Markov Bases for Two-way Subtable Sum Problems

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

Joint work with H. Hara and A. Takemura

www.ms.uky.edu/~ruriko

Dose-response clinical trial

Drug\Usefulness	$--$	$-$	$\pm$	$+$	$++$	$+++$	Total
Placebo	3	6	37	9	15	1	71
AF3mg	7	4	33	21	10	1	76
AF6mg	5	6	21	16	23	6	77

[C. Hirotsu, 1997]

The purpose of this trial is to find out an optimal dose, where a dose level is considered to be optimal if it significantly improves the efficacy over lower doses ($-$: undesirable, $\pm$: not useful, $+$: useful).

In our model we will consider main effects of two factors. The main effects correspond to rows sums and columns sums. In addition we will consider interaction of two factors with a **certain joint threshold**.

Dose-response clinical trial

We propose that the cell $(2, 4)$ is a threshold.

Drug\Usefulness	$--$	$-$	$\pm$	$+$	$++$	$+++$	Total
Placebo	3	6	37	9	15	1	71
AF3mg	7	4	33	21	10	1	76
AF6mg	5	6	21	16	23	6	77

[C. Hirotsu, 1997]

Under this model, we fix the row, column sums and the sum of cells in blue.

Birthday and death day

Table 1: Relationship between birthday and death day

	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
Jan	1	0	0	0	1	2	0	0	1	0	1	0
Feb	1	0	0	1	0	0	0	0	0	1	0	2
March	1	0	0	0	2	1	0	0	0	0	0	1
April	3	0	2	0	0	0	1	0	1	3	1	1
May	2	1	1	1	1	1	1	1	1	1	1	0
June	2	0	0	0	1	0	0	0	0	0	0	0
July	2	0	2	1	0	0	0	0	1	1	1	2
Aug	0	0	0	3	0	0	1	0	0	1	0	2
Sep	0	0	0	1	1	0	0	0	0	0	1	0
Oct	1	1	0	2	0	0	1	0	0	1	1	0
Nov	0	1	1	1	2	0	0	2	0	1	1	0
Dec	0	1	1	0	0	0	1	0	0	0	0	0

Table 1 shows data gathered to test the hypothesis of association between birth day and death day. The table records the month of birth and death for 82 descendants of Queen Victoria. A widely stated claim is that birthday-death day pairs are associated. Columns represent the month of birth day and rows represent the month of death day.

Drawing tables from the hypergeometric distribution

In the first example, this model is called **the two-way change point model** [Hirotsu, 1997] and in the second example, this model is called the **common diagonal effect model**.

In order to compute the **Exact p-value** under the proposed model, we want to sample tables with given row and column sums and an additional sum of subtable from the the hypergeometric distribution.

Question: How can we generate random draws from this distribution with fixed row sums, column sums, and an additional sum?

Answer: Apply Diaconis-Sturmfels algorithm to the MCMC technique. A key of connectivity of the MC is Diaconis-Sturmfels algorithm which is currently the only known method guaranteed to connect the MC.

A **Markov basis** is a set of **moves** which is guaranteed to connect all feasible contingency tables satisfying the given margins [Diaconis and Sturmfels, 1998].

Question: Finding a Markov basis which connects all feasible 2-way contingency tables satisfying the row sums, column sums, and a sum of a subtable.

Answer: Compute a set of generators for a toric ideal accosiate to the **design matrix** of the tables.

Example

Suppose we have the following table and we want to fix the row and column sums, and the sum of cells in blue.

				Total
	2	2	2	6
	2	2	2	6
total	4	4	4	

Exact p-value computation

Fixing the row sums, column sums, and a sum $\sum_{i=1}^{i_0}\sum_{j=1}^{j_0} x_{ij}^{obs}$, we have

				Total
	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	6
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	6
Total	4	4	4	

Note: There are 5 tables satisfying these margins in this example. We counted using a software **LattE**.

From the constraints we can set up the system of linear equations.

e.g. For our 2×3 table, we have:

$$\begin{array}{cccccccc}
x_{1,1} & & & +x_{2,1} & & & = & 4 \\
& x_{1,2} & & & +x_{2,2} & & = & 4 \\
& & x_{1,3} & & & +x_{2,3} & = & 4 \\
x_{1,1} & +x_{1,2} & +x_{1,3} & & & & = & 6 \\
& & & x_{2,1} & +x_{2,2} & +x_{2,3} & = & 6 \\
x_{1,1} & & & & & & = & 2 \\
& & & & & x_{i,j} & \in & \mathbb{Z}_+
\end{array}$$

where $Z_+ = \{0, 1, 2, \cdots\}$.

In general, we can set up a system $\{x \in \mathbb{Z}_+^d | Ax = b\}$ for any tables.

