Contemporary Mathematics
Volume 374, 2005

# Lattice points, contingency tables, and sampling

Yuguo Chen, Ian Dinwoodie, Adrian Dobra, and Mark Huber

ABSTRACT. Markov chains and sequential importance sampling (SIS) are described as two leading sampling methods for Monte Carlo computations in exact conditional inference on discrete data in contingency tables. Examples are explained from genotype data analysis, graphical models, and logistic regression. A new Markov chain and implementation of SIS are described for logistic regression.

## 1. Introduction

This paper is partly a survey of some recent theory on statistical problems of discrete data, and partly a description of new results for problems of sparse contingency tables where existing theory is not adequate.

Counts from statistical experiments are put in contingency tables that may be considered vectors of nonnegative integers. These are typically frequencies of events from an experiment where two or more outcomes are possible in a series of trials. Algebraic and geometric theory of lattice points and polytopes become useful when one wants to make inferences about the statistical model in place during the sampling and the data is multidimensional. One is led to computations over a collection of tables with certain constraints that are often linear and define a polytope $S_0$ whose elements correspond to constrained tables of integers where each cell in a table is a dimension in the space containing the polytope.

The statistical ideas of conditional inference that make polytopes an essential sample space were developed by Ronald Fisher to deal with two fundamental statistical issues: to determine if a family of probabilities (a model) could include the prevailing probability distribution when the family involves several unknown parameters; and to compute measures of distance from the observed data to the collection of tables consistent with the model without using asymptotic approximations. Conditional inference is described in Agresti (1990). The number of lattice points in the polytope $S_0$ representing tables of interest may be $10^{20}$ or larger, and

1991 *Mathematics Subject Classification.* Primary 62F03, 13P10; Secondary 05A99.
*Key words and phrases.* Contingency table, Markov chain, Gröbner basis, toric ideal.
This work was supported by SAMSI under grant DMS-0112069. The first author was supported under NSF grant DMS 0203762. The second author was supported under NSF grant DMS-0200888.

over this set we will want to compute expectations $E_\pi(f(\mathbf{n}))$ for certain functions $f : S_0 \to \mathbf{R}$, and distributions $\pi$ that may be uniform or conditional on sufficient statistics, often the hypergeometric distribution. Sometimes there are formulas or *ad hoc* sampling methods for efficient Monte Carlo computation, but this only happens for certain distributions $\pi$ and special polytopes $S_0$. In general it is not efficient or not possible to list all the elements of $S_0$ for further exact computation.

All expectations of interest can be computed as accurately as desired with a random sample from $S_0$ from a known probability distribution that is positive over all elements of $S_0$. Generating the random sample is the goal. To this end we discuss two particularly useful methods for sampling from polytopes: Markov chains and sequential importance sampling (SIS).

Markov chains are usually easy to program and memory efficient and have been used for Monte Carlo computations for decades. They require a "Markov basis" to make them irreducible–to run through all tables so time averages approximate space averages. The basis can be described as a generating set for a toric ideal, which is one of the fundamental results of Diaconis and Sturmfels (1998). With an irreducible and aperiodic Markov chain in $S_0$ one has the ergodic theorem that allows us to approximate expectations with sample averages:

$$E_\pi(f) = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{n}_i)$$

where $\mathbf{n}_i$ are random tables with stationary distribution $\pi$. The size of $m$ for a good approximation is usually not clear. Theoretical results on the time to stationarity are hard to prove and hard to apply. Perfect sampling methods such as coupling-from-the-past that make unnecessary the analysis of convergence have not yet been found for most applications of conditional inference.

In some cases the Markov basis is very hard or impossible to compute completely. Odds-ratio models such as logistic regression are hard cases like this, and these are the examples we focus on here, beginning in §3. The constraints defining the polytope are Lawrence liftings, and the generators of the toric ideal are hard to enumerate and can have high total degree. We show that slightly larger sets of tables $S_1$, suggested by intuition, saturation, or primary decomposition, may be much easier to sample, and the set $S_0 \subset S_1$ can be studied by "conditioning." This idea is not new, but designing $S_1$ so it is not much larger than $S_0$ can use new technology from algebra. Our main new result is Theorem 3.1, which gives an efficient relaxation of the logistic regression problem that allows easy computations. The relaxation uses ideas of primary decomposition from work of Diaconis, Eisenbud, and Sturmfels (1998) and illustrates results of Hosten and Shapiro (2000).

In §4 we describe sequential importance sampling, with specific application to logistic regression. SIS has proven to be much more efficient than Markov chains for sampling from the uniform distribution from rectangular tables with fixed row and column sums. This has been shown in Chen, Diaconis *et al.* (2003) and in follow-up work. To run the Markov chain for this application, one increments the present table with a random increment of the form

$$\pm \begin{bmatrix} + & - \\ - & + \end{bmatrix}$$

and it can be shown that the resulting sequence of tables will visit all tables eventually. But the time to stationarity is long compared to the time required by SIS to go through the cells in sequence, sampling uniformly from an interval of possible values for each cell computed with up-to-date Fréchet bounds, and keeping track of weights to measure the variation from the uniform distribution. We make some informal connections between SIS and commutative algebra that could be further developed.

## 2. Notation and Examples

A family of positive probabilities $(\mu_\theta)_{\theta \in \mathbf{R}^p}$ on nonnegative integer vectors $\mathbf{n} \in Z_+^d$, whose entries sum to a known sample size $n$ and may have other constraints, is given by

$$\mu_\theta(\mathbf{n}) = h(\mathbf{n}) \frac{e^{\theta' A_0 \mathbf{n}}}{z_\theta}$$

where $A_0$ is a $p \times d$ nonnegative integer matrix, $z_\theta$ is a normalizing constant, $\theta \in \mathbf{R}^p$ is a real parameter, and $h(\mathbf{n}) \geq 0$ may involve multinomial coefficients. The set of integers will be denoted $Z$, and the nonnegative integers in dimension $d$ are denoted $Z_+^d$. The conditional distribution given statistics $A_0\mathbf{n} = \mathbf{t}$ is parameter free:

$$\mu_\theta(\mathbf{n} \mid A_0\mathbf{n} = \mathbf{t}) \propto h(\mathbf{n}),$$

defined on tables or lattice points $\mathbf{n} \in Z_+^d$ that satisfy $A_0\mathbf{n} = \mathbf{t}$ and that may also satisfy *a priori* constraints that combine into a single constraint matrix $A$, say

$$S_0 := \{\mathbf{n} \in Z_+^d : A\mathbf{n} = \begin{bmatrix} \mathbf{t} \\ \mathbf{s} \end{bmatrix}\}.$$

Our main goal is to sample from $S_0$ according to the conditional probability distribution proportional to $h(\mathbf{n})$, in order to compute expectations for tests of goodness-of-fit and parameter significance based on the theory of exact conditional inference.

