Connectivity of Fibers with Positive Margins in Multi-dimensional Contingency Tables

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

polytopes.net

	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.

Imcomplete contingency table

Table 1: 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 $1 \leq I, 1 \leq J$.

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

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 $i = 1, \ldots, I$ 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	

and each cell is bounded by 2, i.e., $x_{i,j} \leq 2$ for $i = 1, 2$ and $j = 1, 2, 3$.

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

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

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

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 2: 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

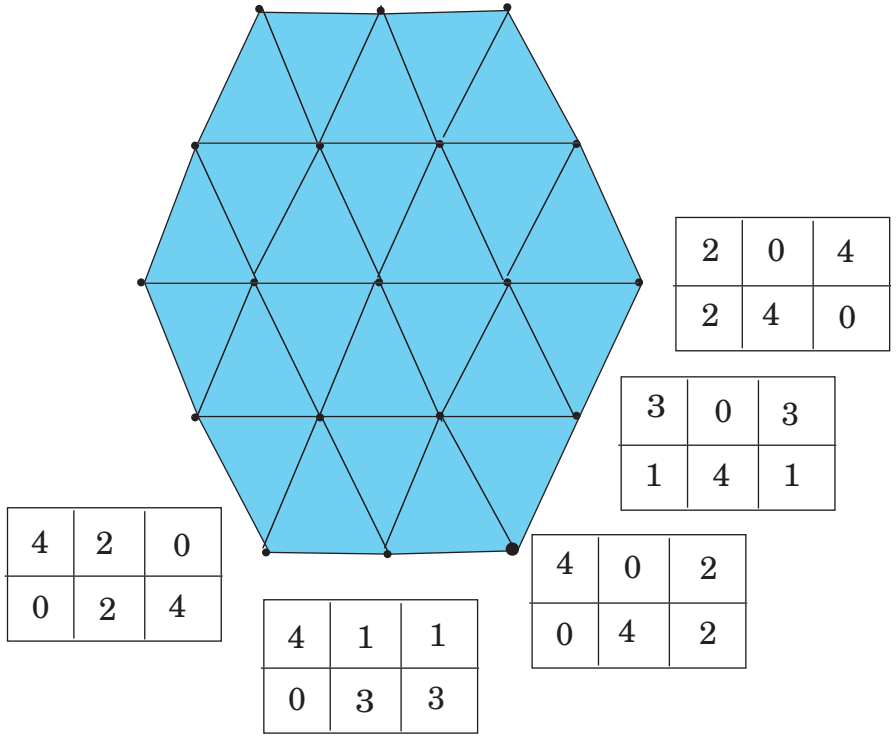
+

-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 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.

Question: Can we just compute a **connecting set** assuming that all margins are positive?

Notation

Without loss of generality, we represent a table by a vector of counts $\mathbf{n} = (n_1, \ldots, n_k)$.

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

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

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

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

Example: Bounded tables

Definition: A Universal Gröbner basis of an ideal is the Gröbner basis with respect to every term order.

Let an $s \times k$-matrix A_T be a design matrix of T and $\mathcal{I}_{A_T}$ be a toric ideal associate with A_T.

Theorem [Rapallo and Rogantin, 2007] A Universal Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ is a Markov basis of bounded tables under the given model.

Note: If we know a Universal Gröbner basis for A_T, then we can compute a MB for incomplete tables.

However, the Universal Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ is, in general, much bigger than a Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ with respect to a given term order. So in general it is very hard to compute.

Just to give the idea of such increase, we present in the following table the number of moves of the standard Markov basis for square $I \times I$ tables for the first I's.

	2	3	4	5	6	7
Standard Markov basis	1	9	36	100	225	441
Universal Gröbner basis	1	15	204	$3,940$	$113,865$	$4,027,161$

Thus, we consider the set of connecting moves.

Markov subbases

Definition: [Chen et. al., 2007] A Markov subbasis $M_{A_T,\mathbf{n}_{\rm obs}}$ for $\mathbf{n}_{\rm obs} \in \mathbb{N}^k$ and integer matrix A_T is a finite subset of $\ker(A_T) \cap \mathbb{Z}^k$ such that, for each pair of vectors $\mathbf{u}$, $\mathbf{v} \in \mathcal{F}_T$, there is a sequence of vectors $\mathbf{m}_i \in M_{A_T,\mathbf{n}_{\rm obs}}, i = 1, \ldots, l$, such that

$$\mathbf{u} = \mathbf{v} + \sum_{i=1}^{l} \mathbf{m}_i,$$

$$0 \leq \mathbf{v} + \sum_{i=1}^{j} \mathbf{m}_i,\ j = 1, \ldots, l.$$

The connectivity through nonnegative lattice points only is required to hold for this specific $\mathbf{n}_{\rm obs}$.

