Statistics and Its Interface Volume 13 (2020) 157–166

Sampling high dimensional tables with applications to assessing linkage disequilibrium*

Robert D. Eisinger, Xiao Su, and Yuguo Chen†,‡

We propose a sequential importance sampling strategy to sample high dimensional tables with fixed one way margins. The proposal distribution for the method is constructed by adapting an approximation to the total number of tables available in the literature. We apply the method to estimating the total number of tables and assessing linkage disequilibrium in multimarker genetic data with the table representing haplotype count data. We demonstrate efficient and accurate performance in these practical, real-world examples. The method may be applied in any situation in which uniformly sampling high dimensional tables with fixed one way margins is of interest. Detailed derivations are provided in the appendix.

Keywords and phrases: Counting problem, Exact test, High dimensional table, Linkage disequilibrium, Monte Carlo method, Sequential importance sampling.

1. INTRODUCTION

We are interested in the problem of sampling high dimensional tables uniformly from the set of all possible tables with fixed one way margins. This problem has a number of applications, including assessing linkage disequilibrium in multimarker genetic data. A marker site is represented by one dimension of a high dimensional table and alleles at a specific marker site are the rows of that margin. The one way marginal sums represent the number of individuals that have a specific allele at a given marker site. This marginal contains no information about recombination between marker sites, so we condition on the marginal sums when assessing linkage disequilibrium. We employ volume measures, which require us to sample high dimensional tables with fixed one way margins uniformly from the set of all possible high dimensional tables. We are also interested in estimating the total number of tables with fixed one way margins.

Several methods exist for these problems. A method for exact enumeration of all tables consistent with general constraints for high dimensional tables has been provided in Dobra and Fienberg [5], and a general method for evaluating the number of tables fulfilling a general set of constraints was provided in Barvinok [1]. Chen et al. [2] developed an importance sampling method for this problem using a uniform proposal distribution for each cell and based on the ideas of computational commutative algebra, however, this method encounters difficulties when sampling large sparse tables with fixed one way margins. Markov chain Monte Carlo (MCMC) methods based on Diaconis and Sturmfels [4] are possible, but it is often difficult to design irreducible Markov chains in high dimensional cases, and they can take a long time to explore the space of possible tables. Lazzeroni and Lange [10] developed an MCMC method for testing linkage and Hardy-Weinberg equilibrium in multidimensional contingency tables.

We will employ the method of importance sampling to sample contingency tables with fixed one way margins. Tables are sampled from a distribution that is close to uniform and then the tables are weighted to correct for the bias. This method allows for the estimation of both the number of tables and the distribution under the null hypothesis of a uniform distribution for any test statistic of interest. We leverage an approximation to the number of tables from Good [8] to develop the proposal distribution for sequential importance sampling (SIS) and demonstrate that the SIS procedure performs well in the task of estimating tables and in genetic applications.

This paper is organized in the following way. Section 2 introduces the basics of SIS, with the proposal distribution for SIS based on the approximation of Good [8] developed in Section 2.2. Section 2.4 describes the problem of assessing linkage disequilibrium in multimarker genetic data when there are more than two alleles at each marker. Section 3 demonstrates results, including estimating the number of tables and the volume measure for assessing linkage disequilibrium. Section 4 provides concluding remarks.

2. MATERIALS AND METHODS

2.1 Sequential importance sampling

Denote by $\mathbf{X} = \{X_1, \ldots, X_k\}$, a vector of k random variables cross-classified in a k dimensional table, T, where X_i takes values in $\{1, \ldots, I_i\}$. Let Σ denote the set of all $I_1 \times \cdots \times I_k$ tables T, with entries $t_{i_1 \ldots i_k}$ and with

*The authors thank Dr. Chia-Ho Lin for sharing the multimarker genetic data for studying bipolar disorder.
†Partially supported by the NSF grant DMS-1406455.
‡Corresponding author.

pre-specified one way marginal sums $n_j^{[i_j]}$, $j = 1, \ldots, k$, $i_j = 1, \ldots, I_j$, i.e.,

$$\sum_{i_l : l \neq j} t_{i_1 \ldots i_{j-1} i_j i_{j+1} \ldots i_k} = n_j^{[i_j]}, \quad \text{for } j = 1, \ldots, k, \text{ and } i_j = 1, \ldots, I_j. \tag{1}$$

Denote by $\mathbf{n}_j = \{n_j^{[1]}, \ldots, n_j^{[I_j]}\}$ the set of one way margins summing over all but dimension j. Let $M = \sum_{i=1}^{I_1} n_1^{[i]} = \cdots = \sum_{i=1}^{I_k} n_k^{[i]}$ be the overall table sum. Let $\pi(T) = 1/|\Sigma|$ be the uniform distribution over Σ, where $|\Sigma|$ is the total number of tables with the specified dimension and marginal sums.

We are interested in estimating $E_\pi[f(T)]$, where $f(T)$ is a function of table T (see an example of f in (19)). Sampling from the uniform distribution $\pi(T)$ is difficult, but if a high dimensional table, T, can be sampled from a proposal distribution, $q(\cdot)$, that is easy to sample from and includes Σ in its support, then $E_\pi[f(T)]$ can be estimated using the weighted average,

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(T)(\pi(T_i)/q(T_i))}{\sum_{i=1}^{N} (\pi(T_i)/q(T_i))} = \frac{\sum_{i=1}^{N} f(T)(\mathbb{1}_{\{T_i \in \Sigma\}}/q(T_i))}{\sum_{i=1}^{N} (\mathbb{1}_{\{T_i \in \Sigma\}}/q(T_i))}, \tag{2}$$

where $T_1, \ldots, T_N$ are independent, identically distributed (iid) samples from $q(T)$, and $\pi(T_i)/q(T_i)$ is the importance weight.

