

Fibers of multi-way contingency tables given conditionals: relation to marginals, cell bounds and Markov bases

Aleksandra Slavković · Xiaotian Zhu ·
Sonja Petrović

Received: 6 November 2012 / Revised: 6 January 2014 / Published online: 25 June 2014
© The Institute of Statistical Mathematics, Tokyo 2014

Abstract A fiber of a contingency table is the space of all realizations of the table under a given set of constraints such as marginal totals. Understanding the geometry of this space is a key problem in algebraic statistics, important for conducting exact conditional inference, calculating cell bounds, imputing missing cell values, and assessing the risk of disclosure of sensitive information. Motivated by disclosure problems, in this paper we study the space of all possible tables for a given sample size and set of observed conditional frequencies. We show that this space can be decomposed according to different possible marginals, which, in turn, are encoded by the solution set of a linear Diophantine equation. Our decomposition has two important consequences: (1) we derive new cell bounds, some including connections to directed acyclic graphs, and (2) we describe a structure for the Markov bases for the given space that leads to a simplified calculation of Markov bases in this particular setting.

Keywords Conditional tables · Contingency tables · Diophantine equations · Disclosure limitation · Directed acyclic graphs · Marginal tables · Markov bases · Optimization for cell entries

A. Slavković (✉) · X. Zhu
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: sesa@stat.psu.edu

X. Zhu
e-mail: xxz131@psu.edu

S. Petrović
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
e-mail: Sonja.Petrovic@iit.edu

1 Introduction

In [Dobra et al. \(2008\)](#), the authors use tools from algebraic statistics to study two related problems: maximum likelihood estimation for log-linear models in multi-way contingency tables, and disclosure limitation strategies to protect against the identification of individuals associated with small counts in the tables; for an overview of disclosure limitation literature see [Doyle et al. \(2001\)](#) and [Hundepool et al. \(2012\)](#). These are linked to the general problem of inference in tables for which only partial information is available (e.g., see [Dobra et al. 2006](#); [Thibaudeau 2003](#); [Marjoram et al. 2003](#)).

Incomplete data commonly arise in surveys or census data which have been modified to limit disclosure of sensitive information. Instead of releasing complete data, summary statistics are often released, even if they may not be the sufficient statistics for the probability model. Examples of summary statistics are marginal tables, or tables of conditional frequencies, e.g., [Slavković \(2009\)](#). Given a set of released statistics, there are a number of ways to assess the disclosure risk and data utility, including computing bounds for cell entries, enumerating all table realizations, and sampling from a fiber to estimate posterior distributions. A fiber is the space of all possible tables consistent with the observed statistics. Since the fibers form the support of the conditional distributions given a set of summary statistics, their properties are important for conducting exact conditional inference; e.g., see [Diaconis and Sturmfels \(1998\)](#) for an algebraic statistics approach to goodness-of-fit testing given the marginal totals, and [Dobra and Fienberg \(2010\)](#) for calculating bounds on the cell entries. Similar techniques that rely on understanding fibers' structure can be used to impute missing data in contingency tables and to create replacement tables; see [Slavković and Lee \(2010\)](#), with focus on tables that arise from preserving conditional frequencies.

In this paper, we study the sample space of contingency tables given observed conditional frequencies and their relations to corresponding marginals. More specifically, we address the following challenge:

Problem 1 (Problem 5.7 in [Dobra et al. 2008](#)) Characterize the difference of two fibers, one for a conditional probability array, and the other for the corresponding margin, and thus simplify the calculation of Markov bases for the conditionals by using the knowledge of the moves of the corresponding marginals.

Here is a general setup. Consider r categorical random variables, X_1, \dots, X_r , where each X_i takes values in the finite set of categories $[d_i] \equiv \{1, \dots, d_i\}$. Let $\mathcal{D} = \bigotimes_{i=1}^r [d_i]$, and $\mathbb{R}^{\mathcal{D}}$ be the vector space of r -dimensional arrays of format $d_1 \times \dots \times d_r$, with a total of $d = \prod_i d_i$ entries. The cross-classification of n independent and identically distributed realizations of (X_1, \dots, X_r) produces a random integer-valued array $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, called a r -way contingency table, whose coordinate entry n_{i_1, \dots, i_r} is the number of times the label combination, or cell, (i_1, \dots, i_r) is observed in the sample (see [Agresti 2002](#); [Bishop et al. 2007](#); [Lauritzen 1996](#) for details). It is often convenient to order the cells in some prespecified way (e.g., lexicographically).

Let A and B be proper subsets of $\{X_1, X_2, \dots, X_r\}$, and $C = \{X_1, X_2, \dots, X_r\} \setminus (A \cup B)$. We can regard A , B and C as three categorical variables with levels A_1, \dots, A_J , B_1, \dots, B_J , and C_1, \dots, C_K . Thus, we can summarize the r -way table \mathbf{n} as a 3-way

table $\mathbf{n}^* := \{s_{ijk}\}$, where s_{ijk} is the count in the cell (A_i, B_j, C_k) . Finally, let c_{ij} be the observed conditional frequency $P(A = i|B = j)$, such that $\sum_i P(A = i|B = j) = 1$. If C is an empty set, we refer to c_{ij} 's as full conditionals, otherwise as small or partial conditionals.

Motivated by Problem 1, we investigate the fiber $\mathcal{F}_{\mathcal{T}}$ for $\mathcal{T} = \{P(A|B), N\}$, that is the space of all possible tables consistent with:

- (a) the observed grand total, $\sum_{i_1 \dots i_r} n_{i_1 i_2 \dots i_r} = N$, and
- (b) a set of observed conditional frequencies, $P(A|B)$.

Note that we do not observe the values of B , and we assume that all of the given frequencies are exact. Then, the space $\mathcal{F}_{\mathcal{T}}$ is the set of integer solutions to the following system of linear equations

$$\left\{ \begin{array}{l} M\mathbf{n} = \mathbf{t} \\ \text{every B marginal} > 0 \end{array} \right\}, \tag{1}$$

where \mathbf{n} and \mathbf{t} are length d column vectors, and M is a $(J + 1) \times d$ matrix that, together with \mathbf{t} , describes the information encoded by the grand total and the given frequencies. When N is clear from the context, we use the shorthand notation $\mathcal{F}_{A|B}$ to denote $\mathcal{F}_{\{P(A|B), N\}}$. The space of tables given the $[AB]$ marginal counts s_{ij+} is denoted by \mathcal{F}_{AB} . For a concrete example, see Sect. 4.1.

The main contributions of this manuscript come from the structural results for the fibers defined above. In particular, we solve a generalization of an open problem posed by Dobra et al. (2008). In Corollary 1 we give conditions for when the two fibers $\mathcal{F}_{A|B}$ and \mathcal{F}_{AB} agree. A decomposition of the table space $\mathcal{F}_{A|B}$ is given in Corollary 2, showing that the space of tables given the conditional is a disjoint union of spaces of tables given distinct marginals. This decomposition of $\mathcal{F}_{\mathcal{T}}$ leads to three important applied results: (1) in Sect. 2.3, we derive new results on computing the exact and approximate cardinality of the given fibers and provide functions to do this in \mathbb{R} , (2) in Sect. 3.1, we derive new cell bounds, some including connections to directed acyclic graphs in Sect. 3.3, and (3) in Sect. 3.2, we describe a structure for the Markov bases for the space $\mathcal{F}_{\mathcal{T}}$ that leads to a simplified calculation of Markov bases in this setting. In Sect. 4, we demonstrate our theoretical results with a series of simple examples and conclude with a brief discussion in Sect. 5.

2 The space of tables with given conditional frequencies

Data examples suggest a connection between the solutions to a Diophantine equation defined below in Eq. (2), and the space of tables $\mathcal{F}_{A|B}$ that we are interested in. Moreover, this connection appears in symbolic computation: points in the fiber are lattice points in polytopes, and their connection to Diophantine equations has a history in mathematics Loera et al. (2004). In what follows, we establish this connection more rigorously from the point of view of marginal and conditional tables. Finding solutions to Diophantine equations is a well-studied classical problem in mathematics, one that is generally hard to solve and with a number of proposed algorithms; (e.g., see Morito

and Salkin 1980; Chen and Li 2007; Eisenbeis et al. 1992; Smarandache 2000), and references therein. But the Eq. (2) here is simple enough that can be analyzed using classical algebra, and as such affords implementations of simple functions in R needed for statistical analyses. Throughout, we use the notation established in Sect. 1.

2.1 Table space decomposition

The table of observed conditional frequencies gives rise to a linear Diophantine Eq. (2) whose solutions correspond to possible marginals B that we condition on in $P(A|B)$. Once we know the corresponding marginals AB , we can decompose the table space $\mathcal{F}_{A|B}$ accordingly.

The observed conditional frequencies c_{ij} can be used to recover marginal values s_{+j+} in the following way.

Theorem 1 Suppose $c_{ij} = \frac{g_{ij}}{h_{ij}}$ for nonnegative and relatively prime integers g_{ij} and h_{ij} . Let m_j be the least common multiple of all h_{ij} for fixed j . Then, each positive integer solution $\{x_j\}_{j=1}^J$ of

$$\sum_{j=1}^J m_j \cdot x_j = N \quad (2)$$

corresponds to a marginal s_{+j+} , up to a scalar multiple. In particular, a table \mathbf{n} consistent with the given information $\{c_{ij}, N\}$ exists if and only if Eq. (2) has a nonnegative integer solution.

Remark 1 If we allow the solutions to be only integers, then an equation of the form (2) is called a linear Diophantine equation.

The proof of the above Theorem can be found in Appendix A (Sect. 6). Since each solution of the Diophantine equation corresponds to a marginal we condition on, we easily obtain the following consequence:

Corollary 1 The following statements are equivalent:

- (a) $\mathcal{F}_{A|B}$ coincides with \mathcal{F}_{AB} .
- (b) Equation (2) has only one positive integer solution.