Note: $M_{A_T,\mathbf{n}_{\rm obs}}$ for every $\mathbf{n}_{\rm obs} \in \mathbb{N}^k$ and for a given A_T is a Markov basis $\mathcal{M}$ for A_T.

To compute a Markov subbasis, recall some definitions from commutative algebra:

An ideal $\mathcal{I} \subset \mathbb{R}[\mathbf{x}]$ is *radical* if

$$\{f \in \mathbb{R}[\mathbf{x}] \mid f^n \in \mathcal{I} \text{ for some } n\} = \mathcal{I}\,;$$

Let $\mathcal{I},\ \mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ be ideals. The quotient ideal $(\mathcal{I} : \mathcal{J})$ is defined by:

$$(\mathcal{I} : \mathcal{J}) = \{f \in \mathbb{R}[\mathbf{x}] \mid f \cdot \mathcal{J} \subset \mathcal{I}\}\ ;$$

Let $\mathcal{I},\ \mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ be ideals. The saturation of $\mathcal{I}$ with respect to $\mathcal{J}$ is the ideal defined by:

$$(\mathcal{I} : \mathcal{J}^{\infty}) = \{f \in \mathbb{R}[\mathbf{x}] \mid g^m \cdot f \in \mathcal{I},\ g \in \mathcal{J},\ \text{for some } m > 0\};$$

Let $Z = \{z_1, \ldots, z_s\} \subset \mathbb{R}^k$. A lattice L generated by Z is defined:

$$L = \mathbb{Z}Z.$$

$M \subset \mathbb{R}^k$ is called a lattice basis of L if each element in L can be written as a linear integer combination of elements in M.

Theorem [Chen, Dinwoodie, and Y., 2008] Suppose $\mathcal{I}_M$ is a radical ideal, and suppose M is a lattice basis. Let $p = x_1 \cdots x_k$. For each index ℓ with $(A_T\mathbf{n})_\ell > 0$, let $\mathcal{I}_\ell = \langle x_h \rangle_{(A_T)_{\ell,h}>0}$ be the monomial ideal generated by indeterminates for cells that contribute to margin ℓ. Let $\mathcal{L}$ be the collection of indices ℓ with $(A_T\mathbf{n})_\ell > 0$. Define

$$\mathcal{I}_\mathcal{L} = \left(\mathcal{I}_M : \prod_{\ell \in \mathcal{L}} \mathcal{I}_\ell \right).$$

If

$$(\mathcal{I}_\mathcal{L} : (\mathcal{I}_\mathcal{L} : p)) = \langle 1 \rangle \tag{1}$$

then the moves in M connect all the tables in $\mathcal{F}_T$.

Markov subbases for tables with positive bounds

We first study Markov subbases $M_{A_T,\mathbf{n}_{\rm obs}}$ for any bounded two-way contingency tables $\mathbf{n}_{\rm obs} \in \mathbb{N}^k$ with positive bounds, i.e., no structural zeros, under independence model.

Theorem [Rapallo and Y., 2010] Consider $I \times J$ tables with row and column sums fixed and with all cells bounded. If these bounds are positive, then a Markov subbasis for the tables is the standard Markov basis for $I \times J$ tables with row and column sums fixed without bounds, i.e., the set of basic moves of all 2×2 minors.

Markov subbases for incomplete tables

Now we study Markov subbases $M_{A_T,\mathbf{n}_{\rm obs}}$ for any incomplete $I \times J$ contingency tables $\mathbf{n}_{\rm obs} \in \mathbb{N}^k$ with positive margins, i.e., $A_T(\mathbf{n}_{\rm obs}) > 0$, under independence model.

Without loss of generality, we can assume that all margins are positive because cell counts in rows and/or columns with zero marginals are necessary zeros and such rows and/or columns can be ignored in the conditional analysis.

Let $\mathcal{X} = \{(i,j) \mid 1 \leq i \leq I, 1 \leq j \leq J\}$ and let S be a non-trivial subset of $\mathcal{X}$.