Additionally, the total number of tables $|\Sigma|$ can be written as

$$|\Sigma| = \sum_{T \in \Sigma} \frac{1}{q(T)} q(T) = E_q\left[\frac{\mathbb{1}_{\{T \in \Sigma\}}}{q(T)}\right], \tag{3}$$

and estimated using

$$\widehat{|\Sigma|} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}_{\{T_i \in \Sigma\}}}{q(T_i)}. \tag{4}$$

The efficiency of the estimator can be assessed using a straightforward application of the Δ-method,

$$\mathrm{se}(\hat{\mu}) \approx \sqrt{\frac{\mathrm{var}_q(\frac{f(T)\pi(T)}{q(T)}) + \mu^2 \mathrm{var}_q(\frac{\pi(T)}{q(T)}) - 2\mu \mathrm{cov}_q(\frac{f(T)\pi(T)}{q(T)}, \frac{\pi(T)}{q(T)})}{N}}, \tag{5}$$

or using the *effective sample size*, $\mathrm{ESS} = N/(1+\mathrm{cv}^2)$, where the *coefficient of variation* (cv) is

$$\mathrm{cv}^2 = \frac{\mathrm{var}_q(\pi(T)/q(T))}{E_q^2(\pi(T)/q(T))}. \tag{6}$$

The *effective sample size* approximates how many iid samples are equivalent to the N weighted SIS samples. The cv^2 is simply the χ^2 distance between the target and proposal distributions, where the sample version of cv^2 is used to evaluate the performance of SIS in practice.

2.2 Sampling high dimensional tables

Sampling tables from Σ is a high dimensional problem, so the strategy is to decompose the table into lower dimensional components and sample sequentially using a suitable proposal distribution. Choosing a proposal distribution that is close to our target distribution for each component will result in an efficient procedure.

The proposal for an entire table $q(T)$ is constructed sequentially cell by cell,

$$\begin{aligned} q(T) =& q(t_{11\ldots1}) q(t_{21\ldots1} | t_{11\ldots1}) \ldots \\ & \ldots q(t_{I_1 I_2 \ldots I_k} | t_{11\ldots1}, \ldots, t_{(I_1-1) I_2 \ldots I_k}). \end{aligned} \tag{7}$$

The cell $(1, 1, \ldots, 1)$ is sampled first, conditional on the observed table margins $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_k\}$. Then the margins are updated and the cell $(2, 1, \ldots, 1)$ is sampled next, conditional on the realization of the first cell. After the first cell has been sampled, the margins $n_j^{[1]}$, $j = 1, \ldots, k$, are updated by subtracting the value of the sampled cell,

$$n_j^{*[1]} = n_j^{[1]} - t_{11\ldots1},$$

and the remaining margins are unchanged, so the updated margins are $\mathbf{n}_j^* = \{n_j^{*[1]}, n_j^{[2]} \ldots, n_j^{[I_j]}\}$ for $j = 1, \ldots, k$.

To motivate the development of the proposal distribution, we begin by writing the true marginal distribution for the first cell $t_{11\ldots1}$,

$$p(t_{11\ldots1} = a_{11\ldots1}) = \frac{|\Sigma^*|}{|\Sigma|}, \tag{8}$$

where Σ^* denotes the number of tables with marginals $\{\mathbf{n}_1^*, \ldots, \mathbf{n}_k^*\}$ and a structural zero in the first cell $t_{11\ldots1}$. Both the numerator and denominator of this expression are difficult to calculate, but Good [8] provided an approximation for high dimensional tables with fixed one way margins.

Good's Approximation. *[8] Let* $\mathcal{I}^{[-j]} = \prod_{i:i\neq j} I_i = I_1 I_2 \cdots I_{j-1} I_{j+1} \cdots I_k$. *Then,*

$$|\Sigma| \approx \Delta^G \equiv \frac{\prod_{i_1=1}^{I_1} \binom{n_1^{[i_1]} + \mathcal{I}^{[-1]} - 1}{n_1^{[i_1]}} \prod_{i_2=1}^{I_2} \binom{n_2^{[i_2]} + \mathcal{I}^{[-2]} - 1}{n_2^{[i_2]}} \cdots \prod_{i_k=1}^{I_K} \binom{n_k^{[i_k]} + \mathcal{I}^{[-k]} - 1}{n_k^{[i_k]}}}{\binom{M + I_1 I_2 \times \cdots \times I_{k-1} I_k - 1}{M}^{k-1}}. \tag{9}$$

This approximation has an informative combinatorial interpretation that will be used to our advantage to construct the sequential importance sampling proposal. Here Δ^G is

the product of the number of ways to arrange each marginal sum divided by the number of $I_1 \times \cdots \times I_k$ tables with sum M, $k-1$ times. So in the uniform probability space on all possible tables with table sum M, it is the product of the probabilities that the j^{th} marginal sum equals $\mathbf{n}_j$ times the total number of k dimensional tables with the prescribed dimensions.

Using a similar approach as the one used to sample two way tables [6], we will leverage this approximation to construct our proposal distribution, denoted by SIS-G. The derivation of this proposal is provided in the appendix.

Proposal 1. *The proposal for the first cell $t_{11\ldots1}$, constructed based on the approximation of Good [8] to $|\Sigma|$, is*

$$q(t_{11\ldots1} = a_{11\ldots1}) \propto \frac{\prod_{j=1}^{k} \binom{n_j^{[1]} - a_{11\ldots1} + \mathcal{I}^{[-j]} - 2}{n_j^{[1]} - a_{11\ldots1}}}{\binom{M - a_{11\ldots1} + I_1 \times \cdots \times I_k - 2}{M - a_{11\ldots1}}^{k-1}}, \tag{10}$$

and the support of this distribution is the set of values that the first cell of all tables in Σ can take.

The first cell will be sampled according to this distribution using multinomial sampling with normalized probabilities. The subsequent cells will be sampled in a similar way, updating the margins and forcing the sampled entries to be structural zeroes. The structural zeros are handled in the approximation (10) by leveraging the combinatorial interpretation, subtracting from $\mathcal{I}^{[-j]}$ the number of cells that have been sampled in the relevant margin. The product $I_1 \times \cdots \times I_k$ in the denominator is updated by subtracting the total number of cells in the overall table that have already been sampled. More precisely, the cell $t_{i_1\ldots i_k}$ will be sampled according to

$$q(t_{i_1\ldots i_k} = a_{i_1\ldots i_k}) \propto \frac{\prod_{j=1}^{k} \binom{n_j^{*[i_j]} - a_{i_1\ldots i_k} + \mathcal{I}^{[-j]} - S_j^{[i_j]} - 1}{n_j^{*[i_j]} - a_{i_1\ldots i_k}}}{\binom{M^* - a_{i_1\ldots i_k} + I_1 \times \cdots \times I_k - S - 1}{M^* - a_{i_1\ldots i_k}}^{k-1}}, \tag{11}$$

where M^* and $n_j^{*[i_j]}$ are the updated table sum and marginal sums before sampling cell $t_{i_1\ldots i_k}$, $S_j^{[i_j]}$ is the number of structural zeros in the i_j-th layer of margin j, corresponding to the total number of cells in the relevant layer that have already been sampled by the importance sampling procedure, and S is the total number of cells in the overall table that have already been sampled. Similar to the proposal for the first cell, the support of the proposal (11) is the set of values that the cell $t_{i_1\ldots i_k}$ of all tables in Σ can take. We will show in the next section that the proposal (11) is well-defined.

