Biometrics 00, 000–000 DOI: xxxx
MMM 0000

Connecting the Latent Multinomial

Matthew R. Schofield[1*], Simon J. Bonner[2]

[1]Department of Mathematics and Statistics, University of Otago, P.O. Box 56 Dunedin 9054, New Zealand

[2]Department of Statistics, University of Kentucky, Lexington KY 40536, U.S.A.

email: mschofield@maths.otago.ac.nz

Summary: Link et al. (2010) define a general framework for analyzing capture-recapture data with potential misidentifications. In this framework, the observed vector of counts, $\boldsymbol{y}$, is considered as a linear function of a vector of latent counts, $\boldsymbol{x}$, such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, with $\boldsymbol{x}$ assumed to follow a multinomial distribution conditional on the model parameters, $\boldsymbol{\theta}$. Bayesian methods are then applied by sampling from the joint posterior distribution of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$. In particular, Link et al. (2010) propose a Metropolis-Hastings algorithm to sample from the full conditional distribution of $\boldsymbol{x}$, where new proposals are generated by sequentially adding elements from a basis of the null space (kernel) of $\boldsymbol{A}$. We consider this algorithm and show that using elements from a simple basis for the kernel of $\boldsymbol{A}$ may not produce an irreducible Markov chain. Instead, we require a Markov basis, as defined by Diaconis and Sturmfels (1998). We illustrate the importance of Markov bases with three capture-recapture examples. We prove that a specific lattice basis is a Markov basis for a class of models including the original model considered by Link et al. (2010) and confirm that the specific basis used by Link et al. (2010) for their example with two sampling occasions is a Markov basis. The constructive nature of our proof provides an immediate method to obtain a Markov basis for any model in this class.

Key words: Capture-recapture; Linear constraint; Markov basis; Markov chain Monte Carlo; Misidentification.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The development of capture-recapture methodology has a long history, allowing estimation of demographic parameters of interest for animal populations (see Amstrup et al. 2005, for a review). Similar methods have also been used to study human populations, including intravenous drug users (King et al. 2009) and human rights abuse victims (Lum et al. 2013). In general, a capture-recapture experiment consists of a series of capture occasions on which overlapping subsets of the population are observed. For animal populations the occasions are usually ordered in time while for human populations they may comprise lists obtained from different sources. It is assumed that each individual has a unique identifying mark that is either given or realized when the individual is first captured and this mark can be used to identify the individual on subsequent occasions. In this paper, we are concerned with fitting capture-recapture models to data that provide an incomplete or inaccurate representation of the true encounters of individuals during the experiment. This may occur if the data consist of incomplete summary statistics or if individuals are misidentified on some occasions. Examples of capture-recapture studies that are prone to identification errors include (i) multi-list studies in which individuals may be matched based on personal information such as name, birth date, medical record number (Seber et al. 2000, Lee et al. 2001, Sutherland and Schwarz 2005, Fienberg and Manrique-Vallier 2009), (ii) animal studies in which individual identity is found from non-invasive sampling, e.g. genetic information from scat or hair (Wright et al. 2009, Link et al. 2010, Yoshizaki et al. 2011) or photographic ID of individuals (Yoshizaki et al. 2009, Bonner and Holmberg 2013, McClintock et al. 2013), and (iii) studies in which (at least) two sources of capture-recapture information are available for the same population with little to no information about how the individual IDs in one source corresponds to individual ID from the other sources (Bonner and Holmberg 2013, McClintock et al. 2013).

Our focus is on the algorithm for a general class of mark-recapture models allowing for

misidentification considered by Link et al. (2010) (hereafter L2010). This class is described by the latent multinomial model, in which an observed data vector, $\boldsymbol{y}$ can be expressed as a linear function of a latent data vector, $\boldsymbol{x}$, modeled by a multinomial distribution with unknown parameters $\boldsymbol{\theta}$, denoted $[\boldsymbol{x}|\boldsymbol{\theta}]$. The notation $[x]$ denotes the probability density function $f_X(x)$ for a continuous random variable X or the probability mass function $\Pr(X = x)$ for a discrete random variable X. The linear function is expressed as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{A}$ is called the configuration matrix (a matrix of known constants that depends on the specific problem) with more columns than rows. We continue to call this modeling setup the latent multinomial model, even though the setup is flexible and can accommodate other probability mass functions $[\boldsymbol{x}|\boldsymbol{\theta}]$, such as the Poisson model considered by Lee (2002).

The goal is to sample from the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}]$ using Markov chain Monte Carlo (MCMC) by alternating between sampling from the full conditional distributions $[\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}]$ and $[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}]$. The difficulty with this approach is in specifying an updating scheme for $\boldsymbol{x}$. That is, how to efficiently sample from $[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}]$ in such a way so that every $\boldsymbol{x}$ vector that satisfies (1) has a positive probability of being reached at some point during the updating. We consider three examples demonstrating that the scheme for updating $\boldsymbol{x}$ proposed by L2010 may not produce an irreducible Markov chain for models within the latent multinomial framework. We then present theory identifying a class of models for which the specific algorithm does produce irreducible Markov chains, and show more generally how these methods fit within the framework of algebraic statistics. This allows us to develop an extension of the algorithm which can be used to generate valid MCMC samplers for the posterior distributions from a broader class of latent multinomial models.

The MCMC algorithm we consider throughout this manuscript is presented in Figure 1. Starting with an initial state $\boldsymbol{x}^0$ satisfying the linear constraint, a proposal is generated on

the first iteration by adding or subtracting an element chosen randomly from a subset of the kernel (or null space) of $\boldsymbol{A}$, $\mathcal{B} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\} \subset \ker(\boldsymbol{A})$, with cardinality m. The proposal is then accepted or rejected with probability determined by the Hasting's ratio, r, and the algorithm continues to the second iteration. This algorithm is a modification of that presented by L2010, with three differences: (i) L2010 steps through all m elements in $\mathcal{B}$ in order instead of selecting an element at random on each iteration, (ii) when stepping through every element in $\mathcal{B}$, L2010 multiplies element $\boldsymbol{a}_i$ by a coefficient $c \in \{-C_i, \ldots, -1, 1, \ldots, C_i\}$ in order to improve convergence, and (iii) L2010 assumes that $\mathcal{B}$ is a basis for $\ker(\boldsymbol{A})$, while we allow $\mathcal{B}$ to be a more general subset that spans $\ker(\boldsymbol{A})$. The first two differences may impact the efficiency of the algorithm but do not change the stationary distribution of the resulting Markov chains, and we do not consider these differences further. Our focus is on the third difference and the effect that the set $\mathcal{B}$ can have on the generated Markov chains and their stationary distributions.

[Figure 1 about here.]

To illustrate the problems that may occur if $\mathcal{B}$ is poorly specified we consider three examples of models which fit into the latent multinomial framework. First we consider the same closed population mark-recapture model with misidentification considered by L2010. This model, called $\mathrm{M}_{\mathrm{t}\alpha}$, assumes that captures occur according to a closed population model with time dependent capture probabilities and that errors in identifying an individual are unique and create ghost histories with single captures. Second, we consider a multi-list modeling problem in which summary statistics are presented in place of the full data set, possibly for privacy reasons. Our aim is to sample from possible complete data sets with the given sufficient statistics. Finally, we consider a more complicated model of misidentification in mark-recapture which allows for one marked individual to be identified as another previously marked individual. Full details of these models and the issues regarding the selection of the

set $\mathcal{B}$ to be used in the algorithm in Figure 1 are provided in sections 3, 4, and 5. As motivation, we consider the output from Markov chains constructed using the algorithm in Figure 1 for each of the three examples. For each example, we defined $\mathcal{B}$ to be a basis for $\ker(\boldsymbol{A})$ as in L2010 and ran two parallel chains, each of which started from a different initial value. For both model $M_{t\alpha}$ and the multi-list model with sufficient statistics, despite strong evidence that each chain has converged, it is clear that the two chains are not sampling from the same distribution for a given quantity of interest (Figure 2). This is even more apparent in the third example where one of the two chains never moves from its initial value.

