The Annals of Statistics
2016, Vol. 44, No. 2, 853–875
DOI: 10.1214/15-AOS1389

ESTIMATION IN EXPONENTIAL FAMILIES ON PERMUTATIONS

BY SUMIT MUKHERJEE

Columbia University

Asymptotics of the normalizing constant are computed for a class of one parameter exponential families on permutations which include Mallows models with Spearmans's Footrule and Spearman's Rank Correlation Statistic. The MLE and a computable approximation of the MLE are shown to be consistent. The pseudo-likelihood estimator of Besag is shown to be $\sqrt{n}$-consistent. An iterative algorithm (IPFP) is proved to converge to the limiting normalizing constant. The Mallows model with Kendall's tau is also analyzed to demonstrate the flexibility of the tools of this paper.

1. Introduction. Analysis of permutation data has a long history in statistics. An early paper is the work of Mallows [34] in 1957, where the author proposed an exponential family of the form

$$e^{-\theta d(\pi,\sigma)-Z_n(\theta,\sigma)},$$

henceforth referred to as Mallows models, to study nonuniform distributions on permutations. In this model, σ is a fixed permutation which is a location parameter, θ is a real valued parameter and $d(\cdot,\cdot) : S_n \times S_n \mapsto \mathbb{R}$ is a "distance" function on the space of permutations. Here, $Z_n(\theta,\sigma)$ denotes the (unknown) log normalizing constant of this family. Mallows mainly considered two distances, namely the Kendall's tau, and the Spearman's rank correlation. Using this modeling approach, Feigin and Cohen [16] analyzed the nature of agreement between several judges in a contest. Critchlow [9] gave some examples where Mallows model gives a good fit to ranking data. Fligner and Verducci [18, 19] generalized the Mallows model with Kendall's tau and Cayley's distance to an n parameter exponential family over permutations, where each parameter inductively determines a permutation in S_k starting from a partition in S_{k-1}. The extra parameters allow more flexibility, and the structure of the model allows one to retain tractability. See also [10] which deal with various aspects of permutation models, Chapters 5 and 6 of [13] which sketch out numerous possible applications and the book length treatment in [35], which covers both theoretical and applied aspects of permutation modeling.

Permutation modeling has also received some recent attention in machine learning literature. The generalized version of the Mallows model with Kendall's tau was considered in [8, 30]. Estimation of the location and scale parameters in this

Received May 2015; revised September 2015.
MSC2010 subject classifications. Primary 62F12, 60F10; secondary 05A05.
Key words and phrases. Permutation, normalizing constant, Mallows model, pseudo-likelihood.

model was carried out in [40]. An infinite version of the generalized Mallows model was proposed in [38, 39], where the authors show existence of sufficient statistics, and compute conjugate priors for this model. If the data has multiple modes, the authors propose to use clustering techniques to separate permutations coming from different modes before applying their estimation scheme. In another direction, [28] propose the use of Fourier analysis on representation theory of finite groups to study probability models on permutations. Representing a probability distribution via its Fourier coefficients, they are able to formulate the update rule in a Bayesian model tracking problem in terms of Fourier coefficients. This representation, along with fast computation of Fourier coefficients allow study of permutation models of moderate size ($n = 30$). For a more detailed exposition on this technique, refer to [23]. Modeling of partially ranked data under the Mallows model with Kendall's tau was studied in [31]. Efficient sampling schemes for learning Mallows models with Cayley distance, Ulam distance and Hamming distance has been studied recently in [24–26] utilizing the structure of the model under consideration. See also the recent work of [3] where the authors study estimation of location and scale mixtures of Mallows model with Kendall's tau in an algorithmic framework.

This paper analyzes a class of exponential families on the space of permutations using the recently developed concept of permutation limits. The notion of permutation limits has been first introduced in [22], and is motivated by dense graph limit theory (see [32] and the references therein). The main idea is that a permutation can be thought of as a probability measure on the unit square with uniform marginals. Multivariate distributions with uniform marginals have been studied widely in probability and statistics (see [20, 27, 33, 37, 45, 46, 48, 51] and references therein) and finance (see [1, 7, 36, 41, 43] and references therein) under the name copula. One of the reasons for their popularity is that copulas are able to capture any dependence structure, as shown in Sklar's theorem [48].

1.1. *Main contributions.* This paper gives a framework for analyzing probability distributions on large permutations. It computes asymptotics of normalizing constants in a class of exponential families on permutations, and explores identifiability of such models. It derives the limit in probability of statistics under such models, and shows the consistency of two estimates including the MLE. It also shows the existence of consistent tests for such models. It gives an Iterative Proportional Fitting Procedure (IPFP) to numerically compute the normalizing constant. It also shows $\sqrt{n}$ consistency of the pseudo-likelihood estimator of Besag. It demonstrates the flexibility of this approach by analyzing the Mallows model with Kendall's tau. For the Mallows model with Kendall's tau, it again shows consistency of two estimates including the MLE. Finally, using results from [6] it computes joint limiting distribution of whole permutation for a class of permutation models which include the Mallows model with Spearman's rank correlation and Kendall's tau.

Even though Mallows models for a class of distance functions have been considered in the literature, not much is available in terms of weak limits, limit distributions, rate of errors in estimation and testing of hypothesis, outside the Mallows model with Kendall's tau. One of the advantages of the proposed framework of this paper is that it allows analysis of permutation models both from a rigorous and a visual point of view. From a rigorous perspective, this approach allows one to study limiting properties of permutation models outside Mallows models with Kendall's tau and its generalizations, which seemingly have no independence structure built in to exploit. In particular, this approach covers the Mallows model with Spearman's correlation, for which not much theory is available in the literature. From a visual and diagnostic perspective, it gives a way to compare permutations, and thus test goodness of fit for permutation models. Another advantage of this approach is that it provides a limiting distribution for the whole permutation viewed as a process for a wide class of models on permutations, which to the best of my knowledge was not known before even for the usual Mallows model with Kendall's tau. This allows for the possibility of studying partially observed rankings as well, by studying finite-dimensional marginals of this permutation process.

The main tool for proving the results is a large deviation principle for a uniformly random permutation. This was first proved in [50], but the supplementary file [42] contains a new proof of this large deviation using permutation limits.

