SAMPLING FOR CONDITIONAL INFERENCE ON CONTINGENCY TABLES, MULTIGRAPHS, AND HIGH DIMENSIONAL TABLES

BY

ROBERT D. EISINGER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Yuguo Chen, Chair
Assistant Professor Steven Culpepper
Emeritus Professor John Marden
Professor Douglas Simpson

Abstract

We propose new sequential importance sampling methods for sampling contingency tables with fixed margins, loopless, undirected multigraphs, and high-dimensional tables. In each case, the proposals for the method are constructed by leveraging approximations to the total number of structures (tables, multigraphs, or high-dimensional tables), based on results in the literature. The methods generate structures that are very close to the target uniform distribution. Along with their importance weights, the data structures are used to approximate the null distribution of test statistics. In the case of contingency tables, we apply the methods to a number of applications and demonstrate an improvement over competing methods. For loopless, undirected multigraphs, we apply the method to ecological and security problems, and demonstrate excellent performance. In the case of high-dimensional tables, we apply the sequential importance sampling method to the analysis of multimarker linkage disequilibrium data and also demonstrate excellent performance.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Importance Sampling

This thesis will focus on applications of importance sampling to the analysis of useful data structures. In particular, algorithms will be proposed to sample two way contingency tables with fixed marginal sums, loopless, undirected, integer-valued networks with fixed degree sequence, and high dimensional tables with fixed one way margins from the uniform distribution. Analysis of these data structures has applications to combinatorial problems, ecology, and sociology. The space of tables, networks and high dimensional tables can be incredibly large, and exhaustive enumeration is often infeasible. Algorithms will be constructed that sample contingecy tables, networks, or high dimensional tables from a distribution that is close to uniform, and samples will be weighted to correct for the bias incurred by sampling.

First, we recall the basics of importance sampling. If we are interested in the quantity

$$\mu = \int h(x)\pi(x)dx = \int \left[h(x)\frac{\pi(x)}{g(x)}\right]g(x)dx, \tag{1.1}$$

we may draw $x_1, \ldots, x_N$ independent, identically distributed (iid) samples from a proposal distribution $g(x)$, and calculate the importance weights

$$w_i = \frac{\pi(x_i)}{g(x_i)}, \tag{1.2}$$

for $i = 1, \ldots, N$. We may estimate μ using

$$\hat{\mu} = \frac{w_1 h(x_1) + \ldots + w_M h(x_N)}{N}. \tag{1.3}$$

The proposal $g(x)$ should be easy to sample from and include the support of $\pi(x)$. If the normalizing constant is not known and we have $\pi(x) \propto l(x)$, we may replace $\pi(x)$ by $l(x)$ and estimate μ using

$$\tilde{\mu} = \frac{w_1 h(x_1) + \ldots + w_N h(x_N)}{w_1 + \ldots + w_N}, \tag{1.4}$$

where $w_i = l(x_i)/g(x_i)$. This results in a biased estimator that still converges to μ.

1.2 Sequential Importance Sampling

In high dimensional problems, it can be difficult to find a reasonable proposal distribution. In these situations, an effective approach is to build up the proposal distribution $g(x)$ sequentially using a procedure called sequential importance sampling (SIS). Write $x = (x_1, \ldots x_d)$, and

$$\begin{aligned} g(x) &= g_1(x_1) g_2(x_2|x_1) \ldots g_d(x_d|x_1, \ldots x_{d-1}) \\ \pi(x) &= \pi_1(x_1) \pi_2(x_2|x_1) \ldots \pi_d(x_d|x_1, \ldots x_{d-1}). \end{aligned} \tag{1.5}$$

Our weight, $w(x)$, is then

$$w(x) = \frac{\pi_1(x_1) \pi_2(x_2|x_1) \ldots \pi_d(x_d|x_1, \ldots x_{d-1})}{g_1(x_1) g_2(x_2|x_1) \ldots g_d(x_d|x_1, \ldots x_{d-1})}. \tag{1.6}$$

Define the current weight as

$$w_t(x) = \frac{\pi(x_1) \pi(x_2|x_1) \ldots \pi(x_t|x_1 \ldots x_{t-1})}{g(x_1) g(x_2|x_1) \ldots g(x_t|x_1 \ldots x_{t-1})}, \tag{1.7}$$

then

$$w_t = w_{t-1}(x) \frac{\pi(x_t|x_1, \ldots, x_{t-1})}{g(x_t|x_1, \ldots, x_{t-1})}. \tag{1.8}$$

The estimator of $\mu = E_\pi[f(x)]$ is

$$\hat{\mu} = \frac{\sum_{i=1}^N h(x_i)(\pi(x_i)/g(x_i))}{\sum_{i=1}^N (\pi(x_i)/g(x_i))}. \tag{1.9}$$

The choice of the proposal distribution is important as it determines the efficiency of the algorithm. Intuitive, *ad hoc*, proposals without theoretical justification can be effective in some cases,

but often the efficiency of the method can be improved by choosing a proposal distribution that is close to the target distribution. The strategy that will be taken to sample two way contingency tables with fixed margins, loopless, undirected multigraphs and high dimensional tables with fixed one way margins will be to use approximations to the total number of these data structures to guide the choice of the proposal distribution and to develop SIS methods.

The thesis is organized in the following way. Chapter 2 provides several effective SIS methods for sampling contingency tables uniformly from the set of all possible tables with fixed marginal sums, based on an approximation to the total number of tables of Good (1976), two asymptotic approximations of Greenhill and McKay (2008), and an asymptotic approximation of Bender (1974). An additional rapid cell by cell sampling method based on an *ad hoc* adaptation of Good (1976) is also proposed and examined. These methods are applied to a number of examples, including estimating the number of tables and analyzing ecological data. Chapter 3 provides an SIS method for sampling undirected, loopless multigraphs uniformly from the set of graphs with fixed degree sequence. An MCMC method based on the random walk moves of Diaconis and Gangolli (1995) is also proposed and evalutated. These methods are applied to ecological and security applications, including the analysis of a chimpanzee grooming network and the resilience of an airline network. Chapter 4 provides an SIS method for analyzing high dimensional tables with fixed marginal sums based on an adaptation of the approximation of Good (1976). This method is motivated by genetic data, and is used to analyze linkage disequilibrium in phase-known multimarker data for a population of individuals with bipolar data from the Central Valley in Costa Rica. Some concluding remarks and future directions are provided in Chapter 5.

Chapter 2

Sampling for Conditional Inference on Contingency Tables

2.1 Introduction

Hypothesis testing problems related to contingency tables have been of interest since Karl Pearson's foundational work in the area. A widely used contribution of his is the χ^2 test of independence. When independence is rejected, there is no information regarding what distribution generated the data. To help interpret the χ^2 statistic, Diaconis and Efron (1985) proposed the uniform distribution as an alternative to independence. In their conditional volume test, the observed table is considered to be a uniform draw from the set of tables with the same marginal sums, and the question is whether or not the χ^2 statistic of the observed table is unusual. The conditional volume test may also be applied in the case where a contingency table has some set of structural zeros (Chen, 2007). Contingency tables can also be used to represent weighted bipartite networks. Comparing the observed table (network) with random tables (networks) from the uniform distribution can be used to detect deviations from randomness in certain properties of the tables (networks).

We are concerned in this chapter with sampling tables uniformly from the set of all possible tables with fixed marginal sums. Based on these sampled tables, the distribution of a test statistic can be approximated. We are additionally interested in estimating the total number of tables with specified margins.

For counting the number of tables, a review is provided in Greenhill and McKay (2008). A breakthrough method was developed in the software LattE (Barvinok, 1994). It performs extremely well for counting the number of tables in small examples, but for larger tables the computation time is prohibitive.

This chapter uses material previously published in Eisinger and Chen (2016)

Several procedures exist for sampling tables. Diaconis and Gangolli (1995) proposed a Markov chain Monte Carlo (MCMC) method, in which tables with specified margins are generated using a random walk which converges to the uniform distribution. Other approaches based on these moves are possible using cycles and universal Gröebner bases (Diaconis and Sturmfels, 1998). These methods are effective in many cases, however, the samples can be highly correlated and it can be difficult to tell if the space has been adequately explored.

Importance sampling is another approach to sampling contingency tables. Tables are generated from a distribution that is close to uniform and each table is assigned a weight to correct for the bias incurred by sampling. Using this method, a reasonable approximation to the null distribution of any test statistic can be obtained and the total number of tables with specified margins can be estimated. Importance sampling for contingency tables was first considered in Chen et al. (2005). Sampling is done cell by cell, with the proposal distribution chosen as uniform over the possible values for that cell. After the first cell has been sampled, the procedure continues after updating the row and column margins. This proposal has performed well in cases where the table is small and moderately dense, but it is not effective when the table is large and sparse. Also, sampling each cell uniformly is an *ad hoc* proposal without any theoretical justification.

In this paper, we use asymptotic approximations of Greenhill and McKay (2008) and Bender (1974), and an approximation of Good (1976) to the total number of tables to justify the choice of new proposal distributions and design sequential importance sampling (SIS) methods. The SIS procedures developed here outperform other approaches. In particular, if the table is large and sparse, competing methods give inaccurate results and the proposed SIS methods perform well in comparison.

The rest of the chapter is organized as follows. Section 2.2 introduces the basic terminology of SIS in the context of sampling tables. Section 2.3 describes how the approximations are incorporated into the proposal to perform SIS. Section 2.4 provides applications, including counting the number of tables, the conditional volume test and testing ecological data. Section 2.5 describes practical details of sampling tables and Section 2.6 compares the proposed SIS methods to competing methods. Section 2.7 provides discussion and concluding remarks.

2.2 Sequential Importance Sampling

Let $\Sigma_{\mathbf{rc}}$ denote the set of all $m \times n$ contingency tables with row margins $\mathbf{r} = (r_1, \ldots, r_m)$, column margins $\mathbf{c} = (c_1, \ldots, c_n)$, $M = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$, and $|\Sigma_{\mathbf{rc}}|$ the total number of tables in the set $\Sigma_{\mathbf{rc}}$. Let $p(T) = 1/|\Sigma_{\mathbf{rc}}|$ be the uniform distribution over $\Sigma_{\mathbf{rc}}$. If we can simulate a table $T \in \Sigma_{\mathbf{rc}}$ from a proposal distribution $q(\cdot)$ that is easy to sample from and has the same support as $\Sigma_{\mathbf{rc}}$, then the total number of tables can be written as

$$|\Sigma_{\mathbf{rc}}| = \sum_{T \in \Sigma_{\mathbf{rc}}} \frac{1}{q(T)} q(T) = E_q \left[\frac{1}{q(T)} \right], \tag{2.1}$$

and $|\Sigma_{\mathbf{rc}}|$ can be estimated using $T_1, \ldots, T_N$, independent, identically distributed (iid) samples drawn from $q(T)$,

$$\widehat{|\Sigma_{\mathbf{rc}}|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(T_i)}. \tag{2.2}$$

If instead we are interested in estimating $\mu = E_p[f(T)]$, then the weighted average,

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(T_i)(p(T_i)/q(T_i))}{\sum_{i=1}^{N} (p(T_i)/q(T_i))} = \frac{\sum_{i=1}^{N} f(T_i)(1/q(T_i))}{\sum_{i=1}^{N} (1/q(T_i))}, \tag{2.3}$$

can be used to estimate μ. If we let $f(T) = \mathbb{1}_{(\chi^2 \text{ statistic of } T \geq S)}$, then (4.2) estimates the p-value of the observed χ^2 statistic S.

The efficiency of the above estimators can be quantified in several ways. The standard error of $\hat{\mu}$ can be estimated by repeatedly running the procedure or using the Δ-method

$$\text{se}(\hat{\mu}) \approx \sqrt{\frac{\text{var}_q(\frac{f(T)p(T)}{q(T)}) + \mu^2 \text{var}_q(\frac{p(T)}{q(T)}) - 2\mu \text{cov}_q(\frac{f(T)p(T)}{q(T)}, \frac{p(T)}{q(T)})}{N}}. \tag{2.4}$$

The *effective sample size* $\text{ESS} = N/(1 + \text{cv}^2)$ is an alternative way to quantify the efficiency of the method (Kong et al., 1994). Here, the *coefficient of variation* (cv) is defined as

$$\text{cv}^2 = \frac{\text{var}_q(p(T)/q(T))}{E_q^2(p(T)/q(T))}. \tag{2.5}$$

The cv^2 is the χ^2 distance between the proposal $q(\cdot)$ and the target $p(\cdot)$. The ESS roughly approximates how many iid samples are equivalent to the N weighted samples obtained using SIS.

In practical implementation, we can use the sample version of cv^2 to evaluate the performance of SIS. A small cv^2 is desired because it indicates that $p(\cdot)$ and $q(\cdot)$ are close to each other and the effective sample size is large.

To check whether we have enough samples to obtain a reliable estimate, we can increase the sample size N to see whether the estimate of cv^2 stabilizes. If the estimate of cv^2 becomes larger as N increases, that indicates that more samples are needed. Another way is to check whether the standard error decreases at the rate of $N^{-1/2}$ as N increases.

The choice of the proposal $q(\cdot)$ is fundamental to an efficient importance sampling procedure. Since this is a high dimensional problem, the strategy that will be employed here is to decompose the proposal into lower-dimensional components. The proposal for an entire table is constructed sequentially component by component conditional on the realization of the previous components. A theoretical framework for SIS is given by Liu and Chen (1998).

2.3 Sampling Contingency Tables

The SIS procedure of Chen et al. (2005) sampled each cell of a table sequentially based on a uniform proposal distribution on the values each entry can take. That is, the first cell entry α_{11} has restrictions $\max(0, r_1+c_1-M) \leq \alpha_{11} \leq \min(r_1, c_1)$, and so the first entry is sampled uniformly from the integers between those two values. After sampling this entry, the remaining cells in the same column are sampled in a similar fashion after updating the row and column sums. The next column is then sampled and the procedure continues until a completed table is obtained. This procedure will be denoted SIS-Uniform. Although this procedure is easily used, there is no theoretical justification for why each cell should be sampled from the uniform distribution. We are proposing a new SIS technique which samples the contingency table column by column and uses approximations to guide the sampling. A similar strategy using a different asymptotic approximation was employed by Zhang and Chen (2013) for symmetric 0-1 tables.

If we denote the columns of T by $t_1, \ldots, t_n$, then the probability of sampling a table T using a proposal $q(\cdot)$ can be written as

$$q(T = (t_1, \ldots, t_n)) = q(t_1)q(t_2|t_1) \ldots q(t_n|t_{n-1}, \ldots, t_1).$$

We begin by sampling the first column of the table, t_1, conditional on $\mathbf{r}$ and $\mathbf{c}$. After t_1 has been sampled, the row and column margins are updated, and we sample the first column of the remaining $m \times (n-1)$ subtable. The row margins are updated by subtracting the value sampled in the corresponding row of t_1 and the column margins are updated by removing the first element of $\mathbf{c}$. Denote the configuration of the jth column by $t_j = (\alpha_{1j}, \ldots, \alpha_{mj})$, and denote by $\mathbf{r}^{(j+1)}$ and $\mathbf{c}^{(j+1)}$ the updated row and column margins after j columns have been sampled, i.e.,

$$\mathbf{r}^{(j+1)} = (r_1^{(j)} - \alpha_{1j}, \ldots, r_m^{(j)} - \alpha_{mj}),$$
$$\mathbf{c}^{(j+1)} = (c_{j+1}, \ldots, c_n).$$

Note that $\mathbf{r}^{(1)} = \mathbf{r}$ and $\mathbf{c}^{(1)} = \mathbf{c}$. The procedure is repeated until all of the columns have been sampled and a completed table is obtained.

We start by writing the true marginal distribution of t_1 under the uniform distribution over $\Sigma_{\mathbf{rc}}$. For a given configuration of the first column $t_1 = (\alpha_{11}, \ldots, \alpha_{m1})$, the true marginal distribution of t_1 is

$$p(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) = \frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{|\Sigma_{\mathbf{rc}}|}.$$

This expression cannot be calculated directly, but asymptotic formulae for $|\Sigma_{\mathbf{rc}}|$ exist under various conditions. O'Neil (1969), Everett and Stein (1970), and Békéssy et al. (1972) all developed asymptotic approximations to $|\Sigma_{\mathbf{rc}}|$ for large sparse tables. However, these approximations can give inaccurate results for small and dense tables. We will focus on two asymptotic approximations of Greenhill and McKay (2008), who gave the number of tables with specified row and column margins under a strong sparsity condition, and an approximation of Good (1976), who provided an approximation to the number of tables with given margins. These approximations work reasonably well in most cases, and will be used to develop proposal distributions for sampling contingency tables. The SIS procedure does not require that the asymptotic approximations be perfect, as they will be used as a guide for sampling and the importance weights will correct for the bias. We also examine an additional approximation of Bender (1974) that works well in sparse cases, but struggles when the table is moderately dense or small.

The first approximation is given in Good (1976) without proof.

Good's Approximation. *(Good, 1976)*

$$|\Sigma_{\mathbf{rc}}| \approx \Delta_{\mathbf{rc}}^{G} \equiv \frac{\prod_{i=1}^{m}\binom{n+r_i-1}{r_i}\prod_{j=1}^{n}\binom{m+c_j-1}{c_j}}{\binom{M+mn-1}{M}}. \tag{2.6}$$

The interpretation of this approximation is informative. It is the product of the number of ways to distribute the row margins in the columns and the column margins in the rows, divided by the number of $m \times n$ tables with sum M (Good, 1976). An additional approximation is taken from Greenhill and McKay (2008). Define $r = \max_{1\le i\le m} r_i$ and $c = \max_{1\le j\le n} c_j$.

Theorem 2.3.1. (Greenhill and McKay 2008) For given $\mathbf{r}$ and $\mathbf{c}$, suppose $1 \le rc = o(M^{2/3})$. Also, define

$$\hat{\mu}_k = \frac{mn}{M(mn+M)}\sum_{i=1}^{m}(r_i - M/m)^k,$$
$$\hat{v}_k = \frac{mn}{M(mn+M)}\sum_{j=1}^{n}(c_j - M/n)^k.$$

Then

$$|\Sigma_{\mathbf{rc}}| = \frac{\prod_{i=1}^{m}\binom{n+r_i-1}{r_i}\prod_{j=1}^{n}\binom{m+c_j-1}{c_j}}{\binom{M+mn-1}{M}}\exp\left\{\boldsymbol{\alpha}(\mathbf{r},\mathbf{c}) + O\left(\frac{r^3c^3}{M^2}\right)\right\} \tag{2.7}$$

as $m, n \to \infty$ and $M \to \infty$, where

$$\boldsymbol{\alpha}(\mathbf{r},\mathbf{c}) = (1-\hat{\mu}_2)(1-\hat{v}_2)\Big(\frac{1}{2}+\frac{3-\hat{\mu}_2\hat{v}_2}{4M}\Big) - \frac{(1-\hat{\mu}_2)(3+\hat{\mu}_2-2\hat{\mu}_2\hat{v}_2)}{4n} - \frac{(1-\hat{v}_2)(3+\hat{v}_2-2\hat{\mu}_2\hat{v}_2)}{4m} + \frac{(1-3\hat{\mu}_2^2+2\hat{\mu}_3)(1-3\hat{v}_2^2+2\hat{v}_3)}{12M}.$$

Define $\Delta_{\mathbf{rc}}^{\mathrm{GM1}}$ to be the approximation (2.7) neglecting the term $O(r^3c^3/M^2)$. Under the conditions of Theorem 2.3.1, Greenhill and McKay (2008) proved that the set of all tables with an entry greater than three will be a "vanishingly small" proportion of $\Sigma_{\mathbf{rc}}$. However, the asymptotic approximation appears to work reasonably well even when there are entries significantly larger than three in the table.

An added condition yields an additional asymptotic approximation.

Theorem 2.3.2. (Greenhill and McKay 2008) Under the conditions of Theorem 2 and with the additional condition that $(1+\hat{\mu}_2)(1+\hat{v}_2) = O(M^{1/3})$, we have

$$|\Sigma_{\mathbf{rc}}| = \frac{\prod_{i=1}^{m}\binom{n+r_i-1}{r_i}\prod_{j=1}^{n}\binom{m+c_j-1}{c_j}}{\binom{M+mn-1}{M}} \exp\left\{\frac{1}{2}(1-\hat{\mu}_2)(1-\hat{v}_2) + O\left(\frac{rc}{M^{2/3}}\right)\right\}. \tag{2.8}$$

Define $\Delta_{\mathbf{rc}}^{\mathrm{GM2}}$ to be the approximation (2.9) neglecting the term $O(rc/M^{2/3})$. Under the uniform distribution over $\Sigma_{\mathbf{rc}}$, three proposal distributions for the first column t_1 are obtained using each of the approximations, $\Delta_{\mathbf{rc}}^{\mathrm{G}}$, $\Delta_{\mathbf{rc}}^{\mathrm{GM1}}$, and $\Delta_{\mathbf{rc}}^{\mathrm{GM2}}$

An additional asymptotic approximation is due to Bender (1974). This approximation performs well in cases where the table is sparse, but is not at all effective when the table is not extremely sparse. It is obtained by specializing Theorem 1 of Bender (1974). The derivation is provided in the appendix.