A general method for constructing an irreducible chain was described in Diaconis and Sturmfels(1998). Suppose

$$G := \{\mathbf{x}^{\mathbf{a}_1} - \mathbf{x}^{\mathbf{b}_1}, \ldots, \mathbf{x}^{\mathbf{a}_g} - \mathbf{x}^{\mathbf{b}_g}\}$$

is a Gröbner basis of monomial differences for the toric ideal

$$I_A := \langle \mathbf{x}^{\mathbf{n}} - \mathbf{x}^{\mathbf{m}} : A\mathbf{n} = A\mathbf{m} \rangle$$

in $Q[\mathbf{x}]$. The vector increments represented by the differences of the exponents

$$M_G := \{\mathbf{a}_1 - \mathbf{b}_1, \ldots, \mathbf{a}_g - \mathbf{b}_g\}$$

chosen randomly with random signs will connect all points of the set $S_0$, eventually, so the process is an irreducible Markov chain. A generating set of binomials is sufficient for irreducibility (Diaconis and Sturmfels (1998), p. 375) , but anything less than a generating set could have two or more connected components within $S_0$, depending on the actual values of the constraints defining the polytope. A fundamental and useful result of Diaconis, Eisenbud, and Sturmfels (see Sturmfels (2002), p. 110) is that two tables $\mathbf{n}$ and $\mathbf{m}$ in the polytope $S_0$ will be connected by the Markov chain based on moves in some collection $C$ whose corresponding ideal $I_C \subset I_A$ if the binomial $\mathbf{x}^{\mathbf{n}} - \mathbf{x}^{\mathbf{m}} \in I_C$. If the collection $C$ is a Gröbner basis, the path between $\mathbf{n}$ and $\mathbf{m}$ can be constructed by long division. The theory of toric ideals, lattice bases, and the connection with Markov chains is explained in

Sturmfels (1996), and Diaconis and Sturmfels (1998). Some aspects of the theory are in Pistone, Riccomagno, and Wynn (2000). The algebra for Markov chains can be quite useful for understanding SIS– arguments in §4 will use notions of square-free lead terms and adjacent minors for establishing implementation details of SIS.

These Markov chains are similar to reflecting random walks, although tight corners of the polytope make a precise analogy difficult. Their convergence rates in simple cases can be estimated based on eigenvalue computations. The strongest results for random walks in lattice points are in Diaconis and Saloff-Coste (1998).

EXAMPLE 2.1 (Genotype Data). Genotype data is a table of counts of unordered pairs of alleles, like the following experimental data on Gaucher disease from a paper of Le Coutre *et al.* (1997). It is known that the genotype pair IVS2+1/IVS2+1 is lethal, and therefore constitutes a structural zero in the triangular table of genotypes (or possibly a missing entry, whose analysis is slightly different, but we will assume a structural zero). The table rows and columns normally are labelled with the allele names, but we omit these since they are not needed here.

$$\begin{array}{ccccccc} 0 & 5 & 2 & 1 & 0 & 0 & 10 \\ & 2 & 0 & 0 & 0 & 1 & 2 \\ & & 0 & 0 & 0 & 0 & 0 \\ & & & - & 0 & 0 & 0 \\ & & & & 0 & 0 & 1 \\ & & & & & 0 & 0 \\ & & & & & & 1 \end{array}$$

The probability model for Hardy-Weinberg equilibrium postulates that the cell probabilities are the result of independent combination of alleles. For parameters $\mathbf{p} = (p_1, \ldots, p_7)$ that give the population proportions of each allele, the (unconditional) probability $\mu_{\mathbf{p}}$ on upper triangular tables $\mathbf{n} = (n_{ij})_{1 \le i \le j \le 7}$ with fixed total sum of $n$ is given by the multinomial formula

$$\mu_{\mathbf{p}}(\mathbf{n}) = \binom{n}{(n_{ij})} p_1^{f_1} \cdot \ldots \cdot p_7^{f_7} \, 2^{\sum_{i<j} n_{ij}}$$

where $f_i$ is the number of times allele $i$ appears in the table. $f_1$ for example is $2 \times 0 + 5 + 2 + 1 + 0 + 0 + 10 = 18$, a sum over the genetic pairs that contain allele 1. These frequency counts are called sufficient statistics. The conditional probability distribution on tables with the same allele frequencies $\mathbf{f} = (f_1 = 18, f_2 = 12, \ldots, f_7 = 15)$, ignoring the structural zero, is given by

$$\mu_{\mathbf{f}}(\mathbf{n}) = \frac{\binom{n}{(n_{ij})} \, 2^{\sum_{i<j} n_{ij}}}{\binom{2n}{\mathbf{f}}},$$

where $n$ is defined by $2n = f_1 + f_2 + \cdots + f_7$ and can be interpreted as the number of individuals in the sample, who contribute a total of $2n$ alleles of 7 types. This model cannot hold with the structural zero as in the data above, which complicates both the model and the analysis. We show how to modify the analysis to handle the structural zero.

