Statistica Sinica **21** (2011), 1591-1609
doi:http://dx.doi.org/10.5705/ss.2008.078

SAMPLING LARGE TABLES WITH CONSTRAINTS

Ian H. Dinwoodie and Yuguo Chen

Portland State University and University of Illinois at Urbana-Champaign

Abstract: We describe a new sequential sampling method for constrained multi-way tables, with foundations in linear programming and sequential normal sampling. The method builds on techniques from other sequential algorithms in a way that scales well and can handle more challenging data sets. We apply the new algorithm to data to demonstrate its efficiency.

Key words and phrases: Conditional inference, contingency table, exact test, hypergeometric distribution, importance sampling, multivariate normal distribution, sequential sampling.

1. Introduction

The problem we address is sampling multi-way tables of counts with linear margin constraints, which relates to multidimensional versions of Fisher's exact test of independence in two-way tables. Current application areas of genetics, medicine, social sciences, and census data are motivating continued interest in sampling large tables. Random sampling from a prescribed distribution π on a space $\mathcal{S}$ of constrained tables is often used for conditional inference, where the constraints come from sufficient statistics. By conditioning on sufficient statistics, one can compute measures of goodness-of-fit that do not require asymptotic approximations; this is valuable for large sparse tables. For conditional inference, often the original table comes from multinomial or Poisson sampling, resulting in random margin values and a conditional hypergeometric distribution. Sampling from tables with fixed margins can also be useful in situations where the original sampling was done in other ways, or where the data is a complete classification of a population (Lehmann (1986, Chap. 4.7)). Other uses of sampling are volume tests (Diaconis and Efron (1985); Chen, Lin and Sabatti (2006)) and approximate enumeration (Chen et al. (2005)).

Methods for sampling constrained tables have evolved steadily with applications. Perfect methods for rectangular tables under the hypergeometric distribution (Fisher-Yates distribution) are old and well-known. Importance sampling is described in Booth and Butler (1999) and is available for use within R (R Development Core Team (2006)) with the package `exactLoglinTest` (Caffo (2006)).

The network method and its extensions are used in StatXact (2004). Markov chain Monte Carlo (MCMC) methods are used in Besag and Clifford (1989), Guo and Thompson (1992), and the fundamental work of Diaconis and Sturmfels (1998). However, all these methods have limitations. For example, none of these methods can do large no-3-way interaction problems. MCMC on constrained tables generally requires computing the Markov basis, a set of moves that guarantee irreducibility of the Markov chain. This algebraic computation can be too time-consuming with certain types of constraints, and even three-way tables can be impossibly difficult (DeLoera and Onn (2006)) despite rapid progress on algorithms and software for computing the Markov basis (see notably 4ti2 (4ti2 team (2006)). Indeed for three examples in Section 7, we were not able to obtain the Markov basis.

The problem of sampling multi-way tables with constraints has received attention in recent years as tables have become more challenging, and algebraic and sequential methods have been developed (Chen, Dinwoodie and Sullivant (2006)). Despite significant advances in methodology, problems of growing size and difficulty continue to challenge existing methods. For example, random graph models for social networks like the p_1 model (Holland and Leinhardt (1981)) can involve hundreds of variables and give rise to very challenging sampling problems for conditional inference. In this paper, we describe the most promising method for problems of increasing size. The new method combines the sequential importance sampling (SIS) method of Chen, Dinwoodie and Sullivant (2006) with a sequentially updated normal proposal distribution similar in flavor, but different in details, from the one in Booth and Butler (1999). After describing the algorithm in Section 3, we show in Section 6 how the details of the algorithm can be unified by the notion of approximating a maximum entropy distribution sequentially.

The importance sampling methods of Booth and Butler (1999) and Chen, Dinwoodie and Sullivant (2006) for constrained tables have complementary strengths. The method of Booth and Butler (1999) uses a sequentially-updated normal proposal distribution, but in some cases it has difficulty producing valid tables. The method of Chen, Dinwoodie and Sullivant (2006) uses SIS with linear programming to generate valid tables quite reliably, but the method as originally described did not include a well-designed proposal distribution. Certain examples, such as the binary 8-way table in Example 4 of Section 7 (Table 9.3.1 of Whittaker (1990, p.280)), were not possible with either method – Booth and Butler produced no valid tables, and Chen, Dinwoodie and Sullivant generated tables that were not at all typical under the target distribution, so estimated expectations were unreliable. The new method can handle this. In terms of speed, the method of Booth and Butler is fast when it works, so the method in this paper is intended for tables where simpler methods fail.

Here are some reasons why the normal distribution is a good proposal for sampling the multivariate hypergeometric distribution on large sparse tables with some high cell counts and some low counts. First, a central limit theorem applies to the multinomial distribution on unconstrained tables. Then conditioning on the margins makes the multinomial distribution into the hypergeometric distribution, and makes the normal approximation from the central limit theorem into another normal multivariate approximation on constrained tables. Those cells with large counts have a distribution that is well-approximated by the normal law, and the moments of the normal proposal can be updated conditionally with simple operations of linear algebra. Now some cell counts have distributions that are poorly approximated by the normal density. But these are the cells with lower counts and hence less room for error. On tables where a central limit theorem applies, on balance the normal approximation with updated moments is a significant improvement over proposals that do not use updated moments.

The paper is organized as follows. In Section 2 we set up the problem. In Section 3 we describe the proposed algorithm, and further implementation details and the computation of standard errors are discussed in Sections 4 and 5. In Section 6, we explain the connection between the normal proposal and the approximation of a maximum entropy distribution. In Section 7, we apply the proposed SIS methods to several data sets that have not been previously analyzed. One of the examples is beyond the capabilities of other existing methods. Section 8 provides concluding remarks.