Proposition [Aoki and Takemura, 2005] Suppose we have $I \times J$ tables with fixed row and column sums. A set of basic moves is a Markov subbasis for $I \times J$ contingency tables, $I,\ J \geq 4$, with structural zeros in only diagonal elements, i.e., (i.e., cells with indices in $S = \{(i,j) : i = j$ for $i = 1, \ldots, \min(I,J)\}$) under the assumption of positive marginals.

Logistic regression and positive margins

In most applications of the logistic regression model, for each combination of covariates, the number of "successes" and the number of "failures" are observed.

The number of trials (i.e. the sum of numbers of "successes" and "failures") for each combination of covariates is usually fixed by a sampling scheme and positive. We call this marginal **the response variable marginal**.

Therefore we are usually interested in the connectivity of fibers with positive response variable marginals for sampling tables via Monte Carlo Markov chain (MCMC).

Univariate Logistic Regression Model

Let $\{1, \ldots, J\}$ be the set levels of a covariate and let X_{1j} and X_{2j}, $j = 1, \ldots, J$, be the numbers of successes and failures, respectively. The probability for success p_j is modeled as

$$\text{logit}(p_j) = \log \frac{p_j}{1 - p_j} = \alpha + \beta j, \qquad j = 1, \ldots, J.$$

The sufficient statistics for the model is $(X_{1+}, X_{+1}, \ldots, X_{+J}, \sum_{j=1}^{J} j X_{1j})$.

A **move** z is a table such that $X + z$ satisfies the given margins.

Moves $z = (z_{ij})$ for the model satisfy $(z_{1+}, z_{+1}, \ldots, z_{+J}) = 0$ and

$$\sum_{j=1}^{J} j z_{1j} = 0.$$

Bivariate Logistic Regression Model

Let $\{1, \ldots, J\}$ and $\{1, \ldots, K\}$ be the sets levels of two covariates. Let X_{1jk} and X_{2jk}, $j = 1, \ldots, J$, $k = 1, \ldots, K$, be the numbers of "successes" and "failures", respectively, for level (j, k). The probability for "success" p_{1jk} is modeled as

$$\text{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \mu + \alpha j + \beta k,$$

$$j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

The sufficient statistics for this model is X_{1++}, $\sum_{j=1}^{J} jX_{1j+}$, $\sum_{k=1}^{K} kX_{1+k}$, X_{+jk}, $\forall j, k$.

Hence moves $Z = (z_{ijk})$ for the model satisfy

$$z_{1++} = 0, \quad \sum_{j=1}^{J} j z_{1j+} = 0, \quad \sum_{k=1}^{K} k z_{1+k} = 0, \quad z_{+jk} = 0, \ \forall j, k.$$

Difficulty: the number of elements in a minimal Markov basis for a model can be exponentially many.

Question: Finding a set of Markov connecting moves, **Markov subbases**, that are much simpler than the full Markov basis with positive response variable marginals.

Markov subbasis for univariate logistic regression

Let $\boldsymbol{e}_j$ denote the contingency table with just 1 frequency in the j-th cell.

$$\mathcal{B} = \{\pm(\boldsymbol{e}_{j_1}+\boldsymbol{e}_{j_4}-\boldsymbol{e}_{j_2}-\boldsymbol{e}_{j_3}) \mid 1 \leq j_1 < j_2 \leq j_3 < j_4 \leq J,\ \ j_2-j_1 = j_4-j_3\}$$

Theorem: [Chen, Dinwoodie, Dobra, Huber, 2005]

The set of moves

$$\mathcal{B}_0 = \{z \in \mathcal{B} \mid j_2 = j_1 + 1, j_3 = j_4 - 1\}$$

connects every fiber satisfying $(X_{+1}, \ldots, X_{+J}) > 0$ for the univariate logistic regression model.