The constraints that come from fixing the sufficient statistics for Hardy-Weinberg proportions are entry-wise dot products with 7 "row vectors" in $A$ like

$$\begin{array}{ccccccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ & 2 & 1 & 1 & 1 & 1 & 1 \\ & & 0 & 0 & 0 & 0 & 0 \\ & & & - & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & & 0 \end{array}$$

which counts the number of alleles of type 2 in the table. We also force entry $(4,4)$ to be zero. With $\mathbf{f} = (f_1 = 18, f_2 = 12, \ldots, f_7 = 15)$ the allele counts, our polytope $S_0$ for the exact analysis with structural zero is the set of triangular tables that have these same allele counts and satisfy the additional linear constraint $n_{4,4} = 0$. The conditional probability distribution on $S_0$ is defined by $\mu_{\mathbf{f}}(\mathbf{n}) \propto \binom{n}{(n_{ij})} 2^{\sum_{i<j} n_{ij}}$. The Markov chain of Guo and Thompson (1992) for the original test of Hardy-Weinberg equilibrium can be described as the collection of moves that arise by folding the traditional $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ minors over the diagonal: $g(\mathbf{n})_{ij} = n_{ij} + n_{ji}$, $i < j$, $g(\mathbf{n})_{ii} = n_{ii}$:

$$x_{11}x_{34} - x_{14}x_{31} = \begin{array}{cccc} + & 0 & 0 & - \\ 0 & 0 & 0 & 0 \\ - & 0 & 0 & + \\ 0 & 0 & 0 & 0 \end{array} \quad g \quad \mapsto \quad \begin{array}{cccc} + & 0 & - & - \\ & 0 & 0 & 0 \\ & & 0 & + \\ & & & 0 \end{array} = x_{11}x_{34} - x_{13}x_{14}.$$

To handle the structural zero at entry $(4,4)$ we can use a Lawrence lifting to get a larger collection of Markov moves. The folded images under $g$ of the well-known "circuit moves" that are the universal Gröbner basis for the independence model (fixed row and column sums) give an irreducible chain in the upper triangular tables with the given constraints, with arbitrarily placed zeros. These moves described differently appear in Takemura and Aoki (2002), but a proof of irreducibility can be made quite simple by using the algebraic description above.

Computationally, it is easiest for this example to just produce the binomials corresponding to the Markov basis from the constraint matrix by using a saturation algorithm, such as the one implemented in Cocoa, because there are fewer than 200 total moves. The calculation requires simply typing the constraint matrix $A$, with $49 = 7 \times 7$ columns and with 7 rows for the allele constraints, one row for forcing the lower triangle to be zeros, and one for forcing entry $n_{44}$ to be zero:

```
Use R::=Q[x[1..7,1..7]];
Toric(A);
```

The work of De Loera, Haws *et al.* (2003) and the Latte software can be used to enumerate the elements of the polytope $S_0$. The value for statistics of enumeration is significant. It can be used to benchmark sequential importance sampling, which requires some fine tuning usually that can be done with an enumeration step. Enumeration can help understand convergence to stationarity of Markov chains, which depends partly on the number of points in the polytope as well as other geometric features such as diameter and shape. And enumeration can show the dependence of the size of the polytope on variations in the constraint matrix and constraint

values. A typical problem of this type is feasibility, or whether the polytope $S_0$ is nonempty for a particular constraint vector $\mathbf{t}$. This is important for applications in SIS and disclosure limitation.

There is an efficient sequential method that fills in entries successively with conditional distributions, which can also handle one diagonal zero. This is described in Huber, Chen *et al.* (2003), and has the advantage that its complexity does not depend on the sizes of the table entries, but only the number of cells. Lazzeroni and Lange (1997) have extensions to multi-locus data and stopping times for exact sampling.

EXAMPLE 2.2 (Graphical Model). We illustrate with an example from the monograph of Sturmfels (2002) which describes many new results on graphical models. Consider the notation $X_1 - -X_2 - -X_3 - -X_4$, where $X_i$ is a random variable that takes values in $\{0,1\}$. The notation is meant to suggest that $X_1$ is independent of $\{X_3, X_4\}$ given $X_2$, and $\{X_1, X_2\}$ are independent of $X_4$ given $X_3$. The data would be a four-way table $\mathbf{n} = (n_{ijkl})$ with $2^4 = 16$ cells each containing a count. The notation $\mathbf{n}_{\{1,2\}}$ used by statisticians is the 2-way table that is the projection onto dimensions 1 and 2, with $\mathbf{n}_{\{1,2\}}(00) = \sum_{kl} \mathbf{n}_{00kl}$, called a margin.

The family of positive probabilities $\left(p_\theta(ijkl)\right)_{\theta \in \mathbf{R}^{12}} \subset \mathbf{R}^{16}$ on the space of $2^4$ outcomes that satisfy these independence conditions is parameterized by an exponential map

$$p_\theta(ijkl) = \frac{e^{(\theta_{12}|\theta_{23}|\theta_{34})A_{(\cdot,ijkl)}}}{z_\theta}$$

where $\theta_{12}$ is a real vector of four components, $z_\theta$ is a normalizing constant, and $A$ is the $12 \times 16$ matrix with columns indexed by $0000, 0001, 0010, \ldots, 1111$ and rows indexed by 2-way margin summaries for each of the 3 cliques $\{1,2\}, \{2,3\}, \{3,4\}$ written by statisticians as $\mathbf{n}_{\{1,2\}}, \mathbf{n}_{\{2,3\}}, \mathbf{n}_{\{3,4\}}$:

$$A = \begin{array}{c} \\ \mathbf{n}_{\{1,2\}}(00) \\ \mathbf{n}_{\{1,2\}}(01) \\ \mathbf{n}_{\{1,2\}}(10) \\ \mathbf{n}_{\{1,2\}}(11) \\ \mathbf{n}_{\{2,3\}}(00) \\ \mathbf{n}_{\{2,3\}}(01) \\ \mathbf{n}_{\{2,3\}}(10) \\ \mathbf{n}_{\{2,3\}}(11) \\ \mathbf{n}_{\{3,4\}}(00) \\ \mathbf{n}_{\{3,4\}}(01) \\ \mathbf{n}_{\{3,4\}}(10) \\ \mathbf{n}_{\{3,4\}}(11) \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} 0000 & 0001 & 0010 & & & & & & & & & & & & & 1111 \end{array} \\ \left( \begin{array}{cccccccccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}$$

