

A sample space of k -way tables given conditionals and their relations to marginals: Implications for cell bounds and Markov bases

Aleksandra B. Slavković*

e-mail: sesa@stat.psu.edu
and

Xiaotian Zhu*

e-mail: xxz131@psu.edu
and

Sonja Petrović

e-mail: petrovic@math.uic.edu

Abstract: The structure of the space of possible contingency table realizations under some constraints such as marginal totals has been studied in the algebraic statistics literature. The properties of the space of tables, i.e., fibers, are important for conducting exact conditional inferences, calculating cell bounds, imputing missing cell values, and assessing the risk of disclosure of sensitive information. We study the space, $\mathcal{F}_{\mathcal{T}}$, of all possible k -way contingency tables for a given sample size and set of conditional frequencies. This space can be decomposed according to different possible marginals, which, in turn, are encoded by the solution set of a linear Diophantine equation, giving $\mathcal{F}_{\mathcal{T}}$ a special structure. We obtain conditions under which two spaces of tables coincide: one is our space, $\mathcal{F}_{\mathcal{T}}$, and the other is the space of tables for a given set of corresponding marginal totals. This characterization of the difference between two fibers provides a solution to a generalization of an open problem posed by [Dobra *et al.* \(2008\)](#). Finally, the decomposition of $\mathcal{F}_{\mathcal{T}}$ has two important consequences (1) it leads to new cell bounds, some including connections to Directed Acyclic Graphs, and (2) it provides a structure for the Markov bases for the space $\mathcal{F}_{\mathcal{T}}$ that leads to a simplified calculation of Markov bases in this particular setting.

AMS 2000 subject classifications: 13P10, 62B05, 62H17, 62P25.

Keywords and phrases: Conditional tables, Diophantine equations, Directed Acyclic Graphs, Marginal tables, Markov bases, Optimization for cell entries..

Received September 2009.

*Supported in part by NSF grant SES-052407 to the Pennsylvania State University.

1. Introduction

Recent advances in the field of algebraic statistics have provided new mathematical tools for statistical inference, in particular for the analysis of categorical data. Algebraic statistics advocates the use of techniques from computational algebraic geometry, commutative algebra, and combinatorics to better understand the structure of statistical models, to improve statistical inference, and to explore new classes of models. Modern algebraic geometry was introduced to the field of statistics by [Diaconis and Sturmfels \(1998\)](#) in work on exact conditional tests, and by [Pistone, Riccomagno and Wynn \(2001\)](#) in work on experimental design. More recent literature includes work on contingency tables and log-linear models, sampling, latent class models, graphical models, and phylogenetic tree models, to name a few. For more details on algebraic statistics see [Drton, Sturmfels and Sullivant \(2009\)](#), [Gibilisco et al. \(2010\)](#), and references given therein.

In a paper by [Dobra et al. \(2008\)](#), the authors focus on the problem of maximum likelihood estimation for log-linear models in k -way contingency tables and a related problem of disclosure limitation strategies to protect against the identification of individuals associated with small counts in the tables. For an overview of the disclosure limitation literature for contingency table data see [Doyle et al. \(2001\)](#). There are a number of ways to assess the disclosure risk including (1) computing bounds for cell entries given a set of released marginals tied to a log-linear model, (2) enumerating all possible table realizations given some partial information extracted from the original table, and (3) sampling from a space of all possible tables, i.e., a fiber, to estimate posterior distributions. The properties of fibers are important for conducting exact conditional inference, and for calculating bounds on the cell entries. If we consider the original data as missing, then these techniques can be used to impute missing data in contingency tables and to create replacement tables; e.g., see [Slavkovic and Lee \(2009\)](#).

Through the underlying mathematics of the same geometric/algebraic framework given a set of released marginals, [Dobra et al. \(2008\)](#) make a more precise connection between the ideas on bounds and the existence of maximum likelihood estimates. They also discuss the more complex problem of releasing marginal and conditional information, and pose a series of open problems, some of which are addressed in this paper. The main open problem we investigate here is a mathematical description of a sample space of contingency tables given observed conditional frequencies, and their relations to corresponding marginals. Our results have implications for calculating bounds on the missing marginals, thus providing a shortcut to computing bounds on the cell entries of the original table. The results of this paper also have implications for characterizing a minimal Markov basis that would allow us to build a connected Markov chain and perform a random walk over all the points in the fiber. More specifically, we address the following challenge posed in [Dobra et al. \(2008\)](#), “*Problem 5.7. Characterize 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 margins.”

We provide extensions to the above-stated problem by considering combinations of conditional arrays and their relations to corresponding marginals via Directed Acyclic Graphs (DAGs); see Section 4 for relevant references on DAGs and for their use in calculating bounds for disclosure risk assessment. This work also relates to the characterizations of joint distributions (e.g., [Kagan, Linnik and Rao \(1973\)](#), [Arnold, Castillo and Sarabia \(1999\)](#), and [Slavkovic and Fienberg \(2009\)](#)). In this paper, we assume that the given sets of conditionals and marginals are compatible. Then we consider if they are sufficient to uniquely identify the existing joint distribution, and if not, we proceed with the description of the related sample space. We assume that the uniqueness theorem as stated in [Arnold, Castillo and Sarabia \(1999\)](#) holds. We allow cell entries to be zero as long as we do not condition on an event of zero probability. For problems on compatibility for categorical and continuous variables, see [Arnold, Castillo and Sarabia \(1999\)](#); on compatibility of full conditionals for discrete random variables, see [Slavkovic and Sullivant \(2006\)](#); and on generalization of compatibility of conditional probabilities in discrete cases, see [Morton \(2008\)](#).

We also derive new results on computing the cardinality of a fiber and enumerating all contingency tables in our setting. This problem is related to counting lattice points in polyhedra, which has 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 \(2007\)](#). We also need to count the solutions to a Diophantine equation (e.g., [Chen and Li \(2007\)](#) and [Eisenbeis, Temam and Wijshoff \(1992\)](#)). However, due to the simpler geometry of our problem, we do not need to use the general algorithms; we derive simpler formulas that estimate the number of lattice points instead.

This paper is organized as follows. In Section 2, we introduce the necessary concepts and notation. Section 3 studies the mathematical aspects of the space of k -way tables given the conditionals. The main observation, the Table-Space Decomposition Corollary 3.3, leads to derivation of formulas for the exact and approximate number of all possible marginal tables and k -way tables consistent with given conditionals, and to efficient description of the Markov bases (moves). Namely, the space of tables given the conditional is a disjoint union of spaces of tables given distinct marginals. Thus, the Markov moves consist of two sets: those that fix the margins, and those that change them. 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 et al. \(2007\)](#), [DeLoera and Onn \(2006\)](#), and references given therein. Less work has been done on studying Markov bases given observed (estimated) conditionals, e.g., see [Slavkovic \(2004\)](#); [Lee \(2009\)](#).