[Figure 2 about here.]

The problem in all three examples is that the stationary distribution reached by the Markov chains produced by the algorithm in Figure 1 may depend on the chosen set, $\mathcal{B}$ and the initial value of $\boldsymbol{x}$. Although the values of $\boldsymbol{x}$ proposed on each iteration are guaranteed to satisfy the linear constraint the resulting Markov chains may not reach all points in the sample space and the stationary distributions may be dependent on the initial values. In the next section we provide a basic introduction to the field of algebraic statistics and the results of Diaconis and Sturmfels (1998) and others who have explored approaches for sampling from $\boldsymbol{x}$ from a linear constraint as in (1) in other application areas. We then consider the implications of this theory to show why the MCMC algorithms failed above (Figure 2), and how valid MCMC samplers can be constructed for each of the three examples.

2. Introduction to algebraic statistics

Sampling $\boldsymbol{x}$ in the presence of the linear constraint in (1) is not unique to capture-recapture problems. In a seminal paper in algebraic statistics, Diaconis and Sturmfels (1998) considered a linear constraint of the same form when developing conditional goodness-of-fit tests for contingency tables (see Karwa and Slavkovic 2013, for a recent review). That is, they

considered how to construct an MCMC algorithm to sample different contingency tables with common (fixed) row and column sums (such ideas can also be extended to multi-way contingency tables).

To consider the problem at hand in more detail we will summarize several definitions and results from linear algebra in this section (basic definitions regarding kernels and bases are provided in the supplementary materials). We will use a 3×3 contingency table example to illustrate many of the ideas. The table is

$$\begin{array}{ccc|c} x_{11} & x_{12} & x_{13} & x_{1\cdot} \\ x_{21} & x_{22} & x_{23} & x_{2\cdot} \\ x_{31} & x_{32} & x_{33} & x_{3\cdot} \\ \hline x_{\cdot 1} & x_{\cdot 2} & x_{\cdot 3} & \end{array}$$

where x_{ij} is the value in the ith row and jth column, $x_{\cdot j}$ refers to the sum of the jth column and $x_{i\cdot}$ refers to the sum of the ith row. The column and row sums are vectorized to give the vector of summary statistics

$$\boldsymbol{y} = (x_{\cdot 1}, x_{\cdot 2}, x_{\cdot 3}, x_{1\cdot}, x_{2\cdot})'.$$

Note that we need not include the third row sum as this is a derived quantity of the other elements of $\boldsymbol{y}$. The individual entries in the table are vectorized to give

$$\boldsymbol{x} = (x_{11}, x_{21}, x_{31}, x_{12}, x_{22}, x_{32}, x_{13}, x_{23}, x_{33})'.$$

The specification is completed with

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

so that the constraints inherent in a contingency table follow (1). If we have column/row

sums given by

$$\boldsymbol{y} = (5, 3, 2, 0, 4)'$$

then two contingency tables compatible with these constraints have entries

$$\boldsymbol{x}_1 = (0, 2, 3, 0, 1, 2, 0, 1, 1)' \quad \text{and} \quad \boldsymbol{x}_2 = (0, 3, 2, 0, 0, 3, 0, 1, 1)'. \tag{2}$$

Our goal is to specify an MCMC algorithm that samples from the set of vectors $\boldsymbol{x}$ that satisfy (1) for a particular $\boldsymbol{y}$. This is defined as the $\boldsymbol{y}$-fiber (or simply fiber) $\mathcal{F}_{\boldsymbol{y}}$,

$$\mathcal{F}_{\boldsymbol{y}} = \{\boldsymbol{x} \in \mathbb{N}^d : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}\},$$

where d is the dimension of $\boldsymbol{x}$ and $\mathbb{N} = \{0, 1, \ldots\}$. L2010 refers to $\mathcal{F}_{\boldsymbol{y}}$ as the feasible set.

To move between elements of the fiber, we make use of the lattice kernel $\ker_{\mathbb{Z}}(\boldsymbol{A})$. The lattice kernel is the integer valued subset of the kernel,

$$\ker_{\mathbb{Z}}(\boldsymbol{A}) = \ker(\boldsymbol{A}) \bigcap \mathbb{Z}^d = \{\boldsymbol{x} \in \mathbb{Z}^d : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

In algebraic statistics, a move is defined to be any element of the lattice kernel, such that the vector $\boldsymbol{v}$ is a move if $\boldsymbol{v} \in \ker_{\mathbb{Z}}(\boldsymbol{A})$. An implication of this is that if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{F}_{\boldsymbol{y}}$ then $\boldsymbol{x}_2 - \boldsymbol{x}_1$ is a move. The idea is that the elements of the lattice kernel can be added to a vector that satisfies the linear constraint and the result is guaranteed to still satisfy the constraint. However, it is not practical to consider all elements of the lattice kernel when updating $\boldsymbol{x}$ as $\ker(\boldsymbol{A})$ is potentially very large and difficult to compute. Instead we want to find a smaller set of moves $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\} \subset \ker_{\mathbb{Z}}(\boldsymbol{A})$ that can be used to update $\boldsymbol{x}$. That is, we require a smaller set of moves so that it is possible to move between all elements of $\mathcal{F}_{\boldsymbol{y}}$ using the algorithm in Figure 1.

The suggestion of L2010 was to use a basis for $\ker(\boldsymbol{A})$ for this set of moves. However, we do not wish to construct a basis for $\ker(\boldsymbol{A})$, but instead a lattice basis for the integer lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$. A lattice basis is a set of linearly independent vectors where every $\boldsymbol{v} \in \ker_{\mathbb{Z}}(\boldsymbol{A})$ can be found as a linear combination of the lattice basis vectors using integer coefficients.

If we insist on using a basis for $\ker(\boldsymbol{A})$, it may not be possible to reach all solutions using only integer values of the coefficients, c, as specified in the algorithm in Figure 1. However, even if we choose to use a lattice basis for $\mathcal{B}$ it may be necessary to pass through one (or more) vectors containing negative elements when applying moves one at a time to transition between elements in the fiber $\mathcal{F}_{\boldsymbol{y}}$. As vectors $\boldsymbol{x}$ containing negative elements can never be accepted, the use of a lattice basis for $\mathcal{B}$ may result in sampling from a subset of the fiber $\mathcal{F}_{\boldsymbol{y}}$ when using the algorithm in Figure 1. This explains the observed results in the three examples shown in Section 1: the two chains are exploring different subsets of the fiber.

