Holes in Semigroups and Their Applications to the Two-Way Common Diagonal Effect Model

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with A. Takemura and P. Thomas

`http://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.

Common Diagonal Effect Model

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 a simple model is to assume a common additional parameter for all the diagonal cells. We call this the **Common Diagonal Effect Model**.

But following a result from Ohsugi and Hibi, the semigroup formed by the columns of the CDEM design matrix is not normal; it has at least one hole. We feel more research on non-normal semigroups is needed, so here we develop some theory for this case.

An example CDEM design matrix

In the CDEM, the sufficient statistics for a table are the row sums, column sums, and the sum of the main diagonal.
Example: Consider a 3×3 table and its marginals

x_{11}	x_{12}	x_{13}	r_1
x_{21}	x_{22}	x_{23}	r_2
x_{31}	x_{32}	x_{33}	r_3
c_1	c_2	c_3	d

where the x_{ij} are the cell counts and each r, c, and d corresponds to a row, column, or diagonal sum.

Now we set up a system of linear equations.

$$
\begin{array}{ccccccccccc}
x_{11} & & & +x_{21} & & & +x_{31} & & & = & c_1 \\
& x_{12} & & & +x_{2,2} & & & +x_{32} & & = & c_2 \\
& & x_{13} & & & +x_{23} & & & +x_{33} & = & c_3 \\
x_{1,1} & +x_{1,2} & +x_{1,3} & & & & & & & = & r_1 \\
& & & x_{2,1} & +x_{2,2} & +x_{2,3} & & & & = & r_2 \\
& & & & & & x_{31} & +x_{32} & +x_{33} & = & r_3 \\
x_{1,1} & & & & +x_{22} & & & & +x_{33} & = & d \\
& & & & & & & & x_{i,j} & \in & \mathbb{Z}_+
\end{array}
$$

where $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$.

The coefficients of these equations provide the model matrix for our marginal sums.

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Suppose we have a given set of margins for contingency tables.

Suppose if we want to decide whether there exists a table satisfying the given margins.

This is called the **multi-dimensional integer planar transportation problem** and it can be applied to **data sequrity problem**.

In terms of Optimization, we can rewrite this problem as an **integral feasibility problem**, that is:

Decide whether there exists an integral solution in the system

$$Ax = b,\ x \geq 0,$$

where $A \in \mathbb{Z}^{d \times n}$ and $b \in \mathbb{Z}^d$.

Observation

Assume the lattice L generated by the columns of A is $\mathbb{Z}^d$. Let cone(A) be the cone generated by the columns of A and $P_b = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$. We assume that cone$(A)$ is pointed.

$$P_b \neq \emptyset \Leftrightarrow b \in \text{cone}(A).$$

Let Q be the semigroup generated by the columns $\boldsymbol{a}_i$ of A, i.e. $Q = \{x \in \mathbb{R}^d : \sum_{i=1}^n \alpha_i \boldsymbol{a}_i,\ \alpha_i \in \mathbb{Z}_+\} \subset \text{cone}(A) \cap \mathbb{Z}^d$.

$$P_b \cap \mathbb{Z}^n \neq \emptyset \Leftrightarrow b \in Q.$$

$$(P_b \neq \emptyset) \bigwedge (P_b \cap \mathbb{Z}^n = \emptyset) \Leftrightarrow b \in \left(\text{cone}(A) \cap \mathbb{Z}^d - Q\right).$$

Some definitions

Let $A = \{a_1, ..., a_n\}, a_i \in \mathbb{Z}^d$ be the design matrix for tables under the CDEM. Let Q denote the commutative semigroup generated by the columns of A.

Let K be the rational polyhedral cone generated by the columns of A and let L be the lattice generated by the columns of A. We will be assuming that $L = \mathbb{Z}^d$.

The saturation Q_{sat} of Q is defined as $Q_{sat} = K \cap L$, and the elements of $H = Q_{sat} \,/\, Q$ are called **holes** of Q

The problem with holes

When sensitive data is released, the marginals are often perturbed to prevent information about individual members of the population from being recovered. If the perturbed marginals are a hole of the semigroup of the design matrix, the data is broken instead of bent.