Each probability point in $\mathbf{R}^{16}$ in the parameterization lies in the toric variety of the toric ideal $I_A \subset Q[x_{0000}, \ldots, x_{1111}]$. The parameterization seems at first to have 12 parameters, but there are in fact only 7 free parameters which is the dimension of the projective variety. To sample from tables with fixed margins $\mathbf{n}_{\{1,2\}}, \mathbf{n}_{\{2,3\}}, \mathbf{n}_{\{3,4\}}$, one can run a uniform Markov chain with vector increments that are the differences of the exponents in a generating set of binomials for the toric ideal $I_A$. An important result of Sturmfels says one can obtain the toric ideal with 20 generators by saturating the smaller ideal of 12 "pairwise independence relations." These are obtained by fixing the values of variables $\{2,4\}$ in 4 ways for

specifying conditional independence of variables $\{1,3\}$, then considering variables $\{1,4\}$ fixing $\{2,3\}$ and finally considering variables $\{2,4\}$ fixing $\{1,3\}$. Then $I_A = IP : p^\infty$ where $IP =$
$\langle x_{0000}x_{1010} - x_{0010}x_{1000}, x_{0001}x_{1011} - x_{0011}x_{1001}, x_{0100}x_{1110} - x_{0110}x_{1100},$
$x_{0101}x_{1111} - x_{1101}x_{0111}, x_{0000}x_{1001} - x_{0001}x_{1000}, x_{0010}x_{1011} - x_{1010}x_{0011},$
$x_{0100}x_{1101} - x_{0101}x_{1100}, x_{0110}x_{1111} - x_{0111}x_{1110}, x_{0000}x_{0101} - x_{0001}x_{0100},$
$x_{0010}x_{0111} - x_{0011}x_{0110}, x_{1000}x_{1101} - x_{1001}x_{1100}, x_{1010}x_{1111} - x_{1011}x_{1110}\rangle$
and $p = \prod_{ijkl} x_{ijkl}$.

An efficient way to build the Markov chain for a large class of graphical models is described in Dobra and Sullivant (2003). Applications of the Markov chains for graphical and more general log-linear models are in conditional statistical inference, and in the developing area of disclosure limitation (Duncan, Fienberg *et al.* (2001)). This area will provide some computational challenges since some applications involve tables of more than ten factors, resulting in at least $2^{10}$ indeterminates in the polynomial ring for simple two-level factors.

## 3. Logistic Regression

Hosmer and Lemeshow (1989, p. 3) present data that relates presence or absence of coronary heart disease of 100 patients to age. The age covariate extends from year 20 to year 69. The data can be summarized in a table that looks like

| Age: | 1 | 2 | 3 | 4 | ..... | $c$ |
|---|---|---|---|---|---|---|
| yes: | $n_{1,1}$ | $n_{1,2}$ | $n_{1,3}$ | $n_{1,4}$ | ..... | $n_{1,c}$ |
| no: | $n_{2,1}$ | $n_{2,2}$ | $n_{2,3}$ | $n_{2,4}$ | ..... | $n_{2,c}$ |
| | $n_{+,1}$ | $n_{+,2}$ | $n_{+,3}$ | $n_{+,4}$ | ..... | $n_{+,c}$ |

The data of Hosmer and Lemeshow has $c = 50$ columns, one for each age level, some of which have 0 counts in both the rows, meaning no one in the study had that age level. A simple statistical model for evaluating the effect of age on the presence of coronary heart disease is the logistic regression model, which specifies that $\mu_{\alpha,\beta}(\mathbf{n}) \propto e^{(\alpha,\beta)\cdot A_0\cdot(n_{1,\cdot})}$ where $A_0 = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & c \end{pmatrix}$ and $(n_{1,\cdot})$ is the top row of data as a column vector. The constraints for conditional inference are: the total number of successes $T_1(\mathbf{n}) := n_{1,+}$ (top row sum) must be fixed at integer $t_1 \geq 0$; and the weighted sum $T_2(\mathbf{n}) := (1,2,\dots,c)\cdot(n_{1,\cdot}) = \sum_{i=1}^{c} i \cdot n_{1,i}$ must be fixed at integer $t_2 \geq 0$; and the column sums $n_{+,i}$ must be fixed at integers $c_i \geq 0$, which comes either from a design constraint on the number of subjects at each age, or a conditioning constraint in an odds-ratio model. With the data ordered $\mathbf{n} = (n_{11},\dots,n_{1,c},n_{2,1},\dots,n_{2,c})$ these can be built into a single constraint matrix

$$A = \begin{pmatrix} A_0 & \mathbf{0} \\ I_{c\times c} & I_{c\times c} \end{pmatrix}. \tag{1}$$

The Markov chains for computing in the set $S_0$ of nonnegative tables with constraints from $A$ have been studied in Diaconis, Graham, and Sturmfels (1996). Their conclusion is that an irreducible chain in the collection of tables with arbitrary fixed column sums (possibly zero), and arbitrary fixed $n_{1,+}, \sum_{i=1}^{c} i\, n_{1,i}$ consists of vector increments that correspond to homogeneous primitive partition identities (hppi's), such as $2+2 = 1+3$. Computing these moves is difficult because their number grows in $c$ faster than any polynomial, and their degree (the total number of +'s in

the vector increment) grows linearly in $c$. The number of moves corresponding to hppi's as a function of $c = 3, 4, 5, \ldots$ is $1, 5, 16, 51, 127, 340, 798, \ldots$ and this number has only been computed up to $c = 20$ using software 4ti2 of Ralf and Raymond Hemmecke. This will not help with a data set of 50 columns. The network method of Mehta, Patel, and Senchaudhuri (2000) will handle this data set, but the memory requirements are large compared to the Markov chain described below, and it is likely that some larger, more complex data sets may not be possible with the network method.

Consider the following collection $M$ of $\binom{c-1}{2}$ vector increments, which together with their negative images form moves for a Markov chain in tables (nonnegative integer lattice points) with the above constraints.

- Choose an ordered pair of columns $1 \leq i < j \leq c$.
- Put a $+$ in the top row in columns $i, j$ and put a $-$ in adjacent columns $i+1$ and $j-1$.
- Put the opposite signs in the bottom row.

With $c = 6$, there are 10 such moves like

| | | | | | |
|---|---|---|---|---|---|
| + | - - | + | 0 | 0 | 0 |
| - | + + | - | 0 | 0 | 0 |

,

| | | | | | |
|---|---|---|---|---|---|
| 0 | + | - | 0 | - | + |
| 0 | - | + | 0 | + | - |

,

etc. These moves can be seen as differences of adjacent minors, and also as a subset of the partition identities. The first table above corresponds to the hppi $1 + 3 = 2 + 2$, and the second corresponds to $2 + 6 = 3 + 5$. The collection $M$ is generally a strict subset of hppi's of degree four, because the adjacent $+, -$ changes do not include three degree 4 hppi's $(1 + 6 = 3 + 4, 2 + 6 = 4 + 4, 1 + 5 = 3 + 3)$. The collection $M$ does not include thirty-eight other hppi's of higher degree.

Let $I_M$ denote the ideal in $Q[x_1, \ldots, x_c, y_1, \ldots, y_c]$ generated by the monomial differences corresponding to the moves described above: $I_M = \langle x_1 x_3 y_2^2 - x_2^2 y_1 y_3, \ldots \rangle$.

