JOURNAL OF ALGEBRAIC STATISTICS
Vol. 1, No. 1, 2010, 13-26
ISSN 1309-3452 – www.jalgstat.com

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

Ruriko Yoshida

Department of Statistics, University of Kentucky, Kentucky, USA

Abstract. Diaconis-Sturmfels developed an algorithm for sampling from conditional distributions for a statistical model of discrete exponential families, based on the algebraic theory of toric ideals. This algorithm is applied to categorical data analysis through the notion of Markov bases. Initiated with its application to Markov chain Monte Carlo approach for testing statistical fitting of the given model, many researchers have extensively studied the structure of Markov bases for models in computational algebraic statistics. In the Markov chain Monte Carlo approach for testing statistical fitting of the given model, a Markov basis is a set of moves connecting all contingency tables satisfying the given margins. Despite the computational advances, there are applied problems where one may never be able to compute a Markov basis. In general, the number of elements in a minimal Markov basis for a model can be exponentially many. Thus, it is important to compute a reduced number of moves which connect all tables instead of computing a Markov basis. In some cases, such as logistic regression, positive margins are shown to allow a set of Markov connecting moves that are much simpler than the full Markov basis. Such a set is called a Markov subbasis with assumption of positive margins.

In this paper we summarize some computations of and open problems on Markov subbases for contingency tables with assumption of positive margins under specific models as well as develop algebraic methods for studying connectivity of Markov moves with margin positivity to develop Markov sampling methods for exact conditional inference in statistical models where the Markov basis is hard to compute.

2000 Mathematics Subject Classifications: 05C81,62H17

Key Words and Phrases: Contingency tables, Markov bases, Toric ideals

1. Introduction

Algebraic Statistics is the field focused on the applications of algebraic geometry and its computational tools in the study of statistical models.

Usually we test a goodness of fit for statistical models of discrete exponential families based on the large sample approximation to the null distribution of test statistics. However, the large sample approximation may be poor when the sample sizes, i.e., the expected cell frequencies, are small [23]. In that case we apply the Markov chain Monte

Email addresses: ruriko.yoshida@uky.edu (R. Yoshida)

Carlo (MCMC) approach for testing statistical fitting of the given model. Sturmfels [33] and Diaconis-Sturmfels [15] developed an algebraic algorithm for sampling from conditional distributions for a statistical model of discrete exponential families, based on the algebraic theory of toric ideals and it is applied to categorical data analysis through the notion of Markov bases. Initiated with its application to MCMC approach for testing statistical fitting of the given model, many researchers have extensively studied the structure of Markov bases for models in computational algebraic statistics [17]. In the MCMC approach for testing statistical fitting of the given model, a Markov basis is defined as a set of moves connecting all contingency tables satisfying the given margins and Diaconis-Sturmfels showed that in fact the Markov basis for the given model with linear sufficient statistics is a set of generators of a toric ideal associate with its design matrix [15]. This theory, then, motivated for obtaining Markov chain moves, such as the genotype sampling method of [22], extensions to graphical models [19] and beyond [26].

There have been several algorithms and software developed for computing generators of a toric ideal such as [33], [28] and the software 4ti2 [1] which is very fast and user friendly [25]. However, despite these significant computational advances, there exist applied problems where one may never be able to compute a Markov basis. In general, computing a Markov basis for a model is NP-hard [14] and for some cases there are exponentially many elements in a Markov basis (e.g., models of no-3-way interaction in [14]). Thus, it is useful to compute a smaller set of moves which connect tables with given constraints rather than all constraints. This problem was already discussed in Section 3 of [15] on "corner minors". In [2] the case of two-way incomplete tables was studied. Connectivity of a set of Markov moves is traditionally studied through primary decomposition [16]. However, as a practical tool, this is problematic because the primary decomposition is very difficult to compute.

Therefore in this paper we will summarize some results and open problems on computing sets of Markov moves that connect tables with positive margins, because sets of Markov moves that work with certain margins may be much simpler than a full Markov basis.

Chen et. al. discussed that in some cases, such as logistic regression, positive margins are shown to allow a set of Markov connecting moves that are much simpler than the full Markov basis [10]. One such example is shown in [24] where a Markov basis for a multiple logistic regression is computed by the Lawrence lifting of this basis.