2. Problem Setting

Let $\mathbf{n}_0$ be the observed data in an ordered vector and A be a nonnegative integer constraint matrix that fixes sufficient statistics and other design constraints. Let $\mathbf{b} = A\mathbf{n}_0$ be the constraint vector. Then the set to be sampled for Monte Carlo computations of expectations is the space of constrained tables

$$\mathcal{S} := \{\mathbf{n} : \ A\mathbf{n} = \mathbf{b}, \quad \mathbf{n} \geq \mathbf{0}\}.$$

The ith column of A will be denoted $\mathbf{a}_i$. In all our examples the sum over all entries of $\mathbf{n}$ will be fixed at the same total count as $\mathbf{n}_0$, denoted s_0. The simplest example of A would be the matrix that fixes row and column sums in a 2×2 table, so

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

with the four cells numbered left to right across the first row, then left to right across the second row, and each row in the matrix computes one constraint. This

is the independence model on two factors, say factors S and C (for Sex at level female and male and Color Blind at levels no or yes, for concreteness), and is commonly specified as [S], [C]. This notation is a list of sets of margins that are the sufficient statistics for the model.

Denote the configurations of the cells in a table $\mathbf{n}$ as integer vectors $\mathbf{n} = (n_1, \ldots, n_c)$, where c is the number of cells or entries in the table, and is also the number of columns in the constraint matrix A. The target distribution π in our examples is the hypergeometric distribution on the constrained tables $\mathcal{S}$, defined by

$$\pi(\mathbf{n} = (n_1, n_2, \ldots, n_c)) = \frac{\kappa}{\prod_{j=1}^{c} n_j!},$$

where κ is the normalizing constant. This is the conditional distribution on tables $\mathbf{n} \in \mathcal{S}$, given $A\mathbf{n} = \mathbf{b}$, of the loglinear multinomial distribution P_θ (on the space of tables of cell counts with total sum s_0) given by

$$P_\theta(\mathbf{n}) = \binom{s_0}{n_1, n_2, \ldots, n_c} \frac{e^{\theta' A\mathbf{n}}}{z_\theta^{s_0}},$$

where $z_\theta = \sum_{i=1}^{c} \exp(\theta' \mathbf{a}_i)$ is the normalizing constant on cell probabilities, and the parameter θ is a real column vector with the same number of coordinates as rows of A. The distribution π also arises in other contexts, such as conditional Poisson regression. The number of free parameters in this exponential family is typically less than the length of θ. Let p_i be the probability of drawing cell i $(i = 1, \ldots, c)$ in the multinomial probability P_θ, so $p_i = \exp(\theta' \mathbf{a}_i)/z_\theta$, and let $\mathbf{p}$ be the vector $(p_1, \ldots, p_c) \in \mathbf{R}^c$. The probability vector $\hat{\mathbf{p}}$ in the algorithm below is an estimate of $\mathbf{p}$.

The importance sampling approach to estimate $\mu := E_\pi[f(\mathbf{n})] = \sum_{\mathbf{n} \in \mathcal{S}} f(\mathbf{n})$ $\pi(\mathbf{n})$ is to simulate tables from a different distribution $q(\cdot)$, where $q(\mathbf{n}) > 0$ for all $\mathbf{n} \in \mathcal{S}$, and estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(\mathbf{n}_i)[\pi(\mathbf{n}_i)/q(\mathbf{n}_i)]}{\sum_{i=1}^{N} [\pi(\mathbf{n}_i)/q(\mathbf{n}_i)]}, \tag{2.1}$$

where $\mathbf{n}_1, \ldots, \mathbf{n}_N$ are independent and identically distributed (i.i.d.) samples from $q(\mathbf{n})$, and $\pi(\mathbf{n}_i)/q(\mathbf{n}_i)$ is the importance weight. In goodness-of-fit tests that we discuss below, the function $f(\mathbf{n})$ takes the form $f(\mathbf{n}) = I_{\{\pi(\mathbf{n}) \le \pi(\mathbf{n}_0)\}}$ for p-value computations. Efficient estimation usually requires that $q(\mathbf{n})$ be close to $\pi(\mathbf{n})$ and this problem is addressed in the proposed algorithm.

3. Sequential Sampling with Linear Programming and Normal Proposal

To generate valid tables, the sequential sampling method samples a table cell by cell, in a way that guarantees that every table in $\mathcal{S}$ can be produced. More precisely, we start by sampling the first cell n_1 of the vector $\mathbf{n}$ conditional on the constraints imposed on the table. Conditional on the realization of the first cell, we sample the second cell n_2 in a similar manner and then move forward conditionally until all the cells are sampled. We can write $q(\cdot)$ as

$$q(\mathbf{n}) = q(n_1)q(n_2|n_1)q(n_3|n_2, n_1)\cdots q(n_c|n_{c-1}, \ldots, n_1).$$

In the SIS algorithm of Chen, Dinwoodie and Sullivant (2006), the lower and upper bounds for the support of each cell are computed by linear programming, and a simple hypergeometric distribution on the interval formed by the lower and upper bounds is used as the proposal distribution to sample each cell. Here we describe a refinement that starts with a normal approximation conditioned on the margin values as an initial proposal distribution. The method then updates the proposal distribution sequentially as cells are filled in and as intervals are computed with linear programming. The detailed procedure is given below.