These ideas are formalized using the concept of connectivity. Elements $\boldsymbol{x}_j, \boldsymbol{x}_k \in \mathcal{F}_{\boldsymbol{y}}$ are connected using the set $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ if there are moves $\boldsymbol{v}_i \in \boldsymbol{V},\ i \in \{1, \ldots, M\}$ so that we can start from $\boldsymbol{x}_j$ and add or subtract these moves one at a time to reach $\boldsymbol{x}_k$ without any element in any of the partial sums ever being negative (note that the elements $\boldsymbol{v}_i,\ i = 1, \ldots, M$ need not be distinct and some elements may be repeated multiple times). That is, there exist $\epsilon_1, \ldots, \epsilon_M \in \{-1, 1\}$ such that

$$\boldsymbol{x}_k = \boldsymbol{x}_j + \sum_{j=1}^{M} \epsilon_j \boldsymbol{v}_j \qquad \text{and} \qquad \boldsymbol{x}_1 + \sum_{k=1}^{L} \epsilon_k \boldsymbol{v}_k \in \mathcal{F}_{\boldsymbol{y}}, \quad L = 1, \ldots, M-1.$$

We then say that the fiber $\mathcal{F}_{\boldsymbol{y}}$ is connected by $\boldsymbol{V}$ if every pair of elements in the fiber are connected.

We can apply the algorithm in Figure 1 to the 3×3 contingency table example using the elements of a lattice basis. A lattice basis can be found using the Hermite normal form (Aoki et al. 2012, pg. 53). Unless otherwise stated, all lattice bases provided in this manuscript are found using this approach. We note that the lattice basis obtained is not unique and a different basis is often found if one reorders the columns of $\boldsymbol{A}$ (and corresponding entries of

$\boldsymbol{x}$). For the contingency table, a lattice basis is given by elements LB1 – LB4 in (3)

	x_{11}	x_{21}	x_{31}	x_{12}	x_{22}	x_{32}	x_{13}	x_{23}	x_{33}
LB1	1	−1	0	−1	1	0	0	0	0
LB2	−1	0	1	1	0	−1	0	0	0
LB3	1	−1	0	0	0	0	−1	1	0
LB4	0	0	0	1	0	−1	−1	0	1

(3)

If we attempt to apply any of the elements LB1 — LB4 to either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ in (2) we immediately find a problem. Either adding or subtracting any of LB1 – LB4 results in at least one negative count in the proposal and will lead to it being automatically rejected. That means there is no way to use the elements LB1 – LB4 as moves in the algorithm in Figure 1 and successfully transition between the two solutions in (2). In fact, we are unable to move between any two valid solutions. As a result, the lattice basis in (3) does not connect the fiber for this example. One solution is to change the algorithm in Figure 1 to use elements of a lattice basis in a linear combination instead of one-at-a-time. While attractively simple, Diaconis and Sturmfels (1998) implemented this for several examples and found that it was inefficient and did not work well in practice. We do not consider this further.

To overcome the shortcomings of constructing moves via integer multiples of an element from a lattice basis, we take a Markov basis for the set $\mathcal{B}$ (Diaconis and Sturmfels 1998). A Markov basis is a larger set of elements in $\ker_{\mathbb{Z}}(\boldsymbol{A})$ that connects all fibers $\mathcal{F}_{\boldsymbol{y}}$ irrespective of the given values in $\boldsymbol{y}$. A finite set $\mathcal{M} \subset \ker_{\mathbb{Z}}(\boldsymbol{A})$ is a Markov basis if, for any $\boldsymbol{y}$ such that $\mathcal{F}_{\boldsymbol{y}} \neq \varnothing$ and for all elements $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{F}_{\boldsymbol{y}}$, $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, there exist $M > 0$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M \in \mathcal{M}$ and $\epsilon_1, \ldots, \epsilon_M \in \{-1, 1\}$ such that

$$\boldsymbol{x}_2 = \boldsymbol{x}_1 + \sum_{j=1}^{M} \epsilon_j \boldsymbol{v}_j \qquad \text{and} \qquad \boldsymbol{x}_1 + \sum_{k=1}^{L} \epsilon_k \boldsymbol{v}_k \in \mathcal{F}_{\boldsymbol{y}}, \quad L = 1, \ldots, M-1.$$

The first condition says that we can use moves from a Markov basis as in the algorithm in Figure 1 to move between any two elements of our fiber. The second condition says that when moving between any two elements in the fiber, we always remain in the fiber (i.e. we never encounter a negative count).

Although Markov bases are relatively easy to describe there is no simple algorithm for their computation. Diaconis and Sturmfels (1998) show how a Markov basis can be computed using techniques from commutative algebra. The theory is based on what is now known as the Fundamental Theorem of Markov Bases which describes how finding a Markov basis is equivalent to finding a set of generators of a toric ideal in a polynomial ring associated with the matrix $\boldsymbol{A}$. We refer the interested reader to Cox et al. (2007) for details on commutative algebra and to Diaconis and Sturmfels (1998), Drton et al. (2009), Aoki et al. (2012) and the references therein for additional information on the generation of Markov bases in algebraic statistics. Unless otherwise stated, we use the freely available software `4ti2` (Hemmecke et al. 2013) to compute the Markov bases for the examples in this manuscript.

For the 3×3 contingency table, a Markov basis consists of the nine elements in (4)

	x_{11}	x_{21}	x_{31}	x_{12}	x_{22}	x_{32}	x_{13}	x_{23}	x_{33}
MB1	0	0	0	0	1	−1	0	−1	1
MB2	0	0	0	1	−1	0	−1	1	0
MB3	0	0	0	1	0	−1	−1	0	1
MB4	0	1	−1	0	−1	1	0	0	0
MB5	0	1	−1	0	0	0	0	−1	1
MB6	1	−1	0	−1	1	0	0	0	0
MB7	1	−1	0	0	0	0	−1	1	0
MB8	1	0	−1	−1	0	1	0	0	0
MB9	1	0	−1	0	0	0	−1	0	1

(4)

It is a straightforward exercise to confirm that we can transition between the two solutions in (2) by adding or subtracting moves from (4) one-at-a-time without encountering a negative count. More importantly, the moves in (4) can be used to connect any two solutions in the same fiber, no matter what value of $\boldsymbol{y}$ is observed.

There is often a need to analytically find a Markov basis for a given problem. Even though tools like `4ti2` are freely available, computation of Markov bases remains challenging. As we discuss later, for many of the capture-recapture examples we have explored, `4ti2` can fail to compute Markov bases for studies with a moderate to large number of sampling occasions.

As we know of no simple test to confirm whether a specified set of moves $\mathcal{B}$ is a Markov basis, we often need to rely on theoretically derived Markov bases to confirm that our MCMC algorithms are valid. In the following section we find such a theoretical result for a class of capture-recapture models including $M_{t\alpha}$.

3. Model $M_{t\alpha}$ and Simple Corruptions

Here, we examine model $M_{t\alpha}$, the specific model of misidentification considered by L2010. We fit this model into a larger class of models in which any identification error results in what we refer to as a simple corruption. We then show that for any model in this class, we can construct a lattice basis that is guaranteed to connect every element of the fiber, irrespective of $\boldsymbol{y}$, i.e. it is also a Markov basis.

Model $M_{t\alpha}$ builds on the standard closed population model with time-dependent capture probabilities, model M_t of Otis et al. (1978), by allowing for individuals to be misidentified when captured. The model assumes that all errors are unique meaning that an individual cannot be identified as another individual and the same error cannot occur multiple times. The result is that an error on the j^{th} capture occasion leads to a ghost observed history containing a single observation on the j^{th} occasion.

For this model, the vector of summary statistics, $\boldsymbol{y}$, contains the counts of the $2^K - 1$ observable capture histories. The vector of latent variables contains the counts of the possible true histories constructed from the events:

- 0 – the individual was not captured,
- 1 – the individual was captured and correctly identified,
- 2 – the individual was captured and incorrectly identified.