2. Preliminary

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 $\mathbf{n}$ be a contingency table with k cells. In order to simplify the notation, we denote by $\mathcal{X} = \{1, \dots, k\}$, 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

$\mathcal{X} = \{1, \ldots, I\} \times \{1, \ldots, J\} = \{(i, j) : i = 1, 2, \ldots, I, \, j = 1, 2, \ldots J\}$.

Let $\mathbb{N}$ be the set of nonnegative integers, i.e., $\mathbb{N} = \{0, 1, 2, \ldots\}$ and let $\mathbb{Z}$ be the set of all integers, i.e., $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Without loss of generality, in this paper, we vectorize a table by a vector of counts $\mathbf{n} = (n_1, \ldots, n_k)$. For example suppose we have the following 2×3 table

$$\begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix}$$

we write this table as a vector of counts

$$\mathbf{n} = (2,\, 3,\, 4,\, 5,\, 6,\, 7)\,.$$

Under this point of view, a contingency table $\mathbf{n}$ can be regarded as a vector $\mathbf{n} \in \mathbb{N}^k$.

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

$$\mathcal{F}_T(\mathbf{n}_{\rm obs}) = \left\{ \mathbf{n} \mid \mathbf{n} \in \mathbb{N}^k \ , \ T(\mathbf{n}) = T(\mathbf{n}_{\rm obs}) \right\} \, . \tag{1}$$

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

In mathematical statistics framework, the map T is usually the minimal sufficient statistic of some statistical model and the usefulness of enumerating the elements in $\mathcal{F}_T(\mathbf{n}_{\rm obs})$ follows from classical theorems such as Rao-Blackwell theorem (see e.g. [32]).

When the map T is linear, T is defined by an $s \times k$-matrix A_T, and the (ℓ, h)th element of the matrix A_T is

$$A_T(\ell, h) = T_\ell(h), \tag{2}$$

where T_ℓ is the ℓ-th component of T. In terms of the matrix A_T, $\mathcal{F}_T$ can be rewritten in the form:

$$\mathcal{F}_T = \left\{ \mathbf{n} \mid \mathbf{n} \in \mathbb{N}^k \ , \ A_T(\mathbf{n}) = A_T(\mathbf{n}_{\rm obs}) \right\} \, . \tag{3}$$

The matrix A_T is called the *design matrix* or *constraint matrix*, and the s rows are the vectors for computing sufficient statistics. For example, for 2×3 tables under the independence model, A_T is the 5×6 matrix given by

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

and the rows of A_T compute row and column sums of the contingency table.

To connect any two tables of the fiber $\mathcal{F}_T$ with a path of nonnegative tables, algebraic statistics suggests an approach based on the notion of Markov moves and Markov bases. A Markov move is any table $\mathbf{m}$ with integer entries that preserves the linear map T, i.e. $T(\mathbf{n} \pm \mathbf{m}) = T(\mathbf{n})$ for all $\mathbf{n} \in \mathcal{F}_T$.

A finite set of moves $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_r\}$ is called a *Markov basis* if it is possible to connect any two tables of $\mathcal{F}_T$ with moves in $\mathcal{M}$. More formally, for all $\mathbf{n}_1$ and $\mathbf{n}_2$ in $\mathcal{F}_T$, there exist a sequence of moves $\{\mathbf{m}_{i_1}, \ldots, \mathbf{m}_{i_A}\}$ and a sequence of signs $\{\epsilon_{i_1}, \ldots, \epsilon_{i_A}\}$ such that

$$\mathbf{n}_2 = \mathbf{n}_1 + \sum_{a=1}^{A} \epsilon_{i_a} \mathbf{m}_{i_a} \tag{4}$$

and

$$\mathbf{n}_1 + \sum_{j=1}^{a} \epsilon_{i_j} \mathbf{m}_{i_j} \geq 0 \quad \text{for all} \;\; a = 1, \ldots, A\,. \tag{5}$$

See [15] for further details on Markov bases.

To actually compute Markov bases, we associate to the problem two distinct polynomial rings. First, we define $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, \ldots, x_k]$, i.e., we associate an indeterminate x_h to any cell of the table; then, we define $\mathbb{R}[\mathbf{y}] = \mathbb{R}[y_1, \ldots, y_s]$, with an indeterminate y_ℓ for any component of the linear map T. In the following we will use some facts from commutative algebra, to be found in, e.g., [12].