Theorem 2.3.3. (Bender, 1974) For given $\mathbf{r}$ and $\mathbf{c}$ and assuming the entries are bounded above by a constant d, then as $M \to \infty$ we have

$$|\Sigma_{\mathbf{rc}}| \sim \Delta_{\mathbf{rc}}^{\mathrm{B}} \equiv \frac{M!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{n} c_j!} \exp\left\{\frac{(\sum_{i=1}^{m} r_i(r_i-1))(\sum_{j=1}^{n} c_j(c_j-1))}{2M^2}\right\}. \tag{2.9}$$

Let $(\alpha_{11}, \ldots, \alpha_{m1})$ denote the entries of the first column and note $\sum_{i=1}^{m} \alpha_{i1} = c_1$. Proposals used to sample the first column t_1 are shown below. The proofs of all proposals are in the appendix.

The proposal in Proposal 2.1 is constructed using $\Delta_{\mathbf{rc}}^{\mathrm{G}}$ and is denoted SIS-G.

Proposal 2.1. $q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{G}}}{\Delta_{\mathbf{rc}}^{\mathrm{G}}} \propto \prod_{i=1}^{m}\binom{n+r_i-\alpha_{i1}-2}{r_i-\alpha_{i1}}$.

The proposal in Proposal 2.2 is constructed using $\Delta_{\mathbf{rc}}^{\mathrm{GM1}}$ and is denoted SIS-GM1.

Proposal 2.2. Define

$$\hat{\mu}_k^{(2)} = \frac{m(n-1)}{(M-c_1)(m(n-1)+M-c_1)} \sum_{i=1}^{m} (r_i - \alpha_{i1} - \frac{M-c_1}{m})^k,$$
$$\hat{v}_k^{(2)} = \frac{m(n-1)}{(M-c_1)(m(n-1)+M-c_1)} \sum_{j=2}^{n} (c_j - \frac{M-c_1}{n-1})^k,$$

and $\boldsymbol{\alpha}(\cdot,\cdot)$ as in Theorem 2. Then,

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta^{\mathrm{GM1}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{GM1}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \exp\{\boldsymbol{\alpha}(\mathbf{r}^{(2)}, \mathbf{c}^{(2)})\}.$$

The proposal in Proposal 2.3 is constructed using $\Delta^{\mathrm{GM2}}_{\mathbf{rc}}$ and is denoted SIS-GM2.

Proposal 2.3.

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta^{\mathrm{GM2}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{GM2}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \exp\left\{\frac{1}{2}(1-\hat{\mu}_2^{(2)})(1-\hat{v}_2^{(2)})\right\}.$$

The proposal in Proposal 2.4 is constructed using $\Delta^{\mathrm{B}}_{\mathbf{rc}}$ and is denoted SIS-B.

Proposal 2.4.

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{(M-c_1)!}{\prod_{i=1}^{m}(r_i - \alpha_{i1})! \prod_{j=2}^{n} c_j!} \times \exp\left\{\frac{(\sum_{i=1}^{m}(r_i - \alpha_{i1})(r_i - \alpha_{i1} - 1))(\sum_{j=2}^{n} c_j(c_j-1))}{2(M-c_1)^2}\right\}.$$

Although $q(t_1)$ in the above two proposals may be sampled directly, this is not feasible for larger tables. In these cases, it is more convenient to sample $q(t_1)$ using the following rejection method.

1. Generate a possible configuration of the first column $\mathbf{x} = (x_1, \ldots, x_m)$ from $g(\mathbf{x})$, where $g(\mathbf{x})$ is the uniform distribution over all possible configurations of the first column. This can be done using the procedure described by Holmes and Jones (1996).
2. Generate $u \sim$ Unif[0,1].
3. Calculate the ratio $q(\mathbf{x})/(cg(\mathbf{x}))$, where $q(\mathbf{x})$ is the proposal of SIS and c is a constant chosen so that $q(\mathbf{x}) \leq cg(\mathbf{x})$.

4. Accept $\mathbf{x}$ if $u \leq q(\mathbf{x})/(cg(\mathbf{x}))$. Otherwise, reject $\mathbf{x}$.

Since $\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}$ will be obtained for every possible configuration of the first column when the normalizing constant for $q(t_1)$ is calculated, it is straightforward to calculate both the number of configurations of the first column and the maximum value of $\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}$ over these configurations. Using these two quantities, a value c such that $q(\mathbf{x}) \leq cg(\mathbf{x})$ is easy to find.

SIS-GM1 and SIS-GM2 both use the approximation of Greenhill and McKay (2008) to design the proposal distribution. The approximation error is $||\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|/|\Sigma_{\mathbf{rc}}| - \Delta^{\mathrm{GM}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}/\Delta^{\mathrm{GM}}_{\mathbf{rc}}|$. In the following theorem, whose proof is in the appendix, the approximation error of SIS-GM1 is quantified. The conclusion for SIS-GM2 is similar.

Theorem 2.3.4. Suppose $1 \leq rc = o(M^{2/3})$. Then

$$\left| \frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{|\Sigma_{\mathbf{rc}}|} - \frac{\Delta^{\mathrm{GM1}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{GM1}}_{\mathbf{rc}}} \right| = O\left(\frac{r^3c^3}{M^2}\right)$$

as $M, n, m \to \infty$.

2.3.1 Cell by cell sampling

The approximations in the last section may be used to derive *ad hoc* cell by cell sampling procedures. Although they are usually not as effective as column by column sampling in terms of cv^2 and standard error, cell by cell sampling is much faster to run because it avoids the calculation of the normalizing constant for the proposal distribution of each column. This makes cell by cell sampling methods a useful method in cases where sampling by column is not feasible.

We consider Good's approximation, which has a combinatorial interpretation that can be leveraged to allow for an *ad hoc* cell by cell sampling procedure. After sampling the first cell with an entry α_{11}, the updated row and column margins are $\mathbf{r}^{*(2)} = (r_1 - \alpha_{11}, r_2, \ldots, r_m)$ and $\mathbf{c}^{*(2)} = (c_1 - \alpha_{11}, c_2, \ldots, c_n)$, respectively. Denote by $|\Sigma_{\mathbf{r}^{*(2)}\mathbf{c}^{*(2)}}|$ the number of tables with these margins and the first entry forced to be zero. Then we can approximate $|\Sigma_{\mathbf{r}^{*(2)}\mathbf{c}^{*(2)}}|$ using a natural extension of Good (1976):

$$|\Sigma_{\mathbf{r}^{*(2)}\mathbf{c}^{*(2)}}| \approx \Delta^{\mathrm{G}^*}_{\mathbf{r}^{*(2)}\mathbf{c}^{*(2)}} = \frac{\binom{n+r_1-\alpha_{11}-2}{r_1-\alpha_{11}}\prod_{i=2}^{m}\binom{n+r_i-1}{r_i}\binom{m+c_1-\alpha_{11}-2}{c_1-\alpha_{11}}\prod_{j=2}^{n}\binom{m+c_j-1}{c_j}}{\binom{M-\alpha_{11}+mn-2}{M-\alpha_{11}}}. \tag{2.10}$$

Here, the remaining first row sum can only be allocated to $n-1$ columns, the remaining first column sum can only be allocated to $m-1$ rows and the remaining table sum can only be allocated to the $mn-1$ elements of the table excluding the first cell.

Using (2.10), the first entry a_{11} may be sampled using the proposal distribution

$$q(a_{11}=\alpha_{11}) \propto \frac{\binom{n+r_1-\alpha_{11}-2}{r_1-\alpha_{11}}\binom{m+c_1-\alpha_{11}-2}{c_1-\alpha_{11}}}{\binom{M-\alpha_{11}+mn-2}{M-\alpha_{11}}}, \tag{2.11}$$

where $\max(0, r_1+c_1-M) \leq \alpha_{11} \leq \min(r_1, c_1)$.

After sampling the first entry, the row and column sums are updated and the next cell in the same column is sampled in a similar way.

If we have already sampled α_{i1} where $1 \leq i \leq k-1$, the proposal for entry a_{k1} is given in Proposal 2.5 and is denoted SIS-G*.

Proposal 2.5.

$$q(a_{k1}=\alpha_{k1}) \propto \frac{\binom{n+r_k-\alpha_{k1}-2}{r_k-\alpha_{k1}}\binom{m-k+c_1-\sum_{i=1}^{k}\alpha_{i1}-1}{c_1-\sum_{i=1}^{k}\alpha_{i1}}}{\binom{M-\sum_{i=1}^{k}\alpha_{i1}+mn-k-1}{M-\sum_{i=1}^{k}\alpha_{i1}}}, \tag{2.12}$$

where $\max(0, c_1-\sum_{i=1}^{k-1}\alpha_{i1}-\sum_{i=k+1}^{m}r_i) \leq \alpha_{k1} \leq \min(r_k, c_1-\sum_{i=1}^{k-1}\alpha_{i1})$, the same bounds as uniform sampling (Chen et al., 2005).

Since after sampling the first column, a smaller $m \times (n-1)$ subtable remains, the procedure continues until a completed table is obtained.

Examining Proposal 2.5 in a specific case provides support for using SIS-G* and also illustrates the disadvantages of using the intuitively derived sampling method based on the uniform proposal

distribution, SIS-Uniform. We consider sampling the first cell of Table 2.1 when the columns and rows are both arranged in increasing order. Clearly $0 \leq \alpha_{11} \leq 10$, and since this is a relatively small example, the true distribution of α_{11} can be calculated explicitly using LattE (Barvinok, 1994). This is shown in Figure 2.1 along with the probability density for the proposal distribution SIS-G* (in red). Figure 2.1 shows that SIS-G* is extremely close to the true target distribution for the first cell. Sampling the first cell uniformly between 0 and 10 would result in a less efficient sampling procedure as this proposal is far from the true target distribution for the first cell.

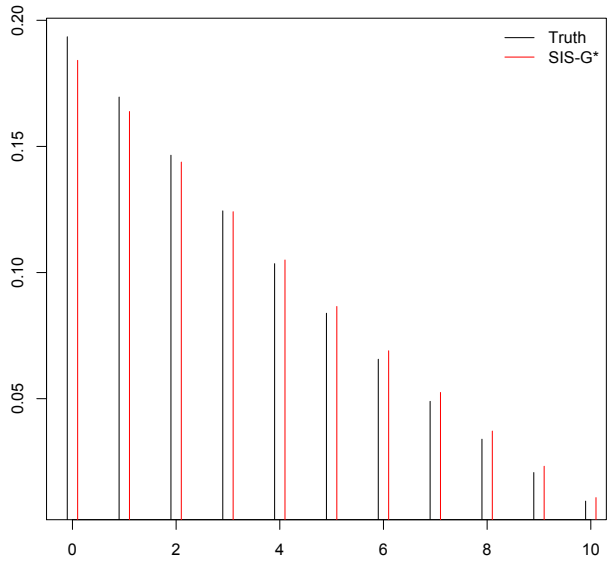


Figure 2.1: Probability densities for α_{11} of Table 2.1

2.4 Applications and Simulations

2.4.1 Estimating the number of tables

The number of contingency tables with fixed row and column margins is difficult to calculate. An exhaustive search may take a prohibitively long time, and asymptotic formulae may not be very accurate. SIS allows us to estimate $|\Sigma_{\mathbf{rc}}|$ based on iid samples from our proposal distribution. We estimate the number of tables in a few examples and compare the performance of SIS-G, SIS-GM1, SIS-GM2 and SIS-Uniform. We also examine the proposal for sparse examples, SIS-B, and the cell

by cell sampling method SIS-G*.

Table 2.1: 5×3 table (Diaconis and Gangolli, 1995)

50	5	7	62
2	30	7	39
3	4	6	13
5	3	3	11
5	3	2	10
65	45	25	135

The simulation results, presented in Table 2.4, Table 2.5, Table 2.6, Table 2.7, Table 2.8, Table 2.9 and Table 2.10, are based on $1,000$ importance samples for each method. Computation was performed on a MacBook Pro with a 2.2 GHz Intel Core i7 processor. Coding was done in C. The number following the $\pm$ sign denotes the standard error.

The first example is estimating the number of 8×8 tables with all margins equal to 6. The second example is the table given in Diaconis and Gangolli (1995) (Table 2.1). The third example is the classification of hair color and eye color in Table 2.2 (Snee, 1974). The fourth example is the birth month and death month for 82 descendants of Queen Victoria (Andrews and Herzberg, 1985) (Table 2.3). We also examine a 30×30 table with all margins equal to 3.

We additionally examine several large and sparse tables. The first two examples of large and sparse tables are 50×50 and 75×75 with all margins equal to 2. The third example is 75×75 with both the row and column margins equal to $(5, 2, \ldots, 2)$. The final example is a 100×100 table with both row and column marginal equal to $(5, 1, \ldots, 1)$.

There are 1.146×10^{20} 8×8 tables with all margins equal to 6 (Good and Crook, 1977). The true number of tables with the same margins as Table 2.1 and Table 2.2 are $239,382,172$ and

Table 2.2: Cross tabulation of hair color and eye color (Snee, 1974)

Hair Color	Eye Color				
	Brown	Blue	Hazel	Green	Total
Black	68	20	15	5	108
Brown	119	84	54	29	286
Red	26	17	14	14	71
Blond	7	94	10	16	127
Total	220	215	93	64	592

Table 2.3: Cross tabulation of birth month and death month (Andrews and Herzberg, 1985)

Birth month	Death month												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
Mar	1	0	0	0	2	1	0	0	0	0	0	1	5
Apr	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
Jun	2	0	0	0	1	0	0	0	0	0	0	0	3
Jul	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sep	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

$1,225,914,276,768,514$, respectively (Diaconis and Gangolli, 1995). The true number of tables with the same margins as Table 2.3 is unknown. The number of tables with the same margins as the 30×30 table is 2.22931×10^{92} (Canfield and McKay, 2010). The true number of tables for the large and sparse examples in Table 2.5 are all unknown.

The free software LattE gives the true number of tables in 0.19 seconds for Table 2.2. However, this method takes a prohibitively long time for larger tables and was not able to run in a reasonable amount of time for larger and sparser examples.

It appears that for all tables, the three new proposals SIS-G, SIS-GM1, and SIS-GM2 are more accurate than SIS-Uniform based on cv^2, indicating that all three new methods are sampling from a distribution that is very close to uniform. The three new proposals gave reasonable approximations to the true number of tables where it is known.

For the 8×8 table, Table 2.3, the 30×30 table and all of the large sparse tables, SIS-Uniform severely underestimates the number of tables and the extremely large cv^2 values indicate that the proposal distribution is far from uniform. This is especially pronounced for the 30×30 table and the results in Table 2.5 which all give extremely inaccurate results. Although SIS-Uniform is faster to run, it fails when the table becomes large and sparse. In these situations, SIS-G, SIS-GM1, and SIS-GM2 give accurate results and outperform SIS-Uniform.

In addition to the approximations SIS-G, SIS-GM1, and SIS-GM2, a sequential sampling proce-

Table 2.4: Performance comparison of methods for estimating the number of tables

Method	Estimated number of tables	cv^2	Time (s)
	8×8 table with all margins = 6		
SIS-G	$(1.1439 \pm 0.0039) \times 10^{20}$	0.0117	2.3
SIS-GM1	$(1.1449 \pm 0.0066) \times 10^{20}$	0.0331	5.9
SIS-GM2	$(1.1485 \pm 0.0057) \times 10^{20}$	0.0245	4.2
SIS-Uniform	$(6.2061 \pm 4.4749) \times 10^{18}$	519.9110	0.003
	5×3 table (Table 2.1)		
SIS-G	$(2.3989 \pm 0.0045) \times 10^{8}$	0.0035	0.6
SIS-GM1	$(2.3843 \pm 0.0107) \times 10^{8}$	0.0200	0.8
SIS-GM2	$(2.3915 \pm 0.0129) \times 10^{8}$	0.0291	0.7
SIS-Uniform	$(2.4477 \pm 0.1457) \times 10^{8}$	3.5421	0.0007
	Hair color vs. eye color (Table 2.2)		
SIS-G	$(1.2314 \pm 0.0059) \times 10^{15}$	0.0227	209.7
SIS-GM1	$(1.2170 \pm 0.0096) \times 10^{15}$	0.0616	259.6
SIS-GM2	$(1.2296 \pm 0.0092) \times 10^{15}$	0.0565	238.2
SIS-Uniform	$(1.1758 \pm 0.0797) \times 10^{15}$	4.5953	0.0008
	Birth month vs. death month (Table 2.3)		
SIS-G	$(6.3027 \pm 0.0206) \times 10^{39}$	0.0107	35.4
SIS-GM1	$(6.2847 \pm 0.0379) \times 10^{39}$	0.0363	97.6
SIS-GM2	$(6.2626 \pm 0.0383) \times 10^{39}$	0.0373	67.9
SIS-Uniform	$(1.1889 \pm 0.9124) \times 10^{32}$	588.9488	0.009
	30×30 table with all margins = 3		
SIS-G	$(2.2373 \pm 0.0093) \times 10^{92}$	0.0174	76.3
SIS-GM1	$(2.2294 \pm 0.0047) \times 10^{92}$	0.0045	198.7
SIS-GM2	$(2.2316 \pm 0.0082) \times 10^{92}$	0.0133	139.2
SIS-Uniform	$(7.9734 \pm 5.5971) \times 10^{51}$	492.7675	0.06

Table 2.5: Performance comparison of methods for estimating the number of large tables

Method	Estimated number of tables	cv^2	Time (s)
	50×50 table with all margins $= 2$		
SIS-G	$(1.2179 \pm 0.0042) \times 10^{128}$	0.0117	48.9
SIS-GM1	$(1.2213 \pm 0.0010) \times 10^{128}$	0.0007	123.7
SIS-GM2	$(1.2212 \pm 0.0030) \times 10^{128}$	0.0059	89.4
SIS-Uniform	$(1.5942 \pm 1.1675) \times 10^{66}$	536.2941	0.21
	75×75 table with all margins $= 2$		
SIS-G	$(6.6499 \pm 0.0201) \times 10^{217}$	0.0091	266.9
SIS-GM1	$(6.6191 \pm 0.0052) \times 10^{217}$	0.0006	633.3
SIS-GM2	$(6.6269 \pm 0.0133) \times 10^{217}$	0.0040	350.6
SIS-Uniform	$(9.3260 \pm 6.7356) \times 10^{94}$	521.6343	0.7
	75×75 table with both margins $= (5, 2, \ldots, 2)$		
SIS-G	$(7.1347 \pm 0.0250) \times 10^{220}$	0.0123	316.8
SIS-GM1	$(7.1438 \pm 0.0061) \times 10^{220}$	0.0007	763.0
SIS-GM2	$(7.1334 \pm 0.0117) \times 10^{220}$	0.0027	541.3
SIS-Uniform	$(9.7180 \pm 9.7130) \times 10^{99}$	998.9555	0.7
	100×100 table with both margins $= (5, 1, \ldots, 1)$		
SIS-G	$(7.2638 \pm 0.0225) \times 10^{161}$	0.0096	227.3
SIS-GM1	$(7.2939 \pm 0.0005) \times 10^{161}$	0.000004	588.0
SIS-GM2	$(7.2939 \pm 0.0012) \times 10^{161}$	0.000028	406.5
SIS-Uniform	$(3.0424 \pm 2.7841) \times 10^{75}$	837.4356	1.5

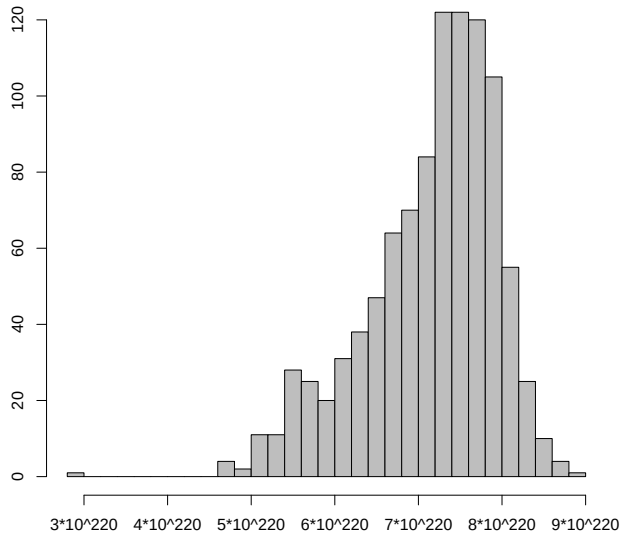


Figure 2.2: Histogram of $1,000$ importance weights for the 75×75 table with both row and column margins $= (5, 2, \ldots, 2)$.

dure may be developed based on the approximation of Bender (1974). It is not as widely applicable as the other three SIS procedures and may only be used effectively in cases where the table is extremely large and sparse. For example, the approximation in Theorem 2.3.3 fails completely for Tables 2.1 and 2.2 and SIS-B will give inaccurate results. In a moderately dense table such as a 4×4 table with margins $\mathbf{r} = \{12, 11, 19, 8\}$ and $\mathbf{c} = \{7, 11, 21, 11\}$, SIS-B estimates the number of tables as $(1.6118 \pm 0.6432) \times 10^3$ with $\text{cv}^2 = 164.2197$. The true value calculated by LattE is 6,846,954, so SIS-B is a severe underestimate in this case (Barvinok, 1994). Other methods do not struggle at all with this small, dense table.