Markov subbasis for univariate logistic regression

if $j_2 \neq j_3$

	j_1	j_2	j_3	j_4
$i=1$	-1	1	1	-1
$i=2$	1	-1	-1	1

	j_1	j_2	j_3	j_4
$i=1$	1	-1	-1	1
$i=2$	-1	1	1	-1

if $j_2 = j_3$

	j_1	j_2	j_4
$i=1$	1	-2	1
$i=2$	-1	2	-1

	j_1	j_2	j_4
$i=1$	-1	2	-1
$i=2$	1	-2	1

Configuration for the bivariate logistic regression model

Consider two configurations $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J)$ and $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)$, where $\boldsymbol{a}_j$ and $\boldsymbol{b}_k$ are column vectors. We assume the homogeneity, i.e., there exist weight vectors w, v such that $\langle w, \boldsymbol{a}_j \rangle = 1, \ \forall j, \ \langle v, \boldsymbol{b}_k \rangle = 1, \ \forall k$.

The configuration $A \otimes B$ of the **Segre product** of A and B is defined as

$$A \otimes B = \Big(\boldsymbol{a}_j \oplus \boldsymbol{b}_k, \ j = 1, \ldots, J, k = 1, \ldots, K \Big), \qquad \boldsymbol{a}_j \oplus \boldsymbol{b}_k = \begin{pmatrix} \boldsymbol{a}_j \\ \boldsymbol{b}_k \end{pmatrix}.$$

Let

$$A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & J \end{pmatrix}, B = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & K \end{pmatrix}.$$

Fact: The configuration for the bivariate logistic regression model is the Lawrence lifting of Segre product $\Lambda(A \otimes B)$.

Markov subbasis

Consider a set of moves which connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.

Let $\boldsymbol{e}_{jk} = (e_{ijk})$ be redefined as an integer array with 1 at the cell $(1jk)$, -1 at the cell $(2jk)$ and 0 everywhere else. Define $\mathcal{B}_{\Lambda(A\otimes B)}$ as the set of moves $z = (z_{ijk})$ satisfying the following conditions,

1. $z = \boldsymbol{e}_{j_1k_1} - \boldsymbol{e}_{j_2k_2} - \boldsymbol{e}_{j_3k_3} + \boldsymbol{e}_{j_4k_4}$;

2. $(j_1, k_1) - (j_2, k_2) = (j_3, k_3) - (j_4, k_4)$.

Theorem [Hara, Takemura, Y., 2010]

$\mathcal{B}_{\Lambda(A\otimes B)}$ connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.

Examples of moves ($i = 1$ layer)

(1) $k_1 = \cdots = k_4$

	j_1	j_2	j_3	j_4
k_1	1	-1	-1	1

(2) $k_1 = \cdots = k_4$ and $j_2 = j_3$

	j_1	j_2	j_4
k_1	1	-2	1

(3) $k_1 = k_2$ and $j_2 = j_3$

	j_1	j_2	j_4
k_1	1	-1	0
k_3	0	-1	1

(4) $(j_2, k_2) = (j_3, k_3)$

	j_1	j_2	j_4
k_1	1	0	0
k_2	0	-2	0
k_4	0	0	1

(5) $k_1 = k_2$ $(k_3 = k_4)$

	j_1	j_2	j_3	j_4
k_1	1	-1	0	0
k_3	0	0	-1	1

(6) $k_1 = k_4$ and $j_2 = j_3$

	j_1	j_2	j_4
k_2	0	-1	0
k_1	1	0	1
k_3	0	-1	0

	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.

Data on coronary heart disease incidence

We examine the goodness-of-fit of the model with $J = 7$ and $K = 8$ by likelihood ratio statistic L_0.

We test the bivariate logistic regression defined above as a null hypothesis vs. ANOVA type logit model, namely:

$$\mathrm{H}_0 : \mathrm{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j + \beta k,$$

for $j = 1, \ldots, J, \quad k = 1, \ldots, K$.

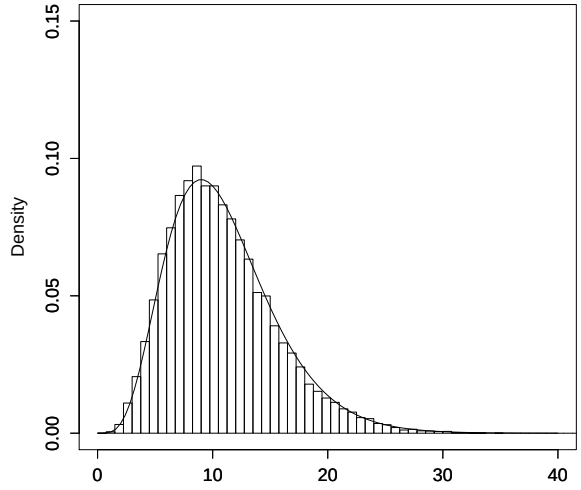
$$\mathrm{H}_1 : \mathrm{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha_j + \beta_k,$$

where $\sum_{j=1}^{J} \alpha_j = 0$ and $\sum_{k=1}^{K} \beta_k = 0$.

