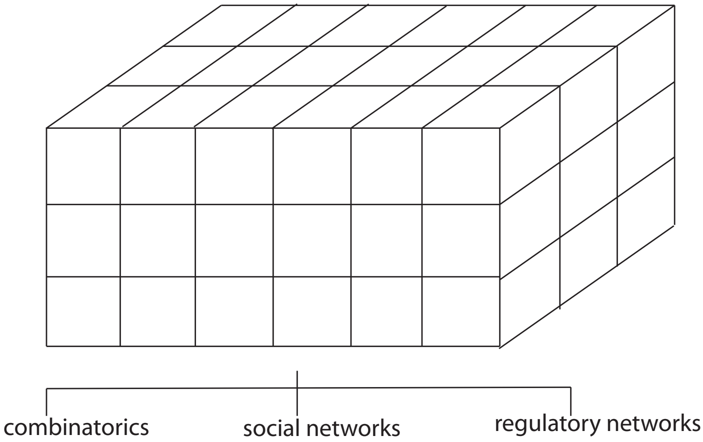
Estimating the Number of Tables via Sequential Importance Sampling

Jing Xi

Department of Statistics
University of Kentucky

Jing Xi[1], Ruriko Yoshida[1], David Haws[1]

Introduction

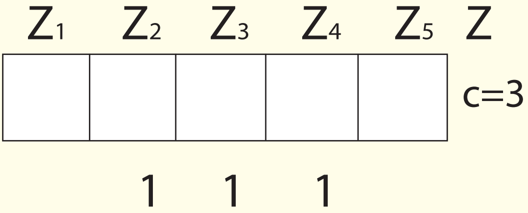


We consider one of the most generally used model: no three way interaction model. ▸ Model

Conditional Poisson Distribution

Let $\mathbf{Z} = (Z_1, \ldots, Z_l)$ be independent Bernoulli trials with probability of successes $p = (p_1, \ldots, p_l)$. Define $w_k = p_k/(1 - p_k)$ for $\forall k$. Then $\mathbf{Z}$ follows a Conditional Poisson Distribution means,

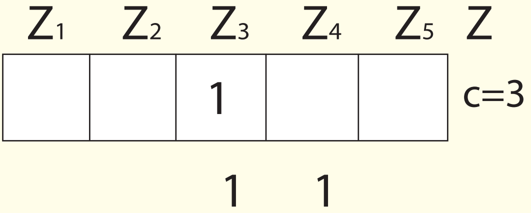
$$P(Z_1 = z_1, \ldots, Z_l = z_l | S_Z = c) \propto \prod_{k=1}^{l} w_k^{z_k}. \tag{1}$$



Conditional Poisson Distribution

Let $\mathbf{Z} = (Z_1, \ldots, Z_l)$ be independent Bernoulli trials with probability of successes $p = (p_1, \ldots, p_l)$. Define $w_k = p_k/(1 - p_k)$ for $\forall k$. Then $\mathbf{Z}$ follows a Conditional Poisson Distribution means,

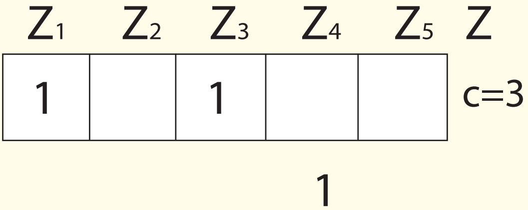
$$P(Z_1 = z_1, \ldots, Z_l = z_l | S_Z = c) \propto \prod_{k=1}^{l} w_k^{z_k}. \tag{1}$$



Conditional Poisson Distribution

Let $\mathbf{Z} = (Z_1, \ldots, Z_l)$ be independent Bernoulli trials with probability of successes $p = (p_1, \ldots, p_l)$. Define $w_k = p_k/(1-p_k)$ for $\forall k$. Then $\mathbf{Z}$ follows a Conditional Poisson Distribution means,

$$P(Z_1 = z_1, \ldots, Z_l = z_l | S_Z = c) \propto \prod_{k=1}^{l} w_k^{z_k}. \quad (1)$$