Algorithm:

1. Number the c cells of the table to determine the order of sampling.
2. Specify the $r \times c$ constraint matrix A for the desired loglinear model.
3. Obtain the normal approximation to the distribution π:
   (a) estimate the $c \times 1$ vector $\mathbf{p}$ of multinomial cell probabilities using data $\mathbf{n}_0$, typically with maximum likelihood estimation, and call it $\hat{\mathbf{p}}$;
   (b) use the multinomial mean $\boldsymbol{\mu}^* = s_0\,\hat{\mathbf{p}}$ and multinomial covariance $\Sigma^* = s_0(\mathrm{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}')$ (here $\mathrm{diag}(\hat{\mathbf{p}})$ is the diagonal matrix with diagonal values $\hat{\mathbf{p}}$) to specify the joint normal distribution $N(\boldsymbol{\mu}^*, \Sigma^*)$ on the table entries $\mathbf{n}$ in $\mathbf{R}^c$, and obtain the joint normal distribution $N\left(\begin{pmatrix} A\boldsymbol{\mu}^* \\ \boldsymbol{\mu}^* \end{pmatrix}, \begin{pmatrix} A\Sigma^*A^T & A\Sigma^* \\ \Sigma^*A^T & \Sigma^* \end{pmatrix}\right)$ on margin values and table entries $\begin{pmatrix} A\mathbf{n} \\ \mathbf{n} \end{pmatrix}$ in $\mathbf{R}^c \times \mathbf{R}^r$;
   (c) based on the joint normal distribution of $\begin{pmatrix} A\mathbf{n} \\ \mathbf{n} \end{pmatrix}$, obtain the conditional normal distribution $N(\boldsymbol{\mu}, \Sigma)$ on table $\mathbf{n}$ in $\mathbf{R}^c$, given the margin values $A\mathbf{n} = \mathbf{b}$.

4. Sample the first cell (and subsequent cells by returning):
   (a) use linear programming to get a lower limit l and upper limit u on the value n_1 of the first cell, subject to the constraints on tables;
   (b) let μ_1 be the first component in the conditional mean vector $\boldsymbol{\mu}$, and let σ_1^2 be the first diagonal entry in the conditional covariance matrix Σ;
   (c) sample from the discrete distribution on the integers in $[l, u]$ proportional to $\exp\{-(j-\mu_1)^2/(2\sigma_1^2)\}$ (i.e., proportional to the density of $N(\mu_1, \sigma_1^2)$), $j = l, l+1, \ldots, u$, to get n_1, the count for the current cell, and record the sampling probability
$$q(n_1) = \frac{e^{-(n_1-\mu_1)^2/(2\sigma_1^2)}}{\sum_{j=l}^{u} e^{-(j-\mu_1)^2/(2\sigma_1^2)}};$$
   (d) if any cells are left, redefine the constraint matrix and constraint vector $A = A[,-1]$, $\mathbf{b} = \mathbf{b} - n_1 \cdot \mathbf{a}_1$, where the notation $A[,-1]$ denotes the matrix A with the first column removed, so it is the constraint matrix on the remaining cells when the initial value is fixed;
   (e) if any cells are left, update the conditional mean vector $\boldsymbol{\mu}$ and the conditional covariance matrix Σ for the remaining cells given n_1;
   (f) if any cells are left, set the next cell to be the first cell and repeat Step 4; otherwise we continue to Step 5 with a table $\mathbf{n} = (n_1, \ldots, n_c)$ as an integer vector.
5. Store the complete unnormalized weight $1/[q(\mathbf{n}) \prod_{j=1}^{c} n_j!]$ for the sampled table, where $q(\mathbf{n})$ is a product of the sampling probabilities from Step 4(c).
6. Repeat the above steps N times to generate N tables $\mathbf{n}_1, \ldots, \mathbf{n}_N$.
7. Normalize the weights over all N tables to have sample mean 1.0.
8. Estimate the p-value with $\hat{\mu}$ at (2.1), and compute the standard error (se) of the estimate and the coefficient of variation (cv) of the importance weights (see (5.1) for definition of cv).

4. Further Discussion of Proposed Algorithm

Here are some comments and clarifications on the outline of the algorithm above. Step 1 can often be done in any natural way on small problems. By natural, we mean proceed through each dimension (or factor) in sequence without jumping across dimensions. The ordering affects the performance of the algorithm in two ways. First, it affects the property of "sequential intervals"

formulated in Chen, Dinwoodie and Sullivant (2006) that plays a major role in generating valid tables. The property of sequential intervals holds when each integer in $[l, u]$ can lead to a valid table, a useful property for efficiency. For most models and most data, any natural order works pretty well in this respect. Second, within the natural orders, some may be much better than others from the point of view of approximating π sequentially with a normal proposal. This is illustrated in Example 4 of Section 7.

A useful rule of thumb is to reorder the cells in such a way that the constraint matrix does not change and the ordering remains natural, but at the same time try to make the cells with larger counts come first, or at least early. This can be done by permuting labels on factors and labels on levels within factors for models with lots of symmetry. The goal is to maintain sequential intervals, but at the same time to put cells first that have large counts so the normal proposal will be good while it matters the most. Then, as cells are filled in and the conditional proposal distribution gets worse as an approximation to the target conditional, the impact on sampling performance is less.

To illustrate more clearly the natural reorderings, consider the simple 2×2 example for testing independence of binary factors at the beginning of Section 2. Suppose the data are

$$\begin{array}{c} \\ \text{Yes} \\ \text{No} \end{array} \begin{array}{c} \begin{array}{cc} \text{Male} & \text{Female} \end{array} \\ \begin{pmatrix} 0 & 1 \\ 10 & 100 \end{pmatrix} \end{array} \text{ or, equivalently, } \begin{array}{c} \\ \text{No} \\ \text{Yes} \end{array} \begin{array}{c} \begin{array}{cc} \text{Female} & \text{Male} \end{array} \\ \begin{pmatrix} 100 & 10 \\ 1 & 0 \end{pmatrix} \end{array}.$$

Then the matrix A on the data vector $(0, 1, 10, 100)$ computes row and column sums and gives the vector $(1, 110, 10, 101)$ of sufficient statistics; the data could be naturally reordered as $(100, 10, 1, 0)$ by swapping the labels of the two binary factors, and the same constraint matrix A would compute a reordered vector of sufficient statistics $(110, 1, 101, 10)$. The second order would be better for sequential sampling:

In the model of no-3-way interaction with k factors all with $\{1, \ldots, \lambda\}$ levels, the symmetry leads to a simple way to generate natural orders. A $(k+1)$-tuple of permutations in the set $S_k \times S_\lambda \times \cdots \times S_\lambda$ with k copies of the permutations S_λ on characters $1, \ldots, \lambda$, gives a transformation on k-way tables $(\mathbf{n_x})$, $\mathbf{x} = (x_1, \ldots, x_k) \in \{1, \ldots, \lambda\}^k$, by

$$(\tau, \sigma_1, \ldots, \sigma_k)(\mathbf{n}_{(x_1,\ldots,x_k)}) = (\mathbf{n}_{(\sigma_1(x_{\tau(1)}),\ldots,\sigma_k(x_{\tau(k)}))},$$

and this rearrangement leaves invariant the constraint matrix A for computing sufficient statistics. Then there are $k!\lambda!^k$ natural reorderings of this type for consideration, generally a small fraction of all $(\lambda^k)!$ vector reorderings many of which destroy sequential intervals and result in poor sampling performance.

