Markov bases of conditional independence models for permutations

Villő Csiszár *

October 22, 2008

Abstract

The L-decomposable and the bi-decomposable models are two families of distributions on the set S_n of all permutations of the first n positive integers. Both of these models are characterized by collections of conditional independence relations. We first compute a Markov basis for the L-decomposable model, then give partial results about the Markov basis of the bi-decomposable model. Using these Markov bases, we show that not all bi-decomposable distributions can be approximated arbitrarily well by strictly positive bi-decomposable distributions.
Keywords: conditional independence, Markov basis, closure of exponential family, permutation, L-decomposable
AMS Subject Classification: 62E10, 62H05, 60C05

1 Introduction

In this paper we are concerned with the study of random permutations. Examples include voters ranking candidates in an election, the order of books on a shelf after many readers have used them, or the order of first appearance of n prescribed words in a word association chain. Permutation data have received most attention in mathematical psychology, in the context of ranking data. For a review of data analysis and stochastic models for random permutations, see the collection of papers Fligner and Verducci [7], or the booklength treatment of Marden [11].

Conditional independence relations are widely used to model the distribution of random vectors $X = (X_1, \ldots, X_n)$ over the product state space $\mathcal{X} = \times_{i=1}^n \mathcal{X}_i$. A conditional independence model is the set of all probability distributions satisfying a collection of conditional independence relations. In the paper [3] we introduced two natural conditional independence models in the setting of random permutations $\Pi = (\Pi_1, \ldots, \Pi_n)$ of the integers $[n] = \{1, \ldots, n\}$. These two

*Eötvös Loránd University, 1117 Budapest, Pázmány Péter s. 1/C, Hungary. e-mail: villo@ludens.elte.hu

models are each other's "inverse" in the sense we will describe later, and distributions belonging to both models are called bi-decomposable. We characterized strictly positive bi-decomposable distributions, but did not deal with the case when some permutations are allowed to have zero probability. This latter case was the motivation of the present paper.

It turns out that Markov bases are a useful tool when dealing with not necessarily strictly positive distributions. A Markov basis essentially lists all polynomial relationships among the probabilities p_π, which are satisfied by all distributions $p = (p_\pi)$ belonging to the model. These relationships remain valid under limits, thus they are satisfied by all distributions belonging to the closure of the model as well.

Algebraic techniques have most extensively been applied to contingency table models, but much less to permutation models. The only work in this direction we are aware of is Diaconis and Eriksson [4]. We think it is interesting and important to broaden the field of application of these algebraic techniques. We mention here that another use of Markov bases is to generate from the conditional distribution of the data, given the sufficient statistics, and thus perform exact tests of model fit, see e.g. [5, 4].

In Section 2 we briefly describe the material from algebraic statistics we need, as well as our models. Section 3 contains our results about Markov bases. In Section 4 we show that there exist bi-decomposable distributions, which cannot be approximated arbitrary well by strictly positive bi-decomposable distributions.

2 Conditional independence models for permutation data

2.1 Algebraic background

The use of commutative algebra in statistics is relatively new, but it has already proved itself very fruitful. Algebraic techniques have first been used in the analysis of contingency tables, by Diaconis and Sturmfels in their 1998 paper [5]. Some recent papers include Geiger et al [9], who analyzed the structure of toric models for discrete sample spaces, with particular attention to graphical models, or Diaconis and Eriksson [4], who use Markov chain methods relying on ideal bases to generate from the conditional distribution of the data, given sufficient statistics. The first book on the subject is Pistone et al [12]. These methods make extensive use of computer algebra packages, which implement various algorithms for computing ideal bases.

We now give a short account of toric models, as defined and studied by Geiger et al [9]. The algebraic background on ideals and varieties can be found in the marvellous textbook by Cox, Little and O'Shea [1]. A more advanced book, with statistical applications is Sturmfels [14]. Let $\mathcal{X} = \{x_1, \ldots, x_s\}$ be a finite set. Define a model, i.e. a family of distributions on $\mathcal{X}$ via a $t \times s$ nonnegative integer matrix $M = (m_{ij})$ as follows: a probability distribution

$p = (p(x_1), \ldots, p(x_s))$ *belongs to the toric model* M, or p *factors* according to the toric model M, if

$$p(x_i) = c(\lambda) \prod_{j=1}^{t} \lambda_j^{m_{ji}} \ \forall i, \text{ for some } \lambda = (\lambda_1, \ldots, \lambda_t) \in [0, \infty)^t. \tag{1}$$

We will denote this model by $\mathrm{To}(M)$. The strictly positive distributions in $\mathrm{To}(M)$ form a discrete exponential family, denoted by $\mathcal{E}(M)$. In the sequel, the row space of M is always assumed to contain the vector $\mathbf{1} = (1, \ldots, 1)^T$, so the normalizing factor $c(\lambda)$ can be omitted in (1). We have the series of inclusions

$$\mathcal{E}(M) \subseteq \mathrm{To}(M) \subseteq \mathrm{cl}(\mathcal{E}(M)), \tag{2}$$

where cl() stands for closure in the usual coordinate-wise topology.