2.3 Calculation of bounds

Sampling by cell requires us to calculate a set of entries to sample from that includes the support of the true conditional distribution of the cell. This can be a difficult and computationally intensive problem. In situations where the support of the cells are intervals, the sequential interval property is said to hold, and instead of calculating a set of viable entries, we may calculate the true lower and upper bounds, sample an integer between these two values and guarantee 100% valid entries.

Even when the support of the cells are not intervals, we can still calculate the lower and upper bounds, extend the support of the distribution in (10) to the interval determined by the lower and upper bounds, and sample from the interval between the two bounds using multinomial sampling with normalized probabilities. Because the interval contains invalid values for the cell when the sequential interval property does not hold, the sampling may generate invalid tables. Those invalid tables will receive zero importance weights (because they are not in the support of $\pi(T)$) and will not contribute to the importance sampling estimate. The sampling procedure is still valid because importance sampling allows the proposal distribution to have a larger support than the target distribution $\pi(T)$.

In the case of two way tables with fixed row and column sums, the lower and upper bounds are easy to calculate for each cell and the sequential interval property holds. For higher dimensional tables, there is generally not an easy way to calculate the bounds, and we must resort to more computationally intensive methods. There are a number of methods for calculating the lower and upper bounds for the cell entries in high dimensional tables. The first of these is integer programming, which always gives the exact integer bounds, but is very slow to implement. Another method is linear programming, implemented in the R package lpSolve. This method must be implemented carefully, as it is possible for linear programming to return wider intervals than the true bounds. Linear programming is computationally intensive, but generally provides accurate results, generating 100% valid tables in each of the tables examined. We suspect, based on this result and extensive testing of a wide range of high dimensional tables with fixed one way margins, that the sequential interval property holds in this situation.

The computation time of integer and linear programming, along with empirical results in favor of the sequential interval property, leads us to pursue a method that calculates bounds for a cell extremely quickly. Although these bounds may be wider than the true bounds and thus risk sampling a value that does not correspond to a valid high dimensional table, the gain in computational efficiency makes the method attractive in practice.

These bounds will be developed by extending standard bounds for high dimensional tables available in the literature. These are the Fréchet bounds for k-way tables with fixed one way margins examined in [7, 9, 13, 15], and repro-

duced below for cell $(i_1, i_2, \ldots, i_k)$,

$$\max\left(0, \sum_{j=1}^{k} n_j^{[i_j]} - (k-1)M\right) \leq t_{i_1 \ldots i_k} \leq \min\left(n_1^{[i_1]}, n_2^{[i_2]}, \ldots, n_k^{[i_k]}\right). \tag{12}$$

These bounds need to be extended to the case where a sequence of cells has already been sampled. If $n_j^{*[i_j]}$ denotes the updated margin after sequentially sampling, and M^* denotes the updated overall table sum, then a natural extension of the Fréchet bounds are

$$\max\left(0, \sum_{j=1}^{k} n_j^{*[i_j]} - (k-1)M^*\right) \leq t_{i_1 \ldots i_k} \leq \min\left(n_1^{*[i_1]}, n_2^{*[i_2]} \ldots, n_k^{*[i_k]}\right). \tag{13}$$

These bounds are denoted by $[l_f, u_f]$, and may be used in a sequential importance sampling procedure, but will generate a certain percentage of invalid tables. An additional, more strict bound is obtained when $i_z = I_z$ for any $z = 1, \ldots, k$, the derivation of which is provided in the appendix. Combining these two bounds yields the following general bounds for $t_{i_1 \ldots i_k}$,

$$[l, u] = \begin{cases} \left[\max\left(0, n_k^{*[i_k]} - \sum\limits_{j \neq k} \sum\limits_{i'_j = i_j + 1}^{I_j} n_j^{*[i'_j]}\right), u_f\right], \\ \qquad \text{if } i_z = I_z \text{ for any } z = 1, \ldots, k, \text{ or} \\ \qquad n_z^{*[i_z+1]} = \cdots = n_z^{*[I_z]} = 0, \text{ for any} \\ \qquad z = 1, \ldots, k, \\ [l_f, u_f], \quad \text{otherwise.} \end{cases} \tag{14}$$

Now we show that the proposal distribution (11) is well-defined within the above bound $[l, u]$. For any integer $a_{i_1 \ldots i_k} \in [l, u]$, we have

$$a_{i_1 \ldots i_k} \leq u_f = \min\left(n_1^{*[i_1]}, n_2^{*[i_2]} \ldots, n_k^{*[i_k]}\right) \leq n_j^{*[i_j]} \quad \text{for } j = 1, \ldots, k. \tag{15}$$

Since $S_j^{[i_j]}$ denotes the total number of cells that have already been sampled in the current layer of margin j, we have $S_j^{[i_j]}$ is at most $\mathcal{I}^{[-j]} - 1$. Therefore

$$n_j^{*[i_j]} - a_{i_1 \ldots i_k} + \mathcal{I}^{[-j]} - S_j^{[i_j]} - 1 \geq n_j^{*[i_j]} - a_{i_1 \ldots i_k} \quad \text{for } j = 1, \ldots, k. \tag{16}$$

Combining (15) and (16) together we have

$$\binom{n_j^{*[i_j]} - a_{i_1 \ldots i_k} + \mathcal{I}^{[-j]} - S_j^{[i_j]} - 1}{n_j^{*[i_j]} - a_{i_1 \ldots i_k}} > 0 \quad \text{for } j = 1, \ldots, k,$$

which means the binomial coefficients in the numerator of the right hand side of (11) are strictly positive. Similarly, we can also show that the binomial coefficient in the denominator of the right hand side of (11) is also positive. Thus, the proposal distribution (11) is well-defined within the above bound $[l, u]$.

These bounds (14) may be wider than the true bounds and thus generate a small percentage of invalid tables, but the gain in method efficiency over other methods of calculating bounds is dramatic, especially for large, high dimensional tables. Extensive simulations indicate that the adapted bounds described in (14) are generally 2 to 3 times more efficient than competing methods. The difference in computation becomes even larger as the dimension of the table increases, and for extremely large tables, linear programming takes too much time to run in practice. Unless otherwise stated, the bounds in (14) will be used for sampling and the percentage of invalid tables will be reported as necessary.