In the 8-way binary data of Example 4, there are $8! \cdot 2^8$ natural orders some of which improve sampling performance significantly. For reordering in practice, the R command `aperm` applies the transformation τ (R Development Core Team (2006)). For models with less symmetry, identifying the natural reorderings may be more difficult.

To obtain the constraint matrix A in Step 2, one can use the `genmodel` command in 4ti2 (4ti2 team (2006)) for many standard loglinear models. The voting model of Example 3 is not possible with this tool however, it must be coded manually.

For the estimation Step 3(a), one can use many existing tools in order to avoid coding the procedure for maximum likelihood estimation. For example, one may normalize the fitted values from regression. This fitting procedure is easiest in R with the `glm` command when the data is in the form of a data frame and the model is standard, and it can be done with `loglin` on a multi-way array using iterative proportional fitting for loglinear models specified with a list of margin sets for sufficient statistics.

In Step 3(c), the mean and variance of the conditional normal distribution can be computed in the standard way. If we use the notation $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ for a pair of normal random vectors, which have joint mean $E \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ and joint covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11}\, \Sigma_{12} \\ \Sigma_{21}\, \Sigma_{22} \end{pmatrix}$, then the conditional mean and covariance of X_2 given $X_1 = \mathbf{x}_1$ are $E(X_2 \mid X_1 = \mathbf{x}_1) = \boldsymbol{\mu}_2 + \Sigma_{12}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and $\mathrm{Cov}(X_2 \mid X_1 = \mathbf{x}_1) = \Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}$ (see for example Whittaker (1990, p.163)). In Step 3(c), the joint normal distribution of $\begin{pmatrix} A\mathbf{n} \\ \mathbf{n} \end{pmatrix}$ has a singular covariance matrix, so one must be careful about the formula for the conditional distribution (Marsaglia (1964)).

Step 4 involves the most intensive numerical work. The linear programming in Step 4(a) finds the maximum u and the minimum l over rational numbers of the objective function $h(\mathbf{x}) = x_1$ subject to $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq 0$. The software `lpsolve` of Berkelaar, Eikland and Notebaert (2004) is used within R (although there is some evidence at this time that `Rsymphony` of Harter, Hornik and Theussl (2008) is faster). The difference between rational and integer programming is a theoretical issue, but in practice rarely becomes a problem with proper rounding (Chen, Dinwoodie and Sullivant (2006)).

In Step 4(b) we have the $c \times 1$ mean $\boldsymbol{\mu}$ and the $c \times c$ covariance matrix Σ conditional on the margins from Step 3(c), and one simply takes the first entries. For the cell sampling in Step 4(c), some further care has to be taken when

the conditional mean μ_1 is outside the interval $[l, u]$, or when the conditional variance σ_1^2 is extremely small. This is rare but may happen when the normal approximation is not good for parts of the table, leading to the situation where the sequentially updated moments of the normal distribution become incompatible with the sequentially updated state space. A simple way to deal with these problems is to move the mean μ_1 to the center of $[l, u]$ and round up σ_1 to $1/2$ when necessary. This *ad hoc* adjustment occurs very rarely and only for cells where the range of values is small, so it has a very small effect on performance; the method is "correct" as long as the adjustment is accounted for in the weights. The updating Step 4(e) is straightforward using the formula for conditional means and variances in Step 3(c). Since the current cell value is one-dimensional, no matrix inversion is needed in the computation. Computing the weights in Step 5 should be done with logarithms and careful numerical methods for large tables.

5. Standard Error Computation and Confidence Interval

We now consider the importance weights that are proportional to ratios of unnormalized target probabilities to proposed probabilities. Step 7 says to normalize the weights and, although the expression (2.1) for computation does not require this, normalizing is good because then the estimate of cv^2 in (5.1) can be computed as the sample variance of the normalized weights, and the normalized weights can be used for diagnostic methods that are awkward when the weights are extremely small, possibly on the order of 10^{-10} in some cases. Define the weight W_i by

$$W_i := \frac{1}{q(\mathbf{n}_i)\prod_{j=1}^{c} n_j!} = \frac{\pi(\mathbf{n}_i)}{\kappa\, q(\mathbf{n}_i)},\ i = 1, \ldots, N,$$

and then $W_1, W_2, \ldots, W_N$ are i.i.d. random variables. To simplify notation in some formulas, let W have the same distribution as any W_i.

