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}.$$

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix} ; \mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

and

$$\mathbf{Z} = \begin{bmatrix} \mathbf{0} \\ \mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ for } \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{BC} \\ \gamma_{ABC} \end{bmatrix}$$

A.2 Four sources

For four sources, A, B, C and D, the incomplete contingency table is given by Table A.2

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	n_{ABCD}	$n_{A\bar{B}CD}$	$n_{\bar{A}BCD}$	$n_{\bar{A}\bar{B}CD}$
	D_{No}	$n_{ABC\bar{D}}$	$n_{A\bar{B}C\bar{D}}$	$n_{\bar{A}B\bar{C}D}$	$n_{\bar{A}\bar{B}C\bar{D}}$
C_{No}	D_{Yes}	$n_{AB\bar{C}D}$	$n_{A\bar{B}\bar{C}D}$	$n_{\bar{A}B\bar{C}D}$	$n_{\bar{A}\bar{B}\bar{C}D}$
	D_{No}	$n_{AB\bar{C}\bar{D}}$	$n_{A\bar{B}\bar{C}\bar{D}}$	$n_{\bar{A}B\bar{C}\bar{D}}$	$n_{\bar{A}\bar{B}\bar{C}\bar{D}} = ?$

Table A.2: Incomplete Contingency Table: Four Sources

$$\mathbf{m} = \begin{bmatrix} m_A \\ m_B \\ m_C \\ m_D \\ m_{AB} \\ m_{AC} \\ m_{AD} \\ m_{BC} \\ m_{BD} \\ m_{CD} \\ m_{ABC} \\ m_{ABD} \\ m_{ACD} \\ m_{BCD} \\ m_{ABCD} \end{bmatrix} ; \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ; \boldsymbol{\mu} = \begin{bmatrix} \mu_{A\bar{B}\bar{C}\bar{D}} \\ \mu_{\bar{A}B\bar{C}\bar{D}} \\ \mu_{\bar{A}\bar{B}C\bar{D}} \\ \mu_{\bar{A}\bar{B}\bar{C}D} \\ \mu_{AB\bar{C}\bar{D}} \\ \mu_{A\bar{B}C\bar{D}} \\ \mu_{A\bar{B}\bar{C}D} \\ \mu_{\bar{A}BC\bar{D}} \\ \mu_{\bar{A}B\bar{C}D} \\ \mu_{\bar{A}\bar{B}CD} \\ \mu_{ABC\bar{D}} \\ \mu_{AB\bar{C}D} \\ \mu_{A\bar{B}CD} \\ \mu_{\bar{A}BCD} \\ \mu_{ABCD} \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 1 & 0 \\ -1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 1 & 1 \\ -2 & 1 & 1 & 1 & 0 \\ -2 & 1 & 1 & 0 & 1 \\ -2 & 1 & 0 & 1 & 1 \\ -2 & 0 & 1 & 1 & 1 \\ -3 & 1 & 1 & 1 & 1 \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \end{bmatrix} = \begin{bmatrix} \log N \\ \log m_A \\ \log m_B \\ \log m_C \\ \log m_D \end{bmatrix}; \boldsymbol{\beta}^* = \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \end{bmatrix}.$$

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{0} \\ \mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \text{ for } \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{AB} \\ \gamma_{AC} \\ \gamma_{AD} \\ \gamma_{BC} \\ \gamma_{BD} \\ \gamma_{CD} \\ \gamma_{ABC} \\ \gamma_{ABD} \\ \gamma_{ACD} \\ \gamma_{BCD} \\ \gamma_{ABCD} \end{bmatrix}$$

Appendix B

Rules of differentiation

The following presentation is in line with that in Melocco (2002). Similar notation, as well as a description of the results presented in this Appendix can be found in Wand (2002) and Magnus and Neudecker (1988). Searle (1982) provides useful general matrix theory for statistics.

Notation

Let $\mathbf{x}$ be a vector of dimension $p \times 1$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

Let u, f_i, for $i = 1, \ldots q$, be scalar functions of a vector and $\mathbf{f}$ a vector function of a vector of dimension $q \times 1$, such as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_q(\mathbf{x}) \end{bmatrix}.$$

The following definitions hold

$$\frac{du(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial u(\mathbf{x})}{\partial x_1} \\ \frac{\partial u(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial u(\mathbf{x})}{\partial x_p} \end{bmatrix}$$

and

$$\begin{aligned} \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} &= \left[\left(\frac{df_1(\mathbf{x})}{d\mathbf{x}} \right) \left(\frac{df_2(\mathbf{x})}{d\mathbf{x}} \right) \cdots \left(\frac{df_q(\mathbf{x})}{d\mathbf{x}} \right) \right] \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_q(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_q(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_q} & \frac{\partial f_2(\mathbf{x})}{\partial x_p} & \cdots & \frac{\partial f_q(\mathbf{x})}{\partial x_p} \end{bmatrix} \end{aligned}$$

Basic Rules

Let $\mathbf{A}$ be an $m \times p$ matrix and $\mathbf{a}$ a p-dimensional vector. We use $D_a = diag(\mathbf{a})$ to denote the diagonal matrix with $\mathbf{a}$ as the diagonal and $\mathbf{a} \circ \mathbf{b}$ to denote the direct product of vectors $\mathbf{a}$ and $\mathbf{b}$. Let $\mathbf{f}$ and $\mathbf{g}$ be vector functions of a vector.
The following results hold:

1. $\frac{d}{d\mathbf{x}}\mathbf{A}(\mathbf{x}) = \mathbf{A}'$

2. $\frac{d}{d\mathbf{x}}\mathbf{a} \circ \mathbf{x} = \frac{d}{d\mathbf{x}} D_{\mathbf{a}}\mathbf{x} = D_{\mathbf{a}}$

3. $\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = \frac{d}{d\mathbf{x}}\mathbf{g}(\mathbf{x}) \frac{d}{d\mathbf{g}(\mathbf{x})}\mathbf{f}(\mathbf{g}(\mathbf{x}))$

4. $\frac{d}{d\mathbf{x}}\left[(\mathbf{f}(\mathbf{x}))'\,\mathbf{g}(\mathbf{x})\right] = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}}\mathbf{g}(\mathbf{x}) + \frac{d\mathbf{g}(\mathbf{x})}{d\mathbf{x}}\mathbf{f}(\mathbf{x})$

5. $\frac{d}{d\mathbf{x}}\left(\mathbf{f}(\mathbf{x}) \circ \mathbf{g}(\mathbf{x})\right) = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} D_{\mathbf{g}(\mathbf{x})} + \frac{d\mathbf{g}(\mathbf{x})}{d\mathbf{x}} D_{\mathbf{f}(\mathbf{x})}$

Useful Results

The following useful results can be obtained using those outlined above.

$$\frac{d}{d\mathbf{x}}\mathbf{A}\mathbf{f}(\mathbf{x}) = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}}\mathbf{A}'$$

Let

$$\mathbf{x}^{-cI} = \begin{bmatrix} \frac{1}{x_1^c} \\ \frac{1}{x_2^c} \\ \vdots \\ \frac{1}{x_p^c} \end{bmatrix}$$

and let $\mathbf{l}$ be a vector function such that

$$\mathbf{l}(\mathbf{x}) = \begin{bmatrix} l(x_1) \\ l(x_2) \\ \vdots \\ l(x_p) \end{bmatrix}$$

Then

$$\frac{d}{d\mathbf{x}}\mathbf{l}(\mathbf{x}) = diag\left(\frac{\partial l(x_i)}{\partial x_i}\right)_i$$

In particular

$$\frac{d}{d\mathbf{x}}\exp(\mathbf{x}) = D_{\exp(\mathbf{x})},$$

$$\frac{d}{d\mathbf{x}}\log(\mathbf{x}) = D_{\mathbf{x}}^{-1} = D_{\mathbf{x}^{-I}}$$

Appendix C

Simulated data set for Chapter 2

Details of the three-source data set

A three-source capture-recapture data set was generated for a population of size 1000, according to a scheme which enforced positive dependence. The theoretical CSDs, deemed reasonable from real data sets observed in the literature, are set at

$$C_{AB} = 1.333, C_{AC} = 1.1304, C_{BC} = 1.1014, \text{ and, } C_{ABC} = 1.623, \tag{C.1}$$

For this scheme, the simulated capture-recapture data set consists of 754 individuals observed out of the population of size 1000. The distribution of these 754 individuals amongst the three sources is given in Table C.1.

The observed cell counts $\mathbf{n}_{\text{cell}}$ and marginal counts $\mathbf{n}_{\text{marg}}$ corresponding to the

	B_{Yes}		B_{No}	
	C_{Yes}	C_{No}	C_{Yes}	C_{No}
A_{Yes}	117	96	64	72
A_{No}	109	134	162	?

Table C.1: Observed sample simulated from population size 1000

incomplete contingency table data of Table C.1 are given by

$$\mathbf{n}_{\text{cell}} = \begin{bmatrix} n_{A\bar{B}\bar{C}} \\ n_{\bar{A}B\bar{C}} \\ n_{\bar{A}\bar{B}C} \\ n_{AB\bar{C}} \\ n_{A\bar{B}C} \\ n_{\bar{A}BC} \\ n_{ABC} \end{bmatrix} = \begin{bmatrix} 72 \\ 134 \\ 62 \\ 96 \\ 64 \\ 109 \\ 117 \end{bmatrix} \text{ and } \mathbf{n}_{\text{marg}} = \mathbf{A}\mathbf{n}_{\text{cell}} = \begin{bmatrix} n_A \\ n_B \\ n_C \\ n_{AB} \\ n_{AC} \\ n_{BC} \\ n_{ABC} \end{bmatrix} = \begin{bmatrix} 349 \\ 456 \\ 452 \\ 213 \\ 181 \\ 226 \\ 117 \end{bmatrix}, \tag{C.2}$$

The pairwise Petersen estimates are given by 747, 872 and 912, for sources A and B, sources A and C, and sources B and C, respectively. These estimates suggest that there is positive dependence between sources A and B, since the Petersen estimate of 747 is lower than the observed number of 754 individuals. For the other two pairs, i.e. A and C, and B and C, the Petersen estimates are larger than the observed number of 754, which tends to suggest negative dependence. Note that this is not in line with the nature of the CSDs which generated the simulated data set, although since the CSDs for these pairs are lower than for the pair A and B, the positive dependence of the generating mechanism is less strong than for sources A and B.

For completeness, we describe the details of the data generation scheme used to generate the data set in Table C.1 according to the CSDs of (C.1). First, each individual was randomly assigned one of two covariates, either level 1 or level 2, with

probabilities

$$Pr[Z = 1] = 0.7, \text{ and, } Pr[Z = 2] = 1 - Pr[Z = 1] = 0.3. \tag{C.3}$$

Source membership was assigned conditional on covariate level to enable us to calculate the true population-level CSDs. The covariate-based CSD is given by Vandal et al. (2005) as follows for a set of sources $\mathcal{Q}$ and a covariate Z

$$C_{\mathcal{Q}} = \sum_{\text{all } z} \frac{\prod_{S \in \mathcal{Q}} \Pr[Z = z|S]}{\Pr[Z = z]^{|\mathcal{Q}|-1}}, \tag{C.4}$$

using the assumption of conditional independence of source membership given covariate level given by

$$\Pr\left[\bigcap_{S \in \mathcal{Q}} S|Z = z\right] = \prod_{S \in \mathcal{Q}} \Pr\left[S|Z = z\right]. \tag{C.5}$$

As a second step in the data generation scheme, source membership was randomly assigned conditional on covariate level, according to the following probabilities (set by us)

$$\begin{bmatrix} Pr[A|Z = 1] & Pr[A|Z = 2] \\ Pr[B|Z = 1] & Pr[B|Z = 2] \\ Pr[C|Z = 1] & Pr[C|Z = 2] \end{bmatrix} = \begin{bmatrix} 0.2 & 0.7 \\ 0.3 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}. \tag{C.6}$$

Conditional on covariate level, membership in a source was mutually independent of membership in any other source. Using the assumption of conditional independence (C.5) and the generating probabilities, given by (C.3) and (C.6), it is straightforward to evaluate the true (unknown) covariate-based coefficients of source dependence, according to (C.4). The population-level CSDs calculated according to this scheme are given above by (C.1).

The theoretical highest-level marginal probabilities are given by

$$p_A = Pr[A] = Pr[A|Z = 1]Pr[Z = 1] + Pr[A|Z = 2]Pr[Z = 2] = 0.35,$$

$$p_B = Pr[B] = Pr[B|Z = 1]Pr[Z = 1] + Pr[B|Z = 2]Pr[Z = 2] = 0.45,$$

and,

$$p_C = Pr[C] = Pr[C|Z=1]Pr[Z=1] + Pr[C|Z=2]Pr[Z=2] = 0.46. \qquad \text{(C.7)}$$

Of course, these probabilities need not sum to 1 since they are not mutually exclusive events. Consequently we obtain the following theoretical values for the source-specific covariate distributions

$$\begin{bmatrix} Pr[Z=1|A] & Pr[Z=1|B] & Pr[Z=1|C] \\ Pr[Z=2|A] & Pr[Z=2|B] & Pr[Z=2|C] \end{bmatrix} = \begin{bmatrix} 0.4 & 0.467 & 0.609 \\ 0.6 & 0.533 & 0.391 \end{bmatrix}.$$

In reality, due to the finite nature of the population of size 1000, we obtain an observed matrix of

$$\begin{bmatrix} \widehat{Pr[Z=1|A]} & \widehat{Pr[Z=1|B]} & \widehat{Pr[Z=1|C]} \\ \widehat{Pr[Z=2|A]} & \widehat{Pr[Z=2|B]} & \widehat{Pr[Z=2|C]} \end{bmatrix} = \begin{bmatrix} 0.372 & 0.4496 & 0.6150 \\ 0.628 & 0.5504 & 0.3850 \end{bmatrix}$$

The observed covariate distributions in the three sources are quite different to each other and to the population generating probabilities of 0.7 and 0.3. Again, due to the finite nature of the simulation of 1000 individuals from such generating probabilities the actual covariate distribution in the 1000 individuals is given by

$$\widehat{Pr[Z=1]} = 0.693, \text{and}, \widehat{Pr[Z=2]} = 1 - \widehat{Pr[Z=1]} = 0.307$$

compared to (C.3). That in the 754 observed individuals is

$$\widehat{Pr[Z=1|\text{obs}]} = 0.6061008, \text{and}, \widehat{Pr[Z=2|\text{obs}]} = 1 - \widehat{Pr[Z=1|\text{obs}]} = 0.3938992.$$

For completeness, the four remaining marginal probabilities can be shown to be given by

$$p_{AB} = 0.21, p_{AC} = 0.182, p_{BC} = 0.228, p_{ABC} = 0.1176.$$

Appendix D

Appendix for Chapter 2

D.1 Proof of Theorem 2.3

Theorem

Let $\mathcal{R}$, $\mathcal{T}$ and $\mathcal{S}$ denote arbitrary sets of sources such that $\mathcal{R} \cap \mathcal{S} = \varnothing$. We recall the HJLLM conventions presented in Section 1.2.1. Then if the groups $[\mathcal{R}, \mathcal{T}]$ and $[\mathcal{S}, \mathcal{T}]$ appear in the HJLLM specification with no $[\mathcal{A}, \mathcal{B}]$ specification where $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{S}$ (so that sources in $\mathcal{R}$ and sources in $\mathcal{S}$ are conditionally independent given sources in $\mathcal{T}$)

$$C_{\mathcal{R} \cup \mathcal{S} \cup \mathcal{T}} = C_{\mathcal{R} \cup \mathcal{T}} + C_{\mathcal{S} \cup \mathcal{T}} - C_{\mathcal{T}}.$$