The distributions in $\mathrm{To}(M)$ and $\mathrm{cl}(\mathrm{To}(M)) = \mathrm{cl}(\mathcal{E}(M))$ can be characterized with a condition on their supports and a set of algebraic constraints. Introduce s indeterminates, also denoted $x_1, \ldots, x_s$, we work in the ring $R[x] = R[x_1, \ldots, x_s]$ of polynomials in s indeterminates with real coefficients. For a nonnegative integer vector $u = (u_1, \ldots, u_s)$, let $x^u = x_1^{u_1} x_2^{u_2} \cdots x_s^{u_s}$ be a monomial in s variables. For the matrix M, we define the nonnegative toric variety X_M as the set of all common solutions in $\mathbb{R}^s_{\geq 0}$ of the set of polynomial equations

$$x^u - x^v = 0 : \quad u, v \in \mathbb{N}^s,\ Mu = Mv. \tag{3}$$

X_M is a nonnegative algebraic variety; it is a toric variety, since every defining polynomial has exactly two terms. In [9] it is shown that a distribution p on $\mathcal{X}$ belongs to $\mathrm{cl}(\mathrm{To}(M))$ if and only if $p \in X_M$. The distribution p belongs to $\mathrm{To}(M)$ as well, if and only if its support is M-feasible. (The set $T \subset \{1, \ldots, s\}$ is called M-feasible, if for every $i \notin T$, $\mathrm{Supp}(m_i)$ is not contained in $\cup_{j \in T} \mathrm{Supp}(m_j)$, where m_i denotes the ith column vector of M and its support is $\mathrm{Supp}(m_i) = \{1 \leq k \leq t : m_{ki} \neq 0\}$.)

The nonnegative variety X_M depends only on the row space of M, but the M-feasible sets depend on the particular row vectors chosen to span this space. Rapallo [13] has shown that for all toric models $\mathrm{To}(M)$, there exists a maximal representation $M_{\max}$ such that $\mathrm{cl}(\mathrm{To}(M)) = \mathrm{To}(M_{\max})$. Therefore, the possible supports of the distributions in $\mathrm{cl}(\mathrm{To}(M))$ are the $M_{\max}$-feasible sets. $M_{\max}$ has the same row space as M, but in addition, $\mathrm{To}(M_{\max})$ is closed. The question whether $p \in X_M$ seems hard at first glance. However, it is often tractable. X_M is the set of common non-negative roots of all polynomials belonging to the toric ideal I_M generated by the binomials (3). The Hilbert basis theorem asserts that every ideal is finitely generated, and more importantly, there exist algorithms for computing a (small) basis of I_M. For example, the computer algebra system 4ti2 [8] does the job. One then only has to check whether all polynomials in this ideal basis vanish at p.

One may want to generate from the conditional distribution of the data, given the value of the sufficient statistics. Such conditional distributions arise in carrying out versions of Fisher's exact test for independence and goodness of

fit. Diaconis and Sturmfels [5] present algebraic algorithms for sampling from conditional distribution in the general setting of toric models. Given the model To(M) defined by (1), and an iid. sample $Z = (Z_1, \ldots, Z_m)$ from it, denote by f^Z the empirical frequency vector on $\mathcal{X}$, i.e $f^Z(x) = |\{1 \leq i \leq m : \ Z_i = x\}|$. Then the statistic $u = Mf^Z$ is sufficient for λ. Define

$$\mathcal{F}_u = \{f : \mathcal{X} \to \mathbb{N} : \ Mf = u\}.$$

We assume throughout that $\mathcal{F}_u$ is finite and non-empty. Then the distribution of f^Z, given u, has a hypergeometric distribution on $\mathcal{F}_u$, i.e. the probability of $f^Z = f$ is proportional to $\prod_x (f(x)!)^{-1}$. It is usually not directly feasible to generate from the hypergeometric distribution on $\mathcal{F}_u$. However, Markov chain techniques can be used, provided we can find a Markov basis for the problem.

Definition 1. A Markov basis for the model To(M) is a set of functions $f_i : \mathcal{X} \to \mathbb{Z}$, such that (i) $Mf_i = 0$ for all i and (ii) for all u, any two elements of the state space $\mathcal{F}_u$ can be connected by a path in $\mathcal{F}_u$, where each step is of form $f \to f \pm f_i$, with f_i an element of the Markov basis.

With a Markov basis in our hand, the usual Metropolis algorithm can be applied to sample from the desired distribution. These ideas, with the necessary modifications, can also be used to solve larger problems, where the calculation of a Markov basis is infeasible, see [5] for details. The basic result (Theorem 3.1 in [5]) states that a collection of functions f_i is a Markov basis if and only if the set $x^{f_i^+} - x^{f_i^-}$ generates the ideal I_M.

