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

Oct 12th, 2007

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**.

- After a certain combination of levels we get extra interaction effect or non-effect, not explained by the two factors separately. Namely, the effect of combination of two treatments is larger or smaller than the sum of the effects of two factors.

- This model describes the interaction effect between the row and column factors in terms of a single parameter γ. We want to test for interaction comparing it to "no interaction".

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]

We want to test the goodness-of-fit of this threshold.

Two-Way Change-Point Model

Let $\mathbf{X} = \{X_{ij}\}$ be a $R \times C$ table $X_{ij} \in \mathbb{N}$, $i = 1, \ldots, R$, $j = 1, \ldots, C$.

$$X_{ij} \sim Poi(\mu_{ij}) \text{ iid}$$

where $\mu_{ij} = \ln(\theta_{ij})$.

Consider the generalized linear model with a canonical linear predictor of the form:

$$\theta_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \lambda_{ij}^{RC}.$$

for $i = 1, \ldots, R$ and $j = 1, \ldots, C$.

Two-Way Change-Point model is a special case in which for some (i_0, j_0), $1 \leq i_0 \leq R, 1 \leq j_0 \leq C$

$$\lambda_{ij}^{RC} = \begin{cases} \gamma & \text{if } 1 \leq i \leq i_0, 1 \leq j \leq j_0 \\ 0 & \text{else.} \end{cases}$$

Hypothesis

As a preliminary step, we test:

$$H_0 : \lambda_{ij}^{RC} = 0. \text{ no interaction.}$$

$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

Lack of fit of the hypothesized no interaction model, one of the possiblity is to look for a change point.

$$H_0 : \lambda_{ij}^{RC} = \begin{cases} \gamma & \text{if } 1 \leq i \leq i_0, 1 \leq j \leq j_0 \\ 0 & \text{else.} \end{cases}$$

$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

Want: the χ^2 goodness-of-fit of this threshold.

The sufficient statistics for the Two-Way Change-Point model include the row and column margins and, in addition, the sum of the cell counts with $1 \leq i \leq i_0, 1 \leq j \leq j_0$. Hence, the conditional distribution of the table counts given the margins is the same regardless of the values of the parameters in the model.

In general let $\mathcal{I} = \{(i,j) \mid 1 \leq i \leq R, 1 \leq j \leq C\}$ and let S be a subset of $\mathcal{I}$. Then

$$\lambda_{ij}^{RC} = \begin{cases} \gamma & \text{if } (i,j) \in S \\ 0 & \text{else.} \end{cases}$$

This model is called the **subtable sum model**. Thus the sufficient statistics for subtable sum model include the row and column margins and, in addition, the sum of the cell counts with $(i,j) \in S$.

Hypothesis

As a preliminary step, we test:

$$H_0 : \lambda_{ij}^{RC} = 0. \text{ no interaction.}$$

$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

The fit of independence is not enough, one of the possiblity is to look for a change point.

$$H_0 : \lambda_{ij}^{RC} = \begin{cases} \gamma & \text{if } (i,j) \in S \\ 0 & \text{else.} \end{cases}$$

$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

Exact p-value computation

Let $\hat{\mathbf{X}}$ be the MLE of the data under the model. Then Pearson's χ^2 statistics is

$$f(X) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(\hat{X_{ij}} - X_{ij})^2}{\hat{X_{ij}}}.$$

An exact permutation test based on the χ^2 statistic is constructed as follows. The p-value of this test is:

$$p = E_{\mathbf{p}}[I_{\{f(\mathbf{X}) \geq f(\mathbf{x})\}} | \text{satisfying margins}]$$

where $\mathbf{x}$ is an observed table and $\mathbf{p}$ is the hypergeometric distribution.

In general we approximate the expected value by generating random draws from the hypergeometric distribution and estimate

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{\{f(\mathbf{x}^i) \geq f(\mathbf{x})\}}$$

where N is the number of draws $\mathbf{x}^1, \cdots, \mathbf{x}^N$ iid from the hypergoemetric conditional on the sufficient statistics under H_0.

Note: This is the only possible method in situations where counts are very small or the number of tables satisfying margins is very small.

Question: How can we generate random draws from this distribution?

Answer: Apply Diaconis-Sturmfels algorithm to the MCMC technique. Diaconis-Sturmfels algorithm is the only method guaranteed to connect the MC.