1.2. *The* 1970 *draft lottery.* To see how permutation data can arise naturally, consider the following example of historical importance. On December 1, 1969, during the Vietnam War the US Government used a random permutation of size 366 to decide the relative dates of when the people (among the citizens of the USA born between the years 1944–1950) will be inducted into the army in the year 1970, based on their birthdays. 366 cylindrical capsules were put in a large box, one for each day of the year. The people who were born on the first chosen date had to join the war first, those born on the second chosen date had to join next, and so on. There were widespread allegations that the chosen permutation was not uniformly random. In [17], Fienberg computed the Spearman's rank correlation between the birthdays and lottery numbers to be -0.226, which is significantly negative at 0.001 level of significance. This suggests that people born in the latter part of the year were more likely to be inducted earlier in the army.

If a permutation is not chosen uniformly at random, then the question arises whether a particular nonuniform model gives a better fit. It might be the case that there is a specific permutation σ toward which the sampling mechanism has a bias, and permutations close to σ have a higher probability of being selected. For example, in the draft lottery example σ is the permutation

$$(366, 365, 364, \ldots, 3, 2, 1).$$

The Mallows models defined above are able to capture such behavior. The hypothesis of uniformity in this setting is equivalent to the hypothesis that $\theta = 0$.

Possibly the most famous and widely used model on permutations is the Mallows model with Kendall's tau as the divergence function. One of the reasons for this is that for this model the normalizing constant is known explicitly (see, e.g., [15], (2.9)), and so analyzing this model becomes a lot simpler. However, when one moves away from the Mallows model with Kendall's tau and its generalizations, not much theory is available in the literature. One reason for this is that normalizing constant is not available in closed form, and there is no straightforward independence assumptions in the model which one can exploit to analyze such models. Even basic properties for such models such as identifiability and consistency of estimates are not well understood in general.

1.3. *Choice of the function d.* Even though the literature on Mallows models mostly mention $d(\cdot,\cdot)$ to be a metric, for the purposes of this paper this is not necessary. As such, the function d will henceforth be referred to as a divergence function. One restriction on the divergence $d(\cdot,\cdot)$ which seems reasonable is that $d(\cdot,\cdot)$ is right invariant, that is,

$$d(\pi,\sigma) = d(\pi\circ\tau,\sigma\circ\tau) \qquad \text{for all } \pi,\sigma,\tau\in S_n.$$

The justification for this last requirement is as follows: suppose the students in a class are labeled $\{1,2,\ldots,n\}$, and let $\pi(i)$ and $\sigma(i)$ denote the rank of student i based on math and physics scores, respectively (assume no tied scores). The divergence $d(\pi,\sigma)$ can be thought of as a measure of the strength of the relationship between math and physics rankings. If students are now labeled differently using a permutation τ, so that student i now becomes student $\tau(i)$, then the math and physics rankings become $\pi\circ\tau$ and $\sigma\circ\tau$, respectively. But this relabeling of students in principle should not change the relation between math and physics rankings, which requires the right invariance of $d(\cdot,\cdot)$.

Some of the common choices of right invariant divergence $d(\cdot,\cdot)$ in the literature are the following ([13], Chapters 5 and 6):

(a) Spearman's foot rule: $\sum_{i=1}^n |\pi(i)-\sigma(i)|$.
(b) Spearman's rank correlation: $\sum_{i=1}^n (\pi(i)-\sigma(i))^2$.
(c) Hamming distance: $\sum_{i=1}^n 1\{\pi(i)\neq\sigma(i)\}$.
(d) Kendall's tau: Minimum number of pairwise adjacent transpositions which converts π^{-1} into σ^{-1}.
(e) Cayley's distance: Minimum number of adjacent transpositions which converts π into σ = n-number of cycles in $\pi\sigma^{-1}$.
(f) Ulam's distance: Number of deletion–insertion operations to convert π into σ = n-length of the longest increasing subsequence in $\sigma\pi^{-1}$.

See [13], Chapters 5 and 6, for more details on these divergences. It should be noted here that barring Spearman's rank correlation, all the other divergences in

this list are metrics on the space of permutations. If $d(\cdot,\cdot)$ is right invariant, then the normalizing constant is free of σ, as

$$\sum_{\pi\in S_n} e^{-\theta d(\pi,\sigma)} = \sum_{\pi\in S_n} e^{-\theta d(\pi\circ\sigma^{-1},e)} = \sum_{\pi\in S_n} e^{-\theta d(\pi,e)},$$

where e is the identity permutation. Also, if π is a sample from the probability mass function $e^{-\theta d(\pi,\sigma)-Z_n(\theta)}$, then $\pi\circ\sigma^{-1}$ is a sample from the probability mass function $e^{-\theta d(\pi,e)-Z_n(\theta)}$. This paper focuses on the case where σ is known, and carries out inference on θ when one sample π is observed from this model. If the location parameter σ is unknown, estimating it from one permutation π seems impossible, unless the model puts very small mass on permutations which are away from σ, in which case π itself is a reasonable estimate for σ. In case σ is known, without loss of generality by a relabeling it can be assumed that σ is the identity permutation.

In an attempt to cover the first two divergences in the above list, consider an exponential family of the form

$$\mathbb{Q}_{n,f,\theta}(\pi) = e^{\theta\sum_{i=1}^n f(i/n,\pi(i)/n)-Z_n(f,\theta)}, \tag{1.1}$$

where f is a continuous function on the unit square. In particular, if $f(x,y) = -|x-y|$ then

$$\sum_{i=1}^n f\big(i/n,\pi(i)/n\big) = -\frac{1}{n}\sum_{i=1}^n |i-\pi(i)|,$$

which is a scaled version of the Foot rule [see (a) in list above]. For the choice $f(x,y) = -(x-y)^2$,

$$\sum_{i=1}^n f\big(i/n,\pi(i)/n\big) = -\frac{1}{n^2}\sum_{i=1}^n \big(i-\pi(i)\big)^2$$

is a scaled version of Spearman's rank correlation statistic [see (b) in the list above]. A simple calculation shows that the right-hand side above is same as

$$\frac{(n+1)(2n+1)}{3n} + \frac{2}{n^2}\sum_{i=1}^n i\pi(i),$$

and so the same model would have been obtained by setting $f(x,y) = xy$. Note that the Hamming distance (third in the list of divergences) is also of this form for the choice $f(x,y) = 1_{x\neq y}$ which is a discontinuous function.

