Hierarchical models for random permutations

Villő Csiszár (villo@ludens.elte.hu)
Eötvös Loránd University, Budapest, Hungary

Algebraic Statistics
15th December, 2008, MSRI, Berkeley, USA

Random permutations

S_n is the set of all permutations of $[n] = \{1, \ldots, n\}$
$\Pi : \Omega \to S_n$ is a random permutation with $p(\pi) = P(\Pi = \pi)$

Examples of permutation data:

→ the order of candidates in an election, as given by voters

→ the order of the strength of association of words to a target word, as given by subjects

→ the assignment of two lists of words, as given by subjects

In these examples, candidates/words are labelled by $1, \ldots, n$.

Hierarchical models for categorical data

Let $I = I_1 \times \cdots \times I_n$ be a finite sample space.
Let $X = (X(1), \ldots, X(n))$ be a random vector taking values in I.
Denote the distribution of X by $P(X = x) = p(x)$.
For a vector $x \in I$ and a subset $A \subseteq [n]$, let $x(A) = (x(i) : i \in A)$.
Let $\mathcal{A}$ be a collection of subsets of $[n]$.

We say that p belongs to the hierarchical model with generator-set $\mathcal{A}$, if p is of form

$$p(x) = \prod_{A \in \mathcal{A}} \theta_A(x(A)) \quad \forall x \in I.$$

Hierarchical models for permutation data

Previous models are applicable with $I = [n]^n$, with structural zeros. Alternatively: Define hierarchical models, where the generators are not subsets $A \subseteq [n]$, but *product partitions* $\mathcal{R} \times \mathcal{C}$ of $[n] \times [n]$.

Let $\mathcal{R} = (R_1, \ldots, R_r)$ and $\mathcal{C} = (C_1, \ldots, C_c)$ be two (ordered) partitions of [n]. The marginal $\pi(A)$ is replaced by the $r \times c$ matrix $\pi(\mathcal{R} \times \mathcal{C})$, whose ijth element is

$$(\pi(\mathcal{R} \times \mathcal{C}))_{ij} = |\{1 \leq k \leq n : \; k \in R_i, \pi(k) \in C_j\}|.$$

For $A = \{a_1, \ldots, a_j\} \subseteq [n]$, the marginal $\pi(A)$ is equivalent to $\pi(\mathcal{R} \times \mathcal{C})$ with $\mathcal{R} = (\{a_1\}, \ldots, \{a_j\}, A^c)$ and $\mathcal{C} = (\{1\}, \ldots, \{n\})$, so traditional hierarchical models fit into this new frame as well.

Example: L-decomposable distributions

Luce's ranking postulate for random orderings:
the ordering of the alternatives is the result of repeated selections of the best alternative from the remaining set of alternatives:

$$p(\pi) = \prod_{k=1}^{n} p(\pi(k) \mid \{\pi(k), \dots, \pi(n)\}),$$

where $p(x \mid C) = P(x$ is chosen from $C)$.

Random permutations (orderings) Π satisfying Luce's ranking postulate are called L-decomposable by Critchlow et al. (1991).

L-decomposable distributions (cont.)

The L-decomposable distributions form a hierarchical model $\mathcal{H}_L$ with $n-2$ generators $\mathcal{R}_i \times \mathcal{C}$, $i = 2, \ldots, n-1$, where

$$\mathcal{R}_i = (\{1, \ldots, i-1\}, \{i\}, \{i+1, \ldots, n\}) \text{ and } \mathcal{C} = (\{1\}, \ldots, \{n\}).$$