Now we remind a reader of a *Markov subbasis.*

Definition 1 ([9]). *A Markov subbasis $\mathcal{M}_{A_T, \mathbf{n}_{\text{obs}}}$ for $\mathbf{n}_{\text{obs}} \in \mathbb{N}^k$ 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 $\mathbf{m}_i \in \mathcal{M}_{A_T, \mathbf{n}_{\text{obs}}}$, $i = 1, \ldots, l$, 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 $\mathbf{n}_{\text{obs}}$.

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

Now we recall some definitions from commutative algebra:

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

$$\{f \in \mathbb{R}[\mathbf{x}] \mid f^a \in \mathcal{I} \;\; \text{for some } a \in (\mathbb{N} - \{0\})\} = \mathcal{I}\,;$$

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

- Let $Z = \{z_1, \ldots, z_s\} \subset \mathbb{R}^k$. A lattice L generated by Z is defined:
$$L = \mathbb{Z}Z = \{x = \alpha \cdot y | \alpha \in \mathbb{Z},\, y \in Z\}.$$
$Z \subset \mathbb{R}^k$ is called a lattice basis of L if each element in L can be written as a linear integer combination of elements in Z. Now a lattice basis for $\ker(A_T)$ has the property that any two tables can be connected by its vector increments with allowing to swing negative in the connecting path (see Chapter 12 of [33] for definitions and properties of a lattice basis).

See [12] more details on the definitions above.

3. Computational methods

There are many ways to fit and evaluate a loglinear model for a multiway table of counts. For example, maximum likelihood fitting and asymptotic measures of goodness of fit are available as part of any generalized linear model package, such as the one in a software R [35]. R command `loglin` fits table, using iterative proportional fitting (IPF). IPF is more convenient than Poisson regression if the data is in a multidimensional array. Both methods rely on χ^2 asymptotics on either the Pearson χ^2 statistic or likelihood ratio statistic for goodness of fit. For sparse tables, we often use exact conditional methods in order to avoid asymptotic doubts. The command `chisq.test` in R has an option for the exact method on two-way tables, called Fisher's exact test introduced by Fisher [18].

For multiway tables, the package `exactLoglinTest` maintained by [7] contains an importance sampling (IS) method in [5] and there are certain examples where it has difficulty generating valid tables.

One can run Markov chains with a set of Markov moves obtained from a set of generators of a toric ideal associate with its design matrix using the algorithm proposed by Diaconis-Sturmfels [15]. One can compute such generators using algebra software packages, including COCOA [11], Macaulay 2 [20], and Singular [21] which implement several algorithms. 4ti2 [1] is one of the most popular software to compute a Markov basis because it is very fast, it has a natural coding language for statistical problems, and it has utilities for filtering output.

A Monte Carlo method, which is extremely flexible and does not require algebraic computations, is sequential importance sampling (SIS) [9].

4. Computing Markov subbases

From the definition of a Markov basis and also from the definition of MCMC, we do not allow all entries in the table to go negative when we are sampling. However, if one allows entries in the table to go negative, connecting Markov chains are easier to find. Let M be a set of Markov moves.

Proposition 1 (Proposition 0.2.1 in [10])**.** *Suppose I_M is a radical ideal, and suppose the moves in M form a lattice basis. Then the Markov chain using the moves in M that allows entries to drop down to -1 connects a set that includes the set $\mathcal{F}_T(\mathbf{n}_{\text{obs}})$.*

The idea of allowing some entries allowed to drop down to -1 appears in [6] and [9]. In high dimensional tables (k large), the enlarged state space that allows entries to drop down to -1 may be much larger than the set of interest $\mathcal{F}_T(\mathbf{n}_{\text{obs}})$, even though each dimension is only slightly extended. Nevertheless, Proposition 1 makes it possible to use the connect tables in a fiber on large tables (see [10]).