2.2 The models

In this section, we introduce our conditional independence models for random permutations. We repeat what is necessary from the paper [3], to which we refer the reader for further details and proofs. Let $v = (v(1), \ldots, v(s))$ be a vector, for the set of the ith to jth coordinates, and for the subvector of the ith to jth coordinates of v, introduce the notations

$$v\{i..j\} = \{v(i), \ldots, v(j)\}, \quad v(i..j) = (v(i), \ldots, v(j)), \quad 1 \leq i \leq j \leq s.$$

Let S_n stand for the symmetric group of all permutations π of $[n] = \{1, \ldots, n\}$. We denote a probability distribution on S_n by $p = \{p(\pi) : \pi \in S_n\}$, and let Π be a random permutation with distribution p, that is $P(\Pi = \pi) = p(\pi)$. The idea of *L-decomposability* first appears in [2], and was motivated by Luce's ranking postulate [10]. It expresses that conditional on the *set* of the first k coordinates of Π, these k coordinates and the remaining $n-k$ coordinates are independent, for all k.

Definition 2. Let Π be a random permutation with probability distribution p on S_n. Π or p is called *L-decomposable*, if for every $1 \leq k \leq n-1$ and $\pi \in S_n$

$$P\left(\Pi(k+1) = \pi(k+1) \mid \Pi(1..k) = \pi(1..k)\right) = \\ = P\left(\Pi(k+1) = \pi(k+1) \mid \Pi\{1..k\} = \pi\{1..k\}\right), \quad (4)$$

if the lefthandside is defined. Equivalently (see [2]), p is L-decomposabe, if there exists a non-negative function Λ and a constant c, such that for all $\pi \in S_n$

$$p(\pi) = c \prod_{k=0}^{n-1} \Lambda(\pi(k+1), \pi\{1..k\}). \tag{5}$$

The function Λ above is defined on the set $\mathcal{D} = \{(x, C) : C \subset [n],\, x \notin C\}$. The following definition was introduced in [3].

Definition 3. Let Π be a random permutation with probability distribution p on S_n. Π or p is called *bi-decomposable*, if both Π and Π^{-1} are L-decomposable.

In this paper, we only consider these two decomposable properties. In [3], we studied some other generalizations as well. The two statistical models we will study are the L-decomposable model, which is the family of all L-decomposable distributions on S_n, and the bi-decomposable model $\mathcal{B}$, which is the family of all bi-decomposable distributions on S_n. From now on, we suppose $n \geq 4$, since for $n \leq 3$, all distributions on S_n are L-decomposable and bi-decomposable as well. The L-decomposable model falls into the framework of toric models, as seen from formula (5). Denoting $a_n = |\mathcal{D}| = n2^{n-1}$, the model matrix L is $a_n \times n!$, whose rows are indexed by the pairs $(x, C) \in \mathcal{D}$, whose columns are indexed by the permutations $\pi \in S_n$, and its entries are given by

$$L((x, C), \pi) = \chi\{\pi\{1..|C|\} = C,\, \pi(|C|+1) = x\}, \tag{6}$$

where χ denotes the indicator function. We avoid denoting the indices of the matrix elements by subscripts for the sake of readability. By definition, the L-decomposable model is closed, so we have $\mathrm{To}(L) = \mathrm{cl}(\mathcal{E}(L))$.

The picture is more complex for the bi-decomposable model, which is again, by definition, closed. It is the intersection of two toric models: the L-decomposable, and the L^{-1}-decomposable models. The latter is obtained by the inversion of the former: a distribution p on S_n is L^{-1}-decomposable, if the distribution q, defined as $q(\pi) = p(\pi^{-1})$, is L-decomposable. We will show that the bi-decomposable model is not toric, i.e. there is no matrix M with non-negative integer elements such that $\mathcal{B} = \mathrm{To}(M)$. In [3] we showed that the family of strictly positive bi-decomposable distributions is an exponential family with a $0-1$ model matrix B. The rows of B are vectors denoted by $\rho_{aq}^{k\ell}$, where k, ℓ, a, q take all possible values. The row-vector $\rho_{aq}^{k\ell}$ is the indicator vector of the event that among the first k coordinates of π there are exactly a which are less than or equal to ℓ, moreover the pair $(\pi(k), \pi^{-1}(\ell)) \in A_q^{k\ell}$, where the pairs (x, y) belong to the sets $A_q^{k\ell}$ according to the following pattern.

$$\begin{array}{lll} A_1^{k\ell} : x > \ell, y < k, & A_2^{k\ell} : x < \ell, y < k, & A_3^{k\ell} : x < \ell, y > k, \\ A_4^{k\ell} : x > \ell, y > k, & A_5^{k\ell} : x = \ell, y = k. & \end{array} \tag{7}$$

In [3], as one of the main results, we also calculated the rank of B and L (the number of free parameters), and identified basis vectors in their row spaces

(possible parametrizations). Thus, in our previous studies, we dealt with $\mathcal{E}(B)$, which we now continue by studying the models $\mathrm{To}(B)$, $\mathrm{cl}(\mathrm{To}(B))$ and $\mathcal{B}$. To see what happens when zero probabilities are allowed is not only of theoretical interest and importance: for a general dataset, a maximum likelihood estimate does not necessarily exist in $\mathcal{E}(B)$ or $\mathrm{To}(B)$, but it uniquely exists in the closures of these models. In practice, when n is moderately large, then most datasets have less than $n!$ elements, thus the empirical distribution necessarily has zeros. Some of these zeros may be present in the maximum likelihood estimate as well.