For example, for a study with $K = 5$ capture occasions the true history 01221 would generate three observed histories: 01001, 00100, and 00010. Including the null history $0 \ldots 0$, the vector

of true counts has length 3^K. The configuration matrix, $\boldsymbol{A}$, has dimension $(2^K - 1) \times 3^K$ and $A_{ij} = 1$ if the j^{th} true history generates the i^{th} observed history and is equal to zero otherwise. For example, the column corresponding to the history 01221 would contain three non-zero entries in the rows associated with the observable histories 01001, 00100, and 00010. A description of the model along with the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and matrix $\boldsymbol{A}$ for $K = 2$ are given in the supplementary materials, with more details in L2010.

A feature of $\mathrm{M}_{\mathrm{t}\alpha}$ is that whenever an error in identification occurs, it involves only one individual and results in one or more observed histories. We define such an error as a simple corruption. For example, the errors in true history 01221 above affect no other true history and lead to three observed histories. Another example of simple corruptions are the errors that occur when multiple marks cannot be matched, as described in Bonner and Holmberg (2013) and McClintock et al. (2013). Suppose that a study uses photographs to identify individuals and that photographs taken from the left or right side cannot be matched without further information. In this case, any individual that is photographed from both the left and right sides on different occasions will contribute two histories to the observed data set. Using the events L and R to denote photographs from the left and right, the true history $0LRRL$ would generate observed histories $0L00L$ and $00RR0$. In this case, each true history will contribute one or two histories to the observed data set.

For a model that contains only simple corruptions, we have the following theorem:

THEOREM 1: *Suppose that: (i) $\boldsymbol{A}$ contains only the values 0 and 1 and (ii) the columns of $\boldsymbol{A}$ contain all of the columns of the identity matrix. Then there exists a lattice basis $\mathcal{L} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$, which is also a Markov basis.*

The first condition (values of 0 and 1) occurs under the assumption of simple corruption, while the second condition (columns of the identity matrix) occurs when every observable history is also a true history in which there is no misidentification. Provided these assumptions hold,

then we can use the algorithm in Figure 1 with a suitable lattice basis $\mathcal{L}$ and connect the fiber. The proof of this theorem is provided in the supplementary materials, along with a description of how to construct the lattice (Markov) basis $\mathcal{L}$.

The conditions of Theorem 1 are satisfied for model $\mathrm{M}_{t\alpha}$, so that for $K = 2$ we obtain the Markov basis in (5)

	x_{00}	x_{01}	x_{02}	x_{10}	x_{11}	x_{12}	x_{20}	x_{21}	x_{22}
MB1	1	0	0	0	0	0	0	0	0
MB2	0	−1	1	0	0	0	0	0	0
MB3	0	−1	0	−1	0	1	0	0	0
MB4	0	0	0	−1	0	0	1	0	0
MB5	0	−1	0	−1	0	0	0	1	0
MB6	0	−1	0	−1	0	0	0	0	1

(5)

The basis in (5) is identical to that presented by L2010 for model $\mathrm{M}_{t\alpha}$ when $K = 2$.

The approach of L2010 to finding a basis involves choosing pivotal (or constraining) variables when solving the set of equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ (a full description is available either in L2010, pg 180–181, or in the supplementary materials). L2010 chose specific pivotal variables (x_{01}, x_{10} and x_{11}) when finding the basis for model $\mathrm{M}_{t\alpha}$ when $K = 2$. However, it was implied that this choice was arbitrary and no guidance was given as to how to select pivotal variables when $K > 2$. It turns out that changing the pivotal variables can lead to different sets of basis vectors which may not be Markov bases. We show in the supplementary materials that for $K = 2$ and a different set of pivotal variables, x_{22}, x_{20} and x_{11}, the resulting basis differs from that in (5). We also show that when the conditions of Theorem 1 are satisfied, there is a specific choice of pivotal variables guaranteed to return the Markov basis $\mathcal{L}$. In particular, if we order $\boldsymbol{x}$ as in L2010 for model $\mathrm{M}_{t\alpha}$ and take the variable corresponding to the leading non-zero entry in each row of $\boldsymbol{A}$ as pivotal (as was done by L2010 for $K = 2$), the basis found will be the Markov basis $\mathcal{L}$.

Theorem 1 ensures that there is at least one lattice basis which is also a Markov basis for model $\mathrm{M}_{t\alpha}$. However, it does not imply that every lattice basis is a Markov basis. For model

$M_{t\alpha}$ and $K = 2$ another lattice basis (found by hand) is given in (6)

	x_{00}	x_{01}	x_{02}	x_{10}	x_{11}	x_{12}	x_{20}	x_{21}	x_{22}
LB1	1	0	0	0	0	0	0	0	0
LB2	0	−1	1	0	0	0	0	1	−1
LB3	0	0	1	0	0	0	1	0	−1
LB4	0	0	0	1	0	0	−1	0	0
LB5	0	0	−1	0	0	1	−1	1	−1
LB6	0	0	0	0	0	0	0	1	−1

(6)

Suppose the observed data are $\boldsymbol{y} = (363, 22, 174)$ (as in L2010), then two elements in the fiber are $\boldsymbol{x}_1 = (0, 363, 0, 22, 174, 0, 0, 0, 0)'$ and $\boldsymbol{x}_2 = (0, 361, 2, 22, 174, 0, 0, 0, 0)'$. We are unable to move between these two using LB1 – LB6 in (6) as moves in the algorithm in Figure 1. In particular, if we start at (the observed history) $\boldsymbol{x}_1$ the moves LB2, LB3, LB5 and LB6 will lead to automatic rejections because they will always propose a negative value. This means that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not connected and thus the fiber is not connected.

We repeated the analysis of L2010 using both the Markov basis in (5) and the lattice basis in (6) using the same prior distributions as in L2010 (we used only one of the priors L2010 considered for α; a beta distribution with parameters 19 and 1). In both cases we implemented the algorithm in Figure 1 using $\boldsymbol{x}_1$ as the starting value with interest in the abundance N. We checked convergence via trace plots and plotted the resulting distribution for $N|y$ in both cases (Figure 3). The lattice basis in (6) leads to a distribution for N that is substantially different from the true posterior distribution and could lead to incorrect decision making.

[Figure 3 about here.]

We note that efficiency gains can be made if there are observable histories with zero count. In particular, we can delete the entries in $\boldsymbol{y}$ and the rows of $\boldsymbol{A}$ corresponding to the zero counts before deleting any columns of $\boldsymbol{A}$ and corresponding entries of $\boldsymbol{x}$ that are known to have zero count. Provided the assumptions of Theorem 1 are still satisfied by the resulting

configuration matrix then we can still find a set of moves guaranteed to connect all elements in the fiber. The resulting set of moves is no longer a Markov basis but a Markov subbasis (Chen et al. 2006) as it is only valid for the observed $\boldsymbol{y}$. This corresponds to the approach taken by both Bonner and Holmberg (2013) and McClintock et al. (2013) for data with multiple marks that could not be matched.

This section shows that we must take care even with simple corruptions to ensure that the lattice basis we are using is also a Markov basis. The following two sections give examples where we do not have simple corruptions (in one of these it does not even make sense to think of corruptions in the sense of model $\mathrm{M_{t\alpha}}$) and a Markov basis has greater cardinality than a lattice basis.