REMARK 1.1. It should be noted here that the model $\mathbb{Q}_{n,f,\theta}$ covers a wide class of models, some of which are not unimodal, for example, if one sets $f(x,y) = x(1-x)y$ then for $n = 7$

$$\sum_{i=1}^7 f(i/7,j/7) = \frac{1}{7^3}\sum_{i=1}^7 i(7-i)\pi(i),$$

which is maximized when

$$\pi(7) = 1, \qquad \{\pi(1), \pi(6)\} = \{2, 3\},$$
$$\{\pi(2), \pi(5)\} = \{4, 5\}, \qquad \{\pi(3), \pi(4)\} = \{6, 7\}.$$

Thus, for $\theta > 0$ this model has $2^3 = 8$ modes. In general, for $\theta > 0$ this model has $2^{(n-1)/2}$ modes for n odd, and $2^{(n-2)/2}$ modes for n even.

If one assumes that for every fixed y the function $y \mapsto f(x, y)$ has a unique global maximum at $y = x$, then the model $\mathbb{Q}_{n,f,\theta}$ is unimodal. Indeed, in this case the mode is the identity permutation $(1, 2, \ldots, n)$ for $\theta > 0$ and the reverse identity permutation $(n, n-1, \ldots, 1)$ for $\theta < 0$. Note that both the functions $f(x, y) = -(x-y)^2$ and $f(x, y) = -|x-y|$ satisfy this condition. In general, if $y \mapsto f(x, y)$ is maximized uniquely at $y = \phi(x)$ for some function $\phi : [0, 1] \mapsto [0, 1]$, then at a heuristic level the mode of the distribution is the permutation π_ϕ given by $\pi_\phi(i) = \phi(ni)$. If the specific problem at hand demands unimodality, one can restrict to functions $f(x, y)$ which are maximized on the diagonal, though nothing changes as this information is not exploited anywhere.

One important comment about the model $\mathbb{Q}_{n,f,\theta}$ is that different choices of the function f may give the same model. Indeed as already remarked above, the function $f(x, y) = -(x-y)^2/2$ and $f(x, y) = xy$ gives rise to the same model. In general, whenever $f(x, y) - g(x, y)$ can be written as $\phi(x) + \psi(y)$ for any two functions $\phi, \psi : [0, 1] \mapsto \mathbb{R}$ the two models are the same. In particular, the function $f(x, y) = x + y$ and $g(x, y) \equiv 0$ gives rise to the same model, which is the uniform distribution on S_n. The following definition restricts the class of functions f to ensure identifiability.

DEFINITION 1.2. Let $\mathcal{C}$ be the set of all continuous functions f on $[0, 1]^2$ which satisfy

$$\int_0^1 f(x, z)\, dz = 0 \qquad \forall x \in [0, 1];$$
$$\int_0^1 f(z, y)\, dz = 0 \qquad \forall y \in [0, 1], \tag{1.2}$$

and f is not identically 0.

PROPOSITION 1.3. *If $f_1, f_2 \in \mathcal{C}$ and $\mathbb{Q}_{n,f_1,\theta}(\pi) = \mathbb{Q}_{n,f_2,\theta}(\pi)$ for all $\pi \in S_n$ for all large n for some $\theta \neq 0$, then $f_1 \equiv f_2$.*

PROOF. Setting $g(x, y) = \theta[f_1(x, y) - f_2(x, y)]$ the given assumption implies $\sum_{i=1}^n g(i/n, \pi(i)/n) = C_n$ for all $\pi \in S_n$, where $C_n := Z_n(f, \theta_1) - Z_n(f, \theta_2)$ is free of π. Thus, for any $1 \le i, j \le n-1$ setting $\sigma_1, \sigma_2 \in S_n$ by

$$\sigma_1(i) = j, \qquad \sigma_1(n) = n, \qquad \sigma_1(i) = i \qquad \text{otherwise},$$
$$\sigma_2(i) = n, \qquad \sigma_2(n) = j, \qquad \sigma_2(i) = i \qquad \text{otherwise},$$

by choosing $\pi = \sigma_1$ and σ_2 in succession one has

$$\sum_{i=1}^{n} g\big(i/n, \sigma_1(i)/n\big) = \sum_{i=1}^{n} g\big(i/n, \sigma_2(i)/n\big) = C_n,$$

giving

$$g(i/n, j/n) + g(1, 1) = g(i/n, 1) + g(1, j/n).$$

Clearly, this holds for $i = n$ or $j = n$ as well by direct substitution. Thus, fixing $x, y \in (0, 1]$ and choosing $i = \lceil nx \rceil$, $j = \lceil ny \rceil$ and letting $n \to \infty$, on invoking continuity of g one gets

$$g(x, y) + g(1, 1) = g(x, 1) + g(1, y).$$

On integrating w.r.t. y and using the condition $g = \theta[f_1 - f_2] \in \mathcal{C}$ this gives $g(x, 1) = g(1, 1)$. By symmetry, one has $g(1, y) = g(1, 1)$, and so $g(x, y) = g(1, 1)$ as well, which forces $g(x, y) \equiv 0$ again invoking the condition $g \in \mathcal{C}$. This completes the proof of the proposition. □

Another set of constraints which would have served the same purpose is $f(x, 0) = 0$, $\forall x \in [0, 1]$; $f(0, y) = 0$, $\forall y \in [0, 1]$. For the sake of definiteness, this paper uses (1.2). This mimics the condition in the discrete setting that the row and column sums of a square matrix are all 0. It should be noted here that the function $f(x, y) = xy$ does not belong to $\mathcal{C}$, and it should be replaced by the function $f(x, y) = (x - 1/2)(y - 1/2)$. However, this is not done in Sections 2 and 3 to simplify notation, on observing that all the proofs and conclusions of this paper go through as long as $f(x, y)$ cannot be written as $\phi(x) + \psi(y)$, which is true for $f(x, y) = xy$.