In Section 4, we describe the implications of the results derived in Section 3 on computing cell bounds and Markov bases. We also describe extensions via DAGs. In Section 5, we demonstrate our theoretical results with a series of simple examples. We show how to use our initial R ([R Development Core Team, 2005](#)) implementation of the formulas from Section 3. We also perform our analysis

using the well-established and free algebraic software LattE macchiato (LattE) which relies on an implementation of the Barvinok's algorithm (Barvinok et al., 2008) for counting and detecting lattice points inside convex polytopes. It is also used for finding solutions to integer programs. In statistical literature, LattE has been predominantly used for counting the number of tables given the margins.

2. Background and Notation

Consider k categorical random variables, X_1, \dots, X_k , where each X_i takes value on the finite set of categories $[d_i] \equiv \{1, \dots, d_i\}$. Let $\mathcal{D} = \bigotimes_{i=1}^k [d_i]$, $\mathbb{R}^{\mathcal{D}}$ be the vector space of k -dimensional arrays of format $d_1 \times \dots \times d_k$, 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_k) produces a random integer-valued array $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, called a k -way contingency table, whose coordinate entry n_{i_1, \dots, i_k} is the number of times the label combination, or cell, (i_1, \dots, i_k) is observed in the sample (see Agresti (2003); Bishop, Fienberg and Holland (2007); Lauritzen (1996) for details). The probability that a given cell appears in the sample is

$$p_{i_1, \dots, i_k} = Pr \{(X_1, \dots, X_k) = (i_1, \dots, i_k)\}, \quad (i_1, \dots, i_k) \in \mathcal{D},$$

and we denote the corresponding array in $\mathbb{R}^{\mathcal{D}}$ with \mathbf{p} . It will often be convenient to order the cells in some prespecified way (e.g., lexicographically) and to treat \mathbf{n} and \mathbf{p} as vectors in \mathbb{R}^d . For example, for a 3-way contingency table \mathbf{n} with $d_1 = d_2 = d_3 = 2$, or a $2 \times 2 \times 2$ table, we will use interchangeably the array notation $\mathbf{n} = (n_{111}, n_{112}, \dots, n_{222})$ and the vector notation $\mathbf{n} = (n_1, n_2, \dots, n_8)$. The marginal sums, or *marginal tables*, are obtained by collapsing the original table; e.g., two-way marginal sums can be denoted as $\{n_{ij+}\}, \{n_{i+k}\}, \{n_{+jk}\}$.

Let A, B be proper subsets of $\{X_1, X_2, \dots, X_k\}$, and $C = \{X_1, X_2, \dots, X_k\} \setminus (A \cup B)$. Then we can regard A, B, C as three categorical variables, and we can summarize \mathbf{n} as a 3-way table $\mathbf{n}^* := \{s_{ijk}\}$, where s_{ijk} is the count in the cell: $A = i, B = j, C = k$. We will use the following notation: if b is a subset of $K = \{1, \dots, k\}$, $|B| = d_b = J$ denotes the number of values B can take. Also let c_{ij} be the observed conditional frequency $P(A = i | B = j)$, such that $\sum_i P(A = i | B = j) = 1$, and suppose it equals $\frac{g_{ij}}{h_{ij}}$ for nonnegative and relatively prime integers g_{ij} and h_{ij} . If $C = \{\}$, we refer to c_{ij} 's as *full conditionals*, otherwise as *small* or *partial conditionals*.

There are two sides of algebraic statistics for contingency tables. One explores the representation of a statistical model by providing an alternative description of the parameter space. The set of all probability points \mathbf{p} lies in the standard $d-1$ simplex, such that $\log \mathbf{p}$ belongs to the row span of a given $m \times d$ design matrix M . For example, for a log-linear model of complete independence, the parameter space is a *toric variety* parametrized by a 0/1 matrix M . Multiplication by M encodes a minimal sufficient statistic: for a table n , $\mathbf{t} = \mathbf{M}\mathbf{n}$ is a vector of marginals: $\{n_{i++}\}, \{n_{i+k}\}, \{n_{++k}\}$. A standard reference for toric varieties from the computational perspective is Sturmfels (1996), and for an introduction to toric varieties of statistical models see Drton, Sturmfels and Sullivant (2009).

The sample space is the set \mathcal{F} of possible observable contingency tables, that is, the set of all non-negative integer-valued arrays in \mathbb{R}^d with entries summing to N . The other side of algebraic statistics is concerned with 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. It has been shown and discussed in the literature that many data-dependent objects encountered in the study of contingency tables are polyhedra; e.g., see (Dobra et al., 2008) in relation to log-linear models, and (Slavkovic and Fienberg, 2009) in relation to conditional probabilities.

In this paper, we focus on investigating the space of all possible tables consistent with the following given information:

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

Note that $P(A|B)$ can be either *full* or *partial* conditionals, and that all of the given frequencies are exact, and we do not observe values of B .

Let $\mathcal{T} = \{P(A|B), N\}$. Then the space of possible tables $\mathcal{F}_{\mathcal{T}}$ can be expressed as the integer solutions to a system of linear equations:

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

where \mathbf{n} and \mathbf{t} are length d column vectors, and M is a $(|B| + 1) \times d$ matrix that together with \mathbf{t} , describes the information encoded by the grand total and the conditional frequencies $P(A|B)$. When N is clear from the context, we use the shorthand notation $\mathcal{F}_{A|B}$ to denote $\mathcal{F}_{P(A|B), N}$.

In this paper, we (1) study the structure of the solution set $\mathcal{F}_{\mathcal{T}}$, (2) calculate the exact and approximate number of tables in $\mathcal{F}_{\mathcal{T}}$ and provide functions to do this in R , (3) explore the links between the space of tables given $\mathcal{F}_{A|B}$ and the space of tables given the corresponding margin, \mathcal{F}_{AB} . We provide extensions to \mathcal{T} in the form of sets of partial conditionals and to their relations to corresponding marginals via Directed Acyclic Graphs (DAG). We also discuss the implications of our results on calculation of bounds on the cell entries of \mathbf{n} given \mathcal{T} , and on the structure of Markov bases.

3. Mathematical Aspects of the Space of Tables Based on Given Conditional Frequencies

Data examples suggest a connection between the solutions to a certain Diophantine equation, defined below (2), and the space of tables $\mathcal{F}_{A|B}$ that we are interested in. In what follows, we establish this connection more rigorously. Finding solutions to Diophantine equations is a well-studied classical problem in mathematics, one that is generally hard to solve; e.g., see Chen and Li (2007) and Eisenbeis, Temam and Wijshoff (1992). But the equation (2) here turns out to be simple enough and can be analyzed using classical algebraic tools.

A solution set of a Diophantine equation (2) identifies the possible marginals B that we condition on in $P(A|B)$. Once we know the corresponding marginals AB , we can decompose the table space accordingly (see Section 3.1). An important question arises: How many marginals can there be for a given conditional? This question can be answered using a straightforward count of lattice points in a polyhedron (Section 3.2). This approach then allows us to derive the exact and approximate sizes of the space of possible tables $|\mathcal{F}_{A|B}|$ in Section 3.3.

Our approach also leads to decomposing of the Markov bases into two sets of moves. One of the sets fixes the marginal, and those types of moves are well understood. The other set of moves changes the marginal, and it turns out to be encoded by solutions of (2) below. As an important consequence, we are able to detect when having partial information in the form of conditionals is equivalent to having partial information in the form of corresponding marginals. This result also affects calculation of bounds and estimation of disclosure risk in data privacy problems.