Proof of Theorem 2.3

Proof. Define $\mathcal{R}_\cap = \cap_{\mathcal{S} \in \mathcal{R}} \mathcal{S}$, and similarly for $\mathcal{S}_\cap$ and $\mathcal{T}_\cap$.

$$
\begin{aligned}
c_{\mathcal{R}\cup\mathcal{S}\cup\mathcal{T}} &= \frac{Pr[\mathcal{R}_\cap \cap \mathcal{S}_\cap \cap \mathcal{T}_\cap]}{\prod_{S\in\mathcal{R},\mathcal{S},\mathcal{T}} Pr[S]}, \text{ by definition} \\
&= \frac{Pr[\mathcal{R}_\cap \cap \mathcal{S}_\cap | \mathcal{T}_\cap] Pr[\mathcal{T}_\cap]}{\prod_{S\in\mathcal{R},\mathcal{S},\mathcal{T}} Pr[S]} \\
&= \frac{Pr[\mathcal{R}_\cap | \mathcal{T}_\cap] Pr[\mathcal{S}_\cap | \mathcal{T}_\cap] Pr[\mathcal{T}_\cap]}{\prod_{S\in\mathcal{R},\mathcal{S},\mathcal{T}} Pr[S]}, \text{ by assumption of conditional independence} \\
&= \frac{(Pr[\mathcal{R}_\cap \cap \mathcal{T}_\cap]/Pr[\mathcal{T}_\cap])\,(Pr[\mathcal{S}_\cap \cap \mathcal{T}_\cap]/Pr[\mathcal{T}_\cap])\,Pr[\mathcal{T}_\cap]}{\prod_{S\in\mathcal{R},\mathcal{S},\mathcal{T}} Pr[S]} \\
&= \frac{Pr[\mathcal{R}_\cap \cap \mathcal{T}_\cap] Pr[\mathcal{S}_\cap \cap \mathcal{T}_\cap]}{Pr[\mathcal{T}_\cap] \prod_{S\in\mathcal{R},\mathcal{S},\mathcal{T}} Pr[S]} \\
&= \frac{Pr[\mathcal{R}_\cap \cap \mathcal{T}_\cap]}{\prod_{S\in\mathcal{R},\mathcal{T}} Pr[S]} \frac{Pr[\mathcal{S}_\cap \cap \mathcal{T}_\cap]}{Pr[\mathcal{T}_\cap] \prod_{S\in\mathcal{S}} Pr[S]} \\
&= c_{\mathcal{R},\mathcal{T}} \frac{Pr[\mathcal{S}_\cap \cap \mathcal{T}_\cap]}{Pr[\mathcal{T}_\cap] \prod_{S\in\mathcal{S}} Pr[S]} \\
&= c_{\mathcal{R},\mathcal{T}} \frac{Pr[\mathcal{S}_\cap \cap \mathcal{T}_\cap]}{\prod_{S\in\mathcal{S},\mathcal{T}} Pr[S]} \frac{\prod_{S\in\mathcal{T}} Pr[S]}{Pr[\mathcal{T}_\cap]} \\
&= \frac{c_{\mathcal{R},\mathcal{T}} c_{\mathcal{S},\mathcal{T}}}{c_{\mathcal{T}}},
\end{aligned}
$$

whence the result follows. □

D.2 Proof of Proposition 2.3

This proposition states the relationship between the CCV and CSD for three sources.

Proposition

Consider three sources, A, B and C. Let ω_{ABC} denote the three-way CCV. Then

$$\omega_{ABC} = (c_{ABC} - 1) - (c_{AB} - 1) - (c_{BC} - 1) - (c_{AC} - 1).$$

If the CSDs are close to zero, then

$$\omega_{ABC} \simeq C_{ABC} - (C_{AB} + C_{BC} + C_{AC}) + (C_A + C_B + C_C)$$

Proof of Proposition 2.3

Proof.

$$\begin{aligned}
\omega_{ABC} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(X_{ia}-\mu_a)(X_{ib}-\mu_b)(X_{ic}-\mu_c)\right]/(\mu_a\mu_b\mu_c) \\
&= \frac{1}{N\mu_a\mu_b\mu_c}\sum_{i=1}^{N}\mathbb{E}\left[X_{ia}X_{ib}X_{ic} - X_{ia}X_{ib}\mu_c - \mu_a X_{ib}X_{ic} + \mu_a\mu_c X_{ib} - \mu_b X_{ia}X_{ic}\right. \\
&\qquad \left. + \mu_b\mu_c X_{ia} + \mu_a\mu_b X_{ic} - \mu_a\mu_b\mu_c\right] \\
&= \frac{1}{N\mu_a\mu_b\mu_c}\left[\sum_{i=1}^{N}\mathbb{E}\left[X_{ia}X_{ib}X_{ic}\right] - \mu_c\sum_{i=1}^{N}\mathbb{E}\left[X_{ia}X_{ib}\right] - \mu_a\sum_{i=1}^{N}E[X_{ib}X_{ic}]\mu_a\mu_c\sum_{i=1}^{N}\mathbb{E}\left[X_{ib}\right]\right. \\
&\qquad \left. -\mu_b\sum_{i=1}^{N}\mathbb{E}\left[X_{ia}X_{ic}\right] + \mu_b\mu_c\sum_{i=1}^{N}\mathbb{E}\left[X_{ia}\right] + \mu_a\mu_b\sum_{i=1}^{N}\mathbb{E}\left[X_{ic}\right] - N\mu_a\mu_b\mu_c\right] \\
&= \frac{1}{N\mu_a\mu_b\mu_c}\left[\sum_{i=1}^{N}Pr\left[A\cap B\cap C|I=i\right] - \mu_c\sum_{i=1}^{N}Pr[A\cap B|I=i]\right. \\
&\qquad - \mu_a\sum_{i=1}^{N}Pr[B\cap C|I=i] + \mu_a\mu_c\sum_{i=1}^{N}Pr[B|I=i] - \mu_b\sum_{i=1}^{N}Pr[A\cap C|I=i] \\
&\qquad \left. + \mu_b\mu_c\sum_{i=1}^{N}Pr[A|I=i] + \mu_a\mu_b\sum_{i=1}^{N}\mathbb{E}\left[C|I=i\right] - N\mu_a\mu_b\mu_c\right] \\
&= \frac{Pr[A\cap B\cap C]}{Pr[A]Pr[B]Pr[C]} - \frac{Pr[A\cap B]}{Pr[A]Pr[B]} - \frac{Pr[B\cap C]}{Pr[B]Pr[C]} - \frac{Pr[A\cap C]}{Pr[A]Pr[C]} + 2 \\
&= (c_{ABC}-1) - (c_{AB}-1) - (c_{BC}-1) - (c_{AC}-1) \\
&\simeq \log c_{ABC} - \log c_{AB} - \log c_{BC} - \log c_{AC}, \text{ from the first order Taylor expansion for log} \\
&= \log\left(\frac{c_{ABC}}{c_{AB}c_{BC}c_{AC}}\right) \\
&= \log\left(\frac{c_{ABC}c_Ac_Bc_C}{c_{AB}c_{BC}c_{AC}}\right), \text{ since by definition } c_A = c_B = c_C = 1 \\
&= C_{ABC} - (C_{AB} + C_{BC} + C_{AC}) + (C_A + C_B + C_C), \text{ by definition of the CSDs}
\end{aligned}$$

□

D.3 Proof of Theorem 2.5

Theorem

Let $\mathcal{Q}$ be any set of sources and denote by $C_{\mathcal{Q}}$ the CSD associated with this set of sources. Then

$$C_{\mathcal{Q}} = \sum_{S \subset \mathcal{Q}} \gamma_S. \tag{D.1}$$

Proof of Theorem 2.5

Lemma required for proof

In order to prove Theorem 2.5, we need the following lemma.

Lemma *Let $i_r = i$ if r is even and $i_r = i + 1$ if r is odd, $i = 1, \ldots, r$. Then*

$$\sum_{i=0}^{r-1} (-1)^{i_r} \binom{r}{i} = -1$$

Proof. Suppose first that r is odd. Then

$$\begin{aligned}
\sum_{i=0}^{r} (-1)^i \binom{r}{i} &= \sum_{i=0}^{(r-1)/2} (-1)^i \binom{r}{i} + \sum_{i=(r+1)/2}^{r} (-1)^i \binom{r}{i} \\
&= \sum_{i=0}^{(r-1)/2} (-1)^i \binom{r}{i} + \sum_{i=(r+1)/2}^{r} (-1)^{r-i+1} \binom{r}{r-i} \qquad \text{(D.2)} \\
&= \sum_{i=0}^{(r-1)/2} (-1)^i \binom{r}{i} + \sum_{j=(r-1)/2}^{0} (-1)^{j+1} \binom{r}{j}, \text{ letting } j = r - i, \\
&= \sum_{i=0}^{(r-1)/2} \left[(-1)^i + (-1)^{i+1}\right] \binom{r}{i} = 0 \qquad \text{(D.3)}
\end{aligned}$$

In this case,

$$\sum_{i=0}^{r-1} (-1)^i \binom{r}{i} = -(-1)^r \binom{r}{r} = (-1)^{r+1} = 1, \tag{D.4}$$

so that

$$\sum_{i=0}^{r-1}(-1)^{ir}\binom{r}{i} = -1,$$

as required.

If r is even, then

$$\begin{aligned}\sum_{i=0}^{r}(-1)^{i}\binom{r}{i} &= (-1)^{0}\binom{r}{0} + \sum_{i=1}^{r-1}(-1)^{i}\binom{r}{i} + (-1)^{r}\binom{r}{r} \\ &= 2 + \sum_{i=1}^{r-1}(-1)^{i}\left[\binom{r-1}{i} + \binom{r-1}{i-1}\right] \\ &= 2 + \sum_{i=1}^{r-1}(-1)^{i}\binom{r-1}{i} + \sum_{i=1}^{r-1}(-1)^{i}\binom{r-1}{i-1}\end{aligned}$$

Now

$$\sum_{i=1}^{r-1}(-1)^{i}\binom{r-1}{i} = \sum_{i=0}^{r-1}(-1)^{i}\binom{r-1}{i} - (-1)^{0}\binom{r-1}{0} = -1$$

by (D.3) since $r-1$ is odd, and

$$\begin{aligned}\sum_{i=1}^{r-1}(-1)^{i}\binom{r-1}{i-1} &= \sum_{i=0}^{r-2}(-1)^{i+1}\binom{r-1}{i} \\ &= -\sum_{i=0}^{r-2}(-1)^{i}\binom{r-1}{i} = -\left[(-1)^{r-1+1}\right] = -1\end{aligned}$$

by (D.4), once again since $r-1$ is odd.

Hence

$$\sum_{i=0}^{r}(-1)^{i}\binom{r}{i} = 0 \tag{D.5}$$

when r is even (or odd, by (D.3), whence

$$\sum_{i=0}^{r-1}(-1)^{ir}\binom{r}{i} = \sum_{i=0}^{r-1}(-1)^{i}\binom{r}{i} = -(-1)^{r}\binom{r}{r} = (-1)^{r+1} = -1,$$

once again. □

Proof of Theorem 2.5 using previous Lemma

In this section we present the proof of Theorem 2.5.

Proof. We have, for $\mathcal{Q}$ a set of sources of size n,

$$\gamma_{\mathcal{Q}} = \sum_{j=1}^{n}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\|\mathcal{R}|=j}} C_{\mathcal{R}} = C_{\mathcal{Q}} + \sum_{j=1}^{n-1}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\|\mathcal{R}|=j}} C_{\mathcal{R}}, \tag{D.6}$$

by Definition 2.4, where $j_n = j$ if n is even and $j+1$ if n is odd.

We proceed by induction. Let $\mathcal{Q}^*$ be a set of sources of cardinality m. Suppose first that $m = 1$. Then $C_{\mathcal{Q}} = C_S = \gamma_S$ by Definition 2.4, and $\gamma_S = \gamma_S + \gamma_\varnothing$, which shows that the induction hypothesis holds true for $m = 1$. Suppose now that $m \geq 1$

$$C_{\mathcal{Q}^*} = \sum_{\mathcal{R}\subset\mathcal{Q}^*} \gamma_{\mathcal{R}} = \sum_{j=1}^{m} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}^*\\|\mathcal{R}|=j}} \gamma_{\mathcal{R}} \text{ for } m = 1, 2, \ldots, n-1. \tag{D.7}$$

We show that if $\mathcal{Q} = \mathcal{Q}^* \cup \{S\}$, then $C_{\mathcal{Q}} = \sum_{\mathcal{S}\subset\mathcal{Q}}\gamma_{\mathcal{S}}$.

From (D.6), we obtain

$$\begin{aligned}
C_{\mathcal{Q}} &= \gamma_{\mathcal{Q}} + \sum_{j=1}^{n-1}(-1)^{j_n+1} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\|\mathcal{R}|=j}} C_{\mathcal{R}} \\
&= \gamma_{\mathcal{Q}} + \sum_{j=1}^{n-1}(-1)^{j_n+1} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\|\mathcal{R}|=j}} \sum_{k=1}^{j} \sum_{\substack{\mathcal{S}\subset\mathcal{R}\\|\mathcal{S}|=k}} \gamma_{\mathcal{S}} \text{ from (D.7)} \\
&= \gamma_{\mathcal{Q}} + \sum_{k=1}^{n-1} \sum_{\substack{\mathcal{S}\subset\mathcal{Q}\\|\mathcal{S}|=k}} \gamma_{\mathcal{S}} \sum_{j=k}^{n-1}(-1)^{j_n+1} \sum_{\substack{\mathcal{R}\supset\mathcal{S}\\|\mathcal{R}|=j}} 1,
\end{aligned}$$

switching the order of summation.

Now the term $\sum_{\substack{\mathcal{R}\supset\mathcal{S}\\|\mathcal{R}|=j}} 1$ is just the number of supersets of $\mathcal{S}$ that are of size $j \geq k$. This is obtained by selecting a further $j-k$ sources in addition to the sources in $\mathcal{S}$ to form $\mathcal{R}$, from a possibility of $n-k$ sources in $\mathcal{Q}\setminus\mathcal{S}$. Thus

$$\sum_{\substack{\mathcal{R}\supset\mathcal{S}\\|\mathcal{R}|=j}} 1 = \binom{n-k}{j-k}.$$

Further,

$$\sum_{j=k}^{n-1}(-1)^{j_n+1}\sum_{\substack{\mathcal{R}\supset\mathcal{S}\\|\mathcal{R}|=j}} 1 = \sum_{j=k}^{n-1}(-1)^{j_n+1}\binom{n-k}{j-k}$$

$$= \sum_{j=0}^{n-k-1}(-1)^{j_n+k+1}\binom{n-k}{j} = -\sum_{j=0}^{n-k-1}(-1)^{j_n+k}\binom{n-k}{j}$$

Now $(j_{n-k} = j_n + k) \mod 2$, since

- if n and k are even, $(j_{n-k} = j = j+k = j_n+k) \mod 2$;
- if n and k are odd, $(j_{n-k} = j = j+1+k = j_n+k) \mod 2$;
- if n is even and k is odd, $(j_{n-k} = j+1 = j+k = j_n+k) \mod 2$;
- if n is odd and k is even, $(j_{n-k} = j+1 = j+1+k = j_n+k) \mod 2$.

