Markov bases and contingency tables (I)

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

polytopes.net

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.

	Blood	Serum Cholesterol (mg/100ml)						
		1	2	3	4	5	6	7
	Pressure	< 200	200-209	210-219	220-244	245-259	260-284	> 284
1	< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source : [Cornfield, 1962]

Data on coronary heart disease incidence in Framingham, Massachusetts [Cornfield, 1962, Agresti, 1990]. A sample of male residents, aged 40 through 50, were classified on blood pressure and serum cholesterol concentration. $2/53$ in the (1,1) cell means that there are 53 cases, of whom 2 exhibited heart disease.

Incomplete contingency table

Table 2: Effects of decision alternatives on the verdicts and social perceptions of simulated jurors.

	Condition						
Alternative	1	2	3	4	5	6	7
First degree	11	[0]	[0]	2	7	[0]	2
Second degree	[0]	20	[0]	22	[0]	11	15
Manslaughter	[0]	[0]	22	[0]	16	13	5
Not guilty	13	4	2	0	1	0	2

Source : [Vidmar, 1972]

This table refers to the possible effects on decision making of limiting the number of alternatives available to the number of a jury panel.

$[0]$ refers to the structural zero on the cell.

Contingency tables

A **contingency table** is a table which records counts of events at combinations of factors, and it is used to study the relationship/correlations between the factors.

All possible combinations of factor labels make **cells** in an array, and the count in each cell may be viewed as the outcome of a multinomial probability distribution.

Let $\mathbf{X}$ be a contingency table with k cells. In order to simplify the notation, we denote by $\mathcal{X} = \{1, \ldots, k\}$, the sample space of the contingency table.

In the special case of two-way contingency tables with I rows and J columns, we also denote the sample space with $\mathcal{X} = \{1, \ldots, I\} \times \{1, \ldots, J\}$.

Example: Independence model

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

An observed table $X^{obs} = \{x_{ij}^{obs}\}$, $x_{ij}^{obs} \in \mathbb{N}$, and $1 \leq I, 1 \leq J$.

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

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 $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Independence model is a special case such that

$$\lambda_{ij}^{RC} = 0 \text{ for } 1 \leq i \leq I, 1 \leq j \leq J.$$

Hypothesis

The sufficient statistics for independence model include the row and column margins. Hence, the conditional distribution of the table counts given the margins is the same regardless of the values of the parameters in the model.

We have the following hypothesis test:

$$H_0 : \lambda_{ij}^{RC} = 0 \text{ no interaction.}$$
$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

Exact p-value computation

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

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

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.

Exact p-value computation

Note that the row sums and column sums 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	

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

e.g. For our 2×3 table, we have: let $Z_+ = \{0, 1, 2, \cdots\}$,

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

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

Consider the independence model,

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

Table 3: 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

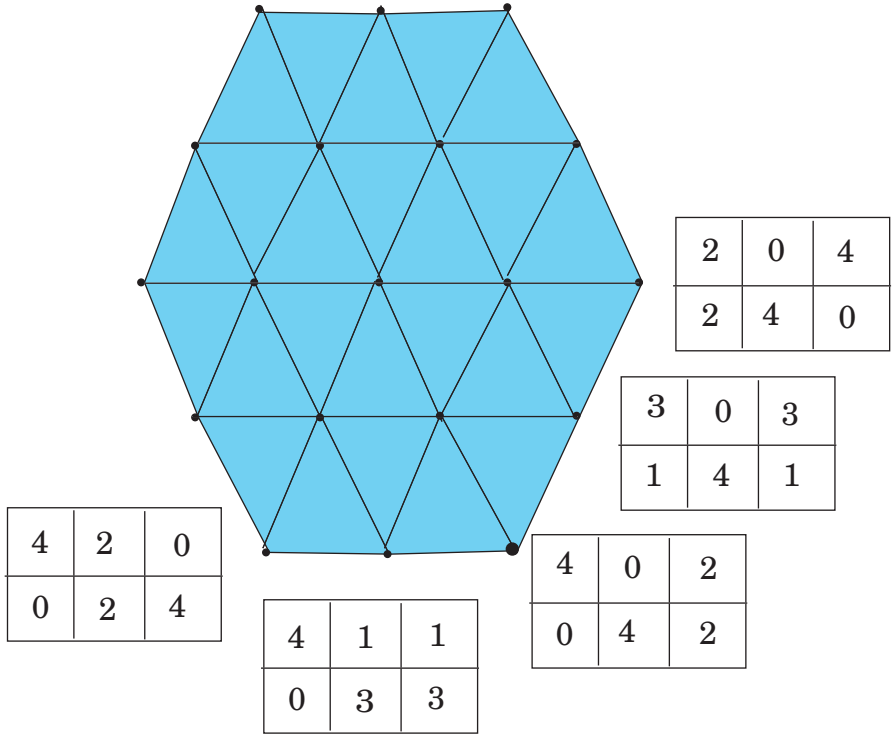
+

-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.