Note that the tables in these fibers form the support of the conditional distributions given some summary statistics. In the case of margins, there has been much work on conditional exact inference given the marginals as sufficient statistics. Also note that a marginal determines the exact (integer) cell bounds of \mathbf{n} : the cell bound for $n_{i_1 i_2 \dots i_r}$ is $[0, s_{+j+} \cdot c_{ij}]$, and a different marginal $\{s_{+j+}\}$ leads to a different cell bound. When Corollary 1 holds, there is only one AB margin. Thus, the support of conditional distribution given $\{A|B, N\}$ is the same as the support given AB and the integer cell bounds are the same, i.e., $0 \leq s_{ijk} \leq s_{ij+}$, that is, $0 \leq n_{i_1, \dots, i_r} \leq n_{ab}$ in the corresponding r -way table.

Let us single out another very important consequence of Theorem 1, which we will refer to as the table-space decomposition result:

Corollary 2 (Table-space $\mathcal{F}_{A|B}$ decomposition) *Suppose that the Diophantine Eq. (2) has m solutions. Denote by \mathbf{p}_i the marginal corresponding to the i th solution, and by $\mathcal{F}_{AB}(\mathbf{p}_i)$ the space of tables given that particular marginal table. Then, we have the following decomposition of the table space, taken as a disjoint union:*

$$\mathcal{F}_{A|B} = \bigcup_{i=1}^m \mathcal{F}_{AB}(\mathbf{p}_i).$$

To conclude this section, note that the proof of Theorem 1 shows that each solution (x_1, \dots, x_J) to the Diophantine Eq. (2) corresponds to a marginal in the following way: $s_{+j+} = m_j x_j$ for $1 \leq j \leq J$; thus, $s_{ij+} = m_j x_j c_{ij}$. We will use this fact often.

2.2 The space of tables and integer points in polyhedra

An important question arises next: how many marginals can there be for a given conditional table? This question can be answered using a straightforward count of lattice points in a polyhedron. Counting lattice points in polyhedra and counting the solutions to a Diophantine equation (e.g., Sertoz 1998; Chen and Li 2007¹) are interesting mathematical problems with a rich history. In particular, there exist polynomial time algorithms for counting the number of lattice points in polyhedra; e.g., see Barvinok (1994) and Lasserre and Zeron (2007). Due to the simpler geometry of our problem, we do not need to use the general algorithms, and, therefore, we derive simpler solutions. We explain the correspondence between solutions of Eq. (2) and nonnegative lattice points \mathbf{p}_i .

Lemma 1 *Suppose that the Diophantine Eq. (2) has a solution \mathbf{x}_0 . Then there exist vectors $\mathbf{v}_1, \dots, \mathbf{v}_{J-1} \in \mathbb{Z}^J$ such that any solution $\mathbf{x} = (x_1, x_2, \dots, x_J)$ of (2) is given as their integral linear combination:*

$$\mathbf{x} = \mathbf{x}_0 + \sum_{i=1}^{J-1} q_i \cdot \mathbf{v}_i.$$

Note that we require that each $q_i \in \mathbb{Z}$, and that $\mathbf{v}_1, \dots, \mathbf{v}_{J-1}$ can be computed from the Diophantine coefficients m_j .

The proof of this result uses elementary algebra (and some number theory). For reader’s convenience, it is included in Appendix A.2. For additional details and a low-dimensional example illustrating this lemma, see Appendix B.

That the set of all solutions to Eq. (2) is a $(J - 1)$ -dimensional lattice is a special case of a classical result that identifies the solution set of any system of linear Diophantine equations with a lattice Lazard (1996). As a subset of that lattice, the set of nonnegative solutions can be expressed as a linear combination of the elements in

¹ We note that our Diophantine equation does not necessarily satisfy the main hypothesis of the main result from Chen and Li (2007).

some basis of the lattice. In the proof of Lemma 1, we give one such combination. We use this construction to write a *solvequick()* function in R (see Appendix B) for quickly finding a solution to (2), and demonstrate its use in Sect. 4. When there is more than one solution, we provide a quick way to count the tables via a *tablecount()* function as explained next.

2.3 Size of table space: exact and approximate

First we derive the exact count formula for the total number of integer-valued r -way tables \mathbf{n} given the marginal $[AB]$. In Corollary 3, this count is combined with the table-space decomposition results from Corollary 2 to derive the number of r -way tables in the fiber $\mathcal{F}_{A|B}$.

Consider a r -way table as a 3-way table of counts s_{ijk} for A , B , and C taking I , J , and K states, respectively. Suppose we marginalize C . One can derive a simple formula for the number of 3-way tables, and, therefore, corresponding r -way tables, all having the same margin $[AB]$.

Lemma 2 (Exact count of data tables given one marginal) *Adopting the above notation, the number of r -way tables (data tables) given one marginal $[AB]$ equals*

$$|\mathcal{F}_{AB}| = \prod_{1 \leq i \leq I, 1 \leq j \leq J} \binom{s_{ij+} + K - 1}{K - 1}. \quad (3)$$

We omit the proof of this lemma, as it follows from the definition of the binomial coefficients. It is simply a count of the number of ways we can write each entry s_{ij+} in the marginal table as a sum of K entries in the data table.

Remark 2 We can find s_{ij+} from the solutions of the Diophantine equation, since $s_{ij+} = x_j m_j c_{ij}$.

With real data in mind, however, we might have to alter the formulas. Specifically, the above formulas assume that the marginals s_{ij+} are integers, but with real data due to possible rounding of observed conditional probabilities, the computed s_{ij+} 's may also be rounded. Recall that the Gamma function is defined so that $\Gamma(n) = (n-1)!$ for all integers n . Since the binomial coefficient in (3) can be written in terms of factorials, if we replace s_{ij+} with a real number instead of an integer, we get:

$$|\mathcal{F}_{AB}| = \prod_{1 \leq i \leq I, 1 \leq j \leq J} \frac{\Gamma(s_{ij+} + K)}{K! \Gamma(s_{ij+} + 1)}. \quad (4)$$

For an example, see Sect. 4.

We can use this formula to derive the exact size of the table space given observed conditionals.

Corollary 3 (Exact count of data tables given conditionals) *The number of possible r -way tables given observed conditionals $[A|B]$ is*

$$|\mathcal{F}_{A|B}| = \sum_{i=1}^m |\mathcal{F}_{AB}(p_i)|, \tag{5}$$

where m is the number of integer solutions to (2), and each $|\mathcal{F}_{AB}(p_i)|$ can be computed using Lemma 2.

Proof The claim follows by Lemma 2 and Corollary 2. □

A `tablecount()` function in R implements the above results and gives the corresponding counts. In practice, however, it may be computationally difficult to obtain the number of solutions to the Diophantine equation exactly. One remedy is provided by approximating the number of those solutions. Then, this approximation can be extended to give an approximate size for the table space $\mathcal{F}_{A|B}$. By approximation we mean a Riemann sum approximation of the integral which calculates the volume of a polytope for fixed N . We deal with the number of marginal tables first, returning to the notation of Lemma 1:

Proposition 1 (Approximate count of marginal tables given conditionals) *Given observed conditionals $[A|B]$, the number of possible marginal tables $[AB]$ is approximately*

$$|\mathcal{F}_{A|B}|_{AB} \approx \frac{N^{J-1} \text{gcd}(m_1, m_2, \dots, m_J)}{(J-1)! \prod_{i=1}^J m_i}. \tag{6}$$

This approximation may also be given by a Dirichlet integral

$$|\mathcal{F}_{A|B}|_{AB} \approx \frac{\text{gcd}(m_1, \dots, m_J)}{m_J} \int_{(x_1, \dots, x_{J-1}) \in \mathcal{M}} 1 dx_1 dx_2 \cdots dx_{J-1}, \tag{7}$$

where \mathcal{M} is the projection of the marginal polygon onto the $x_1 x_2 \dots x_{J-1}$ -plane.

A simple algebraic proof of this result can be found in Appendix A.3. Note that by Theorem 1, the number of possible marginal tables equals the number of positive integer solutions of Eq. (2). Formula in Eq. (6) uses a geometric approach via volumes of cells in the lattice; the second formula in (7) realizes the same approximation using the integral formula for volumes. Section 4 illustrates the use of these approximation formulas.

Corollary 4 (Approximate count of data tables given conditionals) *The number of possible r -way tables in $\mathcal{F}_{A|B}$ is approximately*

$$\frac{\text{gcd}(m_1, \dots, m_J)}{m_J} \int_{(x_1, \dots, x_{J-1}) \in \mathcal{M}} \prod_{i,j} \frac{\Gamma(x_j m_j c_{ij} + |C|)}{\Gamma(|C|) \cdot \Gamma(x_j m_j c_{ij} + 1)} dx_1 dx_2 \cdots dx_{J-1}, \tag{8}$$

where \mathcal{M} is the projection of the marginal polygon onto the $x_1 x_2 \dots x_{J-1}$ -plane.

Proof The claim follows from Lemma 2 and Proposition 1. Note that the total number of r -way tables equals the sum over all possible marginals of the number of tables for a fixed marginal. The approximation comes from using the approximate count in Eq. (8). \square

3 Implications for cell bounds and Markov bases

3.1 Cell bounds

There has been much discussion on calculation of bounds on cell entries given the marginals (e.g., see Dobra and Fienberg 2010 and related references), and to a limited extent the bounds given the observed conditional probabilities; e.g., see Slavković and Fienberg (2004) and Smucker and Slavković (2008). Such values are useful for determining the support of underlying probability distributions. In the context of data privacy, the bounds are useful for assessing disclosure risk; tight bounds imply higher disclosure risk. We can use the structure of the space of possible tables to obtain sharp integer bounds for the cell counts. Recall that we assume that observed conditional probabilities are exact.