In particular, $(-1)^{j_n+k} = (-1)^{j_{n-k}}$, so we obtain

$$-\sum_{j=0}^{n-k-1}(-1)^{j_n+k}\binom{n-k}{j} = -\sum_{j=0}^{n-k-1}(-1)^{j_{n-k}}\binom{n-k}{j} = -(-1) = 1,$$

by the previous Lemma, so

$$C_{\mathcal{Q}} = \gamma_{\mathcal{Q}} + \sum_{k=1}^{n-1}\sum_{\substack{\mathcal{S}\subset\mathcal{Q}\\|\mathcal{S}|=k}}\gamma_{\mathcal{S}} = \sum_{\mathcal{S}\subset\mathcal{Q}}\gamma_{\mathcal{S}},$$

as required, since $\gamma_{\varnothing} = 0$ by definition. □

D.4 Proof of Theorem 2.6

Theorem

Let $\mathcal{Q}$ be a set of sources and $\pi_{\mathcal{Q}} = \log P\left[\cap_{S\in\mathcal{Q}} S\right]$. Then for $|\mathcal{Q}| \geq 2$

$$\gamma_{\mathcal{Q}} = \sum_{j=1}^{n} (-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q} \\ |\mathcal{R}|=j}} \pi_{\mathcal{R}}. \tag{D.8}$$

Proof of Theorem 2.6

Proof. A direct proof is given below. Here we instead write $\dot{\gamma}_{\mathcal{R}} \doteq \gamma_{\mathcal{R}}$ if $|\mathcal{R}| \geq 2$ and$\dot{\gamma}_{\mathcal{R}} \doteq \pi_{\mathcal{R}}$ if $|\mathcal{R}| = 1$. Then from (2.16),

$$\pi_{\mathcal{Q}} = \sum_{\mathcal{R}\subset\mathcal{Q}} \dot{\gamma}_{\mathcal{R}},$$

since $\gamma_{\mathcal{R}} = 0$ for $|\mathcal{R}| = 1$. But this relates $\pi_{\mathcal{Q}}$ and the $\dot{\gamma}_{\mathcal{R}}$'s in a way that is formally identical to (D.1). The inverse of this relationship will thus be formally identical to (D.6), yielding

$$\dot{\gamma}_{\mathcal{Q}} = \sum_{j=1}^{n} (-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q} \\ |\mathcal{R}|=j}} \pi_{\mathcal{R}};$$

but $\dot{\gamma}_{\mathcal{Q}} = \gamma_{\mathcal{Q}}$ for $|\mathcal{Q}| \geq 2$, which completes the proof. □

Alternative Proof of Theorem 2.6

$$\begin{aligned}
\gamma_{\mathcal{Q}} &= \sum_{j=1}^{n}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} C_{\mathcal{R}} \\
&= \sum_{j=1}^{n}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} \left(\pi_{\mathcal{R}} - \sum_{S\in\mathcal{R}} \pi_S\right), \text{ by definition} \\
&= \sum_{j=1}^{n}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} \pi_{\mathcal{R}} - \sum_{j=1}^{n}(-1)^{j_n} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} \sum_{S\in\mathcal{R}} \pi_S.
\end{aligned}$$

It is therefore sufficient to show that the second sum of the right-hand side is 0:

$$\begin{aligned}
\sum_{j=1}^{n}(-1)^{j} \sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} \sum_{S\in\mathcal{R}} \pi_S &= \sum_{S\in\mathcal{Q}} \pi_S \sum_{j=1}^{n}(-1)^{j} \sum_{\substack{\mathcal{R}\ni S\\ \mathcal{R}\subset\mathcal{Q}}} 1 \\
&= \sum_{S\in\mathcal{Q}} \pi_S \sum_{j=1}^{n}(-1)^{j}\binom{n-1}{j-1} \\
&= \sum_{S\in\mathcal{Q}} \pi_S \sum_{j=0}^{n-1}(-1)^{j}\binom{n-1}{j} = 0 \qquad \text{(D.9)}
\end{aligned}$$

by (D.5), whence $\sum_{j=1}^{n}(-1)^{j_n}\sum_{\substack{\mathcal{R}\subset\mathcal{Q}\\ |\mathcal{R}|=j}} \sum_{S\in\mathcal{R}} \pi_S = 0$, since the expression is just (D.9) multiplied by $+1$ or -1.

D.5 Details of the HJLLM of conditional independence of Section 2.4.2.

Under the HJLLM specified by $[AB][AC]$, where $B \perp\!\!\!\perp C|A$, we obtain

$$\begin{aligned}
Pr[B \cap C] &= Pr[B \cap C \cap A] + Pr[B \cap C \cap \bar{A}] \\
&= Pr[B \cap C|A]Pr[A] + Pr[B \cap C|\bar{A}]Pr[\bar{A}] \\
&= Pr[B|A]Pr[C|A]Pr[A] + Pr[B|\bar{A}]Pr[C|\bar{A}]Pr[\bar{A}] \\
&= \frac{Pr[B \cap A]}{Pr[A]}\frac{Pr[C \cap A]}{Pr[A]}Pr[A] + \frac{Pr[B \cap \bar{A}]}{Pr[\bar{A}]}\frac{Pr[C \cap \bar{A}]}{Pr[\bar{A}]}Pr[\bar{A}] \\
&= \frac{Pr[B \cap A]Pr[C \cap A]}{Pr[A]} + \frac{Pr[B \cap \bar{A}]Pr[C \cap \bar{A}]}{Pr[\bar{A}]}
\end{aligned}$$

Thus we have

$$\begin{aligned}
c_{BC} &= \frac{Pr[B \cap A]Pr[C \cap A]}{Pr[A]Pr[B]Pr[C]} + \frac{Pr[B \cap \overline{A}]Pr[C \cap \overline{A}]}{Pr[\overline{A}]Pr[B]Pr[C]} \\
&= Pr[A]c_{AB}c_{AC} + \frac{1}{1 - Pr[A]}\left(1 - \frac{Pr[B \cap A]}{Pr[B]}\right)\left(1 - \frac{PrC \cap A}{Pr[C]}\right) \\
&= Pr[A]c_{AB}c_{AC} + \frac{1}{1 - Pr[A]}\left(1 - Pr[A]c_{AB}\right)\left(1 - Pr[A]c_{AC}\right) \\
&= Pr[A]c_{AB}c_{AC} + \frac{1}{1 - Pr[A]}\left(1 - Pr[A](c_{AB} + c_{AC}) + Pr[A]^2 c_{AB}c_{AC}\right) \\
&= \frac{1}{1 - Pr[A]}\left(Pr[A]c_{AB}c_{AC} - Pr[A]^2 c_{AB}c_{AC} + 1 - Pr[A](c_{AB} + c_{AC})\right. \\
&\qquad \left. + Pr[A]^2 c_{AB}c_{AC}\right) \\
&= \frac{Pr[A]}{1 - Pr[A]}\left(c_{AB}c_{AC} - (c_{AB} + c_{AC}) + \frac{1}{Pr[A]}\right)
\end{aligned}$$

D.6 Example of a non-hierarchical dependence structure

Consider tossing two fair die. The sample space is given by

$$\{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(5,6),(6,6)\}.$$

Let

$$A = \{\text{sum is odd, i.e. 3,5,7,9 or 11}\}, \text{ so } Pr[A] = 18/36 = 1/2$$
$$B = \{\text{sum is 3,5,6 or 12}\}, \text{ so } Pr[B] = 12/36 = 1/3$$
$$C = \{\text{sum is 4,5,10 or 11}\}, \text{ so } Pr[C] = 12/36 = 1/3$$

Then

$$Pr[A \cap B] = Pr[\text{sum is 3 or 5}] = 6/36 = 1/6 = Pr[A]Pr[B]$$
$$Pr[A \cap C] = Pr[\text{sum is 5 or 11}] = 6/36 = 1/6 = Pr[A]Pr[C]$$
$$Pr[B \cap C] = Pr[\text{sum is 5}] = 4/36 = 1/9 = Pr[B]Pr[C]$$

Therefore

$$A \perp\!\!\!\perp B,\ A \perp\!\!\!\perp C \text{ and } B \perp\!\!\!\perp C$$

However,

$$Pr[A \cap B \cap C] = Pr[\text{sum is 5}] = 1/9 \neq 1/18 = Pr[A]Pr[B]Pr[C],$$

which shows that A, B and C are not independent events. Thus, is is possible for three events to be pairwise dependent but jointly dependent i.e. which exhibit three-way dependence.

Appendix E

Simulation of data sets for Chapters 3 and 5

Introduction

We simulate data generated according to a known model and known likelihood in order to demonstrate the performance of the model. The model is the marginal model parameterized in terms of the CIDs, as given by Definition 2.7; the likelihood the multinomial (see Section 1.1.3). We consider the four-source setting.

In order to generate data sets derived from the true model, we first fix the true population size N together with the single source marginal probabilities p_A, p_B, p_C and p_D. The eleven remaining marginal probabilities are calculated according to the assumed dependence structure. Whenever the assumed dependence structure places no constraints on a marginal probability, we fix it at some reasonable value, which must be consistent with the nested ordering of marginal probabilities. This will be demonstrated below for the dependence structures we consider here. Note that, by

specifying the marginal dependence structure, there is an equivalent set of fixed CIDs. Here we describe the steps to be taken to generate such data.

Simulating capture-recapture data under a multinomial likelihood

When employing the multinomial likelihood, inference applies to the parameter N and not $\mathbb{E}[N]$, as is the case under a Poisson likelihood. In the latter case, N is a random variable rather than a parameter, as is the case when a multinomial model is assumed. The following scheme is employed for the general K-source case:

- Fix N and the single-source marginal probabilities $p_A, \ldots, p_K$.
- Specify the non-single source probabilities according to some dependence structure (e.g. independence or conditional independence). In some instances, for certain dependence structures, they will be derived from the single-source marginal probabilities. Fixed values will be chosen if not specified by the structure. Note that using the fully specified dependence structure and corresponding marginal probabilities the corresponding theoretical CIDs can be calculated.
- Obtain the corresponding cell probabilities, $\mathbf{p}_\mu$, using the relationship $\mathbf{p}_\mu = \mathbf{A}^{-1}\mathbf{p}_\mathbf{m}$, where the matrix $\mathbf{A}^{-1}$ transforms marginal counts into cell counts and similarly marginal probabilities into cell probabilities and $\mathbf{p}_\mathbf{m}$ denotes marginal probabilities.
- Using the likelihood assumption that the 2^K cells of the complete contingency table follow a multinomial distribution simulate a 2^K vector of observations from a single realization of the the multinomial distribution with cell probabilities

given by the $2^K - 1$ vector $\mathbf{p}_\mu$ and probability $1 - \mathbf{e}'\mathbf{p}_\mu$ for the remaining cell, where $\mathbf{e}$ is a vector of 1s of length $2^K - 1$.

- Use the first $2^K - 1$ entries of the vector of simulated observations as the capture-recapture data set.

The data sets

We will consider two dependence structures. The first, a conditional independence structure and the second, a nonhierarchical dependence structure. For each, we fix the true population size at $N = 1000$. The single-source marginal probabilities p_A, p_B p_C and p_D are fixed at similar values in each case. We determine the eleven remaining marginal probabilities in order to fully specify the true underlying dependence structure. Finally, in both, we generate a single data set from each of these three known models. The underlying population parameters (i.e. N, p_A, etc.) are thought to be reasonable values based on evidence from the literature. Most epidemiological capture-recapture studies seek to eumerate reasonably sized populations, hence the choice of $N = 1000$ and the marginal source probabilities described below.

Conditional independence

We assume that the true underlying model is that of conditional independence of B and C given A, denoted by $[AB][AC][D]$ in the notation of Christensen (1997). The single-source marginal probabilities are assumed to be given by $p_A = 0.7$, $p_B = 0.4$ $p_C = 0.3$ and $p_D = 0.2$. We fix

$$p_{AB} = 0.3 \text{ and } p_{AC} = 0.2,$$

whilst all other marginal probabilities are obtained using the known probability relationships given in Chapter 2, Examples 2.16 and 2.17 for this particular model of

conditional independence. Thus, the form of p_{BC} depends on the marginal probabilities of the conditioning source, A, and is given by:

$$p_{BC} = \frac{p_{AB}p_{AC}}{p_A} + \frac{(p_B - p_{AB})(p_C - p_{AC})}{1 - p_A} = 0.119047.$$

Notice that this probability is slightly lower than that of 0.12 should B and C be marginally independent. The 3 remaining pairs of sources are marginally independent by assumption. Thus

$$p_{AD} = p_A p_D = 0.14 \quad p_{BD} = p_B p_D = 0.08 \quad p_{CD} = p_C p_D = 0.06.$$

The 4 three-way marginal probabilities are obtained using results from Example 2.17 and are given by

$$p_{ABC} = \frac{p_{AB}p_{AC}}{p_A} = 0.0857, \quad p_{ABD} = p_{AB}p_D = 0.06$$
$$p_{ACD} = p_{AC}p_D = 0.04, \quad p_{BCD} = p_{BC}p_D = 0.0238,$$

and the four-source marginal (which is also the four-source joint probability) by

$$p_{ABCD} = p_D p_{ABC} = 0.0171.$$

Note that, had all sources been marginally independent, the three-way marginal probabilities would be given by $p_{ABC} = 0.084$, $p_{ABD} = 0.056$, $p_{ACD} = 0.042$ and $p_{BCD} = 0.024$, and, the four-way marginal (equivalently, the four-way joint probability) by $p_{ABCD} = 0.0168$. For all of these marginal combinations, we observe that the true three-way and four-way marginal probabilities exceed those under assumptions of marginal independence except for the 2 three-way probabilities p_{ACD} and p_{BCD}. The corresponding cell probabilities sum to 0.893. That is, the theoretical proportion of the population of size 1000 that is observed is equal to 89%.

For the assumed conditional independence structure here, the true underlying CIDs, which correspond to the specified probability distribution are given by

$$\gamma_{AB} = 0.06899 \quad \gamma_{AC} = -0.04879 \quad \gamma_{BC} = -0.00797 \quad \gamma_{ABC} = 0.00797,$$

with all others equal to 0. Note that $\gamma_{ABC} = -\gamma_{BC}$ since, by assumption, $B \perp\!\!\!\perp C|A$, that is B and C are independent, conditionally on A.

Table E.1 presents a data set generated according to this assumed underlying dependence structure.

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	16	24	5	10
	D_{No}	61	92	34	47
C_{No}	D_{Yes}	50	72	16	26
	D_{No}	165	224	50	?

Table E.1: Data generated according to the conditional dependence structure given by $[AB][AC][D]$.

Nonhierarchical dependence

We assume that the true underlying model is that of a nonhierarchical dependence structure with the only dependence present between 2 of the three marginal combinations of sources, namely within each of the two sets given by $\{A, B, C\}$ and $\{A, B, D\}$. Note that such a nonhierarchical dependence structure cannot be expressed in the notation of hierarchical models; it is not represented by $[ABC][ABD]$, since, by assumption, no lower order dependence is present.