1.4. *Statement of main results*. The first main result of this paper is the following theorem which computes the limiting value of the log normalizing constant of models of the form (1.1) for a general continuous function f in terms of an optimization problem over copulas.

DEFINITION 1.4. Let $\mathcal{M}$ denote the space of all probability distributions on the unit square with uniform marginals.

THEOREM 1.5. *For any function $f \in \mathcal{C}$ and $\theta \in \mathbb{R}$, consider the probability model $\mathbb{Q}_{n,f,\theta}(\pi)$ as defined in* (1.1). *Then the following conclusions hold*:

(a)

$$\lim_{n\to\infty} \frac{Z_n(f, \theta) - Z_n(0)}{n} = Z(f, \theta) := \sup_{\mu\in\mathcal{M}} \big\{\theta\mu[f] - D(\mu\|u)\big\},$$

where u is the uniform distribution on the unit square, $\mu[f] := \int f\, d\mu$ is the expectation of f with respect to the measure μ and $D(\cdot\|\cdot)$ is the Kullback–Leibler divergence.

(b) *If $\pi \in S_n$ is a random permutation from the model $\mathbb{Q}_{n,f,\theta}$, then the random probability measure*

$$\nu_\pi := \frac{1}{n}\sum_{i=1}^{n} \delta_{(i/n,\pi(i)/n)}$$

on the unit square converge weakly in probability to the probability measure $\mu_{f,\theta} \in \mathcal{M}$, where $\mu_{f,\theta}$ the unique maximizer of part (a).

(c) *The measure $\mu_{f,\theta}$ of part* (b) *has density*

$$g_{f,\theta}(x, y) := e^{\theta f(x,y)+a_{f,\theta}(x)+b_{f,\theta}(y)}$$

with respect to Lebesgue measure on $[0, 1]^2$, with the functions $a_{f,\theta}(\cdot)$ and $b_{f,\theta}(\cdot) \in L^1[0, 1]$ which are unique almost surely. Consequently, one has

$$\sup_{\mu\in\mathcal{M}} \{\theta\mu[f] - D(\mu\|u)\} = -\int_{x=0}^{1} \big[a_{f,\theta}(x) + b_{f,\theta}(x)\big]\,dx.$$

(d) *The function $Z(f, \theta)$ of part* (b) *is a differentiable convex function with a continuous and strictly increasing derivative $Z'(f, \theta)$ which satisfies*

$$Z'(f, \theta) = \lim_{n\to\infty} \frac{1}{n} Z'_n(f, \theta) = \mu_{f,\theta}[f].$$

REMARK 1.6. Part (b) of the above theorem gives one way to visualize a permutation π as a measure ν_π on the unit square. A somewhat similar way to view a permutation π as a measure is presented in the supplementary file [42]. It also demonstrates how the measure ν_π looks like, when π is a large permutation from $\mathbb{Q}_{n,f,\theta}$. As an example, setting $\theta = 0$ one gets the uniform distribution on S_n, when the limiting measure $\mu_{f,\theta}$ becomes u the uniform distribution on $[0, 1]^2$. Note that the theorem statement uses $Z_n(0)$ instead of $Z_n(f, 0)$. This is because $Z_n(f, 0) = \log n!$ for all choices of the function f, and so the use of the notation $Z_n(0)$ is without loss of generality.

Focusing on inference about θ when an observation π is obtained from the model $\mathbb{Q}_{n,f,\theta}$, then the following corollary of Theorem 1.5 shows consistency of the Maximum Likelihood Estimate (MLE). In this model, MLE for θ is the solution to the equation

$$\left\{\frac{1}{n}\sum_{i=1}^{n} f\big(i/n, \pi(i)/n\big) - \frac{1}{n} Z'_n(f, \theta)\right\} = 0.$$

Since $Z_n(f, \theta)$ and $Z'_n(f, \theta)$ are hard to compute numerically, as an approximation one can replace the quantity $\frac{1}{n} Z'_n(f, \theta)$ above by its limiting value $Z'(f, \theta)$ and then solve for θ. The following corollary shows that this estimate is consistent for θ as well.

COROLLARY 1.7. *For $f \in \mathcal{C}$, consider the model $\mathbb{Q}_{n,f,\theta}$ as in (1.1), and let π be an observation from this model.*

(a) *In this case, one has*

$$\frac{1}{n}\sum_{i=1}^{n} f\big(i/n, \pi(i)/n\big) \xrightarrow{P} Z'(f,\theta) = \mu_{f,\theta}[f]$$

for every $\theta \in \mathbb{R}$.

(b) *Both the expressions*

$$\mathrm{ML}_n(\pi,\theta) := \frac{1}{n}\sum_{i=1}^{n} f\big(i/n, \pi(i)/n\big) - Z_n'(f,\theta),$$

$$\mathrm{LD}_n(\pi,\theta) := \frac{1}{n}\sum_{i=1}^{n} f\big(i/n, \pi(i)/n\big) - Z'(f,\theta)$$

have unique roots $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LD}}$ with probability tending to 1 which are consistent for θ.

(c) *Consider the testing problem of $\theta = \theta_0$ versus $\theta = \theta_1$ with $\theta_1 > \theta_0$. Then the test $\phi_n := 1\{\hat{\theta}_{\mathrm{LD}} > (\theta_0+\theta_1)/2\}$ is consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}_{n,f,\theta_0}} \phi_n = 0, \qquad \lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}_{n,f,\theta_1}} \phi_n = 1.$$

The above corollary shows that it is possible to estimate the parameter θ consistently with just one observation from the model $\mathbb{Q}_{n,f,\theta}$. No error rates have been obtained for the estimates $\{\hat{\theta}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{LD}}\}$ in this paper, as part (a) of Theorem 1.5 does not have any error rates. Thus, a good approximation of the limiting log normalizing constant will lead to an efficient estimator for θ, in the sense that the estimator will be close to the MLE. The definition of $Z(f,\theta)$ is in terms of an optimization problem over $\mathcal{M}$, which is an infinite-dimensional space. In general, such optimization can be hard to carry out. The next theorem gives an iterative algorithm for computing the density of the optimizing measure $\mu_{f,\theta}$ with respect to Lebesgue measure. Intuitively, the algorithm starts with the function $e^{\theta f(x,y)}$ and alternately scales it along x and y marginals to produce uniform marginals in the limit.