There are a number of different ways to get cell bounds: (1) using linear and integer programming to solve the system of linear equations of (1); (2) using the result of equivalence of marginal and conditional fibers (c.f., Corollary 1), the bounds are given by $0 \leq s_{ijk} \leq s_{ij+}$; and (3) using our decomposition result (c.f., Corollary 2) to enumerate all possible marginal tables, and based on those get the cell bounds $\min_l (s_{ij+})_l \leq s_{ijk} \leq \max_l (s_{ij+})_l$, where l is the number of possible marginal tables AB given $A|B$.

Besides the above three methods for computing the exact cell bounds, there is a fourth method that computes approximate cell bounds by allowing arbitrary rounding of $P(A|B) = c_{ij}$. The proof is straightforward: simply recall that $\sum_j x_j m_j = N$ and $s_{ij+} = m_j x_j c_{ij}$.

Theorem 2 Given $\mathcal{T} = \{P(A|B), N\}$, an approximate (relaxation) integer cell bounds are given by

$$m_j \cdot c_{ij} \leq s_{ij+} \leq \left(N - \sum_{t \neq j} m_t \right) \cdot c_{ij}. \quad (9)$$

Furthermore, an approximate number of values that x_i can take is given by

$$\frac{\left(N - \sum_{j \neq i} m_j \right) \cdot (m_1, m_2, \dots, m_j)}{m_i \cdot (m_1, \dots, m_{i-1}, m_{i+1}, \dots, m_j)}. \quad (10)$$

These bounds can be made sharper if we know the rounding scheme of c_{ij} 's. The effect of rounding on bounds and on calculating Markov bases given observed conditionals is of special interest, but we defer that work to a future study. Some preliminary results and discussion are provided in Smucker et al. (2012) and Lee (2009).

3.2 Markov bases

In this section, we describe a structure for the Markov bases for the table space \mathcal{F}_T as defined in (1), resulting from the Corollary 2, which could lead to their simplified computation.

A set of minimal Markov moves allows us to build a connected Markov chain and perform a random walk over all the points in any given fiber. Thus, we can either enumerate or sample from the space of tables via sequential importance sampling (SIS) or Markov Chain Monte Carlo (MCMC) sampling; e.g., see Dobra et al. (2006) and Chen et al. (2006). A Markov basis for a model, or for its design matrix, is a set of moves that are guaranteed to connect all points with the same sufficient statistic. In a seminal paper by Diaconis and Sturmfels (1998), these bases were used for performing exact conditional inference over contingency tables given marginals.

Definition 1 (Diaconis and Sturmfels 1998) Let T be a $d \times n$ matrix whose entries are nonnegative integers. Assume T has no zero columns. In addition, denote by \mathcal{F}_t the fiber for t , that is, the set of all d -tuple preimages of t under the map defined by T :

$$\mathcal{F}_t = \{f \in \mathbb{N}^d : Tf = t\},$$

where t is in $\mathbb{N}^d \setminus \{0\}$.

A Markov basis of T is a set of vectors $f_1, \dots, f_L \in \mathbb{Z}^n$ with the following properties: first, the vectors must be in the kernel of T :

$$Tf_i = 0, \quad 1 \leq i \leq L.$$

Secondly, they must connect all vectors in a given fiber: for any $t \in \mathbb{N}^d \setminus \{0\}$ and any $f, g \in \mathcal{F}_t$, there exist $(\epsilon_1, f_{i_1}), \dots, (\epsilon_K, f_{i_K})$ with $\epsilon_i = \pm 1$, such that

$$g = f + \sum_{j=1}^K \epsilon_j f_{i_j}$$

and, at any step, we remain in the fiber:

$$f + \sum_{j=1}^a \epsilon_j f_{i_j} \geq 0 \quad \text{for all } a \text{ such that } 1 \leq a \leq K.$$

Note that the definition of a Markov basis does not depend on the choice of t ; it must connect each of the fibers.

In our problem, T is the matrix M in Eq. (1). Thus, the fiber \mathcal{F}_t contains the space of possible data tables that satisfy the constraints described in (1) for the given vector t . Theorem 3.1. in Diaconis and Sturmfels (1998) is considered one of the fundamental theorems in algebraic statistics and states that a Markov basis of T can be calculated as a generating set of the toric ideal I_T for the design matrix T of the model; for an introduction to toric varieties of statistical models see Drton et al. (2009).

There are a number of algebraic software packages for computing generating sets of toric ideals, and thus the Markov bases, but the most efficient to date is 4ti2 (4ti2 Team 2014). Sometimes, though, the matrix M can be large, and the computation may take too long. To alleviate some of the computational problems with contingency tables in practice, we use our table-space decomposition result (c.f. Corollary 2) to split the Markov basis into two sets. This could allow for parallel computation of the Markov sub-bases.

Corollary 5 *The Markov basis for the space of tables given the conditional can be split into two sets of moves:*

- (1) *the set of moves that fix the margin, and*
- (2) *the set of moves that change the margin.*

Proof By Corollary 2, the fiber $\mathcal{F}_{A|B}$ of tables given the conditional is a disjoint union of the sub-fibers $\mathcal{F}_{AB}(\mathbf{p}_i)$ given the fixed marginals represented by the points \mathbf{p}_i , for $i = 1, \dots, m$. By definition, the set of Markov moves consisting of the moves that change the margin connect the sub-fibers $\mathcal{F}_{AB}(\mathbf{p}_i)$, for $i = 1, \dots, m$. Thus, the Markov basis connecting all of $\mathcal{F}_{A|B}$ consists of the moves connecting each sub-fiber $\mathcal{F}_{AB}(\mathbf{p}_i)$ (the first set of moves) and the moves connecting each sub-fiber to another (the second set of moves). \square

The moves that fix the margins have been studied in the algebraic statistics literature; for some recent advances in that area, see Aoki and Takemura (2002), Aoki and Takemura (2008), De Loera and Onn (2006), and references given therein. Most recently, Dobra (2012) provided an efficient algorithm to dynamically generate the moves given the margins. Less work has been done on studying Markov bases given observed (estimated) conditionals, (e.g., see Slavković 2004; Lee 2009). Since we know, by Theorem 1, that the margins correspond to solutions to the Diophantine Eq. (2), we can find the latter set of moves by computing the Markov basis for the coefficient matrix of the Diophantine equation.

The number of Markov basis elements for this matrix seems to be small. More specifically, computations suggest the number of Markov basis elements that change the margin is as small as possible:

Conjecture 1 *In the case of small conditionals (i.e., $C \neq \emptyset$), the coefficient matrix of the Diophantine Eq. (2) has a Markov basis consisting of $J - 1$ elements, where $J - 1$ is the dimension of the underlying lattice. In other words, the corresponding toric ideal equals the lattice basis ideal.*

Note that the assumption $C \neq \emptyset$ is necessary, as the Example in Sect. 4.2.2 shows. If the con were true, it would imply the following on the size of the entire Markov basis:

Conjecture 2 *A minimal Markov basis of the matrix M in (1) contains $|B| - 1 + (|C| - 1) \times |B| \times |A|$ elements.*

Supporting examples for the above conjectures are included in Sect. 4.

On a related note, Peter Malkin has shown (in personal communication) that under certain assumptions, the number of solutions to the homogeneous linear Diophantine equation is exactly the dimension of the lattice, where by homogeneous we mean the right-hand side is zero: let D be the minimal size of all $\det(L_i)$, where L_i is the projection of the lattice L onto all variables except the i th variable. In general, a $(k - 1)$ -dimensional lattice in k variables has a Markov basis of size at least $(k - 1)$ and at most $(k - 2)D + 1$. Note that if $D = 1$, then the upper bound is $k - 1$. The size of the Markov basis for the $k - 1$ -dimensional lattice can be obtained as a consequence of a result in [Sturmfels et al. \(1994\)](#) and the Project-and-Lift method from [Hemmecke and Malkin \(2009\)](#). Namely, Proposition 4.1. of [Sturmfels et al. \(1994\)](#) states that the maximal size of a Gröbner and thus a Markov basis for a k -dimensional lattice L in k variables is at most $(k - 1)\det(L) + 1$. They state without proof that $(k - 2)\det(L) + k + 1$ is also an upper bound. The Project-and-Lift method is the one implemented in 4ti2 (4ti2 Team 2014).

Even though we cannot show that $D = 1$ holds, the conjecture above says that the size of the Markov basis is actually as small as possible. It would be of interest to obtain bounds tighter than the general one in the case of a Diophantine equation arising from the study of the table space. For more about the sizes of Markov bases and computing them, see [Malkin \(2007\)](#).

3.3 Extension of relations to marginals via DAGs

Given the marginals only, [Dobra \(2003\)](#) and [Dobra and Fienberg \(2000\)](#) have used graphical models in computing Markov bases and for calculating bounds for disclosure risk assessment. In this section, we provide extensions to those results, to the bounds in Sect. 3.1 and to Problem 1 by considering combinations of multiple conditional arrays and their relations to corresponding marginals via directed acyclic graphs (DAGs); see Sect. 4.2.4 for an example.

A DAG $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of nodes $V = \{v_1, \dots, v_r\}$ and a set of directed edges, $(v_i, v_j) \in E$, that link the ordered pairs of distinct nodes v_i (the parent), and v_j (the child) in V , and there are no cycles. A DAG satisfies the Wermuth condition ([Whittaker 1990](#)) or is perfect ([Lauritzen 1996](#)) if no subgraph has colliders, that is, if no child has parents that are not directly connected. A graph $\mathcal{G}^u = \{\mathcal{V}, \mathcal{E}^u\}$ is called *undirected* if the edges are undirected (lines), that is, if $(v_i, v_j) \in E$ then $(v_j, v_i) \in E$. A moral graph $\mathcal{G}^m = \{\mathcal{V}, \mathcal{E}^m\}$ is the undirected graph on the same vertex set as \mathcal{G} and with the same edge set \mathcal{E} including all edges that would be necessary to eliminate forbidden Wermuth configurations in \mathcal{G} .