The single-source marginal probabilities are given by $p_A = 0.5$, $p_B = 0.4$ $p_C = 0.3$ and $p_D = 0.2$. We fix

$$p_{ABC} = 0.07 \text{ and } p_{ABD} = 0.06,$$

whilst all other marginal probabilities are obtained using the known probability relationships given in Chapter 2, Example 2.17 for this particular model of nonhierarchical

dependence.

All 6 pairwise sets of sources are assumed to be marginally independent and are given by

$$p_{AB} = p_A p_B = 0.2, \quad p_{AC} = p_A p_C = 0.15, \quad p_{AD} = p_A p_D = 0.1,$$
$$p_{BC} = p_B p_C = 0.12, \quad p_{BD} = p_B p_D = 0.08, \quad p_{CD} = p_C p_D = 0.06,$$

Likewise, the remaining 2 three-way sets of sources, $\{A, C, D\}$ and $\{B, C, D\}$, are also assumed to be marginally independent. Thus,

$$p_{ACD} = p_A p_C p_D = 0.03 \text{ and } p_{BCD} = p_B p_C p_D = 0.024,$$

whilst the single four-way marginal (equivalently joint) probability is obtained using the following

$$p_{ABCD} = \frac{p_{ABC} p_{ABD}}{p_A p_B} = 0.021.$$

Notice that this probability exceeds that under marginal independence of the four sources, which would be equal to 0.012. Likewise under marginal independence of each of the sets $\{A, B, C\}$ and $\{A, B, D\}$, the marginal probabilities would be equal to 0.06 and 0.04, respectively, rather than the values of 0.07 and 0.06, respectively, at which they have been fixed. Thus, the nonhierarchical dependence structure assumed in this case tends to exhibit positive dependence throughout whenever there is dependence present. The corresponding cell probabilities sum to 0.853. That is, the theoretical proportion of the population of size 1000 is equal to 85%.

For this nonhierarchical dependence structure, the true underlying CIDs, which correspond to the specified probability distribution, are given by

$$\gamma_{ABC} = 0.1515 \quad \gamma_{ABD} = 0.4055,$$

with all others equal to 0.

Table E.2 presents a data set generated according to this assumed underlying dependence structure.

		A_{Yes}		A_{No}	
		B_{Yes}	B_{No}	B_{Yes}	B_{No}
C_{Yes}	D_{Yes}	21	9	6	26
	D_{No}	61	59	60	70
C_{No}	D_{Yes}	35	27	17	60
	D_{No}	82	195	128	?

Table E.2: Data generated according to the non-hierarchical dependence structure with only dependence within the sets $\{A, B, C\}$ and $\{A, B, D\}$

Appendix F

Appendix for Chapter 3

F.1 Score and Information calculations for Poisson likelihood

Using the rules of matrix and vector differentiation stated in Appendix B, the score vector is derived as follows:

$$\begin{aligned}
U(\boldsymbol{\delta}) &= \frac{\partial l}{\partial \boldsymbol{\delta}} \\
&= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\delta}} \frac{\partial l}{\partial \boldsymbol{\mu}} \\
&= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\delta}} \frac{\partial (\mathbf{n}' \log(\boldsymbol{\mu}) - \mathbf{e}'\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \\
&= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\delta}} \left[\text{diag}(\boldsymbol{\mu})^{-1}\mathbf{n} - \mathbf{e}\right] \\
&= \frac{\partial \mathbf{A}^{-1} \exp(\mathbf{Y}\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \left[\text{diag}(\boldsymbol{\mu})^{-1}\mathbf{n} - \mathbf{e}\right] \\
&= \mathbf{Y}'\text{diag}(\exp(\mathbf{Y}\boldsymbol{\delta}))(\mathbf{A}^{-1})' \left[\mathbf{n} \circ \boldsymbol{\mu}(\boldsymbol{\delta})^{-I} - \mathbf{e}\right],
\end{aligned}$$

whilst the negative $q \times q$ information matrix is given by

$$\begin{aligned}
-I(\boldsymbol{\delta}) &= \frac{\partial^2 l}{\partial \boldsymbol{\delta}' \partial \boldsymbol{\delta}} \\
&= \frac{\partial U(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \\
&= \frac{\partial}{\partial \boldsymbol{\delta}} \left(\exp(\mathbf{Y}\boldsymbol{\delta}) \circ [\mathbf{A}^{-1\prime}(\mathbf{n} \circ \boldsymbol{\mu}^{-1} - \mathbf{e})]\right) \mathbf{Y} \\
&= \left[\frac{\partial}{\partial \boldsymbol{\delta}} \left(\exp(\mathbf{Y}\boldsymbol{\delta})\right) \mathbf{D}_{\mathbf{A}^{-1\prime}(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})} \right. \\
&\qquad \left. + \frac{\partial}{\partial \boldsymbol{\delta}} \left[\mathbf{A}^{-1\prime}(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})\right] \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right] \mathbf{Y} \\
&= \left[\mathbf{Y}' \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})} \mathbf{D}_{(\mathbf{A}^{-1})'(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})} \right. \\
&\qquad \left. + \frac{\partial}{\partial \boldsymbol{\delta}} [\mathbf{n} \circ \boldsymbol{\mu}(\boldsymbol{\delta})^{-I}] \mathbf{A}^{-1} \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right] \mathbf{Y} \\
&= \left[\mathbf{Y}' \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})} \mathbf{D}_{(\mathbf{A}^{-1})'(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})} \right. \\
&\qquad \left. + \frac{\partial}{\partial \boldsymbol{\delta}} [\boldsymbol{\mu}(\boldsymbol{\delta})] \frac{\partial}{\partial \boldsymbol{\mu}} [\boldsymbol{\mu}^{-I}] \mathbf{D}_{\mathbf{n}} \mathbf{A}^{-1} \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right] \mathbf{Y} \\
&= \left[\mathbf{Y}' \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})} \mathbf{D}_{(\mathbf{A}^{-1})'(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})} \right. \\
&\qquad \left. - \left(\mathbf{Y}' \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})} (\mathbf{A}^{-1})' \mathbf{D}_{\mathbf{n} \circ \boldsymbol{\mu}^{-2I}}\right) \mathbf{A}^{-1} \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right] \mathbf{Y} \\
&= \mathbf{Y}' \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})} \left(\mathbf{D}_{(\mathbf{A}^{-1})'(\mathbf{n} \circ \boldsymbol{\mu}^{-I} - \mathbf{e})} - (\mathbf{A}^{-1})' \mathbf{D}_{\mathbf{n} \circ \boldsymbol{\mu}^{-2I}} \mathbf{A}^{-1} \mathbf{D}_{\exp(\mathbf{Y}\boldsymbol{\delta})}\right) \mathbf{Y},
\end{aligned}$$

where, for ease of notation, dependence of $\boldsymbol{\mu}$ on $\boldsymbol{\delta}$ has been suppressed and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\delta})$.

F.2 Derivation of the asymptotic covariance matrix for Poisson sampling of Chapter 3

Lang and Agresti (1994) present the general form of models able to simultaneously model both the joint and marginal distributions of multivariate categorical responses. Such a model, similarly to that introduced by Haber (1985b), takes the form:

$$\mathbf{C} \log \mathbf{B}\boldsymbol{\mu} = \mathbf{W}\boldsymbol{\alpha}, \ \text{ident}(\boldsymbol{\mu}) = \mathbf{0} \tag{F.1}$$

where $\mathbf{C} = \mathbf{C}_J \bigoplus \mathbf{C}_M$, $\mathbf{B}' = (\mathbf{B}'_J, \mathbf{B}'_M)$, $\mathbf{W} = \mathbf{W}_J \bigoplus \mathbf{W}_M$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_J, \boldsymbol{\alpha}'_M)$, and $\text{ident}(\boldsymbol{\mu}) = \mathbf{0}$ denotes the multinomial identifiability constraints, and, J refers to the model on the joint means and M to that on the marginal means. Although, the theory developed in both Lang and Agresti (1994) and Haber (1985b) refers to the general case of sampling from S independent multinomial samples, we will consider $S = 1$ since the capture-recapture setting considered in this dissertation concerns a single incomplete contingency table at a time. We let d denote the length of the vector $\boldsymbol{\mu}$ and q the length of the vector $\boldsymbol{\alpha}$.

Lang and Agresti introduce a constraint reparameterization of the freedom equation form of model (F.1). Such a constraint parameterization is given by

$$\mathbf{U}'\mathbf{C} \log \mathbf{B}\boldsymbol{\mu} = \mathbf{0}, \ \text{ident}(\boldsymbol{\mu}) = \mathbf{0}, \tag{F.2}$$

where the space spanned by the columns of $\mathbf{U}$ is the orthogonal complement of the space spanned by the columns of $\mathbf{W}$ and $\mathbf{U}$ is of full column rank. Equivalently $\mathbf{U}'\mathbf{W} = \mathbf{0}$. From Haber (1985b), $\mathbf{U}$ is computed using:

$$\mathbf{U} = (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')\mathbf{V}, \tag{F.3}$$

where $\mathbf{V}$ is a $d \times (d - q)$ matrix of full column rank. The first constraint equation of (F.2) imposes $d - q$ linear constraints on the cell means $\boldsymbol{\mu}$.

The matrix $\mathbf{V}$, and consequently $\mathbf{U}$, is not unique. It must be noted that, although seemingly not clarified by Haber (1985b), $\mathbf{V}$ is not entirely arbitrary. We must ensure that $\mathbf{U}$ spans the same subspace as $\mathbf{W}'$. An approach to obtaining an appropriate matrix $\mathbf{U}$, is based on using the Singular Value Decomposition (SVD) (Searle, 1982) of $\mathbf{Q} = (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')$ as follows: If $\mathbf{Q}_1 D_{\mathbf{Q}} \mathbf{Q}'_2$ is the SVD of $\mathbf{Q}$, where $D_{\mathbf{Q}}$ has zeros everywhere except for in the first p entries, say of the diagonal, which we denote by $D_{\mathbf{Q}_p}$, then

$$\mathbf{U} = \mathbf{Q}_2 \begin{bmatrix} D_{\mathbf{Q}_p} \\ \mathbf{0} \end{bmatrix}.$$

In practice, the inbuilt **svd** function of the *R Programming Language* (2004) can be used to obtain the SVD of $\mathbf{Q}$ and thus, to obtain $\mathbf{U}$.

Maximum likelihood estimation (MLE) consists in maximizing the kernel of the multinomial, equivalently the Poisson, likelihood given by

$$l(\boldsymbol{\mu}; \mathbf{n}) = \mathbf{n}' \log \boldsymbol{\mu}$$

subject to the model parameter space defined by the freedom equations of (F.1) or equivalently the constraint equations of (F.2). As described by Lang and Agresti (1994), it is most useful to work in terms of the constraint equations. Thus, the model parameter space is given by

$$\{\boldsymbol{\mu} : \mathbf{U}'\mathbf{C} \log \mathbf{B}\boldsymbol{\mu} = \mathbf{0},\ \text{ident}(\boldsymbol{\mu})\} = \{\boldsymbol{\mu} : \mathbf{f}(\boldsymbol{\mu}) = \mathbf{0},\ \text{ident}(\boldsymbol{\mu})\}. \tag{F.4}$$

Lang and Agresti (1994) describe modifications to the Newton-Raphson algorithm proposed by Haber (1985b) to solve the Lagrangian (not presented here) which summarizes the function to be maximimized and the constraints given by (F.2). Their approach deals with a matrix that is much easier to invert than that given in Haber (1985b). In order to ensure that the numerical algorithm used to obtain the maximum likelihood estimates of $\boldsymbol{\alpha}$ does not move to $\boldsymbol{\alpha}$ values that correspond to negative values of $\boldsymbol{\mu}$, a reparameterization from $\boldsymbol{\mu}$ to $\boldsymbol{\zeta} = \log \boldsymbol{\mu}$ is adopted. The model parameter space is thus given by the following reparameterization of (F.4)

$$\{\boldsymbol{\zeta} : \mathbf{U}'\mathbf{C} \log \mathbf{B} \exp \boldsymbol{\zeta} = \mathbf{0},\ \text{ident}(\boldsymbol{\zeta})\} = \{\boldsymbol{\zeta} : \mathbf{h}(\boldsymbol{\zeta}) = \mathbf{0},\ \text{ident}(\boldsymbol{\zeta})\}.$$

The MLE of $\boldsymbol{\zeta}$, $\hat{\boldsymbol{\zeta}}$, is obtained by solving for $\hat{\boldsymbol{\theta}}$ in the likelihood equations

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\partial l(\hat{\boldsymbol{\zeta}}; \mathbf{n})}{\partial \boldsymbol{\zeta}} - \exp(\hat{\boldsymbol{\zeta}}) + \frac{\partial \mathbf{h}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} \hat{\boldsymbol{\lambda}} \\ \mathbf{h}(\hat{\boldsymbol{\zeta}}) \end{bmatrix} = \begin{bmatrix} \mathbf{n} - \exp(\hat{\boldsymbol{\zeta}}) + \mathbf{H}(\hat{\boldsymbol{\zeta}})\hat{\boldsymbol{\lambda}} \\ \mathbf{h}(\hat{\boldsymbol{\zeta}}) \end{bmatrix} = \mathbf{0}$$

where $\boldsymbol{\theta} = vec(\boldsymbol{\zeta}, \boldsymbol{\lambda})$ and $\mathbf{H}(\hat{\boldsymbol{\zeta}}) = \partial \mathbf{h}(\boldsymbol{\zeta}) / \partial \boldsymbol{\zeta}$ is the $d \times (d - q)$ matrix of derivatives of the $(d - q) \times 1$ vector $\mathbf{h}(\boldsymbol{\zeta}) = \mathbf{U}'\mathbf{C} \log \mathbf{A} \exp \boldsymbol{\zeta}$ with respect to the $d \times 1$ vector $\boldsymbol{\zeta}$

given by

$$\mathbf{H} = \frac{\partial \mathbf{h}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} = \begin{bmatrix} \frac{\partial h_1(\boldsymbol{\zeta})}{\partial \zeta_1} & \cdots & \frac{\partial h_{d-p}(\boldsymbol{\zeta})}{\partial \zeta_1} \\ \vdots & & \vdots \\ \frac{\partial h_1(\boldsymbol{\zeta})}{\partial \zeta_d} & \cdots & \frac{\partial h_{d-q}(\boldsymbol{\zeta})}{\partial \zeta_d} \end{bmatrix}, \tag{F.5}$$

where $\mathbf{h}(\boldsymbol{\zeta}) = (h_1(\boldsymbol{\zeta}), \ldots, h_{d-p}(\boldsymbol{\zeta}))'$.

Lang and Agresti (1994) provide the asymptotic normal distributions of $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\zeta}}', \hat{\boldsymbol{\lambda}}')'$ under certain nonrestrictive assumptions. Of particular interest to us is the asymptotic covariance matrix of our model parameters $\boldsymbol{\alpha}$ given for the Poisson sampling scheme by

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}\mathbf{D}(\mathbf{B}\boldsymbol{\mu})^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}}\mathbf{B}'\mathbf{D}(\mathbf{B}\boldsymbol{\mu})^{-1}\mathbf{C}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}, \tag{F.6}$$

where

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}} = \mathbf{D}(\hat{\boldsymbol{\zeta}}) - \mathbf{H}(\mathbf{H}'\mathbf{D}(\hat{\boldsymbol{\zeta}})^{-1}\mathbf{H})^{-1}\mathbf{H}', \tag{F.7}$$

and $\mathbf{D}(\boldsymbol{\zeta})$ is the $d \times d$ vector with $\boldsymbol{\zeta}$ on the diagonal.