However, in cases where the table is large and sparse, SIS-B can be very effective. We examine the performance of SIS-B on the same sparse cases that were used to test SIS-G, SIS-GM1 and SIS-GM2. Results, based on $1,000$ importance samples, are shown in Table 2.6. They indicate an improvement in performance relative to SIS-G, SIS-GM1 and SIS-GM2 in extremely sparse cases.

The efficiency of importance sampling methods are compared by running each method for the same amount of time as $1,000$ iterations of SIS-G and then taking the ratio of the standard errors. For large and sparse tables, the best performance is given by SIS-G, SIS-GM, and SIS-B.

Table 2.6: Performance of SIS-B for estimating the number of tables

Method	Estimated number of tables	cv^2	Time (s)
	8×8 table with all margins $= 6$		
SIS-B	$(1.1407 \pm 0.0276) \times 10^{20}$	0.5843	4.3
	30×30 table with all margins $= 3$		
SIS-B	$(2.2294 \pm 0.0015) \times 10^{92}$	0.0004	137.4
	50×50 table with all margins $= 2$		
SIS-B	$(1.2197 \pm 0.0010) \times 10^{128}$	0.0007	89.9
	75×75 table with all margins $= 2$		
SIS-B	$6.6222 \pm 0.0045) \times 10^{217}$	0.0005	440.2
	75×75 table with both margins $= (5, 2, \ldots, 2)$		
SIS-B	$(7.1449 \pm 0.0046) \times 10^{220}$	0.0004	522.6
	75×75 table with both margins $= (2, \ldots, 2, 5)$		
SIS-B	$(7.1525 \pm 0.0057) \times 10^{220}$	0.0006	415.1
	100×100 table with both margins $= (5, 1, \ldots, 1)$		
SIS-B	$(7.3084 \pm 0.0103) \times 10^{161}$	0.0020	337.7
	100×100 table with both margins $= (1, \ldots, 1, 5)$		
SIS-B	$(7.2902 \pm 0.0031) \times 10^{161}$	0.0002	35.1

Even adjusting for computation time and running SIS-Uniform for a long period does not improve performance, as it consistently underestimates the number of tables in large, sparse cases. However, for small, dense tables (Table 2.2), SIS-Uniform outperforms SIS-G and SIS-GM.

We also tested the cell by cell sampling method SIS-G* on a number of examples. We examine a long 3×49 table and a dense 5×5 table with all margins equal to 50. We also examine a small 5×5 table with rough margins. The true number of tables with the same margins as the 3×49 table is given in Canfield and McKay (2010) as 1.0110×10^{68} and the true number of tables for the 5×5 table with rough margins is about 2.3115×10^{17}, calculated in 13.5 seconds using LattE, and the true number of tables for the 5×5 table with all margins equal to 50 is 7.5063×10^{20}, calculated in 38.2 seconds using LattE.. These results based on $1,000$ samples are in Table 2.7 and indicate that SIS-G* is faster to run than SIS-G and SIS-GM, but also moderately less efficient in

terms of standard error and cv^2.

Estimation of the number of tables in the examples examined already for SIS-G, SIS-GM1, and SIS-GM2 are presented in Table 2.8.

When the table sum is large and the computation is time-intensive, cell sampling can be more efficient than column sampling. For example, using SIS-G* is about 1.5 times as efficient for Table 2.2, which takes over 200 seconds using SIS-G.

To further test performance, SIS-G and SIS-G* are compared on three dense tables. The first example is 4×4 with all margins equal to 8. The second example is 8×8 with all margins equal to 15. The final example has rough margins, with $\mathbf{r} = \{154, 5, 78, 79, 82\}$ and $\mathbf{c} = \{101, 182, 22, 86, 7\}$. Results are presented in Table 2.9. For the 4×4 table with all margins = 8, SIS-G is about 7 times as efficient as SIS-G*, for the 15×15 table SIS-G* is about 3 times as efficient as SIS-G, and for the final table with rough margins, SIS-G* is about thirteen times as efficient as SIS-G. In all of the tables examined, SIS-G* outperforms SIS-Uniform.

We finally examine three extremely dense tables, a 12×12 table with all margins equal to 90, a 10×10 table with all margins equal to 200 and a 5×7 table with margins equal to $\mathbf{r} = \{108, 98, 92, 35, 34\}$ and $\mathbf{c} = \{76, 69, 61, 47, 46, 42, 26\}$. These results are presented in Table 2.10 and indicate effective performance when the table is extremely dense, along with large cv^2 values and underestimates of the number of tables using SIS-Uniform.

Both SIS-G* and SIS-Uniform sample the tables cell by cell and they are both very fast to run. Between these two algorithms, simulations show that SIS-G* always gives smaller standard errors and cv^2 although it takes a little longer to run than SIS-Uniform. Simulation results indicate that SIS-G* still works even for extremely dense tables. For example, SIS-G* works well for the tables described in Table 2.10, while SIS-Uniform severely underestimates the number of tables. This example is challenging for the column by column sampling methods because the tables are extremely dense.

2.4.2 Conditional volume test

Volume tests were developed for regression problems by Hotelling (1939) and were further developed by Diaconis and Efron (1985) to help interpret the χ^2 statistic in the test of independence for two way tables. The conditional volume test of Diaconis and Efron (1985) tests whether or not the

Table 2.7: Performance comparison of methods for estimating the number of tables

Method	Estimated number of tables	cv^2	Time (s)
	3×49 table with all margins $= 98, 6$		
SIS-G*	$(1.0038 \pm 0.0437) \times 10^{68}$	1.8944	1.5
SIS-G	$(1.0079 \pm 0.0136) \times 10^{68}$	0.1831	2.3
SIS-Uniform	$(1.9454 \pm 1.5289) \times 10^{67}$	617.6127	0.01
	5×5 table with all margins $= 50$		
SIS-G*	$(7.2251 \pm 0.1993) \times 10^{20}$	1.7612	1.1
SIS-G	$(7.4822 \pm 0.0198) \times 10^{20}$	0.0070	690.3
SIS-Uniform	$(7.1742 \pm 1.6825) \times 10^{20}$	54.9980	0.0039
	$\mathbf{r} = \{154, 5, 78, 79, 82\}$, $\mathbf{c} = \{101, 182, 22, 86, 7\}$		
SIS-G*	$(2.2057 \pm 0.1067) \times 10^{17}$	2.3396	2.2
SIS-G	$(2.3105 \pm 0.0014) \times 10^{17}$	0.0004	2041.1
SIS-Uniform	$(2.9251 \pm 0.6008) \times 10^{17}$	42.1916	0.003

Table 2.8: SIS-G* results for estimating the number of tables

Estimated number of tables	cv^2	Time (s)
8×8 table with all margins $= 6$		
$(1.1460 \pm 0.04159) \times 10^{20}$	1.3168	0.1
5×3 table (Table 2.1)		
$(2.4045 \pm 0.0246) \times 10^{8}$	0.1043	0.2
Hair color vs. eye color (Table 2.2)		
$(1.2065 \pm 0.0234) \times 10^{15}$	0.3775	3.7
Birth month vs. death month (Table 2.3)		
$(5.3961 \pm 0.5892) \times 10^{39}$	11.9211	0.4
30×30 table with all margins $= 3$		
$(1.8886 \pm 0.1250) \times 10^{92}$	4.3819	1.6
50×50 table with all margins $= 2$		
$(1.1845 \pm 0.1105) \times 10^{128}$	8.7072	3.9

Table 2.9: Performance comparison of SIS-G and SIS-G* for dense tables

Method	Estimated number of tables	cv^2	Time (s)
	4×4 table with all margins = 8		
SIS-G	$(9.8046 \pm 0.0197) \times 10^5$	0.004017	0.04
SIS-G*	$(9.5450 \pm 0.1941) \times 10^5$	0.4137	0.03
	8×8 table with all margins = 15		
SIS-G	$(8.1170 \pm 0.0280) \times 10^{33}$	0.0119	455.2
SIS-G*	$(8.5107 \pm 0.3714) \times 10^{33}$	1.9044	0.5
	5×5 table (described above)		
SIS-G	$(2.4244 \pm 0.0918) \times 10^{17}$	1.4346	942.7
SIS-G*	$(2.3149 \pm 0.1370) \times 10^{17}$	3.5017	2.2

Table 2.10: Performance comparison of SIS-G and SIS-G* for extremely dense tables

Method	Estimated number of tables	cv^2	Time (s)
	12×12 table with all margins = 90		
SIS-G*	$(6.5546 \pm 0.4358) \times 10^{150}$	4.4204	42.5
SIS-Uniform	$(9.6639 \pm 9.6591) \times 10^{116}$	998.9999	0.007
	10×10 table with all margins = 200		
SIS-G*	$(1.3811 \pm 0.0714) \times 10^{133}$	2.6707	119.8
SIS-Uniform	$(3.8062 \pm 3.8043) \times 10^{117}$	998.9983	0.005
	5×7 table (described above)		
SIS-G*	$(8.5705 \pm 1.0007) \times 10^{29}$	13.6329	2.4
SIS-Uniform	$(5.9938 \pm 1.1829) \times 10^{29}$	38.9352	0.002

observed χ^2 statistic is unusual when the observed table is considered to be a uniform draw from the set of all tables with given marginal sums. Given the observed χ^2 statistic S, we are interested in estimating the proportion of tables with $\chi^2 \leq S$.

We begin by describing the basic idea of volume tests, based on Sabatti (2002) and Diaconis and Efron (1985). Given an $m \times n$ contingency table T with bivariate distribution π_{ij}, $i = 1, \ldots m$, $j = 1 \ldots n$, we are interested in a measure of dependency when π_{ij} is unobserved and we observe only the table counts a_{ij}, $i = 1, \ldots m$, $j = 1 \ldots n$.

Adapting a dependency measure defined for π_{ij} to a_{ij}/M is not a solution to the task at hand. Consider the example presented in Sabatti (2002), where $m, n = 2$, and $M = 4$, with all margins equal to 2. The first table has probability 2/3 under independence, and the second two tables each have probability 1/6. The last two tables have a squared correlation coefficient, R^2, equal to 1, so 1/3 of the time an independent model leads to perfect dependence. This is a "spurious result due to sample size" (Sabatti, 2002).

1	1
1	1

2	0
0	2

0	2
2	0

Using a p-value as a measure of dependency also presents problems. Namely, a larger value of χ^2/M does not necessarily imply that the distribution with the larger value has larger dependence (Diaconis and Efron, 1985). Sabatti (2002) considers two 3×2 tables with $n = 100$, where the first table has $\chi^2/M = 0.791$ and the second has $\chi^2/M = 0.854$. These example tables are reproduced in Tables 2.11 and 2.12. One might expect that the second table has higher dependence than the first one, but examining the space of all π_{ij} with the same marginals as what was observed yields a counterintuitive result. This can be seen in Figure 2.3 below, where we examine all tables π_{ij} with margins equal to those reported. The shaded regions are these two spaces, parameterized by π_{11} and π_{21}, the closed circle represents the observed table and the open circle represents the table under independence. The contour lines are the Mahalanobis distance from independence. Table 2.11 has the highest possible distance from independence among tables with fixed margins, whereas Table 2.12 does not, even though Table 2.12 has the larger χ^2/M value.

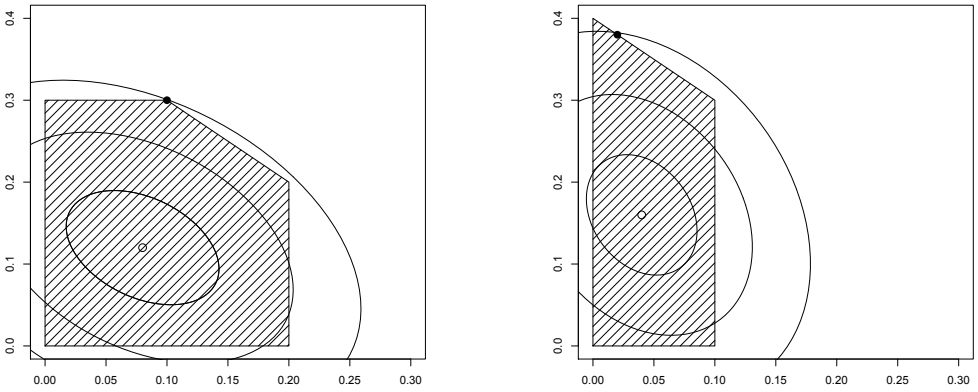
These deficiencies lead us to the conditional volume test, where we consider all possible tables with the same margins as the one observed and calculate what percentage of these possible tables

10	10	20
30	0	30
0	50	50
40	60	100

Table 2.11: Table with $\chi^2/M = 0.791$

2	8	10
38	2	4
0	5	5
40	60	100

Table 2.12: Table with $\chi^2/M = 0.854$



(a) Table 1

(b) Table 2

Figure 2.3: Testing independence and measuring dependency

yield a χ^2 statistic less than or equal to the one observed. A small percentage means that the table is close to independence, and a high percentage means that the observed table is far away from independence. A value of zero corresponds to the table expected under independence.

We observe the conditional volume test in Figure 2.4 following the example in Sabatti (2002). Here, the contour represents a distance measure between independence and the table we observed. When $n = 10$, there are only 2 observations with a smaller distance value from independence so the p-value of the conditional volume test is $2/11$. When $n = 20$, there are 32 possible tables, 11 of which have smaller distance values, so the p-value is $11/32$. The conditional volume test approximates the ratio of the region of the space of tables with smaller distances from independence than the one observed and the space of all possible tables. As n increases, the volume test p-value approaches the true ratio of the two volumes.

The p-value of the test is

$$\frac{\#\{\text{tables } : \chi^2(T) \leq \chi^2 | r_i, c_j\}}{\#\{\text{total tables} | r_i, c_j\}}. \tag{2.13}$$

The conditional volume test is performed on Table 2.1, Table 2.2 and a table given in Jones and

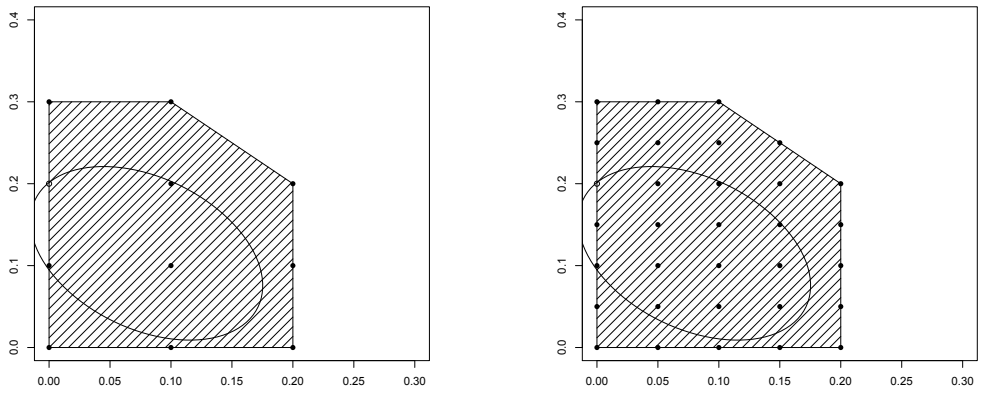


Figure 2.4: Sample size and volume measures

Table 2.13: Cross tabulation of race/ethnicity and weapon (Jones and O'Neil, 2006)

Weapon	Race/Ethnicity			
	White	Black	Hispanic	Total
Firearm	206	608	289	1103
Knife	74	222	130	426
Blunt object	19	49	16	84
Personal weapons	23	54	13	90
Total	322	933	448	1703

O'Neil (1999) depicting race/ethnicity versus type of weapon used for homicides in Los Angeles between 1980 and 1983 derived from an FBI database (Table 2.13). Here, S is equal to 72.18, 138.29 and 13.87 for Tables 2.1, 2.2 and 2.13, respectively. The proportion of Table 2.1 is 0.76086, the proportion of Table 2.2 is estimated to be around 0.154, and the proportion of Table 2.13 is estimated to be 1×10^{-4} (Diaconis and Gangolli, 1995; Jones and O'Neil, 1999). Results for Table 2.1 and Table 2.2 are shown in Table 2.14. SIS-G, SIS-GM1 and SIS-GM2 all perform reasonably well in these situations, and yield lower standard errors than SIS-Uniform. For the FBI data, SIS-G yields 0.00011 ± 0.000064 with cv^2 0.1432, and SIS-Uniform yields 0.000408 ± 0.00029 with cv^2 1.3937.

The conditional volume test can also be performed using the MCMC procedure of Diaconis and Gangolli (1995). At each step of the method, two rows and two columns are chosen randomly, and one of the following moves is made with equal probability:

Table 2.14: Performance comparison of methods for conditional volume test

Method	Estimated proportion $\chi^2 \leq S$
	5×3 table (Table 2.1)
SIS-G	0.7781 ± 0.0132
SIS-GM1	0.7695 ± 0.0131
SIS-GM2	0.7637 ± 0.0146
SIS-Uniform	0.7467 ± 0.0196
	Hair color vs. eye color (Table 2.2)
SIS-G	0.1522 ± 0.0117
SIS-GM1	0.1509 ± 0.0119
SIS-GM2	0.1427 ± 0.0107
SIS-Uniform	0.1171 ± 0.0237

$$\begin{array}{cc} +1 & -1 \\ -1 & +1 \end{array} \qquad \begin{array}{cc} -1 & +1 \\ +1 & -1 \end{array}$$

If a negative entry is obtained, the table remains the same. This method is easy to implement. MCMC is run for each table for $900,000$ iterations with $100,000$ burn-in. For Table 2.1, an estimate of 0.7589 ± 0.0212 is obtained. For Table 2.2, an estimate of 0.1624 ± 0.02 is obtained. For Table 2.13, MCMC yields 0.00085 ± 0.00027. The MCMC procedure has larger standard errors than any of the SIS methods for Table 2.13 and Table 2.2. For Table 2.13, this means that less than 1% of tables with these margins are as close to independence as the observed table, and that we may accept the hypothesis of independence, avoiding conclusions such as "Hispanics are more likely to use knives" (Jones and O'Neil, 1999).

For tables that are large and sparse, the chain is sticky and takes a long time to explore the space. Additionally, the χ^2 test statistic is easy to calculate. If a researcher is interested in a test statistic that is more computationally intensive, MCMC may take a prohibitively long time to run because it requires so many samples relative to SIS.

2.4.3 Sampling Tables with Structural Zeros

The approximation of Good (1976) may be extended to sample tables in other scenarios of interest. For example, tables may contain certain entries that are structural zeros. In this section, we extend SIS-G* to sample tables with structural zeros.

We focus on a common case, a contingency table with structural zeros on the diagonal. Recalling the combinatorial interpretation of Good's Equation, we can adapt the approximation to the case where there are structural zeros. Here, instead of n places to put the first row sum r_1, there are now $n-1$ because of the structural zero in cell $(1,1)$. There are also $n-1$ places to put the first column sum c_1, rather than the original n.

Table 2.15: Example table with structural zeros

[0]							r_1
	[0]						r_2
		[0]					r_3
			$\ddots$				$\vdots$
				[0]			r_{n-2}
					[0]		r_{n-1}
						[0]	r_n
c_1	c_2	c_3	...	c_{n-2}	c_{n-1}	c_n	M

So a natural *ad hoc* extension of Good's Equation for the case of an integer-valued contingency table with structural zeros on the diagonal is

$$\frac{\prod_{i=1}^{n}\binom{n+r_i-2}{r_i}\prod_{j=1}^{n}\binom{n+c_j-2}{c_j}}{\binom{M+n(n-1)-1}{M}}. \tag{2.14}$$

More generally, denote by S the set of structural zeros, $s_r(i) = \sum_{j=1}^{n} \mathbb{1}(\alpha_{ij} \in S)$, $s_c(j) = \sum_{i=1}^{m} \mathbb{1}(\alpha_{ij} \in S)$, and $s_T = \sum_{i=1}^{n} s_r(i) = \sum_{j=1}^{n} s_c(j)$. Denote by $\Sigma_{\mathbf{rc}}^{S}$ the set of tables with S structural zeros. In this case we may approximate $\Sigma_{\mathbf{rc}}^{S}$ by

Theorem 2.4.1.

$$|\Sigma_{\mathbf{rc}}^{S}| \approx \Delta_{\mathbf{rc}}^{\text{G'}} = \frac{\prod_{i=1}^{m}\binom{n+r_i-s_r(i)-1}{r_i}\prod_{j=1}^{n}\binom{m+c_j-s_c(j)-1}{c_j}}{\binom{M+mn-s_T-1}{M}}, \tag{2.15}$$

and we may use this approximation to derive a proposal distribution to sample from $\Sigma_{\mathbf{rc}}^{S}$ using

a similar method as SIS-G* that takes into account structural zeros. In addition to updating the row and column margins, $s_r(i), s_c(j)$, and s_T are updated after sampling each cell.

