Sequential Importance Samping

Ruriko Yoshida

1 Preliminaries

Please review loglinear models for contingency tables (you can read a pdf file from `data.princeton.edu/wws509/notes/c5.pdf` or the book *Categorical Data Analysis* by Agresti).

2 Notation

Suppose, for example, we have four factors A, B, C, D for categorical data, each with several levels, the no-3-way interaction model is the loglinear model described with the common notation [A, B], [A, C], [A, D], [B, C], [B, D], [C, D] (see Christensen (1990) for notation and definitions). That is, the sufficient statistics are given by sums of counts that fix all pairs of factors at specified levels. We can generalize this to s factors $F_1, F_2, \ldots, F_s$ with no-L-way interaction, where $L \leq s$. For now we consider no-L-way interaction model for s factors.

3 Sequential Importance Sampling (SIS)

This is the first method we would like to implement. For more details see "Sequential importance sampling for multi-way tables", by Seth Sullivant, Yuguo Chen, and Ian Dinwoodie, Annals of Statistics (2006) 34 No. 1, 523–545 math.ST/0605615 and "Sequential Monte Carlo Methods for Statistical Analysis of Tables" by Yuguo CHEN, Persi DIACONIS, Susan P. HOLMES, and Jun S. LIU, American Statistical Association, March 2005, Vol. 100, No. 469, Theory and Methods DOI 10.1198/016214504000001303.

Let Σ be the set of all tables satisfying marginal conditions. Here we assume that $\Sigma \neq \emptyset$. Let $P(\mathbf{X})$ for any $\mathbf{X} \in \Sigma$ be the uniform distribution over Σ, so $p(\mathbf{X}) = 1/|\Sigma|$. Let $q(\cdot)$ be a trial distribution such that $q(\mathbf{X}) > 0$

for all $\mathbf{X} \in \Sigma$. Then we have

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

Thus we can estimate $|\Sigma|$ by

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

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

Here this proposed distribution $q(\mathbf{X})$ is the distribution (approximate) to sample tables via the SIS.

Here we vectorized the table $\mathbf{X} = (x_1, \cdots, x_t)$. Then by the multiplication rule we have

$$q(\mathbf{X} = (x_1, \cdots, x_t)) = q(x_1)q(x_2|x_1)q(x_3|x_2, x_1) \cdots q(x_t|x_{t-1}, \dots, x_1).$$

Since we sample each cell count of a table from an interval we can easily compute $q(x_i|x_{i-1}, \dots, x_1)$ for $i = 2, 3, \dots, t$.

When we have rejections, this means that we are sampling tables from a bigger set Σ^* such that $\Sigma \subset \Sigma^*$. In this case, as long as the conditional probability $q(x_i|x_{i-1}, \dots, x_1)$ for $i = 2, 3, \dots$ and $q(x_1)$ are normalized, $q(\mathbf{X})$ is normalized over Σ^* since

$$\begin{aligned} \textstyle\sum_{\mathbf{X}\in\Sigma^*} q(\mathbf{X}) &= \textstyle\sum_{x_1,\dots x_t} q(x_1)q(x_2|x_1)q(x_3|x_2, x_1) \cdots q(x_t|x_{t-1}, \dots, x_1) \\ &= \textstyle\sum_{x_1} q(x_1) \left[\sum_{x_2} q(x_1|x_2) \left[\cdots \left[\sum_{x_t} q(x_t|x_{t-1}, \dots, x_1)\right] \cdots\right]\right] \\ &= 1. \end{aligned}$$

Thus we have

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

where $\mathbb{I}_{\mathbf{X}\in\Sigma}$ is an indicator function for the set Σ. By the law of large numbers this estimator is unbiased.

This is pretty easy algorithm to implement. In order to explain I will show you with an example.

Suppose we have an observed 4×5 table $\mathbf{x_0}$ under the independence model (i.e., row and column sums are fixed) such that

3	2	1	0	8
4	2	3	6	1
0	6	8	3	5
7	2	5	1	10

.

Then first we compute the row and column sums $X_{i\cdot}$ and $X_{\cdot j}$, that is

$$\begin{array}{rcl} X_{1\cdot} &=& 14 \\ X_{2\cdot} &=& 16 \\ X_{3\cdot} &=& 22 \\ X_{4\cdot} &=& 25 \\ X_{\cdot 1} &=& 14 \\ X_{\cdot 2} &=& 12 \\ X_{\cdot 3} &=& 17 \\ X_{\cdot 4} &=& 10 \\ X_{\cdot 5} &=& 24 \end{array} .$$

Now we want to fill up the entries

X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

First we pick an integer uniformly from $[\max\{0, X_{1\cdot}+X_{\cdot 1}-X_{\cdot\cdot}\}, \min\{X_{1\cdot}, X_{\cdot 1}\}]$ for X_{11}. For example, say we pick $X_{11} = 5$.