These vectors $M$ are a lattice basis for the kernel of the constraint matrix $A$, so the ideal $I_M$ saturates to the toric ideal $I_A$. Also, note that if we leave off the bottom row corresponding to the $y$-variables and work in $Q[x_1, \ldots, x_c]$, then the collection of moves on the top row corresponds in fact to a Gröbner basis for lex order for the toric ideal corresponding to $A_0 = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 3 & \ldots & c \end{pmatrix}$ with square-free exponent on the lead indeterminate.

PROPOSITION 3.1. Let $S_0 = \{\mathbf{n} \in Z_+^d : A\mathbf{n} = (t_1, t_2, \mathbf{c})'\}$ be the set of nonnegative integer tables with fixed row sum $n_{1,+}$, fixed weighted sum $\sum i\, n_{1,i}$ and fixed column sums $n_{+,i}$. Let $S_2 = \{\mathbf{z} \in Z^d : A\mathbf{z} = (t_1, t_2, \mathbf{c})', z_{1,i} \geq 0, z_{2,i} \geq -1\}$. Then the Markov chain with moves $M$ connects any two tables in $S_0$ through $S_2$.

The result is a corollary of Theorem 3.1 below, so a careful proof will not be included. Although it appears to be very similar to the Theorem, its computational value is less.

In the data set of Hosmer and Lemeshow with 50 columns the set $S_2$ is about $10^6$ times as big as $S_0$, so calculations are possible by conditioning on $S_0$. The goodness-of-fit calculation required several hours of time on a 48-node Linux cluster.

Proposition 3.1 also follows from a saturation property of the ideal $I_M$. Since

$$I_M : (y_1 y_2 \cdots y_c) = I_A,$$

it follows that any two tables $\boxed{\begin{matrix}\mathbf{m}_1\\ \hline \mathbf{m}_2\end{matrix}}$ and $\boxed{\begin{matrix}\mathbf{n}_1\\ \hline \mathbf{n}_2\end{matrix}}$ in $S_0$ satisfy $(y_1 \cdot y_2 \cdot \ldots \cdot y_c)(\mathbf{x}^{\mathbf{n}_1}\mathbf{y}^{\mathbf{n}_2} - \mathbf{x}^{\mathbf{m}_1}\mathbf{y}^{\mathbf{m}_2}) \in I_M$ implying by results of Sturmfels (1996) that $\boxed{\begin{matrix}\mathbf{m}_1\\ \hline \mathbf{m}_2+\mathbf{1}\end{matrix}}$ (the two tables with 1 added to each cell in the bottom row) can be connected through nonnegative integer points to $\boxed{\begin{matrix}\mathbf{n}_1\\ \hline \mathbf{n}_2+\mathbf{1}\end{matrix}}$. By comparison, the ideal for the Markov moves (that are also a lattice basis) of Bigatti *et al.* (1999) saturate slowly (with 5 columns, saturation occurs at step six).

A more careful look at the Markov chain with moves $M$ above shows that it is irreducible even when only the bottom row entries in columns whose sums are fixed at 0 are allowed to drop down to -1, which is a significant computational advantage. If $S_1$ is this polytope that contains $S_0$ for the data set with 50 columns, we computed numerically by comparing time averages that $|S_1| \approx 10^3|S_0|$, which is a factor of $10^3$ better than $|S_2|/|S_0|$.

THEOREM 3.1. Let

$$S_1 := \{\mathbf{z} \in Z^d : A\mathbf{z} = (t_1, t_2, \mathbf{c})', z_{1,i} \geq 0, z_{2,i} \geq -1 \text{ if } c_i = 0, z_{2,i} \geq 0 \text{ if } c_i \geq 1\}$$

be the set of integer tables with the desired constraints, but allowing bottom row entries to be -1 in columns $i$ where $c_i = 0$. Then the Markov chain with moves $M$ connects any two tables in $S_0$ through $S_1$. It is irreducible in $S_0$ if all column sums $c_i$ are positive.

PROOF. Assume for the moment the second assertion of irreducibility in $S_0$ if the column sums are positive. If this were true, and we wanted to connect nonnegative integer vectors $\mathbf{n}$ and $\mathbf{m}$ in $S_1$, then we could connect $\mathbf{n} + \boxed{\begin{matrix}\mathbf{0}\\ \hline I_{\{i:c_i=0\}}\end{matrix}}$ to $\mathbf{m} + \boxed{\begin{matrix}\mathbf{0}\\ \hline I_{\{i:c_i=0\}}\end{matrix}}$ as points in $S_0$ with constraint values $t_1, t_2, \mathbf{c} + I_{\{i:c_i=0\}}$. Then the path from $\mathbf{n}$ to $\mathbf{m}$ is obtained by subtracting $\boxed{\begin{matrix}\mathbf{0}\\ \hline I_{\{i:c_i=0\}}\end{matrix}}$ from each intermediate table in the connecting path. Thus the first assertion about irreducibility in $S_1$ follows from the second.

The second assertion is proved by showing that the $L^1$ distance between two tables can always be reduced by using one of the moves. One can show that a path between two tables is possible with length at most $c^2 \cdot t_1$.

There is an algorithm for constructing the connecting path, assuming $c_i \geq 1$. The moves of adjacent minors $\{x_i y_{i+1} - x_{i+1} y_i\}$ connect all $2 \times c$ tables with positive column sums and the same row sums. Further, they are a Gröbner basis for their ideal in two term orders: lex order, reading left to right, and weighted term order with weight vector $w = (1, 2, \ldots, c, 0, 0, \ldots, 0)$, with lex for ties. Consider two tables coded as a binomial with $\mathbf{x}^{\mathbf{n}_1}\mathbf{y}^{\mathbf{n}_2} - \mathbf{x}^{\mathbf{m}_1}\mathbf{y}^{\mathbf{m}_2}$ The lead term of one of the adjacent minors divides its lead term, in lex order. Do the division and save the adjacent minor. Now, divide the lead term of the intermediate dividend in weighted term

order, to lower the weight that just increased by 1 using also an adjacent minor with a "−" sign. The two divisions yield a pair of adjacent minors that leave the weights fixed. The division will terminate by connecting the two tables. As an example, consider connecting the two tables

| 1 | 2 | 0 | 0 | 2 |
|---|---|---|---|---|
| 0 | 1 | 3 | 1 | 0 |

,

| 0 | 1 | 3 | 1 | 0 |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 2 |

.

