Sampling Contingency Tables Using Markov Moves Based on Linear Programming

Cox, Lawrence
National Institute of Statistical Sciences
12177 Etchison Road
Ellicott City, MD 21042 USA
cox@niss.org

Introduction

Numerous problems in mathematics and statistics are concerned with contingency tables and other arrangements of integer values, represented here as feasible solutions to a system $\mathbf{Ax} = \mathbf{b}$: $\mathbf{A}$, $\mathbf{b}$, $\mathbf{x}$ integer and $\mathbf{x} \geq \mathbf{0}$. An important statistical application is an iterative MCMC algorithm of Diaconis-Sturmfels for sampling from discrete distributions subject to certain marginal totals ($\mathbf{A}$) held fixed ($\mathbf{b}$): beginning at a current solution $\mathbf{s}$, select an integer move $\mathbf{m}$ (integer solution of $\mathbf{Ay} = \mathbf{0}$, $\mathbf{y}$ unrestricted in sign) subject to appropriate conditions on the Markov chain; then, based on a Metropolis step, either the new solution $\mathbf{s} + \mathbf{m}$ or the current solution $\mathbf{s}$ enters the sample. This method encounters difficulties when the set of candidate moves is large or not fully known, often the case. Here we present an iterative method based on constructing, as opposed to selecting, a proposed next solution to enter the sample. First, for a class of frequency tables known as *tables of network type* (Cox 2007), the selection step is replaced by a computationally efficient construction step based on mathematical networks. We then extend the construction step to Markov moves in any integer linear program—hence any set of contingency tables subject to fixed marginals. This step is based on a more general linear programming (LP) construction (Cox 2011). This is a theoretical advance whose computational performance remains untested. We summarize these developments and present informative examples.

Sampling from discrete distributions

Seminal work of Diaconis and Sturmfels (Diaconis and Sturmfels 1998) blending techniques from mathematical statistics and algebraic geometry provided a theoretical framework for drawing a probability sample of contingency tables from the set of all contingency tables satisfying specified marginal totals. The marginal totals define the minimal sufficient statistics for an associated log-linear model. The set of all conforming contingency tables is represented by a linear generating set for all *moves* between tables (solutions) satisfying the specified marginals—all integer solutions $\mathbf{m}$ (unrestricted in sign) of $\mathbf{An} = \mathbf{0}$—a *Markov basis*. A Markov basis is characterized as follows: each move in the set of all feasible moves between solutions (tables) can be represented as a feasible sequence (sum) of nonnegative integer multiples of basis elements. Henceforth, it will be convenient to express moves $\mathbf{n} = \mathbf{n}^{+} - \mathbf{n}^{-}$ with $\mathbf{n}^{+}, \mathbf{n}^{-} \geq \mathbf{0}$ and $\mathbf{A}(\mathbf{n}^{+} - \mathbf{n}^{-}) = \mathbf{0}$.

Diaconis-Sturmels impose a uniform random distribution on the set of all moves and randomly select a move $\mathbf{m} = \mathbf{m}^+ - \mathbf{m}^-$. If **m** is feasible from the current solution (table) **s**, viz., if $\mathbf{s} + \mathbf{m} \geq \mathbf{0}$, then a Metropolis step (based on a hypergeometric distribution) is invoked to determine whether $\mathbf{s} + \mathbf{m}$ or **s** enters the sample.

There are two difficulties with this elegant procedure. First, the required Markov basis may be unavailable or not computable. Despite strides in the computation of Markov bases—in no small measure due to the Diaconis-Sturmfels work—it is often the case that the basis is unavailable or not computable, particularly for larger tables, and in these cases the difficulty is fatal. The second difficulty is not fatal, but does raise practical concerns—the method may produce infeasible tables, that can lead to burdensome computational waste that can threaten convergence.

Tables of network type

Mathematical networks are a special form of linear program that guarantee integer optimal solutions, thus enabling the solution of certain integer optimization problems at the cost/effort of linear optimization. Mathematically, this is possible because the coefficient (design) matrix **A** of a net work optimization is *totally unimodular*, viz., the determinant of every square sub-matrix of **A** equals -1, 0 or +1. *Tables of network type* by definition correspond to contingency table problems $\mathbf{Ax} = \mathbf{b}$ that are network. Network problems subject to integer capacity constraints: $\mathbf{0} \leq \mathbf{x} \leq \mathbf{c}$, $\mathbf{c}$ = a vector of nonnegative integers, are also network. We are concerned with moves, viz., integer solutions $(\mathbf{n}^+, \mathbf{n}^-) = (\mathbf{y}^+, \mathbf{y}^-)$ of $\mathbf{A}(\mathbf{y}^+ - \mathbf{y}^-) = \mathbf{0}$ on the unit hypercube, viz., $\mathbf{0} \leq \mathbf{y}^+ , \mathbf{y}^- \leq \mathbf{1}$. These are the *square-free moves*, and correspond to extreme points of this network.