If the random variables X_1, \dots, X_r are nodes of the graph \mathcal{G} , then the graph represents dependencies among these variables. More specifically, \mathcal{G} defines the set of probability distributions over the sample space that obeys the directed Markov properties and factorizes the joint distribution,

$$f(x_1, x_2, \dots, x_r) = \prod_{x \in \mathcal{V}} f(x|pa(x)) = f(x_1) f(x_2|x_1) \dots f(x_r|x_{r-1}, x_{n-2}, \dots, x_1). \quad (11)$$

There are many cases when the joint distribution over the contingency table has a graphical representation. In some of these cases, a set of conditionals and marginals will factor the joint according to a DAG representation. Given such a set that also satisfies the Wermuth condition, there is an equivalent undirected graph representation of the same set. In that case, the generalized Problem 1 is reduced to one of knowing a set of marginals, and the bounds are those given by [Dobra and Fienberg \(2000, 2010\)](#). The following results hold for any r -way table.

Theorem 3 *Let \mathcal{T} be a set of conditional and marginal distributions inducing bounds on the cell entries. Let \mathcal{G} be a DAG, and \mathcal{G}^u the undirected graph associated with \mathcal{T} . When \mathcal{G} satisfies the Wermuth condition, the bounds imposed by \mathcal{T} reduce to the bounds imposed by a set of marginals associated with \mathcal{G}^u .*

Proof This result follows from well-known properties of a DAG and more specifically from the Markov theorem for directed independence graphs ([Whittaker 1990](#); [Lauritzen 1996](#)). The theorem states that the DAG possesses the Markov properties of its associated moral graph. Therefore, there is an equivalence of the set of edges for \mathcal{G}^m and \mathcal{G}^u . The directed edges in the DAG carry independence statement information on a sequence of marginal distributions, while the undirected graph describes the independence statements on a single conditional. Since the edge sets are equivalent, the DAG then gives the equivalent information on the joint as its associated undirected graph. \square

Corollary 6 *Let \mathcal{G}^m be the moral graph associated with \mathcal{G} . If $\mathcal{G}^m = \mathcal{G}^u$, then the bounds induced by a set \mathcal{T} are equivalent to the bounds induced by the set of marginals associated with \mathcal{G}^u .*

An interesting link between bounds on cells in the contingency tables, DAGs, and Markov bases is indicated by the next result.

Corollary 7 *Let \mathcal{T} be a set of conditional and marginal distributions. Let \mathcal{G} be a DAG and \mathcal{G}^u the undirected graph associated with \mathcal{T} . When \mathcal{G} satisfies the Wermuth condition, the Markov basis describing \mathcal{T} under the same ordering is the same Markov basis induced by a set of marginals associated with \mathcal{G}^u .*

Proof The claim follows from Corollary 5. \square

It is possible that similar results, with discrete random variables, could be derived for the chain graphs and ancestral graphical models (e.g., [Richardson and Spirites 2002](#)) which are generalization of the directed and undirected graphs. This is an interesting topic for future research.

4 Examples

In this section we illustrate the results described in the preceding sections through analysis of a series of simple contingency tables. We show how to use our initial R (R Development Core Team 2005) implementation of the formulas from Sects. 2 and 3. We also perform our analyses using the well-established and free algebraic software LattE macchiato (LattE 2014) which relies on an implementation of Barvinok’s algorithm (Barvinok et al. 2010) for counting and detecting lattice points inside convex polytopes. In statistical literature, LattE has been mostly used for counting the number of tables given the margins.

4.1 A $2 \times 2 \times 2$ Example

Consider a fictitious $2 \times 2 \times 2$ table that cross-classifies a randomly chosen sample of 50 college students by their *Gender*, illegal *Downloading* of MP3 files, and the dorm *Building* they live in; see counts in Table 1. We use shorthand G for *Gender*, D for *Downloading*, and B for *Building* variable.

The survey administrator has the full information on the $[BGD]$ table, but due to confidential nature of the data, would like to consider releasing only partial information to public such as the marginal counts $[DG]$ as in Table 2 or the grand total 50 and the small conditional $P(\text{Download}|\text{Gender})$ as in Table 3. This requires comparison of the space of tables \mathcal{F}_{DG} , which based on Lemma 2 has $16 \times 11 \times 6 \times 21 = 22176$ possible $[BGD]$ tables, with the space of tables $\mathcal{F}_{D|G}$.

Table 1 A $2 \times 2 \times 2$ table of counts of illegal MP3 downloading by gender and a residing building

Building	Gender	Download		Total
		Yes	No	
I	Male	8 [0, 29.4] [0, 27] [0, 15]	4 [0, 19.6] [0, 18] [0, 10]	12
I	Female	2 [0, 9.8] [0, 9] [0, 5]	9 [0, 39.2] [0, 36] [0, 20]	11
II	Male	7 [0, 29.4] [0, 27] [0, 15]	6 [0, 19.6] [0, 18] [0, 10]	13
II	Female	3 [0, 9.8] [0, 9] [0, 5]	11 [0, 39.2] [0, 36] [0, 20]	14
	Total	20	30	50

The value in the brackets are linear relaxation bounds and sharp integer bounds given released conditional $[D|G]$ and marginal $[DG]$, respectively

Table 2 $[GD]$ Marginal table of illegal MP3 downloading, and integer bounds given released $[D|G]$ and $N = 50$

Gender	Download		Total
	Yes	No	
Male	15 [3, 27]	10 [2, 18]	25
Female	5 [1, 9]	20 [4, 36]	25
Total	20	30	50

Table 3 $[D|G]$ Table of conditional probabilities with reduced fractions and [rounded probability]

Gender	Download	
	Yes	No
Male	$\frac{15}{25} = \frac{3}{5}$ [0.6]	$\frac{10}{25} = \frac{2}{5}$ [0.4]
Female	$\frac{5}{25} = \frac{1}{5}$ [0.2]	$\frac{20}{25} = \frac{4}{5}$ [0.8]
Total	20	30

The reference set $\mathcal{F}_{D|G}$ consists of tables that are solutions to the following:

$$\left\{ \begin{array}{l} \left[\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & -3 & & & 2 & -3 & & \\ & & 4 & -1 & & & 4 & -1 \end{array} \right] \mathbf{n} = \begin{bmatrix} 50 \\ 0 \\ 0 \end{bmatrix} \\ n_1 + n_2 + n_5 + n_6 > 0 \\ n_3 + n_4 + n_7 + n_8 > 0 \\ \text{All } n_i\text{'s are nonnegative integers} \end{array} \right\}.$$

This is part of a 5-dimensional lattice inside the \mathcal{R}^2 . Then Eq. (2) of Theorem 1 for this example is $5x_1 + 5x_2 = 50$, and it has 9 positive integer solutions: $\{(x_1 = i, x_2 = 10 - i) | 1 \leq i \leq 9\}$. Thus, there are 9 different $[DG]$ marginals, which, by Theorem 1, means that the space of tables given the small conditional $[D|G]$ and the grand total is different from the space of tables given the corresponding marginal counts. In fact, the space is larger: $|\mathcal{F}_{D|G}| > |\mathcal{F}_{DG}|$. More specifically, Corollary 3 for $m = 9$ provides the table count: $|\mathcal{F}_{D|G}| = \sum_{m=1}^9 |\mathcal{F}_{DG_m}| = 128676$. In R, we invoke function *tablecount*($M, 2$) where M is any one of 9 possible marginal tables $[DG]$. Notice that this formulation does not allow any row of $[G]$ to have a total of zero counts. If such tables were to be allowed, then the total number of possible 3-way tables would be $128676 + 651 + 451 = 129778$ where 651 and 451 are the numbers of possible 3-way tables given the $[DG]$ when one of the rows of $[G]$ is equal to zero.

To approximate the number of marginal tables $[DG]$, one can use the formula from Eq. (7) in Proposition 1 to count the number of corresponding solutions to the Diophantine equation as $\frac{50\text{gcd}(5,5)}{5 \times 5} = 10$. Then, we can use the integral formula from Corollary 4, which could be evaluated, say, using Maple, to estimate the size of the total table space given the conditionals as $\frac{\text{gcd}(5,5)}{5} \int_0^{10} (3x + 1)(2x + 1)(10 - x + 1)(40 - 4x + 1) = 129676.7$.

Since more than one possible margin is consistent with the given conditional and grand total, clearly \mathcal{F}_{DG} is strictly contained in $\mathcal{F}_{D|G}$. This can also be seen by computing the cell bounds on the cell entries of $[BDG]$ contingency table. In Table 1, given $\mathcal{F}_{D|G}$, the linear relaxation cell bounds and the exact integer bounds are given in the black and blod brackets, respectively. Given \mathcal{F}_{DG} , the exact cell bounds are in italic brackets. The idea is that the wider bounds offer more protection. These bounds are obtained by direct optimization for each given constraint. However, the results of Sect. 3 show a computational shortcut to obtaining bounds given $[D|G]$ and $N = 50$ by using already established results on bounds of cell entries given the marginals. First,

by Theorem 2 we obtain bounds on the missing margin $[DG]$ (see Table 2). Next, we combine this with a well-known fact that given one marginal s_{ij+} , the bounds on each cell entry of the 3-way table are $0 \leq n_{ijk} \leq s_{ij+}$. Thus, the bounds for n_{ijk} are between 0 and the upper bound found for the missing marginal table. For example, for the cell $(1, 1, 1)$, the $3 \leq s_{11+} \leq 27$, and $0 \leq n_{111} \leq 27$; these are the bounds given in the bold brackets in Table 1.