Let $\bar{W} = \sum_{i=1}^{N} W_i/N$. The ratio $W_i/\bar{W}$ is the normalized weight, whose expectation $\mathrm{E}_q(W_i/\bar{W})$ converges to 1 as the sample size N increases. It follows that $\kappa = 1/\mathrm{E}_q(W)$ and

$$\frac{\mathrm{E}_q(W^2)}{[\mathrm{E}_q(W)]^2} = \mathrm{Var}_q\left(\frac{W}{\mathrm{E}_q(W)}\right) + 1.$$

Now define the coefficient of variation (cv), a useful quantity for measuring the variation of weights in importance sampling, by

$$\mathrm{cv}^2 := \frac{\mathrm{E}_q(W^2)}{[\mathrm{E}_q(W)]^2} - 1 \approx \frac{\sum_{i=1}^{N}(W_i - \bar{W})^2}{(N-1)\bar{W}^2} = \frac{\sum_{i=1}^{N}(W_i/\bar{W} - 1)^2}{(N-1)}. \tag{5.1}$$

The quantity cv^2 figures into the definition of "effective sample size (ESS)" (Liu (2001, p.36)) which is used for comparing efficiency of importance sampling with naive Monte Carlo (i.e., direct independent sampling from the target distribution):

$$\mathrm{ESS}(N) = \frac{N}{1+\mathrm{cv}^2}. \tag{5.2}$$

Roughly speaking, we need $\mathrm{ESS}(N)$ i.i.d. samples from the target distribution in order to obtain the same standard error for $\hat{\mu}$ as N importance samples.

The estimate (2.1) for a conditional p-value has $f = I_B$, the indicator function for the set $B = \{\mathbf{n} \in \mathcal{S} : \pi(\mathbf{n}) \leq \pi(\mathbf{n}_0)\}$, and

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} I_B(\mathbf{n}_i) \cdot \frac{W_i}{\bar{W}}. \tag{5.3}$$

The delta method for approximating variances (Liu (2001, p.35)) leads to an approximation of the mean-squared error $\mathrm{MSE}(\hat{\mu})$ by

$$\frac{1}{N \cdot [\mathrm{E}_q(W_i)]^2}\left[\mu^2 \mathrm{Var}_q(W_i) - 2\mu \mathrm{Cov}_q(W_i, I_B(\mathbf{n}_i)\cdot W_i) + \mathrm{Var}_q(I_B(\mathbf{n}_i)\cdot W_i)\right]. \tag{5.4}$$

An approximation to the standard error of the estimate $\hat{\mu}$ is $\sqrt{\mathrm{MSE}(\hat{\mu})}$ with the mean, variance, and covariance in (5.4) replaced by their sample counterparts. This empirical approximation of a delta method approximation is reported in the examples of Section 7. Another method, more direct and reliable but also more time consuming, is to do repeated runs to produce a random sample of estimates of μ, and to use their mean and standard error.

While the estimate of the standard error above is traditional and valuable, consider confidence intervals in more detail for the estimate $\hat{\mu}$ of the probability μ. The traditional interval $\hat{\mu} \pm 2 \cdot \mathrm{se}$ (the Wald interval) is sometimes considered unreliable as a confidence interval for μ when μ is small (Agresti and Coull (1998); Brown, Cai and DasGupta (2001)); the score interval is sometimes recommended instead. The score interval for a binomial proportion μ with n trials and 95% confidence is

$$\frac{\hat{\mu} + z^2_{0.025}/2n \pm z_{0.025}\sqrt{\hat{\mu}(1-\hat{\mu})/n + z^2_{0.025}/4n^2}}{1 + z^2_{0.025}/n}, \tag{5.5}$$

where $z_{0.025} = 1.96$ (Brown, Cai and DasGupta (2001)). We argue below that one may expect the score interval to be a reasonable confidence interval for probabilities μ under SIS with sample size N replaced by $n = \mathrm{ESS}(N) = N/(1+\mathrm{cv}^2)$.

The score interval is derived by starting with the approximation

$$P(-z_{0.025} < \frac{\bar{\mu} - \mu}{\sqrt{\mu(1-\mu)/n}} < z_{0.025}) \approx 0.95,$$

where $\bar{\mu}$ is the proportion of successes in n independent Bernoulli trials with success probability μ. Then one solves a quadratic inequality in μ, treating $\bar{\mu}$ as fixed, to obtain the interval. Using $\mathrm{Var}_\pi[I_B(\mathbf{n})] = \mu(1-\mu)$ (since $\mu = \pi(B)$) and the equation on relative efficiency on p. 36 of Liu (2001), we have

$$\frac{\mu(1-\mu)}{\mathrm{Var}_q\left(I_B \cdot W(\mathbf{n})/\mathrm{E}_q(W)\right)} \approx \frac{1}{1+\mathrm{cv}^2},$$

suggesting that $\mathrm{Var}_q(\hat{\mu}) \approx [\mu(1-\mu)/N]\cdot(1+\mathrm{cv}^2)$. This leads to the approximation

$$P(-z_{0.025} < \frac{\hat{\mu} - \mu}{\sqrt{\mu(1-\mu)/n}} < z_{0.025}) \approx 0.95$$

with $n = N/(1+\mathrm{cv}^2)$. Then the original derivation of the score interval can be applied with the effective sample size.

6. Connection Between Normal Proposal and Maximum Entropy

Our sampling algorithm uses elements of normal sampling theory and linear programming imposed on a discrete state space in a way that may seem contrived; the method can be seen in a more unified way as a sequential way to approximate a maximum entropy distribution. Recall that the entropy of a distribution with density p on a set $\mathcal{S}$ is given by $H(X) = -\sum_{\mathcal{S}} p(x)\log(p(x))$, a nonnegative quantity. Here X represents a random element with distribution p; we are using the notation of Cover and Thomas (1991). The multinomial distribution P_θ on tables can be approximated by a normal distribution using the Central Limit Theorem. The conditional distribution $N(\boldsymbol{\mu}, \Sigma)$ on table cells given margin values $A\mathbf{n} = \mathbf{b}$ (from Step 3(c) of the algorithm), a good approximation to π, has maximum entropy over all densities on $\mathbf{R}^c$ with the same $\boldsymbol{\mu}$ and Σ, being Gaussian (Cover and Thomas (1991, p.270)). Therefore one should use a proposal distribution q on $\mathcal{S}$ that also has maximum entropy, and with the same first and second moments as the maximum entropy approximation $N(\boldsymbol{\mu}, \Sigma)$. Here we are being slightly unclear about the reference measure for the entropy computation on $\mathcal{S}$. If we use the notation $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ for the first cell value $X_1 \in \mathbf{R}^1$ and $X_2 \in \mathbf{R}^{c-1}$, then the proposal distribution should nearly maximize over q the joint entropy written sequentially:

$$H(X_1, X_2) = -\sum_{\mathcal{S}} q(x_1)\log(q(x_1)) - \sum_{\mathcal{S}} q(x_1)q(x_2 \mid x_1)\log(q(x_2 \mid x_1)).$$