After sampling α_{i1} where $2 \leq i \leq k-1$, the proposal for α_{k1} is given in Proposal 2.6. The proposal construction follows the same strategy as the cell by cell sampling method SIS-G* described in Section 2.3.1.

Proposal 2.6. Define $s_c(1)^* = \sum_{i=k+1}^{m} \mathbb{1}(\alpha_{ij} \in S)$. Then

$$q(a_{k1} = \alpha_{k1}) \propto \frac{\binom{n - s_r(i) + r_k - \alpha_{k1} - 2}{r_k - \alpha_{k1}} \binom{m - k - s_c(1)^* + c_1 - \sum_{i=1}^{k} \alpha_{i1} - 1}{c_1 - \sum_{i=1}^{k} \alpha_{i1}}}{\binom{M - \sum_{i=1}^{k} \alpha_{i1} + mn - \sum_{j=2}^{n} s_c(j) - s_c(1)^* - 1}{M - \sum_{i=1}^{k} \alpha_{i1}}}.$$

Unfortunately, the naive implementation using the same bounds as SIS-Uniform will result in a certain percentage of invalid tables. However, Mirsky (1971) provides necessary and sufficient conditions for the existence of an integer matrix with prescribed bounds for its entries and row and columns sums. Chen (2007) used Mirsky's Theorem to derive the necessary and sufficient conditions for the existence of a table with fixed margins and a prescribed set of structural zeros. We focus on the case of a zero diagonal.

In the case where there is at most one structural zero in each column, the condition is simplified and sampling the first column is equivalent to finding an integer vector $(t_{11}, \ldots, t_{m1})$ such that $\sum_{i=1}^{m} t_{i1} = c_1$ and

$$l_{i1} \leq t_{i1} \leq u_{i1}, \; i = 1, \ldots, m, \tag{2.16}$$

where

$$(l_{i1}, u_{i1}) = \begin{cases} (0, 0), & \text{if } (i,1) \in S, \\ (\max\{0, r_i - \sum_{j=2}^{n} c_j \mathbb{1}_{(i,j) \notin S}, \min(r_i, c_1)\} & \text{if } (i,1) \notin S. \end{cases} \tag{2.17}$$

Suppose we have already chosen $t_{i1} = t_{i1}^*$ for $1 \leq i \leq k-1$, then the only restrictions on t_{k1} are

$$\max\left\{ l_{i1}, c_1 - \sum_{i=1}^{k-1} t_{i1}^* - \sum_{i=k+1}^{m} u_{i1} \right\} \leq t_{k1} \leq \min\left\{ u_{i1}, c_1 - \sum_{i=1}^{k-1} t_{i1}^* - \sum_{i=k+1}^{m} l_{i1} \right\}. \tag{2.18}$$

SIS-Uniform will sample an integer uniformly between the lower and upper bounds, while SIS-

Table 2.16: Monkey genital display data (Ploog, 1967)

Active Participant	Passive Participant					
	R	S	T	U	V	W
R	[0]	1	5	8	9	0
S	29	[0]	14	46	6	0
T	0	0	[0]	0	0	0
U	2	3	1	[0]	38	2
V	0	0	0	0	[0]	1
W	9	25	4	6	13	[0]

G* will sample an integer based on Proposal 2.6. Both will use the bounds derived by Chen (2007) from Mirsky (1971).

Ploog (1967) collected data on genital displays in a colony of six squirrel monkeys, labeled R,S,T,U,V, and W. Genital display is a social signal, with an active and passive participant in each display. The diagonal cells are zero since a monkey never displays its genitals to itself. The data is shown in Table 2.16. Fienberg (1980) conducted a test of quasi-independence on these data and rejected the null hypothesis at a small significance level. To help interpret this result, Chen (2007) considered the conditional volume test. Here, we compare their SIS-Uniform procedure to SIS-G* accounting for the structural zeros on the diagonal using Proposal 2.6. We use 1,000 importance samples for estimating the number of tables and $1,000,000$ samples for the conditional volume test. Results are shown in Tables 2.17 and 2.18. It appears that SIS-G* is an improvement over SIS-Uniform, resulting in a lower cv^2 and a smaller standard error for estimating the number of tables. Running $1,000,000$ SIS samples yields an estimate of the number of tables for the genital display data of $(8.76 \pm 0.03) \times 10^{12}$. The conditional volume test also illustrates an improvement in the estimate of the p-value, with a smaller standard error. The SIS-G* procedure is more efficient even adjusting for the moderately increased computation of SIS-G* relative to SIS-Uniform. As another small example consider a 7×7 table with margins $\mathbf{r} = \{17, 15, 32, 15, 9, 3, 14\}$, $\mathbf{c} = \{23, 28, 7, 11, 19, 24, 3\}$ and structural zeros on the diagonal. Here the difference in performance between SIS-G* and SIS-Uniform is even more pronounced, with dramatically higher cv^2 and larger standard error.

Additional methods for performing the conditional volume test on tables with structural zeros include an MCMC method of Aoki and Takemura (2005). This generated $2,000,000$ samples using

Table 2.17: Performance comparison for estimating the number of tables

Method	Estimated number of tables	cv^2
	(Ploog, 1967)	
SIS-G*	$(8.3799 \pm 0.4860) \times 10^{12}$	3.3628
SIS-Uniform	$(8.3503 \pm 0.6953) \times 10^{12}$	6.9324
	7×7 table	
SIS-G*	$(6.6977 \pm 0.4825) \times 10^{18}$	5.1888
SIS-Uniform	$(5.3430 \pm 1.6066) \times 10^{18}$	90.4120

Table 2.18: Performance comparison for the conditional volume test

Method	Estimated proportion $\chi^2 \leq S$
	(Ploog, 1967)
SIS-G*	(0.9568 ± 0.0171)
SIS-Uniform	(0.9465 ± 0.0229)

$500,000$ burn-in in three seconds and estimated the p-value to be 0.93 ± 0.01 (Chen et al., 2005). SIS-Uniform and SIS-G* are both more efficient than this method. In addition, MCMC requires many more samples relative to SIS so if a statistic is difficult to calculate MCMC may be computationally infeasible.

2.4.4 Plant-pollinator networks

Important types of ecological data may be analyzed using sequential importance sampling strategies. For example, plant-pollinator networks are bipartite graphs composed of a set of nodes representing plant species and a set of nodes representing pollinator species. Links between nodes represent the frequency of interaction between a specific plant pollinator pair. These data may be expressed equivalently as a contingency table, with the rows and columns representing plants and pollinators, respectively. See Table 2.19 for an example.

Ecologists are interested in answering research questions concerning phenomena such as patterns of species distribution, biodiversity and coevolutionary processes. Here, we assess the property of nestedness in data collected by Barrett and Helenurm (1987) in New Brunswick, Canada. These data are presented in Table 2.19 with 12 rows and 102 columns representing plants and pollinators, respectively. Nestedness is a pattern in which "specialist pollinator species visit plant species that are subsets of those visited by more generalist pollinators" (Pawar, 2014). The degree of nestedness

in ecological networks has implications for the maintenance of biodiversity and coevolution (Burgos et al., 2007; Bascompte et al., 2003). Specifically, highly nested communities make it less likely for a species to become isolated following the removal of other species from the system. Additionally, the pattern of nestedness allows for rare species to remain in the system (Dormann et al., 2009; Jordano, 1987; Blüthen et al., 2007; Bascompte et al., 2006, 2003).

Table 2.19: The 12×102 plant-pollinator data from New Brunswick, Canada (Barrett and Helenurm, 1987).

	Acmaeopsoides rufula	*Agiotes stabilis*	*Ancistrocerus sp.*	*Andrena melanochroa*	*Andrena miranda*	*Andrena nivalis*	*Andrena rufosignata*	...	*Tropidia quadrata*	*Tychius stephensi*	*Vespula arenaria*	*Xylota bigelowi*	*Xylota hinei*	*Xylota sp.*	*Zeraea americana*	Total
Aralia nudicaulis	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	66
Chimaphila umbellata	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	70
Clintonia borealis	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	24
Cornus canadensis	1	0	2	3	2	1	2	...	2	3	0	10	2	1	1	167
Cypripedium acaule	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	8
Linnaea borealis	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	37
Maianthemum canadense	0	2	0	0	1	1	0	...	2	0	0	0	1	0	0	85
Medeola virginiana	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1
Oxalis montana	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	81
Pyrola secunda	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	4
Trientalis borealis	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	3
Trillium undulatum	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	4
Total	1	2	2	3	3	2	2	...	4	3	1	10	3	1	2	550

To quantify nestedness in weighted bipartite networks, Almeida-Neto and Ulrich (2011) proposed the test statistic WNODF (Weighted Nestedness Metric based on Overlap and Decreasing Fill). This statistic requires that the rows and columns of the table be sorted in decreasing order of the number of nonzero entries. We denote this statistic as

$$S = \frac{2(S_c + S_r)}{m(m-1) + n(n-1)}, \tag{2.19}$$

where

$$S_c = 100 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Big(\frac{\sum_{k=1}^{m} \mathbb{1}(\alpha_{kj} < \alpha_{ki}, \alpha_{kj} > 0)}{F(c_j)} \Big) \mathbb{1}(F(c_j) < F(c_i)), \tag{2.20}$$

where $F(c_i) = \sum_{k=1}^{m} \mathbb{1}(\alpha_{ki} > 0)$ denotes the number of nonzero entries in column c_i. The term S_r is expressed in a similar fashion considering the rows instead of the columns. The statistic S tends to be large when the level of nestedness is high. Calculation of the statistic S is performed in R.

We are interested in assessing whether the observed level of nestedness may be due to chance. A variety of testing procedures are available for these types of ecological data (Ulrich and Gotelli, 2007; Almeida-Neto and Ulrich, 2011; Dormann et al., 2009; Bascompte et al., 2003; Gotelli and Graves, 1996). We will consider the null hypothesis that the observed table is not unusual when considered to be a sample drawn uniformly from the set of all tables with the same margins as the observed table (Ulrich and Gotelli, 2007). The null hypothesis will be rejected if S is too large.

To estimate the p-value of the test, SIS-G was used to generate 125 importance samples which gave an estimated p-value of 0.3367 ± 0.0972. The cv^2 is 5.5986. The p-value indicates that the level of nestedness is not statistically significant, meaning that the data do not suggest a nested structure. MCMC was used to generate $1,100,000$ samples, with $100,000$ burn-in and employing thinning to reduce autocorrelation, and the estimated p-value is 0.3050. The standard error calculated by the batch means method is 0.04738, but running 100 MCMC chains with a different SIS-G generated starting position each time yields a standard error of 0.07281, indicating the standard error based on batch means is an underestimate. While the MCMC procedure is quicker to run than SIS-G, it has a standard error that is an underestimate, and the chain is doing a poor job of exploring the space of tables (SIS-G estimates there are $(2.3116 \pm 0.4892) \times 10^{190}$ tables with the same margins as Table 2.19).

These sampling methods may be extended to other statistics of interest, many of which have been discussed in the literature (Dormann et al., 2009). Additionally, these approaches may be used in a similar fashion for statistical inference on other types of ecological data, including plant-frugivore, host-parasite, plant-herbivore, and plant-seed disperser networks, which have the same bipartite structure as plant-pollinator networks (Guimarâes and Raimundo, 2012). The sampling strategies discussed here may be applied to integer-weighted, directed graphs, allowing for conditional inference on other types of networks.

2.5 Ordering Strategies

For the column by column sampling procedures, the columns may be sampled in any order, and for the cell by cell procedures, both the rows and columns can be sampled in any order. In this section, we compare different ordering strategies and describe specific orderings that will result in effective sampling procedures.

2.5.1 Sampling by Column

For the column by column sampling procedures SIS-G, SIS-GM1, SIS-GM2, and the additional sampling method for sparse tables SIS-B, columns may be sampled in any order. Simulations were run of a wide variety of tables and indicate that sampling the columns in increasing order generally yields the lowest cv^2 and standard errors. In some cases, sampling the columns in increasing order results in a substantially lowered cv^2 relative to sampling the columns in decreasing order. Intuitively, this approach makes sense, because with a small column sum, there are relatively few choices of where to allocate the sum and the proposal will be close to the target. Then, after sampling the first column sum and reducing the row sums, the proposal becomes even closer to the target.

However, there is a time cost to sampling the columns in increasing order, as the normalizing constant for successively larger column sums must be calculated. In the case of sampling in decreasing order, the normalizing constant of the first column only needs to be calculated once and can be reused, and there are less possibilities to examine for the second through n^{th} columns. In some cases the difference in time between increasing and decreasing column sums can be dramatic, and an intermediate position yields an efficient compromise. In this sampling procedure, the largest column is sampled first and then the remaining columns are sampled in increasing order. In this method, the normalizing constant of the largest first column only needs to be calculated once and time savings are accumulated across samples. These are general strategies, however, and it may be advantageous to conduct a small preliminary study to examine which ordering configuration will yield the smallest cv^2 and the most efficient sampling method.

2.5.2 Sampling by Cell

In the case of the cell by cell sampling method, SIS-G*, both the row and column sums may be sampled in increasing or decreasing order. The other cell by cell method examined, SIS-Uniform, was discovered to have the best performance by listing the column sums in decreasing order and the row sums in increasing order (Chen et al., 2005). In the case of SIS-G*, the best performance as judged by cv^2 and standard error was a close tie between sampling columns in increasing order and rows in decreasing order and sampling both rows and columns in increasing order. The worst performance by a wide margin was obtained by sampling both rows and columns in decreasing order.

This intuition behind this results is that if the row and column sums are small, there are not many options of values to put into the first cell and so the proposal is close to the target. After sampling this first cell, the updated row and column sums are reduced, causing the proposal to become even closer to the target uniform distribution. A similar intuition holds for sampling tables in Chen et al. (2005).

2.6 Alternative Methods for Estimating the Number of Tables

There are a number of competing methods for approximating $|\Sigma_{\mathbf{rc}}|$. Asymptotic approximations were provided by Békéssy et al. (1972), O'Neil (1969), Good and Crook (1977) and Bender (1974), however, these methods perform extremely poorly on small and dense tables.

An approximation based on an application of the central limit theorem was provided by Gail and Mantel (1977) and is reported in Theorem 2.6.1.

Theorem 2.6.1.

$$|\Sigma_{\mathbf{rc}}| \approx \Delta^{GM} \equiv \prod_{i=1}^{m} \binom{r_i + n - 1}{n-1} ((n-1)/2\pi\sigma^2 n)^{\frac{n-1}{2}} n^{1/2} \exp[-((n-1)/\sigma^2 n)(\sum_{j=i}^{n} c_j^2 - M^2/n). \tag{2.21}$$

However, this method appears to give inaccurate results in many cases. Diaconis and Efron (1985) provided an approximation which can be effective in cases where m, n are small and M is large, but can still provide misleading results in some situations. It is reported in Theorem 2.6.2.

Table 2.20: Performance comparison of alternative methods for estimating the number of large tables

Method	Estimated number of tables	Relative Error (%)
	5×3 table (Table 2.1)	
Zipunnikov	$85,638,274$	-64.225
Gail Mantel	$220,141,654$	-8.038
Diaconis	$232,034,659$	-3.069
	Hair color vs. eye color (Table 2.2)	
Zipunnikov	1.197054×10^{15}	-2.354
Gail Mantel	1.074267×10^{15}	-12.370
Diaconis	1.261337×10^{15}	+2.889
	8×8 table with all margins = 6	
Zipunnikov	$9.117823e \times 10^{19}$	-20.438
Gail Mantel	1.376972×10^{20}	+20.155
Diaconis	2.113299×10^{20}	84.407
	Birth month vs. death month (Table 2.3)	
Zipunnikov	5.330762×10^{39}	-
Gail Mantel	4.015898×10^{39}	-
Diaconis	7.157447×10^{41}	-
	$\mathbf{r} = \{154, 5, 78, 79, 82\}$, $\mathbf{c} = \{101, 182, 22, 86, 7\}$	
Zipunnikov	2.170806×10^{17}	-6.088
Gail Mantel	7.763618×10^{17}	+235.864
Diaconis	2.762671×10^{17}	+19.517
	5×5 table with all margins = 50	
Zipunnikov	5.923241×10^{20}	-21.091
Gail Mantel	8.529467×10^{20}	13.629
Diaconis	7.439661×10^{20}	-0.889
	3×49 table with all margins = $98, 6$	
Zipunnikov	7.634436×10^{67}	-24.486
Gail Mantel	5.999209×10^{68}	+493.394
Diaconis	1.278121×10^{68}	+26.421

Theorem 2.6.2. Diaconis and Efron (1985) suggest without proof that if $w = \frac{1}{1+mn/2M}$, $k = \frac{n+1}{n\sum \bar{r}_i^2} - \frac{1}{n}$, $\bar{r}_i = \frac{1-w}{m} + \frac{wr_i}{M}$, and $\bar{c}_j = \frac{1-w}{n} + \frac{wc_j}{M}$, then

$$|\Sigma_{\mathbf{rc}}| \sim \left(\frac{2M+mn}{2}\right)^{(m-1)(n-1)} \left(\prod_{i=1}^{m} \bar{r}_i\right)^{n-1} \left(\prod_{j=1}^{n} \bar{c}_j\right)^{k-1} \frac{\Gamma(nk)}{\Gamma(n)^m \Gamma(k)^n}. \tag{2.22}$$

Another set of approximations based on a double saddle point approximation was provided by Zipunnikov et al. (2009), with multiple approximations based on the configuration of the table and a correction. Results for estimating the number of tables for these methods are provided in Tables 2.20 and Tables 2.21. Although there are six approximations for each table, we examine only one and note that results were roughly similar. Results indicate reasonable performance in many cases, but also large relative errors where the true number of tables are known. The Gail and Mantel (1977) approximation has large relative errors in cases where the table is unbalanced or has rough margins, achieving a relative error of over 200% for the 5×5 table with rough margins. The Diaconis and Efron (1985) method also appears to give inaccurate results in large and sparse cases. SIS methods have the advantage of being able to achieve a smaller standard error for additional samples, whereas the methods of Gail and Mantel (1977), Diaconis and Efron (1985) and Zipunnikov et al. (2009) only provide a single estimate.

An additional approximation was provided by Barvinok and Hartigan (2010), which can be very effective. They reported their approximation gives about 1.30×10^{15} for the cross-classified hair color eye color data, a relative error of 6%. However, SIS-G* obtains a relative error of less than 1% in just a few seconds. Holmes and Jones (1996) provided an additional method for estimating $|\Sigma_{\mathbf{rc}}|$, but it requires calculating the coefficients of a product of polynomials and is suspected of underestimating the true number of tables (Chen et al., 2005).

Exhaustively enumerating all tables in $\Sigma_{\mathbf{rc}}$ was explored by Balmer (1988) and Gail and Mantel (1977), see Diaconis and Gangolli (1995) for a review. While exhaustively enumerating all tables in $\Sigma_{\mathbf{rc}}$ is reasonable in some cases, it is not feasible in cases where the number of tables is extremely large. The method LattE is also a useful and groundbreaking tool for calculating the number of tables, and its performance has been described throughout the chapter. It is relatively quick to run, provides an exact value for the number of tables, but takes too long to run for tables that are

Table 2.21: Performance comparison of methods for estimating the number of large tables

Method	Estimated number of tables
	50×50 table with all margins $= 2$
Zipunnikov	5.984609×10^{128}
Gail Mantel	9.72165×10^{128}
	75×75 table with both margins $= (5, 2, \ldots, 2)$
Zipunnikov	9.546187×10^{221}
Gail Mantel	1.4794×10^{219}
	100×100 table with both margins $= (5, 1, \ldots, 1)$
Zipunnikov	1.385168×10^{165}
Gail Mantel	1.856221×10^{163}
	75×75 table with all margins $= 2$
Zipunnikov	9.062334×10^{218}
Gail Mantel	1.479368×10^{219}
	10×10 table with all margins $= 200$
Zipunnikov	1.149967×10^{133}
Gail Mantel	1.765844×10^{133}

moderately large.

Additional computational methods include a Monte Carlo algorithm of Dyer (2003) and a method of Karp et al. (1989) based on dynamic programming. These methods can be effective in many cases but are difficult to implement.

2.7 Discussion

We have developed SIS strategies for sampling tables with fixed margins based on asymptotic approximations of Greenhill and McKay (2008) and an approximation of Good (1976). These methods sample the tables column by column, and provide smaller cv^2 values and standard errors than SIS-Uniform, indicating an improvement over current methods, especially for large sparse tables. We also developed a cell by cell sampling method using (2.11) which provides an improvement when the column sampling procedures are too time-consuming. Although this procedure is less efficient than column sampling, it is much faster and may be used in situations where column-sampling methods

take a prohibitively long time. We also examine an SIS strategy based on an approximation of Bender (1974) that is useful for large, sparse tables.