Definition: Let $A : \mathbb{Z}^n \to \mathbb{Z}^d$. The toric ideal I_A is the ideal

$$\langle p^u - p^v : u, v \in \mathbb{N}^n,\ Au = Av \rangle \subset K[p_1, \ldots, p_n],$$

where $p^u = p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}$.

Fundamental Theorem of Markov Bases (Diaconis-Sturmfels 1998). The set of moves $B \subseteq \ker_Z(A)$ is a Markov basis for A if and only if the set of binomials

$$p^{b+} - p^{b-} : b \in B$$

generates I_A.

$$\begin{array}{|c|c|c|c|} \hline 0 & 1 & -1 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & -1 & 1 & 0 \\ \hline \end{array} \longrightarrow p_{12}p_{33} - p_{13}p_{32}.$$

Fact: For any 2-way contingency tables with fixed row and column sums, we know that a set of basic moves forms a Markov basis.

Note: If you add additional constraints, (for example bounded 2-dimensional tables) then it is not necessarily true anymore.

Note: A Gröbner basis of a toric idea $\mathcal{I}_A$ associate to a design 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 $\mathcal{I}_A$ 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 in general.

Basic moves

Here we consider 2-way tables. Let b be a $r \times c$ table such that

	j	j'
i	1	-1
i'	-1	1

where $1 \leq i, i' \leq r$, $1 \leq j, j' \leq c$, $i \neq i'$, $j \neq j'$ and other cells are all zero. We call $\pm b$ a **basic move**. We denote this table b as $(i, i'; j, j')$.

Fact: A set of basic moves forms a Markov basis for $r \times c$ tables under the independence model.

Notation

Without loss of generality, we represent a table by a vector of counts $\mathbf{x} = (x_1, \ldots, x_k)$.

The fiber of an observed table $\mathbf{x}_{\text{obs}}$ with respect to a function $T : \mathbb{N}^k \longrightarrow \mathbb{N}^s$ is the set

$$\mathcal{F}_T(\mathbf{x}_{\text{obs}}) = \left\{ \mathbf{x} \mid \mathbf{x} \in \mathbb{N}^k \ , \ T(\mathbf{x}) = T(\mathbf{x}_{\text{obs}}) \right\} \ .$$

When the dependence on the specific observed table is irrelevant, we will write simply $\mathcal{F}_T$ instead of $\mathcal{F}_T(\mathbf{x}_{\text{obs}})$.

In mathematical statistics framework, the function T is usually the minimal sufficient statistic of some statistical model.

MCMC procedure

Metropolis-Hastings algorithm

1. Start with the observed table as a current location.

2. Propose a change to another table.

3. Three situations:

   (a) If proposed location is better, move to the new location
   (b) If proposed location is worse, move to the new location with probability equal to ratio of new to old location
   (c) If proposed location is outside of the fiber, then stay at the current location.

4. Record log likelihood value of H_0 and H_1.

Proposing moves

We move from the current state to a new location by the following matter:

- Ratio of the probabilities of new location to old location: $x \rightarrow$ Then accept proposed move with prob. x
- New location higher probability than old location $\rightarrow$ Accept proposed move to a new location.

Example:

Ratio of new location to old location: $1/3 \rightarrow$ Then accept proposed move with prob. $1/3$

New location better than old location $\rightarrow$ Accept proposed move

Estimating the distribution

- Take samples every s steps (e.g., every 100 steps).
  This is mimicking Independence of a sample.

- The Markov chain is started from a random initial value x_0 and the algorithm is run for many iterations until this initial state is "forgotten" (try to get the stationary distribution). These samples, which are discarded, are known as "burn-in".
  Discard the first $x\%$ of steps as "burn-in"

- Plot a histogram from the remaining samples

- This provides an estimate of the distribution of the log likelihood ratios of sampled tables!

But a problem is that the chain might get stuck in a local optima. Solution?

Metropolis-coupling

- Use more than one chain in the analysis
- Additional "heated" chains:
  - More willing to go downhill in the landscape
  - Act as "scouts"
- If one of the additional chains finds a better location, it swaps places with the "cold" chain
- Results in quicker convergence and better mixing
- Reduces chance of being trapped in local optimum

Thank you....