DEFINITION 1.8. For any integer $k \geq 1$, let $\mathcal{M}_k$ denote the set of all $k \times k$ matrices with nonnegative entries with both row and column sums equal to $1/k$.

THEOREM 1.9. (a) *Define a sequence of $k \times k$ matrices by setting $B_0(r,s) := e^{f(r/k,s/k)}$ for $1 \leq r, s \leq k$, and*

$$B_{2m+1}(r,s) := \frac{B_{2m}(r,s)}{k\sum_{l=1}^{m} B_{2m}(r,l)}, \qquad B_{2m+2}(r,s) := \frac{B_{2m+1}(r,s)}{k\sum_{l=1}^{m} B_{2m+1}(l,s)}.$$

Then there exists a matrix $A_{k,\theta} \in \mathcal{M}_k$ *such that* $\lim_{m\to\infty} B_m = A_k$.

(b) $A_{k,\theta} \in \mathcal{M}_k$ *is the unique maximizer of the optimization problem*

$$\sup_{A\in\mathcal{M}_k}\left\{\theta\sum_{r,s=1}^{k} f(r/k,s/k)A(r,s) - 2\log k - \sum_{r,s=1}^{k} A(r,s)\log A(r,s)\right\}.$$

(c) *The function*

$$W_k(f,\theta) := \sup_{A\in\mathcal{M}_k}\left\{\theta\sum_{r,s=1}^{k} f(r/k,s/k)A(r,s) - 2\log k - \sum_{r,s=1}^{k} A(r,s)\log A(r,s)\right\}$$

is a convex differentiable function in θ *for fixed* f *with*

$$W_k'(f,\theta) = \sum_{r,s=1}^{k} A_{k,\theta}(r,s) f(r/k,s/k).$$

(d) *Finally, for any continuous function* $\phi : [0,1]^2 \mapsto \mathbb{R}$ *one has*

$$\lim_{k\to\infty}\sum_{r,s=1}^{k} A_{k,\theta}(r,s)\phi(r/k,s/k) = \int_{[0,1]^2}\phi(x,y)g_{f,\theta}(x,y)\,dx\,dy.$$

In particular, this implies

$$Z(f,\theta) = \lim_{k\to\infty} W_k(f,\theta) = \lim_{k\to\infty}\lim_{m\to\infty}\left\{\theta\sum_{i,j=1}^{k} f(i/k,j/k)B_m(i,j) - 2\log k - \sum_{i,j=1}^{k} B_m(i,j)\log B_m(i,j)\right\}.$$

REMARK 1.10. Since $g_{f,\theta}(x,y)$ has uniform marginals, the functions $a_{f,\theta}(\cdot)$ and $b_{f,\theta}(\cdot)$ are the solutions to the joint integral equations

$$\int_0^1 e^{\theta f(x,z)+a_{f,\theta}(x)+b_{f,\theta}(z)}\,dz = 1,$$

$$\int_0^1 e^{\theta f(z,y)+a_{f,\theta}(z)+b_{f,\theta}(y)}\,dz = 1 \qquad \forall x,y\in[0,1].$$

By Theorem 1.9, it follows that

$$\lim_{n\to\infty}\frac{Z_n(f,\theta)-Z_n(0)}{n} = -\int_{x=0}^{1}\left[a_{f,\theta}(x)+b_{f,\theta}(x)\right]dx.$$

For the limiting normalizing constant in the Mallows model with the Foot-rule or the Spearman's rank correlation, one needs to take $f(x, y) = -|x - y|$ and $f(x, y) = -(x-y)^2$ [or $f(x, y) = xy$], respectively. Even though analytic computation for $a_{f,\theta}(\cdot), b_{f,\theta}(\cdot)$ might be difficult, the algorithm of Theorem 1.9 (known as IPFP) can be used for a numerical evaluation of these functions. Iterative Proportional Fitting Procedure (IPFP) originated in the works of Deming and Stephan [12] in 1940. For more background on IPFP, see [11, 29, 44, 47] and the references therein. Theorem 1.9 gives a way to approximate numerically the limiting log partition function by fixing k large and running the IPFP for m iterations with a suitably large m.

Another approach for estimation in such models can be to estimate the parameter θ without estimating the normalizing constant. The following theorem constructs an explicit $\sqrt{n}$ consistent estimator for θ, for the class of models considered in Theorem 1.5. This estimate is similar in spirit to Besag's pseudo-likelihood estimator [4, 5]. The pseudo-likelihood is defined to be the product of all one-dimensional conditional distributions, one for each random variable. Since in a permutation the conditional distribution of $\pi(i)$ given $\{\pi(j), j \neq i\}$ determines the value of $\pi(i)$, it does not make sense to look at the conditional distribution $(\pi(i)|\pi(j), j \neq i)$. In this case, a meaningful thing to consider is the distribution of $(\pi(i), \pi(j)|\pi(k), k \neq i, j)$, which gives the pseudo-likelihood as

$$\prod_{1 \leq i < j \leq n} \mathbb{Q}_{n,f,\theta}\big(\pi(i), \pi(j)|\pi(k), k \neq i, j\big).$$

The pseudo-likelihood estimate $\hat{\theta}_{\mathrm{PL}}$ is obtained by maximizing the above expression. Taking the log of the pseudo-likelihood and differentiating with respect to θ gives

$$\sum_{1 \leq i < j \leq n} y_\pi(i, j) \frac{1}{1 + e^{\theta y_\pi(i,j)}},$$

where

$$\begin{aligned} y_\pi(i, j) := {} & f\big(i/n, \pi(i)/n\big) + f\big(j/n, \pi(j)/n\big) - f\big(i/n, \pi(j)/n\big) \\ & - f\big(j/n, \pi(i)/n\big). \end{aligned}$$

The pseudo-likelihood estimate can then be obtained by equating this to 0 and solving for θ. One way of computing this estimate is a binary search or other gradient based search methods, and does not require the computation of $Z_n(f, \theta)$. Thus, this gives a fast and practical way for parameter estimation in such models. The next theorem gives error rates for the pseudo-likelihood estimator.

