

Some results on connectivity of fibers with a subset of a Markov basis

Akimichi Takemura and Hisayuki Hara

University of Tokyo

March 28, 2009

Contents

- 1 Introduction: two-way tables
- 2 The case of $I \times J \times 2$ tables
- 3 Bivariate logistic regression
- 4 Numerical example
- 5 Summary

Introduction: two-way tables

Introduction: two-way tables

- For an introduction of Markov bases, as always, we start with a small two-way table (3×3):

| alge. \ stat. | A | B | C | total |
|---------------|----|----|----|-------|
| A | 7 | 5 | 1 | 13 |
| B | 5 | 10 | 6 | 21 |
| C | 2 | 6 | 8 | 16 |
| total | 14 | 21 | 15 | 50 |

Introduction: two-way tables

- The relation between the joint frequencies and the marginal frequencies is written as

$$\begin{pmatrix} 13 \\ 21 \\ 16 \\ 14 \\ 21 \\ 15 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 7 \\ 5 \\ 1 \\ 5 \\ 10 \\ 6 \\ 2 \\ 6 \\ 8 \end{pmatrix}$$

- Write this as

$$\mathbf{b} = \mathbf{A}\mathbf{x}$$

Introduction: two-way tables

- $x \geq 0$: joint frequency vector
- A : configuration
- b : marginal frequency vector
- **Move**: an integer vector z such that $Az = 0$. Then

$$A(x + z) = Ax$$

(We do not change the marginal frequencies.)

- Ex.

$$\begin{pmatrix} 7 & 5 & 1 \\ 5 & 10 & 6 \\ 2 & 6 & 8 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 8 & 4 & 1 \\ 4 & 11 & 6 \\ 2 & 6 & 8 \end{pmatrix}$$

Introduction: two-way tables

- In this way, we can move around the set (“fiber”) of contingency tables with common marginal frequencies.

- Fiber:

$$\mathcal{F}_{\mathbf{b}} = \{\mathbf{x} \geq 0 \mid \mathbf{b} = A\mathbf{x}\}$$

- **Markov basis**: a finite set of moves, by which we can reach **every** table of **every** fiber.
- (But we are often only interested in the **particular** fiber, where a give data belongs. We will talk about this later.)

Introduction: two-way tables

- **FACT:** for $I \times J$ two-way tables, the set of moves of the form

$$\begin{array}{cccccc} & \vdots & & \vdots & & \\ \dots & 1 & \dots & -1 & \dots & \\ & \vdots & & \vdots & & \\ \dots & -1 & \dots & 1 & \dots & \\ & \vdots & & \vdots & & \end{array}$$

forms a Markov basis. Call it a “basic move”.

Introduction: two-way tables

- A simple proof by “distance reducing argument”
- Suppose we have two different tables \mathbf{x}, \mathbf{y} in the same fiber.
- Let

$$d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sum_{i,j} |x_{ij} - y_{ij}|$$

denote the L_1 -distance between \mathbf{x} and \mathbf{y} .

- Because $\mathbf{x} \neq \mathbf{y}$, there exists a cell (i, j) such that

$$x_{ij} < y_{ij}$$

(total sample size n is common to \mathbf{x} and \mathbf{y} .)

Introduction: two-way tables

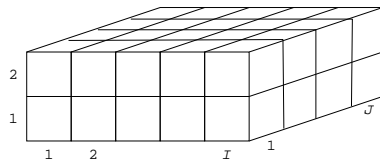
- Therefore there exists a sign pattern of $\mathbf{x} - \mathbf{y}$:

$$\begin{array}{cccccc} & & j & & j' & \\ & & \vdots & & \vdots & \\ i & \dots & - & \dots & + & \dots \\ & & \vdots & & \vdots & \\ i' & \dots & + & \dots & ? & \dots \\ & & \vdots & & \vdots & \end{array}$$

- In particular $x_{i'j} > 0, x_{ij'} > 0$ and we can subtract “1” from both cells of \mathbf{x} . We can always add 1 to x_{ij} and $x_{i'j'}$.
- Irrespective of the sign of “?”, we can reduce $d(\mathbf{x}, \mathbf{y})$ by 2 or 4.

The case of $I \times J \times 2$ tables
(with fixed line sums)

$I \times J \times 2$ tables



- It corresponds to “no-three-factor-interaction model”.
- It also corresponds to the logit model. Let p_{ij} denote the success probability for the pair of levels of covariates (i, j) . It is modeled as

$$\log \frac{p_{ij}}{1 - p_{ij}} = \mu + \alpha_i + \beta_j$$

- It is called the “Lawrence lifting” of $I \times J$ two-way case.