Note: Thus, moves connect all integral points inside a feasible region $P_b = \{x \in \mathbb{R}^d | Ax = b,\ x \geq 0\} \neq \emptyset$.

What is a Markov Basis??

Suppose $P_b = \{x \in \mathbb{R}^d | Ax = b,\ x \geq 0\} \neq \emptyset$ and let M be a finite set such that $M \subset \{x \in \mathbb{Z}^d | Ax = 0\}$.

We define the graph G_b such that:

- Nodes of G_b are the lattice points inside P_b.
- We draw an undirected edge between a node u and a node v iff $u - v \in M$.

Definition : M is called a **Markov basis** if G_b is a connected graph for all b with $P_b \neq \emptyset$.

Note: A Markov basis is a minimum set of generators of a toric ideal, I_A, associate with A.

Fact: It has been well-known that for two-way contingency tables with fixed row sums and column sums, the set of square-free moves of degree two of the form

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

(**basic moves**) constitutes a Markov basis.

However: If you add a constraint of a sum of a subtable, then it is not necessarily true anymore.

For example, if we fix the subtable $x_{1,1}$ and $x_{2,2}$ then there are only three tables such that

$$\begin{array}{|c|c|c|} \hline 2 & 2 & 2 \\ \hline 2 & 2 & 2 \\ \hline \end{array}, \quad \begin{array}{|c|c|c|} \hline 1 & 1 & 4 \\ \hline 3 & 3 & 0 \\ \hline \end{array}, \quad \begin{array}{|c|c|c|} \hline 3 & 3 & 0 \\ \hline 1 & 1 & 4 \\ \hline \end{array}.$$

and these tables are not connected by basic moves.

Question: **When a set of basic moves forms a Markov basis?** Find the necessary and sufficient condition on a subtable.

Note: A Gröbner basis of a toric idea I_A associate to a matrix A with any term order is a Markov basis associate to a matrix A. So one can compute a Markov basis from a Gröbner basis of I_A with any term order.

Note: There are several nice software to compute Gröbner bases (such as **4ti2**). **However**: Computing a Gröbner basis is very hard to compute. **Thus**, it is nice if we know the necessary and sufficient condition on a subtable that a set of basic moves forms a Markov basis.

Note: A minimal Markov basis associate to a matrix A is not unique in general while the minimal Gröbner basis of I_A with the given term order is unique. **However**, for 2-way tables with fixed row sums, column sums, and a sum of a subtable, A minimal Markov basis associate to a matrix A is unique if a set of basic moves forms a Markov basis.

Notation

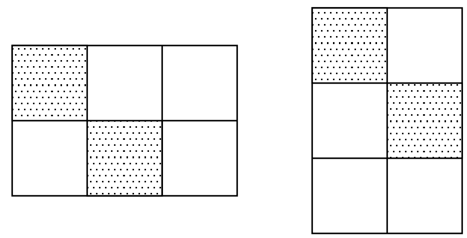
Suppose we have a $R \times C$ table, $X = \{x_{ij}\}$, $x_{ij} \in \mathbb{N}$, $i = 1, \ldots, R$, $j = 1, \ldots, C$.

Let $\mathcal{I} = \{(i, j) \mid 1 \leq i \leq R, 1 \leq j \leq C\}$.

Let S be a subset of $\mathcal{I}$ and S^c is the complement of S.

Necessary and sufficient condition

Here, we give a necessary and sufficient condition on the subtable sum problem so that a Markov basis consists of basic moves.



$\mathcal{P}$ $\mathcal{P}^t$

Figure 1: The pattern $\mathcal{P}$ and $\mathcal{P}^t$

A shaded area shows a cell belonging to S.

We call these two patterns in Figure 1 the pattern $\mathcal{P}$ and $\mathcal{P}^t$, respectively.

Necessary and sufficient condition

Theorem: [Hara, Takemura, Y, 2007]

Let I_S be a toric ideal associate with A for fixing row, column sums and the sum of cells with index in S.

Then I_S is generated by quadratic binomials if and only if there exist no patterns of the form $\mathcal{P}$ or $\mathcal{P}^t$ in any 2×3 and 3×2 subtable of S or S^c after any interchange of rows and columns.

i.e., the set of square-free moves of degree two of the form

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

(**basic moves**) constitutes a Markov basis if and only if there exist no patterns of the form $\mathcal{P}$ or $\mathcal{P}^t$ in any 2×3 and 3×2 subtable of S or S^c after any interchange of rows and columns.

Go back to example....

If we have the first example,

				Total
	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	6
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	6
Total	4	4	4	

then there is no such a subtable in Figure 1 in the subtable of S, thus a set of basic moves forms a Markov basis.

In fact, There is one basic move in the Markov basis.

Go back to example....