2.4 Linkage disequilibrium

Linkage disequilibrium refers to the association between quantitative random variables corresponding to alleles at different loci on a chromosome. We say the loci are in linkage disequilibrium if the observed frequency of a particular combination of alleles is different from the expected for random association. Measuring linkage disequilibrium assists in testing genetic hypotheses, mapping the genome and understanding genome structure. A number of measures exist for assessing linkage disequilibrium for pairs of biallelic markers (markers with only two possible alleles at a specific locus), and several of these measures have been extended to assess linkage disequilibrium for pairs of multiallelic markers [3, 12]. Chen et al. [3] extended methods for assessing linkage disequilibrium in biallelic markers to the multiallelic marker case using volume measures. We will extend this method further to encompass the case where there are more than two multiallelic markers.

The basic idea of volume measure is that given some quantity that measures the divergence between the observed table S and the table expected under linkage equilibrium, a volume measure is defined as the proportion of tables $T \in \Sigma$ that lead to a smaller divergence value. The volume measure will be zero if all other tables have larger divergences, and the volume measure will be close to one if the observed divergence is the largest possible [14].

We first give a brief discussion of linkage disequilibrium for two markers where each marker can take one of two possible alleles, following Chen et al. [3]. The haplotype distribution, ρ, of two markers with alleles $\{A, a\}$ and $\{B, b\}$ is

	B	b	
A	x	$p - x$	p
a	$q - x$	$1 - p - q + x$	$1 - p$
	q	$1 - q$	1

.

If the marginals are fixed, ρ is determined by the probability x, and the magnitude of the disparity between ρ and linkage equilibrium can be quantified by $x - pq$. Various standardized measures of this quantity have been proposed and examined, such as

$$D = x - pq \quad \text{and} \quad r^2 = \frac{D^2}{pq(1-p)(1-q)}.$$

Generally, these measures of linkage disequilibrium are defined on ρ, but this population distribution is usually unknown. A practical and effective solution to this problem that also allows us to examine multiple markers with more than two alleles at each polymorphic site is provided by volume measures [3]. Volume measures naturally account for sample size, have a simple intuitive interpretation and can be readily applied to this situation [14].

To apply volume measures, we evaluate the total number of tables out of all possible tables with fixed margins that have a smaller divergence from linkage equilibrium than what was observed. Since ρ is unknown, volume measures are defined on the sample haplotype data, and we evaluate the proportion of high dimensional tables that have a lower level of divergence than what was observed in our data. The one way margins are fixed because this quantity corresponds to the total number of individuals that have a specific allele at a given marker site. This quantity provides no information about the amount of recombination, so we condition on these marginals [14]. If all other tables have larger divergences, the volume measure will be zero, and if the observed divergence is one of the largest possible, the volume measure will be near one.

The volume measure *Mvol* was defined for pairs of markers in Chen et al. [3], and can be readily extended to the case where there are more than two markers. *Mvol* is

$$Mvol(S) = \frac{1}{|\Sigma|} \sum_{T \in \Sigma} \mathbb{1}_{\{M(T) < M(S)\}}, \tag{17}$$

where

$$M(T) = \sum_{i_1,\ldots,i_k} \frac{(t_{i_1\ldots i_k} - \prod_{j=1}^{k} n_j^{[i_j]}/M^{k-1})^2}{\prod_{j=1}^{k} n_j^{[i_j]}/M^{k-1}}, \tag{18}$$

and M is the table sum and k is the dimension of the table.

If all markers with multiple alleles are independent (linkage equilibrium), then the divergence $M(S)$ for the observed table S tends to be small, which will lead to a volume measure $Mvol(S)$ that is close to 0. If there is strong linkage disequilibrium among all markers, the divergence $M(S)$ tends to be large, which will lead to a volume measure $Mvol(S)$ that is close to 1. In the case of two way tables, the volume measure *Mvol* is one minus the p-value for the χ^2 test of independence.

Similar to the traditional linkage disequilibrium measure r^2, the volume measure *Mvol* also takes values between 0 and 1, with larger values indicating strong linkage disequilibrium. However, unlike D and r^2, the volume measure *Mvol* does not require the knowledge of the population haplotype distribution ρ (which is usually unknown). In addition, the extended version of *Mvol* can handle multiple markers with multiple alleles at each polymorphic site, which is more general than D and r^2.

However, assessing *Mvol* requires examining all tables in Σ, the set of all $I_1 \times \cdots \times I_k$ tables with margins $n_j^{[i_j]}$, $j = 1, \ldots, k$, and $i_j = 1, \ldots, I_j$. This is generally not feasible, so instead we use our proposal SIS-G to sample tables $T_1, \ldots, T_N$ from Σ, assign each sampled table an importance weight, and estimate *Mvol* by (2) with

$$f(T) = \mathbb{1}_{\{M(T) < M(S)\}}. \tag{19}$$

Results for assessing linkage disequilibrium for real genetic data is provided in Section 3.2.

3. NUMERICAL RESULTS

3.1 Estimating the number of tables

The number of high dimensional tables with fixed one way margins is difficult to calculate, and exhaustive enumeration is generally infeasible. The sequential sampling algorithm proposed in Section 2 can be used to estimate $|\Sigma|$. In all examples in this section, we sample high dimensional tables cell by cell as illustrated in (7). For each cell $t_{i_1\ldots i_k}$, we first calculate the lower and upper bounds for that cell using (14), and then sample a value for that cell from the proposal distribution based on Good's approximation in (10). We keep track of the sampling probability for each cell to compute the proposal $q(T)$ after the whole table T is sampled. Finally $|\Sigma|$ can be estimated using (4). We demonstrate the performance in a few examples. All simulation was done on a MacBook Pro with a 2.6 GHz processor with coding performed in R.

First, we examine some small high dimensional tables with equal margins. The first is a $3 \times 3 \times 3$ table with all one way margins equal to 3, and the second is a $3 \times 3 \times 3$ table with all one way margins equal to 20. We also examine a $3 \times 3 \times 5$ table with all one way margins equal to 30 or 50. The simulation results based on 1,000 importance samples are presented in Table 1. The number following the $\pm$ sign denotes the standard error. The estimated number of tables are close to the true number of tables for the first two examples, which are 22,620 and 642,635,414,923,248, respectively, calculated using LattE [1]. However LattE is much slower than the sequential sampling algorithm. For the third example, it is not feasible to calculate the true number of tables using LattE, but SIS-G can give an estimate in a few seconds.