The proposed SIS methods give more reliable results than MCMC for testing statistical hypotheses on contingency tables or bipartite graphs. The proposed methods are extremely flexible in the sense that the distribution of any test statistic related to the structure or pattern of a contingency table with fixed margins can be approximated and a p-value estimated.

2.8 Proofs of the Main Results

Proof of Theorem 2.3.3

Denote by (m_{ij}) an $m \times n$ 0-1 matrix, where $m_{ij} = 0$ denotes a structural zero at position (i, j). Let $|\Sigma_{\mathbf{rc}}|$ be the number of $m \times n$ integer matrices over $[0, d]$ such that $\alpha_{ij} = 0$ whenever $m_{ij} = 0$, $\sum_j \alpha_{ij} = r_i$ and $\sum_i \alpha_{ij} = c_j$. According to Theorem 1 of Bender (1974),

$$|\Sigma_{\mathbf{rc}}| \sim \frac{M!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{n} c_j!} \exp\{\epsilon a - b\} \tag{2.23}$$

uniformly, where $\epsilon = -1$ if $d = 1$ and $\epsilon = 1$ if $d > 1$, $a = (\sum_{i=1}^{m} r_i(r_i - 1))(\sum_{j=1}^{n} c_j(c_j - 1))/2M^2$, and $b = \sum_{m_{ij}=0} r_i c_j / M$.

In the case of integer-valued tables bounded above by a constant d, $\epsilon = 1$ and there are no structural zeros, so $b = 0$. Substituting these values into equation 2.23 yields Theorem 2.3.3.

Proof of Proposal 2.1

The approximation of Good (1976) implies the number of tables after sampling the first column is approximately

$$|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}| \approx \Delta^{\mathrm{G}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}} = \frac{\prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \prod_{j=2}^{n} \binom{m + c_j - 1}{c_j}}{\binom{M - c_1 + m(n-1) - 1}{M - c_1}},$$

and the approximation to the total number of tables $|\Sigma_{\mathbf{rc}}| \approx \Delta^{\mathrm{G}}_{\mathbf{rc}}$ is given in (2.6). Consequently, the proposal SIS-G is

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta^{\mathrm{G}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{G}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}}.$$

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta^{\mathrm{G}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{G}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}}.$$

Proof of Proposal 2.2

Using the approximation $\Delta^{\mathrm{GM1}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}$ from Greenhill and McKay (2008), we have

$$|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}| \sim \Delta^{\mathrm{GM1}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}} = \frac{\prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \prod_{j=2}^{n} \binom{m + c_j - 1}{c_j}}{\binom{M - c_1 + m(n-1) - 1}{M - c_1}} \exp\{\boldsymbol{\alpha}(\mathbf{r}^{(2)}, \mathbf{c}^{(2)})\},$$

and

$$|\Sigma_{\mathbf{rc}}| \sim \Delta^{\mathrm{GM1}}_{\mathbf{rc}} = \frac{\prod_{i=1}^{m} \binom{n + r_i - 1}{r_i} \prod_{j=1}^{n} \binom{m + c_j - 1}{c_j}}{\binom{M + mn - 1}{M}} \exp\{\boldsymbol{\alpha}(\mathbf{r}, \mathbf{c})\}.$$

So the proposal SIS-GM1 is

$$q(t_1 = (\alpha_{11}, \ldots, \alpha_{m1})) \propto \frac{\Delta^{\mathrm{GM1}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{GM1}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \exp\{\boldsymbol{\alpha}(\mathbf{r}^{(2)}, \mathbf{c}^{(2)})\}.$$

Proof of Proposal 2.3

Using the approximation $\Delta^{\mathrm{GM2}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}$ from Greenhill and McKay (2008), we have

$$|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}| \sim \Delta^{\mathrm{GM2}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}} = \frac{\prod_{i=1}^{m} \binom{n + r_i - \alpha_{i1} - 2}{r_i - \alpha_{i1}} \prod_{j=2}^{n} \binom{m + c_j - 1}{c_j}}{\binom{M - c_1 + m(n-1) - 1}{M - c_1}} \exp\left\{\frac{1}{2}(1 - \hat{\mu}_2^{(2)})(1 - \hat{v}_2^{(2)})\right\},$$

and

$$|\Sigma_{\mathbf{rc}}| \sim \Delta^{\mathrm{GM2}}_{\mathbf{rc}} = \frac{\prod_{i=1}^{m}\binom{n+r_i-1}{r_i}\prod_{j=1}^{n}\binom{m+c_j-1}{c_j}}{\binom{M+mn-1}{M}} \exp\left\{\frac{1}{2}(1-\hat{\mu}_2)(1-\hat{v}_2)\right\}.$$

So the proposal SIS-GM2 is

$$q(t_1 = (\alpha_{11},\ldots,\alpha_{m1})) \propto \frac{\Delta^{\mathrm{GM2}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}}{\Delta^{\mathrm{GM1}}_{\mathbf{rc}}} \propto \prod_{i=1}^{m}\binom{n+r_i-\alpha_{i1}-2}{r_i-\alpha_{i1}} \exp\left\{\frac{1}{2}(1-\hat{\mu}_2^{(2)})(1-\hat{v}_2^{(2)})\right\}.$$

Proof of Proposal 2.4

Using the approximation $\Delta^{\mathrm{B}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}$ from Bender (1974), we have

$$|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}| \sim \Delta^{\mathrm{B}}_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}} = \frac{(M-c_1)!}{\prod_{i=1}^{m}(r_i-\alpha_{i1})!\prod_{j=2}^{n}c_j!} \exp\left\{\frac{(\sum_{i=1}^{m}(r_i-\alpha_{i1})(r_i-\alpha_{i1}-1))(\sum_{j=2}^{n}c_j(c_j-1))}{2(M-c_1)^2}\right\},$$

and

$$|\Sigma_{\mathbf{rc}}| \sim \Delta^{\mathrm{B}}_{\mathbf{rc}} = \frac{M!}{\prod_{i=1}^{m}r_i!\prod_{j=1}^{n}c_j!} \exp\left\{\frac{(\sum_{i=1}^{m}r_i(r_i-1))(\sum_{j=1}^{n}c_j(c_j-1))}{2M^2}\right\}.$$

So the proposal SIS-B is

$$q(t_1 = (\alpha_{11},\ldots,\alpha_{m1})) \propto \frac{1}{\prod_{i=1}^{m}(r_i-\alpha_{i1})!} \exp\left\{\frac{(\sum_{i=1}^{m}(r_i-\alpha_{i1})(r_i-\alpha_{i1}-1))(\sum_{j=2}^{n}c_j(c_j-1))}{2(M-c_1)^2}\right\}.$$

Proof of Theorem 2.3.4

Since the conditions of Theorem 2 hold, we have

$$\left|\frac{|\Sigma_{\mathbf{rc}}|}{\Delta_{\mathbf{rc}}^{\mathrm{GM1}}}-1\right|=\left|\exp\left\{O\left(\frac{r^3c^3}{M^2}\right)\right\}-1\right|=O\left(\frac{r^3c^3}{M^2}\right).$$

The approximation error can be written as

$$\begin{aligned}\left|\frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{|\Sigma_{\mathbf{rc}}|}-\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{\Delta_{\mathbf{rc}}^{\mathrm{GM1}}}\right|&=\left|\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{rc}}|}\left(\frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}-1\right)-\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{rc}}|}\left(\frac{|\Sigma_{\mathbf{rc}}|}{\Delta_{\mathbf{rc}}^{\mathrm{GM1}}}-1\right)\right|\\&\leq\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{rc}}|}\left(\left|\frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}-1\right|+\left|\frac{|\Sigma_{\mathbf{rc}}|}{\Delta_{\mathbf{rc}}^{\mathrm{GM1}}}-1\right|\right).\end{aligned}$$

Combine the above results with the fact that

$$\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{rc}}|}=\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}\frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{|\Sigma_{\mathbf{rc}}|}\leq\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}\leq 2$$

for large M, then we have

$$\left|\frac{|\Sigma_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}|}{|\Sigma_{\mathbf{rc}}|}-\frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}^{\mathrm{GM1}}}{\Delta_{\mathbf{rc}}^{\mathrm{GM1}}}\right|=O\left(\frac{r^3c^3}{M^2}\right).$$

The proof is similar for $\Delta_{\mathbf{rc}}^{\mathrm{GM2}}$.

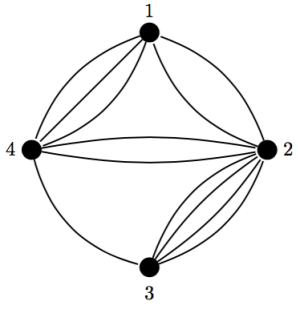
Chapter 3

Sampling for Conditional Inference on Multigraphs

3.1 Introduction

Network data is extremely common and there is currently a huge interest in statistical methods for analyzing networks. Fields as diverse as ecology, sociology, and economics deal with networks on a regular basis and require statistical approaches and analysis strategies. Substantial literature is available on methods for graphs with only a single edge between nodes (simple graphs), however, relatively less time has been spent on the case where the network may have multiple links between edges. A network of this type is commonly called a multigraph. Performing statistical inference on multigraphs is of interest to researchers. For example, they may be interested in the number of emails sent between pairs of people in a social group or the number of interactions observed between pairs of animals. As a small, toy example, consider Figure 3.1, which shows an undirected, loopless multigraph and the equivalent adjacency matrix.

We are interested in testing whether or not some pattern or property of a multigraph deviates from random. To perform the task of testing, a common procedure is to condition on the degree sequence and compare the observed graph to the set of graphs with the same degree sequence. This



(a) Undirected Multigraph

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

(b) Adjacency Matrix

Figure 3.1: Undirected multigraph and its associated adjacency matrix

is an application of exact inference, which requires no potentially inaccurate asymptotic approximations and additionally eliminates nuisance parameters (Agresti, 1992*a*; Cochran, 1952; Lehmann, 1959). Conditioning on the degree sequence also creates a probabilistic basis for a test in situations where the subjects were not obtained by sampling but are the only ones available (Lehmann, 1959).

Several Markov chain Monte Carlo (MCMC) algorithms for sampling simple graphs from the uniform distribution have been proposed (Roberts, 2000; Milo et al., 2002; McDonald et al., 2007; Handcock et al., 2008), and importance sampling methods were considered in Snijders (2006), Blitzstein and Diaconis (2010), Bayati et al. (2010) and Zhang and Chen (2013). Relatively less attention has been paid to the problem of sampling multigraphs.

Here, we are concerned with sampling multigraphs with no self loops uniformly from the set of all such multigraphs with fixed degree sequence. Based on these sampled graphs, the distribution of a test statistic may be approximated. Additionally, we are interested in estimating the total number of multigraphs with the same, fixed degree sequence.

Sampling from the uniform distribution over multigraphs with fixed degree is difficult. Here, we propose a new sequential importance sampling (SIS) method that uses the asymptotic approximation of Bender and Canfield (1978) to guide the sampling. A multigraph is generated and its associated importance weight is used to correct for the bias incurred by sampling. Using these graphs and weights, the distribution of any test statistic may be estimated, and we may additionally obtain an approximation to the number of multigraphs. We also propose an MCMC method for sampling multigraphs with fixed degree.

This chapter is organized in the following way. Section 3.2 introduces the basics of SIS. Section 3.3 describes how the approximation is incorporated into the proposal to perform SIS. Section 3.4 proposes an MCMC method for sampling multigraphs. Section 3.5 provides applications, including an analysis of the clustering of a primate social network and the resilience of an airline network, as well as counting the number of graphs. Section 3.6 provides concluding remarks.

3.2 Sequential Importance Sampling

Multigraphs can be expressed equivalently as a symmetric integer-valued adjacency matrix with a zero diagonal, so to sample multigraphs we may equivalently sample adjacency matrices. Let $\Sigma_{\mathbf{d}}$

denote the set of all $n \times n$ symmetric tables with row margins $\mathbf{d} = (d_1, \ldots, d_n)$, non-negative integer entries, and a zero diagonal, $M = \sum_{i=1}^{n} d_i$, and $|\Sigma_{\mathbf{d}}|$ the total number of tables in the set. Denote by $p(T) = 1/|\Sigma_{\mathbf{d}}|$ the uniform distribution over $\Sigma_{\mathbf{d}}$.

If we are interested in estimating $\mu = E_p[f(T)]$, and a table $T \in \Sigma_{\mathbf{d}}$ can be simulated from a proposal distribution $q(\cdot)$ that can be easily sampled from and includes the support of $\Sigma_{\mathbf{d}}$, then we may estimate μ using the weighted average of $T_1, \ldots, T_N$, independent and identically distributed (iid) samples drawn from $q(T)$,

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(T_i) \frac{p(T_i)}{q(T_i)}}{\sum_{i=1}^{N} \frac{p(T_i)}{q(T_i)}} = \frac{\sum_{i=1}^{N} f(T_i) \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}}{\sum_{i=1}^{N} \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}}. \tag{3.1}$$

Additionally, the total number of graphs can be written as

$$|\Sigma_{\mathbf{d}}| = \sum_{T \in \Sigma_{\mathbf{d}}} \frac{1}{q(T)} q(T) = E_q \left[\frac{\mathbb{1}_{\{T \in \Sigma_{\mathbf{d}}\}}}{q(T)} \right], \tag{3.2}$$

so if we are interested in estimating $|\Sigma_{\mathbf{d}}|$, we may use the estimator

$$\widehat{|\Sigma_{\mathbf{d}}|} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}. \tag{3.3}$$

The efficiency of the above estimators may be quantified in several ways. The standard error of $\hat{\mu}$ can be estimated by either repeatedly running the procedure or using an approximation based on the Δ-method:

$$\mathrm{se}(\hat{\mu}) \approx \sqrt{\frac{\mathrm{var}_q \left(\frac{f(T)p(T)}{q(T)} - \mu \frac{p(T)}{q(T)} \right)}{N}}. \tag{3.4}$$

The *effective sample size*, $\mathrm{ESS} = N/(1 + \mathrm{cv}^2)$, is another way to assess method efficiency (Kong et al., 1994). Here, the *coefficient of variation* (cv) is given by

$$\mathrm{cv}^2 = \frac{\mathrm{var}_q(p(T)/q(T))}{E_q^2(p(T)/q(T))}. \tag{3.5}$$

The ESS approximates how many iid samples are equivalent to the N weighted samples obtained through SIS, and the cv^2 is the χ^2 distance between proposal and target. A small cv^2 and a large ESS indicates that we are sampling from a distribution that is close to the desired target uniform

distribution. The theoretical value of cv^2 is unknown, so the sample version is used.

The choice of the proposal $q(\cdot)$ determines the efficiency of the importance sampling procedure. This is a high dimensional problem, so the strategy that will be employed here is to decompose the proposal into lower dimensional components. The first component of the table is sampled, and then the second component of the table is sampled conditional on the realization of the first component. The remainder of the table is sampled sequentially in a similar way conditional on the realization of all previous components.

3.3 Sampling Multigraphs

We are proposing a new SIS technique which samples the table column by column and uses an asymptotic approximation of Bender and Canfield (1978) to guide the sampling.

If we denote the columns of T by $t_1, \ldots, t_n$, then the probability of sampling a table T using a proposal $q(\cdot)$ can be written as

$$q(T = (t_1, \ldots, t_n)) = q(t_1) \times q(t_2|t_1) \times \cdots \times q(t_n|t_{n-1}, \ldots, t_1). \tag{3.6}$$

We begin by sampling the first column of the table, t_1, conditional on $\mathbf{d}$. After t_1 has been sampled, the degree sequence is updated, the first column is removed, and we sample the first column of the remaining $(n-1) \times (n-1)$ subtable. Denote the configuration of the first column by $t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})$, and denote by $\mathbf{d}^{(2)}$ the updated margins of the $(n-1) \times (n-1)$ subtable after the first column has been sampled, i.e.,

$$\mathbf{d}^{(2)} = (d_2 - \alpha_{21}, d_3 - \alpha_{31}, \ldots, d_n - \alpha_{n1}). \tag{3.7}$$

This procedure is repeated until all of the columns have been sampled and a completed table is obtained.

We start by writing the true marginal distribution of t_1 under the uniform distribution over $\Sigma_{\mathbf{d}}$. For a given configuration of the first column, $t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})$, the true marginal distribution of t_1 is

$$p(t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})) = \frac{|\Sigma_{\mathbf{d}^{(2)}}|}{|\Sigma_{\mathbf{d}}|}. \tag{3.8}$$

This expression cannot be calculated directly, but an asymptotic formula for $|\Sigma_{\mathbf{d}}|$ was given by Bender and Canfield (1978). We will use SIS to generate tables and then assign each sampled table an importance weight to correct for the bias incurred by sampling.

The asymptotic approximation that will be employed was obtained by specializing Theorem 1 of Bender and Canfield (1978) to our setting.

Theorem 3.3.1. Given $\mathbf{d} = (d_1, \ldots, d_n)$ and $M = \sum_{i=1}^{n} d_i$,

$$|\Sigma_{\mathbf{d}}| \sim \Delta_{\mathbf{d}} \equiv \frac{f(M)}{\prod_{i=1}^{n} d_i!} \exp\{\boldsymbol{a}(\mathbf{d})\}, \tag{3.9}$$

where $f(M) = M!/[(M/2)!2^{M/2}]$ and $\boldsymbol{a}(\mathbf{d}) = (\sum_i \binom{d_i}{2}/M)^2 - \sum_i \binom{d_i}{2}/M$.

The proof is given in the appendix. This approximation assumes that all marginal sums are bounded above by a constant d^* and that $M \to \infty$.

The proposal used to sample the first column t_1 is based on the approximation in Theorem 3.3.1 and is shown in Proposal 3.1. Denote this method SIS-BC.

Proposal 3.1. The proposal for the first column based on Bender and Canfield (1978) approximation is

$$q(t_1 = (0, \alpha_{21}, \ldots, \alpha_{n2})) \propto \frac{\Delta_{\mathbf{d}^{(2)}}}{\Delta_{\mathbf{d}}} \propto \frac{1}{\prod_{i=2}^{n}(d_i - \alpha_{i1})!} \exp\{\boldsymbol{a}(\mathbf{d}^{(2)})\}, \tag{3.10}$$

where $\boldsymbol{a}(\cdot)$ is defined as in Corollary 3.3.1.

The proof is provided in the Appendix. Although $q(t_1)$ in the above proposal may be sampled directly using enumeration, this is not feasible for larger tables. In these cases, enumeration takes a long time and it is more convenient to sample $q(t_1)$ using the following rejection method. This is the strategy that will be employed in this paper.

1. Generate a configuration of the first column $\mathbf{a} = (a_1, \ldots, a_n)$ from $g(\mathbf{a})$, where $g(\mathbf{a})$ is the uniform distribution over all possible configurations of the first column. This can be done using the procedure described by Holmes and Jones (1996).
2. Generate a $u \sim$ Unif[0,1].

3. Calculate the ratio $q(\mathbf{a})/(cg(\mathbf{a}))$, where $q(\mathbf{a})$ is the proposal of SIS and c is a constant chosen so that $q(\mathbf{a}) \leq cg(\mathbf{a})$ for any $\mathbf{a}$.
4. Accept $\mathbf{a}$ if $u \leq q(\mathbf{a})/(cg(\mathbf{a}))$. Otherwise, reject $\mathbf{a}$.

Note that $\Delta_{\mathbf{d}^{(2)}}$ will be obtained for every possible configuration of the first column when the normalizing constant for $q(t_1)$ is calculated, so both the number of configurations of the first column and the maximum value of $\Delta_{\mathbf{d}^{(2)}}$ over these configurations are relatively easy to calculate. These quantities may be used to obtain a value c such that $q(\mathbf{a}) \leq cg(\mathbf{a})$ for all $\mathbf{a}$.

3.3.1 Valid Sampling

While the above procedure will yield reasonable estimates, there will be a certain percentage of tables generated that are invalid. This may occur after some of the columns have been sampled because there is no multigraph that corresponds to the updated degree sequence of the subtable. Consider the following small example. If the margins are $\{2, 2, 2\}$ and the first column sampled is $t_1 = \{0, 2, 0\}$, then the updated margins for the 2×2 subtable are $\{0, 2\}$, and the sampling cannot proceed because this degree sequence does not correspond to a valid multigraph. A sequential importance sampling procedure that guarantees the existence of every table takes into account an existence condition of Hakimi (1962), cited in Meierling and Volkmann (2008), to guarantee that every generated table is valid. This is the procedure that will be used in Section 3.5.

Theorem 3.3.2. Hakimi (1962) A degree sequence $d_n \geq d_{n-1} \geq \cdots \geq d_1$ is multigraphical if and only if $\sum_{i=1}^{n} d_i$ is even and $d_n \leq \sum_{i=1}^{n-1} d_i$.

This condition is incorporated through an additional rejection step. Only those columns that guarantee the existence of a multigraph are sampled so that the sampling method generates 100% valid tables. This approach takes longer to run than sampling without generating valid tables, however, it provides an advantage in terms of cv^2 and standard error, as well as guaranteeing that every generated table will be valid.