Using a software **4ti2**, we found out that the minimum Markov basis consists of one move such that:

$$\begin{array}{|c|c|c|}\hline 0 & -1 & 1 \\ \hline 0 & 1 & -1 \\ \hline \end{array}\,.$$

This move (multiplied by a sign) connects all three tables such that:

$$\begin{array}{|c|c|c|}\hline {\color{blue}2} & 2 & 2 \\ \hline 2 & 2 & 2 \\ \hline \end{array}\,,\ \begin{array}{|c|c|c|}\hline {\color{blue}2} & 3 & 1 \\ \hline 2 & 1 & 3 \\ \hline \end{array}\,,\ \begin{array}{|c|c|c|}\hline {\color{blue}2} & 1 & 3 \\ \hline 2 & 3 & 1 \\ \hline \end{array}\,,\ \begin{array}{|c|c|c|}\hline {\color{blue}2} & 4 & 0 \\ \hline 2 & 0 & 4 \\ \hline \end{array}\,,\ \begin{array}{|c|c|c|}\hline {\color{blue}2} & 0 & 4 \\ \hline 2 & 4 & 0 \\ \hline \end{array}\,.$$

However....

If we have the subtable fixed such that,

				Total
	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	6
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	6
Total	4	4	4	

then, a pattern $\mathcal{P}$ is in the subtable of S. Thus, a set of basic moves does not form a Markov basis.

Using a software **4ti2**, we found out that a minimum Markov basis consists of one move such that:

1	1	-2
-1	-1	2

This move (multiplied by a sign) connects all three tables such that:

$$\begin{array}{|c|c|c|}\hline 2 & 2 & 2 \\ \hline 2 & 2 & 2 \\ \hline \end{array}\,,\quad \begin{array}{|c|c|c|}\hline 1 & 1 & 4 \\ \hline 3 & 3 & 0 \\ \hline \end{array}\,,\quad \begin{array}{|c|c|c|}\hline 3 & 3 & 0 \\ \hline 1 & 1 & 4 \\ \hline \end{array}\,.$$

Updates...

Theorem: [Ohsugi and Hibi (2008)]

The followings are equivalent:

(i) the toric ideal I_S is generated by quadratic binomials;

(ii) the toric ideal I_S possesses a squarefree initial ideal;

(iii) the toric ideal I_S possesses a quadratic Gröbner basis;

(iv) the semigroup ring R_S is normal;

(v) the semigroup ring R_S is Koszul;

(vi) the subset S does not contain pattern $\mathcal{P}$ or $\mathcal{P}^t$.

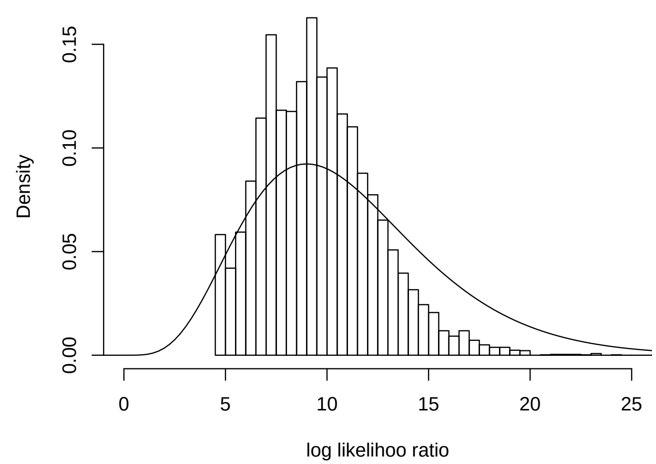
Updates...

Common Diagonal Effect Model: [Hara, Takemura, Y (2008)] We computed a Markov basis for a 2-way table with fixed row and column sums and the diagonal sum.

Relationship between birthday and death day: We now test CDEM against the quasi-independence model. The value of the loglikelihood ratio for the observed table in Table 1 is 6.18839 and the corresponding asymptotic p-value is 0.860503 from the asymptotic distribution χ^2_{11}.

Histogram of sampled tables via MCMC with a Markov basis

We estimated the p-value 0.8934 via MCMC with the Markov Basis computed in this paper. There exists a large discrepancy between the asymptotic distribution and the distribution estimated by MCMC due to the sparsity of the table.



Holes of Semigroup

We also study the difference between the semigroup generated by columns of the design matrix and its saturation.

Theorem [Thomas, Takemura, Y, (2008)]

Let $R, C \in \mathbb{Z}$ be positive integers such that $\min\{R, C\} \geq 2$ and $\max\{R, C\} \geq 3$. The semigroup generated by columns of the design matrix of a $R \times C$ table with fixed row, column sums and the diagonal sum has infinitely many holes.

Thank you....

The paper is available at http://arxiv.org/abs/0708.2312.