Conditional Poisson Distribution

Let $\mathbf{Z} = (Z_1, \ldots, Z_l)$ be independent Bernoulli trials with probability of successes $p = (p_1, \ldots, p_l)$. Define $w_k = p_k/(1 - p_k)$ for $\forall k$. Then $\mathbf{Z}$ follows a Conditional Poisson Distribution means,

$$P(Z_1 = z_1, \ldots, Z_l = z_l | S_Z = c) \propto \prod_{k=1}^{l} w_k^{z_k}. \tag{1}$$

Z_1	Z_2	Z_3	Z_4	Z_5	Z
1		1	1		c=3

Conditional Poisson Distribution

Let $\mathbf{Z} = (Z_1, \ldots, Z_l)$ be independent Bernoulli trials with probability of successes $p = (p_1, \ldots, p_l)$. Define $w_k = p_k/(1 - p_k)$ for $\forall k$. Then $\mathbf{Z}$ follows a Conditional Poisson Distribution means,

$$P(Z_1 = z_1, \ldots, Z_l = z_l | S_Z = c) \propto \prod_{k=1}^{l} w_k^{z_k}. \tag{1}$$

Z_1	Z_2	Z_3	Z_4	Z_5	Z
1	0	1	1	0	c=3

Sequential Importance Sampling (SIS)

$\Sigma \neq \emptyset$, the set of all tables satisfying marginal conditions.
$p(\mathbf{X})$: $\Sigma \to [0, 1]$, target distribution, the uniform distribution over Σ, $p(\mathbf{X}) = 1/|\Sigma|$.
$q(\mathbf{X}) > 0$ for all $\mathbf{X} \in \Sigma$, the proposal distribution for sampling.
We have

$$\mathbb{E}\left[\frac{1}{q(\mathbf{X})}\right] = \sum_{\mathbf{X} \in \Sigma} \frac{1}{q(\mathbf{X})} q(\mathbf{X}) = |\Sigma|$$

which can be estimated $|\Sigma|$ by

$$\widehat{|\Sigma|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(\mathbf{X_i})},$$

where $\mathbf{X_1}, \ldots, \mathbf{X_N}$ are tables drawn iid from $q(\mathbf{X})$.

Sequential Importance Sampling (SIS)

How to get the probability of the whole table $\mathbf{X}$?

Denote the columns of the table $\mathbf{X}$ as $x_1, \cdots, x_t$. By the multiplication rule we have

$$q(\mathbf{X} = (x_1, \cdots, x_t)) = q(x_1)q(x_2|x_1)\cdots q(x_t|x_{t-1}, \ldots, x_1).$$

We can easily compute $q(x_i|x_{i-1}, \ldots, x_1)$ for $i = 2, 3, \ldots, t$ using Conditional Poisson distribution.

What if we have rejections?

Having rejections means that $q(\mathbf{X})$: $\Sigma^* \to [0, 1]$ where $\Sigma \subsetneq \Sigma^*$. The SIS estimator is still unbiased and consistent:

$$\mathbb{E}\left[\frac{\mathbb{I}_{\mathbf{X}\in\Sigma}}{q(\mathbf{X})}\right] = \sum_{\mathbf{X}\in\Sigma^*} \frac{\mathbb{I}_{\mathbf{X}\in\Sigma}}{q(\mathbf{X})} q(\mathbf{X}) = |\Sigma|,$$

where $\mathbb{I}_{\mathbf{X}\in\Sigma}$ is an indicator function for the set Σ.

SIS for 2-way Table

Theorem [Chen et. al., 2005]

For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1, \ldots, r_m$ and first column sum c_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = c_1$ with $p_i = r_i/n$.

				2
				1
				3
2				

SIS for 2-way Table

Theorem [Chen et. al., 2005]

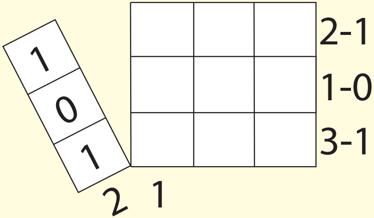
For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1, \ldots, r_m$ and first column sum c_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = c_1$ with $p_i = r_i/n$.