Now the margin $q(x_1)$ appears as a weighting factor in the second term, but if we ignore its presence in the second term we get the approximate maximum entropy solution

$$q(x_1) = \text{argmax}_Q\{-\sum_{\mathcal{S}} Q(x_1)\log(Q(x_1)) : \text{E}_Q(X_1) = \mu_1, \text{Var}_Q(X_1) = \sigma_1^2\}.$$

Now by the well-known exponential form of the maximum entropy density under constraints, it follows that the one-dimensional margin $q(x_1)$ has the form $q(x_1) \propto \exp(\alpha x_1 + \beta x_1^2)$ on the support $[l, u]$ of the first coordinate from $\mathcal{S}$ (Cover and Thomas (1991, p.267)) – that is, we get the representation in Step 4(c) of the algorithm, where α and β are chosen so the mean is approximately μ_1 and the variance is approximately σ_1^2 based on heuristics, rather than computation, for speed. Now once the first margin is found, the computation is repeated again in the same way, with a new joint mean and joint covariance on the remaining cells providing moment constraints for maximum entropy:

$$\begin{aligned} \text{E}(X_2 \mid X_1 = x_1) &= \boldsymbol{\mu}_2 + \Sigma_{12}\Sigma_{11}^{-1}(x_1 - \mu_1), \\ \text{Cov}(X_2 \mid X_1 = x_1) &= \Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}. \end{aligned}$$

Therefore the sampling algorithm can be viewed as a sequence of maximum entropy approximations on the projected first coordinate of a polytope $\mathcal{S} \in \mathbf{R}^c$, under sequential moment constraints that come from a normal, maximum entropy approximation.

7. Examples

In examples we compare different methods of computing p-values for goodness-of-fit. The p-value is defined by $\pi\{\mathbf{n} : \pi(\mathbf{n}) \leq \pi(\mathbf{n}_0)\}$, where $\mathbf{n}_0$ is still the observed data. All examples were coded in R (R Development Core Team (2006)) and run on a laptop with a 1.66 GHz Intel Pentium processor and 1 GB of memory. The computations were done so that the actual running times were similar for different methods. The estimate of the standard error is based on the square root of (5.4), except in Example 4.

Example 1. Deresiewicz et al. (1997, Table 2) present data comparing computed tomographic (CT) scans and magnetic resonance imaging (MRI) scans on 32 patients with equine encephalitis. Our concern is whether the two methods are equally sensitive for detection. The null hypothesis is that they are equally sensitive, with a view to supporting the claim that one is more sensitive than the other under rejection (indeed, it is claimed that MRI is more sensitive). The data in Table 1 are counts of abnormal findings which make up a $2 \times 2 \times 8$ sparse three-way table.

Table 1. Deresiewicz et al.'s (1997) data of neuroradiographic studies on 32 patients with equine encephalitis.

Anatomical Site or Abnormality	CT Scan (32 patients)	MRI Scan (14 patients)
Basal gangli	18	10
Thalamus	8	10
Brain stem	3	6
Cortex	4	5
Periventricular area	0	2
Meninges	2	0
Hydrocephalus	1	0
Any abnormality	21	13

Two Poisson regression models can be fit and evaluated with standard generalized linear model asymptotics. First, a two-factor model, with factor Scan type at levels CT and MRI and factor Site at eight levels, has a response which is the abnormal count at each combination of factors. The significance of the parameter on Scan type is what we want, and its p-value is reported to be 1.1×10^{-5} in R with a reasonable fit judging by residual deviance. Second, if one fits a smaller model by leaving out the Scan type factor, the model does not fit well with residual deviance giving p-value of approximately 1.0×10^{-4}. Both methods confirm a very significant difference in sensitivity between the CT and MRI scans.

For conditional inference, the $2 \times 2 \times 8$ table has factors Status (at two levels normal and abnormal), Scan Type (at two levels CT and MRI) and Site (at eight levels). A model that leaves out Scan Type in a formula for the probability of abnormal Status has sufficient statistics [Status, Site], and 16 fixed combinations of [Scan Type, Site] from the design. Thus the constraint matrix has 32 rows of constraints and 32 columns for the cells.

Table 2 has the analysis. We compared the proposed SIS with the normal proposal with the SIS with the hypergeometric proposal in Chen, Dinwoodie and Sullivant (2006). Both methods were run for about 40 minutes. The p-value is quite small, so the computation is delicate. SIS delivered 100% valid tables with both proposals.

The cv^2 value for SIS was 0.11 with the normal proposal, but 12.62 with the hypergeometric proposal, which indicates that the normal proposal is much closer to the target distribution than the hypergeometric proposal. Although both methods took about the same amount of time to generate $25,000$ importance samples, the effective sample size for the normal proposal is about 12 times larger than that for the hypergeometric proposal. These results show that the normal proposal is much more efficient than the hypergeometric proposal in this example.

Table 2. Comparison of SIS with different proposal distributions on the data in Table 1.

	SIS normal	SIS hypergeometic
Estimate	3.4×10^{-5}	3.8×10^{-6}
Standard error	2.4×10^{-5}	1.6×10^{-7}
Sample size	$25,000$	$25,000$
cv^2	0.11	12.62
ESS	$22,523$	220

Because of the discrepancy between the two estimates, we also implemented the MCMC algorithm (Diaconis and Sturmfels (1998)) for this problem. The Markov basis consists of 8 moves. A long simulation of 4,000,000 MCMC samples gave an estimate of order 10^{-5} which agreed with SIS with the normal proposal. The result for the hypergeometric proposal is biased low by a factor of 10, and the standard error approximation is unrealistically low, possibly because of the bias towards 0. This type of phenomenon is not unusual in importance sampling when a not well-designed proposal is used.