3.4 MCMC method

Sampling and testing multigraphs may also be performed using an MCMC procedure based on the Diaconis and Gangolli (1995) method for sampling contingency tables. At each step, two rows i_1 and i_2 are chosen from $\{1, \ldots, n\}$, and two columns j_1 and j_2 are chosen from $\{1, \ldots, n\}$, where $j_1 \neq i_1, i_2$ and $j_2 \neq i_1, i_2$. One of the following two moves is made with equal probability on the four cells at the intersection of rows i_1 and i_2 and columns j_1 and j_2:

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \qquad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix} .$$

The same move is then made on the cells opposite the ones sampled to maintain the symmetry constraint. More specifically, the move is performed on both cells $(i_1, j_1), (i_2, j_1), (i_2, j_2), (i_1, j_2)$, and cells $(j_1, i_1), (j_1, i_2), (j_2, i_2), (j_2, i_2)$. If a negative entry is obtained, the new table is rejected and the Markov chain stays at the current table.

Theorem 3.4.1. Choosing two rows i_1 and i_2 from $\{1, \ldots, n\}$, and two columns j_1 and j_2 from $\{1, \ldots, n\}$, where $j_1 \neq i_1, i_2$ and $j_2 \neq i_1, i_2$, and performing $\begin{smallmatrix} -1 & +1 \\ +1 & -1 \end{smallmatrix}$ or $\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}$ with equal probability and the corresponding move on the cells opposite the diagonal constitutes an irreducible Markov Chain on $\Sigma_{\mathbf{d}}$.

The proof is given in the appendix and follows Diaconis and Gangolli (1995). This method is relatively easy to implement and also allows sampling of larger and denser tables relative to SIS-BC. However, the chain is sticky and it may take a long time to explore the space.

3.5 Applications and Simulations

We illustrate the efficacy of the methods by describing a number of applications and simulations. For SIS-BC, the refined sampling procedure is used in all cases. Computation was performed on a MacBook Pro with a 2.2 GHz Intel Core i7 processor. Coding was done in C with calculation of statistics performed in R.

3.5.1 Estimating the number of multigraphs

Calculating the total number of multigraphs with a prescribed, fixed degree sequence is difficult. An exhaustive search is feasible for very small tables, but will take a prohibitively long time for tables that are even moderately large. Using an SIS strategy, we may estimate $|\Sigma_{\mathbf{d}}|$ using (4.4), based on iid samples from our asymptotically-guided proposal distribution.

We estimate the number of tables in a few examples. We consider a 9×9 table with all margins equal to 4, a 14×14 table with all margins equal to 2, a 26×26 table with all margins equal to 5, a 30×30 table with all margins equal to 3, and a real 15×15 table of chimpanzee grooming behavior (Sugiyama, 1969). To further test the method we also consider a 20×20 table with moderately rough margins equal to $\{15, 5, 5, 5, 5, 5, 5, 5, 5, 5, 1, \ldots, 1\}$, a large and sparse 100×100 table with all margins equal to 2, and an extremely large and sparse 200×200 table with margins equal to $\{3, 1, \ldots, 1\}$. The simulation results, along with the exact number of multigraphs calculated by McKay and McLeod (2012) when they are available are given in Table 3.1. Estimates are based on $1,000$ samples and the number following the $\pm$ sign denotes the standard error calculated using the Δ-method (4.5). Simulation results indicate that SIS-BC is performing well in the task of estimating the total number of multigraphs, producing accurate estimates in a relatively short amount of time. There are exactly $170,816,680$ 9×9 tables with margins equal to 4, $10,157,945,044$ 14×14 tables with all margins equal to 2, 1.2836×10^{56} 26×26 tables with all margins equal to 5, and 1.5998×10^{45} 30×30 tables with all margins equal to 3. The exact number of tables for the remaining examples were not feasible to calculate. We conclude that the method is working even for tables that are large and sparse. Sampling using SIS-BC without guaranteeing validity for the 9×9 table with all margins equal to 4 yields an estimate of $(1.7173 \pm 0.0378) \times 10^8$ with $\text{cv}^2 = 0.4853$ and 73.2% valid samples in 0.2 seconds. For the 30×30 table with all margins equal to 3, an estimate of $(1.5994 \pm 0.0197) \times 10^{45}$ is obtained with $\text{cv}^2 = 0.1521$ and 89.5% valid samples in 23.9 seconds.

3.5.2 Primate social network data

The use of network analysis tools to answer research questions related to the social interactions of animals is currently a growing area of study (Lusseau and Newman, 2004; Croft et al., 2004; Sundaresan et al., 2007; Croft et al., 2008). Here, we will consider grooming data collected by

Table 3.1: Performance of SIS-BC for estimating the number of tables

Table	Estimated # tables	cv^2	Time (sec)
9×9 with margins $= 4$	$(1.7468 \pm 0.0199) \times 10^8$	0.1297	0.2
14×14 with margins $= 2$	$(1.0205 \pm 0.0058) \times 10^{10}$	0.0247	0.2
15×15 chimpanzee data	$(1.0089 \pm 0.0543) \times 10^{25}$	2.8990	3.3
20×20 with rough margins	$(1.0813 \pm 0.0079) \times 10^{20}$	0.0538	56.5
26×26 with margins $= 5$	$(1.2839 \pm 0.0108) \times 10^{56}$	0.0703	386.7
30×30 with margins $= 3$	$(1.5890 \pm 0.0080) \times 10^{45}$	0.0253	26.1
50×50 with margins $= 3$	$(7.4774 \pm 0.0355) \times 10^{91}$	0.0225	350.1
100×100 with margins $= 2$	$(4.1248 \pm 0.0571) \times 10^{56}$	0.0191	$\approx$1hr
200×200 with margins $= \{3, 1, \ldots, 1\}$	$(2.1984 \pm 0.0059) \times 10^{188}$	0.0007	$\approx$1hr

Sugiyama (1969) in the Budongo Forest in Uganda. This data, pictured in Figure 3.2, represents a symmetrized version of grooming interactions among fifteen chimpanzees. The counts represent the number of grooming interactions between a pair of chimpanzees and the labels are the chimpanzee names. While the original data was directed, we summed across the diagonal to obtain a symmetric matrix to represent sociopositive interactions between pairs of chimpanzees (Kasper and Voelkl, 2009). The diagonal is zero since chimpanzees are not able to groom themselves. These data are used to assess the property of group cohesiveness, which was considered in the context of primate data in Lehmann and Boesch (2009). This will be quantified using the average of the weighted clustering coefficients for each node as defined by Barrat et al. (2004),

$$C_w = \frac{1}{N} \sum_i c_i^w, \tag{3.11}$$

where the weighted clustering coefficient for node i is

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}. \tag{3.12}$$

Here $a_{ij} = 1$ if there is an edge between nodes i and j and zero otherwise, w_{ij} is the number of edges between nodes i and j, $s_i = \sum_{j=1}^N a_{ij} w_{ij}$ and $k_i = \sum_j a_{ij}$.

High values of C_w indicate a large degree of overall clustering in the network. The null hypothesis

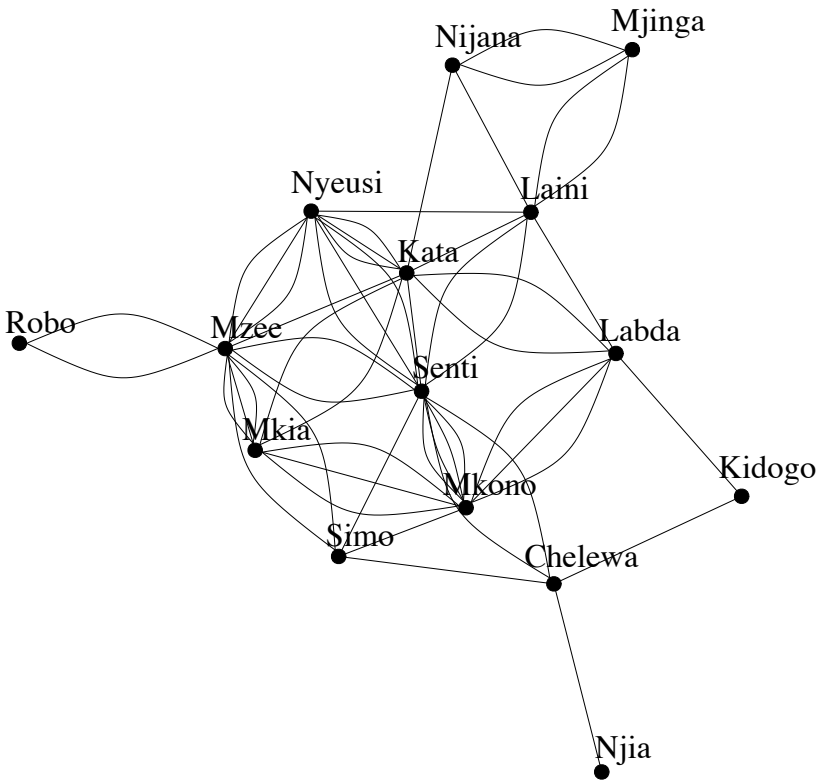


Figure 3.2: The fifteen node multigraph of chimpanzee grooming relations (Sugiyama, 1969)

is that C_w is not unusual when the observed network is considered to be a uniform draw from the set of networks with the same degree. We obtain a p-value of 0.5737 ± 0.02797 with $\text{cv}^2 = 2.8990$ in a few seconds, indicating no statistical significance. Using the alternative MCMC method for these data with $1,000,000$ iterations and $100,000$ burn-in, with standard error calculated using the batch means method results in a p-value of 0.5562 ± 0.00108. Although this standard error is smaller than the one obtained using SIS-BC, it appears to be an underestimate. Running the MCMC procedure 100 times with a different SIS-BC generated starting position each time yields a standard error of 0.03820. It appears that the chain is sticky and takes a long time to explore the space, resulting in an underestimate of the standard error.

3.5.3 Airline network resilience

The airline network is an important aspect of the national transportation system, responsible for moving millions of passengers every year according to the Bureau of Transportation Statistics (2015). History has shown that the airline network is vulnerable to disruption by both targeted attacks and random events. For example, the terrorist attacks in 2001, the eruption of Eyjafjallajökull in 2010, and the United technical glitches in July 2015 all resulted in delays and grounded flights. These consequences impose huge costs on both passengers and the airline industry.

We use sequential importance sampling to examine the resilience of an airline network to a targeted attack, where resilience refers to the ability of the network to maintain short weighted paths between nodes in response to the removal of an important airport. Nodes in this network represent airports and the number of edges between the nodes represents the number of flights between two airports for the month of December 2010 Csardi (2014). For simplicity and computational efficiency, we examine only the flights of PSA Airlines, a regional airline headquartered in Ohio. There are 68 total airports in the network and 135 edges.

To measure the resilience of the network we use the average of the closeness centrality values for all of the nodes Opsahl et al. (2010). The closeness centrality for a node i is the sum of the inverse of the shortest weighted paths between all other nodes and node i, i.e.,

$$C(i) = \sum_{j:j\neq i} \frac{1}{d(i,j)}, \tag{3.13}$$

where $d(i,j)$ represents the shortest weighted path between nodes i and j. Because two airports may be considered to be 'close' if they have a large number of flights between them, we define the distance as

$$d(i,j) = \frac{1}{w_{ii_2}} + \frac{1}{w_{i_2i_3}} + \cdots + \frac{1}{w_{i_kj}}, \tag{3.14}$$

where $\{i, i_2, i_3, \ldots, i_k, j\}$ are the nodes on the shortest weighted path between nodes i and j, and $w_{i_{k-1}i_k}$ is the number of edges between nodes i_{k-1} and i_k.

The overall closeness of the network is the average of the closeness values for all nodes (i.e., $\sum_{i=1}^{n} C(i)/n$). We are interested in determining if the airline network is less resilient to targeted attacks than would be expected by chance. Eliminating the airport with the second largest degree

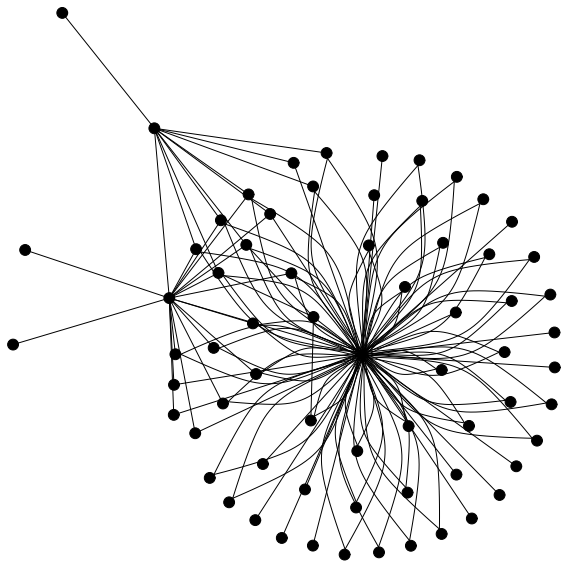


Figure 3.3: The PSA Airlines network. Nodes represent airports and each edge represents a flight

causes the closeness value to decrease by 2.93%, indicating that it is harder to traverse the network following the removal of the airport with the second largest degree. To test the significance of this change, we generated $1,000$ random graphs with the same degree sequence using SIS-BC, eliminated the node corresponding to the airport with the second largest degree and calculated the percent change in closeness. Based on these samples, the probability of seeing a 2.93% or more decrease in average closeness is 0.04427 ± 0.01889, indicating that the airline network formed in such a way that it is less resilient to attacks than we would expect by chance. Computation was completed in under an hour and a cv^2 of 9.4476 was obtained.

3.6 Discussion

We have developed an SIS strategy for sampling multigraphs with fixed degree based on an asymptotic approximation of Bender and Canfield (1978). This method samples column by column and performs best in cases where the graph is at least moderately sparse. As the graph becomes denser, performance decreases as judged by cv^2. We have also proposed an MCMC method based on the moves described by Diaconis and Gangolli (1995) for sampling contingency tables. This method

performs well even in cases where the graph is extremely dense. The methods we have proposed are extremely flexible, as the distribution of any test statistic of interest related to multigraphs may be approximated and a p-value estimated.

The approximation of Bender and Canfield (1978) may be used to approximate the number of graphs where a set of entries in the adjacency matrix are forced to be structural zeros. This additional feature of the approximation may be leveraged to allow for cell by cell sampling of the adjacency matrix of a multigraph with fixed degree. A cell is first sampled and then forced to be a structural zero. While this approach is appealing, in practice it tends to perform poorly and column by column sampling is preferred.

3.7 Proofs of the Main Results

Proof of Theorem 3.3.1

Denote by $(m_{ij})_{n\times n}$ an $n \times n$ symmetric 0-1 matrix, where $m_{ij} = 0$ denotes a structural zero at position (i, j). Let $\Delta_{\mathbf{d}}$ be the number of $n \times n$ symmetric matrices over $[0, t]$ such that $g_{ij} = 0$ whenever $m_{ij} = 0$ and $\sum_j g_{ij} = d_i$. According to Theorem 1 of Bender and Canfield (1978),

$$\Sigma_{\mathbf{d}} \sim \Delta_{\mathbf{d}} \equiv T(M, \delta)\exp\{\epsilon a - b\}/\prod_{j=1}^{n} d_j!,$$

where $M = \sum_{j=1}^{n} d_j$, $\delta = \sum_{m_{ii}=0} d_i$, $\epsilon = 1$ if $t > 1$ and $\epsilon = -1$ if $t = 1$, $a = \left(\sum_{j=1}^{n} \binom{d_j}{2}/M\right)^2$, $b = \left(\sum_{i<j, m_{ij}=0} d_i d_j + \sum_{j=1}^{n} \binom{d_j}{2}\right)/M$, $T(M, \delta) = \sum_j \binom{M-\delta}{j} C_{M-j}$, and $C_j = j!/((j/2)!2^{j/2})$ if j is even and 0 if j is odd.

In the case of multigraphs with no self-loops, the diagonal is zero and we have $\epsilon = 1$, $a = (\sum_{j=1}^{n} \binom{d_j}{2}/M)^2$, $b = \sum_{j=1}^{n} \binom{d_j}{2}/M$, $\delta = M$, and also $T(M, \delta) = M!/((M/2)!2^{M/2})$. Plugging these values in yields the expression in Corollary 3.3.1.

Proof of Proposal 3.1

The approximation of Bender and Canfield (1978) implies that the number of multigraphs after sampling the first column is approximately

$$|\Sigma_{\mathbf{d^{(2)}}}| \sim \Delta_{\mathbf{d^{(2)}}} \equiv \frac{f(M-2d_1)}{\prod_{i=2}^{n}(d_i-\alpha_{i1})!}\exp\{\boldsymbol{a}(\mathbf{d^{(2)}})\}, \tag{3.15}$$

and the approximation to the total number of multigraphs $|\Sigma_{\mathbf{d}}|$ is given in Corollary 3.3.1. Combining the two expressions above yields the proposal SIS-BC:

$$q(t_1=(0,\alpha_{21},\ldots,\alpha_{n2})) \propto \frac{\Delta_{\mathbf{d}^{(2)}}}{\Delta_{\mathbf{d}}} \propto \frac{1}{\prod_{i=2}^{n}(d_i-\alpha_{i1})!}\exp\{\boldsymbol{a}(\mathbf{d}^{(2)})\},$$

where $\boldsymbol{a}(\cdot)$ is as in Corollary 3.3.1.

Proof of Theorem 3.4.1

We need to show that for every $A, B \in \Sigma_{\mathbf{d}}$, there is a sequence of moves of type

$$\begin{array}{cc} +1 & -1 \\ -1 & +1 \end{array} \qquad\qquad \begin{array}{cc} -1 & +1 \\ +1 & -1 \end{array}$$

leading from A to B. We will use induction.

First define $d(A,B)=\sum_{i,j}|a_{ij}-b_{ij}|$ and note that $d(A,B)$ is divisible by 4 and has minimum nonzero value equal to 8.

Assume the induction hypothesis, that if $0 \le d(A,B) \le 4k$ there is a path joining A and B.

This is true for $d(A,B)=8$ because then there are only 8 elements (4 on either side of the diagonal) for which $|a_{ij}-b_{ij}|=1$. Call these cells (i_1,j_1), (i_2,j_1),(i_1,j_2),(i_2,j_2) and (j_1,i_1), (j_2,i_1), (j_1,i_2), (j_2,i_2), where $i_1<i_2$ and $j_1<j_2$. Then we can make an appropriate move to decrease $d(A,B)$ by 8 and obtain $A=B$.

Next, suppose $d(A,B)=4(k+1)$. We will show there is a move from $A \to A'$ where $d(A',B) \le 4k$ or a move $B \to B'$ where $d(A,B') \le 4k$.

Suppose A and B have different elements in the first column (otherwise we can remove the first column and first row and check the second column). Suppose $a_{i_1 1}$ is the first element at which A and B differ and that $a_{i_1 1} < b_{i_1 1}$. Then there exists an i_2 such that $a_{i_2 1} > b_{i_2 1}$ and a j_2 such that $a_{i_1 j_2} > b_{i_1 j_2}$, where $i_2 > i_1$ since $a_{i_1 1}$ is the first element at which A and B differ and $j_2 \neq i_1$ since $a_{i_1 i_1}$ is a structural zero.

There are two cases. The first is that $i_2 \neq j_2$ and the second is that $i_2 = j_2$. In both of these

cases we will show that there is a move $A \to A'$ where $d(A', B) \leq 4k$ or a move $B \to B'$ where $d(A, B') \leq 4k$.

In the first case where $i_2 \neq j_2$ make the move

$$\begin{aligned} a'_{i_1 1} &= a_{i_1 1} + 1 \quad a'_{i_1 j_2} = a_{i_1 j_2} - 1 \\ a'_{i_2 1} &= a_{i_2 1} - 1 \quad a'_{i_2 j_2} = a_{i_2 j_2} + 1 \end{aligned}$$

and the corresponding symmetric move

$$\begin{aligned} a'_{1 i_1} &= a_{1 i_1} + 1 \quad a'_{j_2 i_1} = a_{j_2 i_1} - 1 \\ a'_{1 i_2} &= a_{1 i_2} - 1 \quad a'_{j_2 i_2} = a_{j_2 i_2} + 1 \end{aligned} .$$

We know $a_{i_2 1}, a_{1 i_2}, a_{i_1 j_2}, a_{j_2, i_1} > 0$ since $a_{i_2 1} > b_{i_2 1}$ and $a_{i_1 j_2} > b_{i_1 j_2}$. Since $i_2 \neq j_2$ there are no structural zeros. Moving from $A \to A'$ results in a decrease in the difference of A' with respect to B of 6 on $(i_1, 1)$, (i_1, j_2), $(i_2, 1)$, $(1, i_1)$, (j_2, i_1), $(1, i_2)$. The difference on (i_2, j_2) and (j_2, i_2) may increase by 2, but the net change is at least 4.

So $d(A', B) \leq d(A, B) - 4 \leq 4(k + 1) - 4 = 4k$.

In the second case, $i_2 = j_2$. Here (i_2, i_2) is a structural zero, so we cannot make any move with rows i_1 and i_2 and columns 1 and i_2.