Next we see how to apply this theory to our marginal log-linear model.

Application of approach to frequentist MLLM of Chapter 3

It is clear that the marginal model introduced in Section 2.4 (see Definition 2.7) fits in the class of models of the form (F.1) with $\mathbf{C}$ equal to the identity matrix and $\mathbf{B}$ equal to the $\mathbf{A}$ matrix of the marginal model so that the $\mathbf{B}_J$ matrix of $\mathbf{B}$ is not included for the marginal model. For the marginal model forms considered in Chapter 3, in which the CIDs are treated as fixed parameters to be estimated, the $\boldsymbol{\alpha}$ vector of (F.1) is precisely that denoted by $\boldsymbol{\delta}$, the vector of length $q \leq d$, where $d = 2^K - 1$ in the K-source capture-recapture setting.

In order to derive the asymptotic covariance matrix of the parameter vector $\boldsymbol{\delta}$ of our model using (F.6), we first derive the form of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}}$ using (F.7). Using the definition of $\mathbf{h}(\boldsymbol{\zeta})$ from (F.2) together with that of $\mathbf{H}(\boldsymbol{\zeta})$ from (F.5) and the chain rule for differentiation of a vector function yields

$$\begin{aligned}
\mathbf{H} &= \frac{\partial(\mathbf{U}' \log \mathbf{B}e^{\boldsymbol{\zeta}})}{\partial \boldsymbol{\zeta}} \\
&= \frac{de^{\boldsymbol{\zeta}}}{d\boldsymbol{\zeta}} \frac{d\mathbf{B}e^{\boldsymbol{\zeta}}}{de^{\boldsymbol{\zeta}}} \frac{d \log \mathbf{B}e^{\boldsymbol{\zeta}}}{d\mathbf{B}e^{\boldsymbol{\zeta}}} \frac{d\mathbf{U}' \log \mathbf{B}e^{\boldsymbol{\zeta}}}{d \log \mathbf{B}e^{\boldsymbol{\zeta}}} \\
&= \mathbf{D}_{\exp(\boldsymbol{\zeta})} \mathbf{B}' \mathbf{D}^{-1}_{\mathbf{B}\exp(\boldsymbol{\zeta})} \mathbf{U} \\
&= \mathbf{D}_{\boldsymbol{\mu}} \mathbf{B}' \mathbf{D}^{-1}_{\mathbf{B}\boldsymbol{\mu}} \mathbf{U},
\end{aligned}$$

where $\mathbf{B} = \mathbf{A}$, for the marginal model and $\mathbf{U}$ is obtained as described above.

Appendix G

R code for Chapter 3

Code for algorithm to fit frequentist MLLM

```
-----------------------------------------------------------------------
FOUR SOURCE FREQUENTIST CODE # THIS CODE FITS ALL 2047 JLL (HIER
AND NON-HIER) TO FOUR SOURCES # AS WELL AS ALL FREQUENTIST MLLM
WITH CIDS TREATED AS FIXED EFFECTS # WITH ALL DIFFERENT
COMBINATIONS SET TO ZERO # # FOR VARIANCE ESTIMATION FROM LANG AND
AGRESTI (JASA 1994) # #
-----------------------------------------------------------------------


K<-4

S4<-matrix(0,nrow=15,ncol=4) S4[,1] <- c(1, 0, 0, 0, 1, 1, 1, 0,
0, 0, 1, 1, 1, 0, 1) S4[,2] <- c(0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1,
1, 0, 1, 1) S4[,3] <- c(0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1,
1) S4[,4] <- c(0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1)


Mmat<-t(S4) Mmat<-rbind(Mmat,S4[,1]+S4[,2]==2)
Mmat<-rbind(Mmat,S4[,1]+S4[,3]==2)
Mmat<-rbind(Mmat,S4[,1]+S4[,4]==2)
Mmat<-rbind(Mmat,S4[,2]+S4[,3]==2)
Mmat<-rbind(Mmat,S4[,2]+S4[,4]==2)
Mmat<-rbind(Mmat,S4[,3]+S4[,4]==2)
Mmat<-rbind(Mmat,S4[,1]+S4[,2]+S4[,3]==3)
```

```
Mmat<-rbind(Mmat,S4[,1]+S4[,2]+S4[,4]==3)
Mmat<-rbind(Mmat,S4[,1]+S4[,3]+S4[,4]==3)
Mmat<-rbind(Mmat,S4[,2]+S4[,3]+S4[,4]==3)
Mmat<-rbind(Mmat,S4[,1]+S4[,2]+S4[,3]+S4[,4]==4)

A4<-Mmat
A4<-t(A4)


dimnames(A4)<-list(
c("Abcd","aBcd","abCd","abcD","ABcd","AbCd","AbcD","aBCd","aBcD","abCD",
"ABCd","ABcD","AbCD","aBCD","ABCD"),
c("A","B","C","D","AB","AC","AD","BC","BD","CD","ABC","ABD","ACD","BCD","ABCD"))
# X matrices will be extracted from this one
bigm<-cbind(1-apply(A4[,1:4],1,sum),A4)

A4<-t(A4)

# This will be used for column extraction # from bigm

# Note that there are 2047 models (hierarcal and non-hierarchical)
for four sources with the four main effects included in all #
Calculate using a<-choose(11,1); or(i in 2:11){a <- a +
choose(11,i)}; a

models<-matrix(0,nrow=2047,ncol=11) for (i in 1:2047) {
 modn<-i-1
 for (j in 1:11){
    if (modn\%\%2) models[i,j]<-1 # i.e. modn\%\%2 = modn mod 2
    modn<-modn\%/\%2}}# integer division

models<-cbind(1,1,1,1,1,models)

models<-models==1 # changes to TRUE/FALSE


satform4<-(apply(models,1,sum)==15)*1 # indicator of whether model
is one of 11 saturdated models


#----- indicator for hierarchical models

hierform4<-rep(0,2047) hierform4[1:64]<-1
```

```
ind.hier<-c(76,80,92,96,108,112,124,128,
150,152,158,160,182,184,190,192, 295,296,303,304,311,312,319,320,
569,570,571,572,573,574,575,576, 224,256, 368,384, 636,640,
440,448, 702,704, 831,832, 512,768,896,960,1024)
hierform4[ind.hier]<-1


basmod<-"tot~A+B+C+D"
tokens<-c("AB","AC","AD","BC","BD","CD","ABC","ABD","ACD","BCD","ABCD")
formulas<-list() for (i in 1:dim(models)[1])
formulas<-c(formulas,as.formula(paste(c(basmod,
          tokens[models[i,-(1:5)]]),collapse="+")))

# The usual suspects
ll<-function(n,beta,X,A) {
mu<-as.vector(solve(A,exp(X%*%beta)))
sum(n*log(mu)-mu)}

score<-function(n,beta,X,A) {
eta<-as.vector(exp(X%*%beta))
mu<-as.vector(solve(A,eta))
t(X)\%*\%diag(eta)\%*\%solve(t(A),n/mu-1)}

finf<-function(n,beta,X,A) {
eta<-as.vector(exp(X\%*\%beta))
mu<-as.vector(solve(A,eta))
temp<-solve(A,diag(eta)\%*\%X)
t(temp)\%*\%diag(1/mu)\%*\%temp}



# This is really Fisher scoring mixed with steepest ascent,
# all dampened & with a few checks on feasibility

mle.old<-function(n,betastart,X,A,tol=1e-7,maxiter=100,maxhi=5,fac=10,ns=4)
{ diff<-1 oldbeta<-betastart iter<-0 whup<-F hyperiter<-0 while
(diff>tol && hyperiter<maxhi) {
    while (diff>tol && iter<maxiter) {
      dir1<-solve(finf(n,oldbeta,X,A),score(n,oldbeta,X,A))
      top1<-min(-(oldbeta[1:(ns+1)]/dir1[1:(ns+1)])[dir1[1:(ns+1)]<0],1)
      if (dir1[1]<0)
        top1<-min((oldbeta[2:(ns+1)][dir1[2:(ns+1)]>0]-oldbeta[1])/
                      (dir1[1]-dir1[2:(ns+1)][dir1[2:(ns+1)]>0]),top1)
```

```
      else
        top1<-1
      dir2<-score(n,oldbeta,X,A)
      top2<-min(-(oldbeta[1:(ns+1)]/dir2[1:(ns+1)])[dir2[1:(ns+1)]<0],1)
      if (dir2[1]<0)
        top2<-min((oldbeta[2:(ns+1)][dir2[2:(ns+1)]>0]-oldbeta[1])/
                (dir2[1]-dir2[2:(ns+1)][dir2[2:(ns+1)]>0]),top2)
      else
        top2<-1
      foo.fun<-function(lambda,oldbeta,dir,n,X,A)
        -ll(n,oldbeta+lambda*dir,X,A)
      opt1<-optimize(f=foo.fun,interval=c(0,top1),oldbeta=oldbeta,dir=dir1,n=n,X=X,A=
      opt2<-optimize(f=foo.fun,interval=c(0,top2),oldbeta=oldbeta,dir=dir2,n=n,X=X,A=
      if (is.na(opt1$objective)) opt1$objective<-Inf
      if (is.na(opt2$objective)) opt2$objective<-Inf
      if (whup) {
        opt1$minimum<-min(fac*opt1$minimum,1)
        opt2$minimum<-min(fac*opt2$minimum,1)
        whup<-F}
      if (opt1$objective<opt2$objective)  #objective function is -loglhd
          beta<-oldbeta+opt1$minimum*dir1
      else
          beta<-oldbeta+opt2$minimum*dir2
      diff1<-beta-oldbeta
      diff<-max(abs(diff1))
      oldbeta<-beta
      iter<-iter+1}
    if (diff>tol) whup<-T
    hyperiter<-hyperiter+1}
return(list(beta,ll(n,beta,X,A),diff1,score(n,beta,X,A),iter))}

# Here the parameters of beta are (beta0,betastar,CIDs)

mle<-function(n,betastart,X,A,tol=1e-8,maxiter=100,ns=4) { diff<-1
oldbeta<-betastart iter<-0
    while (diff>tol && iter<maxiter) {
      dir<-solve(finf(n,oldbeta,X,A),score(n,oldbeta,X,A))
      top<-min(-(oldbeta[1:(ns+1)]/dir[1:(ns+1)])[dir[1:(ns+1)]<0],1)
      if (dir[1]<0)
        top<-min((oldbeta[2:(ns+1)][dir[2:(ns+1)]>0]-oldbeta[1])/
                        (dir[1]-dir[2:(ns+1)][dir[2:(ns+1)]>0]),top)
      else
        top<-1
```

```
      foo.fun<-function(lambda,oldbeta,dir,n,X,A)
        -ll(n,oldbeta+lambda*dir,X,A)
      opt<-optimize(f=foo.fun,interval=c(0,top),
                    oldbeta=oldbeta,dir=dir,n=n,X=X,A=A)
      if (is.na(opt$objective)) opt$objective<-Inf
          beta<-oldbeta+opt$minimum*dir
      diff1<-beta-oldbeta
      diff<-max(abs(diff1))
      oldbeta<-beta
      iter<-iter+1}
return(list(beta,ll(n,beta,X,A),diff1,score(n,beta,X,A),iter))}

fitted.marg<-function(beta,X,A) {
    mtot<-exp(X%*%beta)
    solve(A,mtot)}



# asymptotic covariance matrix of model parameters for a single
model
# U is any d by (d-p) matrix (where p is number of
parameters in model) such that U'X = 0

# Haber (1985) states that U can be calculated using U = (I -
X'(X'X)^(-1)X)W,
# where W is a d by (d-p) matrix with independent
columns

asymp.cov.mle.fun<-function(beta.MLE,mu.MLE,X,A,U) {
   p<-length(beta.MLE)
   eta <-log(mu.MLE)
   marg.MLE <- A%*%mu.MLE
   diag.eta <-diag(as.vector(eta))
   diag.inv.eta <- diag(as.vector(1/eta))
   diag.inv.Amu <-diag(as.vector(1/(marg.MLE)))
   H <- diag(as.vector(mu.MLE))%*%t(A)%*%diag(as.vector(1/marg.MLE))%*%U
   Sigma.mu.MLE <- diag.eta - H%*%solve(t(H)%*%diag.inv.eta%*%H)%*%t(H)
   temp <- solve(t(X)%*%X)%*%t(X)%*%diag.inv.Amu%*%A
   Sigma.alpha.MLE <- temp%*%Sigma.mu.MLE%*%t(temp)
}

# to extract diagonal just need to do diag(matrix.name)
```