The procedure is a sequence of pairs of divisions that proceeds as below, with the divisors on the right:

```
0)lead  12002   lex 11102  weight   ;  0-+00,  000+-
        00310    -> 01210    ->        0+-00   000-+

1)lead  11111   lex 02111  weight   ;  -+000,  00+-0
        01201    -> 10201    ->       +-000   00-+0

2)lead  02201   lex 01301  weight   ;  0-+00,  000+-
        10111    -> 11011    ->       0+-00   000-+

3)end   01310
        11002
```

□

The result above can be seen in the primary decomposition of $I_M$ as presented (in edited form) by Singular (Greuel *et al.* (2001)) below for five columns. At this time there do not appear to be general theorems that would give a proof based on the primary decomposition, but for five columns it is clear. The primary decomposition has four components, the toric part with 16 generators corresponding to the hppi's, and three others. Recall the basic result of Diaconis, Eisenbud, and Sturmfels: if the binomial $\mathbf{x}^{\mathbf{n}_1}\mathbf{y}^{\mathbf{n}_2} - \mathbf{x}^{\mathbf{m}_1}\mathbf{y}^{\mathbf{m}_2}$ corresponding to the two tables $\begin{array}{|c|}\hline \mathbf{n}_1 \\ \hline \mathbf{n}_2 \\ \hline \end{array}$, $\begin{array}{|c|}\hline \mathbf{m}_1 \\ \hline \mathbf{m}_2 \\ \hline \end{array}$ belongs to the ideal $I_M$ generated by the moves $M$, then the moves in $M$ connect the two tables. If the column sum $c_i$ is positive for all $i$, then either $x_i$ or $y_i$ is present in each monomial, which implies membership in each of the primary components, and hence membership of the binomial in the ideal $I_M$.

```
i;
i[1]=x(1)*x(4)*y(2)*y(3)-x(2)*x(3)*y(1)*y(4)
i[2]=x(1)*x(5)*y(2)*y(4)-x(2)*x(4)*y(1)*y(5)
i[3]=x(2)*x(5)*y(3)*y(4)-x(3)*x(4)*y(2)*y(5)
i[4]=x(1)*x(3)*y(2)^2-x(2)^2*y(1)*y(3)
i[5]=x(2)*x(4)*y(3)^2-x(3)^2*y(2)*y(4)
i[6]=x(3)*x(5)*y(4)^2-x(4)^2*y(3)*y(5)
> primarydecomposition;
[1]:
   _[1]=x(3)*x(5)*y(4)^2-x(4)^2*y(3)*y(5)
   _[2]=x(2)*x(5)*y(3)*y(4)-x(3)*x(4)*y(2)*y(5)
   _[3]=x(2)*x(5)^2*y(4)^3-x(4)^3*y(2)*y(5)^2
   _[4]=x(2)*x(4)*y(3)^2-x(3)^2*y(2)*y(4)
   _[5]=x(2)^2*x(5)*y(3)^3-x(3)^3*y(2)^2*y(5)
   _[6]=x(1)*x(5)*y(3)^2-x(3)^2*y(1)*y(5)
   _[7]=x(1)*x(5)*y(2)*y(4)-x(2)*x(4)*y(1)*y(5)
   _[8]=x(1)*x(5)^2*y(3)*y(4)^2-x(3)*x(4)^2*y(1)*y(5)^2
   _[9]=x(1)*x(5)^3*y(4)^4-x(4)^4*y(1)*y(5)^3
   _[10]=x(1)*x(4)*y(2)*y(3)-x(2)*x(3)*y(1)*y(4)
   _[11]=x(1)*x(4)^2*y(3)^3-x(3)^3*y(1)*y(4)^2
   _[12]=-x(2)^2*x(5)*y(1)*y(4)^2+x(1)*x(4)^2*y(2)^2*y(5)
   _[13]=x(1)*x(3)*y(2)^2-x(2)^2*y(1)*y(3)
   _[14]=x(1)^2*x(5)*y(2)^2*y(3)-x(2)^2*x(3)*y(1)^2*y(5)
   _[15]=x(1)^2*x(4)*y(2)^3-x(2)^3*y(1)^2*y(4)
   _[16]=x(1)^3*x(5)*y(2)^4-x(2)^4*y(1)^3*y(5)
[2]:
   _[1]=y(3)
   _[2]=x(3)
   _[3]=x(1)*x(5)*y(2)*y(4)-x(2)*x(4)*y(1)*y(5)
```

```
[3]:
     _[1]=y(4)
     _[2]=x(4)
     _[3]=x(1)*x(3)*y(2)^2-x(2)^2*y(1)*y(3)
[4]:
     _[1]=y(2)
     _[2]=x(3)*x(5)*y(4)^2-x(4)^2*y(3)*y(5)
     _[3]=x(2)
> quit;
```

Logistic regression is a special type of odds-ratio model, a challenging class of models for which the constraint matrices are the higher Lawrence liftings of Santos and Sturmfels (2002).

## 4. Sequential Importance Sampling

Consider the three $2 \times 3$ tables below with the same row and column sums. It is clear that the marginal distribution of the count in cell $(1,1)$ is not uniform when the tables are equally likely.

$$\begin{array}{|c|c|c|}\hline 0&0&1\\\hline 1&1&0\\\hline\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 0&1&0\\\hline 1&0&1\\\hline\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&0&0\\\hline 0&1&1\\\hline\end{array}\ .$$

One can generate tables randomly from a known distribution as follows. First, recall the well-known Fréchet bounds on entries: $\max\{0, r_i + c_j - n\} \le n_{ij} \le \min\{r_i, c_j\}$ where the row sums are $r_i$, the column sums are $c_j$ and $n$ is the total count over all cells. In particular $n_{11} \in \{0,1\}$. Choose uniformly from this interval to get a value for $n_{11}$, say 0, to get a partial table $\begin{array}{|c|c|c|}\hline 0&&\\\hline 1&&\\\hline\end{array}$ chosen with probability $\frac{1}{2}$. Now the bounds applied to the remaining part of the table for entry $(1,2)$ are $n_{2,1} \in [1+1-2, \min\{1,1\}]$, so 1 can be chosen with probability $\frac{1}{2}$ to get a further partial table $\begin{array}{|c|c|c|}\hline 0&1&\\\hline 1&0&\\\hline\end{array}$. The final column is then determined. The procedure generates the three tables above with probabilities $q_1 = \frac{1}{4}, q_2 = \frac{1}{4}, q_3 = \frac{1}{2}$ respectively, and sample averages can be reweighted with the reciprocals $1/q_i$ for approximate expectations with respect to the uniform distribution. A description of SIS can be found in Liu (2001) with many applications.