Sequential importance sampling faces a similar problem; if a sampled table contains a hole, it must be rejected. This can cause an increase in sampling time in cases where holes abound.

As the result from Ohsugi and Hibi guarantees the existence of holes under the CDEM, we will investigate their distribution.

An example

Consider American football. With one rare exception, the list of all possible score changes is

$$\begin{pmatrix} 2 & 0 & 3 & 0 & 7 & 0 & 8 & 0 \\ 0 & 2 & 0 & 3 & 0 & 7 & 0 & 8 \end{pmatrix}$$

In this case, $Q_{sat} = K \cap L = \mathbb{Z}_+^2$ and $Q = \mathbb{Z}_+^2 / \{(k,1), (1,k) \mid \forall k \in \mathbb{Z}_+\}$. So the holes in this semigroup are all scores where at least one team has exactly one point. You can get there with rational additions of the permitted score changes, but not with integral additions.

Example

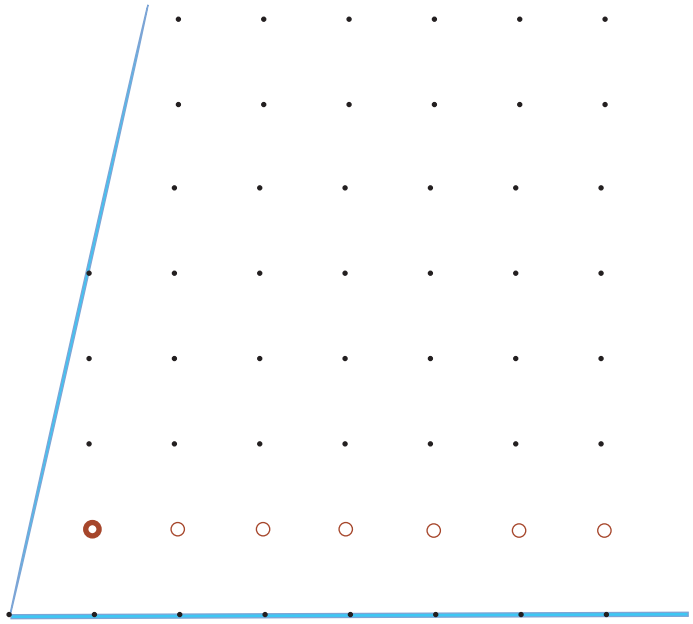

Figure 1: Non-holes, and holes for Example.

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 3 & 4 \end{pmatrix}.$$

Example cont.

Q has infinitely many holes

$$H = \{(1,1)^\intercal + \alpha \cdot (1,0)^\intercal : \alpha \in \mathbb{Z}_+\},$$

out of which only $(1,1)^\intercal$ is in the min. Hilbert basis.

A condition for finiteness of H

Thm. [Takemura and Y., 2008]

Let $B = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L\}$ denote the Hilbert basis of Q_{sat}. If $\boldsymbol{b}_l + \lambda \boldsymbol{a}_i \in Q$ for some $\lambda \in \mathbb{Z}$, let

$$\bar{\mu}_{li} = \min\{\lambda \in \mathbb{Z} \mid \boldsymbol{b}_l + \lambda \boldsymbol{a}_i \in Q\}$$

and $\bar{\mu}_{li} = \infty$ otherwise.

Then H is finite if and only if $\bar{\mu}_{li} < \infty$ for all $l = 1, \ldots, L$ and all $i = 1, \ldots, n$.