3 Markov bases

In this section we give some results regarding Markov bases in the two toric models $\mathrm{To}(L)$ and $\mathrm{To}(B)$.

3.1 L-decomposable model

In the case of $\mathrm{To}(L)$, the constraints defining conditional independence can be translated into an equivalent set of polynomial constraints. Let C be a subset of $[n]$ of size k, where $2 \leq k \leq n-2$ and let ρ_1, ρ_2 be two distinct permutations of C, and τ_1, τ_2 be two distinct permutations of $[n] \setminus C$. For all such choices, define the binomial

$$x_{(\rho_1,\tau_1)} x_{(\rho_2,\tau_2)} - x_{(\rho_1,\tau_2)} x_{(\rho_2,\tau_1)}, \tag{8}$$

where (ρ, τ) is the concatenation of the two permutations. These binomials obviously belong to I_L. According to Theorem 1, no other polynomial constraints are needed, since these generate the ideal I_L. This theorem can be proved either directly, or by referring to more general results. First we give a simple direct proof.

Theorem 1. *The system of polynomials in* (8) *generates the ideal* I_L.

Proof. It is enough to show that the functions defined by system (8) constitute a Markov basis of moves, i.e. any two datasets with the same set of sufficient statistics can be connected by a path using these moves. In the proof, we will construct such a path. Define the index set

$$\mathcal{I} = \{i = i(C, \pi_1, \pi_2, \rho_1, \rho_2) : C \subset [n], \pi_1, \pi_2 \in S_C,\ \pi_1 \neq \pi_2, \\ \rho_1, \rho_2 \in S_{[n]\setminus C},\ \rho_1 \neq \rho_2\}, \tag{9}$$

where S_C denotes the set of permutations of the elements of C. For $i \in \mathcal{I}$, define $f_i : S_n \to \mathbb{Z}$ as

$$f_i(\pi_1, \rho_1) = f_i(\pi_2, \rho_2) = -1,\ f_i(\pi_1, \rho_2) = f_i(\pi_2, \rho_1) = 1,\ f_i(\sigma) = 0 \text{ otherwise.} \tag{10}$$

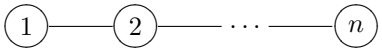
It is obvious that $Lf_i = 0$ for all i, so the only thing to show is connectedness. Let $u, v \in \mathbb{N}^{n!}$ be two frequency vectors such that $Lu = Lv$. We will construct a path from u to v using the moves f_i. For any vector $\mathbf{j} = (j_1, \ldots, j_k)$ of

distinct integers between 1 and n, denote by $B_{\mathbf{j}} \subset S_n$ the set of permutations beginning with the specified string, i.e. $\pi(s) = j_s$ for $1 \le s \le k$. We will use the shorthand notation $u(B_{\mathbf{j}}) = \sum_{\pi \in B_{\mathbf{j}}} u(\pi)$. First of all, $u(B_j) = v(B_j)$ for all j, since $u(B_j) = (Lu)(j, \emptyset)$, one of the sufficient statistics supposed to be equal for u and v. Suppose by induction that using the steps f_i, we have already arrived at a frequency vector u^k, such that for all $\mathbf{j}$ of length (less than or) equal to k, $u^k(B_{\mathbf{j}}) = v(B_{\mathbf{j}})$.

Now take any C with $|C| = k$, and construct a $k! \times (n-k)$ two-way contingency table for both u^k and v, where the rows are labelled by the permutations of C and the columns by the elements in $[n] \setminus C$. In the cell labelled by $(\mathbf{j}, x)$, enter the value $u^k(B_{\mathbf{j},x})$ and $v(B_{\mathbf{j},x})$ respectively. The u^k-table and the v-table have the same marginals: the row sums are equal by the induction hypothesis, while the column sums are equal since they are just the sufficient statistics $(Lu^k)(C, x) = (Lv)(C, x)$.

It is well-known (see e.g.[5]), that for two-way contingency tables, a Markov basis is provided by the following moves. There is a move for all $r_1 \neq r_2$ and $c_1 \neq c_2$, namely, we subtract 1 from the (r_1, c_1) and to the (r_2, c_2) entries of the table, and add 1 to the (r_1, c_2) and to the (r_2, c_1) entries. Applying this to our contingency table, we can transform the u^k-table into the v-table by such moves. These moves are realised on a finer scale: if we choose rows $\mathbf{j}$ and $\mathbf{j}'$, and columns x and x', such that $u^k(B_{\mathbf{j},x}) > 0$ and $u^k(B_{\mathbf{j}',x'}) > 0$ (the contingency-table move can be performed), then there exist permutations $\mathbf{g}$ and $\mathbf{g}'$ of the sets $[n] \setminus (C \cup x)$ and $[n] \setminus (C \cup x')$ respectively, such that $u^k(\mathbf{j}, x, \mathbf{g}) > 0$ and $u^k(\mathbf{j}', x', \mathbf{g}') > 0$. Then with $i = i(C, \mathbf{j}, \mathbf{j}', (x, \mathbf{g}), (x', \mathbf{g}'))$, making move f_i on u^k results in the desired move on the contingency table. Performing this contingency table transformation for all C of cardinality k, we get the new frequency vector u^{k+1} on the path. Finally, $u^n = v$. □