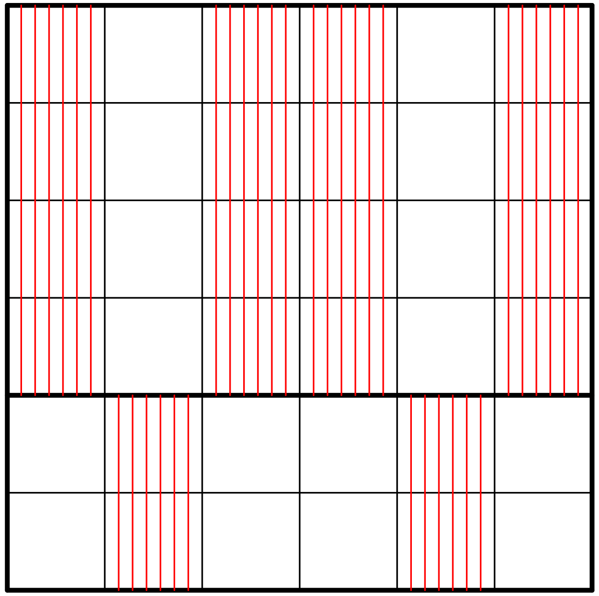
Theorem. A random permutation Π is L-decomposable, if and only if for every k, the first k and last $n-k$ elements of Π are conditionally independent, given the set of the first k elements. The distribution of Π is parametrized by the conditional probabilities

$$P(\Pi(k+1) = x \,|\, \{\Pi(1), \ldots, \Pi(k)\} = C),$$

where $|C| = k$ is a subset of $[n]$ and $x \notin C$. The ML estimate of these parameters is given by the corresponding empirical conditional probabilities.

L-decomposability: illustration

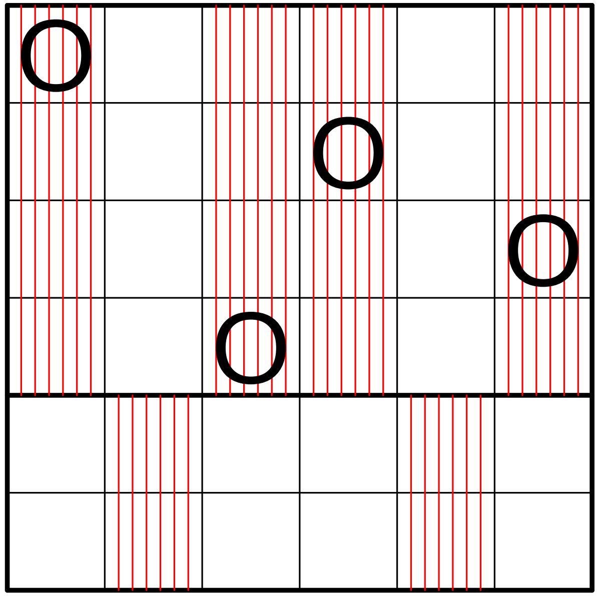
chessboard: rows = ranks, columns = alternatives



Given that $\{\Pi(1), \Pi(2), \Pi(3), \Pi(4)\} = \{1, 3, 4, 6\}$,

L-decomposability: illustration

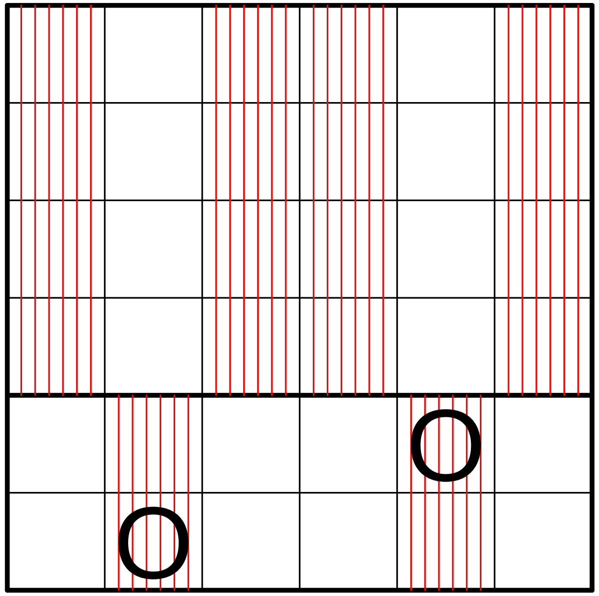
chessboard: rows = ranks, columns = alternatives



Given that $\{\Pi(1), \Pi(2), \Pi(3), \Pi(4)\} = \{1, 3, 4, 6\}$, $(\Pi(1), \ldots, \Pi(4))$

L-decomposability: illustration

chessboard: rows = ranks, columns = alternatives



Given that $\{\Pi(1), \Pi(2), \Pi(3), \Pi(4)\} = \{1, 3, 4, 6\}$,

and $(\Pi(5), \Pi(6))$ are independent.