Example 2. Data on frequency of livestock breeds (Table 3) is presented in Hall and Ruane (1993), classified three ways by type of animal, presence (common, rare, extinct), and region. We removed the common level to focus on rare and extinct combinations. Our interest is how well the model of no-3-way interaction, sometimes called all-2-way interaction, fits the data. This is the model with sufficient statistics that are the sums of counts over the third factor at all fixed combinations of pairs of levels of two factors.

Using glm with Poisson regression gives a likelihood ratio statistic of 36.1 on 36 degrees of freedom. These numbers make the model fit look decent because the degrees of freedom includes even the contributions of cells that are in the domain of vanishing margins, and therefore fixed at 0 in the exact sampling.

Based on 1,000 samples that took about 9 minutes to generate, SIS with normal proposal estimated a p-value of 0.012 with standard error 0.005. SIS gave 100% valid tables, with a cv^2 value of 0.28. After manually removing the cells forced to be 0 by vanishing margins, `exactLoglinTest` can be used and reports a p-value for the Pearson χ^2 statistic of 0.013. This number is based on the hypergeometric distribution π for the Pearson χ^2 distance function, not the asymptotic $\chi^2(36)$ distribution, and is computed with parameter `maxiter` set equal to 10^6. For this problem, the Markov basis for MCMC was not found after 24 hours of running time.

Example 3. Voting data on five candidates has been analyzed in Diaconis (1989) and Eriksson and Diaconis (2005). The table presents counts for each of the 5! orderings of candidates. The first-order model, whose sufficient statistics

Table 3. Hall and Ruane's (1993) data on frequency of livestock breeds.

		Ass	Water Buffalo	Cattle	Goat	Horse	Pig	Sheep
Africa	Rare	0	0	10	0	2	0	4
	Extinct	0	0	22	0	2	0	1
Asia	Rare	0	2	8	4	14	2	1
	Extinct	0	0	5	1	3	8	2
Europe	Rare	10	0	101	29	49	37	109
	Extinct	5	0	154	19	58	79	98
North and Central America	Rare	0	0	8	4	9	5	7
	Extinct	0	0	1	1	4	17	10
South America	Rare	1	0	4	0	0	0	1
	Extinct	0	0	19	0	0	0	0
Oceania	Rare	0	0	1	0	1	1	2
	Extinct	0	0	2	0	1	1	5
ex-U.S.S.R.	Rare	0	0	9	4	23	2	11
	Extinct	0	0	21	6	20	21	32

are counts of frequencies with fixed value at each coordinate, does not fit at all well. Establishing this result is more than fitting an additive Poisson regression model on 5-way data, because the permutation data makes structural zeros in a 5-way classification.

Now we are concerned with a large model that may fit better. A model with more parameters that includes some second-order terms is given by

$$p_{(i_1,i_2,i_3,i_4,i_5)} \propto e^{\theta_{1,i_1}+\theta_{2,i_2}+\theta_{3,i_3}+\theta_{4,i_4}+\theta_{5,i_5}+\gamma_{1,(i_1,i_2)}+\gamma_{2,(i_2,i_3)}+\gamma_{3,(i_3,i_4)}+\gamma_{4,(i_4,i_5)}},$$

which makes the sufficient statistics all the first-order totals fixing each variable (5 levels) at each coordinate (5 coordinates), together with the second-order sums of counts at consecutive pairs of coordinate values. This model is supposed to include some of the candidate grouping effects discovered in Diaconis (1989). The model leads to $5 \times 5 + 4 \times (5 \times 4) = 105$ constraints on $120 = 5!$ variables. Each sequence $(i_1, i_2, i_3, i_4, i_5)$ is a permutation of $(1, 2, 3, 4, 5)$, so no repetition is allowed. Therefore if one considers the data to be a 5-way table with 5 levels in each dimension, one must force structural zeros on $5^5 - 5!$ entries which makes generic loglinear model fitting difficult. For the estimate $\hat{\mathbf{p}}$ in the normal proposal, we used the normalized fitted values from Poisson regression on a model of no interaction, ignoring structural zeros, which seemed to work as well as any other choice.

It looks like there are 105 parameters, but because of redundancy in the parameterization, the 105×120 constraint matrix from sufficient statistics has

Table 4. Whittaker's (1990) survey data on women's economic activity.

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0
5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

rank 58, the degrees of freedom is only 62, and the model has dimension 57 (57 free parameters). Whereas it is now possible to obtain the Markov basis for the first-order model in a few hours with 4ti2, we were not able to get the basis for the second-order model above. This requires at least several days of computing time.

Based on 1,000 samples, which took about 15 minutes to generate, SIS with normal proposal estimated a p-value of 1.78×10^{-5} with standard error 1.30×10^{-5}. The cv^2 value is 18.4. The p-value estimate indicates that the second-order model that includes effects for consecutive pairs of candidates, does not fit well. Other models have been proposed in Chung and Marden (1993).

Example 4. The vector of counts of 8-way binary data in Table 4 is reported in Whittaker (1990, p.280). It concerns a survey of 665 households with each response classified in one of 2^8 ways based on 8 "no or yes" responses to economic and employment questions. It is given in lexicographic order, that is the cells are numbered 00000000, 00000001, 00000010, etc. where 0 codes for "no" on a particular question.

Fitting the "all two-way interaction" model in Whittaker's terminology (i.e., no-3-way interaction) in R gives a likelihood ratio statistic of 144.6 on 219 residual degrees of freedom, and p-value over 0.99 on the $\chi^2(219)$ scale. We will see that the exact test gives a much smaller value, although one that does not contradict reasonable model fit. For the exact analysis, the Markov basis is quite difficult to find, and we were not able to compute it in 4ti2.