We now outline an alternative proof. It was shown in Dobra [6] that Markov bases for decomposable graphical models can always be built up from the basic $\pm\mp$ moves. Moreover, Geiger et al. [9] proved that for a graphical model, the existence of a degree-2 Markov basis is equivalent to the decomposability of the model. In addition, in the case of decomposable graphical models, the polynomial relations expressing the global Markov property form a Markov basis. Now, Π is an L-decomposable random permutation, if and only if the variables $Z_k = \Pi\{1..k\}$, $1 \le k \le n$ form a Markov chain, i.e. the variables Z_k satisfy the global Markov property with respect to the following graph:



If we allowed the vector $Z = (Z_1, \ldots, Z_n)$ to take its values in the whole product of the sets $\mathcal{Z}_k = \{A_k \subseteq [n] : |A_k| = k\}$, then the corresponding Markov basis $\mathcal{M}$ could be obtained from the above results of Dobra and Geiger et al. However, we now have a certain pattern of zeros, namely, only the cells $(A_1, \ldots, A_n)$ with $A_1 \subset \cdots \subset A_n$ are allowed to have positive probability. Let us call these cells legal, the other cells illegal. The idea is to restrict the

Markov basis $\mathcal{M}$ to the legal cells, i.e. to throw away every basis polynomial, which contains variables corresponding to illegal cells. In general, the remaining polynomials $\mathcal{N}$ are not a Markov basis. However, it is easy to show that if the pattern of zeros is admissible, then $\mathcal{N}$ is a Markov basis. By admissible we mean the following. It is well-known that the vector Z satisfies the global Markov property with respect to the above graph, if and only if its distribution factorizes:

$$P(Z_1 = A_1, \ldots, Z_n = A_n) = \prod_{j=1}^{n-1} \theta_j(A_j, A_{j+1}).$$

The pattern of zeros is admissible, if it can be achieved by setting some parameters $\theta_j(A_j, A_{j+1})$ to zero. In our case, a cell is illegal if and only if $A_j \not\subset A_{j+1}$ for some j, so the pattern of zeros is achieved by setting $\theta_j(A_j, A_{j+1}) = 0$ for all such pairs.

The advantage of the direct proof of Theorem 1 is that it shows how two frequency vectors with the same sufficient statistics can be connected, using the Markov moves. In addition, this proof can be carried over to other situations, where the pattern of zeros is not admissible. For example, take the toric model of distributions p of the form

$$p(\pi) = \prod_{k=1}^{n} \Lambda(\pi\{1..k\}).$$

A Markov basis of this model can be described using an argument similar to the one in the proof of Theorem 1.

3.2 Minimality

We say that a Markov basis is minimal, if omitting any of the moves, the remaining ones no longer form a Markov basis. Though in practical applications a non-minimal Markov basis may produce a more rapidly mixing Markov chain, determining the minimal Markov bases for a problem can be of theoretic interest. Takemura and Aoki [16] have given some characterizations of minimal Markov bases, and a necessary and sufficient condition for the uniqueness of the minimal Markov basis. It is straightforward to apply their results to the L-decomposable model, as follows. Given two permutations $\pi_0 \neq \pi_1$, we say that they *part* at k, if $\pi_0\{1..k\} = \pi_1\{1..k\}$, but $\pi_0(k+1) \neq \pi_1(k+1)$, where the possible parting places are $0 \leq k \leq n-2$. If π_0 and π_1 part at exactly $h \geq 2$ places, then they can be partitioned as $\pi_i = (\pi_i^1, \ldots, \pi_i^h)$, $i = 0, 1$, where for each $1 \leq k \leq h-1$, the sub-permutations $(\pi_0^1, \ldots, \pi_0^k)$ and $(\pi_1^1, \ldots, \pi_1^k)$ permute the same elements, but the first elements of π_0^{k+1} and π_1^{k+1} are different. For such a pair π_0, π_1, let f^0 be the frequency vector on S_n with

$$f^0(\pi_0) = f^0(\pi_1) = 1, \quad f^0(\sigma) = 0 \text{ if } \sigma \neq \pi_0, \pi_1.$$

Then there are exactly 2^{h-1} frequency vectors f with $Lf = Lf^0$. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_h)$ be a vector with $\varepsilon_k = 0, 1$ $(1 \leq k \leq h)$, and let $\pi_\varepsilon = (\pi_{\varepsilon_1}^1, \ldots, \pi_{\varepsilon_h}^h)$.