THEOREM 1.11. *For $f \in \mathcal{C}$, consider the model $\mathbb{Q}_{n,f,\theta}$ of (1.1), and let π be a sample from $\mathbb{Q}_{n,f,\theta}$. Setting*

$$y_\pi(i,j) := f\big(i/n, \pi(i)/n\big) + f\big(j/n, \pi(j)/n\big) - f\big(i/n, \pi(j)/n\big) - f\big(j/n, \pi(i)/n\big)$$

the expression

$$\mathrm{PL}_n(\pi,\theta) := \sum_{1\le i<j\le n} y_\pi(i,j)\frac{1}{1+e^{\theta y_\pi(i,j)}}$$

has a unique root in θ with probability tending to 1. *Further, denoting this root by $\hat{\theta}_n$ one has $\sqrt{n}(\hat{\theta}-\theta)$ is $O_P(1)$.*

The estimating equations $\mathrm{LD}_n(\pi,\theta)$ and $\mathrm{ML}_n(\pi,\theta)$ of Corollary 1.7 and $\mathrm{PL}_n(\pi,\theta)$ of Theorem 1.11 are stated when a single permutation π is observed from $\mathbb{Q}_{n,f,\theta}$. If i.i.d. samples $\pi^{(1)},\dots,\pi^{(m)}$ are observed from $\mathbb{Q}_{n,f,\theta,}$ one should use the equations

$$\sum_{l=1}^{m}\mathrm{LD}_n\big(\pi^{(l)},\theta\big), \qquad \sum_{l=1}^{m}\mathrm{ML}_n\big(\pi^{(l)},\theta\big), \qquad \sum_{l=1}^{m}\mathrm{PL}_n\big(\pi^{(l)},\theta\big)$$

instead.

So far all results relate to the model $\mathbb{Q}_{n,f,\theta}$ as defined in (1.1). To demonstrate that the tools used to prove these results are quite robust, the next proposition analyzes the Mallows model with Kendall's tau [item (d) in the original list of divergences].

PROPOSITION 1.12. *Consider the Mallows model on S_n with Kendall's tau, defined by*

$$M_{n,\theta}(\pi) := e^{(\theta/n)\,\mathrm{Inv}(\pi)-C_n(\theta)}, \qquad \mathrm{Inv}(\pi) := \sum_{i<j} 1_{\pi(i)>\pi(j)},$$

where $C_n(\theta)$ is the normalizing constant. Also, let $h:[0,1]^4\mapsto\mathbb{R}$ denote the function

$$h\big((x_1,y_1),(x_2,y_2)\big) := 1_{(x_1-x_2)(y_1-y_2)<0}.$$

(a) *In this case, one has*

$$\lim_{n\to\infty}\frac{C_n(\theta)-C_n(0)}{n} = C(\theta) := \sup_{\mu\in\mathcal{M}}\left\{\frac{\theta}{2}(\mu\times\mu)(h) - D(\mu\|u)\right\}.$$

Further, the supremum above is attained at a unique measure on the unit square given by the density

$$\rho_\theta(x, y) := \frac{(\theta/2)\sinh(\theta/2)}{[e^{-\theta/4}\cosh(\theta(x-y)/2) + e^{\theta/4}\cosh(\theta(x+y-1)/2)]^2},$$

and consequently $C(\theta) = \int_0^1 \frac{e^{\theta x}-1}{\theta x}\,dx$.

(b) *If* π *is a sample from* $M_{n,\theta}$, *then both the expressions*

$$\widetilde{\mathrm{ML}}_n(\pi, \theta) := \frac{1}{n^2}\,\mathrm{Inv}(\pi) - C_n'(\theta),$$

$$\widetilde{\mathrm{LD}}_n(\pi, \theta) := \frac{1}{n^2}\,\mathrm{Inv}(\pi) - C'(\theta)$$

have unique roots $\widetilde{\theta}_{\mathrm{ML}}$ *and* $\widetilde{\theta}_{\mathrm{LD}}$ *with probability tending to* 1 *which are consistent for* θ.

REMARK 1.13. The connection between the Mallows model with Kendall's tau and the density $\rho_\theta(\cdot,\cdot)$ was first demonstrated in [49]. Since for this model the log normalizing constant $C_n(\theta)$ is explicitly known, the formula for $C(\theta)$ can be computed easily. In this case by a direct argument one can show that $\widetilde{\theta}_{\mathrm{LD}}, \widetilde{\theta}_{\mathrm{ML}}$ are $\sqrt{n}$ consistent. The theorem shows that the general tools developed in this paper can also be used to show consistency, even though establishing optimal rates requires finer results.

As mentioned before, the techniques of this paper can also be used to comment on the joint limiting distribution of $\{\pi(1), \ldots, \pi(n)\}$. This is demonstrated in Proposition 2.2 contained in the supplementary file [42] for academic interest, and is not used in the rest of the paper.

1.5. *Scope of future research.* Even though the Mallows model with Kendall's tau is not in the setting of Theorem 1.5, estimation of the log normalization constant is still possible using results of this paper. This is because the function

$$\mu \mapsto \int_{[0,1]^4} 1_{(x_1-x_2)(y_1-y_2)<0}\,d\mu(x_1, y_1)\,d\mu(x_2, y_2)$$

is continuous on $\mathcal{M}$ with respect to weak convergence, and is a natural extension for the number of inversions of a permutation to a general probability measure in $\mathcal{M}$. Thus, to explore other nonuniform models on permutations, one needs to understand the continuous real valued functionals on $\mathcal{M}$. For an example of a natural function on permutations which is not continuous, let $N(\pi)$ denote the number of fixed points of π. Then the function $\pi \mapsto N(\pi)/n$ is not continuous on $\mathcal{M}$. Indeed, its natural analogue on $\mathcal{M}$ is the function

$$\mu \mapsto \int_{[0,1]^2} 1_{x=y}\,d\mu(x, y),$$

which is not continuous with respect to weak topology on $\mathcal{M}$. Extending the framework to cover examples of such noncontinuous functions is under consideration by the author.