Table 1. Results for estimating the number of small high dimensional tables

Estimated number of tables	cv^2	Time (sec)
$3 \times 3 \times 3$ table with margins $= 3$		
$(2.2259 \pm 0.0474) \times 10^4$	0.4548	0.7
$3 \times 3 \times 3$ table with margins $= 20$		
$(6.5931 \pm 0.1827) \times 10^{14}$	0.7728	1.0
$3 \times 3 \times 5$ table with margins $= 30,\ 50$		
$(5.3472 \pm 0.1643) \times 10^{32}$	0.9444	1.8

Table 1 also shows that the cv^2 is smaller than 1 for all three tables, which indicates that the proposal distribution is quite close to the target uniform distribution in terms of the χ^2 distance. It also implies that the effective sample size of the 1,000 importance samples is quite large. For example, the cv^2 of the first table is 0.4548, so the effective sample size is about $1000/(1 + 0.4548) \approx 687$, which means the $1,000$ importance samples from the proposal distribution are roughly equivalent to 687 iid samples from the target distribution. The percentage of valid tables is 100% for all three tables.

We also examined a few more challenging high dimensional tables. They are a $5 \times 5 \times 5 \times 5$ table with margins $\{4, 4, 3, 1, 2\}$, $\{4, 3, 3, 2, 2\}$, $\{4, 3, 3, 2, 2\}$, $\{1, 1, 2, 4, 6\}$, a $3 \times 3 \times 3 \times 3 \times 2 \times 2 \times 2 \times 2$ table with margins $\{11, 3, 2\}$, $\{8, 4, 4\}$, $\{8, 4, 4\}$, $\{8, 4, 4\}$, $\{9, 7\}$, $\{9, 7\}$, $\{8, 8\}$, $\{11, 5\}$, a $5 \times 3 \times 3 \times 4 \times 3 \times 2$ table with margins $\{2, 2, 2, 2, 2\}$, $\{3, 3, 4\}$, $\{2, 4, 4\}$, $\{3, 3, 3, 1\}$, $\{2, 4, 4\}$, $\{5, 5\}$, and finally an eight-dimensional $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 \times 2$ table with margins $\{10, 10\}$, $\{7, 13\}$, $\{12, 8\}$, $\{9, 11\}$, $\{10, 10\}$, $\{8, 12\}$, $\{6, 7, 7\}$, $\{6, 14\}$. The simulation results based on $1,000$ importance samples are presented in Table 2.

Table 2. Results for estimating the number of challenging high dimensional tables

Estimated number of tables	cv^2	Time (sec)
$5 \times 5 \times 5 \times 5$ table		
$(2.5223 \pm 0.1132) \times 10^{17}$	2.0129	14.7
$3 \times 3 \times 3 \times 3 \times 2 \times 2 \times 2 \times 2$ table		
$(1.3323 \pm 0.0696) \times 10^{25}$	2.7256	43.7
$5 \times 3 \times 3 \times 4 \times 3 \times 2$ table		
$(5.2420 \pm 0.1818) \times 10^{15}$	1.8006	32.9
$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 \times 2$ table		
$(2.2704 \pm 0.1046) \times 10^{25}$	2.1241	15.8

We continue to observe in Table 2 that the number of tables are estimated well based on the standard error and the cv^2. The sequential sampling algorithm can obtain the estimate in less than a minute in all cases, while it is not feasible to use LattE to calculate the true number of tables with these margins. The percentage of valid tables is 100% for all four high dimensional tables.

3.2 Linkage disequilibrium

We apply the volume measures described in Section 2.4 to assess linkage disequilibrium for multimarker genetic data. The data are 157 phase-known non-transmitted chromosomes 2 of parents of BP-I persons from Costa Rica's Central Valley. The chromosomes were genotyped with 85 markers [11]. We examine all possible sets of marker triplets for the first ten markers along the chromosome, and use $Mvol$ to evaluate the level of disequilibrium among these marker triplets. For convenience, we denote the ten markers by $\{1, 2, \ldots, 10\}$, and denote the number of different observed alleles at marker j by I_j, $j = 1, \ldots, 10$. For a marker triplet (i, j, h), we compute the volume measure $Mvol$ for the $I_i \times I_j \times I_h$ contingency table using (17).

Representing the volume measures for all marker triplets (i, j, h) would require a 3 dimensional figure, which is not easy to display. Instead, in Figure 1, we present each slice of the 3 dimensional figure which corresponds to the volume measures of marker triplets (i, j, h) with h fixed. So the x-axis represents the index for marker i, the y-axis represents the index for marker j, and the title of the plot represents the index of marker h. For example, the value with x-axis equal to 6 and y-axis equal to 3 in plot 5 denotes the volume measure for marker triplet $(6, 3, 5)$. Note that in plot h, the value h does not appear in the x-axis and y-axis because we are considering three different markers. Also in each plot, we did not compute the volume measures for the diagonal because they correspond to $i = j$. We only presented nine plots because the volume measure for triplet $(i, j, 10)$ is the same as the volume measure for $(10, i, j)$ which is already reported in the first nine plots.

For all sets of three markers, we used $1,000$ importance samples which took less than an hour. In the course of sampling, no invalid tables were generated. These results in Figure 1 indicate marker triplets that have high levels of $Mvol$, suggesting dependency and linkage disequilibrium. The results are consistent with previous analyses of these data [3], and also identify additional regions with high levels of linkage disequilibrium, suggesting high levels of recombination and a larger distance between the marker sites [14].

4. DISCUSSION

We have developed a sequential importance sampling strategy for sampling high dimensional tables with fixed one way margins based on an approximation of Good [8]. Applications to estimating the number of tables and assessing linkage disequilibrium have been examined and effective performance has been demonstrated.