However, there exists j'_2 such that $a_{i_2 j'_2} < b_{i_2 j'_2}$ where $j'_2 \neq 1$ and $j'_2 \neq i_1$ because $a_{i_2 1} > b_{i_2 1}$ and $a_{i_2 i_1} > b_{i_2 i_1}$.

Make the below move on B

$$\begin{aligned} b'_{i_1 1} &= b_{i_1 1} - 1 \quad b'_{i_1 j'_2} = b_{i_1 j'_2} + 1 \\ b'_{i_2 1} &= b_{i_2 1} + 1 \quad b'_{i_2 j'_2} = b_{i_2 j'_2} - 1 \end{aligned}$$

and the corresponding symmetric move

$$\begin{aligned} b'_{1 i_1} &= b_{1 i_1} - 1 \quad b'_{j'_2 i_1} = b_{j'_2 i_1} + 1 \\ b'_{1 i_2} &= b_{1 i_2} + 1 \quad b'_{j'_2 i_2} = b_{j'_2 i_2} - 1 \end{aligned} .$$

Moving from $B \to B'$ results in $d(A, B') \leq d(A, B) - 4 \leq 4k$.

The case where $a_{i_1 1} > b_{i_1 1}$ is symmetric, simply reverse roles of A and B.

Chapter 4

Sampling High Dimensional Tables with Applications to Assessing Linkage Disequilibrium

4.1 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, which will be used to assess linkage disequilibrium in multimarker genetic data. 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 multiway contingency tables was provided in Dobra and Fienberg (2009), and a general method for evaluating the number of tables fulfilling a general set of constraints was provided in Barvinok (1994). MCMC methods based on Diaconis and Sturmfels (1998) 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. Chen, Dinwoodie and Sullivant (2006) 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 tables with fixed one way margins and when the table is large and sparse. Lazzeroni and Lange (1997) 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 (1976) 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 chapter is organized in the following way. Section 4.2 introduces the basics of SIS, with the proposal distribution for SIS based on the approximation of Good (1976) developed in Section 4.3. Section 4.4 describes the problem of assessing linkage disequilibrium in multimarker genetic data when there are more than two alleles at each marker. Section 4.5 demonstrates applications, including estimating the number of tables $|\Sigma|$ and the volume test for assessing linkage disequilibrium. Section 4.6 provides concluding remarks.

4.2 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 high dimensional tables with one way marginal sums

$$n_j^{[i_j]} = \sum_{i_l : l \neq j} t_{i_1 \ldots i_{j-1} i_j i_{j+1} \ldots i_k}, \text{ for } j = 1, \ldots, k, \text{and } i_j = 1, \ldots, I_j, \tag{4.1}$$

where $\sum_{i=1}^{I_1} n_1^{[i]} = \cdots = \sum_{i=1}^{I_k} n_k^{[i]} = M$, and M is the overall table sum. 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 $\pi(T) = 1/|\Sigma|$ be the uniform distribution over Σ.

We are interested in sampling uniformly from Σ. This is a difficult problem, but if a high dimensional table, T, can be sampled from a proposal distribution, $q(\cdot)$, that is easy to sample from and includes the support of Σ, then $E_\pi[f(T)]$ can be estimated using the weighted average,

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(T)(\pi(T_i)/q(T_i))}{\sum_{i=1}^{N} (\pi(T_i)/q(T_i))} = \frac{\sum_{i=1}^{N} f(T)(\mathbb{1}_{\{T_i \in \Sigma\}}/q(T_i))}{\sum_{i=1}^{N} (\mathbb{1}_{\{T_i \in \Sigma\}}/q(T_i))}, \tag{4.2}$$

where $T_1, \ldots, T_N$ are independent, identically distributed (iid) samples from $q(T)$.

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

$$|\Sigma| = \sum_{T \in \Sigma} \frac{1}{q(T)} q(T) = E_q \left[\frac{\mathbb{1}_{\{T_i \in \Sigma\}}}{q(T)} \right], \tag{4.3}$$

and estimated using

$$\widehat{|\Sigma|} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{\{T_i\in\Sigma\}}}{q(T_i)}. \tag{4.4}$$

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

$$\mathrm{se}(\hat{\mu}) \approx \sqrt{\frac{\mathrm{var}_q(\frac{f(T)\pi(T)}{q(T)}) + \mu^2\mathrm{var}_q(\frac{\pi(T)}{q(T)}) - 2\mu\mathrm{cov}_q(\frac{f(T)\pi(T)}{q(T)}, \frac{\pi(T)}{q(T)})}{N}}, \tag{4.5}$$

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

$$\mathrm{cv}^2 = \frac{\mathrm{var}_q(\pi(T)/q(T))}{E_q^2(\pi(T)/q(T))}. \tag{4.6}$$

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.

4.3 Sampling High Dimensional Tables

This 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,

$$q(T) = q(t_{11\ldots1})q(t_{21\ldots1}|t_{11\ldots1})\ldots q(t_{I_1I_2\ldots I_k}|t_{11\ldots1},\ldots,t_{(I_1-1)I_2\ldots I_k}). \tag{4.7}$$

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 $\mathbf{n}_j^* = \{n_j^{*[1]}, \ldots, n_j^{[I_j]}\}$ 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 $\alpha_{11\ldots1}$,

$$p(t_{11\ldots1} = a_{11\ldots1}) = \frac{|\Sigma^*|}{|\Sigma|}, \tag{4.8}$$

where Σ^* denotes the number of tables with marginals $\{\mathbf{n}_1^*, \ldots, \mathbf{n}_k^*\}$ and a structural zero in the first cell. Both the numerator and denominator of this expression are difficult to calculate, but Good (1976) provided an approximation for high dimensional tables with fixed one way margins.

Good's Approximation. *(Good, 1976)*

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

$$|\Sigma| \approx \Delta^G \equiv \frac{\prod_{i_1=1}^{I_1} \binom{n_1^{[i_1]} + \mathcal{I}^{[-1]} - 1}{n_1^{[i_1]}} \prod_{i_2=1}^{I_2} \binom{n_2^{[i_2]} + \mathcal{I}^{[-2]} - 1}{n_2^{[i_2]}} \cdots \prod_{i_k=1}^{I_K} \binom{n_k^{[i_k]} + \mathcal{I}^{[-k]} - 1}{n_k^{[i_k]}}}{\binom{M + I_1 I_2 \times \cdots \times I_{k-1} I_k - 1}{M}^{k-1}} \tag{4.9}$$

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 in Eisinger and Chen (2016), we will leverage this approximation to construct our proposal distribution, denoted SIS-G. The derivation of this proposal is provided in the appendix.

Proposal 4.1. The proposal, constructed based on the approximation of Good (1976) to $|\Sigma|$, is

$$q(t_{11\ldots1} = a_{11\ldots1}) \propto \frac{\prod_{j=1}^{k} \binom{n_j^{[1]} - a_{11\ldots1} + \mathcal{I}^{[-j]} - 2}{n_j^{[1]} - a_{11\ldots1}}}{\binom{M - a_{11\ldots1} + I_1 \times \cdots \times I_k - 2}{M - a_{11\ldots1}}^{k-1}}. \tag{4.10}$$

The first cell will be sampled according to this density using multinomial sampling and 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 (4.10) 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.

4.3.1 Calculation of Bounds

Sampling by cell requires us to calculate a set of entries to sample from that includes the support of the true marginal distribution of the first cell. This can be a difficult and computationally intensive problem, and for some high dimensional tables, the support of the cell may not be an interval. 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 instead calculate the true lower and upper bound, sample an integer between these two values and guarantee 100% valid entries.

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 while sequentially sampling, and we must resort to more computationally intensive methods. Additionally, in situations where we are sampling from a larger set of values for each cell than the true values that will yield a valid high dimensional table, we may generate an invalid table.

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 justifies generating some invalid tables.

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 Fienberg (1999), Kwerel (1988), Warmuth (1988), and Rüschendorf (1991), and reproduced below for cell $(i_1, i_2, \ldots, i_k)$,

$$\max\left(0, \sum_{j=1}^{k} n_j^{[i_j]} - (k-1)M\right) \leq t_{i_1 \ldots i_k} \leq \min\left(n_1^{[i_1]}, n_2^{[i_2]}, \ldots, n_{k-1}^{[i_{k-1}]}, n_k^{[i_k]}\right). \tag{4.11}$$

These bounds need to be extended to the case where a sequence of cells have 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

$$\max\left(0, \sum_{j=1}^{k} n_j^{*[i_j]} - (k-1)M^*\right) \leq t_{i_1 \ldots i_k} \leq \min\left(n_1^{*[i_1]}, \ldots, n_k^{*[il]}\right). \tag{4.12}$$

These bounds are denoted $[l_f, u_f]$, 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 bounds for $t_{i_1 \ldots i_k}$,

$$[l,u] = \begin{cases} [l_f, u_f], & \text{if } i_z \neq I_z \text{ for all } z = 1, \ldots, k, \\ [\max\left(0, n_k^{*[i_k]} - \sum_{j \neq k} \sum_{i'_j = i_j + 1}^{I_j} n_j^{*[i'_j]}\right), u_f] & \text{if } i_z = I_z \text{ for any } z = 1, \ldots, k. \end{cases} \tag{4.13}$$

These bounds 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 (4.13) are generally 2 to 3 times as efficient as competing methods. Unless otherwise stated, the bounds in (4.13) will be used for sampling and the percentage of invalid tables will be reported as necessary.

4.4 Linkage Disequilibrium

Linkage disequilibrium refers to the association between random variables whose realizations represent alleles at different loci on a chromosome. 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 (Pritchard and Przeworski, 2001; Chen, Lin and Sabatti, 2006).

Chen, Lin and Sabatti (2006) extended methods for assessing linkage disequilibrium in biallelic markers to the multiallelic marker case using volume measures. This paper will extend this method further to encompass the case where there are more than two multiallelic markers.

The basic idea of volume measures 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 (Sabatti, 2002).

To motivate this problem, we begin with a brief discussion of linkage disequilibrium for two

markers where each marker can take one of two possible alleles, following Chen, Lin and Sabatti (2006). 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. 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 (Chen, Lin and Sabatti, 2006). Volume measures naturally account for sample size, have a simple intuitive interpretation and can be readily applied to this situation (Sabatti, 2002).

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 (Sabatti, 2002). 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.

Mvol was defined for pairs of markers in Chen, Lin and Sabatti (2006), and can be readily extended to the case where there are more than two markers. *Mvol* is

$$Mvol(S) = \frac{1}{|\Sigma|} \sum_{T \in \Sigma} \mathbb{1}_{\{M(T) < M(S)\}}, \tag{4.14}$$

where

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

Assessing *Mvol* using volume measures 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 a sample of tables $T_1, \ldots, T_N$ is taken from Σ instead. We will use our proposal SIS-G to sample tables from a distribution that is close to uniform, and then assign each sampled table an importance weight. Results for assessing linkage disequilibrium for real genetic data is provided in Section 4.5.2.

4.5 Applications and Simulations

4.5.1 Estimating the number of tables

Exhaustively enumerating the exact number of high dimensional tables with fixed one way margins, $|\Sigma|$, is generally not feasible. Using our proposal to generate samples and importance weights yields an effective method to estimate $|\Sigma|$ using (4.4). We estimate this quantity in a few examples.

First, we examine some small tables with equal margins. The first high dimensional table is $3 \times 3 \times 3$ with all marginal sums equal to 3, and the second is a $3 \times 3 \times 3$ table with all marginal sums equal to 20. We also examine a $3 \times 3 \times 5$ table with all marginals equal to $30, 50$. The true number of tables for the first two examples are $22,620$ and $642,635,414,923,248$, respectively, which was calculated using Latte (Barvinok, 1994). The exact number of tables in the final case is not feasible to calculate using LattE.

Additional challenging tables are examined, a $5 \times 5 \times 5 \times 5$ table with margins $\{4,4,3,1,2\}$, $\{4,3,3,2,2\}$, $\{4,3,3,2,2\}$, and $\{1,1,2,4,6\}$, a $3 \times 3 \times 2 \times 3 \times 3 \times 3 \times 2$ table with margins $\{2,4,6\}$, $\{1,2,9\}$, $\{4,8\}$, $\{4,2,6\}$, $\{3,2,7\}$, $\{2,2,8\}$, and $\{1,11\}$, a $2 \times 2 \times 2 \times 3 \times 2 \times 3 \times 3 \times 3$ table with margins $\{7,9\}$, $\{8,8\}$, $\{7,9\}$, $\{4,4,8\}$, $\{5,11\}$, $\{4,4,8\}$, $\{4,4,8\}$, $\{2,3,11\}$, $5 \times 3 \times 3 \times 4 \times 3 \times 2$ 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 table with margins $\{10,10\}$, $\{7,13\}$, $\{12,8\}$, $\{9,11\}$, $\{10,10\}$, $\{8,12\}$, $\{6,7,7\}$, $\{6,14\}$. The true number of tables with these margins are not feasible to calculate.

Table 4.1: Results for estimating the number of high dimensional tables

Estimated number of tables	cv^2	Time (s)
$3 \times 3 \times 3$ table with all margins $= 3$		
$(2.1939 \pm 0.0509) \times 10^4$	0.5384	30.0
$3 \times 3 \times 3$ table with all margins $= 20$		
$(6.3603 \pm 0.1836) \times 10^{14}$	0.8334	45.0
$3 \times 3 \times 5$ table with all margins $= 30, 50$		
$(5.5462 \pm 0.1812) \times 10^{32}$	1.0677	120.0

Table 4.2: Challenging results for estimating the number of high dimensional tables

Estimated number of tables	cv^2	Time (min)
$5 \times 5 \times 5 \times 5$ table		
$(2.4791 \pm 0.1050) \times 10^{17}$	1.7954	7.0
$3 \times 3 \times 2 \times 3 \times 3 \times 3 \times 2$ table		
$(2.5950 \pm 0.1975) \times 10^{13}$	5.7951	20.7
$2 \times 2 \times 2 \times 3 \times 2 \times 3 \times 3 \times 3$ table		
$(1.2895 \pm 0.0667) \times 10^{25}$	2.6732	47.3
$5 \times 3 \times 3 \times 4 \times 3 \times 2$ table		
$(5.5281 \pm 0.2180) \times 10^{15}$	1.5548	66.6
$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 \times 2$ table		
$(2.2217 \pm 0.1133) \times 10^{25}$	2.6002	13.8

The simulation results are based on $1,000$ importance samples and are presented in Table 4.1 and Table 4.2. Computation was done on a MacBook Pro with a 2.2 GHz processor with coding performed in R. The number following the $\pm$ sign denotes the standard error. In these examples, SIS-G gives reasonable estimates of the number of tables in a relatively short amount of time, and the small cv^2 indicates that we are sampling from a distribution that is very close to uniform.

4.5.2 Linkage Disequilibrium

We apply the volume measures described in Section 4.4 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 typed with 85 markers Ophoff et al. (2002). We examine all possible sets of marker triplets for the first ten markers along

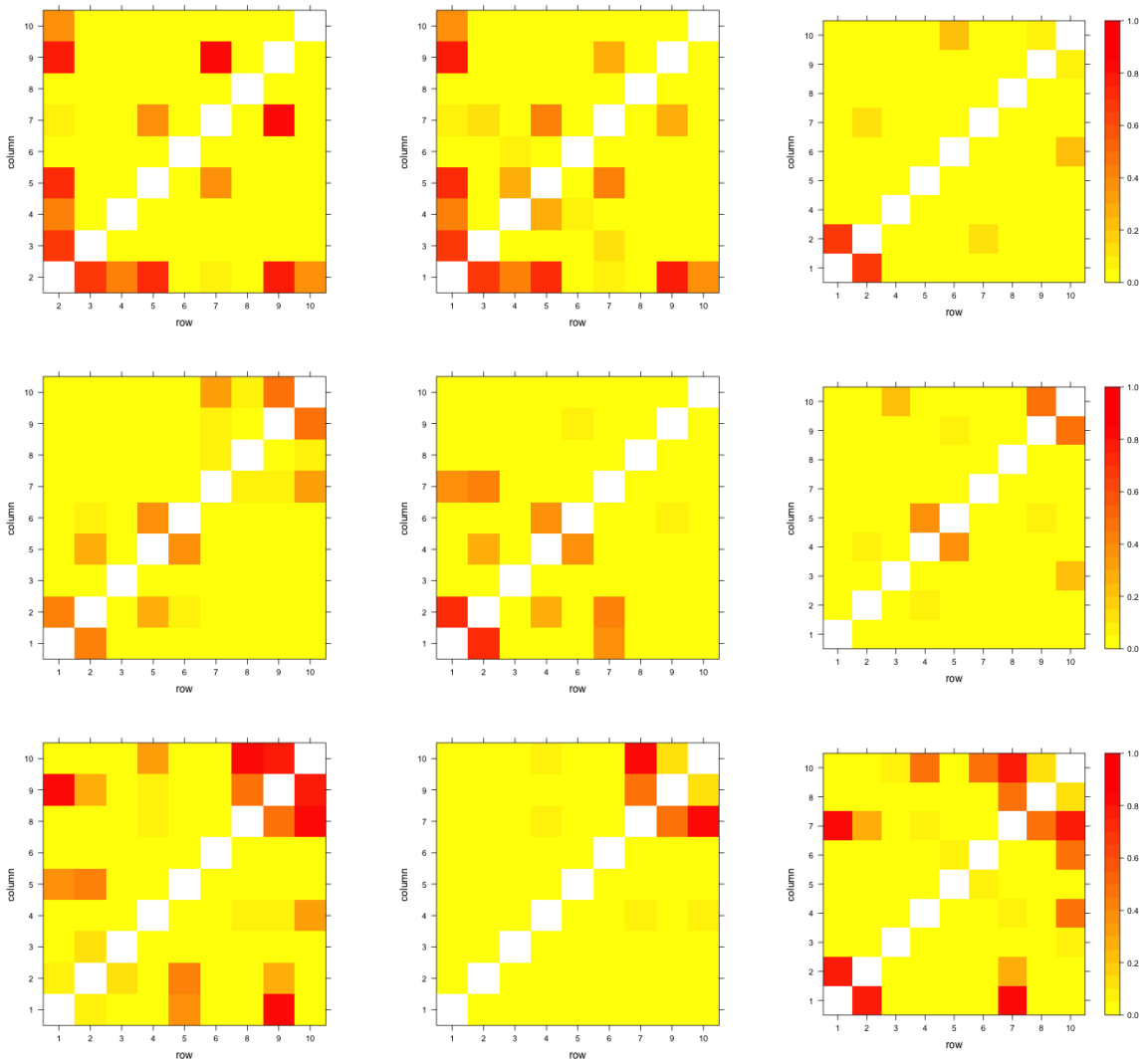


Figure 4.1: Linkage disequilibrium for all marker triplets for first ten markers

the chromosome. For all sets of three markers, $1,000$ importance samples are used. In the course of sampling, no invalid tables were generated. Results are plotted in Figure 4.1.

4.6 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 (1976). 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 obtained by ordering the table in increasing order of marginal sums. This ordering has the advantage of resulting in an extremely small percentage of invalid tables when sampling using $[l, u]$, even for extreme cases. Usually there are no generated invalid tables when the marginal sums are arranged in increasing order. In the case of the genetic data described in Section 4.5.2, there were no invalid tables generated at all. If the columns are arranged in decreasing order, the percentage of invalid tables can be greater than zero, but is generally very small.

Future work may include further analysis of multimarker genetic data from different sources, and also improving the running of the importance sampling algorithm. Only $1,000$ importance samples were used for the analysis in Section 4.5.2 because of limitations on computation time. Taking more samples will allow for a larger effective sample size, which is important as some of the cv^2 values of the marker triplets were moderately large.

4.7 Proofs of the Main Results

Proof of Proposal 4.1

Recall the approximation to $|\Sigma|$ is

$$|\Sigma| \approx \Delta^G \equiv \frac{\prod_{i_1=1}^{I_1} \binom{n_1^{[i_1]} + \mathcal{I}^{[-1]} - 1}{n_1^{[i_1]}} \prod_{i_2=1}^{I_2} \binom{n_2^{[i_2]} + \mathcal{I}^{[-2]} - 1}{n_2^{[i_2]}} \cdots \prod_{i_k=1}^{I_k} \binom{n_k^{[i_k]} + \mathcal{I}^{[-k]} - 1}{n_k^{[i_k]}}}{\binom{M + I_1 I_2 \times \cdots \times I_{k-1} I_k - 1}{M}^{k-1}}. \quad (4.15)$$

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

$$|\Sigma^*| \approx \Delta^{G*} \equiv \frac{\prod_{j=1}^{k} \binom{n_j^{[1]} - a_{11\ldots1} + \mathcal{I}^{[-j]} - 2}{n_j^{[1]} - a_{11\ldots1}} \prod_{j=1}^{k} \prod_{i=2}^{I_j} \binom{n_j^{[i]} + \mathcal{I}^{[-j]} - 1}{n_j^{[i]}}}{\binom{M - a_{11\ldots1} + I_1 \times \cdots \times I_k - 2}{M - a_{11\ldots1}}^{k-1}}. \quad (4.16)$$

Consequently, our proposal for the first cell is

$$q(t_{11\ldots1} = a_{11\ldots1}) \propto \frac{\Delta^{G*}}{\Delta^{G}} \propto \frac{\prod_{j=1}^{k} \binom{n_j^{[1]} - a_{11\ldots1} + \mathcal{I}^{[-j]} - 2}{n_j^{[1]} - a_{11\ldots1}}}{\binom{M - a_{11\ldots1} + I_1 \times \cdots \times I_k - 2}{M - a_{11\ldots1}}^{k-1}}. \tag{4.17}$$

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 $(i_1, \ldots, i_k)$.