1				2
0				1
1				3
2				

SIS for 2-way Table

Theorem [Chen et. al., 2005]

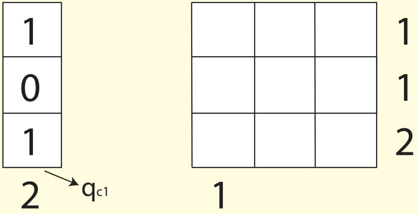
For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1, \ldots, r_m$ and first column sum c_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = c_1$ with $p_i = r_i/n$.



SIS for 2-way Table

Theorem [Chen et. al., 2005]

For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1, \ldots, r_m$ and first column sum c_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = c_1$ with $p_i = r_i/n$.



SIS for 2-way Table

Theorem [Chen et. al., 2005]

For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1, \ldots, r_m$ and first column sum c_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = c_1$ with $p_i = r_i/n$.

1	0
0	1
1	0
2	1

$q_{c1} * q_{c2}$

		1
		1
		2
1		

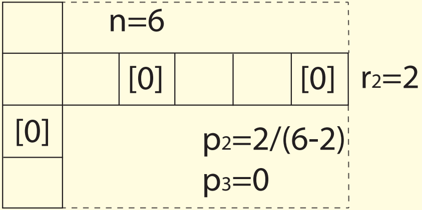
SIS for 2-way Tables with Structural Zero's

A structural zero means a cell in the table that is fixed to be 0.

Example

Theorem [Yuguo Chen, 2007] ["Hand Waving" version]

Key: If $(i, 1)$ is not a structural 0: change $p_i = r_i/n$ to $p_i = r_i/(n - g_i)$ where g_i is the number of structural zeros in the ith row; otherwise $p_i = 0$. Theorem



SIS for 3-way Tables

Theorem ["Hand Waving" version]

For a cell (i_0, j_0, k_0), 3 columns will go through it: $(i_0, j_0, \cdot)$, $(i_0, \cdot, k_0)$, $(\cdot, j_0, k_0)$. Key: When generating $(i_0, j_0, \cdot)$, let $\mathbf{r} = (i_0, \cdot, k_0)$, $\mathbf{c} = (\cdot, j_0, k_0)$, define $r_{k_0} = X_{i_0+k_0}$ and $c_{k_0} = X_{+j_0k_0}$, then set:

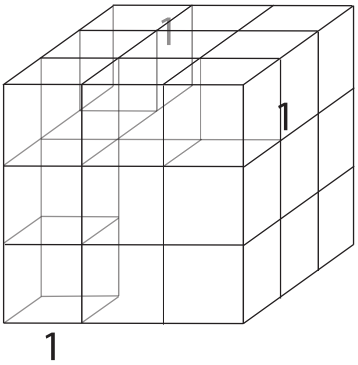
$$p_{k_0} = \frac{r_{k_0} \cdot c_{k_0}}{r_{k_0} \cdot c_{k_0} + (n - r_{k_0} - g^{\mathbf{r}}_{k_0})(m - c_{k_0} - g^{\mathbf{c}}_{k_0})},$$

where $g^{\mathbf{r}}_{k_0}$, and $g^{\mathbf{c}}_{k_0}$ are the numbers of structural zeros in $\mathbf{r}$ and $\mathbf{c}$, respectively. ▸ Theorem