Consider now the problem of sampling from the $2 \times c$ tables for logistic regression introduced in §3. Recall the set of tables

$$S_0 := \{\mathbf{n} = (n_{11}, n_{12}, \ldots, n_{1c}, n_{21}, \ldots, n_{2,c}) : A\mathbf{n} = (t_1, t_2, \mathbf{c})'\}$$

where $A$ is the matrix of equation (1) of §3. SIS attempts to 1) choose $n_{11}$ for the first cell in a range of values, say an interval $[l_1, u_1]$, that allows ultimate completion to a table in $S_0$, then 2) choose $n_{1,2}$ from a new updated interval $[l_2(n_{11}), u_2(n_{11})]$ that allows ultimate completion, etc. It is useful if these intervals can be accurately computed during the sampling, and it is better if the range of values are intervals without gaps. That is, at each stage it is useful if the projection onto the next dimension of the polytope, after having fixed the first coordinate values, is an unbroken interval.

PROPOSITION 4.1. If the column sums $c_i$ are constrained to be positive, then every integer in a subinterval of the Fréchet bounds for $n_{11}$ can yield a valid table in $S_0$.

PROOF. With positive column sums, all tables in $S_0$ are connected with the moves in $M$, which correspond to a lex Gröbner basis with square-free exponent on first indeterminate. Since entries $n_{1,1}$ for cell $(1,1)$ are incremented by 1 in a connecting path between tables, the set of feasible values for cell $(1,1)$ is an interval of integers with no gaps. □

By comparison, consider the two tables $\begin{array}{|c|c|c|c|}\hline 0&3&0&0\\\hline 2&0&0&1\\\hline\end{array}$, $\begin{array}{|c|c|c|c|}\hline 2&0&0&1\\\hline 0&3&0&0\\\hline\end{array}$. There are no tables with the same constraint values and the value 1 in cell (1,1), but these two tables are not in the same connected component of the Markov chain built with the moves in $M$. The simplest example of this connectivity issue has $A := \begin{pmatrix}1 & 3\end{pmatrix}$. The set $S_0 = \{(n_1, n_2) : A\mathbf{n} = 3\}$ includes points $(0,1),(3,0)$ so the first coordinate is not in a complete interval. Here the Gröbner basis for $I_A$ is $\{x^3 - y\}$ and does not have the square-free property . When structural zeros are imposed on tables, the sequential-interval property may be lost, because the Markov chain for irreducibility may no longer have the square-free property, which occurs for example in the genotype data of §2.

For logistic regression there is a recurring structure that reduces the problem of determining sampling intervals $[l_i, u_i]$ to the case of $[l_1, u_1]$, the first column. Recall that we need to fix the row sum $T_1(\mathbf{n}) := n_{1,+} = t_1$, and the weighted row sum $T_2(\mathbf{n}) := \sum_i i \cdot n_{1,i} = t_2$, and the column sums $n_{+,i} = c_i$. The following diagram illustrates how the problem of interval values recurs with different constraints after the first column is filled:

$$\begin{array}{|c|c|c|c|c|c|}\hline [l_1,u_1] & n_{1,2} & n_{1,3} & n_{1,4} & \cdots & n_{1,c}\\\hline c_1-[l_1,u_1] & n_{2,2} & n_{2,3} & n_{2,4} & \cdots & n_{2,c}\\\hline\end{array} : \begin{array}{l} T_1(\mathbf{n}) = t_1\\ T_2(\mathbf{n}) = t_2\end{array}$$

$$\begin{array}{|c|}\hline x\\\hline c_1-x\\\hline\end{array}\ \begin{array}{|c|c|c|c|c|}\hline [l_2(x),u_2(x)] & n_{1,3} & n_{1,4} & \cdots & n_{1,c}\\\hline c_2-[l_2(x),u_2(x)] & n_{2,3} & n_{2,4} & \cdots & n_{2,c}\\\hline\end{array} : \begin{array}{l} T_1(\mathbf{n}_{(\cdot,2:c)}) = t_1 - x\\ T_2(\mathbf{n}_{(\cdot,2:c)}) = t_2 - t_1\end{array}$$

Computing the interval $[l_1, u_1]$ is made more efficient with the following feasibility test.

PROPOSITION 4.2. Assume $c_i \geq 1$. Let $n = \sum_{i=1}^{c} c_i$ be the table sum. For each $x \in [0, c_1]$ define tables $\mathbf{n}_L(x), \mathbf{n}_U(x)$ by

$$\mathbf{n}_L(x) := \begin{array}{|c|c|c|c|}\hline x & \min\{c_2, r_1 - x\} & \min\{c_3, r_1 - x - n_{12}\} & \cdots\\\hline c_1 - x & c_2 - n_{1,2} & c_3 - n_{1,3} & \cdots\\\hline\end{array}$$

$$\mathbf{n}_U(x) := \begin{array}{|c|c|c|c|}\hline x & \max\{c_2 + r_1 - x - n, 0\} & \max\{c_3 + r_1 - x - n_{12} - n, 0\} & \cdots\\\hline c_1 - x & c_2 - n_{1,2} & c_3 - n_{1,3} & \cdots\\\hline\end{array}$$

Then a nonnegative integer $x \in [l_1, u_1]$ if and only if $T_2(\mathbf{n}_L(x)) \leq t_2 \leq T_2(\mathbf{n}_U(x))$.