3.1. The space of tables and a linear Diophantine equation

We begin by using the solution set of a linear Diophantine equation to identify marginals.

Theorem 3.1. *Adopt the notation of Section 2. Let m_j be the least common multiple of all h_{ij} for fixed j , and let $J = |B|$, the number of values that B takes. 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 Equation (2) has a nonnegative integer solution.

Remark 3.1. 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 (Section A). Since each solution of the Diophantine equation corresponds to a marginal we condition on, we easily obtain the following consequence:

Corollary 3.1. *Let $\mathcal{F}_{A|B}$ be the space of tables given $\mathcal{T} = \{P(A|B), N\}$, where we allow for full conditionals. In addition, let \mathcal{F}_{AB} and the space of tables given the corresponding $[AB]$ marginal counts s_{ij+} . Then, 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 $n_{i_1 i_2 \dots i_k}$ is $[0, s_{+j+} \cdot c_{ij}]$, and a different marginal $\{s_{+j+}\}$ leads to a different cell bound. When Corollary 3.1 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.

Corollary 3.2. *In the situation of Corollary 3.1, the integer cell bounds on \mathbf{n} based on these two spaces are the same. More specifically, $0 \leq s_{ijk} \leq s_{ij+}$, that is, $0 \leq n_{i_1, \dots, i_k} \leq n_{ab}$.*

Let us single out another very important consequence of Theorem 3.1, which we will refer to as the table-space decomposition result:

Corollary 3.3. *As in Corollary 3.1, let $\mathcal{F}_{A|B}$ be the space of tables given $\mathcal{T} = \{P(A|B), N\}$, where we allow for full conditionals. Suppose that the Diophantine equation (2) has m solutions. Denote by \mathbf{p}_i the marginal corresponding to the i^{th} solution. Thus, we will denote the space of tables given that particular marginal table by $\mathcal{F}_{AB}(\mathbf{p}_i)$.*

Then, we have the following decomposition of the table space taken as a disjoint union:

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

Remark 3.2. *We will see shortly that each \mathbf{p}_i for $1 \leq i \leq m$, representing a marginal, is actually an integer lattice point.*

To conclude this section, note that the proof of Theorem 3.1 shows that each solution to the Diophantine equation (2) corresponds to a marginal in the following way: $s_{+j+} = m_j x_j$; thus, $s_{ij+} = m_j x_j c_{ij}$. We will use this fact often.

3.2. The space of tables and integer points in polyhedra

By Theorem 3.1, solutions of Equation (2) correspond to marginals. To see the correspondence between these solutions and lattice points \mathbf{p}_i , we begin with two low-dimensional examples. They are generalized at the end of this section.

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

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

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

There are two important observations to be made. First, note that if we allow x and y to be any real number, then the equation (*) is just an equation of a

line in 2-space. Thus, the real solutions $(x, y) \in \mathbb{R}^2$ to the equation form a line L in \mathbb{R}^2 . Then, clearly, the integer solutions $(x, y) \in \mathbb{Z}^2$ are just the integral points, $L \cap \mathbb{Z}^2$, on the line L . These points are called the *lattice points on the line* L . In short, the solutions of the Diophantine equation (*) are in one-to-one correspondence with the lattice points on L . However, it is important to note that we are only interested in the set of *nonnegative* integral solutions to (*), and, therefore, the set of lattice points on the line L in the first quadrant only. These are the points depicted in Figure 1.



FIG 1. 2d lattice

Secondly, there is a nice algebraic way to view these lattice points. For $x, y \in \mathbb{Z}$, by definition, the expression $ax + by$ is an element of the *ideal* generated by a and b . In algebraic literature, this ideal is denoted by (a, b) . If $ax + by = N$ for some integer N , then this means that N is an element of the ideal (a, b) . It is an easy algebraic exercise to show that 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 $\gcd(a, b)$. Then, N being an element of the ideal (a, b) is equivalent to N being a multiple of $\gcd(a, b)$. In particular, it follows that the equation (*) 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 s is an arbitrary integer:

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

In fact, we can also estimate the *number of solutions* of (*). From the geometry of the line, we see that $x \in [0, N/a]$. From (**), 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 $\{x : ax + by = N \text{ for some } y\} \cap \mathbb{Z}$.

Example 2. Now suppose that $J = 3$, so that the Diophantine equation is of the form

$$ax + by + cz = N, \quad (\dagger)$$

where a, b, c , and N are positive integers. If (\dagger) is an equation in 3-space, and allow x, y , and z to be any real numbers, then it defines a plane P . Thus, any real solution of (\dagger) represents a point on P , while the integer solutions, $(x, y, z) \in \mathbb{Z}^3$ correspond to the *lattice points* in P . As we are interested in nonnegative solutions only, the solutions correspond to the finitely many integer lattice points in the part of the plane that lies in the *positive orthant* $\{(x, y, z) \in \mathbb{R}_{\geq 0}^3\}$.

The algebraic arguments from the bivariate case generalize, and give us an algebraic description of the solution set. More precisely, the equation (\dagger) has an integer solution (x, y, z) if and only if the greatest common divisor of a, b , and c , denoted by $\gcd(a, b, c)$, divides N (i.e., N is in the ideal (a, b, c)). Furthermore, if (x_0, y_0, z_0) is a particular solution, then we can obtain all other solutions as linear combinations using this one. More precisely, suppose x_1 and y_1 are such that $ax_1 + by_1 = \gcd(a, b)$ (such values can always be found by the Euclidean algorithm). Then all integer solutions of (\dagger) are given by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + s \cdot \begin{pmatrix} \frac{-x_1 c}{\gcd(a, b, c)} \\ \frac{-y_1 c}{\gcd(a, b, c)} \\ \frac{\gcd(a, b)}{\gcd(a, b, c)} \end{pmatrix} + t \cdot \begin{pmatrix} \frac{b}{\gcd(a, b)} \\ \frac{-a}{\gcd(a, b)} \\ 0 \end{pmatrix},$$

where s and t are integers. To simplify notation, let us name the vectors appearing in the above equation, respectively, as follows:

$$v = v_0 + s \cdot v_1 + t \cdot v_2.$$

Geometrically, if we let s and t vary over all integers, the above equation describes a sublattice \mathcal{L} of the 3-dimensional integer lattice \mathbb{Z}^3 whose generators are v_0, v_1 and v_2 . Thus \mathcal{L} is precisely the intersection of the plane $ax+by+cz = N$ with the integral points in \mathbb{R}^3 , while the positive solutions we are after are precisely those points in \mathcal{L} that lie in the positive orthant. To estimate the number of these solutions, we can argue similarly to the two-dimensional case, dividing the area of the triangle by the area of the parallelogram spanned by v_1 and v_2 . Thus, we get at most

$$\begin{aligned} & \frac{\sqrt{(\frac{N^2}{2ab})^2 + (\frac{N^2}{2ac})^2 + (\frac{N^2}{2bc})^2}}{|a \times b|} = \\ & \frac{\frac{N^2}{2abc} \sqrt{a^2 + b^2 + c^2}}{\frac{\sqrt{a^2 + b^2 + c^2}}{\gcd(a, b, c)}} = \\ & \frac{N^2 \cdot \gcd(a, b, c)}{2abc} \end{aligned}$$

points representing integral solutions of (\dagger) .