What is a set of **moves** which connect all feasible contingency tables satisfying these margins?

In this particular example, this problem under the two-way change point model is called **the two-way change point problem** [Hirotsu, 1997].

Note: We can generalize this problem by fixing the sum of any subtable in addition to row and column sums.

Question: Finding a set of moves which connect all feasible 2-way contingency tables satisfying the row sums, column sums, and a sum of a subtable.

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

Note that the row sums, column sums, and a sum $\sum_{i=1}^{i_0}\sum_{j=1}^{j_0} x_{ij}^{obs}$ are the sufficient statistics under H_0. For example, 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 $P_b \neq \emptyset$.

Why do we care?: A Markov basis is the only known set of moves which guarantees to connect all tables with any constraints.

Example

To make it simple we just removed a constraint, that is, a sum of colored cells.

				Total
	? ? ?	? ? ?	? ? ?	6
	? ? ?	? ? ?	? ? ?	6
Total	4	4	4	

Table 1: 2×3 tables with 1-marginals.

There are 19 tables satisfying these margins. We counted using a software **LattE**.

$$\pm \begin{array}{|c|c|c|} \hline 1 & -1 & 0 \\ \hline -1 & 1 & 0 \\ \hline \end{array} \qquad \pm \begin{array}{|c|c|c|} \hline 0 & 1 & -1 \\ \hline 0 & -1 & 1 \\ \hline \end{array}$$

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

There are 3 elements in a Markov basis modulo signs.

In fact such moves are called **basic moves**.

4	0	2
0	4	2

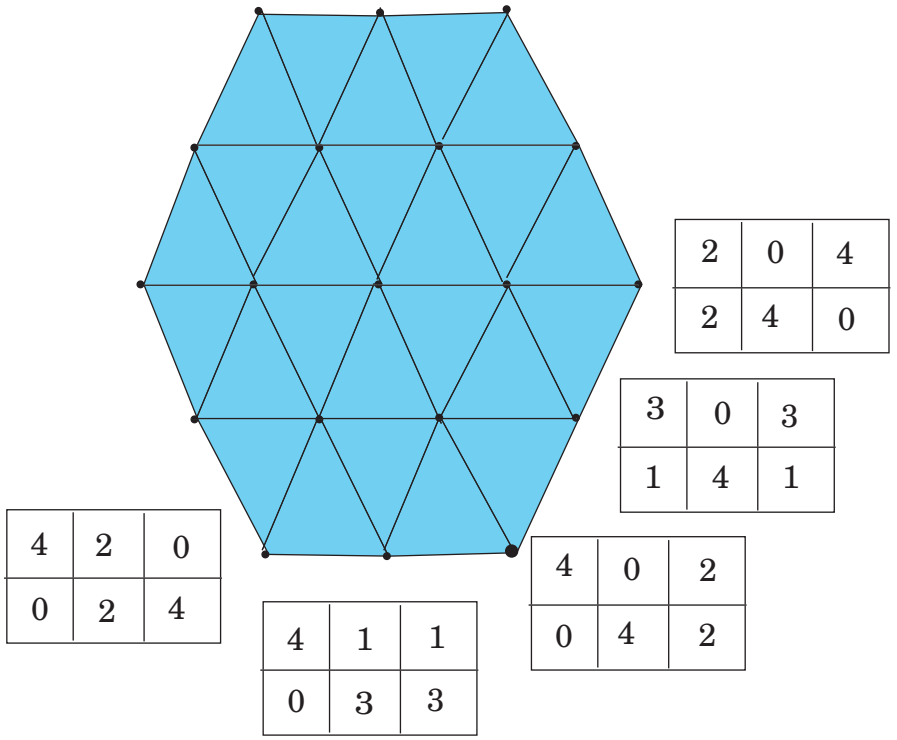
+

-1	0	1
1	0	-1

=

3	0	3
1	4	1

A table with the marginals plus an element of a Markov basis is also a table with the given marginals.



A Markov basis for 2×3 tables. An element of the Markov basis is a undirected edge between integral points in the polytope.

Fact: For any 2-way contingency tables with row and column sums (without a sum of a subtable), we know that a set of basic moves forms 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.

Notes

There always exists a Markov basis for tables.

One can compute a Markov basis using algebraic geometry and there are several nice software to compute a Markov basis, such as **4ti2**.