It has been observed in the literature already that the above-described bounds have gaps. That is, not all values within the interval are possible. This observation is particularly important for assessing disclosure risk with contingency tables. By enumerating all possible marginal tables, we learn both the number of all possible r -way tables, and the values in the cell counts of those tables. We can obtain such tables quickly by using the `solvequick()` function. For example, `solvequick(c(5, 5), 50)` gives a vector of all possible G margins that we conditioned on in $[D|G]$. To get $[DG]$ margins, compute $m_j \times b \times c_{ij}$.

Next, we calculate a Markov basis for fixed $[D|G]$ using 4ti2. If Conjecture 1 is true, then so is Corollary 2, and there should be $5 = |G| - 1 + (|B| - 1) \times |G| \times |D| = 1 + 1 \times 2 \times 2$ Markov moves. Our computation finds exactly 5 moves:

$$\begin{pmatrix} 3 & 2 & -1 & -4 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

In accordance with Corollary 5, the last 4 moves correspond to a set of moves that fix the $[DG]$ margin, while the first move changes the margin $[DG]$, but keeps the N fixed. From the first element, $n_1^3 n_2^2 - n_5^1 n_6^4$, by summing the exponents in each monomial, we can deduce exactly the amount by which a count in each level of the margin we condition on changes. In this example, each marginal count of $[G]$ changes by a count of 5. Thus, with the sample size $N = 50$, the upper bound for the solution to equation (2) for the number of possible marginals $[G]$, and thus of $[DG]$, is 10.

A related example, providing more details and implications of when a Diophantine equation has only one solution is available in the supplementary documents at <http://www.stat.psu.edu/~sesa/cctable>.

4.2 A $3 \times 2 \times 2$ table with zero counts

In this section, we apply our derived results to a $3 \times 2 \times 2$ table (see Table 4) with zero counts, and show the convergence of exact and approximate results.

4.2.1 Small conditional $B|A$ and N

Consider that we do not observe the original table, and the only available information is $\mathcal{T} = \{Pr(B|A), N = 240\}$; the sample values are given in Table 5.

Table 4 A $3 \times 2 \times 2$ Table

		C = 1	C = 2	Total
A = 1	B = 1	10	20	30
A = 1	B = 2	10	20	30
A = 2	B = 1	20	0	20
A = 2	B = 2	0	40	40
A = 3	B = 1	0	30	30
A = 3	B = 2	30	60	90
Total		70	170	240

Table 5 Left panel: observed counts of the $[AB]$ marginal table, and notation for when those counts are missing, right panel: observed conditional probabilities $[B|A]$ based on values in Table 4

	B = 1	B = 2		B = 1	B = 2
A = 1	30 [x]	30 [x]	A = 1	1/2	1/2
A = 2	20 [y]	40 [2y]	A = 2	1/3	2/3
A = 3	30 [z]	90 [3z]	A = 3	1/4	3/4

By Theorem 1, the linear Diophantine equation that characterizes all possible missing $[AB]$ margins is

$$2x + 3y + 4z = 240. \tag{12}$$

Using our R code, e.g., `solvecount(c(2, 3, 4), 240)`, we learn that there are 1141 possible A marginals consistent with the provided information. Since the triplets (x, y, z) are in 1-to-1 correspondence to $[AB]$ margins (see Table 5), there are 1141 missing $[AB]$ marginals consistent with the provided information. Furthermore, `solvequick(c(2, 3, 4), 240)` lists all positive integer solutions to Eq. (12), and from there we easily obtain all corresponding $[AB]$ margins.

We are ultimately interested in finding all possible 3-way tables consistent with given information, i.e, solutions to the following system

$$\left\{ \begin{array}{l} \left[\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & & & & & & & & \\ & & & & 2 & -1 & 2 & -1 & & & & \\ & & & & & & & & 3 & -1 & 3 & -1 \end{array} \right] X = \begin{bmatrix} 240 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\ n_1 + n_2 + n_3 + n_4 > 0 \\ n_5 + n_6 + n_7 + n_8 > 0 \\ n_9 + n_{10} + n_{11} + n_{12} > 0 \\ \text{All } n_i\text{'s are nonnegative integers} \end{array} \right\},$$

which is part of a 8-dimensional lattice inside the \mathcal{R}^{12} . The exact number of possible 3-way tables can be obtained by Corollary 3, $|\mathcal{F}_{B|A}| = \sum_{m=1}^{1141} |\mathcal{F}_{AB_m}|$. In R, we invoke `format(tablecount(M, 2), digits = 22)`, which gives 1187848498271 possible $[ABC]$ contingency tables.

Next, we demonstrate in a little more detail and following the proof of Proposition 1, how to set up the integrals to calculate the approximate number of solutions. Recall that

a marginal table $[AB]$ corresponds to a triple (x, y, z) . Note that $z = (240 - 2x - 3y)/4$. Thus, for each marginal table, the number of possible tables that have this margin is

$$(x + 1)^2(y + 1)(2y + 1) \left(\frac{240 - 2x - 3y}{4} + 1 \right) \left(3 \frac{240 - 2x - 3y}{4} + 1 \right).$$

After summing over all possible (x, y) , we get the count of all possible $[ABC]$ tables:

$$\sum_{(x,y) \in \mathcal{M}} (x + 1)^2(y + 1)(2y + 1) \left(\frac{240 - 2x - 3y}{4} + 1 \right) \left(3 \frac{240 - 2x - 3y}{4} + 1 \right)$$

where \mathcal{M} is the projection of all possible triple (x, y, z) onto the xy -plane. As discussed in the proof of Proposition 1, notice that \mathcal{M} is a part of a lattice whose unit cell has an area of $4/\text{gcd}(2,3,4)$. Thus, the number of possible solutions is approximately $1.188479935 \times 10^{12}$ by solving the following

$$\frac{1}{4} \int_0^{80} \int_0^{\frac{240-3y}{2}} (x + 1)^2(y + 1)(2y + 1) \left(\frac{240 - 2x - 3y}{4} + 1 \right) \left(3 \cdot \frac{240 - 2x - 3y}{4} + 1 \right) dx dy.$$

The ratio of the exact solution to the approximate solution, for either counting the missing margin or the r -way table, is $1 + O(1/N)$. For this example, we compute exact and approximate number of tables while varying the grand total N . Table 6 summarized the results for the missing marginal $[AB]$, and Table 7 lists the exact number and approximate number of $[ABC]$ tables for different values of the total sample size. Numerical experiments show evidence that our approximation is sharper for equations with fewer unknowns, and/or when N is much larger than the coefficients in the equation. For the small number of margins, the approximation does not work well.

Next, we calculate a Markov basis for fixed $[B|A]$ using 4ti2. According to Corollary 2, there should be 8 elements in this basis. A Markov basis for this example is given below. In accordance to Corollary 5, the last 6 moves correspond to a set of moves that fix the $[AB]$ margin, and the first two moves change the margin $[AB]$ while keeping N fixed. As noted before, the sum of the exponents in the monomial tells us by how much the margin $[A]$ can change.

Table 6 Exact and approximate number of missing marginal tables $[AB]$

	Exact count	Approximation
$N = 24$	7	12
$N = 240$	1141	1200
$N = 2400$	119401	120000
$N = 24000$	11994001	12000000

Table 7 Exact and approximate number of missing tables $[ABC]$

	Exact count	Approximation
$N = 24$	52937	65150
$N = 240$	1187848498271	$1.188479935 \times 10^{12}$
$N = 2400$	96999660430647444101	$9.699971869 \times 10^{19}$
$N = 24000$	9501190342113804461451781001	$9.501190349 \times 10^{27}$

$$\begin{pmatrix} -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 \\ -3 & -3 & 0 & 0 & 2 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

4.2.2 Full conditional $A|BC$ and N

Before considering the release of other partial conditionals, we next demonstrate how some of our results also hold for the full conditional. First, if the only information available about the original table are the observed conditional rates, e.g., $[A|BC]$, and N , as indicated in Sect. 3.2, we only need to solve a linear Diophantine equation to find the total number of possible 3-way tables, e.g.,

$$3x_1 + 4x_2 + 5x_3 + 6x_4 = 240.$$

We would typically count the number of possible solutions by setting up the full constraint matrix in LattE (e.g., see Appendix B), but now we can simply apply *solvequick*($c(3, 4, 5, 6), 240$) in R. The number of possible tables is 5715, which corresponds to the number of possible $[BC]$ margins. Second, notice that the $[A|BC]$ conditional rates have zero values, e.g., cell (2, 2, 1) since the original cell has a zero count. However, the presence of zeros does not affect our computation since we are not conditioning on margins with zero counts.

Last, the Markov basis has the following 4 elements, all of which change the $[ABC]$ margin:

$$\begin{pmatrix} -2 & 0 & 0 & 1 & -4 & 0 & 0 & 3 & 0 & 0 & 0 & 2 \\ -3 & 2 & 1 & 0 & -6 & 0 & 0 & 0 & 0 & 3 & 3 & 0 \\ -2 & 4 & -1 & 0 & -4 & 0 & 0 & 0 & 0 & 6 & -3 & 0 \\ -1 & -2 & 2 & 0 & -2 & 0 & 0 & 0 & 0 & -3 & 6 & 0 \end{pmatrix}.$$

Conjecture 1 about the number of elements in the basis, however, does not hold here because we are using full conditionals, that is, $C = \emptyset$. As supported by other examples, this conjecture seems true for *small* conditionals only.