4. Example: Sufficient Statistics

Next we consider the problem of modeling data from a closed population when sufficient statistics from one or more models are provided in place of the raw data. The raw data may not be available for a variety of reasons, e.g. privacy concerns. Here we assume that the population is closed and that we have the sufficient statistics associated with three commonly used models $\mathrm{M_t}$, $\mathrm{M_b}$ and $\mathrm{M_h}$ (Otis et al. 1978). From model $\mathrm{M_h}$ we have the statistics $f_1, \ldots, f_K$, where f_j is the number of individuals who were caught j times from a total of K sampling occasions; from model $\mathrm{M_t}$ we have the statistics $n_1, \ldots, n_K$, where n_j is the number of individuals captured in the jth sample; and from model $\mathrm{M_b}$ we have the statistic $M_\cdot = \sum_{j=1}^{t} M_j$, with M_j the number of marked individuals in the population in sample j. Note that we do not include the other sufficient statistics for model $\mathrm{M_t}$ and $\mathrm{M_b}$ noted by Otis et al. (1978) as they are deterministic functions of $f_1, \ldots, f_K$.

All of these statistics are linear functions of the data which means that this problem can be expressed using the linear constraint in (1). In this example, $\boldsymbol{x}$ represents the vector of counts for the $2^K - 1$ true histories; $\boldsymbol{y}$ represents the vector of counts for the $2K + 1$ sufficient

statistics; and the configuration matrix, $\boldsymbol{A}$, is a $(2K+1) \times (2^K - 1)$ matrix. Details of how to find $\boldsymbol{A}$ along with an example for a study with $K = 4$ occasions are provided in the supplementary materials.

Here we explore this scenario using multi-list data from a South Auckland, New Zealand, diabetes study from the Ph.D. research of Huakau (2001) and included in the Ph.D. research of Sutherland (2003). We ignore the potential errors in matching individuals between lists and assume that each individual is correctly matched (see Lee (2002) for how such errors could also be accounted for using the linear constraint (1)). There are $K = 4$ lists: general practitioners records (G), pharmacy records (P), outpatient records (O) and inpatient discharge records (D) that we assume are ordered as written. We use the data for males and reduce the full data (which is available in Sutherland 2003) to the statistics: $\boldsymbol{n} = (n_G, n_P, n_O, n_D)' = (629, 622, 6279, 1623)'$, $\boldsymbol{f} = (f_1, f_2, f_3, f_4)' = (6030, 1312, 161, 4)'$ and $M_\cdot = 8680$ to give

$$\boldsymbol{y} = (6030, 1312, 161, 4, 629, 622, 6279, 1623, 8680)'.$$

As well as $\boldsymbol{y}$ being sufficient for models $\mathrm{M_t}$, $\mathrm{M_h}$ and $\mathrm{M_b}$, it is also sufficient for the two-factor quasi-symmetric version of model $\mathrm{M_{th}}$ that is induced by a Rasch model (see Agresti 1994, for details of this model).

The vector $\boldsymbol{x}$ is indexed by $\boldsymbol{\omega} = (\omega_G, \omega_P, \omega_O, \omega_D)$, where $\omega_j = 1$ denotes inclusion on list j with $\omega_j = 0$ otherwise, so that x_{1101} is the number of individuals on lists G, P and D and not on list O. Our focus here is to attempt to make inference about x_{1000}, the number of individuals who appear only in list G. We may also wish to fit a model to $\boldsymbol{x}$ for which $\boldsymbol{y}$ are not sufficient statistics. By definition, the resulting model would be nonidentifiable, but this does not necessarily mean that there is no information about parameters of this model, including the abundance N. The latent multinomial model can be used in either of these situations.

A lattice basis found using the Hermite normal form is

	x_{0001}	x_{0010}	x_{0011}	x_{0100}	x_{0101}	x_{0110}	x_{0111}	x_{1000}	x_{1001}	x_{1010}	x_{1011}	x_{1100}	x_{1101}	x_{1110}	x_{1111}
LB1	0	0	0	0	0	0	0	0	−1	0	1	1	0	−1	0
LB2	0	0	0	0	0	0	0	0	−1	1	0	0	1	−1	0
LB3	0	−1	1	0	0	0	0	1	−1	0	0	0	0	0	0
LB4	1	−2	0	1	0	0	0	0	0	0	1	0	−2	1	0
LB5	1	−2	0	0	1	0	0	1	−1	0	1	0	−2	1	0
LB6	1	−2	0	0	0	1	0	1	−1	0	1	0	−1	0	0
LB7	1	−2	0	0	0	0	1	1	0	0	0	0	−2	1	0

Using the seven moves LB1 – LB7 in the algorithm in Figure 1 it is impossible to move between the two solutions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

$$\boldsymbol{x}_1 = (652, 4865, 794, 253, 18, 234, 62, 260, 26, 221, 67, 19, 0, 32, 4)'$$

$$\boldsymbol{x}_2 = (684, 4901, 694, 253, 31, 154, 161, 192, 49, 365, 0, 19, 0, 0, 4)'.$$

If we are currently at $\boldsymbol{x}_2$, it is clear that all moves (except LB3) will lead to at least one negative cell count and will be automatically rejected. The vector LB3 can be used to update $\boldsymbol{x}_2$, but we are unable to get to $\boldsymbol{x}_1$ using LB3 alone. Again, we have at least two sets of elements in the fiber that we can move within, but are unable to move between.

A Markov basis for this problem can be constructed in `4ti2` and is made up of the 16 elements given in the supplementary materials. Since (i) `4ti2` finds a minimal Markov basis, and (ii) the cardinality of the Markov basis is larger than that of a lattice basis, we can be certain that a lattice basis can never be a Markov basis for this problem. Even though it is likely possible to construct another lattice basis that can move between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ there will be either (i) another two elements in the fiber that are not connected, or (ii) another two elements in the fiber for a different $\boldsymbol{y}$ that we cannot move between with such a lattice basis.

Here we fit model $\mathrm{M_t}$ and run the algorithm in Figure 1 with both the Markov basis given in the supplementary materials and the lattice basis specified above (details of the model are given in the supplementary materials). We make use of the factorization theorem (e.g, see Casella and Berger 2002, pg. 276) that states that a model $f(\boldsymbol{x}|\boldsymbol{\theta})$ with sufficient statistics

$\boldsymbol{y}$ can be expressed as

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{x}|\boldsymbol{y})h(\boldsymbol{y}|\boldsymbol{\theta}).$$

A practical implication is that only $g(\boldsymbol{x}|\boldsymbol{y})$ is required if interest is in a function of $\boldsymbol{x}$ such as x_{1000}, and the parameters $\boldsymbol{\theta} = (N, p_1, \ldots, p_K)$ need not be specified. A related implication is that if we do choose to update $\boldsymbol{\theta}$ the resulting chains will converge to the correct posterior $[\boldsymbol{\theta}|\boldsymbol{y}]$ even if we (i) do not update $\boldsymbol{x}$, or (ii) update $\boldsymbol{x}$ using a set of moves that is unable to connect the fiber, such as the lattice basis above; provided we specify an appropriate MCMC sampler for $\boldsymbol{\theta}$.

Using the lattice basis and starting at $\boldsymbol{x}_2$ the resulting distributions for x_{1000} are qualitatively different from the posterior distribution found using the Markov basis even though the individual chains appear to have converged to the stationary distribution (Figure 4). The true value of $x_{1000} = 260$ has some posterior mass when using a Markov basis (despite being in the tail). If we were to believe the results when using the lattice basis $x_{1000} = 260$ is so far in the tail, we would conclude it has negligible posterior mass.

[Figure 4 about here.]

5. Example: Band Misreading in Mark-Resight

As a final example we consider a mark-resight model which allows for the possibility that individuals are misidentified when resighted in the field. Imagine that there are K_1 distinct occasions, on which researchers capture a number of unmarked individuals, mark them, and release them back into the population. Along with that are a series of K_2 resighting occasions, on which the researchers conduct visual surveys to identify previously marked individuals. Data from the experiment consist of the observed resighting histories for each individual. If there were no errors then standard mark-resight models could be used to estimate survival or movement rates (e.g. Hestbeck et al. 1991); or abundance (e.g. McClintock et al. 2006).