The SIS method does not work well using the original cell ordering. More than one reordering can be used to get the cv^2 value down to a quantity generally

Table 5. Comparison of SIS with normal proposal on Whittaker's (1990) survey data with different orders of the cells.

	SIS normal, reordered	SIS normal, not reordered
Estimate	0.186	0.223
Standard error	0.041	0.091
Sample size	8,000	8,000

in the hundreds. The fraction of valid tables was 99.7%. In our analysis, we chose to do eight SIS runs of size 1,000 each, in order to get a more reliable estimate of the standard error than the expression (5.4), which can be unstable and misleading on these data. The p-values in the table for SIS are the averages of the eight runs, and the standard error is the sample standard deviation divided by $\sqrt{8}$. The total running time over all eight runs was about 7 hours. Each run had its own cv^2 value, and those for the reordered data with larger counts near the beginning were better– the cv^2 values on the original order were generally twice as high and with more variability across runs. The comparison of the standard errors shows that the algorithm on the reordered data is about $(0.091/0.041)^2 \approx 5$ times more efficient than the algorithm on the data in the original order.

8. Conclusions

High dimensional contingency tables are common structures for data from research in biology and the social sciences. Despite significant advances, these tables remain challenging to analyze. Further research into sampling methods for Monte Carlo computations will enable researchers to analyze larger and more complex tables.

Certain types of constraints and features like structural zeros overwhelm existing methods. That is, examples persist where computing the Markov basis is not possible in reasonable time, and where other existing methods cannot generate valid tables. Methods like the one implemented in `exactLoglinTest` work fine on some standard loglinear models, but cannot handle large problems of no-3-way interaction.

The method that is most promising is a type of sequential sampling where cell intervals are computed in sequence using linear programming, and a normal proposal distribution is updated sequentially. This method generates nearly 100% valid tables in almost all examples and can deliver good answers on the hardest problems. In particular, its ability to generate valid tables seems to "scale well" in the number of cells within difficult model families like no-3-way interaction. Another good property is that the quality of the Monte Carlo estimates can be judged using standard tools of importance sampling like cv^2. The SIS sampling

method can also be distributed easily over many processors because random tables are independent, unlike Markov chains.

Certain difficulties remain in the implementation of the SIS algorithm: clearer diagnostics on the accuracy of the results are needed; the order of the cell sampling is critically important but hard to systematize; the speed of the algorithm needs to be improved. The speed of the algorithm depends very much on large amounts of linear programming, and improvements in this aspect of the implementation may be key to pushing on to larger and larger tables. Other methods for obtaining bounds may provide improvements, such as the generalized shuttle algorithm (Dobra (2002)). Choosing the normal proposal in an optimal way, and also choosing the cell ordering in an automated and optimal fashion are interesting technical problems that could improve the algorithm.

Acknowledgements

The authors thank the Editor, an associate editor, and two referees for many helpful suggestions. Yuguo Chen's research was partly supported by the National Science Foundation grants DMS-0503981 and DMS-0806175.

References

4ti2 team (2006). 4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces. Available at `www.4ti2.de`.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52**, 119-126.

Berkelaar, M., Eikland, K. and Notebaert, P. (2004). *lpsolve: Open Source (Mixed-Integer) Linear Programming System.* GNU LGPL (Lesser General Public Licence).

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633-642.

Booth, J. G. and Butler, J. W. (1999). An importance sampling algorithm for exact conditional tests in loglinear models. *Biometrika* **86**, 321-332.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci.* **16**, 101-133.

Caffo, B. S. (2006). The `exactLoglinTest` Package. Available at `cran.r-project.org`.

Chen, Y., Diaconis, P., Holmes, S. P. and Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.* **100**, 109-120.

Chen, Y., Dinwoodie, I. H. and Sullivant, S. (2006). Sequential importance sampling for multiway tables. *Ann. Statist.* **34**, 523-545.

Chen, Y., Lin, C. H. and Sabatti, C. (2006). Volume measures for linkage disequilibrium. *BMC Genetics* **7**, 54.

Chung, L. and Marden, J. I. (1993). Extensions of Mallows' ϕ model. In *Probability Models and Statistical Analyses for Ranking Data* (edited by Fligner and Verducci), 108-139. Springer-Verlag, New York.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* Wiley, New York.

DeLoera, J. and Onn, S. (2006). Markov bases of three-way tables are arbitrarily complicated. *J. Symbolic Comput.* **41**, 173-181.

Deresiewicz, R. L., Thaler, S. J., Hsu, L. and Zamani, A. A. (1997). Clinical and neuroradiographic manifestations of Eastern equine encephalitis. *New England J. Medicine* **336**, 1867-1874.

Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17**, 949-979.

Diaconis, P. and Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *Ann. Statist.* **13**, 845-874.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26**, 363-397.

Dobra, A. (2002). Statistical tools for disclosure limitation in multi-way contingency tables. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University.

Eriksson, N., and Diaconis, P. (2005). Markov bases for noncommutative Fourier analysis of ranked data. arXiv:math.AC/0405060 v2.

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

Hall, S. J. G. and Ruane, J. (1993). Livestock breeds and their conservation: a global overview. *Conservation Biology* **7**, 815-825.

Harter, R., Hornik, K. and Theussl, S. (2008). *Rsymphony: An R interface to the SYMPHONY MILP solver.* GNU GPL (General Public License).

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76**, 33-50.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses.* 2nd Edition. Springer, New York.

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

Marsaglia, G. (1964). Conditional means and covariances of normal variables with singular covariance matrix. *J. Amer. Statist. Assoc.* **59**, 1203-1204.

R Development Core Team (2006). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org.`

StatXact (2004). `www.cytel.com.`

Whittaker, J. (1990). *Graphical Models in Applied Mathematical Statistics.* Wiley, New York.

Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, PO Box 751, Portland, OR 97207-0751, USA.

E-mail: ihd@pdx.edu

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu

(Received March 2008; accepted May 2010)