Then for the frequency vector

$$f^\varepsilon(\pi_\varepsilon) = f^\varepsilon(\pi_{1-\varepsilon}) = 1, \quad f^\varepsilon(\sigma) = 0 \text{ otherwise},$$

we have $Lf^\varepsilon = Lf^0$, and these are the only frequency vectors with this property. It follows from Theorem 2.1 in [16] that a minimal Markov basis has to contain $2^{h-1}-1$ moves of the form $f^\varepsilon - f^{\tilde{\varepsilon}}$, which connect the frequency vectors f^ε into a tree. For $h = 2$, this can only be done in one way, however, for $h > 2$, there are as many ways as there are (undirected) trees on 2^{h-1} labelled vertices. Of course, not all moves $f^\varepsilon - f^{\tilde{\varepsilon}}$ are of form (10), since they can contain several "crossovers". However, even if we are only allowed to choose from (10), we can construct many trees. For $n = 4, 5$, the minimal Markov basis is unique, since $h \leq 2$ in these cases. For $n \geq 6$, there is no unique minimal Markov basis. Thus we have shown the following.

Theorem 2. *In the L-decomposable toric model, for $n = 4, 5$ the unique minimal Markov basis is given by the binomials in* (8)*, while for $n \geq 6$, this Markov basis is not minimal. In the latter case, there is no unique minimal Markov basis.*

Regarding the number of elements in a minimal Markov basis, we note that for $n = 4$ and 5, the minimal basis contains 6 and 270 elements respectively.

3.3 The toric model To(B)

For this model, in the case $n = 4$, the package 4ti2 [8] arrived at a minimal Markov basis very quickly, which we now describe. For this n, the six binomials in (8) are the unique minimal basis of I_L. By inverting the problem, we get six analogous binomials, of which only four are new. This gives altogether 10 binomials, which are satisfied by all bi-decomposable distributions $p \in \mathcal{B}$. The minimal basis of I_B contains 18 binomials, the 10 degree two binomials defining $\mathcal{B}$, plus 8 binomials of degree four. One of these additional binomials is

$$x_{1324}x_{2431}x_{3241}x_{4123} - x_{1423}x_{2341}x_{3124}x_{4231}. \tag{11}$$

We do not list the others, since these eight binomials form a complete orbit in the following sense. We have shown in [3] that cl(To(B)) is invariant under several transformations, namely under the reordering of the probabilities $p(\pi)$ according to the inversion of permutations, and right or left multiplication by the subgroup of S_n generated by the leading transposition $(213\ldots n)$ and the reversing permutation $(n(n-1)\ldots 21)$. We will denote this subgroup by H_8, since it has eight elements. More formally, define the transformation group $\mathcal{T}$ acting on S_n, with

$$\mathcal{T} = \{\tau_{(\rho_1,\varepsilon,\rho_2)} : (\rho_1, \varepsilon, \rho_2) \in H_8 \times \{-1, +1\} \times H_8\},$$

and action $\tau_{(\rho_1,\varepsilon,\rho_2)}(\pi) = \rho_1\pi^\varepsilon\rho_2$. The elements of $\mathcal{T}$ act on distributions by permuting the probabilities: $p^\tau(\pi) = p(\tau(\pi))$ if $\tau \in \mathcal{T}$. Invariance of the model means that

$$p \in \text{cl(To}(B)) \Rightarrow p^\tau \in \text{cl(To}(B)) \ \forall \tau \in \mathcal{T}.$$

The transformation group also acts naturally on the polynomials in the ring $R[x_\pi : \pi \in S_n]$: a transformation τ takes the polynomial f to the polynomial f^τ by replacing all variables x_π by $x_{\tau(\pi)}$. The orbit of f is the set $\mathcal{O}(f) = \{f^\tau : \tau \in \mathcal{T}\}$. Of course, if a polynomial constraint is satisfied in cl(To(B)), then, by invariance, all polynomials in its orbit are satisfied as well. It can be checked that the orbit of the binomial in (11) consists of eight elements. Transforming (11) in ratio form, we get

$$\left(\frac{x_{1324}}{x_{3124}}\right) / \left(\frac{x_{1423}}{x_{4123}}\right) = \left(\frac{x_{2341}}{x_{3241}}\right) / \left(\frac{x_{2431}}{x_{4231}}\right)$$

Thus, the binomial in (11) can be interpreted as the denominator-cleared version of the equality of the ratios of certain odds-ratios. All other binomials in the orbit inherit this interpretation.

We will see in Section 4 that the Markov basis for $n = 4$ makes it possible to prove some results for all $n \geq 4$. However, one wants to find a Markov basis also for larger values of n. After the first version of this paper was submitted, Johannes Rauh managed to calculate a minimal Markov basis for $n = 5$ with 4ti2. The calculations took about 195 hours on a 2.6 GHz machine. The Markov basis contained almost fifty thousand elements, and the largest degree was 8. This suggest that the problem becomes very complicated as n grows: even for $n = 5$, much more work would be needed to understand at least partially the structure of the Markov basis.

4 Strict inclusions

In this section we study the relationship between the models To(B), cl(To(B)), and $\mathcal{B}$. The main result is the following.