The Markov basis is a powerful tool to construct an irreducible Markov chain for any margins. However, in general, it is very hard to compute a Markov basis and also there can be arbitrary many elements in the Markov basis [14]. It is possible that a smaller set of moves may connect tables if margins are strictly positive. In order to study the connecting sets of moves in a fiber for certain values of margins, Chen et. al., in [9], introduced a notion of Markov subbases and here we are interested in computing a Markov subbasis under models with certain values of margins directly without computing a Markov basis nor computing the radical of the ideal of binomials from the lattice basis. In [10] Chen et. al. developed algorithms to compute such a Markov subbasis under some assumptions.

Theorem 1 (Proposition 0.4.1 in [10])**.** *Let A_T be a 0-1 matrix. Suppose there is an integer lower bound $b > 0$ on all the constraint values:*

$$t_\ell \geq b, \ \ell = 1, 2, \ldots, s,$$

where t_ℓ is the ℓ-th coordinate of the vector $\mathbf{t} = A_T(\mathbf{n}_{\text{obs}})$.

Let $\mathcal{I}_\ell = \langle x_h \rangle_{A_T(\ell,h)>0}$ be the monomial ideal generated by all the indeterminates for the cells that contribute to margin ℓ. If

$$\mathcal{I}_{A_T} \cap \bigcap_{\ell=1}^{s} \mathcal{I}_\ell^b \subset \mathcal{I}_M$$

where $\mathcal{I}_{A_T}$ is the toric ideal associate with the matrix A_T, $\mathcal{I}_M$ is a toric ideal generated by elements in M, and $\mathcal{I}_\ell^b = \langle x_{i_1} x_{i_2} \cdots x_{i_b} \rangle_{A_T(\ell,i_j)>0}$, then the moves in M connect all tables in $\mathcal{F}_T$.

This result can establish connectivity in examples where the primary decomposition is hard to compute. It does not require that I_M be radical. If we assume that I_M be radical, then we have the following theorem.

Theorem 2 (Theorem 0.4.1 in [10])**.** *Suppose $\mathcal{I}_M$ is a radical ideal, and suppose M is a lattice basis. Let $p = x_1 \cdots x_k$, let $\mathbf{t} = A_T(\mathbf{n}_{\text{obs}})$, and let t_ℓ be the ℓ-th coordinate of* $\mathbf{t}$. *For each index ℓ with $t_\ell > 0$, 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 $t_\ell > 0$. Define*

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

If

$$(\mathcal{I}_{\mathcal{L}} : (\mathcal{I}_{\mathcal{L}} : p)) = \langle 1 \rangle \tag{6}$$

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

One can find examples and applications of Theorems 1 and 2 in [10].

Assuming I_M be radical might be too strict thus we would like to remove this assumption. Here we have the following conjecture:

Conjecture 1. *Suppose M is a lattice basis. Let $p = x_1 \cdots x_k$, let $\mathbf{t} = A_T(\mathbf{n}_{\text{obs}})$, and let t_ℓ be the ℓ-th coordinate of $\mathbf{t}$. For each index ℓ with $t_\ell > 0$, 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 $t_\ell > 0$. Define $\mathcal{I}_{\mathcal{L}}$ as in Theorem 2. If the equation in (6) satisfies, then the moves in M connect all the tables in $\mathcal{F}_T$.*

Algorithms used in Theorem 1 and Theorem 2 do not compute a minimum set of moves connecting all tables in $\mathcal{F}_T$ with assumption of positive margins. Also note that 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. Thus we have the following problem:

Problem 1. *Assume that all margins are positive. Find an algorithm to compute a minimum set of moves connecting all tables in $\mathcal{F}_T$ in terms of inclusions.*

More generally,

Problem 2. *With $t_\ell > 0$ for some row index ℓ, find an algorithm to compute a minimum set of moves M in terms of inclusion connecting all tables in $\mathcal{F}_T$.*

From our simulations, the algorithms used in Theorem 2 and Theorem 1 seem slower than the algorithms to compute a full Markov basis using geometry, such as implemented in `4ti2`. Therefore to make an algebraic method more practical we have the following problem.

Problem 3. *With $t_\ell > 0$ for some row index ℓ, develop an algorithm to compute a minimum set of moves M in terms of inclusion which connects all tables in $\mathcal{F}_T$ faster than computing a full Markov basis via algorithms using geometry such as implemented in* `4ti2`.