The above examples suggest that, in general, our focus should be on finding the integral lattice points in the polytope, which is the intersection of the hyperplane defined by Equation (2) and the positive orthant.

Remark 3.3. As noted in the Introduction, counting lattice points in polyhedra and counting the solutions to a Diophantine equation are interesting mathematical problems with a rich history. Due to the simpler geometry of our problem, we do not need to use the general algorithms, and, therefore, we derive simpler solutions. Also, we note that our Diophantine equation does not necessarily satisfy the main hypothesis of the main result from Chen and Li (2007).

The above two examples can be generalized to higher dimensions.

Lemma 3.2. *Suppose that the Diophantine equation (2) has a solution. Then there exist vectors $v_0, v_1, \dots, v_{j-1} \in \mathbb{Z}^j$ such that any solution $v = (x_1, x_2, \dots, x_j)$ of (2) is given as their integral linear combination:*

$$v = v_0 + \sum_{i=1}^{j-1} a_i \cdot v_i.$$

Note that we require that each $a_i \in \mathbb{Z}$.

The proof of this result which uses basic algebra (and some number theory), is elementary. For reader's convenience, it is included in Appendix A.

As mentioned previously, the set of all solutions to equation (2) is a $(J - 1)$ -dimensional lattice; this is a special case of a classical result that identifies the solution set of any system of linear Diophantine equations with a lattice Lazebnik (1996). As a subset of that lattice, the set of *nonnegative* solutions can be expressed as a linear combination of the elements in some basis of the lattice. In the proof of Lemma 3.2, 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). Given integers a and b , the Euclidean algorithm produces integers x and y such that $ax + by = (a, b)$. Repeatedly using this process, we get z_i 's such that $\sum_{i=1}^k m_i \cdot z_i = (m_1, m_2, \dots, m_k)$. Then set $x_i = z_i \cdot \frac{N}{(m_1, \dots, m_k)}$,

we get an integer solution of (2). The algorithm performs at most $\sum_{i=1}^k m_i$ steps of calculations. Function *solvedioph()* (Appendix B) uses a direct approach and it is slower. Section 5 includes examples of how these functions can be used, and the relevant code is available at <http://www.stat.psu.edu/~sesa/cctable>.

3.3. Size of table space: exact and approximate

First we derive the exact count formula for the total number of integer-valued k -way tables \mathbf{n} given the marginal $[AB]$. This count is combined, in Corollary 3.4, with the table-space decomposition results to derive the number of k -way tables in the fiber $\mathcal{F}_{A|B}$.

Consider a k -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 k -way tables, all having the same margin $[AB]$.

Lemma 3.3 (Exact count of data tables given one marginal). *Adopt the above notation. If $K = 2$, the number of 3-way tables that can have a given margin is*

$$\prod_{1 \leq i \leq I, 1 \leq j \leq J} s_{ij+} + 1.$$

In general, if $K \geq 3$, the number of possible data tables is

$$|\mathcal{F}_{AB}| = \prod_{1 \leq i \leq I, 1 \leq j \leq J} \left(\sum_{t_2=1}^{s_{ij+}+1} \sum_{t_3=1}^{t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} t_{k-1} \right).$$

Note that this expression simply counts the number of ways to write each entry s_{ij+} in the marginal table as a sum of k entries in the data table. As such, it equals

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

For the reader's convenience, a technical proof of Lemma 3.3 can be found in Appendix A.4

Remark 3.4. We can find s_{ij+} from the solutions of the Diophantine equation, since $s_{ij+} = x_j m_j c_{ij}$.

With this notion and real data in mind, we might have to alter the formula. 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 above can be written in terms of factorials, if we replace s_{ij+} with a real number instead of an integer, we get:

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

For an example, see Section 5.

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

Corollary 3.4 (Exact count of data tables given conditionals). *The number of possible k -way tables given observed conditionals $[A|B]$ is*

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

where $\mathcal{F}_{AB}(\mathbf{p}_i)$ is as defined in Corollary 3.3, and m is the number of integer solutions to (2). Each $|\mathcal{F}_{AB}(\mathbf{p}_i)|$ can be computed using Lemma 3.3.

Proof. When the linear Diophantine equation (2) has only one integer solution, we conclude using Lemma 3.3. If, on the other hand, the linear Diophantine equation (2) has more than one integer solution, we conclude by Corollary 3.3. \square

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 $\mathcal{F}_{A|B}$. We deal with the number of marginal tables first.

Proposition 3.4 (Approximate count of marginal tables given conditionals). *Returning to the notation of Lemma 3.2 given observed conditionals $[A|B]$, the number of possible marginal tables is approximately*

$$|\mathcal{F}_{A|B}| \approx \frac{N^{j-1} \gcd(m_1, m_2, \dots, m_j)}{(j-1)! \prod_{i=1}^j m_i}.$$

This number can also be approximated by an integral:

$$|\mathcal{F}_{A|B}| \approx \frac{\gcd(m_1, \dots, m_j)}{m_j} \int_{(x_1, \dots, x_{j-1}) \in \mathcal{M}} 1 dx_1 dx_2 \cdots dx_{j-1}.$$

A simple algebraic proof of this result can be found in Appendix A.3. Note that by Theorem 3.1, the number of possible marginal tables equals the number of positive integer solutions of Equation (2). The first formula uses a geometric approach via volumes of cells in the lattice; the second realizes the same approximation using the integral formula for volumes. Section 5 illustrates the use of these approximation formulas.

Corollary 3.5 (Approximate count of data tables given conditionals). *When the notation of Proposition 3.4 is adopted, the number of possible tables in $\mathcal{F}_{A|B}$ is approximately*

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

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 the ideas of Lemma 3.3 and Proposition 3.4. Note that the total number of tables equals the sum over all possible marginals of the number of tables for a fixed marginal. Since 3.4 gives an approximation to the latter, we obtain an approximate count of possible data tables given the conditional. \square

4. Implications for cell bounds and Markov bases

4.1. Cell bounds

There has been much discussion on calculation of bounds on cell entries given the marginals (e.g., see [Dobra and Fienberg \(2009\)](#) 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 Slavkovic \(2008\)](#). Such values are useful for determining the support of underlying probability distributions. In the context of data privacy and confidentiality, 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, and thus the bounds given in Corollary 3.1; and (3) using our decomposition result (c.f., Corollary 3.3) 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 omitted, but the idea is really simple. The integer cell bounds of \mathbf{n} are easily determined once we know the values the x_i 's of the linear Diophantine equation ((2)) take, and the i^{th} row of the collapsed $I \times J$ table sums up to at least the value of m_i . Based on the structure of the solution set, we know that that x_i is less than $\frac{1}{m_i}(N - \sum_{j \neq i} m_j)$. Thus, an approximate number of values that x_i can take is given by

$$\frac{(N - \sum_{j \neq i} m_j) \cdot (m_1, m_2, \dots, m_k)}{m_i \cdot (m_1, \dots, m_{i-1}, m_{i+1}, \dots, m_k)},$$

and the following result holds:

Theorem 4.1. *Given $\mathcal{T} = \{P(A|B), N\}$, an approximate (relaxation) integer cell bounds on are given by*

$$m_j \cdot c_{ij} \leq s_{ij+} \leq (N - \sum_{t \neq i} m_t) \cdot c_{ij}.$$

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, Slavkovic and Zhu \(2009\)](#) and [Lee \(2009\)](#).