$I \times J \times 2$ tables

- For the Markov bases of $I \times J \times 2$ tables, we need longer “loops”:

| | | | |
|----|----|----|---|
| 1 | -1 | 0 | 0 |
| 0 | 1 | -1 | 0 |
| -1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 |

| | | | |
|----|----|----|---|
| -1 | 1 | 0 | 0 |
| 0 | -1 | 1 | 0 |
| 1 | 0 | -1 | 0 |
| 0 | 0 | 0 | 0 |

- Observation: these longer loops are needed when some of the “vertical line sum” is equal to 0 ($x_{ij+} = 0$).
- Consider a fiber that all vertical line sums are positive

$$x_{ij+} > 0, \quad \forall i, j.$$

$I \times J \times 2$ tables

- **FACT:** The fibers with $x_{ij+} > 0, \forall i, j$ are connected by “basic moves”

$$\begin{array}{cccccccccc} & & j & & j' & & & \vdots & & \vdots & \\ i & \dots & 1 & \dots & -1 & \dots & \dots & 1 & \dots & -1 & \dots \\ & & \vdots & & \vdots & & & \vdots & & \vdots & \\ i' & \dots & -1 & \dots & 1 & \dots & \dots & -1 & \dots & 1 & \dots \\ & & \vdots & & \vdots & & & \vdots & & \vdots & \end{array}$$

only. (We do not need longer loops!)

$I \times J \times 2$ tables

- Consider the sign pattern of difference $x - y$ of two $I \times J \times 2$ tables x, y :

| | | | | | | | | | | |
|------|-----|----------|-----|----------|-----|-----|----------|-----|----------|-----|
| | | j | | j' | | | j | | j' | |
| | | \vdots | | \vdots | | | \vdots | | \vdots | |
| i | ... | - | ... | + | ... | ... | + | ... | - | ... |
| | | \vdots | | \vdots | | | \vdots | | \vdots | |
| i' | ... | + | ... | ? | ... | ... | - | ... | ?? | ... |
| | | \vdots | | \vdots | | | \vdots | | \vdots | |

- If “??” is positive in x , we can apply a basic move to x .
- Therefore we can assume that $?? = 0$ in x .
- If “?” is positive in y , we can reverse the roles of x and y .
- After some inductive argument we can show that we can make the distance smaller by a basic move!

Bivariate logistic regression

Preliminary: Univariate logistic regression

- Univariate logistic regression

$$\log \left(\frac{p_i}{1 - p_i} \right) = \mu + \alpha i, \quad i = 1, \dots, l$$

- Let $\mathbf{e}_i = (e_{ik})$ denote an $l \times 2$ integer array with

$$\begin{cases} 1 & \text{in the } (i) \text{ cell} \\ -1 & \text{in the } (i) \text{ cell} \\ 0 & \text{everywhere else.} \end{cases}$$

- Let

① $\mathbf{z} = \mathbf{e}_{i_1} - \mathbf{e}_{i_2} - \mathbf{e}_{i_3} + \mathbf{e}_{i_4}$;

② $i_1 - i_2 = i_3 - i_4$

- **Theorem (Chen et al.(2005)):** The set of these moves connects every fiber with $x_{i+} > 0$.

Bivariate logistic regression

*Reference: "On connectivity of fibers with positive marginals in multiple logistic regression" by Hisayuki Hara, Akimichi Takemura and Ruriko Yoshida.
arXiv:0810.1793v1*

- Bivariate logistic regression

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \mu + \alpha i + \beta j,$$

$$i = 1, \dots, I, \quad j = 1, \dots, J.$$

(linear in i and j).

Bivariate logistic regression

- Sufficient statistics

$$x_{++1}, x_{ij+}, \sum_{i=1}^I i x_{i+1}, \sum_{j=1}^J j x_{+j1}$$

- Configuration $\Lambda(A \otimes B)$

$$\Lambda(A \otimes B) = \begin{pmatrix} A \otimes B & 0 \\ E_{IJ} & E_{IJ} \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & I \end{pmatrix}, B = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & J \end{pmatrix},$$

$$A \otimes B = \left(\mathbf{a}_i \oplus \mathbf{b}_j, i = 1, \dots, I, j = 1, \dots, J \right), \mathbf{a}_i \oplus \mathbf{b}_j = \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_j \end{pmatrix}.$$

E_{IJ} : $IJ \times IJ$ identity matrix

Bivariate logistic regression

- This is a “Lawrence lifting” of bivariate Poisson regression with the configuration $A \otimes B$.
- Again consider the case that $x_{ij+} > 0$ for all i, j .
- This assumption is natural, because x_{ij+} are often positive and fixed by a sampling scheme.
(binomial sampling for the pair of levels (i, j) .)
- Again Markov bases for the bivariate logistic regression are very complicated without the assumption $x_{ij+} > 0, \forall i, j$.