Now we consider computation on Markov subbases for some specific models. Firstly we consider a two-dimensional tables with upper bounds. Contingency tables with upper bounds on the cell counts have recently been considered in, e.g., [13]. In general a Markov basis for unbounded contingency table under a certain model differs from a Markov basis for bounded tables. In [29, 30] Rapallo applied Lawrence lifting to compute a Markov basis for contingency tables whose cell entries are bounded. However, in the process, one has to compute the universal Gröbner basis of the ideal associated with the design matrix for a model which is, in general, larger than any reduced Gröbner basis. Thus, this is also infeasible in small- and medium-sized problems. Here we consider bounded two-way contingency tables under independence model. For $i \neq i'$ and $j \neq j'$, consider the square-free move of degree two with $+1$ at cells (i, j), (i', j') and -1 at cells (i, j') and (i', j) :

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

For simplicity we call this a *basic move.* It is well known that the set of all basic moves forms a unique minimal Markov basis under the independence model (in terms of algebra, extending the notion of indispensable binomials of a toric ideal, in [3] Aoki et. al. define indispensable monomials of a toric ideal and establish some of their properties), i.e. the problem on the tables with fixed rows sums and column sums.

However, the Markov basis for for bounded tables (i.e. all of the cells of a table have upper bounds) is much bigger that the set of all 2×2 minors. Then in [31] Rapallo and Yoshida showed that if these bounds on cells are positive, i.e., they are not structural zeros, the set of basic moves of all 2×2 minors connects all tables with given margins. Contingency tables with structural zero cells are called incomplete contingency tables ([4, Chapter 5]). Properties of Markov bases for incomplete tables are studied in [2, 27, 29].

Theorem 3 ([31])**.** *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 bounded tables is the set of basic moves of all 2×2 minors.*

Now we assume that the given margins are positive for bounded $I \times J$ tables, i.e., we assume that all row and column sums are positive. 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.

Now we consider $I \times J$ contingency tables with only diagonal elements being structural zeros under assumption of positive conditions on row and column sums. In [2] Aoki and Takemura showed the following propositions.

Proposition 2 ([2])**.** *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 under the assumption of positive marginals.*

Motivated by the proposition above we have the following problem to solve:

Problem 4. *Let $S \subset \mathcal{X}$ be the set of structural zeros. Suppose we have $I \times J$ tables with fixed row and column sums. What is the necessary and sufficient condition on S so that a* *set of basic moves is a Markov subbasis for $I \times J$ contingency tables with structural zeros in S under the assumption of positive marginals.*

5. Connectivity of fibers of positive marginals in bivariate logistic regression

In the case of bivariate logistic regression, in [24] Hara et. al. showed a simple subset of the Markov basis which connects all fibers with a positive sample size for each combination of levels of covariates. First, we consider univariate Poisson regression [16] with the set of levels $\{1,\dots,J\}$ of a covariate. The mean μ_j of independent Poisson random variables X_j, $j = 1,\dots,J$, is modeled as

$$\log \mu_j = \alpha + \beta j, \quad j = 1,\dots,J.$$

The sufficient statistic for the models is

$$(\sum_{j=1}^{J} X_j, \sum_{j=1}^{J} jX_j).$$

The first component is the total sample size

$$n = \sum_{j=1}^{J} X_j.$$

The design matrix A_T for this model is given by

$$A_T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & J \end{pmatrix}. \tag{7}$$

First we focus on the minimum-fiber Markov basis for the univariate Poisson regression. The minimum-fiber Markov basis is defined as the union of all minimal Markov bases [34].

Proposition 3 ([24])**.** *Let $\boldsymbol{e}_j$ denote the contingency table with just 1 frequency in the j-th cell. The set of moves*

$$\mathcal{B} = \{\pm(\boldsymbol{e}_{j_1} + \boldsymbol{e}_{j_4} - \boldsymbol{e}_{j_2} - \boldsymbol{e}_{j_3}) \mid 1 \le j_1 < j_2 \le j_3 < j_4 \le J, \quad j_2 - j_1 = j_4 - j_3\} \tag{8}$$

forms the minimum-fiber Markov basis for the univariate Poisson regression.