Markov basis of the L-decomposable model

Every hierarchical model has a 0-1 model matrix, thus it is also a toric model. The strictly positive (s.p.) part of a hierarchical model is an exponential family.

Define the index set

$$\begin{aligned}\mathcal{I} = \{i = i(C, \pi_1, \pi_2, \rho_1, \rho_2) : C \subset [n],\ 2 \le |C| \le n-2, \\ \pi_1, \pi_2 \in S_C,\ \pi_1 \ne \pi_2,\ \rho_1, \rho_2 \in S_{[n]\setminus C},\ \rho_1 \ne \rho_2\},\end{aligned}$$

where S_C denotes the set of permutations of the elements of C. Let (π, ρ) denote the concatenation of two permutations.

Theorem. The moves f_i form a Markov basis of the L-decomposable model, where for $i \in \mathcal{I}$,

$$\begin{gathered} f_i(\pi_1, \rho_1) = f_i(\pi_2, \rho_2) = -1, \\ f_i(\pi_1, \rho_2) = f_i(\pi_2, \rho_1) = 1, \\ f_i(\sigma) = 0 \text{ for all other } \sigma \in S_n. \end{gathered}$$

Theorem. The Markov basis $\{f_i : i \in \mathcal{I}\}$ is unique minimal for $n = 4, 5$. For $n \geq 6$, it is not minimal, and there is no unique minimal Markov basis.

Proof by direct application of the condition in Takemura and Aoki (2004).
Number of moves in the unique minimal Markov basis:
6 moves for $n = 4$, 270 moves for $n = 5$.

Word association data, L-decomposable ML estimate

Data: 129 college students ranked the words score, instrument, solo, benediction, suit according to the strength of association with the target word song.
24 orderings observed, the two most frequent:
solo, instrument, score, benediction, suit (34 times)
solo, instrument, benediction, score, suit (24 times)

Fitting the L-decomposable model:
ML estimate allocates positive probability to 33 permutations.
Chi-square statistic is $\chi^2 = 14.58$. Is this small or large?

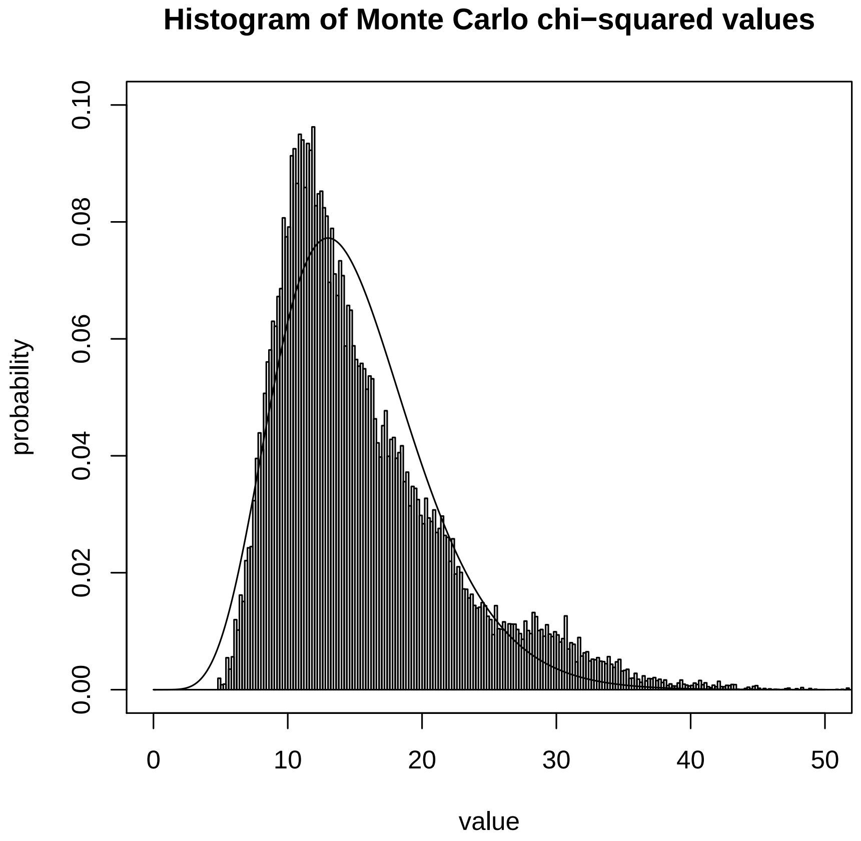
Word association: Testing L-decomposability

