Statistical analysis of random permutations

Summary of the Ph.D. dissertation

By: Villő Csiszár

Mathematics Doctorate School
Applied Mathematics Doctorate Program

School leader:	Miklós Laczkovich
Program leader:	György Michaletzky
Thesis Supervisor:	Gábor Tusnády, member of the Hungarian Academy of Sciences

Eötvös Loránd University, Budapest
Faculty of Science
2008

1 Preliminaries

The dissertation deals with some aspects of the statistical analysis of random permutations. Random permutations most frequently appear as orderings of the elements of a finite set. In sociological studies, individuals may be asked to rank a number of choices according to importance or preference. In other cases, voters, judges, or exam boards rank candidates, tenders, or applicants, and make decisions based on these orderings. Moreover, a permutation can describe the pairing of the elements of two sets: one can pair jobs with employees or students with tutors. Finally, any real dataset may be analysed based on only the ranks.

My aim was to give an overview of the various parametric models applicable to permutation data, to study estimation of the parameters and assess of fit, and to develop new models. Part of my motivation for this research was the experience that we – my supervisor and I – could not find a well-fitting simple model for the 1980 election data of the American Psychological Association. This dataset is one of the most well-studied in the literature, see for example [1, 3, 10, 11, 13]. We could formulate a simple model, which did not provide a satisfactory fit for these data, but performed significantly better than the other models. This new model led me to consider conditional independence models and factorizing models for random permutations. In the dissertation, borrowing terminology from the theory of contingency table analysis, I call these models hierarchical models.

The book by Marden [10] gives a thorough overview of the existing models for random rankings. The author was partly inspired by the conference entitled "Probability Models and Statistical Analyses for Ranking Data," held in Amherst in 1990, the proceedings of which was edited by Fligner and Verducci [5]. The paper by Critchlow, Fligner and Verducci [2] also deserves mentioning, which, while shorter, contains important results about the properties of the models (reversibility, label-invariance, L-decomposability, unimodality, complete consensus).

2 Methods

One of the tools I used is algebraic statistics. As the name suggests, this field is concerned with the application of algebraic tools in statistics. These tools are especially suited for studying closures of exponential families, which is of both theoretical and practical importance. The following specific results, used in the dissertation, can be found in the papers by Diaconis and Sturmfels [4], Geiger, Meek and Sturmfels [6], and Rapallo [12].

Let $\mathcal{X} = \{x_1, \ldots, x_s\}$ denote a finite set, and let $M = (m_{ij})$ be a $t \times s$ matrix with nonnegative integer entries. The probability distribution $p = (p(x_1), \ldots, p(x_s))$ is said to belong to the so-called *toric model* $\mathbf{F}(M)$, if there exist nonnegative parameters $\lambda_1, \ldots, \lambda_t$, with which

$$p(x_i) = c(\lambda) \prod_{j=1}^{t} \lambda_j^{m_{ji}}, \quad 1 \leq i \leq s.$$

The set T is called *M-feasible*, if for each $i \notin T$, we have $\mathrm{Supp}(m_i) \not\subseteq \cup_{j \in T} \mathrm{Supp}(m_j)$, where m_i denotes the ith column vector of M, and $\mathrm{Supp}(\cdot)$ denotes the support of the vector in the argument. The nonnegative *toric variety* associated with M is the set

$$X_M = \{x \in \mathbb{R}^s_{\geq 0} : x^u - x^v = 0 \quad \forall u, v \in \mathbb{N}^s \text{ such that } Mu = Mv\},$$

where $x^u = \prod_i x_i^{u_i}$. The toric ideal generated by the polynomials $x^u - x^v$ above is denoted by I_M.