For the Mallows model with Kendall's tau, the limiting normalizing constant is available in closed form, which is not the case for the Mallows model with Spearman's correlation. To make the answer explicit, one needs to solve the optimization problem of Theorem 1.5. An explicit answer will not only give the limiting log normalizing constant, but also provide a more explicit description for the joint limiting distribution of π via Proposition 2.2 in [42].

Another interesting problem is to compute the limiting distribution of $\sum_{i=1}^{n} f(i/n, \pi(i)/n)$ under the model $\mathbb{Q}_{n,f,\theta}$. Under uniform distribution on S_n, this statistic has a limiting normal distribution if $f \in \mathcal{C}$, by Hoeffding's combinatorial central limit theorem ([21], Theorem 3). Theorem 1.5 shows that $\frac{1}{n}\sum_{i=1}^{n} f(i/n, \pi(i)/n)$ converges to a constant, and gives a characterization of this constant in terms of permutation limits. It, however, fails to find nondegenerate limit distribution for this statistic. If one is interested in the testing problem of $\theta = \theta_0$ versus $\theta = \theta_1$ as in Corollary 1.7, then this distribution will be useful in determination of exact cut-offs under null hypothesis, and evaluation of power under the alternative. Also, using such distribution results, it should be possible to find out limit distributions of the estimators considered in this paper.

Finally, this paper explores the asymptotics of parametric models on permutations. Viewing a permutation as a measure, one can study nonparametric models on permutations as well, and in fact one class of models was introduced and studied in [22]. Such models can be used to fit permutations. This technique can also be used for comparing permutations in a nonparametric manner, such as in a classification problems on permutations. Section 3 gives a visual comparison, but comparisons can also be carried out in a more precise manner using a "suitable" metric for bivariate probability measures.

1.6. *Outline*. Section 2 explores the Mallows model with Spearman's rank correlation as sufficient statistic, using the results of this paper. Section 3 analyzes the draft lottery data of 1971. The supplemental article [42] describes in brief the concept of permutation limits introduced in [22], and proves a large deviation principle for permutations in Theorem 1.1 in [42]. It also carries out the proofs of the main results of this paper using this large deviation.

2. An example: Spearman's rank correlation divergence. This section illustrates the conclusions of Theorem 1.5 and Theorem 1.9 with a concrete example, the Spearman's rank correlation model. This is number (b) in the list of divergences in the Introduction. The Spearman's rank correlation statistic is given by

$$\|\pi - \sigma\|_2^2 = \sum_{i=1}^{n} \big(\pi(i) - \sigma(i)\big)^2.$$

This divergence function is right invariant, and has received attention in Statistics literature (see [9, 13, 16, 34] and references therein). The reason for its nomenclature is that if π and σ are two permutations of size n, then the simple correlation coefficient of the points $\{\pi(i), \sigma(i)\}_{i=1}^n$ has the formula

$$r(\pi, \sigma) = 1 - \frac{6\|\pi - \sigma\|_2^2}{n(n^2-1)},$$

which is a one–one function of $\|\pi - \sigma\|_2^2$.

Even for this simple divergence the normalizing constant for the corresponding Mallows model is not available in closed form. As observed in the Introduction, the Spearman's rank correlation model is obtained by setting $f(x, y) = -(x-y)^2$ or $f(x, y) = xy$ in the model of Theorem 1.5.

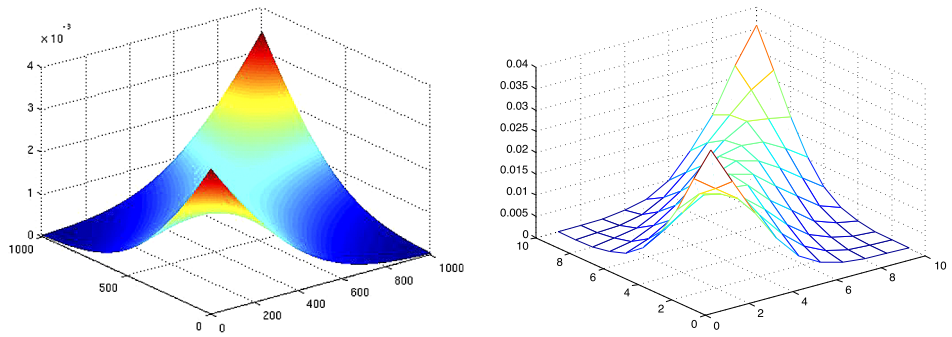
This section will work with the choice $f(x, y) = xy$. To be precise, the p.m.f. of this model is

$$\mathbb{Q}_{n,f,\theta} = e^{(\theta/n^2)\sum_{i=1}^n i\pi(i) - Z_n(f,\theta)},$$

where $Z_n(f, \theta)$ is the appropriate log normalizing constant as before. By the discussion after equation (1.2), it follows that Theorem 1.5 is applicable for $f(x, y) = xy$. Thus, if π is a random permutation from $\mathbb{Q}_{n,f,\theta}$ for this f, then the empirical measure $\nu_\pi = \frac{1}{n}\sum_{i=1}^n \delta_{(i/n,\pi(i)/n)}$ converges weakly to a measure in $\mathcal{M}$ with density of the form

$$g_{f,\theta}(x, y) = e^{\theta xy + a_{f,\theta}(x) + a_{f,\theta}(y)},$$

where the symmetry of f has been used to deduce $b_{f,\theta}(\cdot) = a_{f,\theta}(\cdot)$. However, analytic solution of $g_{f,\theta}(\cdot)$ seems intractable and is not attempted here. Instead, Figure 1(a) plots the density $g_{f,\theta}(x, y) = e^{\theta xy + a_{f,\theta}(x) + a_{f,\theta}(y)}$ on a discrete grid of size $k \times k$ with $k = 1000$. The values of the function are computed by the algorithm



(a) (b)

FIG. 1. (a) *Density of limiting measure* $\mu_{f,\theta}$ *for* $f(x, y) = xy, \theta = 20$, (b) *Histogram of* ν_π *with* $n = 10{,}000$ *with* 10×10 *bins.*

of Theorem 1.9 starting with the $k \times k$ matrix $B_0(i, j) = e^{(\theta/k^2)ij}$, where $\theta = 20$. Part (d) of Theorem 1.9 implies that $k^2 B_m(i, j)$ can be taken as an approximation of the limiting density $g_{f,\theta}(i/k, j/k)$.