The table may be sampled in any order. The best performance is usually obtained by first arranging the dimension size in decreasing order, and then ordering the marginal sums from the highest to lowest in each dimension. This or-

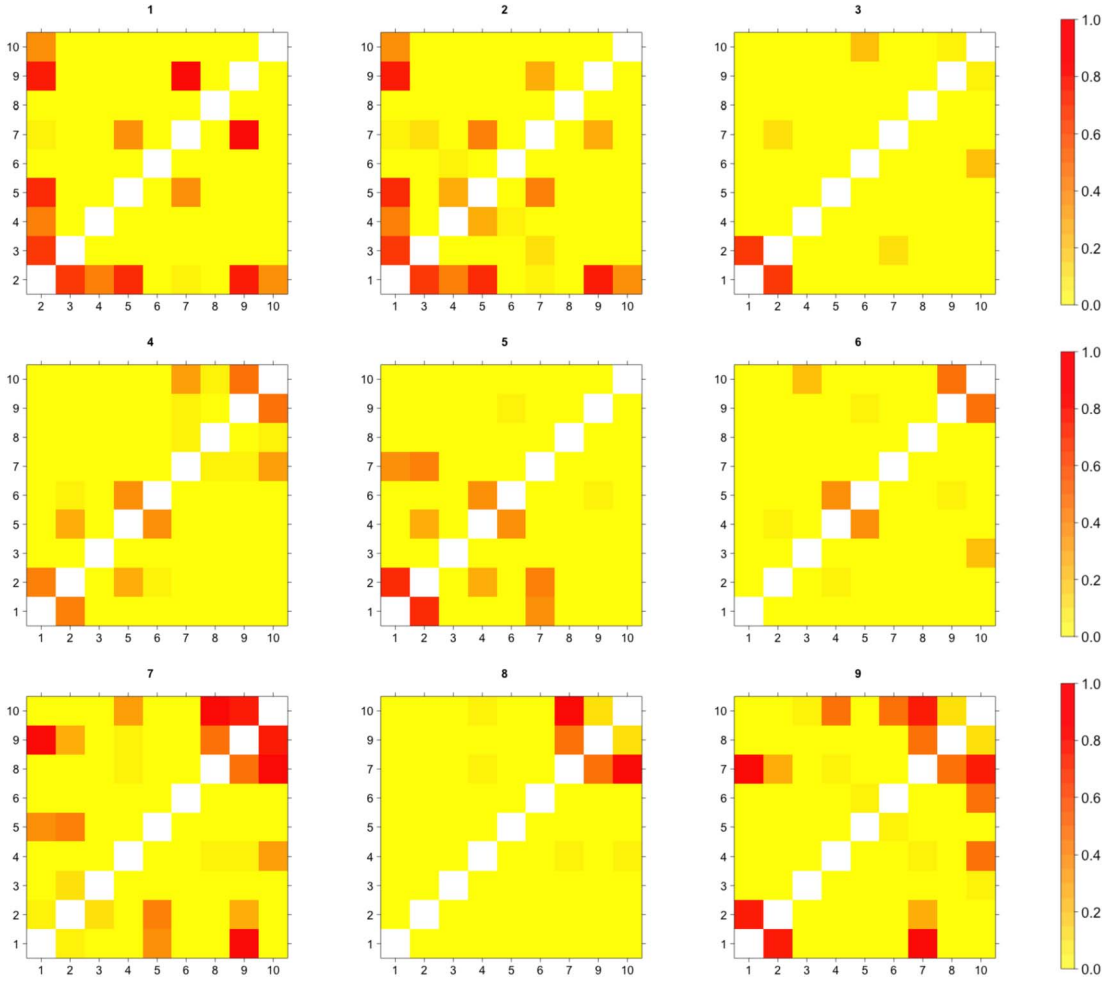

Figure 1. Measures of linkage disequilibrium using $Mvol$ between all possible multiallelic marker triplets for ten markers. The markers are denoted by $\{1, 2, \ldots, 10\}$. The x-axis represents the index for marker i, the y-axis represents the index for marker j, and the title of the plot represents the index of marker h. So the value with x-axis equal to i and y-axis equal to j in plot h denotes the volume measure for marker triplet (i, j, h).

dering has the advantage of often resulting in a relatively small cv^2 and standard error. The ordering may also affect the percentage of invalid tables. We obtained 100% valid tables in all simulation studies. In the case of the genetic data described in Section 3.2, there were no invalid tables generated when the marginal sums are arranged in increasing order. If the columns are arranged in decreasing order, the percentage of invalid tables can be greater than zero, but is generally very small. Further research on the effect of ordering will be useful.

APPENDIX A

A.1 Derivation of Proposal 1

Recall the approximation to $|\Sigma|$ is

(20)
$$|\Sigma| \approx \Delta^G \equiv \frac{\prod_{i_1=1}^{I_1} \binom{n_1^{[i_1]}+\mathcal{I}^{[-1]}-1}{n_1^{[i_1]}} \prod_{i_2=1}^{I_2} \binom{n_2^{[i_2]}+\mathcal{I}^{[-2]}-1}{n_2^{[i_2]}} \cdots \prod_{i_k=1}^{I_k} \binom{n_k^{[i_k]}+\mathcal{I}^{[-k]}-1}{n_k^{[i_k]}}}{\binom{M+I_1 I_2 \times \cdots \times I_{k-1} I_k - 1}{M}^{k-1}}.$$

The approximation to $|\Sigma^*|$ is obtained by using the combinatorial interpretation of Δ^G. The new margins are given by $\mathbf{n}_j^*$, and instead of $\mathcal{I}^{[-1]}$ places for the margin with sum $n_j^{[i]}$, there are now $\mathcal{I}^{[-1]} - 1$. So a natural approximation Δ^{G*} to $|\Sigma^*|$ is

(21)
$$|\Sigma^*| \approx \Delta^{G*} \equiv \frac{\prod_{j=1}^{k} \binom{n_j^{[1]}-a_{11\ldots1}+\mathcal{I}^{[-j]}-2}{n_j^{[1]}-a_{11\ldots1}} \prod_{j=1}^{k} \prod_{i=2}^{I_j} \binom{n_j^{[i]}+\mathcal{I}^{[-j]}-1}{n_j^{[i]}}}{\binom{M-a_{11\ldots1}+I_1 \times \cdots \times I_k - 2}{M-a_{11\ldots1}}^{k-1}}.$$

Consequently, our proposal for the first cell is

(22)
$$q(t_{11\ldots1}=a_{11\ldots1}) \propto \frac{\Delta^{G*}}{\Delta^{G}} \propto \frac{\prod_{j=1}^{k}\binom{n_j^{[1]}-a_{11\ldots1}+\mathcal{I}^{[-j]}-2}{n_j^{[1]}-a_{11\ldots1}}}{\binom{M-a_{11\ldots1}+I_1\times\cdots\times I_k-2}{M-a_{11\ldots1}}^{k-1}}.$$

A.2 Derivation of bounds

Recall Σ is the set of all k dimensional tables with dimensions $I_1\times\cdots\times I_k$ and one way margins $n_j^{[i_j]}$, $j=1,\ldots,k$, $i_j=1,\ldots,I_j$, and $\mathbf{n}_j=\{n_j^{[1]},\ldots,n_j^{[I_j]}\}$. Also recall M^* and $n_j^{*[i_j]}$ are the updated table sum and marginal sums after sampling up to cell $(i_1,\ldots,i_k)$.