D.f.: If the support of the ML estimate is taken as known a priori, then df$= 33 - 17 - 1 = 15$ and $p = 0.48$. But: Chi-square approximation probably not very good due to small sample sizes.

Monte Carlo assessment of fit: Generate data from the conditional distribution of the data, given the sufficient statistics.

1. Direct generation: we generated 1000 datasets, and obtained $p = 0.44$.

2. Markov chain: after an initial warm up (10^5 steps), we ran 10^6 steps, and obtained $p = 0.45$. This p-value seemed quite stable (varying between 0.43 and 0.45).



Histogram of χ^2 statistics for data generated by the Markov chain approach.

Example: bi-L-decomposable distributions

Ordering data: we can work either with orderings, or rankings: group-theoretically, these are each other's inverse.

Usually, one supposes that the orderings are L-decomposable. However, if there is a natural ordering of the alternatives, then L-decomposability of the rankings is also plausible.

Assignment data: the permutation and its inverse are totally symmetric.

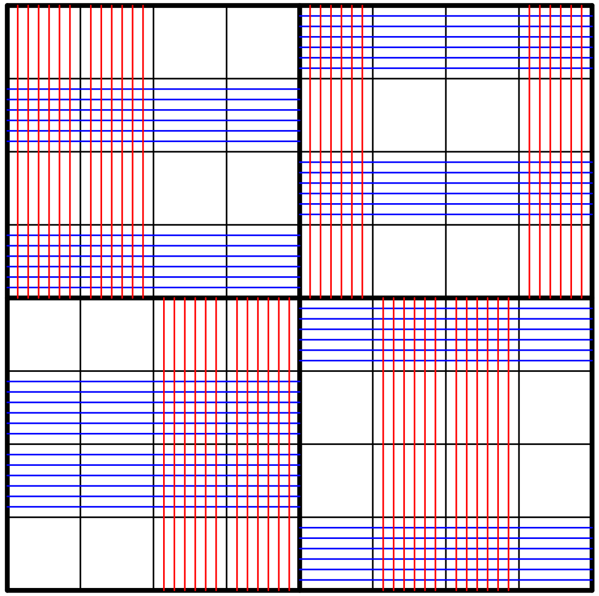
Definition. The random permutation Π is called bi-L-decomposable, if both Π and Π^{-1} are L-decomposable.

Question: Does the family of all bi-L-decomposable distributions form a hierarchical model?

Answer: No, but almost.

Equivalent definition by independence

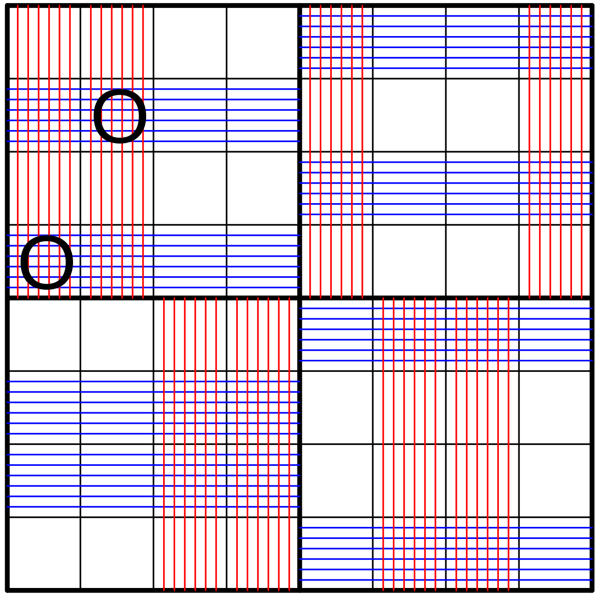
Fix a partition of both rows and columns into intervals.