4.2. Markov bases

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, Tebaldi and West \(2006\)](#) and [Chen, Dinwoodie and Sullivant \(2006\)](#). 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. Markov bases were introduced to the statistical community in a seminal paper by [Diaconis and Sturmfels \(1998\)](#) in which they used these bases for performing exact conditional inference over contingency tables given marginals. Let us recall the definition.

Definition 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:

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

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 Equation 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 . Diaconis and Sturmfels proved a fundamental theorem in algebraic statistics (Theorem 3.1. in [Diaconis and Sturmfels \(1998\)](#)): 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.

Of the computational algebra software packages for computing generating sets of toric ideals, the most efficient to date is 4ti2 ([4ti2 team](#)). Sometimes,

though, the matrix M can be large, and the computation may take too long. To alleviate some of the computational problems in practice, we use our table-space decomposition result (c.f. Corollary 3.3) to split the Markov basis into two sets. This could allow for parallel computation of the Markov sub-bases.

Corollary 4.2. *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 3.3, 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

Note that the first set of moves has been studied; for references to the literature, see Section 1. Since we know, by Theorem 3.1, that the margins correspond to solutions to the Diophantine equation (2), we can find the latter set of moves by computing the Markov basis for the coefficient matrix M of the Diophantine equation.

The number of Markov basis elements for the matrix M seems to be small. More specifically, computations suggest the following:

Conjecture 4.3. *In the case of small conditionals (i.e., $C \neq \emptyset$), the coefficient matrix M of the Diophantine Equation (2) has a Markov basis consisting of $k - 1$ elements, where $k - 1$ 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 5.2.2 shows. If the conjecture were true, it would imply the following:

Corollary 4.4 (Conjecture). *A minimal Markov basis of M in (1) contains $|B| - 1 + (|C| - 1) \times |B| \times |A|$ elements.*

On a related note, Peter Malkin has showed (personal communication, Malkin (2009)) 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, Weismantel and Ziegler (1995) and the Project-and-Lift method by Hemmecke and Malkin (Hemmecke and Malkin (2005)). Namely, Proposition 4.1. of Sturmfels, Weismantel and Ziegler (1995) 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).

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). Examples of Markov bases for our problem are included in Section 5.

4.3. Extension of relations to marginals via DAGs

In this section we explore Directed Acyclic Graphs (DAGs) as a tool for finding the bounds on missing cell counts. A DAG $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of nodes $V = \{v_1, \dots, v_k\}$ and a set of directed edges, $(v_i, v_j) \in E$, that link the ordered pairs of distinct nodes v_i, v_j in V . The vertex v_i is called the *parent*, and v_j the *child*. Let $pa(v_j)$ denote the *parent set* of v_j . A DAG does not contain a *cycle* if there does not exist a sequence of distinct vertices v_1, \dots, v_m for which $(v_i, v_{i+1}) \in E$ for each $i = 1, \dots, m-1$ that begins and ends in the same node. For more detailed definitions and properties of graphs, see Lauritzen (1996), Whittaker (1990), and Edwards (2000). 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} (Whittaker (1990); Lauritzen (1996)).

Let $\{X_1, X_2, \dots, X_k\}$ be a set of random variables with the joint distribution $f(X_1, X_2, \dots, X_k)$. If the random variables X_1, \dots, X_k are nodes of the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, then the graph represents dependencies among these random variables. More specifically, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ defines the set of probability distributions over the sample space that obeys the *directed Markov properties* and factorizes the joint distribution according to

$$f(x_1, x_2, \dots, x_n) = \prod_{x \in \mathcal{V}} f(x|pa(x)) = f(x_1)f(x_2|x_1)\dots f(x_n|x_{n-1}, x_{n-2}, \dots, x_1). \quad (3)$$

While in undirected graphs the missing edge represents the conditional independence of two variables given *all* other variables, in DAGs the missing edge represents the conditional independence given *all prior* variables. If there are no prior variables, this implies *marginal* independence in DAGs.

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

a set of marginals and conditional distributions with a DAG representation that satisfies the Wermuth condition, there is an equivalent undirected graph representation of the same set. In this case, the problem is reduced to one of knowing a set of marginals. Bounds for these problems then are those given by [Dobra and Fienberg \(2000, 2009\)](#). The following results hold for any k -way table.

Theorem 4.5. *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 4.6. *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 4.1. *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 4.2](#). \square

We demonstrate this and other results in the next section.

5. Examples

In this section we illustrate the results described in the preceding sections through analysis of a series of simple contingency tables.

5.1. 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.

TABLE 1

A $2 \times 2 \times 2$ table of counts of illegal MP3 downloading by gender and a residing building. The value in the brackets are linear relaxation bounds and sharp integer bounds given released conditional $[D|G]$ and marginal $[DG]$, respectively.

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

Consider a case where the only information we have about the original $2 \times 2 \times 2$ table is the marginal table of counts $[Download, Gender]$; e.g., see Table 2. Based on Lemma 3.3 the space of tables \mathcal{F}_{DG} has $16 \times 11 \times 6 \times 21 = 22176$ possible 3-way tables. The sharp integer bounds are given in red brackets in Table 1.

TABLE 2

$[Gender, Download]$ Marginal table of illegal MP3 downloading, and integer bounds given released $[Download|Gender]$ 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

We compare this space with the space of tables $\mathcal{F}_{D|G}$ defined by the grand total 50 and the small conditional $P(Download|Gender)$, e.g., see Table 3. The linear relaxation bounds are given in black brackets and the sharp integer bounds in blue brackets in Table 1.

TABLE 3

$[Download|Gender]$ Table of conditional probabilities with [rounded probability].

Gender	Download		Total
	Yes	No	
Male	$\frac{15}{25} = \frac{3}{5}$ [0.6]	$\frac{10}{25} = \frac{2}{5}$ [0.4]	25
Female	$\frac{5}{25} = \frac{1}{5}$ [0.2]	$\frac{20}{25} = \frac{4}{5}$ [0.8]	25
Total	20	30	50

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

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

This is part of a 5-dimensional lattice inside the \mathcal{R}^2 . Then equation (2) of Theorem 3.1 for this example is:

$$5x_1 + 5x_2 = 50.$$

It has 9 positive integer solutions: $\{(x_1 = i, x_2 = 10 - i) | 1 \leq i \leq 9\}$. Thus, there are 9 different [Download, Gender] marginals, which, by Theorem 3.1, means that the space of tables given the small conditional 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 3.4 for $m = 9$ provides the table count: $|\mathcal{F}_{D|G}| = \sum_{m=1}^9 |\mathcal{F}_{DG_m}| = 128767$. 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 $128767 + 651 + 451 = 129778$ where the latter two numbers before the equals sign present the number of possible 3-way tables given the $[DG]$ marginal where one of the rows of $[G]$ is equal to zero.