Suppose now that individuals may be misidentified when they are resighted. In direct contrast to model $M_{t\alpha}$, which assumes that errors are unique and never match other individuals, we assume that errors may be repeated and always match the identity of previously marked individuals. The justification for this assumption is that the available set of marks is known on each occasion when individuals are identified by man-made marks instead of natural markers (e.g., genotypes or photo-id). Erroneous sightings of marks which have not been released can then be identified and removed from the data prior to the analysis. The only time an error cannot be detected and discarded is when one previously marked individual is misidentified as another previously marked individual. We note that removal of erroneous sightings is only justified when estimating survival. Removing erronous sightings when including unmarked individuals would lead to biased estimators of abundance (McClintock et al. 2014).

For the remainder of the section, we assume that the capture and resighting occasions occur simultaneously so that $K = K_1 = K_2$. The true capture histories for each individual can now be constructed in terms of four possible events. On each occasion, individual i may be:

- not captured or resighted (event 0),
- captured or resighted and correctly identified (event 1), or
- resighted and incorrectly identified (event 2).

Further to this, another individual may be resighted and incorrectly identified as individual i (event 3). Events 2 and 3 represent false negative and false positive resightings. For example, the history 123 for individual i would indicate that i was captured and marked on the first occasion, was resighted and misidentified on the second occasion, and that another individual was resighted and identified as i on the third occasion of a study with $K = 3$ occasions. To simplify the example, we assume that individuals cannot be misidentified when they are first captured and that multiple events involving the same individual cannot occur on a single

occasion (e.g., it is not possible to resight i and incorrectly identify another individual as i on the same occasion). This assumption may be unrealistic in some situations and was made to make the approach tractable. Developing methodology to relax this assumption is ongoing research.

For an experiment with K occasions, the model has $(4^K - 1)/3$ possible true histories and the usual $2^K - 1$ observable histories. Further to this, there are $K - 1$ extra constraints that equate the number of false negatives and false positives (2s and 3s) on occasions 2 through K. As a result, $\boldsymbol{A}$ has dimension $(2^K + K - 2) \times (4^K - 1)/3$ and a basis for $\ker_{\mathbb{Z}}(\boldsymbol{A})$ has $(4^K - 1)/3 - (2^K + K - 2)$ elements.

To make this more concrete, we consider the specific case of an experiment comprising $K = 3$ occasions. In this case, there are $(4^3 - 1)/3 = 21$ possible true histories, $2^3 - 1 = 7$ observable histories, and $3 - 1 = 2$ extra constraints on the number of false positive and negative resightings (2s and 3s) on occasions 2 and 3. Details of how to construct $\boldsymbol{A}$ along with $\boldsymbol{x}$ and $\boldsymbol{y}$ for a study with $K = 3$ capture occasions are provided in the supplementary materials. In this case, a basis for $\ker_{\mathbb{Z}}(\boldsymbol{A})$ has 12 elements and the specific lattice basis obtained using the Hermite normal form is provided in the supplementary materials, along with the Markov basis, computed using `4ti2`, that has 63 elements.

To illustrate the problems that can occur with this model we first consider the analysis of a single (fake) data set. Suppose that each observable history is recorded one time so that

$$\boldsymbol{y} = (1, 1, 1, 1, 1, 1, 1).$$

An exhaustive search confirms that the fiber defined by $\boldsymbol{y}$ contains exactly 120 unique elements. However, the lattice basis given in the supplementary materials does not connect all of the elements in the fiber. Instead, the lattice basis divides the fiber into two distinct pieces including a large set of 87 connected elements; and a further set of 33 isolated elements which connect to nothing else. As a result, the distribution of the sample generated by the

algorithm in Figure 1 using the elements of the lattice basis in the supplementary materials as moves will depend on the starting point.

To show this, we have investigated the output from the algorithm in Figure 1 when using a lattice basis as our set of moves. We have chosen a starting point that lies in the largest part of the fiber and connects with 86 other elements:

$$\boldsymbol{x}_1 = (1,0,1,1,0,1,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0)'.$$

Assuming a multinomial distribution for $[\boldsymbol{x}|\boldsymbol{\theta}]$ is not appropriate to account for the band misreading process and specification of a more complex $[\boldsymbol{x}|\boldsymbol{\theta}]$ is ongoing research. As our goal is to show that a lattice basis is unable to connect the fiber, we simplify the model by setting $[\boldsymbol{x}|\boldsymbol{\theta}] \propto 1$. A valid sampler should then sample uniformly from the 120 elements in the fiber. For comparison, we have also run a chain using the full Markov basis starting at $\boldsymbol{x}_1$. As expected, the first chain visits 87 unique solutions and the second visits all 120. To visualize the impact this can have on inference, Figure 5 compares the distributions of the number of errors in the solutions identified by each chain. Using the lattice basis, the first chain oversamples the solutions with too few errors, placing too much mass on solutions with one or two errors and not enough on solutions with three, four, or five errors. In comparison, the distribution generated using the full Markov basis matches the true distribution of the number of errors in the 120 elements almost exactly.

[Figure 5 about here.]

6. Discussion

Here we have presented examples of capture-recapture models that show the importance of using a Markov basis when sampling from a linearly constrained vector of counts. In particular, we have demonstrated the danger of using elements of a lattice basis as one-at-a-time moves in an algorithm as in Figure 1. In many situations a set referred to as a Markov

basis is needed to ensure we can move between various elements of the fiber without passing through invalid (negative) counts. Even when a Markov basis is a lattice basis, we must take care because not every lattice basis is a Markov basis.

For a given matrix $\boldsymbol{A}$ the need for a Markov basis over a simpler lattice basis depends on the lattice basis chosen, as well as the data observed. If we consider the lattice basis for the 3×3 contingency table in section 2, difficulties arose because our data had a row sum of 0. A related issue is that even when a lattice basis is unable to connect the fiber, it may still be able to connect nearly all elements in the fiber. In such a case, using a lattice basis may lead to a distribution that is an acceptable approximation of the true posterior distribution. This is especially the case if the elements of the fiber that are not connected to the initial value are in areas of low probability in the model $[\boldsymbol{x}|\boldsymbol{\theta}]$. This can be seen in the example from Section 4: using the lattice basis and starting at the second starting value (Figure 2; right panel) results in an estimated posterior density that is practically indistinguishable from the true posterior distribution (Figure 4). However, there is no guarantee that any given lattice basis will provide a good approximation to the fiber. It is possible that even with multiple starting values we may choose values that only connect a small proportion of the fiber.

One important aspect that we have only briefly mentioned is the difficulty in constructing Markov bases. For the purposes of this manuscript we have overcome this difficulty through (i) analytical results, or (ii) the use of the software package `4ti2` (Hemmecke et al. 2013). While the latter is possible for the examples we explored, it is unable to evaluate a Markov basis for some capture-recapture examples with a moderate to large number of sampling occasions. For example, `4ti2` was unable to compute a Markov basis (on the lead authors work machine) for the band read error model in section 5 for $K > 4$. If we were to use `4ti2` for model $\mathrm{M}_{t\alpha}$ (ignoring the theorem presented in section 3), `4ti2` was unable to compute a Markov basis for $K > 5$. The implication of this is that for an algorithm in the spirit of

Figure 1 to be implemented for problems not involving simple corruptions, methodological work is likely to be necessary to ensure a potential set of moves is a Markov basis.