In lieu of sampling from a minimal Markov basis of moves (which may be unavailable), we construct *proposal moves* iteratively using linear programming. At each stage, we construct a random cost function of a specialized form and run the linear program to produce an optimal solution. Owing to the network structure, this solution corresponds to a unique square-free move--the proposal move. The proposal is subjected to the Metropolis step to determine whether the current or proposal solution enters the sample. As we see below, this method offers the additional advantage of never constructing an infeasible move. The validity of the method is supported by the following theorems. Proofs are suppressed here but may be found in analogous form in Cox (2007).

Theorem N.1: The set of square-free moves is a Markov basis for the set of all moves between tables satisfying $\mathbf{Ax} = \mathbf{b}$.

Network arc capacities and costs are assigned as follows at each stage (current solution = **s**):

- $c_i^+ = 1$ and $c_i^- = \min\{1, s_i\}$
- $M >> 0$
- Select r_i from $\{-M, 0, M\}$ with equal probabilities = 1/3
- Cost function is $d(\mathbf{y}^+ , \mathbf{y}^-) = \sum_i (r_i (y_i^+ - y_i^-) + (y_i^+ + y_i^-))$

Theorem N.2: For each square-free move $\mathbf{n} = \mathbf{n}^+ - \mathbf{n}^-$ there exists a cost function **r** with capacities **c** of this form that optimizes uniquely at $(\mathbf{n}^+, \mathbf{n}^-)$.

Thus, all square-free moves have nonzero probability of being constructed. By Theorem N.1, all solutions (tables) have nonzero probability of being sampled. In addition, all constructed moves are feasible. The Markov chain is *irreducible* and, by virtue of the zero move, is *aperiodic*.

Theorem N.3: The proposal density function thus defined is *symmetric*, viz., the probability of moving from a solution $s^{(1)}$ to a solution $s^{(2)}$ equals the probability of moving from $s^{(2)}$ to $s^{(1)}$.

Symmetry assures that the Markov transition probabilities achieve a uniform distribution. The Metropolis step assures that the sample is random with respect to a proper (hypergeometric) distribution over the set of feasible tables satisfying the specified marginals.

In terms of log-linear models, tables of network type include two-way tables; hierarchies of two-way tables along one, but not both, dimensions; thin ($p x q x 2^{k-2}$) tables subject to a *no k-factor effects* log-linear model; and, any log-linear model with two or fewer configurations of minimal sufficient statistics. A second category of tables exhibiting square-free Markov bases correspond to *decomposable graphical models* (Dobra 2003). In terms of log-linear models, these include complete independence models. Beyond two dimensions, the two classes are distinct.

Networks can be implemented using standard mathematical programming software, commercial or open-source. Networks compute in time quadratic to cubic in the number of arcs.

General tables and general integer linear programs

Not all tables are network or decomposable. For example, the coefficient matrix **A** of a 3x3x3 table subject to fixed 2-dimensional marginals (the no 3-factor effects model) contains submatrices with determinant = 2, corresponding to a move not expressible as a positive sum of square-free moves, viz.,

2	0	0	0	0	1	0	1	0
0	0	1	0	0	0	1	0	0
0	1	0	1	0	0	0	0	1

Example 1a: Table ***s*** *underlying a 3x3x3 no 3-factor effects model*

-2	1	1	1	0	-1	1	-1	0
1	0	-1	0	0	0	-1	0	1
1	-1	0	-1	0	1	0	1	-1

Example 1b: Minimal but non-square-free move ***m*** *from* ***s***

$$\begin{array}{ccc} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{array} \qquad \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{array} \qquad \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}$$

Example 1c: $\boldsymbol{t} = \boldsymbol{m} + \boldsymbol{s}$

Such tables cannot be represented by a network; and it is no longer sufficient to restrict attention to integer extreme points on the restriction of $\mathbf{Ay} = \mathbf{0}$ to the unit hypercube. Our approach is to define a suitable system of linear constraints and a procedure for constructing random cost functions, and demonstrate that optimal solutions to this linear program are related to moves in arbitrary integer linear systems—and, consequently, arbitrary tables and tabular systems.