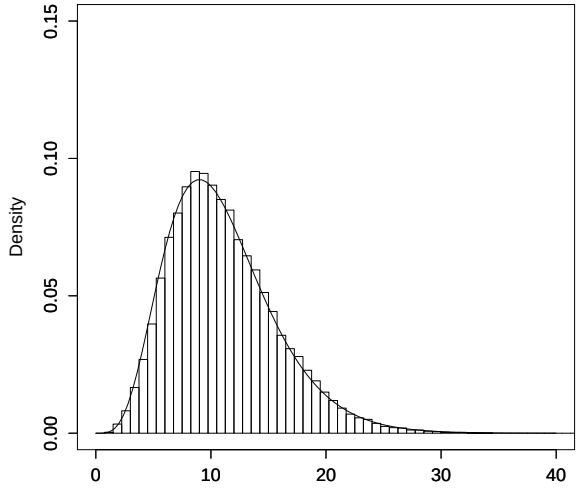
Data on coronary heart disease incidence

The value of L_0 is 13.07587 and the asymptotic p-value is 0.2884 from the asymptotic distribution χ^2_{11}. We computed the exact distribution of L_0 via MCMC with $\mathcal{B}_{\Gamma(A\otimes B)}$ defined. As an extension of $\mathcal{B}_0$ to the bivariate model, we define $\mathcal{B}_0^2$ by the set of moves $z = \boldsymbol{e}_{j_1k_1} - \boldsymbol{e}_{j_2k_2} - \boldsymbol{e}_{j_3k_3} + \boldsymbol{e}_{j_4k_4}$ satisfying $(j_1, k_1) - (j_2, k_2) = (j_3, k_3) - (j_4, k_4)$ is either of $(\pm 1, 0)$, $(0, \pm 1)$, $(\pm 1, \pm 1)$ or $(\pm 1, \mp 1)$.

The estimated p-values are 0.2706 with $\mathcal{B}_{\Gamma(A\otimes B)}$ and 0.2958 with $\mathcal{B}_0^2$. Therefore bivariate logistic regression model is accepted.



(a) A histogram with $\mathcal{B}_{\Lambda(A\otimes B)}$



(b) A histogram with $\mathcal{B}_0^2$

Figure 1: Histograms of L_0 via MCMC with $\mathcal{B}_{\Lambda(A\otimes B)}$ and $\mathcal{B}_0^2$

Data on occurrence of esophageal cancer

Table 3: Data on occurrence of esophageal cancer

		Age					
	Alcohol	1	2	3	4	5	6
	Consumption	25-34	35-44	45-54	55-64	65-74	75+
0	Low	0/106	5/169	21/159	34/173	36/124	8/39
1	High	1/10	4/30	25/54	42/69	19/37	5/5

Source : [Breslow and Day, 1980]

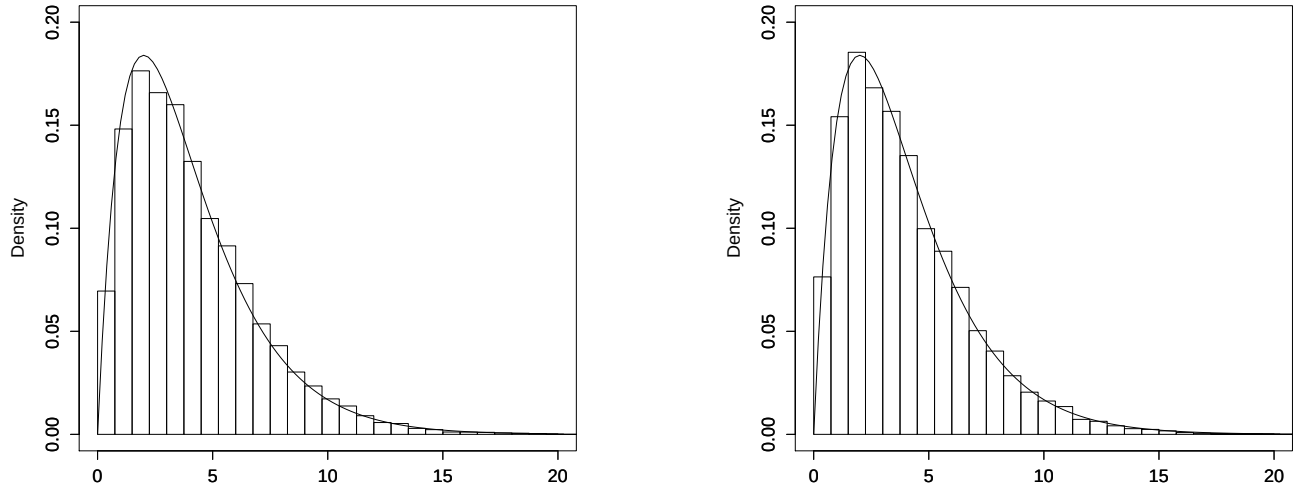
This table refers to the occurrence of esophageal cancer in Frenchmen which were classified on ages and dummy variable on alcohol consumption.