However: In general computing a Markov basis is very hard. **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 but 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.

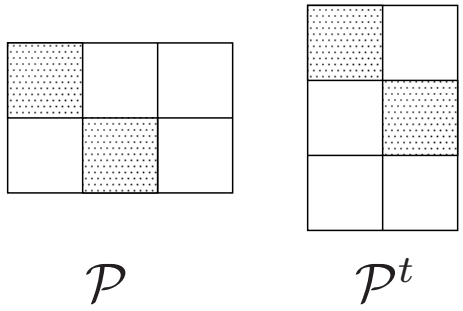


Figure 1: The pattern $\mathcal{P}$ and $\mathcal{P}^t$

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]

A set of basic moves forms 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:

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

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:

2	2	2
2	2	2

,

1	1	4
3	3	0

,

3	3	0
1	1	4

.

Example

Suppose we want to move from the left table to the right table such that:

1	**0**	0	0
0	**0**	0	1
0	0	1	0
0	1	0	0

$\Rightarrow$

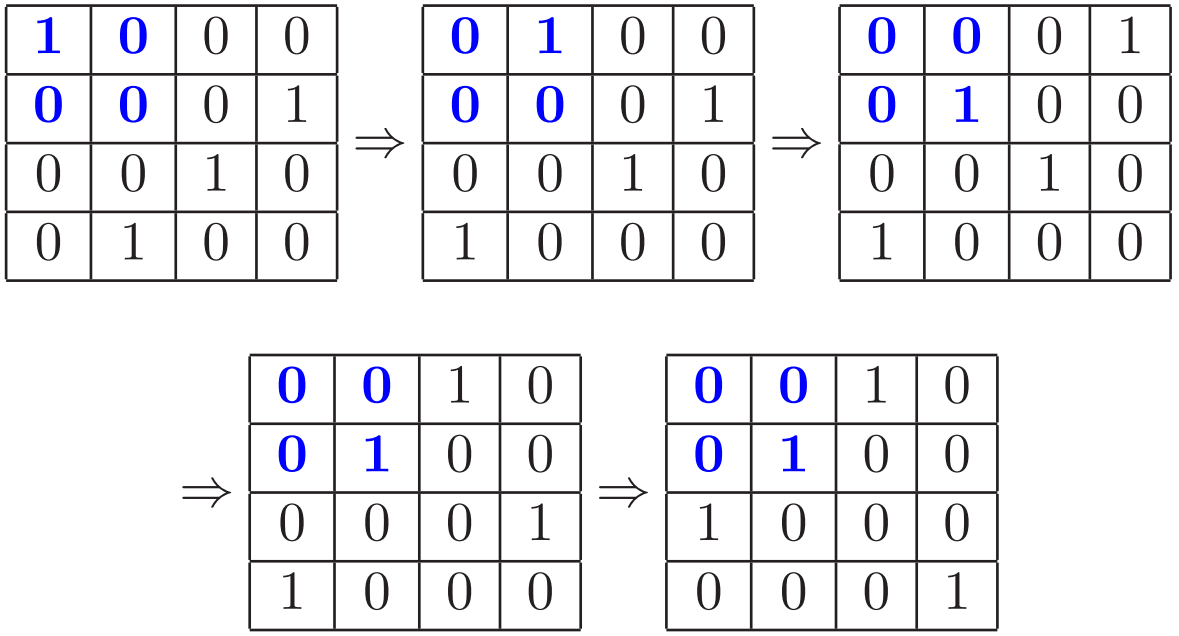
0	**0**	1	0
0	**1**	0	0
1	0	0	0
0	0	0	1

where we fix the row sums, column sums and the sum of blue cell counts.

By our theorem, we should be able to move from the left table to the right table by basic moves only.

Example

Indeed we can move as follows:



Notes

From theoretical viewpoint, it is interesting to study the structure of Markov bases for cases if S or S^c contains a pattern $\mathcal{P}$ or $\mathcal{P}^t$.

A structural zero problem is a particular case of the subtable sum problems. Various properties of Markov bases are known for structural zero problems. It is of interest to investigate which properties of Markov bases for structural zero problem for S can be generalized to subtable sum problem for S.

We did not use algebraic geometry to prove our theorem. Thus, it would be also very interesting to investigate subtable sum problem from the viewpoint of algebraic geometry.

Questions??

Thank you....

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