4.2.3 Partial conditional $B|C$ and N

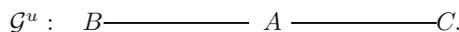
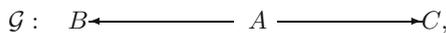
Here we briefly consider a case where the missing marginal has more than two levels. Let the available information be the sample size and the small conditional $[B|C]$ with the missing variable $[A]$ that has 3 levels. The following Diophantine equation captures the information preserved by the sample size and $[B|C]$:

$$7x_1 + 17x_2 = 240.$$

In R, the `solvequick(c(7, 17), 240)` function obtains only two possible non-negative integer solutions, that is, only two possible marginal tables $[BC]$. Then, running `tablecount(M, 3)`, where M is one of the $[BC]$ margins, tells us that there are total of 6130182419416 $[ABC]$ tables. In this example, it is easy to check via LattE that Corollary 3 holds. We compute the number of ABC tables given each BC margin, and see that their sum is equal to the number we obtained via the `tablecount()` function. According to this corollary, $|\mathcal{F}_{B|C}| = \sum_{m=1}^2 |\mathcal{F}_{BC_m}| = 4179685045536 + 1950497373880 = 6130182419416$. It should be noted here that the function `tablecount(M, 3)` gives the total number of $[ABC]$ tables regardless of which compatible $[BC]$ margin we use. The conjectures for the size of Markov bases hold here as well. We observe that there are 9 elements in a basis: 8 fix the $[BC]$ margin, and 1 changes the $[BC]$ margin.

4.2.4 Combinations of partial conditionals and N

Let's assume that we observe $\mathcal{T} = \{P(B|A), P(C|A), P(A), N\}$, and recall that we assume that there exists a joint distribution from which we observed these compatible pieces. Then this collection can be graphically represented by a DAG \mathcal{G} that satisfies the Wermuth condition. This DAG and its corresponding undirected graph \mathcal{G}^u are given in the picture below. By Theorem 3 the bounds on the cell counts are the same as in the case of given margins $[AB]$ and $[AC]$. Based on Corollary 7, the Markov bases will be the same, and so will the fibers \mathcal{F}_τ and $\mathcal{F}_{AB,AC}$. Note that these results capture the following special case: if the model according to DAG is true, that is B and C are conditionally independent given A , then by the Wermuth condition we can uniquely specify the joint distribution, $P(A, B, C) = P(AB)P(AC)$:



Now assume that marginal $[A]$ is missing or hidden, and we only have partial information in the form of observed conditional frequencies $[B|A]$ and $[C|A]$, and sample size N . If there is a unique solution for the margin $[A]$, then there are unique two-way margins $[AB]$ and $[AC]$. By Theorem 3 and Corollary 7 then this is equivalent to having information on two margins, and we can proceed by calculating the cell bounds, counting tables, and by sampling given the marginals.

Consider our running example from Table 4 but with $N = 24$. Let $\mathcal{T} = \{P(B|A), P(C|A), N = 24\}$, where the observed conditional values are the same as with $N = 240$; e.g., for $P(B|A)$, see Table 5. By Theorem 1, the equation that characterizes the missing marginal $[A]$ and thus $[AB]$ for $[B|A]$ is

$$2x + 3y + 4z = 24. \tag{13}$$

Based on $\text{solvecount}(c(2, 3, 4), 24)$, we learn that there are 7 possible $[A]$ margins. Furthermore, there are 52937 possible 3-way $[ABC]$ tables. The linear Diophantine equation that characterizes the missing marginal $[A]$ and thus $[AC]$ based on knowledge of $[C|A]$ is

$$3x + 3y + 4z = 24, \tag{14}$$

and from the running $\text{solvecount}(c(3, 3, 4), 24)$, we learn that there are 3 possible A margins. There are 22440 possible 3-way ABC tables.

We are interested in the intersection of the two solution spaces. Using our function $\text{intersect}()$ in R, we learn that there is only one $[A]$ that satisfies both equations, and it takes values (6, 6, 12). Since there is only one $[A]$, this implies that there is only one $[AB]$ and one $[AC]$ margin, and thus the space of 3-way tables $[ABC]$ is the same as the space given these two margins. More specifically, $|\mathcal{F}_\tau| = |\mathcal{F}_{AB,AC}| = 36$. Our analysis shows that the results from Sect. 3.3 hold, and we do get the same bounds and Markov bases as would if we only consider the marginal information. A Markov basis for fixed $[B|A]$ and $[C|A]$ has 5 elements: 3 fix the missing $[A]$ margin, and 2 change it:

$$\begin{pmatrix} -4 & -2 & 0 & -6 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 9 \\ -2 & -1 & 0 & -3 & 2 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since in this example $[A]$ is unique, that would be like adding an additional constraint, and the actual minimal basis that describes our system of polynomial equations reduces to:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We get the same Markov basis if we calculate it based on fixing $[AB]$ and $[AC]$ margins.

If $N = 240$, the Markov bases based on fixing $[B|A]$ and $[C|A]$ will be the same as with $N = 24$; that is, they will have 5 elements shown above. However, now there are 361 possible $[A]$ margins consistent with both $[B|A]$ and $[C|A]$, and the Theorem 3 and Corollary 7 are not satisfied, and the Markov basis will not reduce to the Markov basis given the corresponding marginals. Furthermore, the space of tables given the conditional is signifi-

cantly larger than the space of tables given the corresponding marginals: $|\mathcal{F}_T| = 3066315 \geq |\mathcal{F}_{AB,AC}| = 13671$. Thus, the bounds on the cell entries are different, as is the support for the sampling distribution over the space of tables $[ABC]$.

Similar analysis can be done for other arbitrary collections of conditionals and marginals. For example, $\mathcal{T} = \{P(B|A), P(A|C), P(C)\}$ will also satisfy the results from Sect. 3.3. If margin $[C]$ is missing, but it is unique based on the solution to a linear Diophantine equation, we would again have a reduction of results; that is, the space \mathcal{F}_T will be equivalent to the space $\mathcal{F}_{AB,AC}$. For additional examples, see <http://www.stat.psu.edu/~sesa/cctable>.

5 Conclusions

We have used algebraic statistics to solve an open problem posed by [Dobra et al. \(2008\)](#). One area of this expanding field is concerned with the study and characterization of portions of the sample space and, in particular, of all datasets (i.e., tables) having the same observed margins and/or conditionals. In this paper, we describe the space of all possible r -way contingency tables for a given sample size and set of observed (estimated) conditional frequencies. This space of contingency tables can be decomposed according to different possible marginals, which, in turn, are encoded by the solution set to a linear Diophantine equation, giving the table space a special structure. As a consequence, we obtain conditions under which two spaces of tables coincide: one is the space of tables for a given set of marginals, and the other is our space—for a given sample size and set of conditionals. This characterization of the difference between two fibers has thus provided a solution to an open problem in the literature.

In general, these fibers can be quite large. We provide formulas for computing the approximate and exact cardinality of the fibers in question, and we implemented those in R. The knowledge of the structure of the space of tables also enables us to enumerate all the possible data tables. This, in turn, leads to new cell bounds, some including connections to DAGs with combinations of conditionals and marginals. In this paper, we assumed that the given sets of conditionals and marginals are compatible; for problems on compatibility for categorical and continuous variables, see [Arnold et al. \(1999\)](#); on compatibility of full conditionals for discrete random variables, see [Slavković and Sullivant \(2006\)](#); and on generalization of compatibility of conditional probabilities in discrete cases, see [Morton \(2013\)](#). Consistent with the literature on the characterizations of joint discrete distributions, we allow cell entries to be zero as long as we do not condition on an event of zero probability, and we assumed that the uniqueness theorems as stated in [Arnold et al. \(1999\)](#) and [Slavković and Fienberg \(2010\)](#) hold. Then we considered if the given summary statistics are sufficient to uniquely identify the existing joint distribution, and if not, we proceed with the description of the related sample space.

Another application of the main observation, the table-space decomposition result, is that it allows us to describe the Markov bases given the conditionals. We observe that

the moves consist of two sets: those that fix the margins, and those that change them. This result could lead to a simplified calculation of Markov bases in this particular setting. However, this remains to be studied more carefully. We raised a number of conjectures, and in particular we hope to prove Conjecture 1.

The properties of fibers, and, therefore, the results of this paper, are important in determining the support of sampling distributions, for conducting exact conditional inference, calculating cell bounds in contingency tables, and imputing missing cells in tables. The degree of Markov moves for given conditionals is arbitrary in the sense that it depends on the values of observed conditional probabilities, unless we use the observed cell counts directly. In practice, however, the conditional values are reported as real numbers. Depending on the rounding point, the bounds, the moves and the fibers will differ from each of its kind. This has implications for statistical inference; in particular, in assessing “true” disclosure risk in data privacy problems. The effect of rounding needs more careful investigation. This problem is related to characterizing when the integral approximation of the number of tables is correct up to rounding, and when the error is “too large.”

6 Proofs

6.1 Proof of Theorem 1

Proof Assume \mathbf{n} is a table consistent with the given conditional $\{c_{ij}\}$ and grand total N . We can summarize the table using \mathbf{n}^* as described in the Introduction. Thus $\frac{g_{ij}}{h_{ij}} = \frac{s_{ij+}}{s_{+j+}}$. Since g_{ij} and h_{ij} are relatively prime, it follows that s_{+j+} is an integer multiple of h_{ij} . Furthermore, this is true for any i . By definition of m_j , s_{+j+} is an integer multiple of m_j . In other words, we can write s_{+j+} as $m_j \cdot x_j$ where x_j is a positive integer. Now Eq. (2) is satisfied since by definition $\sum_j s_{+j+} = N$. Conversely, assume (2) holds for the positive integers x_j 's. Then we construct \mathbf{n} by letting s_{ij+} to be $m_j \cdot x_j \cdot c_{ij}$. Then let s_{ijk} to be nonnegative integers according to the equation $s_{ij+} = \sum_k s_{ijk}$. Then construct \mathbf{n} according to \mathbf{n}^* in a similar way.