Several alternative algorithms and methods have been proposed for sampling from the fiber that avoid the calculation of a full Markov basis. We anticipate that such approaches may be useful for a range of capture-recapture examples. These include independent sampling of elements of the fiber (e.g., see Chen et al. 2005), extending the algorithm to allow limited travel through vectors $\boldsymbol{x}$ that contain negative values while using a set of moves that is not guaranteed to connect the fiber (e.g., see Bunea and Besag 2000) and approaches that dynamically find a Markov basis as the algorithm runs (e.g., see Dobra 2012). While promising, we expect these approaches will require adapting to the particular challenges faced in problems involving misidentification in capture-recapture data.

Acknowledgements

We thank three annonymous referees for comments on a previous version of this manuscript as well as Ruriko Yoshida for participating in valuable discussion on algebraic statistics.

Supplementary Materials

Web Appendices referenced in Sections 2, 3, 4 and 5 are available with this paper at the Biometrics website on Wiley Online Library.

References

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics **50,** 494–500.

Amstrup, S. C., McDonald, T. L., and Manly, B. F. J. (2005). Handbook of Capture-Recapture Analysis. Princeton University Press.

Aoki, S., Hara, H., and Takemura, A. (2012). Markov Bases in Algebraic Statistics. Springer.

Bonner, S. J. and Holmberg, J. (2013). Mark-recapture with multiple non-invasive marks. Biometrics **69,** 766–775.

Bunea, F. and Besag, J. (2000). MCMC in $i \times j \times k$ contingency tables. Fields Institute Communications **26,** 23–36.

Casella, G. and Berger, R. L. (2002). Statistical Inference. Duxbury Pacific Grove, CA.

Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. Journal of the American Statistical Association **100,** 109–120.

Chen, Y., Dinwoodie, I., and Sullivant, S. (2006). Sequential importance sampling for multiway tables. The Annals of Statistics **34,** 523–545.

Cox, D., Little, J., and O'Shea, D. (2007). Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra. Springer.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. The Annals of Statistics **26,** 363–397.

Dobra, A. (2012). Dynamic Markov bases. Journal of Computational and Graphical Statistics **21,** 496–517.

Drton, M., Sturmfels, B., and Sullivant, S. (2009). Lectures on Algebraic Statistics. Springer.

Fienberg, S. E. and Manrique-Vallier, D. (2009). Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. AStA Advances in Statistical Analysis **93,** 49–60.

Hemmecke, R., Hemmecke, R., Koeppe, M., Malkin, P., and Walter, M. (2013). User's guide for `4ti2` version 1.6.

Hestbeck, J. B., Nichols, J. D., and Malecki, R. A. (1991). Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. Ecology **72,** 523–533.

Huakau, J. T. (2001). New methods for analysis of epidemiological data using

capture-recapture methods. PhD thesis, The University of Auckland.

Karwa, V. and Slavkovic, A. (2013). Conditional inference given partial information in contingency tables using Markov bases. Wiley Interdisciplinary Reviews: Computational Statistics **5,** 207–218.

King, R., Bird, S. M., Hay, G., and Hutchinson, S. J. (2009). Estimating current injectors in Scotland and their drug-related death rate by sex, region and age-group via Bayesian capture-recapture methods. Statistical Methods in Medical Research **18,** 341–359.

Lee, A. (2002). Effect of list errors on the estimation of population size. Biometrics **58,** 185–191.

Lee, A. J., Seber, G. A. F., Holden, J. K., and Huakau, J. T. (2001). Capture–recapture, epidemiology, and list mismatches: several lists. Biometrics **57,** 707–713.

Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. Biometrics **66,** 178–185.

Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. The American Statistician **67,** 191–200.

McClintock, B. T., Conn, P., Alonso, R., and Crooks, K. R. (2013). Integrated modeling of bilateral photo-identification data in mark-recapture analyses. Ecology **94,** 1464–1471.

McClintock, B. T., Hill, J. M., Fritz, L., Chumbley, K., Luxa, K., and Diefenbach, D. R. (2014). Mark-resight abundance estimation under incomplete identification of marked individuals. Methods in Ecology and Evolution **5,** 1294 – 1304.

McClintock, B. T., White, G. C., and Burnham, K. P. (2006). A robust design mark-resight abundance estimator allowing heterogeneity in resighting probabilities. Journal of Agricultural, Biological, and Environmental Statistics **11,** 231–248.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference

565 from capture data on closed animal populations. Wildlife Monographs **62,** 1–135.

566 Seber, G. A., Huakau, J. T., and Simmons, D. (2000). Capture-recapture, epidemiology, and
567 list mismatches: two lists. Biometrics **56,** 1227–1232.

568 Sutherland, J. and Schwarz, C. (2005). Multi-list methods using incomplete lists in closed
569 populations. Biometrics **61,** 134–140.

570 Sutherland, J. M. (2003). Multi-list methods in closed populations with stratified or
571 incomplete information. PhD thesis, Simon Fraser University.

572 Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson,
573 D. M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for
574 estimating abundance using DNA samples. Biometrics **65,** 833–840.

575 Yoshizaki, J., Brownie, C., Pollock, K. H., and Link, W. A. (2011). Modeling misidentification
576 errors that result from use of genetic tags in capture–recapture studies. Environmental
577 and Ecological Statistics **18,** 27–55.

578 Yoshizaki, J., Pollock, K. H., Brownie, C., and Webster, R. A. (2009). Modeling misidenti-
579 fication errors in capture-recapture studies using photographic identification of evolving
580 marks. Ecology **90,** 3–9.

Received MMM 2015. *Revised MMM* 2015.

Accepted MMM 2015.

1: Initialize $\boldsymbol{x}^0$ so that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^0$
2: **for** $i = 1 : n$ **do**
3: Sample $k \in \{1, 2, \ldots, m\}$ with equal probability
4: Sample $c \in \{-1, 1\}$ with equal probability
5: Set $\boldsymbol{x}_{\text{cand}} = \boldsymbol{x}^{i-1} + c\boldsymbol{a}_k$
6: Calculate the metropolis acceptance probability: $r = \min\left(1, \frac{[\boldsymbol{x}_{\text{cand}}|\theta]}{[\boldsymbol{x}^{i-1}|\theta]}\right)$
7: Accept $\boldsymbol{x}_{\text{cand}}$ with probability r (if accepted $\boldsymbol{x}^i = \boldsymbol{x}_{\text{cand}}$; otherwise $\boldsymbol{x}^i = \boldsymbol{x}^{i-1}$)
8: **end for**

Figure 1. Algorithm for updating the latent counts $\boldsymbol{x}$. The value n is the number of iterations in the algorithm and the vectors $\mathcal{B} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ are a subset of the kernel of $\boldsymbol{A}$.