A similar strategy can be used in multi-way tables. ▸ Theorem

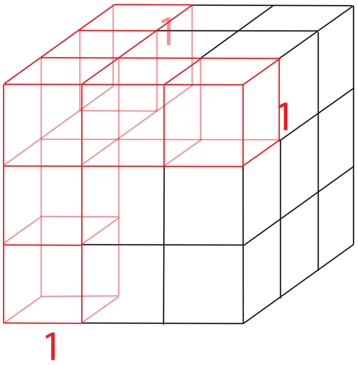
Algorithm

Example: $3 \times 3 \times 3$ Semimagic Cube



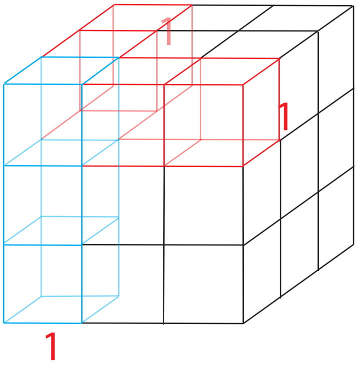
Algorithm

Example: $3 \times 3 \times 3$ Semimagic Cube



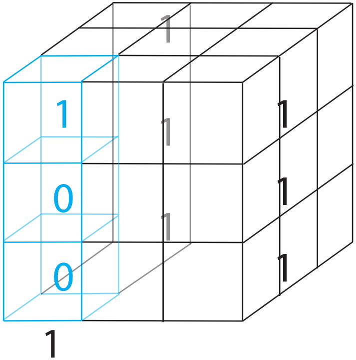
Algorithm

Example: $3 \times 3 \times 3$ Semimagic Cube



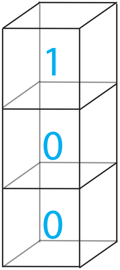
Algorithm

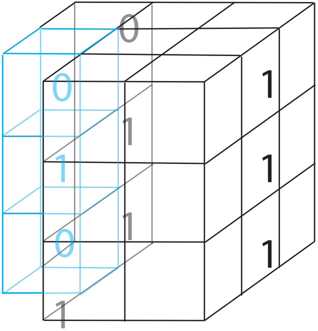
Example: $3 \times 3 \times 3$ Semimagic Cube



Algorithm

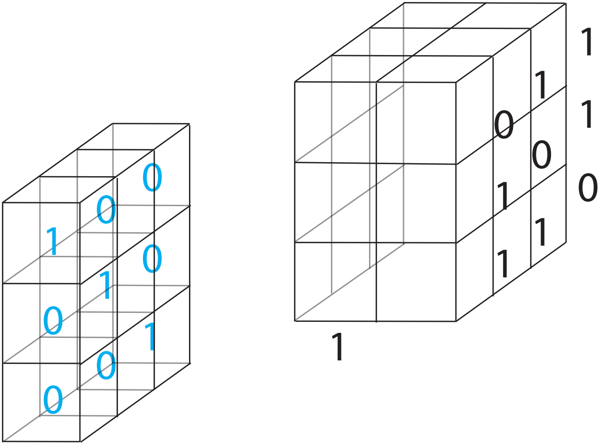
Example: $3 \times 3 \times 3$ Semimagic Cube





Algorithm

Example: $3 \times 3 \times 3$ Semimagic Cube



Simulations - Semimagic Cubes (Table 2)

This tables lists the results from $m \times m \times m$ tables with all marginals equals to 1.

Dim m	$\#$ *tables*	Estimation	cv^2	δ
4	576	571.1472	0.27	100%
5	161280	161439.3	0.18	99.2%
6	812851200	819177227	0.45	98.8%
7	6.14794e+13	6.146227e+13	0.64	97.7%
8	1.08776e+20	1.099627e+20	1.00	96.5%
9	5.52475e+27	5.684428e+27	1.59	95.3%
10	9.98244e+36	9.73486e+36	1.73	93.3%

δ: acceptance rate

Simulations - Semimagic Cubes (Table 3)

We can also change marginal s.

Dimension m	s	Estimation	cv^2	δ
6	3	1.269398e+22	2.83	96.5%
7	3	2.365389e+38	25.33	96.7%
8	3	3.236556e+59	7.05	94.5%
	4	2.448923e+64	11.98	94.3%
9	3	7.871387e+85	15.23	91.6%
	4	2.422237e+97	14.00	93.4%
10	3	6.861123e+117	26.62	90%
	4	3.652694e+137	33.33	93.8%
	5	1.315069e+144	46.2	91.3%

δ: acceptance rate

Experiment - Sampson's Dataset

- It is a dataset about the social interactions among a group of monks recorded by Sampson.
- Data structure:
  - Dimension: $18 \times 18 \times 10$
  - Rows/Columns: the 18 monks.
  - Levels: 10 questions: liking (3 timepoints), disliking, esteem, disesteem, positive influenc, negative influence, praise and blame.
  - Values: answers: 3 top choices were listed in original dataset, ranks were recorded. We set these ranks as an indicator (1 if in top three choices, 0 if not).
- $N = 100000$, estimator is $1.704774e + 117$, $cv^2 = 621.4$, acceptance rate is 3%.