Given the "grids" inside the rectangles of the product partition,

Equivalent definition by independence

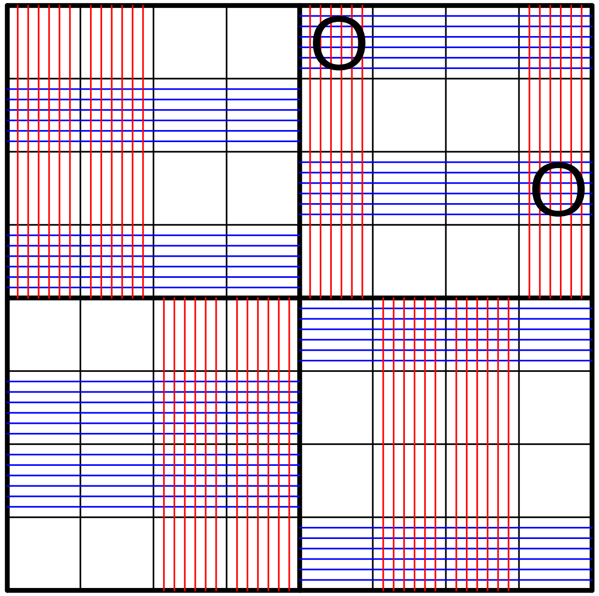
Fix a partition of both rows and columns into intervals.



Given the "grids" inside the rectangles of the product partition, the placement of the rooks in these rectangles are independent.

Equivalent definition by independence

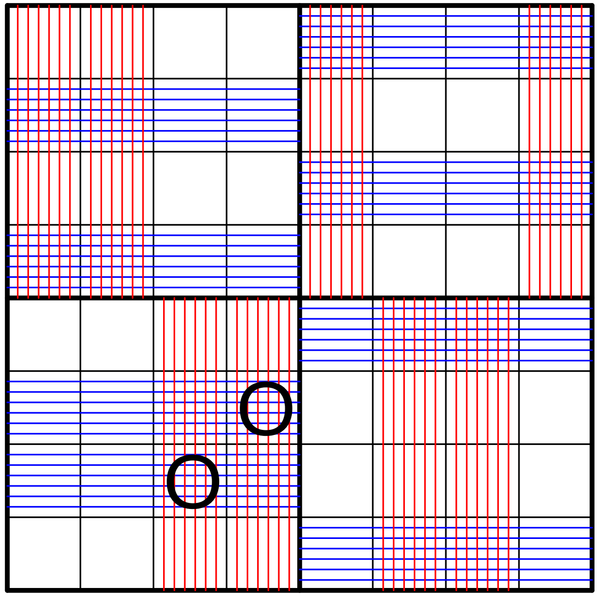
Fix a partition of both rows and columns into intervals.



Given the "grids" inside the rectangles of the product partition, the placement of the rooks in these rectangles are independent.