We define $\tilde{I}_j=\max\{1\le i'_j\le I_j : n_j^{*[i'_j]}>0\}$. If any $i_j\ge\tilde{I}_j$, the lower bound in (13) can be made more strict. For the case $i_j>\tilde{I}_j$, then $n_j^{*[\tilde{i}_j]}=0$, which indicates that the upper bound u_f in (13) is 0, so $t_{i_1\ldots i_k}$ has to be 0.

For the case $i_j=\tilde{I}_j$, we are interested in calculating bounds for the cell $(i_1,\ldots,\tilde{I}_j,\ldots,i_k)$. There are $k-1$ additional lower bounds, the first of which is obtained considering the remaining marginal sum to be sampled in the first dimension with index i_1, $n_1^{*[i_1]}$,

$$\begin{aligned}
n_1^{*[i_1]} = &\sum_{i'_k=i_k+1}^{I_k}\sum_{i'_{k-1}=1}^{I_{k-1}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_k}\\
&+\sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}}\sum_{i'_{k-2}=1}^{I_{k-2}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_{k-1} i_k}+\\
&\vdots\\
&+\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}}\sum_{i'_j=1}^{I_j}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_{j+1} i_{j+2}\ldots i_k}\\
&+\sum_{i'_j=\tilde{I}_j+1}^{I_j}\sum_{i'_{j-1}=1}^{I_{j-1}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_j i_{j+1}\ldots i_k}\\
&+\sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}}\sum_{i'_{j-2}=1}^{I_{j-2}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_{j-1}\tilde{I}_j\ldots i_k}+\\
&\vdots\\
&+\sum_{i'_2=i_2+1}^{I_2} t_{i_1 i'_2 i_3\ldots i_k}+t_{i_1\ldots i_k}. \qquad (23)
\end{aligned}$$

Since $n_j^{*[i'_j]}=0$ for $i'_j=\tilde{I}_j+1,\ldots,I_j$, we have $\sum_{i'_j=\tilde{I}_j+1}^{I_j}\sum_{i'_{j-1}=1}^{I_{j-1}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_j i_{j+1}\ldots i_k}=0$. When $i'_k>i_k$, we have

$$\begin{aligned}
n_k^{*[i'_k]} &= \sum_{i'_1=1}^{I_1}\sum_{i'_2=1}^{I_2}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}} t_{i'_1\ldots i'_{k-1} i'_k}\\
&= \sum_{i'_1\neq i_1}\sum_{i'_2=1}^{I_2}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}} t_{i'_1 i'_2\ldots i'_{k-1} i'_k}+\\
&\qquad \sum_{i'_2=1}^{I_2}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}} t_{i_1 i'_2\ldots i'_{k-1} i'_k},
\end{aligned}$$

so

$$\sum_{i'_2=1}^{I_2}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}} t_{i_1 i'_2\ldots i'_{k-1} i'_k}\le n_k^{*[i'_k]}. \qquad (24)$$

The first component on the right hand side of (23) can be bounded above using (24),

$$\sum_{i'_k=i_k+1}^{I_k}\sum_{i'_{k-1}=1}^{I_{k-1}}\cdots\sum_{i'_2=1}^{I_2} t_{i_1 i'_2\ldots i'_k}\le\sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]}. \qquad (25)$$

Employing a very similar approach for the remaining terms of (23) yields

$$\begin{aligned}
n_1^{*[i_1]} \le &\sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]}+\cdots+\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]}\\
&+\sum_{i'_j=\tilde{I}_j+1}^{I_j} n_j^{*[i'_j]}+\sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]}+\ldots\\
&+\sum_{i'_2=i_2+1}^{I_2} n_2^{*[i'_2]}+t_{i_1\ldots i_k}.
\end{aligned}$$

From the definition of $\tilde{I}_j$, we have $\sum_{i'_j=\tilde{I}_j+1}^{I_j} n_j^{*[i'_j]}=0$. So a lower bound for $t_{i_1\ldots i_k}$ is

$$\begin{aligned}
&n_1^{*[i_1]}-\sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]}-\cdots-\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]}-\\
&\sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]}-\cdots-\sum_{i'_2=i_2+1}^{I_2} n_2^{*[i'_2]}\le t_{i_1\ldots i_k}.
\end{aligned}$$

A similar procedure may be used to obtain bounds based on the remaining marginal sums for $\{n_2^{*[i_2]},\ldots,n_{j-1}^{*[i_{j-1}]},n_{j+1}^{*[i_{j+1}]},\ldots,n_k^{*[i_k]}\}$. There are $k-1$ of these bounds in total:

$$n_1^{*[i_1]}-\sum_{i'_2=i_2+1}^{I_2} n_2^{*[i'_2]}-\cdots-\sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]}-$$

$$
\begin{aligned}
&\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \cdots - \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]},\\
n_2^{*[i_2]} - &\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} -\\
&\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \cdots - \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]},\\
&\vdots\\
n_{j-1}^{*[i_{j-1}]} - &\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-2}=i_{j-2}+1}^{I_{j-2}} n_{j-2}^{*[i'_{j-2}]} -\\
&\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \cdots - \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]},\\
n_{j+1}^{*[i_{j+1}]} - &\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} -\\
&\sum_{i'_{j+2}=i_{j+2}+1}^{I_{j+2}} n_{j+2}^{*[i'_{j+2}]} - \cdots - \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]},\\
&\vdots\\
n_k^{*[i_k]} - &\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} -\\
&\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \cdots - \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]},
\end{aligned}
\tag{26}
$$

which can be written more concisely as

$$
n_z^{*[i_z]} - \sum_{m\neq j,z} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}
$$
$$
\text{for } z = 1,\ldots,j-1,j+1,\ldots,k.
$$