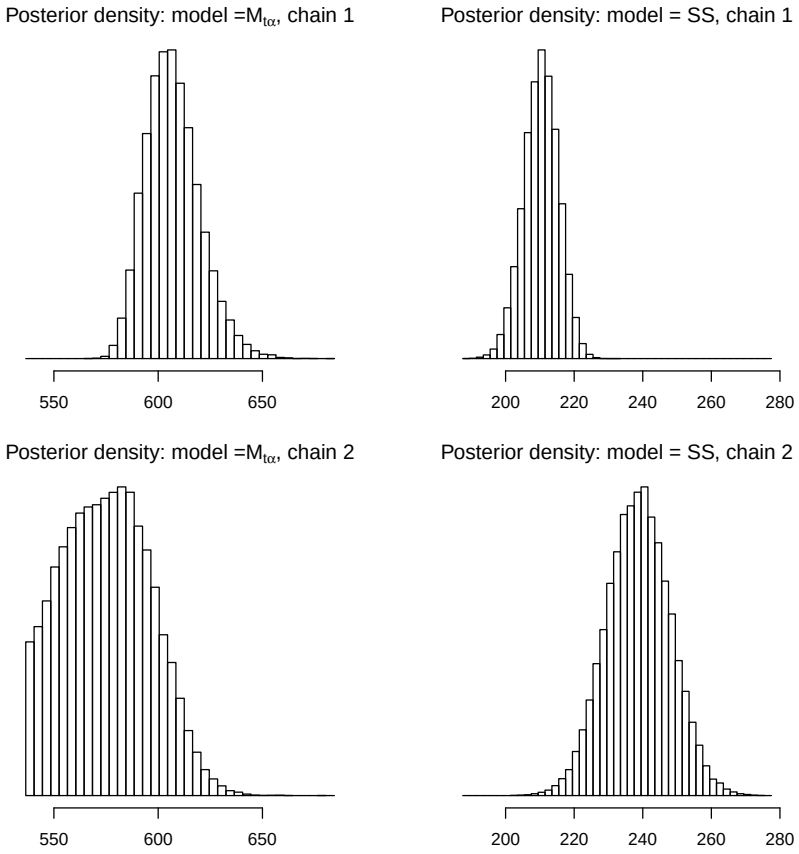

Figure 2. Estimated posterior densities of a quantity of interest for model $M_{t\alpha}$ (left panel) and a multi-list model where summary statistics are presented in place of full data (SS; right panel). Within each model, the resulting density estimates are plotted separately from the output of two parallel MCMC algorithms (for each model) with different starting values.

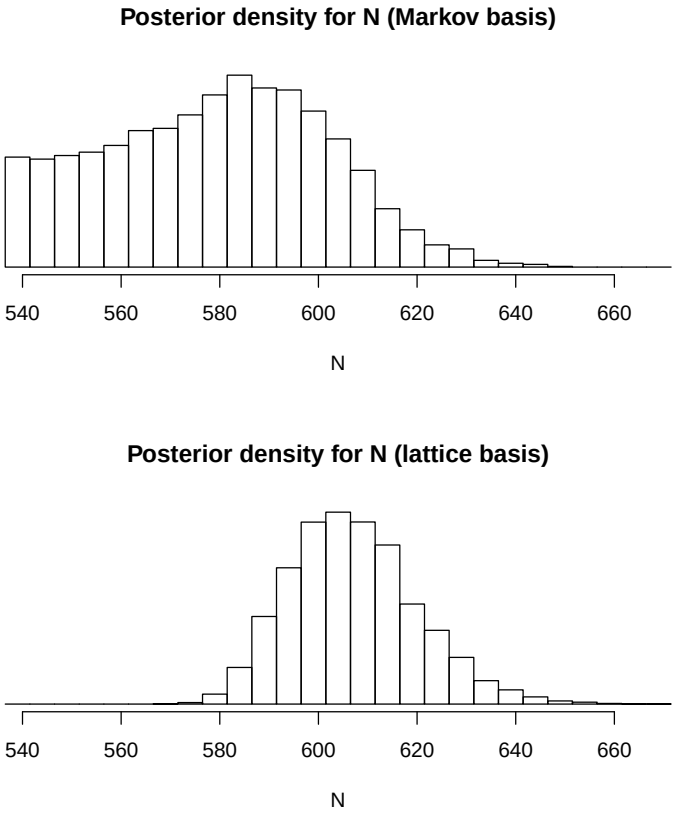

Figure 3. Histograms of the estimated posterior density of $N|y$ when using the Markov basis from (5) (top) and the lattice basis from (6) (bottom) when starting from $\boldsymbol{x}_1$.

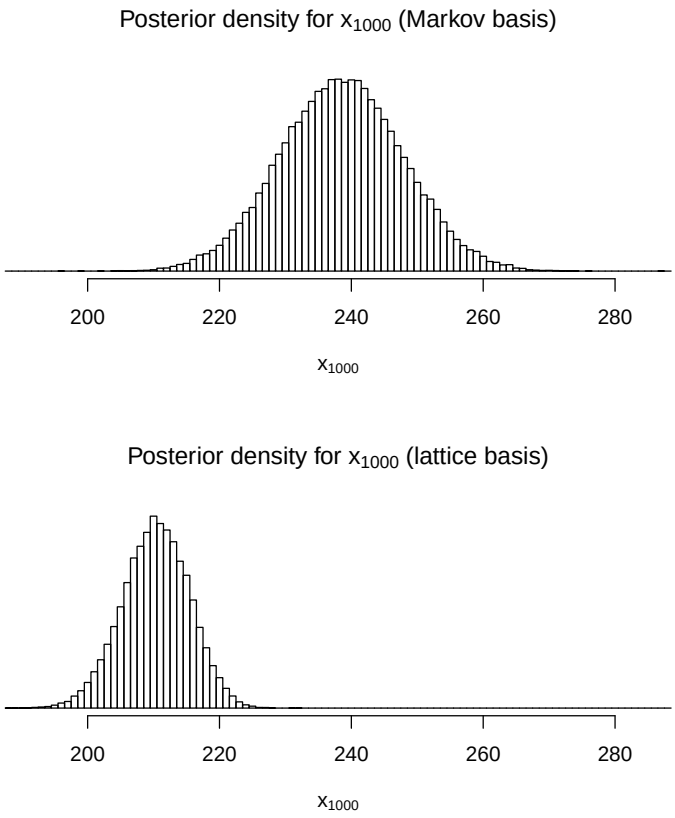

Figure 4. Posterior densities of x_{1000} when using the Markov basis from the supplementary materials (top) and the lattice basis specified in section 4 (bottom) when starting at $\boldsymbol{x}_2$.

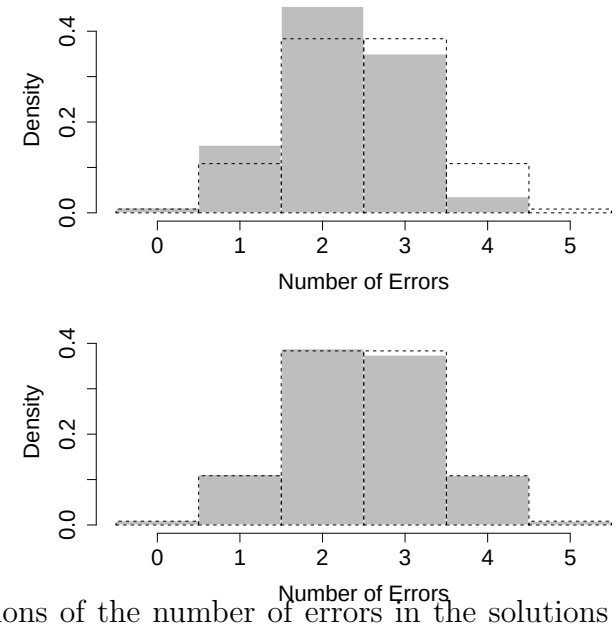

Figure 5. Distributions of the number of errors in the solutions sampled given the data $\boldsymbol{y}$. The top histogram illustrates the distribution generated using the lattice basis with the starting value $\boldsymbol{x}_1$. The bottom plot illustrates the distribution obtained using the full Markov basis with the same starting value. In each plot, the gray bars represent the distribution of the number of errors while the dashed bars represent the true distribution over all 120 unique solutions.