Doing so brings us to define a new type of basis. Whereas a Markov basis contains enough moves to construct the move between any two solutions **s** and **t**, we focus on a basis for all moves from a fixed **s** to arbitrary **t**, and call this a *local Markov basis* (from **s**). The union of all local Markov bases is, of course, a Markov basis. The linear constraint system used to construct a minimal local Markov basis from **s** is as follows.

$$\mathbf{A(y^{+} - y^{-}) = 0}$$
$$\mathbf{0 \leq L \leq (y^{+}, y^{-}) \leq U}$$
$$\mathbf{s + (y^{+} - y^{-}) \geq 0}$$

Figure 1: Linear constraint system for constructing a minimal local Markov move from ***s***

L and **U** are random integer vectors of dimension twice dim (**y**). The linear system is optimized: $\mathbf{min} \sum_i \mathbf{c_i} \, (\mathbf{y_i^{+} - y_i^{-}})$ with costs c_i drawn from a uniform distribution over {-1, 0, +1}. For technical reasons, the last (feasibility) constraint is replaced by substituting zero for corresponding elements of **L**, **U**.

A number of technical lemmas and theorems—in Cox (2011) and not presented here—allow us to draw a one-to-one correspondence between extreme points of this constraint system and "potential" minimal Markov moves from **s**. The precise correspondence is between the first integer point on the ray from the origin through the extreme point. This integer point corresponds to a "move" from **s** to another integer solution of $\mathbf{Ax} = \mathbf{b}$**--**but not necessarily a nonnegative (feasible) solution—hence the term "potential move". Using the same Metropolis step as before, Cox (2011) shows that the Markov chain thus defined is irreducible, aperiodic and reversible, and consequently that the resulting sample is random with respect to a proper distribution over the set of contingency tables satisfying the marginal constraints. Indeed, the method is entirely general in that the constraint system can be any integer linear system—not simply tabular—and the construction can be performed in many ways—probabilistic and deterministic—for purposes beyond the construction of a random sample.

Illustrative examples

We conclude with examples from the literature to which we apply our LP method.

The first example relates to failure of the integer rounding property in tables (Cox 2004).

(k,l = 1,1		k,l = 2,1		k,l = 1,2		k,l = 2,2)	
0	2	2	0	2	0	0	0
2	0	0	0	0	0	0	2

Example 2a: 2x2x2x2 table ***s*****=*****s****(i,j,k,l) subject to 2-dimensional marginals*

2	-1	-2	1	-1	0	1	0
-1	0	1	0	0	1	0	-1

Example 2b: One of 5 minimal local moves ***m*** *from* ***s***

2	1	0	1	1	0	1	0
1	0	1	0	0	1	0	1

Example 2c: ***t*** *=* ***m*** *+* ***s***

The second example deals with tables subject to additional (subtotal) constraints (Hara et al. 2009).

2	0	0
0	**0**	2
0	2	0

Example 3a: 3x3 table ***s*** *with subtotal constraint* ***(bold)***

-1	1	0
1	1	-2
0	-2	2

Example 3b: One of 5 local moves ***m*** *respecting table and subtotal constraints*

1	1	0
1	1	0
0	0	2

Example 3c: ***t = m + s***

The final example is a 2x2x2 complete independence model. The minimal local move **m** we generate is not square-free, illustrating a difference between our method and the decomposable graphical method of Dobra (2003).

2	0	0	0
0	0	0	2

Example 4a: 2x2x2 complete independence model ***s***

-2	1	1	0
1	0	0	-1

Example 4b: One of 9 minimal local moves ***m***

$$\begin{matrix}0 & 1 & \quad & 1 & 0\\ 1 & 0 & \quad & 0 & 1\end{matrix}$$

Example 4c: ***t* = *m* + *s***

REFERENCES

Cox, LH (2004). Inference control problems in statistical data base query systems. In: **Research Directions in Data and Applications Security** (C Farkas and P Samarati, eds.) Boston: Kluwer, 1-13.

Cox, LH (2007). Contingency tables of network type: Models, Markov basis and applications. *Statistica Sinica* **17**, 1371-1393.

Cox, LH (2011). A methodology for constructing Markov moves based on linear programming. Submitted for publication.

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

Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli* **9**, 1093-1108.

Hara, H, A Takemura and R Yoshida (2009). Markov bases for subtable sum problems. *Journal of Pure and Applied Algebra* **213**, 1507-1521.