Appendix H

C++ code for MCMC scheme of Chapter 5

```
/*
Code to run MCMC simulation of CID formulation of Bayesian
random effects model for K sources using data from incomplete
contingency table at fixed level of RE variance with no centering
of RE and ability to set any RE equal to 0

Procedure:
   1. Metropolis-Hastings update for N
   2. Metropolis-Hastings update for all betastar
   3. Metropolis-Hastings update for all pairwise RE centered on 0
   4. Metropolis-Hastings update for all threeway RE centered on 0
   5. Etc. for RE

In order to set a specific RE equal to 0 the tool we use is to fix
the prior RE variance for that component equal to 0

*/

using namespace std; #include <iostream> #include <fstream> //
Needed for file input/output #include <math.h>

/* Library containing random/statistical functions */ #include
<gsl/gsl_rng.h> #include <gsl/gsl_randist.h> #include
<gsl/gsl_statistics_double.h> #include <gsl/gsl_vector.h> #include
<gsl/gsl_sf_gamma.h> // Needed for gamma function #include
<gsl/gsl_matrix.h> #include <gsl/gsl_blas.h> #include
<gsl/gsl_statistics_double.h> #include <gsl/gsl_sf_gamma.h>
```

```
double multinom_loglike_fun(const double totalN, gsl_vector*
cell_num_vec, const double nobs,
    gsl_vector* cell_mean_vec){

      double n_unobs = totalN - nobs;
      int j;
      double loglike=0,  sumprobs=0;

      //cout<<"$$$$$$"<<endl;
      loglike+=gsl_sf_lnfact((unsigned int)totalN);
      //cout<<loglike<<endl;
      loglike -= gsl_sf_lnfact((unsigned int) n_unobs);
      //cout<<loglike<<endl;
      for(j=0;j<cell_num_vec->size;j++){  //Trick to avoid passing length of vector
         loglike+=cell_num_vec->data[j]*log(cell_mean_vec->data[j]/totalN) -
                                 gsl_sf_lnfact((unsigned int) cell_num_vec->data[j]);
         sumprobs+=cell_mean_vec->data[j]/totalN;
         //cout<<loglike<<endl;
      }
      loglike += n_unobs * log(1 - sumprobs);
      //cout<<loglike<<endl;
      //cout<<"$$$$$$"<<endl;
      return(loglike);
}


int main(int argc, char* argv[]) {

  /*Creates file streams for file output */


  ifstream the_data_file, the_param_file;
  ofstream the_output_file;

  /*Checks to see that enough arguments have been given*/

  if(argc < 4){
    cout<<"Not enough arguments"<<endl;
    return(0);
  }

  /* Opens file for writing */
```

```
/*Assumes first argument is data file*/

the_data_file.open(argv[1], ios::in);
if(!the_data_file){
  cout<<"Parameter file not found"<<endl;
}

/*Assumes second argument is the parameter file*/

the_param_file.open(argv[2], ios::in);
if(!the_param_file){
  cout<<"Parameter file not found"<<endl;
  }


/*Assumes third argument is output file*/

the_output_file.open(argv[3], ios::out);
if(!the_output_file){
  cout<<"File a not found"<<endl;
}

/*Set up for RNG*/

 gsl_rng *r;
 const gsl_rng_type * T;

 gsl_rng_env_setup();

 T = gsl_rng_default;
 r = gsl_rng_alloc (T);



/*Assumes that fourth argument is the seed*/

long int seed;

seed = atoi(argv[4]);
gsl_rng_set(r,seed);

 /*Maintenanace vars*/
```

```
 int ctr;
 int feasible_flag, pwise_loop, levelwise_loop, pwise_ctr, pwise_ctr2;
 double current_like=0, prop_like=0;
 double current_prior=0, prop_prior=0;

/*Assumes that the number of iterations is first*/

int number_of_samples;
the_param_file >> number_of_samples;

/*Assumes that the number of data points is second in
  parameter file i.e. K*/
int numsources;
the_param_file >> numsources;

/*Assumes that the number by which to thin is third in
  parameter file */

int thin;
the_param_file >> thin;

 cout<<number_of_samples<<" "<<numsources<<endl;

 int nobs=0;
 int n_rand_effect=-numsources;
 // counts how many of the RE, i.e. the two or more source gammas are to be estimat
 // i.e. how many ahve non-zero prior variance in param_file
 // need to initialise at -K since in loop where it is updated below
 // the loop includes the betastar terms and there are K of them
 int num_observed_counts = (int)pow(2,(double)numsources) - 1;

 int data_loop;
 gsl_vector* observed_counts;
 observed_counts = gsl_vector_alloc(num_observed_counts);

 /* Read in cell entries of incomplete contingency table from data file
  ordered according to order laid out by Ainv matrix*/

 for(data_loop=0;data_loop<num_observed_counts;data_loop++){
     the_data_file >> observed_counts->data[data_loop];
     nobs += (int) observed_counts->data[data_loop];
     cout<<gsl_vector_get(observed_counts,data_loop)<<" ";
```

```
}
cout<<endl; // to return to a new line

/*Declare beta0*/

double beta0, beta0_prop;
double N;
int beta_loop;

/* 4th entry in param file is starting value for N  */
the_param_file >> N;
beta0 = log(N);

/*Declare gammas*/
gsl_vector *gammas, *gammas_prop;
int gammas_loop;
gammas = gsl_vector_calloc(num_observed_counts); // Sets equal to zero
gammas_prop = gsl_vector_calloc(num_observed_counts); // Sets equal to zero
/* Put in loop to read gammas here from param file
   5th entry is beta vec
   6th entry is RE vector
  Note that the K beta terms and d-K RE terms are all
  called gammas here in this code for convenience.
  They will all be generated similarly centered on previous value
  Under independence all REs are zero */
cout<<"***"<<endl;
for(gammas_loop = 0; gammas_loop < num_observed_counts; gammas_loop++){
      the_param_file >> gammas->data[gammas_loop];
      gammas_prop->data[gammas_loop] = gammas->data[gammas_loop];
      cout<<gammas->data[gammas_loop]<<" ";
}
cout<<endl;


/*Declare acceptance indicators*/
int accept_N;
gsl_vector *accept_gammas;
accept_gammas = gsl_vector_calloc(numsources);
/* this is of length equal to number of sources, i.e. K, since we accept/reject
  at each level so have acceptance indicator at each level of source
  combinations, i.e. single, pair, threeway etc. */

/*Declare large vector multiplication*/
```

```
gsl_vector *current_logmargmeanvec, *prop_logmargmeanvec;
gsl_vector *current_margmeanvec, *prop_margmeanvec;
gsl_vector *current_cellmeanvec, *prop_cellmeanvec, *output_cellmeanvec;

current_logmargmeanvec=gsl_vector_calloc(num_observed_counts);
prop_logmargmeanvec=gsl_vector_calloc(num_observed_counts);
current_margmeanvec=gsl_vector_calloc(num_observed_counts);
prop_margmeanvec=gsl_vector_calloc(num_observed_counts);
current_cellmeanvec=gsl_vector_calloc(num_observed_counts);
prop_cellmeanvec=gsl_vector_calloc(num_observed_counts);
output_cellmeanvec=gsl_vector_calloc(num_observed_counts);


/*Set up Ainv matrix
   Read in 7th entry in param file as Ainv matrix*/

gsl_matrix *Ainv;
Ainv = gsl_matrix_alloc(num_observed_counts,num_observed_counts);
int matrowloop,matcolloop;
double tempval;
for(matrowloop=0;matrowloop<num_observed_counts;matrowloop++){
     for(matcolloop=0;matcolloop<num_observed_counts;matcolloop++){
         the_param_file >> tempval;
         gsl_matrix_set(Ainv,matrowloop,matcolloop,tempval);
     }
 }

/* Read in 8th entry
  Set up X matrix
  the design matrix for all gammas,
  i.e. betastar (not including beta0) and RE (CIDs) centered on 0*/

gsl_matrix* X;
X = gsl_matrix_calloc(num_observed_counts, num_observed_counts);
for(matrowloop=0;matrowloop<num_observed_counts;matrowloop++){
     for(matcolloop=0;matcolloop<num_observed_counts;matcolloop++){
         the_param_file >> tempval;
         gsl_matrix_set(X,matrowloop,matcolloop,tempval);
         cout<<gsl_matrix_get(X, matrowloop, matcolloop)<<" "; // prints out X mat
     }
     cout<<endl;
}
```

```
/*Setting vector for beta0 constants i.e. vector of
values which mulitply beta0 in model */
  gsl_vector *beta0constvec;
  beta0constvec = gsl_vector_alloc(num_observed_counts);

  pwise_ctr=0;

  for(pwise_loop=1;pwise_loop<=numsources;pwise_loop++){
    for(levelwise_loop = pwise_ctr;
         levelwise_loop < pwise_ctr+(int) gsl_sf_choose(numsources, pwise_loop);
         levelwise_loop++){
            beta0constvec->data[levelwise_loop] = 1 - pwise_loop;
            cout<<levelwise_loop<<" "<<1-pwise_loop<<endl;
            // prints out level and corresponding factor to multiply beta0
     }
    pwise_ctr = pwise_ctr+(int) gsl_sf_choose(numsources, pwise_loop);
 }


/* Calculate marginal mean and cell vectors*/



/* Calculate current log marg mean*/

// First multiply X by vector of gammas
gsl_blas_dgemv(CblasNoTrans, 1.0, X, gammas, 0.0, current_logmargmeanvec);
// Second add in the correct number of beta0 terms to each entry in vector
gsl_blas_daxpy(beta0,beta0constvec,current_logmargmeanvec);

/* Turn into current marginal mean's*/
// Note on dereferencing vectors: (*betas).data is equivalent to betas->data

for(pwise_ctr = 0; pwise_ctr < num_observed_counts; pwise_ctr++){
    current_margmeanvec->data[pwise_ctr] =
       exp(current_logmargmeanvec->data[pwise_ctr]);
}

/* Turn into Cell means by premultiplying by Ainv*/

gsl_blas_dgemv(CblasNoTrans,1.0,Ainv, current_margmeanvec, 0.0,
    current_cellmeanvec);
```

```
/* Turn into Cell means by premultiplying by Ainv*/

gsl_blas_dgemv(CblasNoTrans,1.0,Ainv, current_margmeanvec, 0.0,
    output_cellmeanvec);

// Evaluate likelihood from starting values
current_like = multinom_loglike_fun(N, observed_counts, nobs, current_cellmeanvec)

/*Read in prior values*/
double priormu_beta0, priorsd_beta0;
gsl_vector *priormu_gamma, *priorsd_gamma, *indicator_sd_gamma;
priormu_gamma = gsl_vector_alloc(num_observed_counts);
priorsd_gamma = gsl_vector_alloc(num_observed_counts);
indicator_sd_gamma = gsl_vector_alloc(num_observed_counts);
/*a vector of indicators of whether the corresponding gamma
should be estimated or left set to 0
clearly the first K entries which correspond to the
single sources betastar vector should be estimated
and these entries should be equal to 1 */

// 9th and 10th entry in param file
the_param_file >> priormu_beta0;
the_param_file >> priorsd_beta0;
/* 11th series of entries are prior mean and sd first for betastar vector
and then in the 12th series for each of the d-K gamma RE (usually centered on 0) *
for(gammas_loop=0;gammas_loop<num_observed_counts;gammas_loop++){
   the_param_file >> priormu_gamma->data[gammas_loop];
   the_param_file >> priorsd_gamma->data[gammas_loop];
   n_rand_effect = n_rand_effect + (priorsd_gamma->data[gammas_loop] > 0);
   indicator_sd_gamma->data[gammas_loop] = (priorsd_gamma->data[gammas_loop] > 0);
   // set equal to 1 if gamma param to be estimated or to 0 if not
}

cout << " Number random effects    "  << n_rand_effect    << endl;

 /*Prior calculations*/


int priorloop;

current_prior = log(gsl_ran_gaussian_pdf(beta0 - priormu_beta0, priorsd_beta0));
```

```
/*Set up current stuff*/
/*Set proposal size for N and beta*/

double N_jump, N_prop;
gsl_vector *gammas_jump;
gammas_jump = gsl_vector_alloc(numsources);
//double beta_jump;
// 13th and 14th entries  in param file
the_param_file >> N_jump;
cout << "N jump:"<< N_jump << endl;
//the_param_file >> beta_jump;
for(gammas_loop=0;gammas_loop<numsources;gammas_loop++){
   the_param_file >> gammas_jump->data[gammas_loop];
   cout<<"Gammas jump:"<< gammas_jump->data[gammas_loop]  <<endl;
 }


/* Set random effects variance to initial value.
15th entry in file. 16th and 17th entry
are prior scale s20 and prior nu0 of RE variance distn*/
double sd_re;
double nu0, nu , df, scale;
// set degrees of freedom parameter of prior & posterior dist of random
// effects variance.
double s20, s2;
// set scale parameter of prior &
// posterior dist of random effects variance
// start only from numsources since don't want
//  to look at variance of betastar
// terms which are the first K gamma terms in our gamma vector
/*for(gammas_loop=numsources;gammas_loop<num_observed_counts;gammas_loop++){
   if(priorsd_gamma->data[gammas_loop]>0)
         var_re = priorsd_gamma->data[gammas_loop];
}*/
// actually this loop goes through all RE variances and so ends up
// starting from the last values which is nonzero.
// All RE variance which are non-zero should be set to same value
// as each other since that is what would be required for fixed variance case.

the_param_file >> sd_re; //initial value
the_param_file >> s20;
the_param_file >> nu0;
```

```
df = nu0 + (double)n_rand_effect; // known from calculation of Gibbs update

// now reset all RE sd equal to initial value of RE sd if
// they are supposed to be estimated as given by indicator_sd_gamma vector
for(gammas_loop=numsources;gammas_loop<num_observed_counts;gammas_loop++){
   priorsd_gamma->data[gammas_loop] = indicator_sd_gamma->data[gammas_loop]*.sd_re
// this is under the assumption that all RE have same variance
 cout << " Initial RE sd  " <<  priorsd_gamma->data[gammas_loop] << endl;
}


/*Main loop*/

double accept_ratio;
int level_flag=0;
for(ctr=0; ctr<number_of_samples; ctr++){
   // reset acceptance indicators accept_N and accept_gammas
   accept_N = 0;
   for(pwise_ctr2 = 0; pwise_ctr2 < numsources; pwise_ctr2++){
       accept_gammas->data[pwise_ctr2] = 0;
    }


   /*Betas together */
   // use this next trick to correctly have probability of going +-N_jump,
   // need to add 1 to upper bound so that taking floor will get to +N_jump
   N_prop = floor(gsl_ran_flat(r, -N_jump, N_jump+1) + N);
   beta0_prop = log(N_prop);

   /* Calculate prop log marg mean*/

    gsl_blas_dgemv(CblasNoTrans, 1.0, X, gammas_prop, 0.0, prop_logmargmeanvec);
    // prop_logmargmeanvec = X%*%gammas_prop
    gsl_blas_daxpy(beta0_prop,beta0constvec,prop_logmargmeanvec);
    //  prop_logmargmeanvec= beta0_prop*beta0constvec + prop_logmargmeanvec

   /* Turn into prop marginal means*/
   for(pwise_ctr2 = 0; pwise_ctr2 < num_observed_counts; pwise_ctr2++){
      prop_margmeanvec->data[pwise_ctr2] = exp(prop_logmargmeanvec->data[pwise_ctr
   }

   /*Turn into prop cell means*/
```

```
gsl_blas_dgemv(CblasNoTrans,1.0,Ainv, prop_margmeanvec, 0.0,
                prop_cellmeanvec);
// prop_cellmeanvec =  Ainv\%*\%prop_margmeanvec

feasible_flag = (gsl_vector_min(prop_cellmeanvec) > 0) &&
      (N_prop > max((double)nobs,
gsl_stats_mean(prop_cellmeanvec->data, 1,
       num_observed_counts)*(double)(num_observed_counts)));

if(feasible_flag){

 /*Accept/reject*/

prop_like = multinom_loglike_fun(N_prop,observed_counts,
      nobs, prop_cellmeanvec);

/*Calculate prior distribution for beta0 */
current_prior = log(gsl_ran_gaussian_pdf(beta0 -
       priormu_beta0, priorsd_beta0));
prop_prior = log(gsl_ran_gaussian_pdf(beta0_prop -
       priormu_beta0, priorsd_beta0));


/*Calculate accept/reject ratio for beta0*/

accept_ratio = prop_like + prop_prior - (current_like + current_prior);

if(gsl_ran_flat(r,0.0,1.0)<exp(accept_ratio)){
      current_like = prop_like;
      beta0 = beta0_prop;
      N = N_prop;
      accept_N = 1;
      //cout<<"###  "<<accept_N<<endl;
 }
} // end of feasible_flag check on N_prop


/*Update for the proposed beta/eta vector */

pwise_ctr=0;
gsl_vector_memcpy(gammas_prop,gammas);
for(pwise_loop=1;pwise_loop<=numsources;pwise_loop++){
 level_flag=0;
```

```
  for(levelwise_loop = pwise_ctr;
      levelwise_loop < pwise_ctr+(int) gsl_sf_choose(numsources, pwise_loop);
      levelwise_loop++){

      if(priorsd_gamma->data[levelwise_loop] >0){
       // Only update gammas_prop different to 0 if
       // the prior variance for that component > 0.
       // In this way do not update Re that we want to set equal to 0

       /* Note on use of gsl_vector_set(gsl_vector * v, size_t i, double x)
        This function sets the value of the i-th element of a vector v to x.
        If i lies outside the allowed range of 0 to n-1
        then the error handler is invoked */

      gsl_vector_set(gammas_prop,levelwise_loop,
      gammas->data[levelwise_loop]+gsl_ran_gaussian(r,
           gammas_jump->data[pwise_loop-1]));
      level_flag=1;
       // this indicates that we've reached the end of the series of marginal ga
       // corresponding to the same number of sources
       // e.g. single source, then pairs etc.
       }
 } // end of 'for' loop over levelwise loop

 if(level_flag){
 // START level_flag: so for the same level of margins,
 // i.e. single then pairs then triples etc.
 feasible_flag =0;
 // reset feasible flag to 0 so that can make
 // a check at each level of sources for gammas

 /* Calculate prop log marg mean*/

gsl_blas_dgemv(CblasNoTrans, 1.0, X, gammas_prop, 0.0, prop_logmargmeanvec);
gsl_blas_daxpy(beta0,beta0constvec,prop_logmargmeanvec);

/* Turn into prop marginal means by just updating those margins
 which are being dealt with,
 i.e. single source, then pairs then triples etc. All other entries remain the sa
for(pwise_ctr2 = 0; pwise_ctr2 < num_observed_counts; pwise_ctr2++){
prop_margmeanvec->data[pwise_ctr2] = exp(prop_logmargmeanvec->data[pwise_ctr2]);
}
```

```
/*Turn into prop cell means*/
gsl_blas_dgemv(CblasNoTrans,1.0,Ainv, prop_margmeanvec, 0.0,prop_cellmeanvec);

feasible_flag = (gsl_vector_min(prop_cellmeanvec) > 0) && (N > max((double)nobs,
gsl_stats_mean(prop_cellmeanvec->data, 1, num_observed_counts)*
     (double)(num_observed_counts)));

if(feasible_flag){

    /*Accept/reject by level of number of sources in margin
    i.e. single source then pairs, triples, etc.*/

  prop_like = multinom_loglike_fun(N, observed_counts, nobs, prop_cellmeanvec);

  /*Calculate prior distribution for gammas*/
  current_prior=0;
  prop_prior=0;
  for(levelwise_loop = pwise_ctr;
   levelwise_loop < pwise_ctr+(int) gsl_sf_choose(numsources, pwise_loop);
        levelwise_loop++){
  if(priorsd_gamma->data[levelwise_loop] >0){
     current_prior += log(gsl_ran_gaussian_pdf(gammas->data[levelwise_loop] -
     priormu_gamma->data[levelwise_loop], priorsd_gamma->data[levelwise_loop]));
     prop_prior += log(gsl_ran_gaussian_pdf(gammas_prop->data[levelwise_loop] -
     priormu_gamma->data[levelwise_loop], priorsd_gamma->data[levelwise_loop]));
    } // end of 'if' priorsd_gamma->data[levelwise_loop] >0
 }


  /*Calculate accept/reject ratio*/

 accept_ratio = prop_like + prop_prior - (current_like + current_prior);

 if(gsl_ran_flat(r,0.0,1.0)<exp(accept_ratio)){

  gsl_blas_dgemv(CblasNoTrans,1.0,Ainv, prop_margmeanvec, 0.0,
     output_cellmeanvec);

 current_like = prop_like;
 accept_gammas->data[pwise_loop-1] = 1;
  for(levelwise_loop = pwise_ctr;
      levelwise_loop < pwise_ctr+(int) gsl_sf_choose(numsources, pwise_loop);
```

```
      levelwise_loop++){
         gammas->data[levelwise_loop] = gammas_prop->data[levelwise_loop];
      }
   } // end 'if' accept_ratio
 } // END feasible_flag
} // END level_flag:

pwise_ctr = pwise_ctr+ (int) gsl_sf_choose(numsources,pwise_loop);
// To shift down gamma vector by the number of at the specific level
// of margin, i.e. single then pair, etc.
} // END for(pwise_loop=1;pwise_loop<=numsources;pwise_loop++){



/* Gibbs step for RE variance */
nu = 0.0;
for(gammas_loop = numsources; gammas_loop < num_observed_counts; gammas_loop++){
// add up over RE not including betastar
   nu =  nu +  gammas->data[gammas_loop]*gammas->data[gammas_loop];
}
scale = (nu0*s20 + nu)/df;
sd_re = sqrt((df*scale)/gsl_ran_chisq(r,df));


// now update sd of each RE according to whether or not it should be estimated

for(gammas_loop=numsources;gammas_loop<num_observed_counts;gammas_loop++){
  priorsd_gamma->data[gammas_loop] = indicator_sd_gamma->data[gammas_loop]* sd_re;
   // this is under the assumption that all RE have same variance
  }


 // THINNING
 if( (ctr % thin)==0){
    the_output_file<< beta0<<" ";
    for(pwise_ctr2=0;pwise_ctr2<num_observed_counts;pwise_ctr2++){
        the_output_file << gammas->data[pwise_ctr2]<<" ";
    }

    the_output_file << accept_N<<" ";
    for(pwise_ctr2=0;pwise_ctr2<numsources;pwise_ctr2++){
        the_output_file << accept_gammas->data[pwise_ctr2]<<" ";
    }
```

```
        the_output_file<< sd_re<<" ";

        for(pwise_ctr2=0;pwise_ctr2<num_observed_counts;pwise_ctr2++){
            the_output_file << output_cellmeanvec->data[pwise_ctr2]<<" ";
        }



        the_output_file << endl;
       }

    }


    system("PAUSE");
   /*Allocate space for the data*/

    gsl_vector_free(prop_margmeanvec);
    gsl_vector_free(current_margmeanvec);
    gsl_vector_free(prop_cellmeanvec);
    gsl_vector_free(current_cellmeanvec);
    gsl_vector_free(output_cellmeanvec);
    gsl_vector_free(gammas);
    gsl_vector_free(prop_logmargmeanvec);
    gsl_vector_free(current_logmargmeanvec);
    gsl_vector_free(observed_counts);
    gsl_vector_free(beta0constvec);
    gsl_vector_free(priormu_gamma);
    gsl_vector_free(priorsd_gamma);
    gsl_rng_free(r);
    the_output_file.close();
    the_param_file.close();
    the_data_file.close();
    return(0);

}
```