Then we have a table:

5	X_{12}	X_{13}	X_{14}	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

For X_{12} we pick an integer uniformly from $[\max\{0, X_{1\cdot}+X_{\cdot 1}-X_{\cdot\cdot}\}, \min\{X_{1\cdot}-X_{11}, X_{\cdot 2}\}] = [0, \min\{14-5, 12\}]$. For example, say we pick $X_{12} = 2$.

Then we have a table:

5	2	X_{13}	X_{14}	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

For X_{13} we pick an integer uniformly from $[\max\{0, X_{1\cdot}+X_{\cdot 1}-X_{\cdot\cdot}\}, \min\{X_{1\cdot}-X_{11}-X_{12}, X_{\cdot 3}\}] = [0, \min\{14-5-2, 17\}]$. For example, say we pick $X_{13} = 3$.

Then we have a table:

5	2	3	X_{14}	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

For X_{14} we pick an integer uniformly from $[\max\{0, X_{1\cdot}+X_{\cdot 1}-X_{\cdot\cdot}\}, \min\{X_{1\cdot} - X_{11} - X_{12} - X_{13}, X_{\cdot 4}\}] = [0, \min\{14 - 5 - 2 - 3, 10\}]$. For example, say we pick $X_{14} = 3$.

Then we have a table:

5	2	3	3	X_{15}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

.

For X_{15}, we assign $X_{15} = X_{1\cdot} - X_{11} - X_{12} - X_{13} - X_{14} = 14-5-2-3-3 = 1$. Then we have: Then we have a table:

5	2	3	3	1
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

.

Similarly we fill out other entries. Please

Problem 3.1. *Fill out other entries and make a table which satisfies the given row and column sums.*

4 Implementation in R

I want you to start implementing the SIS for contingency tables with 2 factors A, B. For now we assume for the independence model (i.e., row and column sums are fixed). Here is the algorithm.

Algorithm 4.1. *(SIS for two way tables)*

- *INPUT the number of rows m, the number of columns n, the observed table $\mathbf{x_0}$.*
- *OUTPUT A table $\mathbf{X}$ with the same row and column sums with $\mathbf{x_0}$ sampled via SIS.*
- *ALGORITHM*

```
1. Compute the row sums $X_{i\cdot}$ and column sums $X_{\cdot j}$.
2. For $i = (m - 1)$ do:
    (a) For $j = (n - 1)$ do:
            i. Pick integer $x$ uniformly from $[\max\{0, ((X_{i\cdot} - \sum_{k=1}^{j-1} X_{ik}) - \sum_{k=j+1}^{m} (X_{\cdot j} - (\sum_{k=1}^{i-1} X_{kj}))))\}, \min\{X_{i\cdot} - (\sum_{k=1}^{j-1} X_{ik}), X_{\cdot j} - (\sum_{k=1}^{i-1} X_{kj})\}]$, where we define $\sum_{k=1}^{0} X_{ik} = \sum_{k=1}^{0} X_{kj} = 0$.
3. For $i = 1, \ldots, m$ do:
    (a) Set $X_{in} = X_{i\cdot} - \sum_{k=1}^{n-1} X_{ik}$.
4. For $j = 1, \ldots, n$ do:
    (a) Set $X_{mj} = X_{\cdot j} - \sum_{k=1}^{m-1} X_{kj}$.
5. Return $\mathbf{X}$.
```

Remark 4.2. *We will generalize this algorithm to multi-dimensional contingency tables with no-L-way interaction model for s factors. So please implement the code so that it will be easy to generalize it.*

5 Useful functions in R

We will not use it for now but there are some useful functions in R. We might not use them in future but I think it is good for you to know.

A loglinear model for a multiway table of counts can be fit and evaluated many ways. Maximum likelihood fitting and asymptotic measures of goodness-of-fit are available from Poisson regression on a data frame, part of any generalized linear model package such as the one in R (R Development Core Team, 2004). The R command `loglin` also does table fitting, using iterative proportional fitting (IPF), and this is more convenient than Poisson regression when the data is in a multidimensional array. Both these methods rely on χ^2 asymptotics on either the Pearson χ^2 statistic or likelihood ratio statistic for goodness-of-fit. For sparse tables, one often wants exact conditional methods to avoid asymptotic doubts. The basic command `chisq.test` in R has an option for the exact method on two-way tables, usually called Fisher's exact test. For theoretical details of IPF, the Pearson χ^2 statistic and likelihood ratio statistic for goodness-of-fit, see the book *Categorical Data Analysis* by Agresti).