PROOF. The tables $\mathbf{n}_L(x), \mathbf{n}_U(x)$ minimize and maximize the value of $T_2$ over the collection of $2 \times c$ tables with top row sum $t_1$ and column sums $c_i$. $\mathbf{n}_L(x)$ puts the largest values possible to the left in the upper row, consistent with the Fréchet bounds for rectangular tables. Similarly, $\mathbf{n}_U(x)$ puts the smallest values possible to the left, and hence the largest on the right, consistent with the Fréchet bounds. Then it is clear that both inequalities must hold if a value $x$ is consistent with

$T_2 = t_2$. Conversely, if both inequalities hold, then there are two tables $\mathbf{m}$ and $\mathbf{n}$ with the right row and column sums, but $T_2(\mathbf{m}) \le t_2, T_2(\mathbf{n}) \ge t_2$. With column sums positive the table $\mathbf{m}$ can be connected to $\mathbf{n}$ with a sequence of adjacent minor moves (Sturmfels (2002), p. 64) $\begin{matrix} \cdots & - & + & \cdots \\ \cdots & + & - & \cdots \end{matrix}$, each of which changes the value of $T_2$ by $\pm 1$. Therefore the intermediate value theorem proves that some table will have the right value $T_2 = t_2$. □

EXAMPLE 4.1 (Cancer data from Sugiura and Otake (1974)). Binary data on death from leukemia is classified by dose at 6-levels.

| 5 | 4 | 6 | 1 | 3 | 6 |
|---|---|---|---|---|---|
| 5973 | 11811 | 2620 | 771 | 792 | 820 |

It took only a few seconds to obtain $10^6$ samples with SIS, which gave an estimate of 3053 with standard error 27.8 for the total number of tables with the same constraints. Based on these samples, the exact $p$-value for goodness-of-fit can be estimated at 0.0875 with standard error 0.005.

When some column sums are zero, SIS can still work without the sequential-interval property. The results for the 50 column cancer data of Hosmer and Lemeshow agreed closely with the Markov chain analysis.

For decomposable graphical models, it is known that there is a Markov basis with square free initial terms (Dobra and Sullivant (2003)) , and there are sharp bounds analogous to the Fréchet bounds (Dobra and Fienberg (2000)) so some of the elements for SIS are in place and a general theory may be possible. The use of algebra in SIS is relatively undeveloped compared to its use in designing Markov chains.

**Acknowledgements.** We thank Chong Tu for computing the $p$-value for Example 4.1 with sequential importance sampling, Seth Sullivant and Raymond Hemmecke for comments and discussions on this paper, and referees for valuable comments and suggestions.

## References

[1] Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

[2] Aoki, S. (2001). Network algorithm for the exact test of Hardy-Weinberg proportion for multiple alleles. METR Technical Report 01-06. http://www.stat.t.u-tokyo.ac.jp/~aoki/research.html.

[3] Bigatti, A. M., La Scala, R., and Robbiano, L. (1999). Computing toric ideals. *Journal of Symbolic Computation*, **27**, 351-365.

[4] Capani, A., Niesi, G., and Robbiano, L. (2002). *Cocoa: a system for doing computations in commutative algebra.* ftp://cocoa.dima.unige.it.

[5] Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2003). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, in press.

[6] De Loera, J. A., Haws, D., Hemmecke, R., Huggins, P., Sturmfels, B., and Yoshida, R. (2003). Short rational functions for toric algebra and applications. Manuscript.

[7] Diaconis, P., Eisenbud, D., and Sturmfels, B. (1998). *Lattice walks and primary decomposition.* In *Mathematical Essays in Honor of Gian-Carlo Rota*, eds. B. Sagan and R. Stanley. Birkhauser, Boston, pp. 173-193.

[8] Diaconis, P., Graham, R. L., and Sturmfels, B. (1996). Primitive partition identities. *Combinatorics: Paul Erdős is Eighty, Vol. 2*, 173-192.

[9] Diaconis, P., and Saloff-Coste, L. (1998). Nash inequalities for finite Markov chains. *Journal of Theoretical Probability*, **9**, 459-510.

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

[11] Dobra, A., and Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the NAS*, **97**, 11885-11892.

[12] Dobra, A., and Sullivant, S. (2003). A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Computational Statistics*, to appear August 2004.

[13] Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P Doyle, J. Lane, J. Theeuwes, L. Zayatz. Elsevier, New York, pp. 135-166.

[14] Greuel, G.-M., Pfister, G., and Schönemann, H. (2001). SINGULAR 2.0. A Computer Algebra System for Polynomial Computations. Centre for Computer Algebra, University of Kaiserslautern. http://www.singular.uni-kl.de.

[15] Guo, S. W., and Thompson, E. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361-372.

[16] Hemmecke, R., and Hemmecke, R. (2003). *4ti2 Version 1.1 – Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more*. http://www.4ti2.de.

[17] Hosmer, D. W., and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[18] Hosten, S., and Shapiro, J. (2000). Primary decomposition of lattice basis ideals. *Journal of Symbolic Computation*, **29**, 625-639.

[19] Huber, M., Chen, Y., Dobra, A., Dinwoodie, I. H., and Nicholas, M. (2003). Monte Carlo algorithms for Hardy-Weinberg proportions. Discussion Paper 03-09, Institute of Statistics and Decision Sciences, Duke University.

[20] Lazzeroni, L. C., and Lange, K. (1997). Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Annals of Statistics*, **25**, 138-168.

[21] Le Coutre, P., Demina, A., Beutler, E., Beck, M., and Petrides, P. E. (1997). Molecular analysis of Gaucher disease: distribution of eight mutations and the complete gene deletion in 27 patients from Germany. *Human Genetics*, **99**, 816-821.

[22] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

[23] Mehta, C., Patel, N. R., and Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic legression. *Journal of the American Statistical Association*, **95**, 99-108.

[24] Pistone, G., Riccomagno, and Wynn, H. P. (2000). *Algebraic Statistics: computational commutative algebra in statistics*. Chapman and Hall, New York.

[25] Santos, F., and Sturmfels, B. (2002). Higher Lawrence liftings. Manuscript.

[26] Sturmfels, B. (1996). *Gröbner Bases and Convex Polytopes*. AMS, Providence Rhode Island.

[27] Sturmfels, B. (2002). *Solving Systems of Polynomial Equations*. AMS, Providence Rhode Island.

[28] Sugiura, N., and Otake, M. (1974). An extension of the Mantel-Haenszel procedure to $K$ $2 \times c$ contingency tables and the relation to the logit model. *Communications in Statistics*, **A 3**, 829-842.

[29] Takemura, A., and Aoki, S. (2002). Some characterizations of minimal Markov basis for sampling from discrete conditional distributions. METR Technical Report 02-04, http://www.stat.t.u-tokyo.ac.jp/~aoki/research.html.

ISDS, 214 OLD CHEMISTRY BUILDING, BOX 90251, DUKE UNIVERSITY, DURHAM, NC 27708