Theorem 2.1 (Geiger et al. [6]) *$cl(\mathbf{F}(M)) = X_M$, where $cl(\cdot)$ stands for closure. Moreover, for $p \in X_M$, we have $p \in \mathbf{F}(M)$ if and only if the support of p is M-feasible.*

Theorem 2.2 (Rapallo [12]) *For every M, there exists a maximal representation $M_{\max}$, for which $cl(\mathbf{F}(M)) = \mathbf{F}(M_{\max})$.*

The functions $f_1, \ldots, f_L : \mathcal{X} \to \mathbb{Z}$ are said to be a *Markov basis* for the model $\mathbf{F}(M)$, if for every u, the steps f_i generate a strongly connected graph on the frequency vectors $g : \mathcal{X} \to \mathbb{N}$ satisfying $Mg = u$ (where the step f_i takes

g to $g + f_i$). Markov bases can be used to run Monte Carlo procedures for assessing goodness of fit in the case of small datasets, where the asymptotics of the χ^2 test is not applicable.

Theorem 2.3 (Diaconis és Sturmfels [4]) *The functions $f_1, \ldots, f_L$ constitute a Markov basis if and only if the polynomials $x^{f_i^+} - x^{f_i^-}$ are a generating set of the ideal I_M, where f_i^+ (f_i^-) is the positive (negative) part of f_i.*

I also used the theory of hierarchical and log-linear models, or more generally the theory of discrete exponential families. The relevant results can be found in the book by Lauritzen [8]. Keeping the previous setting, let $A \in \mathcal{A}$ be partitions of the set $\mathcal{X}$, and let the rows of the matrix $M_{\mathcal{A}}$ be the indicator vectors of the classes of these partitions (so for a partition with k classes, we have k rows in $M_{\mathcal{A}}$). Then in the model $\mathrm{cl}(\mathbf{F}(M_{\mathcal{A}}))$, the maximum likelihood estimate exists uniquely, and it can be found by e.g. *iterative proportional scaling* (IPS). For any $x \in \mathcal{X}$, let $x(A)$ be the class of partition A containing x, and for any probability distribution p on $\mathcal{X}$, let $p(x(A)) = \sum_{y \in \mathcal{X}: y(A) = x(A)} p(y)$ be the p-probability of this class. Suppose we have a sample with empirical distribution r, and we want to find the element of $\mathrm{cl}(\mathbf{F}(M_{\mathcal{A}}))$, which maximizes the likelihood of the sample. Let $p^{(0)}$ be an arbitrary strictly positive element of the model $\mathbf{F}(M_{\mathcal{A}})$ (e.g. the uniform distribution). Then the $(t+1)$st iteration step of the IPS algorithm updates $p^{(t)}$ as

$$p^{(t+1)}(x) = \frac{r(x(A))}{p^{(t)}(x(A))} p^{(t)}(x), \quad x \in \mathcal{X},$$

where A runs cyclically over the set $\mathcal{A}$.

3 Results

3.1 The inversions model of McCullagh

The following model was introduced by Peter McCullagh [11]. Let $C \subseteq [n]$ be a k-element subset. A permutation σ_C of the elements of C is called a $(k-1)$*st order inversion*, if none of its coordinates is in its own place with respect to the monotone increasing order (where $[n] = \{1, \ldots, n\}$). We say that the permutation $\pi \in S_n$ contains the inversion σ_C (in notation $\sigma_C \subseteq \pi$), if the elements of C appear in π in the order σ_C. Then the model defined by the inversions $\sigma^1_{C_1}, \ldots, \sigma^s_{C_s}$ consists of distributions satisfying

$$\log p_\theta(\pi) = \sum_{i:\sigma^i_{C_i} \subseteq \pi} \theta_i, \quad \theta = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s. \tag{1}$$

In the dissertation, I prove McCullagh's following conjecture.

Theorem 3.1 *In the model* (1)*, if $\theta \neq \tau$, then $p_\theta \neq p_\tau$.*

The theorem is equivalent to the following combinatorical reformulation. Define a graph G_H on the vertex set S_n as follows. From the permutation π, there is a directed edge to all permutations obtained from π by moving one element to its proper position (other elements are shifted, if necessary). As an example, for $n = 5$, from (24351) there are directed edges to the permutations (12435), (42351), (23541), (24315).