Theorem 3. *For all $n \geq 4$, To(B) $\subsetneq$ cl(To(B)) $\subsetneq \mathcal{B}$.*

Proof. In both cases, we first give a counterexample for $n = 4$, which can then be extended to any $n > 4$ using Lemma 1 after the theorem. The first statement is proved by the following example for $n = 4$. Let $T = T_4 \subset S_4$ consist of the following 16 permutations:

$$\begin{array}{cccccccc} 1234, & 1243, & 1324, & 1342, & 1423, & 2134, & 2143, & 2341, \\ 3241, & 3412, & 3421, & 4231, & 4213, & 4321, & 4312, & 4123. \end{array} \tag{12}$$

It is easily seen that this set is not B-feasible, since $\cup_{\pi \in T} \mathrm{Supp}(b(\cdot, \pi))$ is the whole set of rows of B. Therefore, the distribution p, which assigns probability $1/16$ to every $\pi \in T$ does not factor according to B. However, it can be checked directly that the distributions $p_m \in \mathcal{E}(B)$, form a convergent sequence with limit p, where $\log p_m = mv + c(m)\mathbf{1}$, and v is given by

$$v = \nu_1^{11} - \nu_1^{12} - \nu_1^{21} + 2\nu_2^{22} - \rho_{25}^{22} + \rho_{25}^{23} + \rho_{25}^{32} - \nu_3^{33} + \rho_{35}^{33},$$

where for any k, ℓ, a, we define $\nu_a^{k\ell} = \sum_{q=1}^5 \rho_{aq}^{k\ell}$.

Let us turn to the second relation, still fixing $n = 4$. It is enough to give explicitly a distribution $x = (x_1, \ldots, x_{n!})$, which is a root of all of the degree two binomials in the Markov basis of I_B, but fails to satisfy at least one of the other binomials, for example (11). Such an example is the uniform distribution on the set of the seven permutations

$$1324, \quad 2341, \quad 2431, \quad 3241, \quad 3124, \quad 4231, \quad 4123. \tag{13}$$

Turning to general $n > 4$, fix any permutation σ of the numbers $5, \ldots, n$. First let $\rho_i = (\pi_i, \sigma)$ for $1 \leq i \leq 16$, where the π_i are the sixteen permutations in (12). It is easy to see that the uniform distribution on ρ_i is not an element of $\mathrm{To}(B)$, since the B-feasibility closure of the set $T_n = \{\rho_i : 1 \leq i \leq 16\}$ is the set $S_4 \times \sigma = \{(\pi, \sigma) : \pi \in S_4\}$. Observe that for $k, \ell \leq 4$,

$$\{(a^{k\ell}(\pi), q^{k\ell}(\pi)) : \pi \in T_n\} = \{(a^{k\ell}(\pi), q^{k\ell}(\pi)) : \pi \in S_4 \times \sigma\},$$

while for all other pairs k, ℓ, the statistics $(a^{k\ell}, q^{k\ell})$ are constant on $T_4 \times \sigma$, only depending on σ:

$$(a^{k\ell}(\pi), q^{k\ell}(\pi)) = c(k, \ell, \sigma) \, \forall \pi \in S_4 \times \sigma, \text{ if } k > 4 \text{ or } \ell > 4.$$

We use this fact to show that the uniform distribution on the ρ_i is in $\mathrm{cl}(\mathcal{E}(B))$. Consider the vector

$$v_n = \nu_1^{11} - \nu_1^{12} - \nu_1^{21} + 2\nu_2^{22} - \rho_{25}^{22} + \rho_{25}^{23} + \rho_{25}^{32} - \nu_3^{33} + \rho_{35}^{33},$$

where the vectors on the right are of length $n!$, and the coordinates are indexed by $\pi \in S_n$. We have $v_n(\pi, \sigma) = v_4(\pi)$ for every $\pi \in S_4$. For $k, \ell > 4$, define the vectors

$$w_n^{k\ell} = \mathbf{1} - \rho^{k\ell}_{a^{k\ell}(\pi,\sigma), q^{k\ell}(\pi,\sigma)},$$

with $\pi \in S_4$ arbitrary. Thus, these vectors are zero on all permutations in $S_4 \times \sigma$, but for any other permutation in $\pi \in S_n$, there is at least one pair k, ℓ, such that $w_n^{k\ell}(\pi) = 1$. Thus letting p_m be the distribution in $\mathcal{E}(B)$ with

$$\log p_m = m(v_n - c \sum_{k,\ell > 4} w_n^{k\ell}) + c(m)\mathbf{1},$$

with c large enough, we see that p_m converges to the uniform distribution on T_n.

Now let $\rho_i = (\pi_i, \sigma)$ for $1 \leq i \leq 7$, where the π_i are the seven permutations in (13). By the first part of Lemma 1, the uniform distribution on ρ_i is in $\mathcal{B}$. However, it is not in $\mathrm{cl}(\mathcal{E}(B))$. Suppose, indirectly, that it is. Then there is a convergent sequence to it in $\mathcal{E}(B)$, and by Lemma 1, its σ-restriction to S_4 is in $\mathcal{E}(B)$ for $n = 4$. However, this sequence of restricted distributions converges to the uniform distribution on the π_i's, which is a contradiction. □

In the following lemma, whose proof we leave to the reader, $*$-decomposable can mean any of the two types (L, bi), but is the same within one statement.