Next we show that the last one of these bounds (26), denoted by l, is the sharpest. Note that $M^* = \sum_{i'_k=i_k}^{I_k} n_k^{*[i'_k]}$. Since $M^* = \sum_{i'_z=1}^{I_z} n_z^{*[i'_z]}$, we know $M^* \geq \sum_{i'_z=i_z}^{I_z} n_z^{*[i'_z]}$ for all z, so $\sum_{i'_k=i_k}^{I_k} n_k^{*[i'_k]} \geq \sum_{i'_z=i_z}^{I_z} n_z^{*[i'_z]}$, and rearranging yields

$$
n_k^{*[i_k]} - \sum_{m\neq j,k} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]} \geq n_z^{*[i_z]} - \sum_{m\neq j,z} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}. \tag{27}
$$

So when $i_j = \tilde{I}_j$, the sharpest lower bound is $n_k^{*[i_k]} - \sum_{m\neq j,k} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}$. If we are currently sampling $i_k = \tilde{I}_k$, then the sharpest lower bound is given by $n_{k-1}^{*[i_{k-1}]} - \sum_{m\neq j,k-1} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}$. If $i_p = \tilde{I}_p$ for more than one p, the bound will be the same since summing over any dimension in which $i_p = \tilde{I}_p$, will not contribute anything to the summation $\sum_{m\neq j,k} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}$.

Next, we show that $l \geq l_f$, so when $i_j = \tilde{I}_j$, the Fréchet bound is not as strict as l. Recall

$$
\begin{aligned}
l = n_k^{*[i_k]} - &\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-1}=i_{i-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} -\\
&\sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \cdots - \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]}.
\end{aligned}
$$

Since $\sum_{i'_z=i_z}^{I_z} n_z^{*[i'_z]} \leq M^*$ and $n_j^{*[i_j]} \leq M^*$ we have

$$
\begin{aligned}
&\sum_{i'_1=i_1}^{I_1} n_1^{*[i'_1]} + \cdots + \sum_{i'_{j-1}=i_{j-1}}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} + n_j^{*[i_j]} +\\
&\sum_{i'_{j+1}=i_{j+1}}^{I_{j+1}} n_{j+1}^{*[i'_j+1]} + \cdots + \sum_{i'_{k-1}=i_{k-1}}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]} \leq (k-1)M^*.
\end{aligned}
$$

That means

$$
\begin{aligned}
&\sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} + \cdots + \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]}\\
&+ \sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_j+1]} + \cdots + \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]}\\
&\leq (k-1)M^* - n_1^{*[i_1]} - n_2^{*[i_2]} - \cdots - n_{k-1}^{*[i_{k-1}]},
\end{aligned}
$$

which leads to

$$
\begin{aligned}
&n_k^{*[i_k]} - \sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \cdots - \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]}\\
&- \sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_j+1]} - \cdots - \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]}\\
&\geq n_1^{*[i_1]} + n_2^{*[i_2]} + \cdots + n_k^{*[i_k]} - (k-1)M^*.
\end{aligned}
$$

Therefore $l \geq l_f$. Putting these together yields the bounds in (14).

Received 12 April 2019

REFERENCES

[1] Barvinok, A. I. (1994). A Polynomial Time Algorithm for Counting Integral Points in Polyhedra When the Dimension is Fixed. *Mathematics of Operations Research* **19** 769-779. MR1304623

[2] Chen, Y., Dinwoodie, I. H. and Sullivant, S. (2006). Sequential Importance Sampling for Multiway Tables. *The Annals of Statistics* **34** 523-545. MR2275252

[3] Chen, Y., Lin, C. H. and Sabatti, C. (2006). Volume Measures for Linkage Disequilibrium. *BMC Genetics* **7**.

[4] Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling from Conditional Distributions. *The Annals of Statistics* **26** 363-397. MR1608156

[5] Dobra, A. and Fienberg, S. E. (2008). The generalized shuttle algorithm. *Working Paper no. 83, Center for Statistics and the Social Sciences*. University of Washington. MR3632013

[6] Eisinger, R. D. and Chen, Y. (2017). Sampling for Conditional Inference on Contingency Tables. *Journal of Computational and Graphical Statistics* **26** 79-87. MR3610409

[7] Fienberg, S. E. (1999). Fréchet and Bonferroni bounds for multiway tables of counts with applications to disclosure limitation. In *Statistical Data Protection Proceedings. SDP'98* 115-129. Eurostat.

[8] Good, I. J. (1976). On the Application of Symmetric Dirichlet Distributions and their Mixtures to Contingency Tables. *The Annals of Statistics* **4** 1159-1189. MR0428568

[9] Kwerel, S. M. (1988). Fréchet Bounds. In *Encyclopedia of Statistical Sciences* (S. Kotz and N. L. Johnson, eds.) 202-209. Wiley, New York.

[10] Lazzeroni, L. C. and Lange, K. (1997). Markov Chains for Monte Carlo Tests of Genetic Equilibrium in Multidimensional Contingency Tables. *The Annals of Statistics* **25** 138-168. MR1429920

[11] Ophoff, R., Escamilla, M., Service, S., Spesny, M., Meshi, D., Poon, W., Molina, J., Fournier, E., Gallegos, A., Mathew, C., Neylan, Y., Batki, S., Roche, E., Ramirez, M., Silva, S., De Mille, M., Dong, P., Leon, P., Reus, V., Sandkuifl, L. and Freimer, N. (2002). Genomewide Linkage Disequilibrium Mapping of Severe Bipolar Disorder in a Population Isolate. *American Journal of Human Genetics* **71** 565-574.

[12] Pritchard, J. K. and Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data. *American Journal of Human Genetics* **69** 1-14.

[13] Rüschendorf, L. (1991). Bounds for Distributions with Multivariate Margins. In *Stochastic Order and Decision under Risk* (K. Mosler and M. Scarsini, eds.). *IMS Lecture Notes-Monograph Series* **19** 285-310.

[14] Sabatti, C. (2002). Measuring Dependence with Volume Tests. *The American Statistician* **50** 191-195. MR1963264

[15] Warmuth, W. (1988). Marginal-Fréchet-bounds for multidimensional distribution functions. *Statistics* **19** 283-294. MR0945385

Robert D. Eisinger
Department of Statistical Science
Duke University
Durham, NC 27708
USA
E-mail address: robert.eisinger@duke.edu

Xiao Su
Wells Fargo
Charlotte, NC 28282
USA
E-mail address: xiao.su@wellsfargo.com

Yuguo Chen
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA
E-mail address: yuguo@illinois.edu