Theorem 3.2 *There are no directed cycles in the graph G_H.*

3.2 EM algorithms for Plackett-Luce-type models

Take n players, the overall ability of the ith one is expressed by the parameter λ_i. According to the Plackett-Luce model, if the players $I \subseteq [n]$ take part in a competition, then the probability of the ordering π is

$$p(\pi) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}, \tag{2}$$

where $\pi(1)$ is the overall winner, $\pi(|I|)$ is the overall loser. Luce [9] derived this model from the ranking postulate and the choice axiom. It is easy to check that if Z_i ($i = 1, \ldots, n$) are independent random variables, exponentially distributed with parameters λ_i, then the righthandside of (2) is just the probability $P(Z_{\pi(1)} < \ldots < Z_{\pi(|I|)})$. This observation leads to the following EM algorithm, which iteratively finds the maximum likelihood estimate of the parameters λ_i. Suppose we have m observations, the rth of which consists of the ordering π_r of the players in I_r. For all $i \in I_r$, let $\alpha_r(i)$ be the rank of i in the ordering π_r. Finally, let m_i denote the number of observed orderings containing player i. With these notations, one EM-step is given by

$$\lambda_i^{(t+1)} = m_i \left[\sum_{r:i\in I_r} \sum_{k=1}^{\alpha_r(i)} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} \right]^{-1} \qquad 1 \le i \le n.$$

Hunter [7] derived MM algorithms for this model and its generalizations. He also gave conditions under which the algorithms converge to the unique maximum likelihood estimate. I showed that EM algorithms are also a natural choice for this estimation problem, although, according to my simulation studies, their convergence is slower than the convergence of the MM algorithms.

3.3 L-decomposability

For a permutation $\pi = (\pi(1), \ldots, \pi(n))$, let $\pi\{i..j\} = \{\pi(i), \ldots, \pi(j)\}$ and $\pi(i..j) = (\pi(i), \ldots, \pi(j))$. The random permutation Π (and its distribution) is *L-decomposable*, if the sets $\Pi\{1..k\}$, $k = 1, \ldots, n$ form a Markov chain. "L" stands for Luce, since these are exactly the distributions satisfying Luce's ranking postulate. Denote by (π, ρ) the concatenation of two partial permutations, and for a subset $C \subseteq [n]$, let S_C consist of all permutations of the elements of C.

Theorem 3.3 *The L-decomposable distributions form a closed toric model. The toric ideal corresponding to the model is generated by all polynomials of form $x_{(\pi_1,\rho_1)}x_{(\pi_2,\rho_2)} - x_{(\pi_1,\rho_2)}x_{(\pi_2,\rho_1)}$, where $\pi_1, \pi_2 \in S_C$ and $\rho_1, \rho_2 \in S_{[n]\setminus C}$ for some C.*

Theorem 3.4 *For $n = 4$ and $n = 5$, the L-decomposable model has a unique minimal Markov basis, which is equal to the one described in the previous theorem. For $n \geq 6$, the basis of the previous theorem is not minimal, and the minimal basis is not unique.*

The following theorem is about the properties of the maximum likelihood (ML) estimate.

Theorem 3.5 *In the L-decomposable model, the ML estimate always exists uniquely, it has an explicite form, and its exact distribution can be calculated. Moreover, the following hyper Markov property holds: for all k, the random distributions $\{\hat{P}(\Pi(1..k) = u)\}_u$ and $\{\hat{P}(\Pi(k+1..n) = v)\}_v$ are conditionally independent, given the random distribution $\{\hat{P}(\Pi\{1..k\} = C)\}_C$, where $\hat{P}$ denotes the ML estimate.*

3.4 Bi-L-decomposability

The following property was not studied in the literature before. Let us call the random permutation Π (and its distribution) *bi-L-decomposable*, if both Π and Π^{-1} are L-decomposable. I introduced hierarchical models for random permutations to study bi-L-decomposability. Let $\mathcal{D}$ (resp. $\mathcal{R}$) be partitions of $[n]$ with d (resp. r) classes. The *coarsening* of π on the product partition $\mathcal{P} = \mathcal{D} \times \mathcal{R}$ is the $d \times r$ matrix

$$|\pi(\mathcal{P})| = (t_{ij}), \quad t_{ij} = |\{1 \le s \le n : s \in D_i,\, \pi(s) \in R_j\}|.$$