Remark. For each $1 \leq i \leq n$, let $\tilde{Q}_{(i)} = \{\sum_{j \neq i} \lambda_j \boldsymbol{a}_j \mid \lambda_j \in \mathbb{Z}_+, \ j \neq i\}$ be the semigroup spanned by $\boldsymbol{a}_j, j \neq i$. For each extreme $\boldsymbol{a}_i$ and for each $\boldsymbol{b}_l \notin Q$, we only have to check

$$\boldsymbol{b}_l \in (-\mathbb{Z}_+ \boldsymbol{a}_i) + \tilde{Q}_{(i)}, \text{ for } l = 1, \ldots, L.$$

Applications to contingency tables under the CDEM

We will now apply the theorem to a 3×3 table under the CDEM. After removing redundant rows with cddlib, we get the 6×9 matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The Hilbert basis of the cone generated by these nine vectors (computed via normaliz) consists of the nine vectors and three more.

$$b_{10} = (1\ 1\ 0\ 1\ 1\ 1)^t$$
$$b_{11} = (1\ 0\ 1\ 1\ 0\ 1)^t$$
$$b_{12} = (0\ 1\ 1\ 0\ 1\ 1)^t$$

These vectors are holes of the semigroup. And solving the semigroup membership problem via lrs, we determined that

$$b_{10} \notin (-\mathbb{N}a_1) + Q_1.$$

Thus by Theorem above, there are an infinite number of holes.

The 2×3 case

Consider the 2×3 table with the following marginals, with $y_{11} + y_{22} = c$. Here we can use combinatorics in lieu of applying the theorem.

y_{11}	y_{12}	y_{13}	c
y_{21}	y_{22}	y_{23}	c
c	c	0	

The unique solution for these equations is

$$\frac{c}{2} = y_{11} = y_{12} = y_{21} = y_{22},\ 0 = y_{13} = y_{23}$$

which is clearly not an integral for any odd positive integer c.

As suggested in Ohsugi and Hibi, the element of the lattice

$$\begin{pmatrix} c & c & 0 \\ 0 & 0 & 2c \end{pmatrix} - \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & c \end{pmatrix}$$

has the above marginals. So the previous table is a hole for all positive odd integer c; you can get there with the lattice, but not with integral combinations of the columns of the design matrix.

More generally

In the case of a 2×2 table we have a unique solution to $b = Ax$ and thus no holes, but for larger cases we will always have infinitely many holes.

Thm. [Takemura, Thomas and 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 with fixed row sums, column sums, and diagonal sum has infinitely many holes.

This arises from the fact that every such table has a 2×3 subtable, and every hole in the subtable corresponds to a unique hole in the larger table.

Distribution of holes

Saturation points

We call $s \in Q$ a **saturation point** of Q, if $s + Q_{\text{sat}} \subseteq Q$. The set of all saturation points of Q is denoted by S.

Note. If cone(A) is pointed then $S \neq \emptyset$.

Ex. In the previous example, any score where both teams have two or more points would be a saturation point. Any score where either team has zero points but no team has one point would be a non-saturation point (but not a hole).

Thm.[Takemura and Y., 2007] A face F of K is nowhere saturated (i.e., contains no saturation points) if and only if for some element b of the Hilbert basis B

$$b = x_1 a_1 + ... + x_n a_n, x_j \in \mathbb{Z} \text{ and } x_j \geq 0 \text{ for } a_j \notin F$$

does not have a feasible solution. Otherwise F is almost saturated (i.e., not nowhere saturated).

Results on hole distribution in the 3×3 case

Using allfaces_gmp from cddlib, we calculated which faces of the polyhedral cone defined by the design matrix for a 3×3 table are almost saturated and which are nowhere saturated.

Dimension	# of faces	# of nowhere	# of almost
6	1	0	1
5	16	0	16
4	54	3	51
3	67	13	54
2	36	18	18
1	9	9	0

Conclusions

So we have an entire family of semigroups with an infinite number of holes. Most theoretical studies on semigroups have been assuming normality, an absence of holes. In practice there are cases where this assumption does not hold, as with the CDEM. We think that we need more research on semigroups with holes.

In addition, non-normality of the semigroups causes difficulty for sequential importance sampling. While a Markov basis has been obtained for this model, it is still of interest to consider how to perform SIS under the CDEM.

Questions?

Thank you....