6.2 Proof of Lemma 1

In the following, let (m_1, \dots, m_l) denote the greatest common divisor of m_1, \dots, m_l for any arbitrary $1 \leq l \leq J$. Notice that the standard Euclidean algorithm produces integers x_1^0, \dots, x_l^0 such that $m_1 x_1 + \dots + m_l x_l = (m_1, \dots, m_l)$. Repeatedly using this process, we get x_i 's such that $\sum_{i=1}^{J-j} m_i x_i^{(j)} = (m_1, m_2, \dots, m_{J-j})$ for any j . In particular, we can set $x_i = x_i^{(0)} \cdot \frac{N}{(m_1, \dots, m_J)}$ to obtain one of the integer solutions of (2). Note that this algorithm performs at most $\sum_{i=1}^J m_i$ calculations. Similarly, every solution of the Diophantine equation can be obtained by integers linear combinations, generalizing the two basic examples.

Proof Elementary arguments allow us to express the vectors $\mathbf{v}_1, \dots, \mathbf{v}_{J-1}$ in terms of the coefficients m_1, \dots, m_J . By the Euclidean algorithm, the gcd's (m_1, \dots, m_l) can be expressed as a linear combination of the m_j 's:

$$\sum_{i=1}^{J-j} m_i x_i^{(j)} = (m_1, m_2, \dots, m_{k-j})$$

for $j = 1, 2, \dots, J$. Then we can express all integer solutions of Eq. (2) as:

$$x_l = x_l^{(0)} - \sum_{h=1}^{J-l} \frac{m_{J+1-h} x_l^{(h)}}{(m_1, \dots, m_{J+1-h})} \cdot q_h + \frac{(m_1, \dots, m_{l-1})}{(m_1, \dots, m_l)} \cdot q_{J-l+1} \quad \text{for } l=2, \dots, J,$$

$$x_1 = x_1^{(0)} - \sum_{h=1}^{J-1} \frac{m_{J+1-h} x_1^{(h)}}{(m_1, \dots, m_{J+1-h})} \cdot q_h,$$

where $q_i \in \mathbb{Z}$ for all i with $1 \leq i \leq J-1$. Then the vectors \mathbf{v}_i , for $i = 1, \dots, J-1$, are determined from these expressions as follows: the l th coordinate of \mathbf{v}_i is the coefficient of q_i in the expression for x_l .

6.3 Proof of Proposition 1

Proof To approximate the number of nonnegative solutions, define a vector $u := [m_1, m_2, \dots, m_k]^T$, and a matrix $A := [u, v_1, v_2, \dots, v_{k-1}]$. Recall that vectors v_1, \dots, v_k come from Lemma 1. From the expressions above, we see that

$$A = \begin{bmatrix} m_1 & \frac{m_k x_1^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_1^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \dots & \frac{m_2 x_1^{(k-1)}}{(m_1, m_2)} \\ m_2 & \frac{m_k x_2^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_2^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \dots & \frac{-m_1}{(m_1, m_2)} \\ m_3 & \frac{m_k x_3^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_3^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \frac{-(m_1, m_2)}{(m_1, m_2, m_3)} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{k-1} & \frac{m_k x_{k-1}^{(1)}}{(m_1, \dots, m_k)} & \frac{-(m_1, \dots, m_{k-2})}{(m_1, \dots, m_{k-1})} & 0 & \dots & 0 \\ m_k & \frac{-(m_1, \dots, m_{k-1})}{(m_1, \dots, m_k)} & 0 & 0 & \dots & 0 \end{bmatrix}.$$

One readily checks that u is orthogonal to any column v_i . Thus the absolute value of $(\det A)/\|u\|$ is the $(k - 1)$ -dimensional volume of the parallelepiped spanned by v_1, v_2, \dots, v_{k-1} . Let's compute this value:

$$\begin{aligned}
 & \frac{\det A}{\|u\|} \\
 &= \frac{1}{\sqrt{m_1^2 + m_2^2 + \dots + m_k^2}} \cdot \det A \\
 &= \frac{1}{\sqrt{m_1^2 + \dots + m_k^2}} \cdot \det \begin{bmatrix} m_1 & \frac{m_k x_1^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_1^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \dots & \frac{m_2 x_1^{(k-1)}}{(m_1, m_2)} \\ m_2 & \frac{m_k x_2^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_2^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \dots & \frac{-m_1}{(m_1, m_2)} \\ m_3 & \frac{m_k x_3^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_3^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \frac{-(m_1, m_2)}{(m_1, m_2, m_3)} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{k-1} & \frac{m_k x_{k-1}^{(1)}}{(m_1, \dots, m_k)} & \frac{-(m_1, \dots, m_{k-2})}{(m_1, \dots, m_{k-1})} & 0 & \dots & 0 \\ m_k & \frac{-(m_1, \dots, m_{k-1})}{(m_1, \dots, m_k)} & 0 & 0 & \dots & 0 \end{bmatrix} \\
 &= \frac{1}{m_1 \sqrt{m_1^2 + \dots + m_k^2}} \cdot \det \begin{bmatrix} \sum_{i=1}^k m_i^2 & 0 & 0 & \dots & \dots & 0 \\ m_2 & \frac{m_k x_2^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_2^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \dots & \frac{-m_1}{(m_1, m_2)} \\ m_3 & \frac{m_k x_3^{(1)}}{(m_1, \dots, m_k)} & \frac{m_{k-1} x_3^{(2)}}{(m_1, \dots, m_{k-1})} & \dots & \frac{-(m_1, m_2)}{(m_1, m_2, m_3)} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{k-1} & \frac{m_k x_{k-1}^{(1)}}{(m_1, \dots, m_k)} & \frac{-(m_1, \dots, m_{k-2})}{(m_1, \dots, m_{k-1})} & 0 & \dots & 0 \\ m_k & \frac{-(m_1, \dots, m_{k-1})}{(m_1, \dots, m_k)} & 0 & 0 & \dots & 0 \end{bmatrix} \\
 &= \frac{(-1)^{k-1} \sqrt{m_1^2 + \dots + m_k^2}}{(m_1, m_2, \dots, m_k)}.
 \end{aligned}$$

Thus the volume of the parallelotope spanned by v_1, v_2, \dots, v_{k-1} is

$$\frac{\sqrt{m_1^2 + m_2^2 + \dots + m_k^2}}{(m_1, m_2, \dots, m_k)}.$$

Next, define

$$G = \{(x_1, \dots, x_k)^T \mid m_1 x_1 + m_2 x_2 \dots + m_k x_k = N, \quad x_1 \geq 0, x_2 \geq 0, \dots, x_k \geq 0\}.$$

Lets refer to G as the marginal polytope. The volume of G is easily calculated to be

$$\frac{N^{k-1}}{(k-1)!(m_1 \cdot m_2 \cdot \dots \cdot m_k)} \sqrt{m_1^2 + m_2^2 + \dots + m_k^2}.$$

The approximation to the number of lattice points in G , that is, the number of positive integer solutions of (2) is obtained by dividing the volume of G by the volume of the parallelotope above. This proves the first claim.

For the second claim, let \mathcal{P} be the projection of the set of positive integer solutions onto the $x_1 \dots x_{j-1}$ -plane. Then there are exactly

$$\sum_{x_1 x_2 \dots x_{j-1} \in \mathcal{P}} 1$$

positive integer solutions of the Diophantine Eq. (2). Let a be the area of the unit cell of the lattice spanned by \mathcal{P} .

Then

$$\int_{(x_1, \dots, x_{j-1}) \in \mathcal{M}} 1 dx_1 \dots dx_j \approx a \cdot \sum_{x_1 x_2 \dots x_{j-1} \in \mathcal{P}} 1,$$

where the right hand side is, by definition, the Riemann sum approximation of the integral. In particular, one easily concludes that the error of this approximation is given by the difference in the volume of the polytope \mathcal{M} and the volume of the polyhedron which is the union of all the unit cells anchored at the lattice points \mathcal{P} .

To complete the proof, we calculate the area of the unit cell a . Let \mathcal{L} be the lattice of all integer solutions to Eq. (2). Since $\mathcal{P} \subseteq \mathcal{L} \cup \{x_j = 0\}$, we can choose its unit cell to be the projection of the unit cell of \mathcal{L} onto $\{x_j = 0\}$. This projection, in turn, is a parallelopiped whose $(j - 1)$ -dimensional volume is the absolute value of

$$\det \begin{bmatrix} \frac{m_j x_2^{(1)}}{\gcd(m_1, \dots, m_j)} & \frac{m_{j-1} x_2^{(2)}}{\gcd(m_1, \dots, m_{j-1})} & \dots & \dots & \frac{-m_1}{\gcd(m_1, m_2)} \\ \frac{m_j x_3^{(1)}}{\gcd(m_1, \dots, m_j)} & \frac{m_{j-1} x_3^{(2)}}{\gcd(m_1, \dots, m_{j-1})} & \dots & \frac{-\gcd(m_1, m_2)}{\gcd(m_1, m_2, m_3)} & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \frac{m_j x_{k-1}^{(1)}}{\gcd(m_1, \dots, m_j)} & \frac{-\gcd(m_1, \dots, m_{j-2})}{\gcd(m_1, \dots, m_{j-1})} & 0 & \dots & 0 \\ \frac{-\gcd(m_1, \dots, m_{j-1})}{\gcd(m_1, \dots, m_j)} & 0 & 0 & \dots & 0 \end{bmatrix}$$

which is $\frac{m_1}{\gcd(m_1, m_2, \dots, m_j)}$.

7 Code and Examples

Example 1 Let us consider a bivariate ($J = 2$) Diophantine equation

$$ax + by = N, \tag{15}$$

where $a := m_1, b := m_2$, and N are positive integers. Note that we have renamed the variables $x := x_1$ and $y := x_2$ for simplicity of notation.

Let L be the line defined by (15) for $(x, y) \in \mathbb{R}^2$. We are only interested in the set of nonnegative integer solutions to (15), that is, nonnegative lattice points $L \cap \mathbb{Z}_{\geq 0}^2$ on the line L . Every ideal in \mathbb{Z} can be generated by one element; in our case, this element is the greatest common divisor of a and b , which we will denote by $\gcd(a, b)$.