Definition 3.6 *Let $\mathcal{P}_1, \ldots, \mathcal{P}_s$ be product partitions of $[n] \times [n]$. The strictly positive distribution p on S_n belongs to the hierarchical model with generators $\mathcal{P}_1, \ldots, \mathcal{P}_s$, in notation $p \in \mathcal{L}(\mathcal{P}_1, \ldots, \mathcal{P}_s)$, if there exist functions θ_i such that*

$$\log p(\pi) = \sum_{i=1}^{s} \theta_i(|\pi(\mathcal{P}_i)|) \quad \forall \pi \in S_n.$$

We write $\mathcal{D}' \succeq \mathcal{D}$ if the partition $\mathcal{D}'$ is finer than $\mathcal{D}$.

Theorem 3.7 *Let $\mathcal{L}(\mathcal{D}_i \times \mathcal{R} : i = 1, \ldots, s)$ and $\mathcal{L}(\mathcal{D} \times \mathcal{R}_j : j = 1, \ldots, t)$ be two hierarchical models, where $\mathcal{D} \succeq \mathcal{D}_i$ and $\mathcal{R} \succeq \mathcal{R}_j$ for all $1 \le i \le s$, $1 \le j \le t$. Then the intersection of the two models is the hierarchical model $\mathcal{L}(\mathcal{D}_i \times \mathcal{R}_j : i = 1, \ldots, s,\, j = 1, \ldots, t)$.*

This theorem is applicable to the L-decomposable hierarchical model and its inverse, whose intersection is the bi-L-decomposable hierarchical model. Some further calculations yield the following.

Theorem 3.8 *The family of strictly positive bi-L-decomposable distributions has $\sum_{i=1}^{n-1} i^2$ free parameters.*

In the dissertation, I give two parametrizations, which correspond to two bases of the subspace in $\mathbb{R}^{n!}$ spanned by the logarithms of bi-L-decomposable distributions. One basis is orthogonal, the other is $0-1$.

Each hierarchical model has a matrix $M_{\mathcal{A}}$ of the type described in the Methods section. Denote by M_L the matrix of the L-decomposable model, and by M_B the matrix of the bi-L-decomposable model. I could determine the Markov basis of $\mathbf{F}(M_B)$ only for $n=4$.

Theorem 3.9 *The minimal Markov basis of $\mathbf{F}(M_B)$ for $n=4$ consists of 10 degree-2 polynomials (from the Markov bases of $\mathbf{F}(M_L)$ and its inverse), and 8 degree-4 polynomials.*

This observation led to the following result, valid for all n.

Theorem 3.10 *The model $\mathbf{F}(M_B)$ is not closed, and even its closure is a strict subset of all bi-L-decomposable distributions.*

3.5 S-decomposability

In the analysis of bi-L-decomposable distributions, a stronger property, S-decomposability played an important role. The random permutation Π and its distribution p is *S-decomposable*, if there exist parameters $\Lambda(C) \geq 0$ ($C \subseteq [n]$) such that $p(\pi) = \prod_{k=1}^{n} \Lambda(\pi\{1..k\})$. Π is *bi-S-decomposable*, if both Π and Π^{-1} are S-decomposable.

Theorem 3.11 *A strictly positive distribution p is S-decomposable, if and only if it is L-decomposable, and there exist parameters $\Lambda'(C) > 0$ such that*

$$P(\Pi(k+1) = x | \, \Pi\{1..k\} = C) = \frac{\Lambda'(C \cup x)}{\sum_{y \notin C} \Lambda'(C \cup y)}.$$

Strictly positive S-decomposable distributions form a hierarchical model with model matrix M_S. The corresponding toric model $\mathbf{F}(M_S)$ is not closed, however, its Markov basis can be characterized.