To approximate the number of marginal tables, one can use the formula from the second part of Proposition 3.4 to count the number of corresponding solutions to the Diophantine equation. Then, we can use the integral formula from Corollary 3.5, which could be evaluated, say, using Maple, to estimate the size of the total table space given the conditionals. Note that Proposition 3.4 gives an approximation of the number of marginal tables $[DG]$ by $\frac{50 * gcd(5,5)}{5 * 5}$, which is 10. Then the Corollary 3.5 gives an approximation of the total number of 3-way tables by $\frac{gcd(5,5)}{5} \int_0^{10} (3x + 1)(2x + 1)(10 - x + 1)(40 - 4x + 1) = 129676.7$.

Since more than 1 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 blue brackets, respectively. Given \mathcal{F}_{DG} , the exact cell bounds are in red brackets. These bounds are obtained by direct optimization for each given constraint. However, the results of Section 4, 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 4.1 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 blue 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 k -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, $b = \text{solvequick}(c(5, 5), 50)$ is a vector of all possible B margins that we conditioned on in $[A|B]$; in our case $B = \text{Gender}$. 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 4.3 is true, then so is Corollary 4.4, and there should be $5 = |G| - 1 + (|A| - 1) \times |G| \times |D| = 1 + 1 \times 2 \times 2$ 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 4.2, 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 3.1 for the number of possible marginals $[G]$, and thus of $[DG]$, is 10.

5.1.1. Modified MP3 Example

TABLE 4
A new $2 \times 2 \times 2$ Table on illegal MP3 downloading, and integer bounds.

Building	Gender	Download		Total
		Yes	No	
I	Male	9 [0,16]	4[0,10]	11
I	Female	1[0,4]	9[0,20]	10
II	Male	7[0,16]	6[0,10]	13
II	Female	3[0,4]	11[0,20]	14
Total		20	30	50

Suppose now that the original table is Table 4, and we only have information on the grand total 50 and the small conditional $P(\text{Download}|\text{Gender})$:

$$\begin{bmatrix} \frac{8}{13} & \frac{5}{13} \\ \frac{1}{6} & \frac{5}{6} \end{bmatrix}$$

Equation (2) of Theorem 3.1 for this example is:

$$13x_1 + 6x_2 = 50,$$

and it has only 1 positive integer solution: (2, 4); that is, there is only one $[\text{Download}, \text{Gender}]$ marginal. Then, by Theorem 3.1, the two table spaces, $\mathcal{F}_{D|G}$ and \mathcal{F}_{DG} , coincide. The $[DG]$ marginal has the following cell entries:

$s_{11+} = 16, s_{12+} = 10, s_{21+} = 4, s_{22+} = 20$ that we can use in the formula from Lemma 3.3 to get the exact size of this space:

$$|\mathcal{F}_{DG}| = |\mathcal{F}_{D|G}| = 17 \times 11 \times 5 \times 21 = 19635,$$

or we can use $\text{tablecount}(M, 2)$, where M is the $[DG]$ table of counts.

By Proposition 3.4 an approximate number of $[DG]$ tables is $\frac{50 * \text{gcd}(13,6)}{13*6} \approx 0.641$. In situations when the integral is less than equal to 1, we claim that only one marginal table will correspond to the released information. However, this claim needs more careful investigation, as we have not studied in detail the error in the integral approximation. Note that Corollary 3.5 would give an approximation to the total number of 3-way tables by

$$\frac{\text{gcd}(13,6)}{6} * \int_0^{50/13} (8x+1)(5x+1)((50-13x)/6+1)((50-13x)*5/6+1) = 7398.637.$$

In this case, the approximation formula performs poorly. In the next section, we explore this issue in more detail.

A Markov basis for this example has the following elements:

$$\begin{pmatrix} 48 & 30 & -13 & -65 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

Again, the first element indicates the amount by which we need to change the margin we condition on and the underlying margin $[DG]$. The sum of the exponents in the monomials of the first element is 78. Since this number is larger than the given $N = 50$, for this example we can conclude that we have only one possible contingency table that satisfies the sample size and the observed conditional rates. That is, the Markov move found is not applicable to our table. This deep issue is one of the difficulties of the field in regard to instances in which the statistical model in question is not the full toric variety for which we calculate Markov moves, but instead corresponds to its real positive part.

5.2. 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 5) with zero counts, and show the convergence of exact and approximate results.

5.2.1. 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 6. By Theorem 3.1, the linear Diophantine equation that characterizes all possible missing $[AB]$ margins is

$$2x + 3y + 4z = 240. \quad (4)$$

TABLE 5
A 3 × 2 × 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 6

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 5.

	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

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. Note that the triplets (x, y, z) are in 1 to 1 correspondence to $[AB]$ margins (see Table 6), and thus there are 1141 missing $[AB]$ marginals consistent with the provided information. Furthermore, `solvequick(c(2, 3, 4), 240)` lists all positive integer solutions to Equation (4), 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 3.4, $|\mathcal{F}_{B|A}| = \sum_{m=1}^{1141} |\mathcal{F}_{AB_m}|$. 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 3.4, 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:

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

where \mathcal{M} is the projection of all possible triple (x, y, z) onto the xy -plane. As discussed in the proof of Proposition 3.4, notice that \mathcal{M} is a part of a lattice whose unit cell has an area of $4/\gcd(2,3,4)$. Thus, the number of possible solutions is approximately

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

this is about $1.188479935 \times 10^{12}$ (computed by Maple).

Our simulations show that the ratio of the exact solution to the approximate solution, for either counting the missing margin or the k -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 7 summarized the results for the missing marginal $[AB]$, and Table 8 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. This supports our earlier observation in Example 1 that for a small number of margins, the approximation does not work well.

TABLE 7
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 8
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	$9.501190349 \times 10^{27}$

Next, we calculate a Markov basis for fixed $[B|A]$ using $4ti2$. According to Corollary 4.4, there should be 8 elements in this basis. A Markov basis for this example is given below. In accordance to Corollary 4.2, 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.

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

5.2.2. 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 Section 4.2, we only need to solve a linear Diophantine equation to find the total number of possible 3-way tables. Up until now we would typically count the number of possible solutions by setting up the full constraint matrix in LattE. However, now we can solve a much simpler equation in either LattE or R, e.g.,

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

using `solvequick(c(2, 3, 4, 5), 240)` in R; for LattE code see Appendix B. 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:

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

Note that Conjecture 4.3 about the number of elements in the basis does not hold here. We think that this is because we are using full conditionals, that is, $C = \emptyset$. As supported by other examples, this conjecture seems true for *small* conditionals.

5.2.3. 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 3.4 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.