Data on occurrence of esophageal cancer

We test the goodness-of-fit of the bivariate logistic regression model with $J = 6$ and $K = 2$ by likelihood ratio statistics L_0 via MCMC. Then the value of L_0 is 20.89 and the asymptotic p-value is 0.0003330 from the asymptotic distribution χ^2_4.

We computed the exact distribution of L_0 via MCMC with $\mathcal{B}_{\Gamma(A\otimes B)}$ and $\mathcal{B}^2_0$. Figure 2 represents the histograms of L_0. The estimated p-values are 0.00011 with $\mathcal{B}_{\Gamma(A\otimes B)}$ and 0.00055 with $\mathcal{B}^2_0$. Therefore the model is rejected at the significance level of 1%.

Data on occurrence of esophageal cancer



(a) a histogram with $\mathcal{B}_{\Lambda(A\otimes B)}$ (b) a histogram with $\mathcal{B}_0^2$

Figure 2: Histograms of L_0 via MCMC with $\mathcal{B}_{\Lambda(A\otimes B)}$ and $\mathcal{B}_0^2$

The smooth line is asymptotic chi-square density, which shows a good fit.

Conjectures

The current proof for bivariate case is already very difficult and the general multivariate case remains to be a conjecture.

Conjecture: The set of moves $\mathcal{B}_{\Lambda(A_1 \otimes \cdots \otimes A_m)}$ connects every fiber with positive response marginals for the logistic regression with m covariates.

Conjecture: The subset of moves from $\mathcal{B}_{\Lambda(A_1 \otimes \cdots \otimes A_m)}$ such that the elements of $\boldsymbol{j}_1 - \boldsymbol{j}_2 = \boldsymbol{j}_3 - \boldsymbol{j}_4$ are ± 1 or 0 connects every fiber with positive response marginals for the logistic regression with m covariates. This is still conjecture for even $m = 2$.

Conjecture

Suppose M is a lattice basis. Let $p = x_1 \cdots x_k$. For each index ℓ with $(A_T\mathbf{n})_\ell > 0$, let $\mathcal{I}_\ell = \langle x_h \rangle_{(A_T)_{\ell,h}>0}$ be the monomial ideal generated by indeterminates for cells that contribute to margin ℓ. Let $\mathcal{L}$ be the collection of indices ℓ with $(A_T\mathbf{n})_\ell > 0$. Define

$$\mathcal{I}_\mathcal{L} = \left(\mathcal{I}_M : \prod_{\ell \in \mathcal{L}} \mathcal{I}_\ell \right).$$

If

$$(\mathcal{I}_\mathcal{L} : (\mathcal{I}_\mathcal{L} : p)) = \langle 1 \rangle \tag{2}$$

then the moves in M connect all the tables in $\mathcal{F}_T$.

Advertisement 1

Journal of Algebraic Statistics

The first issue will be published very soon (June 30th 2010).

http://www.jalgstat.com/

Advertisement 2

SIAM Annual Meeting on July 12th to 15th, Pittsburg

Minisymposium on Algebraic Statistics

`http://cophylogeny.net/SIAM_AN10.php`

Special Issue on Minisymposium on Algebraic Statistics in J of Algebraic Statistics

`http://www.jalgstat.com/`

Thank you....

The summary paper is in the first issue of J of Algebraic Statistics.