Theorem 3.12 *Let the sets $C_1, D_1, C_2, D_2, \ldots, C_j, D_j \subseteq [n]$ satisfy $|C_i| = k$, $|D_i| = k + 1$, and $C_i, C_{i+1} \subset D_i$ (with $C_{j+1} := C_1$). Let $\pi_i \in S_{C_i}$ and $\rho_i \in S_{[n]\setminus D_i}$. For all such choices, create the polynomial*

$$\prod_{i=1}^{j} x_{(\pi_i, D_i \setminus C_i, \rho_i)} - \prod_{i=1}^{j} x_{(\pi_i, D_{i-1} \setminus C_i, \rho_{i-1})},$$

where $D_0 = D_j$ and $\rho_0 = \rho_j$. These polynomials, together with the ones in the Markov basis of $\mathbf{F}(M_L)$, form a Markov basis of the model $\mathbf{F}(M_S)$.

In summary, we found that the S- and bi-S-decomposable families are more complex algebraically than the L- and bi-L-decomposable families.

3.6 Label-invariance

Suppose the elements of a set are labelled with the integers $1, \ldots, n$, and an ordering of these elements is given by π. If we relabel the elements, i.e. change label i to label $\sigma(i)$, then the same ordering is given by the permutation $\sigma\pi$ (the operation here is group multiplication). Similarly, if π^{-1} is a ranking expressed with the original labelling, then the same ranking with the new labelling becomes $\pi^{-1}\sigma^{-1}$. This motivates the question, whether a model for random permutations is invariant under multiplications from the left and right.

Theorem 3.13 *Let $n \geq 4$. The family of L-decomposable distributions is invariant under left multiplications. It is invariant under right multiplication by σ, if and only if σ belongs to the eight-element subgroup of S_n generated by the permutations $(n\, n{-}1 \ldots 21)$ and $(2134 \ldots n)$.*

It is natural to ask which subfamily of the L-decomposable distributions is invariant under all right multiplications. More precisely, we are looking for those distributions on S_n, which remain L-decomposable after any right multiplication.

Theorem 3.14 *Let $n \geq 4$. A strictly positive distribution p on S_n remains L-decomposable after all right multiplications, if and only if it is* *quasi-independent, i.e. there exist parameters* $c_i(x)$, $1 \leq i, x \leq n$ *such that* $p(\pi) = \prod_{i=1}^{n} c_i(\pi(i))$.

4 Conclusions

The research reported in the dissertation showed that while there are many simple, elegant, practical and realistic models to describe random permutations, there is still room for the development of new models. Such could be hierarchical models, some of which can be interpreted as conditional independence models. I would like to characterize "simple" hierarchical models, which would be an analogue of decomposable graphical models in the classical theory. Greater insight could be gained by calculating the Markov basis of other hierarchical models, the main difficulty is that current algorithms quickly become infeasible as n grows. It would be useful to give general upper bounds for the degree of these Markov bases. A characterization of the intersection of hierarchical models in the general case is also open.

The dissertation is based on the following papers

- Conditional independence relations and log-linear models for random matchings. *Acta Math. Hungar.*, Online First (2008).
- (with Rejtő, L. and Tusnády, G.) Statistical Inference on Random Structures. In: *Horizons of Combinatorics*, Bolyai Society Mathematical Studies **17**, Springer (2008), 37-66.
- Markov bases of conditional independence models for permutations. To appear in *Kybernetika.*
- On L-decomposability of random permutations. Submitted to *J. Math. Psych.*, under revision.
- An acyclic operation on the symmetric group. Submitted.

References

[1] Chung, L. and Marden, J. I.: Extensions of Mallows' ϕ model. In: [5], 108-139.

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

[3] Diaconis, P.: A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17** (1989), 949-979.

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

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

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

[7] Hunter, D. R.: MM algorithms for generalized Bradley-Terry models. *Ann. Statist.* **32** (2004), 384-406.

[8] Lauritzen, S.: *Graphical Models.* Clarendon Press, Oxford (1996).

[9] Luce, R. D.: *Individual choice behavior.* Wiley, New York (1959).

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

[11] McCullagh, P.: Permutations and regression models. In: [5], 196-215.

[12] Rapallo, F.: Toric statistical models: parametric and binomial representations. *Ann. Inst. Statist. Math.*, Online First (2006).

[13] Stern, H.: Probability models on rankings and the electoral process. In: [5], 173-195.