Problems Still Open

Our code performs good when marginals are close to each other. But for the opposite case, the acceptance rate can become very low.

How can we reduce rejection rate?

Possible idea: arrange the order of columns in different ways?

How can Gale-Ryser Theorem be used for 3-way tables?

Thank you!

Questions?

Website for this paper:
`http://arxiv.org/abs/1108.5939`

Model

Let $\mathbf{X} = (X_{ijk})$ of size (m, n, l), where $m, n, l \in \mathbb{N}$ and $\mathbb{N} = \{1, 2, \ldots\}$, be a table of counts whose entries are independent Poisson random variables with parameters, $\{\theta_{ijk}\}$. Here $X_{ijk} \in \{0, 1\}$. Consider the loglinear model,

$$\log(\theta_{ijk}) = \lambda + \lambda_i^M + \lambda_j^N + \lambda_k^L + \lambda_{ij}^{MN} + \lambda_{ik}^{ML} + \lambda_{jk}^{NL} \qquad (2)$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, l$ where M, N, and L denote the nominal-scale factors. This model is called *no three-way interaction model*.

Notice that the sufficient statistics under the model in (2) are the *two-way marginals*.

Back

Example for Structural Zero's

How can structural zero's come?

Different types of cancer separated by gender for Alaska in year 1989:

Type of cancer	Female	Male	Total
Lung	38	90	128
Melanoma	15	15	30
Ovarian	18	[0]	18
Prostate	[0]	111	111
Stomach	0	5	5
Total	71	221	292

The structural zeros's are denoted by "[0]" Back

Theorem [2-way Tables with Structural Zero's]

Define the set of structural zeros Ω as: $\Omega = \{(i,j) : (i,j)$ is a structural zero, $i = 1,\dots,m,\ j = 1,\dots,n\}$

Theorem [Yuguo Chen, 2007]

For the uniform distribution over all $m \times n$ 0-1 tables with given row sums $r_1,\dots,r_m$, first column sum c_1, and the set of structural zeros Ω, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_{\mathbf{Z}} = c_1$ with $p_i = I_{[(i,1)\notin\Omega]} r_i/(n - g_i)$ where g_i is the number of structural zeros in the ith row.

Back

Theorem [SIS for 3-way Tables]

Theorem

For the uniform distribution over all $m \times n \times l$ 0-1 tables with structural zeros with given marginals $r_k = X_{i_0+k}$, $c_k = X_{+j_0k}$ for $k = 1, 2, \ldots, l$, and a fixed marginal for the factor L, l_0, the marginal distribution of the fixed marginal l_0 is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = l_0$ with

$$p_k := \frac{r_k \cdot c_k}{r_k \cdot c_k + (n - r_k - g_k^{r_0})(m - c_k - g_k^{c_0})},$$

where $g_k^{r_0}$ is the number of structural zeros in the (r_0, k)th column and $g_k^{c_0}$ is the number of structural zeros in the (c_0, k)th column.

Back

Theorem [SIS for Multi-way Tables]

Theorem

For the uniform distribution over all d-way 0-1 contingency tables $\mathbf{X} = (X_{i_1 \dots i_d})$ of size $(n_1 \times \cdots \times n_d)$, where $n_i \in \mathbb{N}$ for $i = 1, \dots d$ with marginals $l_0 = X_{i_1^0, \dots i_{d-1}^0 +}$, and $r_k^j = X_{i_1^0 \dots i_{j-1}^0 + i_{j+1}^0 \dots i_{d-1}^0 k}$ for $k \in \{1, \dots, n_d\}$, the marginal distribution of the fixed marginal l_0 is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = l_0$ with

$$p_k := \frac{\prod_{j=1}^{d-1} r_k^j}{\prod_{j=1}^{d-1} r_k^j + \prod_{j=1}^{d-1} (n_j - r_k^j - g_k^j)}$$

where g_k^j is the number of structural zeros in the $(i_1^0, \dots, i_{j-1}^0, i_{j+1}^0, \dots, i_{d-1}^0, k)$th column of $\mathbf{X}$. Back