5.2.4. 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. Then by Theorem 4.5 the bounds on the cell counts are the same as in the case of given margins $[AB]$ and $[AC]$; for bounds given marginals see Dobra and Fienberg (2009). Based on Corollary 4.1, the Markov bases will be the same, and so will the fibers \mathcal{F}_t and $\mathcal{F}_{AB, AC}$. Note that these theorems 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, $P(A, B, C) = P(AB)P(AC)$:

$$\mathcal{G} : \quad B \longleftarrow A \longrightarrow C$$

$$\mathcal{G}^u : \quad B \text{-----} A \text{-----} C$$

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 Theorems 4.5 and 4.1 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 5 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 6. By Theorem 3.1, the linear Diophantine equation that characterizes the missing marginal $[A]$ and thus $[AB]$ for $[B|A]$ is

$$2x + 3y + 4z = 24. \quad (5)$$

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

$$3x + 3y + 4z = 24, \quad (6)$$

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, $|\mathcal{F}_t| = |\mathcal{F}_{AB,AC}| = 36$. We can also solve the system of linear Diophantine equations in Latte and obtain the same result for the number of missing margins $[A]$. Our analysis shows that Theorems 4.1 and 4.5 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 Theorems 4.1 and 4.5 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 significantly 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 Theorems 4.1 and 4.5. 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>.

6. Conclusions

In this paper, we describe the space of all possible k -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. 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 4.3.

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.”

Appendix A: Proofs

A.1. Proof of Theorem 3.1

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 Equation (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. \square

A.2. Proof of Lemma 3.2

Proof. Elementary arguments, generalizing those of the previous two examples, allow us to express the vectors v_1, \dots, v_k in terms of the coefficients m_1, \dots, m_k .

As before, let (m_1, \dots, m_l) denote the greatest common divisor of m_1, \dots, m_l . Then the gcd can be expressed as a linear combination of the m_i 's; let us denote the coefficients by x_i^j , so that

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

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

$$x_l = x_l^{(0)} - \sum_{h=1}^{k-l} \frac{m_{k+1-h} x_l^{(h)}}{(m_1, \dots, m_{k+1-h})} \cdot s_h + \frac{(m_1, \dots, m_{l-1})}{(m_1, \dots, m_l)} \cdot s_{k-l+1} \text{ for } l = 2, \dots, k,$$

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

where $s_i \in \mathbb{Z}$ for all i with $1 \leq i \leq k-1$. Then the vectors v_h , for $h = 1, \dots, k$, are determined from these expressions as follows: the l^{th} coordinate of v_h is the coefficient of s_h in the expression for x_l . \square

A.3. Proof of Proposition 3.4

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

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

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 parallelotope spanned by v_1, v_2, \dots, v_{k-1} . Let's compute this value:

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

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

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

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.

Now we turn to the integral approximation formula. The exact number of positive integer solutions of the Diophantine equation (2) is

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

where \mathcal{P} is the projection of the set of positive integer solutions onto the $x_1 \dots x_{j-1}$ -plane. Recall that \mathcal{P} is part of a lattice. Let \mathfrak{a} be the area of the unit cell of this lattice. Multiplying the above sum by the area of the unit cell gives, by definition, a Riemann sum approximation of the following integral:

$$\int_{(x_1, \dots, x_{j-1}) \in \mathcal{M}} 1 dx_1 \dots dx_{j-1}.$$

Therefore, the sum we are interested in is approximated by this integral, divided by the area of the unit cell. Since \mathcal{P} is a part of a lattice which is a projection of another lattice, \mathcal{L} , we can choose its unit cell to be the projection of the unit cell of the lattice \mathcal{L} onto the $x_1 \dots x_{j-1}$ -plane. (Note that \mathcal{L} is the lattice of all integer solutions to Equation (2).)

To complete the proof, note that the projection of the unit cell onto the $x_1 x_2 \dots x_{j-1}$ -plane is a parallelepiped 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 $\frac{m_1}{\gcd(m_1, m_2, \dots, m_j)}$. \square

A.4. Proof of Lemma 3.3

Proof. The reader may simply use the definition of the binomial coefficient for this formula. But, for completeness, we prove the general formula. The proof relies on a simple partition count. We will use induction.

Consider the case when C is ternary (that is, $K = 3$). To count the number of 3-way tables that can have the given marginal table s_{ij+} , we need to count the number of ways each cell s_{ij+} , for $1 \leq i \leq I$ and $1 \leq j \leq J$, can be decomposed as a sum

$$s_{ij+} = s_{ij1} + s_{ij2} + s_{ij3},$$

subject to

$$0 \leq s_{ij1}, s_{ij2}, s_{ij3} \leq s_{ij+}.$$

Fix i and j . Notice that the value of s_{ij3} is determined by the other two values: $s_{ij3} = s_{ij+} - s_{ij1} - s_{ij2}$. Clearly, the number of choices for s_{ij1} is $s_{ij+} + 1$. Once the value for s_{ij1} is chosen, there are less choices for s_{ij2} ; in fact, s_{ij2} must be chosen in the range $0 \leq s_{ij2} \leq s_{ij+} - s_{ij1}$. Let us summarize:

- when $s_{ij1} = 0$, there are $s_{ij+} + 1$ ways to choose s_{ij2} ,
- when $s_{ij1} = 1$, there are s_{ij+} ways to choose s_{ij2} ,
- when $s_{ij1} = 2$, there are $s_{ij+} - 1$ ways to choose s_{ij2} ,
- and so on, until $s_{ij1} = s_{ij+}$, when there is no choice for s_{ij2} .

Thus the total number is $(s_{ij+} + 1) + s_{ij+} + \cdots + 2 + 1$. Letting i and j vary, we get the count for each cell in s_{ij+} . Since each cell can be decomposed in any of the $\sum_{t=1}^{s_{ij+}+1} t$ ways, the number of 3-way tables is

$$\prod_{1 \leq i \leq I, 1 \leq j \leq J} \sum_{t=1}^{s_{ij+}+1} t,$$

as predicted by the claim.

Assume the formula holds for $|C| = k - 1$. Then the number of ways to write

$$s_{ij+} = s_{ij2} + \cdots + s_{ijk}$$

subject to $0 \leq s_{ij2}, \dots, s_{ijk} \leq s_{ij+}$, for $1 \leq i \leq I$ and $1 \leq j \leq J$ is

$$\prod_{1 \leq i \leq I, 1 \leq j \leq J} \left(\sum_{t_3=1}^{s_{ij+}+1} \sum_{t_4=1}^{t_3} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} t_{k-1} \right)$$

(note one less summation, the indices start with t_3 instead of t_2 as in the general formula). Now we must count the number of ways to write

$$s_{ij+} = s_{ij1} + s_{ij2} + \cdots + s_{ijk},$$

subject to $0 \leq s_{ij1}, \dots, s_{ijk} \leq s_{ij+}$. Generalizing the strategy for the case $K = 3$, we track this number by choosing s_{ij1} for fixed i and j :

- when $s_{ij1} = 0$, induction hypothesis says that there are

$$\sum_{t_3=1}^{(s_{ij+}+1)-0} \sum_{t_4=1}^{t_3} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} t_{k-1}$$

ways to choose the remaining s_{ij2}, \dots, s_{ijk} values.

- Similarly, when $s_{ij1} = 1$, there are $\sum_{t_3=1}^{(s_{ij+}+1)-1} \sum_{t_4=1}^{t_3} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} t_{k-1}$ ways,
- and so on, until $s_{ij1} = s_{ij+}$.