References

Aaron, D., Chang, Y., Markovic, N., & LaPorte, R. (2003). Estimating the lesbian population: a capture-recapture approach. *Journal of Epidemiology and Community Health*, *57*(3), 207–209.

Abeni, D., Brancato, G., & Perucci, C. (1994). Capture-recapture to estimate the size of the population with human immunodeficiency virus type I infection. *Epidemiology*, *5*, 410–414.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* Wiley.

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, *50*(2), 494–500.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Wiley.

Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse.* Academic Press.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, *46*(3), 623–635.

Anderson, M., & Fienberg, S. (2002). Why is there still a controversy about adjusting the census for undercount? *Political Science and Politics*, *35*(1), 83–85.

Aslan, D., Ozcebe, H., Bertan, M., & Karaagaoglu, E. (2004). Capture-recapture methods for estimation of fertility and mortality in a rural district of Turkey. *Eastern Mediterranean Health Journal*, *10*(1-2), 56–63.

Baker, S. (1994). The multinomial-Poisson transformation. *The Statistician*, *43*, 495–504.

Ball, P., & Asher, J. (2002). Statistics and Slobodan: Using data analysis and statistics in the war crimes trial of former President Milosevic. *Chance*, *15*, 17–24.

Bartolucci, F., & Forcina, A. (2001). Analysis of capture-recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality.

Biometrics, 57, 714–719.

Bartolucci, F., & Forcina, A. (2002). Extended rc association models allowing for order restrictions and marginal modeling. *Journal of the American Statistical Asociation, 97*, 1192–1199.

Bartolucci, F., & Forcina, A. (2006). A class of latent marginal models for capture-recapture data with continuous covariates. *Journal of the American Statistical Association, 101*, 786–794.

Basu, S., & Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika, 88*, 269–279.

Berger, J., Liseo, B., & Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science, 14*, 1–28.

Berger, J. O., & Wolpert, R. (1984). *The Likelihood Principle.* Institute of Mathematical Statistics.

Bergsma, W., & Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics, 30*, 140–159.

Bertsekas, D. (1995). *Nonlinear Programming.* Athena Scientific.

Bishop, Y., Fienberg, S., & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge Massachusetts, MIT.

Borchers, D., Buckland, S., & Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations* (First ed.). London: Springer.

Borg, J., & Greenacre, M. (1998). *Visualization of Categorical Data.* Academic Press.

Brenner, H. (1994). Application of capture-recapture methods for disease monitoring - potential effects of imperfect record linkage. *Methods of Information in Medicine, 33*(5), 502–506.

Brenner, H. (1996). Effects of misdiagnoses on disease monitoring with capture-recapture methods. *Journal of Clinical Epidemiology, 49*(11), 1303–1307.

Brooks, S. (1998). Markov chain Monte Carlo methods and its application. *The Statistician, 47*, 69–100.

Brooks, S., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Brooks, S., & Morgan, J. (1995). Optimisation using simulated annealing. *The Statistician, 44*, 241–257.

Brown, J., Diamond, I., Chambers, R., Buckner, L., & Teague, A. (1999). A methdological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society, Series A, 162*(2), 247–267.

Bruno, G., Biggeri, A., LaPorte, R. E., McCarty, D., Merletti, F., & Pagano, G. (1994). Application of capture-recapture to count diabetes? *Diabetes Care, 17*, 548–556.

Buckland, S. (1984). Monte Carlo confidence intervals. *Biometrics, 40*, 811–817.

Buckland, S., Burnham, K., & Augustin, N. (1997). Model selection: An integral part of inference. *Biometrics, 53*, 603–618.

Buckland, S., & Garthwaite, P. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics, 47*, 255–268.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association, 88*, 364 – 373.

Burnham, K., White, G., & Anderson, D. (1995). Model selection strategy in the analysis of capture-recapture data. *Biometrics, 51*, 888 – 898.

Carroll, J., & Green, P. (1997). *Mathematical Tools for Applied Multivariate Analysis.* Academic Press.

Casella, G., & George, E. (1992). Explaining the Gibbs Sampler. *The American Statistician, 46*(3), 167-174.

Castledine, B. (1981). A Bayesian analysis of multiple-recapture sampling for a closed

population. *Biometrika, 67*, 197–210.

Chang, Y., LaPorte, R., Aaron, D., & Songer, T. (1999). The importance of source selection and pilot study in the capture-recapture application. *Journal of Clinical Epidemiology, 52*(10), 927–928.

Chao, A., Lee, S.-M., & Jeng, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics, 48*, 201–216.

Chao, A., & Tsay, P. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association, 93*, 283–293.

Chao, A., Tsay, P., Lin, S., Shau, W., & Chao, D. (2001). The application of capture-recapture models to epidemiological data. *Statistics in Medicine, 20*, 3123–3157.

Chapman, D. (1951). Some properties of the hypergeometric ditribution with applications to zoological censuses. *University of California Publications in Statistics, 1*, 131–160.

Chaudhuri, A., & Stenger, H. (2005). *Survey Sampling: Theory and Methods* (2nd ed.). Chapman and Hall \CRC.

Chib, S., & Carlin, B. P. (1999). On MCMC sampling in hierarchical models. *Statistics and Computing, 9*, 17–26.

Christensen, R. (1996). *Plane Answers to Complex Questions* (2nd ed.). Springer.

Christensen, R. (1997). *Log-linear Models and Logistic Regression* (2nd ed.). New York: Springer.

Coffman, C. J., Horner, R., Grambow, S. C., & Lindquist, J. (2005). Estimating the occurrence of amyotrophic lateral sclerosis among Gulf War (1990-1991) veterans using capture-recapture methods. *Neuroepidemiology, 24*, 141–150.

Congdon, P. (2005). *Bayesian Models for Categorical Data.* New York: Wiley.

Cormack, R. (1989). Log-linear models for capture-recapture. *Biometrics*, *45*, 395–413.

Cormack, R. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, *48*(2), 567–576.

Cormack, R. (1999). Problems with using capture-recapture in epidemiology: An example of a measles epidemic. *Journal of Clinical Epidemiology*, *52*(10), 909–914.

Cormack, R. (2000). On the need for a 16th and 17th recommendation for capture-recapture analysis - response. *Journal of Clinical Epidemiology*, *53*(12), 1276–1277.

Cormack, R., Chang, Y., & Smith, G. (2000). Estimating deaths from industrial injury by capture-recapture: a cautionary tale. *International Journal of Epidemiology*, *29*(6), 1053–1059.

Cormack, R., & Jupp, P. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, *78*(4), 911–916.

Corona, T., & Romàn, G. (2006). Multiple sclerosis in Latin America. *Neuroepidemiology*, *26*(6), 1–3.

Coull, B., & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, *55*, 294–301.

Coull, B., & Agresti, A. (2003). Generalized log-linear models with random effects, with application to smoothing contingency tables. *Statistical Modelling*, *3*, 251–271.

Cox, D. (2003). Conditional and marginal association for binary random variables. *Biometrika*, *90*, 982–984.

Cox, L. (2003). On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, *117*, 251–273.

D'Agostino, R., & Rubin, D. (2000). Estimating and using propensity scores with

partially missing data. *Journal of the American Statistical Association, 95*, 749–759.

Darroch, J. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika, 45*.

Darroch, J., Fienberg, S., Gloneck, G., & Junker, B. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association, 88*, 1137–1148.

de Greef, S., Spanjaard, L., Danker, J., Hoebe, C., Nagelkerke, N., & de Melker, H. (2006). Underreporting of meningococcal disease incidence in The Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false positive diagnoses. *European Journal of Epidemiology, 21*(4), 315–321.

Debrock, C., Preux, P., & Houinato, D. (2000). Estimation of the prevalence of epilepsy in the Benin region of Zinvié using capture-recapture method. *International Journal of Epidemiology, 29*, 330–335.

Dellaportas, P., & Forster, J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika, 86*, 615–633.

Dendukuri, N., & Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics, 57*, 158–167.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of Longitudinal Data* (2nd ed.). New York: Oxford University Press.

Dobra, A., Tebaldi, C., & West, M. (2005). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference, 136*(2), 355–372.

Domingo-Salvany, A., Hartnoll, R., Maguire, A., Brugal, M., Albertín, P., Caylà, J., et al. (1998). Analytical considerations in the use of capture-recapture to estimate prevalence: Case studies of the estimation of opiate use in the metropolitan area

of Barcelona, Spain. *American Journal of Epidemiology, 148*, 732–740.

Draper, D. (1995). Assessment and propagation of model uncertainty [with discussion]. *Journal of the Royal Statistical Society, Series B, 57*, 45–70.

Egeland, G., Perham-Hester, K., & Hook, E. (1995). Use of capture-recapture analyses in fetal alcohol syndrome surveillance in Alaska. *American Journal of Epidemiology*(4), 335–341.

Eriksson, N., Fienberg, S. E., Rinaldo, A., & Sullivant, S. (2005). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *arXiv.*

Evans, M., & Bonett, D. (1994). Bias reduction for multiple-recapture estimators of closed population size. *Biometrics*, 388–395.

Farrington, C. (2002). Interval estimation for Poisson capture-recapture models in epidemiology. *Statistics in Medicine*, 3079–3092.

Fienberg, S. (1972). Multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika, 59*(3), 591–603.

Fienberg, S. (1977). *The Analysis of Cross-Classified Categorical Data.* MIT Press.

Fienberg, S. (1992). Bibliography on capture-recapture modeling with application to census undercount adjustment. *Survey Methodology, 18*, 143–154.

Fienberg, S. (2000). Contingency tables and log-linear models: basic results and new developments. *Journal of the American Statistical Association, 95*, 643–647.