In particular, it follows that the Eq. (15) has integer solutions if and only if $\gcd(a, b)$ divides N . In addition, the description of all integral solutions readily follows by elementary algebra. Namely, suppose that $(x_0, y_0) \in \mathbb{Z}^2$ is one integer solution of $ax + by = N$. Then all other integer solutions are given by the following equation where q is an arbitrary integer:

$$\begin{cases} x = x_0 + \frac{b}{\gcd(a,b)} \cdot q \\ y = y_0 - \frac{a}{\gcd(a,b)} \cdot q \end{cases} \quad (16)$$

In fact, we can also estimate the number of solutions of (15). The geometry of the line provides that $x \in [0, N/a]$. From (16), it follows that x varies by multiples of $b/\gcd(a, b)$. Therefore, there are at most

$$\frac{N/a}{b/\gcd(a, b)} = \frac{N \cdot \gcd(a, b)}{ab}$$

points in $L \cap \mathbb{Z}_{\geq 0}^2$. Note that this is only an estimate, albeit a good one, since we are essentially counting only $\{x : ax + by = N \text{ for some } y\} \cap \mathbb{Z}$.

The code used for the analysis in this paper and additional examples are available at <http://www.stat.psu.edu/~sesa/cctable>.

The examples suggest that, in general, we are interested in the lattice points of the polytope obtained by intersecting the hyperplane defined by Eq. (2) and the positive orthant.

The code includes:

- A collection of functions we wrote in R for enumerating and counting the number of missing marginal and k -way tables given the partial information described in the paper. There are functions for (1) finding the greatest common divisor, (2) solving Diophantine equations, and (3) counting the number of tables.
- A sample R and LattE code for the examples in this paper, and some additional related examples.
- An additional example in support of Lemma 1.

Acknowledgments A. Slavković and X. Zhu supported in part by NSF grants SES-052407 and BCS-0941553 from the Pennsylvania State University. S. Petrović supported in part by grant FA9550-12-1-0392 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA). The authors would like to thank the reviewers for their valuable insights and comments.

References

- 4ti2 Team. (2014). 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. <http://www.4ti2.de>.
- Agresti, A. (2002). *Categorical data analysis*. New Jersey: Wiley.
- Aoki, S., Takemura, A. (2002). Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics*, 45, 229–249.

- Aoki, S., Takemura, A. (2008). Minimal invariant Markov basis for sampling contingency tables with fixed marginals. *Annals of the Institute of Statistical Mathematics*, 60, 229–256.
- Arnold, B., Castillo, E., Sarabia, J. (1999). *Conditional specification of statistical models*. Berlin: Springer.
- Barvinok, A. I. (1994). A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Mathematics of Operations Research*, 19, 769–779. doi:10.1287/moor.19.4.769.
- Barvinok, A., Luria, Z., Samorodnitsky, A., Yong, A. (2010). An approximation algorithm for counting contingency tables. *Random Structures and Algorithms*, 37, 25–66.
- Bishop, Y., Fienberg, S., Holland, P. (2007). *Discrete multivariate analysis*. New York: Springer.
- Chen, S., Li, N. (2007). On a conjecture about the number of solutions to linear diophantine equations with a positive integer parameter. arXiv:0710.0177.
- Chen, Y., Dinwoodie, I., Sullivan, S. (2006). Sequential importance sampling for multiway tables. *Annals of Statistics*, 34, 523–545.
- De Loera, J. A., Onn, S. (2006). Markov bases of three-way tables are arbitrarily complicated. *Journal of Symbolic Computation*, 41, 173–181. doi:10.1016/j.jsc.2005.04.010.
- De Loera, J. A., Hemmecke, R., Tauzer, J., Yoshida, R. (2004). Effective lattice point counting in rational convex polytopes. *Journal of Symbolic Computation*, 38, 1273–1302.
- Diaconis, P., Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26, 363–397.
- Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli*, 9(6), 1093–1108.
- Dobra, A. (2012). Dynamic Markov bases. *Journal of Computational and Graphical Statistics*, 21, 496–517.
- Dobra, A., Fienberg, S. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97, 11885–11892.
- Dobra, A., Fienberg, S. E. (2010). The generalized Shuttle algorithm. In M. R. P. Gibilisco, E. Riccomagno, H. Wynn, (Eds.), *Algebraic and geometric methods in statistics* (pp. 135–156). UK: Cambridge University Press.
- Dobra, A., Tebaldi, C., West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference*, 136(2), 355–372.
- Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., Zhou, Y. (2008). Algebraic statistics and contingency table problems: log-linear models, likelihood estimation and disclosure limitation. In M. Putinar, S. Sullivan, (Eds.), *IMA volumes and in mathematics and its applications: emerging applications of algebraic geometry*, (Vol. 149, pp. 63–88). New York: Springer.
- Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (2001). *Confidentiality, disclosure and data access*. USA: North Holland.
- Drton, M., Sturmfels, B., Sullivan, S. (2009). *Lectures on algebraic statistics. Oberwolfach seminars*, Vol. 40. Basel: Birkhäuser.
- Eisenbeis, C., Temam, O., Wijshoff, H. (1992). On efficiently characterizing solutions of linear Diophantine equations and its application to data dependence analysis. Technical Report No. RUU-CS-92-01. Utrecht University.
- Hemmecke, R., Malkin, P. N. (2009). Computing generating sets of lattice ideals and Markov bases of lattices. *Journal of Symbolic Computation*. 44, 1463–1476. <http://dblp.uni-trier.de/db/journals/jsc/jsc44.html#HemmeckeM09>.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). *Statistical disclosure control*. West Sussex: Wiley.
- Lassere, J. B., Zeron, E. S. (2007). Simple explicit formula for counting lattice points of polyhedra. In *IPCO '07: Proceedings of the 12th international conference on integer programming and combinatorial optimization* (pp. 367–381). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-72792-7_28.
- LattE. (2014). LattE machiato—lattice point enumeration. <http://www.math.ucdavis.edu/~mkoeppel/latte/>.
- Lauritzen, S. (1996). *Graphical models*. USA: Oxford University Press.
- Lazebnik, F. (1996). On systems of linear Diophantine equations. *Mathematics Magazine*, 69, 261–266. <http://www.jstor.org/stable/2690528>.
- Lee, J. (2009). Sampling contingency tables given sets of conditionals and marginals in the context of statistical disclosure limitation. PhD thesis, Penn State University.
- Malkin, P. (2007). Computing Markov bases, Grbner bases, and extreme rays. PhD thesis. <http://edoc.bib.ucl.ac.be:81/ETD-db/collection/available/BelNucetd-06222007-144602/unrestricted/thesis.pdf>.
- Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), 15324–15328.

- Morito, S., Salkin, H. M. (1980). Using the Blankinship algorithm to find the general solution of a linear Diophantine equation. *Acta Informatica*, 13, 379–382.
- Morton, J. (2013). Relations among conditional probabilities. *Journal of Symbolic Computation*, 50, 478–492. doi:10.1016/j.jsc.2012.02.005.
- R Development Core Team. (2005). R: a language and environment for statistical computing ISBN 3-900051-07-0. <http://www.R-project.org>.
- Richardson, T., Spirites, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 30(4), 962–1030.
- Sertoz, S. (1998). On the number of solutions of a Diophantine equation of Frobenius. *Discrete Mathematics and Applications*, 8, 153–162.
- Slavković, A. (2004). Statistical disclosure limitation beyond the margins: characterization of joint distributions for contingency tables. PhD thesis, Carnegie Mellon University.
- Slavković, A. (2009). Partial information releases for confidential contingency table entries: present and future research efforts. *Journal of Privacy and Confidentiality*, 1(2), 253–264.
- Slavković, A. B., Fienberg, S. E. (2004). Bounds for cell entries in two-way tables given conditional relative frequencies. In J. Domingo-Ferrer, V. Torra (Eds.) *Privacy in statistical databases—PSD 2004, lecture notes in computer science No. 3050* (pp. 30–43). Berlin: Springer.
- Slavković, A. B., Fienberg, S. E. (2010). Algebraic geometry of 2×2 contingency tables. In M. R. P. Gibilisco E. Riccomagno, H. Wynn (Eds.) *Algebraic and geometric methods in statistics* (pp. 63–81). UK: Cambridge University Press.
- Slavković, A., Lee, J. (2010). Synthetic two-way contingency table preserving conditional frequencies. *Statistical Methodology*, 7, 225–239.
- Slavković, A., Sullivant, S. (2006). The space of compatible full conditionals is a unimodular toric variety. *Journal of Symbolic Computation*, 41, 196–209.
- Smarandache, F. (2000). Integer algorithms to solver Diophantine linear equations and systems. <http://arxiv.org/abs/math/0010134>.
- Smucker, B., Slavković, A. (2008). Cell bounds in two-way contingency tables based on conditional frequencies. In J. Domingo-Ferrer, S. Ycel (Eds.) *Proceedings of the UNESCO Chair in Data Privacy International Conference—PSD 2008, lecture notes in computer science No. 5262* (pp. 64–76). Berlin: Springer.
- Smucker, B., Slavković, A., Zhu, X. (2012). Cell Bounds in multi-way contingency tables based on conditional frequencies. *Journal of Official Statistics*, 28, 121–140.
- Sturmfels, B., Weismantel, R., Ziegler, G. (1994). *Gröbner bases of lattices, corner polyhedra, and integer programming*. Berlin: Konrad-Zuse-Zentrum für Informationstechnik.
- Thibaudeau, Y. (2003). An algorithm for computing full rank minimal sufficient statistics with applications to confidentiality protection. In *Monographs of official statistics, work session on statistical data confidentiality*, Vol. 1. Luxembourg: Eurostat.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York: Wiley.