Now we focus on the univariate logistic regression [8]. We first show a brief review of Markov bases of univariate logistic regression model. Let $\{1,\dots,J\}$ be the set levels of a covariate and let X_{1j} and X_{2j}, $j = 1,\dots,J$, be the numbers of successes and failures, respectively. The probability for success p_j is modeled as

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

The sufficient statistics for the model is

$$(X_{1+}, X_{+1}, \ldots, X_{+J}, \sum_{j=1}^{J} jX_{+j}).$$

Hence moves $\mathbf{m} = (m_{ij})$ for the model satisfy $(m_{1+}, m_{+1}, \ldots, m_{+J}) = 0$ and

$$\sum_{j=1}^{J} jm_{+j} = 0. \tag{9}$$

The design matrix for this model is the Lawrence lifting $\Lambda(A_T)$ of A_T in (7):

$$\Lambda(A_T) = \begin{pmatrix} A_T & 0 \\ E_J & E_J \end{pmatrix}, \qquad A_T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & J \end{pmatrix}, \tag{10}$$

where E_J denotes the $J \times J$ identity matrix.

In general, Markov bases of the Lawrence lifting of A_T, $\Lambda(A_T)$, become more complicated than Markov bases for A_T. In usual applications of the logistic regression model, however, $X_{+j} := X_{1j} + X_{2j}$ is fixed by a sampling scheme and positive. In [8] Chen et. al. showed that a simple subset of Markov bases of $\Lambda(A_T)$ guarantees connectivity of all fibers satisfying $(X_{+1}, \ldots, X_{+J}) > 0$.

Let $\boldsymbol{e}_j$ be redefined by a $2 \times J$ integer array with 1 in the $(1, j)$-cell and -1 in the $(2, j)$-cell. Then in [24] Hara et. al. showed that the set of moves in (8) connects all fibers with $(X_{+1}, \ldots, X_{+J}) > 0$. More strongly, the set of moves is norm-reducing [34] for any two tables x, y in any fiber with positive marginals.

Proposition 4 ([24]). *The set of moves*

$$\mathcal{B}_{\Lambda(A_T)} = \{\pm(\boldsymbol{e}_{j_1} + \boldsymbol{e}_{j_4} - \boldsymbol{e}_{j_2} - \boldsymbol{e}_{j_3}) \mid 1 \le j_1 < j_2 \le j_3 < j_4 \le J, \;\; j_2 - j_1 = j_4 - j_3\} \tag{11}$$

is norm-reducing for all fibers with $(X_{+1}, \ldots, X_{+J}) > 0$ for the univariate logistic regression model.

In [8] Chen et. al. introduced a subset of $\mathcal{B}$ which still connects all fibers with $X_{+j} > 0, \forall j$.

Theorem 4 ([8]). *The set of moves*

$$\mathcal{B}_0 = \{\mathbf{m} \in \mathcal{B} \mid j_2 = j_1 + 1, j_3 = j_4 - 1\} \tag{12}$$

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

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

$$\text{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \mu + \alpha j + \beta k, \tag{13}$$

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

Then the likelihood is written by

$$\begin{aligned} L(\alpha, \beta, \gamma) &\propto \prod_{j=1}^{J} \prod_{k=1}^{K} (1 + \exp(\alpha + \beta j + \gamma k))^{-n_{+jk}} \\ &\quad \times \prod_{j=1}^{J} \prod_{k=1}^{K} \exp\left(\alpha n_{1jk} + \beta j n_{1jk} + \gamma k n_{1jk}\right) \\ &= \prod_{j=1}^{J} \prod_{k=1}^{K} (1 + \exp(\alpha + \beta j + \gamma k))^{-n_{+jk}} \\ &\quad \times \exp\left(\alpha n_{1++} + \beta \sum_{j=1}^{J} j n_{1j+} + \gamma \sum_{k=1}^{K} k n_{1+k}\right) \end{aligned}$$

Thus the sufficient statistics for this model is X_{1++}, $\sum_{j=1}^{J} j X_{1j+}$, $\sum_{k=1}^{K} k X_{1+k}$, X_{+jk}, $\forall j, k$. Hence moves $\mathbf{m} = (m_{ijk})$ for the model satisfy

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

Let

$$B = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & K \end{pmatrix}.$$

Then the design matrix for the bivariate logistic regression model is Lawrence lifting of Segre product $\Lambda(A_T \otimes B)$. Here we consider a set of moves which connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.