In summary, as s_{ij1} varies from 0 to s_{ij+} , the top index of the first summation varies accordingly. Thus for fixed i and j , there are

$$\sum_{t_2=1}^{s_{ij+}+1} \sum_{t_3=1}^{t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} t_{k-1}$$

ways to decompose s_{ij+} into an (i, j) -slice of our 3-way table, and the claim follows. \square

Appendix B: Code

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

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.

References

- 4TI2 TEAM, 4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.
- AGRESTI, A. (2003). *Categorical data analysis*. Wiley-Interscience.
- AOKI, S. and 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*. v45 229–249.
- AOKI, S., AOKI, S., TAKEMURA, A. and TAKEMURA, A. (2007). Invariant minimal Markov basis for sampling contingency tables with fixed marginals. *Annals of the Institute of Statistical Mathematics*.

- ARNOLD, B., CASTILLO, E. and SARABIA, J. (1999). *Conditional specification of statistical models*. Springer Verlag.
- BARVINOK, A. I. (1994). A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math. Oper. Res.* **19** 769–779. Available at <http://dx.doi.org/10.1287/moor.19.4.769>
- BARVINOK, A., LURIA, Z., SAMORODNITSKY, A. and YONG, A. (2008). An approximation algorithm for counting contingency tables. *Random Structures & Algorithms, to appear; preprint*. Available at [arXiv:0803.3948](https://arxiv.org/abs/0803.3948)
- BISHOP, Y., FIENBERG, S. and HOLLAND, P. (2007). *Discrete multivariate analysis*. Springer.
- CHEN, Y., DINWOODIE, I. and SULLIVANT, S. (2006). Sequential Importance Sampling for Multiway Tables. *Annals of Statistics* **34** 523–545.
- CHEN, S. and LI, N. (2007). On a Conjecture about the Number of Solutions to Linear Diophantine Equations with a Positive Integer Parameter. Available at <http://www.citebase.org/abstract?id=oai:arXiv.org:0710.0177>
- DELOERA, J. A. and ONN, S. (2006). Markov bases of three-way tables are arbitrarily complicated. *J. Symb. Comput.* **41** 173–181. Available at <http://dx.doi.org/10.1016/j.jsc.2005.04.010>
- DIACONIS, P. and STURMFELS, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* 363–397.
- DOBRA, A. and 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.
- DOBRA, A. and FIENBERG, S. E. (2009). The Generalized Shuttle Algorithm. In *Algebraic and geometric methods in statistics* (M. R. P. Gibilisco E. Riccomagno and H. Wynn, eds.) Cambridge University Press, to appear.
- DOBRA, A., TEBALDI, C. and WEST, M. (2006). Data augmentation in multiway contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference* **136** 355–372.
- DOBRA, A., FIENBERG, S., RINALDO, A., SLAVKOVIC, A. and ZHOU, Y. (2008). Algebraic statistics and contingency table problems: Estimations and disclosure limitation. *Emerging Applications of Algebraic Geometry*.
- DOYLE, P., LANE, J., THEEUWES, J. and ZAYATZ, L. (2001). *Confidentiality, disclosure and data access*. North Holland.
- DRTON, M., STURMFELS, B. and SULLIVANT, S. (2009). Lectures on Algebraic Statistics. Oberwolfach Seminars, Vol. 40.
- EDWARDS, D. (2000). *Introduction to graphical modelling*. Springer Verlag.
- EISENBEIS, C., TEMAM, O. and WIJSHOFF, H. (1992). On efficiently characterizing solutions of linear Diophantine equations and its application to data dependence analysis Technical Report.
- GIBILISCO, P., RICCOMAGNO, E., ROGANTIN, M. and WYNN, H. (2010). *Algebraic and geometric methods in statistics*. Cambridge University Press.
- HEMMECKE, R. and MALKIN, P. (2005). Computing generating sets of lattice ideals. Available at <http://www.citebase.org/abstract?id=oai:arXiv.org:math/0508359>
- KAGAN, A., LINNIK, Y. and RAO, C. (1973). Characterization Problems in

- Mathematical. *Statistics*, Wiley, New York.
- LASSERRE, J. B. and 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* 367–381. Springer-Verlag, Berlin, Heidelberg. Available at http://dx.doi.org/10.1007/978-3-540-72792-7_28
- LATTE, LattE machiato—Lattice point Enumeration. Available at <http://www.math.ucdavis.edu/~mkoeppel/latte/>.
- LAURITZEN, S. (1996). *Graphical models*. Oxford University Press, USA.
- LAZEBNIK, F. (1996). On Systems of Linear Diophantine Equations. *Mathematics Magazine* **69** 261–266. Available at <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.
- MALKIN, P. (2007). PhD thesis. Available at <http://edoc.bib.ucl.ac.be:81/ETD-db/collection/available/BelnUcetd-06222007-144602/>
- MALKIN, P. (2009). Personal communication.
- MORTON, J. (2008). Relations among conditional probabilities. *Arxiv preprint arXiv:0808.1149*.
- PISTONE, G., RICCOMAGNO, E. and WYNN, H. (2001). *Algebraic statistics: computational commutative algebra in statistics*. CRC Press.
- R DEVELOPMENT CORE TEAM, (2005). R: A Language and Environment for Statistical Computing ISBN 3-900051-07-0. Available at <http://www.R-project.org>
- SLAVKOVIC, A. (2004). Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables.
- SLAVKOVIĆ, A. B. and FIENBERG, S. E. (2004). Bounds for cell entries in two-way tables given conditional relative frequencies. In *Privacy in Statistical Databases – PSD 2004, Lecture Notes in Computer Science No. 3050* (J. Domingo-Ferrer and V. Torra, eds.) 30–43. Springer-Verlag.
- SLAVKOVIC, A. B. and FIENBERG, S. E. (2009). Algebraic geometry of 2×2 contingency tables. In *Algebraic and geometric methods in statistics* (M. R. P. Gibilisco E. Riccomagno and H. Wynn, eds.) Cambridge University Press, to appear.
- SLAVKOVIC, A. and LEE, J. (2009). Synthetic Two-Way Contingency Table Preserving Conditional Frequencies. *Statistical Methodology, under revision*.
- SLAVKOVIC, A. and SULLIVANT, S. (2006). The space of compatible full conditionals is a unimodular toric variety. *Journal of Symbolic Computation* **41** 196–209.
- SMUCKER, B. and SLAVKOVIC, A. (2008). Cell Bounds in Two-Way Contingency Tables Based on Conditional Frequencies. In *Proceedings of the UNESCO Chair in data privacy international conference on Privacy in Statistical Databases* 64–76. Springer.
- SMUCKER, B., SLAVKOVIC, A. and ZHU, X. (2009). Cell Bounds in Multi-Way Contingency Tables Based on Conditional Frequencies. *Journal of Official Statistics—to be submitted*.

- STURMFELS, B. (1996). *Gröbner bases and convex polytopes*. American Mathematical Society.
- STURMFELS, B., WEISMANTEL, R. and ZIEGLER, G. M. (1995). Grbner Bases of Lattices, Corner Polyhedra, and Integer Programming.
- WHITTAKER, J. (1990). *Graphical models in applied multivariate statistics*. Wiley New York.