Equivalent definition by independence

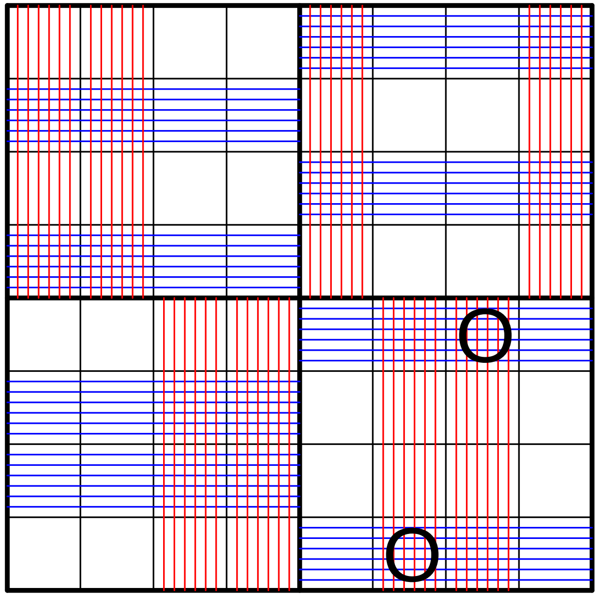
Fix a partition of both rows and columns into intervals.



Given the "grids" inside the rectangles of the product partition, the placement of the rooks in these rectangles are independent.

Equivalent definition by independence

Fix a partition of both rows and columns into intervals.



Given the "grids" inside the rectangles of the product partition, the placement of the rooks in these rectangles are independent.

The strictly positive (s.p.) part

Theorem. The family of s.p. bi-L-decomposable distributions is just the s.p. part of the hierarchical model $\mathcal{H}_{bL}$, whose generators are the product partitions $\mathcal{R}_i \times \mathcal{R}_j$, where $i = 2, \ldots, n-1$ and (as before)

$$\mathcal{R}_i = (\{1, \ldots, i-1\}, \{i\}, \{i+1, \ldots, n\}).$$

Non-trivial statement:

Theorem. The number of free parameters of $\mathcal{H}_{bL}$ is $\sum_{i=1}^{n-1} i^2$.

Strict inclusions

Question: what is the relationship between the family of all bi-L-decomposable distributions (denoted by $\mathcal{BL}$) and $\mathcal{H}_{bL}$?

Theorem. For all $n \geq 4$, $\mathcal{H}_{bL} \subsetneq \mathrm{cl}(\mathcal{H}_{bL}) \subsetneq \mathcal{BL}$.

Implications:

→ The first inclusion shows that $\mathcal{H}_{bL}$, unlike $\mathcal{H}_L$, is not closed.

→ The second inclusion shows that $\mathcal{BL}$ is not a hierarchical model.

→ For any dataset, a unique ML estimate is guaranteed in cl$(\mathcal{H}_{bL})$, which can be calculated via iterative scaling.

→ ML estimation in $\mathcal{BL}$ is not clear.

The proof relies on the calculation of the Markov basis of $\mathcal{H}_{bL}$ for $n = 4$.

Markov basis of $\mathcal{H}_{bL}$

$H_8 \subset S_n$ is the 8-element subgroup generated by the leading transposition $(213\ldots n)$, and the reversing permutation $(n(n-1)\ldots 21)$. The symmetry group of the Markov basis is $T = H_8 \times \{-1,+1\} \times H_8$, which acts on S_n by $(\rho_1, \varepsilon, \rho_2)(\pi) = \rho_1 \pi^{\varepsilon} \rho_2$.

We calculated a minimal Markov basis of $\mathcal{H}_{bL}$ by 4ti2 for $n = 4, 5$. The number of orbits under T of various degrees are

n	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$
4	2	–	1	–	–	–	–
5	16	1	210	13	578	50	40

Problem: T does not grow with n. We know that H_8 cannot be replaced by a larger group.

Example: S-decomposable distributions

A distribution p on S_n is S-decomposable, if it can be written as

$$p(\pi) = \prod_{k=1}^{n} \theta_k(\{\pi(1), \pi(2), \ldots, \pi(k)\}).$$

The S-decomposable distributions form a hierarchical model $\mathcal{H}_S$ with $n-1$ generators $\mathcal{R}_i \times \mathcal{C}$, $i = 1, \ldots, n-1$, where

$$\mathcal{R}_i = (\{1, \ldots, i\}, \{i+1, \ldots, n\}) \text{ and } \mathcal{C} = (\{1\}, \ldots, \{n\}).$$

Remark: $\mathcal{H}_S$ is a submodel of $\mathcal{H}_L$.