Definition 2. *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 $\mathcal{B}_{\Lambda(A_T \otimes B)}$ as the set of moves $\mathbf{m} = (m_{ijk})$ satisfying the following conditions.*

1. $\mathbf{m} = \boldsymbol{e}_{j_1 k_1} - \boldsymbol{e}_{j_2 k_2} - \boldsymbol{e}_{j_3 k_3} + \boldsymbol{e}_{j_4 k_4}$
2. $(j_1, k_1) - (j_2, k_2) = (j_3, k_3) - (j_4, k_4)$

Theorem 5 ([24]). *$\mathcal{B}_{\Lambda(A_T \otimes B)}$ connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.*

Theorem 5 shows the connectivity result for fibers with positive response variable marginals for bivariate logistic regression. There is a natural extension of Theorem 5 to m covariates: Let $\boldsymbol{j} = (j_1, \dots, j_m)$ denote the combination of m levels and let $\boldsymbol{e_j}$ denote an array with 1 at the cell $(1, \boldsymbol{j})$ and -1 at the cell $(2, \boldsymbol{j})$. Define $\mathcal{B}_{\Lambda(A_1 \otimes \cdots \otimes A_m)}$ as the set of the following moves $\mathbf{m}$:

1. $\mathbf{m} = \boldsymbol{e}_{\boldsymbol{j}_1} - \boldsymbol{e}_{\boldsymbol{j}_2} - \boldsymbol{e}_{\boldsymbol{j}_3} + \boldsymbol{e}_{\boldsymbol{j}_4}$

2. $\boldsymbol{j}_1 - \boldsymbol{j}_2 = \boldsymbol{j}_3 - \boldsymbol{j}_4$.

Then we have the following conjecture.

Conjecture 2. *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.*

In [24] Hara et. al. conjectured that we can further restrict to the set of moves $z = \boldsymbol{e}_{\boldsymbol{j}_1} - \boldsymbol{e}_{\boldsymbol{j}_2} - \boldsymbol{e}_{\boldsymbol{j}_3} + \boldsymbol{e}_{\boldsymbol{j}_4}$, where the elements of $\boldsymbol{j}_1 - \boldsymbol{j}_2 = \boldsymbol{j}_3 - \boldsymbol{j}_4$ are ±1 or 0.

Conjecture 3. *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.*

We have here an assumption that response marginals are positive. However this assumption may be too strict in practice. As discussed in Chen et al. [8], the connectivity under this assumption means that, if entries in columns in which response marginals are zeros are allowed to drop down to -1, any two tables with zero response marginals are connected by the set of moves proposed above. Hence it is possible to implement MCMC theoretically. Thus it would be interesting to investigate the cases if we have that the coefficient of one of the covariates is zero in the bivariate logistic regression.

References

[1] 4ti2 Team. 4ti2 – a software package for algebraic, geometric and combinatorial problems on linear spaces, 2006. Available at `www.4ti2.de`.

[2] S. Aoki and A. Takemura. Markov chain Monte Carlo exact tests for incomplete two-way contingency table. *Journal of Statistical Computation and Simulation*, 75 (10):787–812, 2005.

[3] S. Aoki, A. Takemura, and R. Yoshida. Indispensable monomials of toric ideals and markov bases. *J of Symbolic Computations*, 43:490–509, 2008.

[4] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge, Massachusetts, 1975.

[5] J. G. Booth and J. W. Butler. An importance sampling algorithm for exact conditional tests in loglinear models. *Biometrika*, 86:321–332, 1999.

[6] F. Bunea and J. Besag. Mcmc in $i \times j \times k$ contingency tables. *Monte Carlo Methods. N. Madras ed. Communications, American Mathematical Society*, pages 25–36, 2000.

[7] B. Caffo. exactloglintest: A program for monte carlo conditional analysis of log-linear models, 2006. Available at `http://www.cran.r-project.org/src/contrib/Descriptions/exactLoglinTest.html`.

