Markov Bases for Conditional Test of Common Diagonal Effect in Quasi-Independence Model for Two-Way Contingency Tables

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

Joint work with H. Hara and A. Takemura

www.ms.uky.edu/~ruriko

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 two-way contingency tables we sometimes find that frequencies along the diagonal cells are relatively larger (or smaller) compared to off-diagonal cells, particularly in square tables with the common categories for the rows and the columns, such as the previous example.

In this case the quasi-independence model with an additional parameter for each of the diagonal cells is usually fitted to the data.

A simpler model than the quasi-independence model is to assume a common additional parameter for all the diagonal cells. We call this **Common Diagonal Effect Model**.

We consider testing the goodness of fit of the common diagonal effect by Markov chain Monte Carlo (MCMC) method via Markov bases.

Quasi-Independence Model

Consider an $R \times C$ two-way contingency table $\boldsymbol{x} = \{x_{ij}\}$.

In the quasi-independence model, the cell probabilities $\{p_{ij}\}$ are modeled as

$$\log p_{ij} = \mu + \alpha_i + \beta_j + \gamma_i \delta_{ij},$$

where δ_{ij} is Kronecker's delta.

Here each diagonal cell (i, i), $i = 1, \ldots, \min(R, C)$, has its own free parameter γ_i. This implies that in the maximum likelihood estimation each diagonal cell is perfectly fitted:

$$\hat{p}_{ii} = \frac{x_{ii}}{n},$$

where $n = \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij}$ is the total frequency.

Common Diagonal Effect Model

As a simpler submodel of the quasi-independence model we consider the null hypothesis

$$H_0: \ \gamma = \gamma_i, \quad i = 1, \ldots, \min(R, C),$$

in the quasi-independence model.

We call this model a **Common Diagonal Effect Model (CDEM)**.

Note: Under CDEM the sufficient statistics consists of the row sums, column sums and the sum of the diagonal frequencies.

Want: We want to sample tables from the hypergeometric distribution given the sufficient statistics via MCMC.

In order to run MCMC on all tables satisfying the row sums, column sums, and the diagonal sum, we want to compute a **Markov basis** for this model.

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 diagonal cells.

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 the diagonal sum, then it is not necessarily true anymore.

For example, with a 2×3 tables with the fixed sum of x_{11} and x_{22} then there are only three tables such that

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

and these tables are not connected by basic moves.

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

$$\begin{array}{|c|c|c|}\hline 1 & 1 & -2 \\ \hline -1 & -1 & 2 \\ \hline \end{array}$$

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

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

Compute Markov bases

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.

Note: To compute a Markov basis for the CDEM, we did not use computational algebraic techniques.

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 $S = \{(i, j) \mid i = j\} \subset \mathcal{I}$ and S^c is the complement of S.

An **indispensable move** is a move which belongs every Markov bases.

A **dispensable move** is a move such that there exists a Markov Basis, not containing this particular move.

Note Basic moves are indispensable.

Markov basis for CDEM

We define 5 different types of moves:

Type I (basic moves in S^C for $\max(R, C) \geq 4$):

$$\begin{array}{ccc} & j & j' \\ i & +1 & -1 \\ i' & -1 & +1 \end{array}$$

where i, i', j, j' are all distinct.

Type II (indispensable moves of degree 3 in S^C for $\min(R, C) \geq 3$):

$$\begin{array}{cccc} & i & i' & i'' \\ i & 0 & +1 & -1 \\ i' & -1 & 0 & +1 \\ i'' & +1 & -1 & 0 \end{array}$$

Markov basis for CDEM

Type III (dispensable moves of degree 3 for $\min(R, C) \geq 3$):

$$\begin{array}{cccc} & i & i' & i'' \\ i & +1 & 0 & -1 \\ i' & 0 & -1 & +1 \\ i'' & -1 & +1 & 0 \end{array}$$

Type IV (indispensable moves of degree 4 which are non-square free):

$$\begin{array}{cccc} & j & j' & j'' \\ i & +1 & +1 & -2 \\ i' & -1 & -1 & +2 \end{array}$$

where $i = j$ and $i' = j'$, i.e., two cells are on the diagonal. Note that we also include the transpose of this type as Type IV moves.

Markov basis for CDEM

Type V (square free indispensable move of degree 4 for $\max(R, C) \geq 4$):

$$\begin{array}{ccccc} & j & j' & j'' & j'''' \\ i & +1 & +1 & -1 & -1 \\ i' & -1 & -1 & +1 & +1 \end{array}$$

where $i = j$ and $i' = j'$. Type V includes the transpose of this type.

Theorem [Hara, Takemura, Y (2008)]

The above moves of Types I-V form a Markov basis for the diagonal sum problem with $\min(R, C) \geq 3$ and $\max(R, C) \geq 4$.

Relationship between birthday and death day

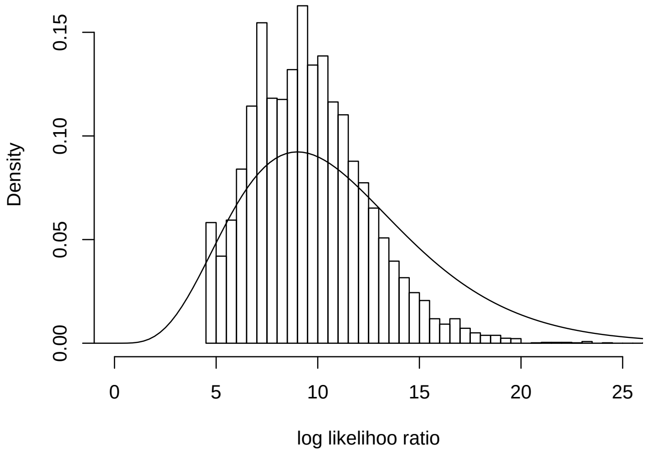
We now test CDEM against the quasi-independence model.

$$H_0 : \text{CDEM fits}$$
$$H_1 : \text{QI model fits}$$

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.



Thank you....

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