Fienberg, S., Johnson, M., & Junker, B. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, part A, 162*, 383–406.

Fingleton, B. (1984). *Models of Category Counts.* CUP.

Forbes, R., & Swingler, R. (1999). Estimating the prevalence of multiple sclerosis in the United Kingdom by using capture-recapture methodology. *American Journal of Epidemiology, 149*(11), 1016–1024.

Foreman, E. K. (1991). *Survey Sampling Principles* (2nd ed.). Dekker.

Freedman, D. (1991). Adjusting the 1990 census. *Science, 252*(5010), 1233–1236.

Garthwaite, P., Yu, K., & Hope, P. (1995). Bayesian analysis of a multiple-recapture model. *Communications in Statistics - Theory and Methods, 24*, 2229–2247.

Gelfand, A., Sahu, S., & Carlin, B. (1995). Efficient parametrization for normal linear mixed models. *Biometrika, 82*, 476–488.

Gelfand, A., Sahu, S., & Carlin, B. (1996). Efficient parametrization for generalised linear models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 5* (p. 479-488). Oxford University Press.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis* (2nd ed.). Boca Raton: Chapman & Hall

Gentle, J. (1998). *Numerical Linear Algebra for Applications in Statistics.* Springer.

Gentleman, R., & Vandal, A. (2001). Computational algorithms for censored data problems using intersection graphs. *Journal of Computational and Graphical Statistics, 10*, 403–421.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo Methods in Practice.* Chapman and Hall, New York.

Good, I. (1953). On the population frequencies of species and the estimation of population parameters. *Biometrika, 40*, 237–264.

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

Grijalva, C., Craig, A., Dupont, W., Bridges, C., Schrag, S., Iwane, M., et al. (2006). Estimating influenza hospitalizations among children. *Emerging Infectious Diseases [serial on the Internet]*, Available from http://www.cdc.gov/ncidod/EID/vol12no01/05-0308.htm.

Gurgel, R., da Fonseca, J., Neyra-Castaeda, D., Gill, G., & Cuevas, L. (2004). Capture-recapture to estimate the number of street children in a city in Brazil. *Archives of Disease in Childhood*, 222–224.

Haber, M. (1985a). Log-linear models for correlated marginal totals of a contingency table. *Communications in Statististics- Theory and Methods*, *14*(12), 2845–2856.

Haber, M. (1985b). Maximum likelihood methods for linear and log-linear models in categorical data. *Computational statistics and data analysis*, *3*(1), 1–10.

Haber, M., & Brown, M. (1986). Maximum likliehood methods for log-linear models when expected frequencies are subjected to linear constraints. *Journal of the American Statistical Association*, *81*, 477–482.

Hagan, J., Schoenfeld, H., & Palloni, A. (2006). The science of human rights, war crimes, and humanitarian emergencies. *Annual Review of Sociology*, *32*, 329–349.

Hay, G. (1997). The selection from multiple data sources in epidemiological capture-recapture studies. *The Statistician*, *46*(4), 515–520.

Hickman, M., Higgins, V., Hope, V., Bellis, M., Tilling, K., Walker, A., et al. (2004). Injecting drug use in Brighton, Liverpool, and London: best estimates of prevalence and coverage of public health indicators. *Journal of Epidemiology and Community Health*, *58*, 766–771.

Hook, E., & Regal, R. (1980). Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York state. *American Journal of Epidemiology*, *112*, 750–758.

Hook, E., & Regal, R. (1992). The value of capture-recapture methods even for apparently exhaustive surveys. *American Journal of Epidemiology*, *135*, 1060–1067.

Hook, E., & Regal, R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiological Reviews*, *17*(2), 243–264.

Hook, E., & Regal, R. (1997). Validity of methods for model selection, weight-

ing for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology, 145*(12), 1138–1144.

Hook, E., & Regal, R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology, 52*(10), 917–926.

Hook, E., & Regal, R. (2000). On the need for a 16th and 17th recommendation for capture-recapture analysis. *Journal of Clinical Epidemiology, 53*(12), 1275–1276.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika, 76*, 133–140.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995a). Capture-recapture and multiple-record systems estimation. I. History and theoretical developments. *American Journal of Epidemiology, 142*, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995b). Capture-recapture and multiple-record systems estimation. II. Applications in human diseases. *American Journal of Epidemiology, 142*, 1059–1068.

Ismail, A., Beeching, N., Gill, G., & Bellis, M. (2000). How many data sources are needed to determine diabetes prevalence by capture-recapture? *International Journal of Epidemiology, 29*(3), 536–541.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman and Hall.

Joffe, M., & Rosenbaum, P. (1999). Propensity scores. *American Journal of Epidemiology, 150*, 327–333.

Joseph, L., Gyorkos, T., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology, 141*, 263–272.

Kadane, J., Meyer, M., & Tukey, J. (1999). Yule's association paradox and ignored

stratum heterogeneity in capture-recapture studies. *Journal of the American Statistical Association, 94*, 855–859.

Kim, D. K., & Taylor, J. M. G. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *J. Amer. Statist. Assoc., 90*, 708–716.

King, R., & Brooks, S. (2001a). On the Bayesian analysis of population size. *Biometrika, 88*(2), 317–336.

King, R., & Brooks, S. (2001b). Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics, 29*(3), 715–747.

King, R., & Brooks, S. (2002). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika, 89*(4), 785–806.

Kollo, T., & von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices.* Springer.

Lang, J. (1996a). Maximum likliehood methods for a generalized class of log-linear models. *The Annals of Statistics, 24*, 726–752.

Lang, J. (1996b). On the comparison of multinomial and Poisson log-linear models. *Journal of the Royal Statistical Society. Series B (Methodological), 58*, 253–266.

Lang, J., & Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association, 89*, 625–632.

Lang, J., McDonald, J., & Smith, P. (1999). Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *Journal of the American Statistical Association, 94*, 1161–1171.

Lange, J., Chang, Y.-F., & LaPorte, R. (2004). Use of the capture-recapture method for epidemiological studies in determining prevalence. *Acta Neurologica Scandinavica, 109*, 79.

Lange, J., Chang, Y.-F., LaPorte, R., & Mastrangelo, G. (2003). Hazardous waste

site frequency: use of the capture-recapture method. *Toxicology and Industrial Health, 19*, 109–113.

Lange, J., & LaPorte, R. (2003). capture-recapture method should be used to count how many cases of SARS really exist. *British Medical Journal, 326*, 1396.

Laporte, R. (1994). Assessing the human condition - capture-recapture techniques. *British Medical Journal, 308*, 6920.

Laska, E. (2002). The use of capture - recapture methods in public health. *Bulletin of the World Health Organization, 80*(11), 845.

Lauritzen, S. (1995). *Graphical Models.* New York: Oxford Science Publications.

Lee, A. (2002). Effect of list errors on the estimation of population size. *Biometrics, 58*(1), 185–191.

Lee, A., Seber, G., Holden, J., & Huakau, J. (2001). Capture-recapture, epidemiology, and list mismatches: Several lists. *Biometrics, 57*(3), 707–713.

Lesperance, M., & Kalbfleisch, J. (1992). An algorithm for computing the non-parametric MLE of a mixing distribution. *Journal of the American Statistical Association, 87*(417), 120–126.

Lindsey, J. K. (1995). *Modelling Frequency and Count data.* Oxford Science Publications.

Link, W. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics, 59*, 1123–1130.

Madigan, D., & York, J. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika, 84*, 19–31.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review, 63*, 215–232.

Magnus, J. R., & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons.

McCullagh, P., & Nelder, J. A. (1999). *Generalized Linear Models* (2nd ed.). Chap-

man & Hall \CRC.

McCulloch, C., & Searle, S. (2001). *Generalized, Linear and Mixed Models.* New York: Wiley.

Melocco, M. (2002). *Modeling Heterogeneity of Capture Probabilities in Capture-Recapture Studies.* Unpublished master's thesis, McGill University.

Mitchell, R., Dorling, D., Martin, D., & Simpson, L. (2002). Bringing the missing million home: correcting the 1991 small area statistics for undercount. *Environment and Planning A, 34*(6), 1021–1035.

Molenberghs, G., & Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine, 18*(17-18), 2237–2255.

Nocedal, J., & Wright, S. J. (1999). *Numerical Optimization.* Springer Series in Operations Research.

Norris, J., & Pollock, K. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics, 52*, 639–649.

O'Callaghan, F., Shiell, A., Osborne, J., & Martyn, C. (1998). Capture-recapture analysis to estimate the prevalence of tuberous sclerosis. *Lancet, 352*(9124), 318–319.

Papaspiliopoulos, O., Roberts, G., & Sköld, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In J. M. Bernardo et al. (Eds.), *Bayesian Statistics 7* (p. 307-326). Oxford University Press.

Papoz, L., Balkau, B., & Lellouch, J. (1996). Case counting in epidemiology: Limitations of methods based on multiple data sources. *International Journal of Epidemiology, 25*(3), 474–478.

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station (1895), 6*, 5–84.

Plante, N., Rivest, L.-P., & Tremblay, G. (1998). Stratified capture-recapture estimation of the size of a closed population. *Biometrics, 54*, 47–60.

Pledger, S. (2000). Unified maximum likelihood estimation for closed capture-recapture models using mixtures. *Biometrics*, *56*, 434–442.

Pollock, K., Hines, J., & Nichols, J. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, *40*, 329–340.

Preux, P.-M., Druet-Cabanac, M., Couratier, P., Debrock, C., Truong, T., Marcharia, W., et al. (2000). Estimation of the amyotrophic lateral sclerosis incidence by capture-recapture method in the Limousin region of France. *Journal of Clinical Epidemiology*, *53*(5), 1025–1029.

Rasch, G. (1969). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

Redfern, P. (2004). An alternative view of the 2001 census and future census taking. *Journal of the Royal Statistical Society Series A*, *167*(2), 209–228.

Regal, R., & Hook, E. (1984). Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Statistics in Medicine*, *3*, 287–291.

Regal, R., & Hook, E. (1998). Marginal versus conditional versus structural source models: A rationale for an alternative to log-linear methods for capture-recapture estimates. *Statistics in Medicine*, *17*, 69–74.

Regal, R., & Hook, E. B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, *10*, 717–721.

Rivest, L.-P., & Lévesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics*, *29*(4), 555-572.

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (Second ed.). Springer.

Roberts, H. (1967). Informative stopping rules and inference about population size. *Journal of the American Statistical Association*, *62*, 763–775.

Robinson, J., West, K., & Adlakha, A. (2002). Coverage of the population in cen-

sus 2000: Results from demographic analysis. *Population Research and Policy Review*, *21*(1-2), 19–38.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

R programming language [Programming language]. (2004). Available free online at http://jbr.org/articles.html. The R Development Core Team.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

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

Sakamoto, Y., Ishiguro, M., & Kitigawa, G. (1986). *Akaike Information Criterion Statistics.* Tokyo, Japan: KTK Scientific.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, *43*(1), 142–152.

Sanchez, O., Buritica, O., Pineda, D., Uribe, C., & Palacio, L. (2004). Prevalence of Parkinson's disease and Parkinsonism in a Colombian population using the capture-recapture method. *International Journal of Neuroscience*, *114*(2), 175–182.

Sanderson, M., Benjamin, J., Lane, M., C.B., C., & Davis, D. (2003). Application of capture-recapture methodology to estimate the prevalence of dementia in South Carolina. *Annals of Epidemiology*, *13*, 518–524.

Sandland, R. L., & Cormack, R. (1984). Statistical inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, *71*(1), 27–33.

Schwarz, C., & Seber, G. (1999). Estimating animal abundance: review III. *Statistical Science*, *14*(4), 427–456.

Searle, S. (1982). *Matrix Algebra Useful for Statistics.* Wiley.

Seber, G. (1982). *The Estimation of Animal Abundance and Related Parameters* (2nd ed.). London: Griffin.

Seber, G., Huakau, J., & Simmons, D. (2000). Capture-recapture, epidemiology, and list mismatches: Two lists. *Biometrics, 56*(4), 1227–1232.

Sekar, C., & Deming, W. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association, 44*, 101–115.

Silva, R., & Ball, P. (2005). *On the Use of Sample Surveys and Multiple Systems Estimations in Assessing Large-Scale Human Rights Violations: Recent Experiences from Timor-Leste.* American Statistical Association.

Silvapulle, M. J., & Sen, P. (2005). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions.* Wiley.

Smith, A., & Roberts, G. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B, 55*(1), 2–23.

Smith, P. (1991). Bayesian analysis for a multiple capture-recapture model. *Biometrika, 78*(2), 399–407.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B, 64*(4), 583–639.

Stanghellini, E., & van der Heijden, P. (2004). A multiple-record systems estimation method that takes observed and unobserved heterogenity into account. *Biometrics, 60*, 510–516.

Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics, 43*, 243–258.

Sutherland, J. M. (2003). *Multi-List Methods in Closed Populations with Stratified or Incomplete Information.* Unpublished doctoral dissertation, Simon Fraser

University.

Tanner, M. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (3rd ed.). Springer.

Tilling, K. (2001). Capture-recapture methods - useful or misleading? *International Journal of Epidemiology, 30*(1), 12–14.

Tilling, K., & Sterne, J. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology, 149*, 392–400.

Tilling, K., Sterne, J., & Wolfe, C. (2001). Estimation of the incidence of stroke using a capture-recapture model including covariates. *International Journal of Epidemiology, 30*, 1351–1359.

Totaro, R., Marini, C., Cialfi, A., Giunta, M., & Carolei, A. (2000). Prevalence of multiple sclerosis in the L'Aquila district, central Italy. *Journal of Neurology, Neurosurgery & Psychiatry, 68*(3), 349–352.

Turner, E. (2002). *Analysis of Prostate Specific Antigen "Trajectories": Statistical Challenges.* Unpublished master's thesis, McGill University.

Vandal, A., Walker, N., & Pearson, J. (2005). *A covariate-based coefficient of source dependence for capture-recapture models.* Technical report, McGill University.

Walker, N., Vandal, A., Holden, J., Rodgers, A., Birchall, N., Norton, R., et al. (2002). Does capture-recapture analysis provide more reliable estimates of the incidence and prevalence of leg ulcers in the community? *Australia and New Zealand Journal of Public Health, 26*(5), 451–455.

Wand, M. (2002). Vector differential calculus in statistics. *The American Statistician, 56*(1), 1–7.

Wang, X., He, C., & Sun, D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *Journal of Statistical Planning and Inference, 137*(4), 1099–1118.

Warren, S., & Warren, K. (1992). Prevalence of multiple sclerosis in Barrhead county,

Alberta, Canada. *Canadian Journal of Neurological Sciences, 19*(1), 72–75.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* Springer.

Whittaker. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley.

Wittes, J. (1969). Applications of a multinomial capture-reapture model to epidemiological data. *Journal of the American Statistical Association, 69*, 93–97.

Wittes, J., Colton, T., & Sidel, V. (1974). Capture-reapture models for assessing the completeness of case ascertainment using multiple information sources. *Journal of Chronic Diseases, 27*, 25–36.

Zhao, Y., Staudenmayer, J., Coull, B., & Wand, M. (2006). General design Bayesian generalized linear mixed models. *Statistical Science, 21*(1), 35–51.

Zwane, E., & van der Heijden, P. G. M. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling, 5*, 39–52.