Bivariate logistic regression

- For the connectivity of fibers with $x_{ij+} > 0$, $\forall i, j$ we obtained the following subset Markov basis.
- Let $\mathbf{e}_{ij} = (e_{ijk})$ denote an integer array with

$$\begin{cases} 1 & \text{in the } (ij1) \text{ cell} \\ -1 & \text{in the } (ij2) \text{ cell} \\ 0 & \text{everywhere else.} \end{cases}$$

- Let
 - 1 $\mathbf{z} = \mathbf{e}_{i_1 j_1} - \mathbf{e}_{i_2 j_2} - \mathbf{e}_{i_3 j_3} + \mathbf{e}_{i_4 j_4}$;
 - 2 $(i_1, j_1) - (i_2, j_2) = (i_3, j_3) - (i_4, j_4)$
- **Theorem:** The set of these moves connects every fiber with $x_{ij+} > 0$.

Examples of moves

(Showing one slice of the moves only.)

(1) $i_1 = \dots = i_4$

| | j_1 | j_2 | j_3 | j_4 |
|-------|-------|-------|-------|-------|
| i_1 | 1 | -1 | -1 | 1 |

(2) $i_1 = \dots = i_4$ and $j_2 = j_3$

| | j_1 | j_2 | j_4 |
|-------|-------|-------|-------|
| i_1 | 1 | -2 | 1 |

(3) $i_1 = i_2$ ($i_3 = i_4$)

| | j_1 | j_2 | j_3 | j_4 |
|-------|-------|-------|-------|-------|
| i_1 | 1 | -1 | 0 | 0 |
| i_3 | 0 | 0 | -1 | 1 |

(4) $i_1 = i_2$ and $j_2 = j_3$

| | j_1 | j_2 | j_4 |
|-------|-------|-------|-------|
| i_1 | 1 | -1 | 0 |
| i_3 | 0 | -1 | 1 |

(5) $(j_2, i_2) = (j_3, i_3)$

| | j_1 | j_2 | j_4 |
|-------|-------|-------|-------|
| i_1 | 1 | 0 | 0 |
| i_2 | 0 | -2 | 0 |
| i_4 | 0 | 0 | 1 |

(6) $i_1 = i_4$ and $j_2 = j_3$

| | j_1 | j_2 | j_4 |
|-------|-------|-------|-------|
| i_2 | 0 | -1 | 0 |
| i_1 | 1 | 0 | 1 |
| i_3 | 0 | -1 | 0 |

(7)

| | j_1 | j_2 | j_3 | j_4 |
|-------|-------|-------|-------|-------|
| i_1 | 1 | 0 | 0 | 0 |
| i_2 | 0 | -1 | 0 | 0 |
| i_3 | 0 | 0 | -1 | 0 |
| i_4 | 0 | 0 | 0 | 1 |

Bivariate logistic regression

- The result on the bivariate case is a natural extension of the univariate case by Chen et al.
- Our proof of the bivariate case is already very difficult.
- We can formulate a natural conjecture for general multiple logistic regression.
- However we do not have a proof yet.

Numerical example

Data on coronary heart disease incidence

| | Blood Pressure | Serum Cholesterol (mg/100ml) | | | | | | |
|---|----------------|------------------------------|--------------|--------------|--------------|--------------|--------------|------------|
| | | 1 < 200 | 2 200-209 | 3 210-219 | 4 220-244 | 5 245-259 | 6 260-284 | 7 > 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)

- A sample of male residents, aged 40 through 50, were classified on blood pressure and serum cholesterol concentration
- 2/53 in (1, 1) cell means that there are 53 cases, of whom 2 exhibited heart disease

Exact test for bivariate logistic regression

- H_0 : bivariate logistic model

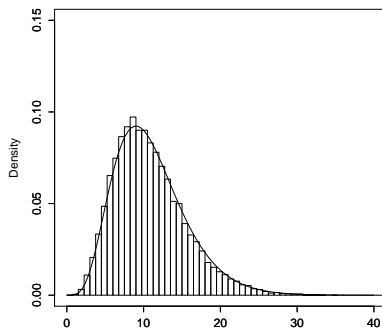
$$H_0 : \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \mu + \alpha_i + \beta_j,$$

- H_1 : No three-way interaction model

$$H_1 : \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \mu + \alpha_i + \beta_j,$$

- test statistics : likelihood ratio statistics L_0

Results



L_0

- $L_0 = 13.076$

- p -value

| | | |
|---------------|--------|----|
| χ_{11}^2 | MC | MC |
| 0.2884 | 0.2706 | |

Summary

Summary

- We have discussed that a subset of Markov basis connects $I \times J \times 2$ tables with all positive vertical sums.
- We have shown a similar result for the bivariate logistic regression.
- We have shown a numerical example of the bivariate logistic regression.
- In general, the problem of determining a subset of Markov basis for connecting one particular fiber remains a difficult question