Lemma 1. *(i) Let $n < m$, and let p be a $*$-decomposable distribution on S_n. For any permutation σ of the numbers $n+1, \ldots, m$, let q be the following distribution on S_m, called the σ-lifting of p:*

$$q(\pi) = \begin{cases} p(\pi(1..n)) & \text{if } \pi\{1..n\} = \{1..n\} \text{ and } \pi(n+1..m) = \sigma \\ 0 & \text{otherwise.} \end{cases}$$

Then q is a $*$-decomposable distribution on S_m.

(ii) Let $n < m$ and q a $*$-decomposable distribution on S_m with $\sum_{\rho\{1..n\}=\{1..n\}} q(\rho) > 0$. For any permutation σ of the numbers $n+1, \ldots, m$ such that $\sum_{\rho(n+1..m)=\sigma} q(\rho) > 0$, let p be the following distribution on S_n, called the σ-restriction of q: $p(\pi) = c \cdot q(\pi, \sigma)$, where c is a normalizing constant. Then p is a $*$-decomposable distribution on S_n.

We have seen that $\mathrm{To}(B)$ is not closed, but as we mentioned earlier, there exists a maximal representation $B_{\max}$ such that $\mathrm{To}(B_{\max}) = \mathrm{cl}(\mathrm{To}(B))$. There is a function in 4ti2 for calculating a maximal representation of a model matrix M. The idea is the following. The vectors with nonnegative integer coordinates in the row space of M form a lattice $\mathcal{L}$. A set $\mathcal{H} = \{v_1, \ldots, v_h\}$ is a Hilbert basis of the lattice, if every $v \in \mathcal{L}$ can be written in the form $v = \sum_{i=1}^{h} c_i v_i$ with nonnegative integer coefficients c_i. Arranging the vectors in $\mathcal{H}$ as the rows of a matrix, we get a maximal representation $M_{\max}$, as shown in [13].

For $n = 4$, the 4ti2 output was a $0-1$ matrix with 32 rows. Out of these 32 rows, 24 correspond to sufficient statistics $\rho_{aq}^{k\ell}$ for some values of k, ℓ, a, q. The remaining eight rows have 8 ones each, for example, one row has zeros at the 16 permutations of the counterexample (12), and ones at the remaining 8 permutations. This shows that the set (12) is indeed $B_{\max}$-feasible, and thus the uniform distribution on it is in $\mathrm{To}(B_{\max})$. This maximal representation provides another proof that the uniform distribution on the set of permutations in (13) is not in $\mathrm{cl}(\mathrm{To}(B))$, since this set is not $B_{\max}$-feasible, in fact it becomes feasible only after adding the permutation 1423 to it. It remains for future work to understand the eight new rows of $B_{\max}$) combinatorically, as well as to calculate a maximal representation of B for $n = 5$.

Acknowledgements I would like to thank Johannes Rauh for calculating the Markov basis of the bi-decomposable model for $n = 5$, and the two referees for their helpful suggestions and comments.

References

[1] Cox, D., Little, J., O'Shea, D. (1992) *Ideals, Varieties, and Algorithms.* Springer, New York.

[2] Critchlow, D. E., Fligner, M. A. and Verducci, J. S. (1991). Probability models on rankings. *J. Math. Psych.* **35** 294–318.

[3] Csiszár, V. Conditional independence relations and log-linear models for random matchings. *Acta Mathematica Hungarica*, Online First.

[4] Diaconis, P., Eriksson, N. (2006) Markov bases for noncommutative Fourier analysis of ranked data. *Journal of Symbolic Computation* **41** 173–181.

[5] Diaconis, P., Sturmfels, B. (1998) Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397.

[6] Dobra, A. (2003) Markov bases for decomposable graphical models. *Bernoulli* **9** 1093–1108.

[7] Fligner, M.A., Verducci, J.S (eds.) (1993) *Probability Models and Statistical Analyses for Ranking Data.* Springer-Verlag, New York.

[8] 4ti2 team. *4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces.* Available at www.4ti2.de.

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

[10] Luce, R. D. (1959). *Individual Choice Behavior*, Wiley, New York.

[11] Marden, J.I. (1995) *Analyzing and Modelling Rank Data.* Chapman&Hall, London.

[12] Pistone, G., Riccomagno, E. and Wynn, H. P. (2000) *Algebraic Statistics.* Chapman&Hall/CRC, Bocan Raton.

[13] Rapallo, F. (2007) Toric statistical models: parametric and binomial representations. *Annals of the Institute of Statistical Mathematics* **59** 727–740

[14] Sturmfels, B. (1996) *Gröbner bases and convex polytopes.* Amer. Math. Soc., Providence, RI.

[15] Sullivant, S. (2005) *Toric Ideals in Algebraic Statistics.* PhD thesis, University of California, Berkeley.

[16] Takemura, A. and Aoki, S. (2004) Some characterizations of minimal Markov basis for sampling from discrete conditional distributions. *Ann. Inst. Statist. Math.* **56** 1–17.