[8] Y. Chen, I. Dinwoodie, A. Dobra, and M. Huber. Lattice points, contingency tables, and sampling. In *Integer points in polyhedra—geometry, number theory, algebra, optimization*, volume 374 of *Contemp. Math.*, pages 65–78. Amer. Math. Soc., Providence, RI, 2005.

[9] Y. Chen, I. H. Dinwoodie, and S. Sullivant. Sequential importance sampling for multiway tables. *The Annals of Statistics*, 34:523–545, 2006.

[10] Y. Chen, I. Dinwoodie, and R. Yoshida. Markov chains, quotient ideals, and connectivity with positive margins. Algebraic and Geometric Methods in Statistics *dedicated to Professor Giovanni Pistone (P. Gibilisco, E. Riccomagno, M.-P. Rogantin, H. P. Wynn, eds.)*, 2008. To appear.

[11] CoCoATeam. Cocoa: a system for doing computations in commutative algebra, 2007. Available at `http://cocoa.dima.unige.it`.

[12] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms.* Springer, New York, 2nd edition edition, 1997.

[13] M. Cryan, M. Dyer, and D. Randall. Approximately counting integral flows and cell-bounded contingency tables. In *Proc. STOC'05*, pages 413–422, Baltimore, Maryland, USA, May 2005.

[14] J. De Loera and S. Onn. Markov bases of three-way tables are arbitrarily complicated. *Journal of Symbolic Computation*, 41:173–181, 2005.

[15] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998. ISSN 0090-5364.

[16] P. Diaconis, D. Eisenbud, and B. Sturmfels. Lattice walks and primary decomposition. In *Mathematical essays in honor of Gian-Carlo Rota (Cambridge, MA, 1996)*, volume 161 of *Progr. Math.*, pages 173–193. Birkhäuser Boston, Boston, MA, 1998.

[17] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics.* Springer, New York, 2009. ISBN 978-3-7643-8904-8.

[18] R. A. Fisher. On the interpretation of 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[19] D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34(3):1463–1492, 2006.

[20] D. Grayson and M Stillman. Macaulay 2, a software system for research in algebraic geometry, 2006. `http://www.math.uiuc.edu/Macaulay2/`.

[21] G.-M. Greuel, G. Pfister, and H. Schoenemann. Singular: A computer algebra system for polynomial computations, 2006. `http://www.singular.uni-kl.de`.

[22] S. W. Guo and E. A. Thompson. Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics*, 48:361–372, 1992.

[23] S. J. Haberman. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.*, 83(402):555–560, 1988. ISSN 0162-1459.

[24] H. Hara, A. Takemura, and R. Yoshida. On connectivity of fibers with positive marginals in multiple logistic regression. *J of Multivariate Analysis*, 2009. In press.

[25] R. Hemmecke and P. Malkin. Computing generating sets of lattice ideals. *Journal of Symbolic Computation*, 44(10):1463–1476, 2009.

[26] S. Hosten and S. Sullivant. Ideals of adjacent minors. *Journal of Algebra*, 277:615–642, 2004.

[27] M. Huber, Y. Chen, I. Dinwoodie, A. Dobra, and M. Nicholas. Monte carlo algorithms for Hardy-Weinberg proportions. *Biometrics*, 62:49–53, 2006.

[28] M. Kreuzer and L. Robbiano. *Computational Commutative Algebra*. Springer, New York, 2000.

[29] F. Rapallo. Markov bases and structural zeros. *Journal of Symbolic Computation*, 41:164–172, 2006.

[30] F. Rapallo and M. P. Rogantin. Markov chains on the reference set of contingency tables with upper bounds. *Metron*, 65(1), 2007.

[31] F. Rapallo and R. Yoshida. Markov bases and subbases for bounded contingency tables. *Annals of Institution of Statistical Mathematics*, 2010. In press. Available at arxiv:0905.4841.

[32] J. Shao. *Mathematical Statistics*. Springer Verlag, New York, 1998.

[33] B. Sturmfels. *Gröbner Bases and Convex Polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI, 1996. ISBN 0-8218-0487-1.

[34] A. Takemura and S. Aoki. Distance reducing Markov bases for sampling from a discrete sample space. *Bernoulli*, 11(5):793–813, 2005.

[35] R Development Core Team. R: A language and environment for statistical computing, 2004. `http://www.R-project.org`.