Interpretation: when you choose the next alternative, $\pi(k)$, you focus only on the emerging set $\{\pi(1), \pi(2), \ldots, \pi(k)\}$.

The Markov basis of $\mathcal{H}_S$

Define a "rewiring move" as follows, e.g. for $n = 5$.

→ Choose two neighboring numbers e.g. 2 and 3.

→ Choose an alternating cycle of 2-sets and 3-sets, where for adjacent sets, the 3-set contains the 2-set, e.g.

$$\{1,2\} \Rightarrow \{1,2,3\} \rightarrow \{1,3\} \Rightarrow \{1,3,4\} \rightarrow \{1,4\} \Rightarrow \{1,2,4\} \rightarrow \{1,2\}.$$

→ For each 2-set, choose a permutation of its elements, and for each 3-set, choose a permutation of its complement, e.g.

$$(12), (31), (41); \quad (54), (25), (53).$$

→ Connect these in the two possible ways (forward or backward):

$$(12354), (31425), (41253) \text{ or } (12453), (31254), (41325).$$

Then the move assigns $+1$ to the first set of permutations (12354), (31425), and (41253), and it assigns -1 to the second set of permutations (12453), (31254), and (41325).

Theorem. A Markov basis of $\mathcal{H}_S$ is formed by the basis moves of $\mathcal{H}_L$, plus the above "rewiring" moves.

Question: What is the degree of the minimal Markov basis?
Remark: For $n = 4$, the minimal Markov basis (calculated by 4ti2) has 6 moves of degree two, 64 moves of degree three, and 93 moves of degree four.

Example: Bi-$\mathcal{S}$-decomposable distributions

We can define bi-$\mathcal{S}$-decomposable distribution in the obvious way. Definition. The random permutation Π is called bi-$\mathcal{S}$-decomposable, if both Π and Π^{-1} are $\mathcal{S}$-decomposable.

Theorem. The family of s.p. bi-$\mathcal{S}$-decomposable distributions is just the s.p. part of the hierarchical model $\mathcal{H}_{bS}$, whose generators are the product partitions $\mathcal{R}_i \times \mathcal{R}_j$, where $i = 1, \ldots, n-1$ and

$$\mathcal{R}_i = (\{1, \ldots, i\}, \{i+1, \ldots, n\}).$$

The minimal Markov basis of $\mathcal{H}_{bS}$ contains 10 moves of degree two, 104 moves of degree three, and 33 moves of degree four.

Remark: $\mathcal{H}_{bS}$ is invariant under the same group T as $\mathcal{H}_{bL}$.

Example: Quasi-independent distributions

A distribution p on S_n is quasi-independent, if it can be written as $p(\pi) = \prod_{k=1}^{n} \theta_k(\pi(k))$.
The quasi-independent distributions form a hierarchical model $\mathcal{H}_Q$ with n^2 generators $\mathcal{R}_i \times \mathcal{R}_j$, $i, j = 1, \ldots, n$, where $\mathcal{R}_i = (\{i\}, [n] \setminus \{i\})$.

The Markov basis of $\mathcal{H}_Q$ was computed by Diaconis and Eriksson (2006) for $n \leq 6$. They show that the degree of the basis is $\leq n - 1$, and conjecture it to be 3.

Theorem. For $n \geq 4$, a strictly positive distribution p on S_n is quasi-independent, if and only if p is bi-L-decomposable, and remains so after any right or left multiplication.

Summary

1) We have defined hierarchical models for random permutations.
2) The models are based on statistics $\pi(\mathcal{R} \times \mathcal{C})$, which are two-dimensional non-negative integer tables with fixed marginals.
3) We looked at five examples of such models.

Properties to look for:

Is the hierarchical model closed?

Is it (or at least the s.p. part) characterized by conditional independence?

Does the model have an explicit Markov basis?

What is the degree of the Markov basis, and how many elements does it have?

What is the symmetry group of the model?

References

[1] 4ti2 team. *4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces.* Available at www.4ti2.de.

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

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

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

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

[6] 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.