If any $i_j = I_j$, the lower bound in (4.12) can be made more strict. Say $i_j = I_j$, and we are interested in calculating bounds for the cell $(i_1, \ldots, I_j, \ldots, i_k)$. 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, $n_1^{*[i_1]}$,

$$\begin{aligned}
n_1^{*[i_1]} = & \sum_{i'_k=i_k+1}^{I_k} \sum_{i'_{k-1}=1}^{I_{k-1}} \cdots \sum_{i'_2=1}^{I_2} t_{i_1 i'_2 \ldots i'_k} \\
& + \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} \sum_{i'_{k-2}=1}^{I_{k-2}} \cdots \sum_{i'_2=1}^{I_2} t_{i_1 i'_2 \ldots i'_{k-1} i_k} \\
& \quad \vdots \\
& + \sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} \sum_{i'_j=1}^{I_j} \cdots \sum_{i'_2=1}^{I_2} t_{i_1 i'_2 \ldots i'_{j+1} i_{j+2} \ldots i_k} \\
& + \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} \sum_{i'_{j-2}=1}^{I_{j-2}} \cdots \sum_{i'_2=1}^{I_2} t_{i_1 i'_2 \ldots i'_{j-1} I_j \ldots i_k} \\
& \quad \vdots \\
& + \sum_{i'_2=i_2+1}^{I_2} t_{i_1 i'_2 i_3 \ldots i_k} \\
& + t_{i_1 \ldots i_k}.
\end{aligned} \tag{4.18}$$

When $i'_k > i_k$, we have

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

so

$$\sum_{i'_2=1}^{I_2} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} t_{i_1 i'_2 \ldots i'_{k-1} i_k} \leq n_k^{*[i_k]}. \tag{4.19}$$

The first component of (4.18) can be bounded above using (4.19),

$$\sum_{i'_k=i_k+1}^{I_k} \sum_{i'_{k-1}=1}^{I_{k-1}} \cdots \sum_{i'_2=1}^{I_2} t_{i_1 i'_2 \ldots i'_k} \leq \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]}. \tag{4.20}$$

Employing a very similar approach for the remaining terms of (4.18) yields

$$n_1^{*[i_1]} \leq \sum_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]} + \cdots + \sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} + \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} t_{i'_1 i'_2 \ldots i'_{k-1} i_k} + \cdots + \sum_{i'_2=i_2+1}^{I_2} n_2^{*[i'_2]} + t_{i_1 \ldots i_k}.$$

So a lower bound for $t_{i_1 \ldots i_k}$ is

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

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,

$$\begin{array}{l}
n_1^{*[i_1]} - \sum\limits_{i'_2=i_2+1}^{I_2} n_2^{*[i'_2]} - \ldots - \sum\limits_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} - \sum\limits_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \ldots - \sum\limits_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]} \\
n_2^{*[i_1]} - \sum\limits_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \ldots - \sum\limits_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} - \sum\limits_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \ldots - \sum\limits_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]} \\
\vdots \\
n_{j-1}^{*[j_1]} - \sum\limits_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \ldots - \sum\limits_{i'_{j-2}=i_{j-1}+1}^{I_{j-2}} n_{j-2}^{*[i'_{j-2}]} - \sum\limits_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \ldots - \sum\limits_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]} \\
n_{j+1}^{*[i_{j+1}]} - \sum\limits_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \ldots - \sum\limits_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} - \sum\limits_{i'_{j+2}=i_{j+1}+1}^{I_{j+2}} n_{j+2}^{*[i'_{j+2}]} - \ldots - \sum\limits_{i'_k=i_k+1}^{I_k} n_k^{*[i'_k]}
\end{array}$$

$$\vdots$$

$$n_k^{*[i_k]} - \sum_{i'_1=i_1+1}^{I_1} n_1^{*[i'_1]} - \ldots - \sum_{i'_{j-1}=i_{j-1}+1}^{I_{j-1}} n_{j-1}^{*[i'_{j-1}]} - \sum_{i'_{j+1}=i_{j+1}+1}^{I_{j+1}} n_{j+1}^{*[i'_{j+1}]} - \ldots - \sum_{i'_{k-1}=i_{k-1}+1}^{I_{k-1}} n_{k-1}^{*[i'_{k-1}]},$$

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 final of these bounds 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

$$n_k^{*[i'_k]} - \sum_{m\neq j,k} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]} \geq n_z^{*[i'_z]} - \sum_{m\neq j,z} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}. \tag{4.21}$$

So when $i_j = I_j$ 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 $i_k = I_k$, 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 $i_p = I_p$ for more than one p, the bound will be the same since summing over any dimension in which $i_p = I_p$, will not contribute anything to the summation $\sum_{m\neq j,k} \sum_{i'_m=i_m+1}^{I_m} n_m^{*[i'_m]}$. Call this new bound l.

Next, we show that $l \geq l_f$, so when $i_j = I_j$, the Fréchet bound is not as strict as l. Recall

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

Since $\sum_{i'_z=i_z+1}^{I_z} n_z^{*[i'_z]} \leq M^*$ and $n_j^{*[i_j]} \leq M^*$ we have

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

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

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

So $l \geq l_f$.

This yields the bounds in (4.13).

Chapter 5

Conclusion

In this thesis, we have developed methods for sampling contingency tables with fixed marginal sums, undirected, loopless multigraphs, and high dimensional tables with fixed one way margins. These methods have been applied to a number of examples. In particular, we have illustrated these approaches with data from biology, ecology and genetics and have demonstrated excellent performance.

One area of future work is additional applications of these methods to the sciences. Network data is becoming more and more prevalent, and tools for conditional inference on networks are needed to answer research questions and to develop our understanding of scientific phenomena. Tables under different constraints are also of interest to researchers, and exhaustive enumeration is often not feasible. Examples include two way tables under the assumption of quasi-symmetry or quasi-independence. These tools are useful for analyzing marriage preferences between ethnicities and the relations between social classes of fathers and sons. Sampling tables under these constraints is difficult, and importance sampling with a carefully chosen proposal distribution can be employed to quickly generate tables and estimate a p-value.

For higher dimensional tables, marginal constraints other than the one examined in this thesis are of interest to researchers and importance sampling methods provide a possible tool to conduct conditional inference (Chen, Dinwoodie and Sullivant, 2006). Statistical methods that quickly provide answers to computationally difficult research questions deserve future study, and the examples reported in this thesis support further research into this topic.

References

Agresti, A. (1992*a*), *Categorical Data Analysis*, New York: Wiley.

Agresti, A. (1992*b*), "A survey of exact inference for contingency tables," *Statistical Science*, 7, 131–177.

Almeida-Neto, M., and Ulrich, W. (2011), "A straightforward computational approach for measuring nestedness using quantitative matrices," *Environmental Modelling and Software*, 26, 173–178.

Andrews, D. F., and Herzberg, A. M. (1985), "Relationships between birthday and death day," in *Data*, New York: Springer-Verlag.

Aoki, S., and Takemura, A. (2005), "Markov Chain Monte Carlo Exact Tests for Incomplete Two-Way Contingency Tables," *Journal of Statistical Computation and Simulation*, 75, 787–812.

Balmer, D. W. (1988), "Algorithm AS 236: Recursive enumeration of r x c tables for exact likelihood evaluation," *Journal of the Royal Statsitical Society, Series C*, 37, 290–301.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004), "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, 101, 3747–3752.

Barrett, S. C. H., and Helenurm, K. (1987), "The reproductive-biology of boreal forest herbs. 1. Breeding systems and pollination," *Canadian Journal of Botany*, 65, 2036–2046.

Bartomeus, I., Vilà, M., and Santamaria, L. (2008), "Contrasting effects of invasive plants in plant-pollinator networks," *Oecologia*, 155, 761–770.

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.

Barvinok, A. I., and Hartigan, J. A. (2010), "Maximum entropy Gaussian approximations for the number of integer points and volumes of polytopes," *Advances in Applied Mathematics*, 45, 252–289.

Bascompte, J., Jordano, P., Melían, C. J., and Olesen, J. M. (2003), "The nested assembly of plant-animal mutualistic networks," *Proceedings of the National Academy of Sciences*, 100, 9383–9387.

Bascompte, J., Jordano, P., and Olesen, J. M. (2006), "Asymmetric coevolutionary networks facilitate biodiversity maintenance," *Science*, 312, 431–433.

Bascompte, J., and Pedro, J. (2006), "The structure of plant-animal mutualistic networks," in *Ecological Networks*, Oxford,US: Oxford University Press.

Bayati, M., Kim, J. H., and Saberi, A. (2010), "A sequential algorithm for generating random graphs," *Algorithmica*, 58, 860–910.

Békéssy, A., Békéssy, P., and Kómlos, J. (1972), "Asymptotic enumeration of regular matrices," *Studia Scientarium Mathematicarum Hungarica*, 7, 343–353.

Bender, E. A. (1974), "The asymptotic number of non-negative integer matrices with given row and column sums," *Discrete Mathematics*, 10, 217–223.

Bender, E. A., and Canfield, E. R. (1978), "The asymptotic number of labeled graphs with given degree sequences," *Journal of Combinatorial Theory, Series A*, 25, 296–307.

Blitzstein, J., and Diaconis, P. (2010), "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Mathematics*, 6, 489–522.

Blüthen, N., Menzel, F., Hovestadt, T., Fiala, B., and Blüthen, N. (2007), "Specialization, constraints, and conflicting interests in mutualistic networks," *Current Biology*, 17, 341–346.

Bureau of Transportation Statistics (2015), "Airline Activity: National Summary (U.S. Flights),", http://www.transtats.bts.gov/.

Burgos, E., Deva, H., Perazzo, R. P. et al. (2007), "Why nestedness in mutualistic networks?," *Journal of Theoretical Biology*, 249, 307–313.

C Kasper, B. V. (2009), "A social network analysis of primate groups," *Primates*, 50, 343–356.

Canfield, E. R., and McKay, B. D. (2010), "Asymptotic enumeration of integer matrices with large equal row and column sums," *Combinatorica*, 30, 655–680.

Chen, Y. (2007), "Conditional inference on tables with structural zeros," *Journal of Computational and Graphical Statistics*, 16, 445–467.

Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2005), "Sequential Monte Carlo methods for statistical analysis of tables," *Journal of the American Statistical Association*, 100, 109–120.

Chen, Y., Dinwoodie, I. H., and Sullivant, S. (2006), "Sequential importance sampling for multiway tables," *The Annals of Statistics*, 34, 523–545.

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

Cochran, W. G. (1952), "The χ^2 test of goodness of fit," *The Annals of Mathematical Statistics*, 23, 315–345.

Conti, E., Cao, S., and Thomas, A. (2013), "Disruptions in the U.S. Airport Network," , ArXiv e-prints.

Croft, D. P., James, R., and Krause, J. (2004), "Social networks in the guppy (*Poecilia reticulata*)," *Proceedings of the Royal Society B*, 271, S516–S519.

Croft, D. P., James, R., and Krause, J. (2008), *Exploring Animal Social Networks*, Princeton, New Jersey: Princeton University Press.

Csardi, G. (2014), *igraphdata: a collection of network data sets for the igraph package.* R package version 0.2.

Devlin, B., and Risch, N. (1995), "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics*, 29, 311–322.

Diaconis, P., and Efron, B. (1985), "Testing for independence in a two-way table: new interpretations of the chi-square statistic," *The Annals of Statistics*, 1, 834–874.

Diaconis, P., and Gangolli, A. (1995), "Rectangular arrays with fixed margins," in *Discrete Probability and Algorithms*, eds. D. Aldous, P. Diaconis, J. Spencer, and J. Steele, New York: Springer-Verlag.

Diaconis, P., and Sturmfels, B. (1998), "Algebraic algorithms for sampling from conditional distributions," *The Annals of Statistics*, 26, 363–397.

Dobra, A., and Fienberg, S. E. (2009), "The generalized shuttle algorithm," in *Algebraic and Geometric Methods in Statistics*, eds. P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, Cambridge: Cambridge University Press, pp. 395–407.

Dormann, C. F., Fründ, J., Blüthgen, N., and Gruber, B. (2009), "Indices, graphs and null models: analyzing bipartite ecological networks," *The Open Ecology Journal*, 2, 7–24.

Dyer, M. (2003), Approximate Counting by Dynamic Programming,, in *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC '03, ACM, New York, NY, USA, pp. 693–699.

Eisinger, R., and Chen, Y. (2016), "Sampling for conditional inference on contingency tables," *Journal of Computational and Graphical Statistics*, 25. In press.

Everett, C. J., and Stein, P. R. (1970), "The asymptotic number of integer stochastic matrices," *Discrete Mathematics*, 1, 55–72.

Feller, W. (1957), *An introduction to probability theory and its applications*, New York: Wiley.

Fienberg, S. E. (1980), *The analysis of cross-classified data (2nd ed.)*, Cambridge, MA: M.I.T. Press.

Fienberg, S. E. (1999), Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation,, in *Statistical Data Protection Proceedings*, SDP'98, Eurostat, pp. 115–129.

Gail, M., and Mantel, N. (1977), "Counting the number of contingency tables with fixed margins," *Journal of the American Statistical Association*, 72, 859–862.

Good, I. J. (1976), "On the application of symmetric Dirichlet distributions and their mixtures to contingency tables," *The Annals of Statistics*, 4, 1159–1189.

Good, I. J., and Crook, J. F. (1977), "The enumeration of arrays and a generalization related to contingency tables," *Discrete Mathematics*, 19, 23–45.

Gotelli, N. J., and Graves, G. R. (1996), *Null Models in Ecology*, Washington, D.C.: Smithsonian Institution Press.

Greenhill, C., and McKay, B. D. (2008), "Asymptotic enumeration of sparse nonnegative integer matrices with specified row and column sums," *Advances in Applied Mathematics*, 41, 459–481.

Guimarâes, P. R., and Raimundo, R. L. (2012), "Interaction web Database,".

Hakimi, S. L. (1962), "On realizability of a set of integers as degrees of optimally edge-connected multigraphs," *Journal of the Society for Industrial and Applied Mathematics*, 10, 496–506.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), "statnet: software tools for the representation, visualization, analysis and simulation of network data," *Journal of Statistical Software*, 24, 1–11.

Holmes, R. B., and Jones, L. K. (1996), "On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron," *The Annals of Statistics*, 24, 64–68.

Hossain, M., Alam, S., Rees, T., and Abbass, H. (2013), Australian airport network robustness analysis: a complex network approach,, in *Australasian Transport Research Forum 2013 Proceedings*, The Planning and Transport Research Centre.

Hotelling, H. (1939), "Tubes and spheres in n-spaces, and a class of statistical problems," *American Journal of Mathematics*, 61, 440–460.

Jones, L. K., and O'Neil, P. J. (1999), "Contingency Tables: Diaconis-Efron Conditional Volume Test," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz, D. L. Banks, and C. B. Read, New York: Wiley, pp. 123–125.

Jordano, P. (1987), "Patterns of mutualistic interactions in pollination and seed dispersal: conductance, dependence asymmetry, and coevolution," *American Naturalist*, 129, 657–677.

Karp, R. M., Luby, M., and Madras, N. (1989), "Monte-Carlo approximation algorithms for enumeration problems," *Journal of Algorithms*, 10, 429–448.

Kasper, C., and Voelkl, B. (2009), "A social network analysis of primate groups," *Primates*, 50, 343–356.

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential imputations and Bayesian missing data problems," *Journal of the American Statistical Association*, 89, 278–288.

Kwerel, S. M. (1988), "Fréchet Bounds," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz, and N. L. Johnson, Vol. 3, New York: Wiley, pp. 202–209.

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.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, New York: Wiley.

Lehmann, L., and Boesch, C. (2009), "Sociality of the dispersing sex: the nature of social bonds in West African female chimpanzees, Pan trogolodytes," *Animal Behavior*, 77, 377–387.

Liu, J. S., and Chen, R. (1998), "Sequential Monte Carlo methods for dynamic systems," *Journal of the American Statistical Association*, 93, 1032–1044.

Lusseau, D., and Newman, M. E. J. (2004), "The emergent properties of a dolphin social network," *Proceedings of the Royal Society B*, 270, S186–S188.

McDonald, J. W., Smith, P. W. F., and Forster, J. J. (2007), "Markov Chain Monte Carlo exact inference for social networks," *Social Networks*, 29, 127–136.

McKay, B. D., and McLeod, J. C. (2012), "Asymptotic enumeration of symmetric integer matrices with uniform row sums," *Journal of the Australian Mathematical Society*, 92, 367–384.

Meierling, D., and Volkmann, L. (2008), "A remark on the degree sequences of multigraphs," *Mathematical Methods of Operations Research*, 69, 369–374.

Milo, R., Shen-Orr, S., Itzkovitz, S. et al. (2002), "Network motifs: simple building blocks of complex networks," *Science*, 298, 824–827.

Mirsky, L. (1971), *Transversal Theory*, New York: Academic Press.

Mosquin, T., and Martin, J. E. H. (1967), "Observations on the pollination biology of plants on Melville Island, N.W.T., Canada," *Canadian Field Naturalist*, 81, 201–205.

Newman, M. E. J. (2001), "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical Review E*, 64, 016132.

O'Neil, P. E. (1969), "Asymptotics and random matrices with row sum and column sum restrictions," *Bulletin of the American Mathematical Society*, 75, 1276–1283.

Ophoff, R., Escamilla, M., Service, S. et al. (2002), "Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate," *American Journal of Human Genetics*, 71, 565–574.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010), "Node centrality in weighed networks: Generalizing degree and shortest paths," *Social Networks*, 32, 245–251.

Opsahl, T., and Panzarasa, P. (2009), "Clustering in weighted networks," *Social Networks*, 31, 155–163.

Pawar, S. (2014), "Why are plant-pollinator networks nested?," *Science*, 345, 383.

Ploog, D. W. (1967), "The behavior of squirrel monkeys (*Saimiri sciureus*) as revealed by sociometry, bioacoustics, and brain stimulation," in *Social Communication Among Primates*, ed. S. Altmann, Chicago: University of Chicago Press.

Pritchard, J. K., and Przeworski, M. (2001), "Linkage disequilibrium in humans: models and data," *American Journal of Human Genetics*, 69, 1–14.

Roberts, J. M. (2000), "Simple methods for simulation sociomatrices with given marginal totals," *Social Networks*, 22, 273–283.

Rüschendorf, L. (1991), Bounds for distributions with multivariate margins,, in *Stochastic Order and Decision under Risk*, eds. K. Mosler, and M. Scarsini, Vol. 19 of *IMS Lecture Notes-Monograph Series*, pp. 285–310.

Ryder, T. B., McDonald, D. B., Blake, J. G., Parker, P. G., and Loiselle, B. A. (2008), "Social networks in the lek-mating wire-tailed manakin (*Pipra filicauda*)," *Proceedings of the Royal Society B*, 274, 1367–1374.

Sabatti, C. (2002), "Measuring dependency with volume tests," *The American Statistician*, 56, 191–195.

Sabatti, C., and Risch, N. (2002), "Homozygosity and linkage disequilibrium," *Genetics*, 160, 1707–1719.

Snee, R. D. (1974), "Graphical display of two-way contingency tables," *The American Statistician*, 28, 9–12.

Snijders, T. A. B. (2006), "Enumeration and simulation methods for 0-1 matrices with given marginals," *Psychometrika*, 56, 397–417.

Sugiyama, Y. (1969), "Social behavior of chimpanzees in the Budongo forest, Uganda," *Primates*, 10, 197–225.

Sundaresan, S. R., Fischoff, I. R., Dushoff, J., and Rubenstein, D. I. (2007), "Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager," *Oecologia*, 151, 140–149.

Teare, M., Dunning, A., Durocher, F., Rennart, G., and Easton, D. (2002), "Sampling distribution of summary linkage disequilibrium measures," *Annals of Human Genetics*, 66, 223–233.

Ulrich, W., and Gotelli, N. J. (2007), "Null model analysis of species nestedness patterns," *The Ecological Society of America*, 88, 1824–1831.

Warmuth, W. (1988), "Marginal Fréchet-bounds for multidimensional distribution functions," *Statistics*, 19, 283–294.

Zhang, J., and Chen, Y. (2013), "Sampling for conditional inference on network data," *Journal of the American Statistical Association*, 108, 1295–1307.

Zipunnikov, V., Booth, J. G., and Yoshida, R. (2009), "Counting tables using the double-saddlepoint approximation," *Journal of Computational and Graphical Statistics*, 18, 915–929.