From Figure 1, it is easy to see that $g_{f,\theta}$ has higher values on the diagonal $x = y$, which also follows from the fact that for $\theta > 0$ the identity permutation has the largest probability under this model. The function $g_{f,\theta}(\cdot, \cdot)$ is symmetric about the diagonal $x = y$, which follows from the fact that $f(\cdot, \cdot)$ is symmetric. Another way to see this is by noting that if π converges to a probability measure on $[0, 1]^2$ with limiting density $g_{f,\theta}(x, y)$, then π^{-1} converges to a measure on $[0, 1]^2$ with limiting density $g_{f,\theta}(y, x)$. But since

$$\sum_{i=1}^{n} i\pi(i) = \sum_{i=1}^{n} i\pi^{-1}(i),$$

the law of π and π^{-1} are same under $\mathbb{Q}_{n,f,\theta}$, and so π^{-1} has the limiting density $g_\theta(x, y)$ as well, thus giving $g_{f,\theta}(x, y) = g_{f,\theta}(y, x)$.

The function is also symmetric about the other diagonal $x + y = 1$. A similar reasoning as above justifies this:

Define $\sigma \in S_n$ by $\sigma(i) := n+1-\pi^{-1}(n+1-\pi(i))$ and note that if π converges to a probability on $[0, 1]^2$ with density $g_{f,\theta}(x, y)$, then σ converges to a probability on $[0, 1]^2$ with density $g_{f,\theta}(1 - y, 1 - x)$. But since

$$\sum_{i=1}^{n} i\pi(i) = \sum_{i=1}^{n} (n + 1 - i)\big(n + 1 - \pi(i)\big) = \sum_{i=1}^{n} i\sigma(i),$$

it follows that under $\mathbb{Q}_{n,f,\theta}$ the distribution of π is same as the distribution of σ. Thus, σ has limiting density $g_{f,\theta}(x, y)$ as well, which implies $g_{f,\theta}(x, y) = g_{f,\theta}(1 - y, 1 - x)$, and so $g_{f,\theta}$ is symmetric about the line $x + y = 1$.

To compare how close the empirical measure ν_π is to the limit, a random permutation π of size $n = 10{,}000$ is drawn from $\mathbb{Q}_{n,f,\theta}$ via MCMC. The algorithm used to simulate from this model is adopted from [2], and is explained below:

1. Start with π chosen uniformly at random from S_n.

2. Given π, simulate $\{U_i\}_{i=1}^n$ mutually independent with U_i uniform on $[0, e^{(\theta/n^2)i\pi(i)}]$.

3. Given U, let $b_j := \max\{(n^2/\theta j)\log U_j, 1\}$. Then $1 \le b_j \le n$. Choose an index i_1 uniformly at random from set $\{j : b_j \le 1\}$, and set $\pi(i_1) = 1$. Remove this index from $[n]$ and choose an index i_2 uniformly from $\{j : b_j \le 2\} - \{i_1\}$, and set $\sigma(i_2) = 2$. In general, having defined $\{i_1, \ldots, i_{l-1}\}$, remove them from $[n]$, and choose i_l uniformly from $\{j : b_j \le l\} - \{i_1, i_2 \cdots i_{l-1}\}$, and set $\pi(i_l) = l$. (That this step can be always carried out completely was proved in [14].)

4. Iterate between the steps 2 and 3 until convergence.

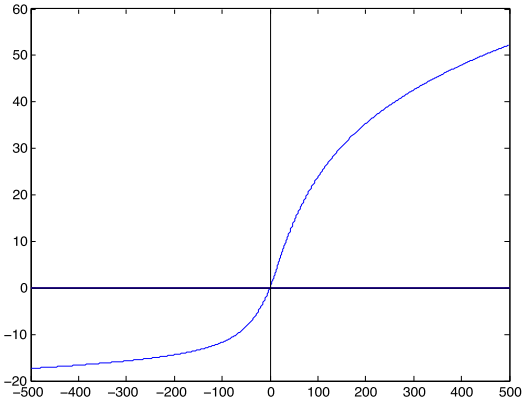


FIG. 2. *Plot of θ versus $Z(f,\theta)$ for rank correlation model.*

The above iteration is run 10 times to obtain a single permutation π, and then the frequency histogram of the points $\{i/n, \pi(i)/n\}_{i=1}^{n}$ are computed with $k \times k$ bins, where $k = 10$. The mesh plot of the frequency histogram is given in Figure 1(b).

The pattern of the histogram in Figure 1(b) is very similar to the function plotted in Figure 1(a), showing that the probability assigned by the random permutation π has a similar pattern as that of the limiting density $g_{f,\theta}(x, y)$. The histogram has been drawn with k^2 squares, each of size 0.1 as $k = 10$.

Using Theorem 1.9 gives an approximation to $\frac{1}{n}[Z_n(f,\theta) - Z_n(0)]$ as

$$\frac{\theta}{k^2}\sum_{r,s=1}^{k} ij\,B_m(r,s) - 2\log k - \sum_{r,s=1}^{k} B_m(r,s)\log B_m(r,s).$$

Figure 2 gives a plot of θ versus $\lim_{n\to\infty}\frac{1}{n}[Z_n(f,\theta) - Z_n(0)]$, where the limiting value is estimated using the above approximation. For this plot, k has been chosen to be 100, and the range of θ has been taken to be $[-500, 500]$. The number of iterations for the convergence of the iterative algorithm for each θ has been taken as 20. The curve passes through $(0, 0)$, and goes to $\pm\infty$ as θ goes to $\pm\infty$, as expected.

The above method can be used to approximate the limiting log normalizing constant for any model of permutations described in the setting Theorem 1.5.

3. Analysis of the 1970 draft lottery data. This section analyses the 1970 draft lottery data using the methods developed in this paper. The data for this lottery is taken from http://www.sss.gov/LOTTER8.HTM. This lottery was used to determine the relative order in which male US citizens born between 1944–1950 will join the army, based on their birthdays. As an example, September 14th was the first chosen day, which means that people born on this date had to join first.

Assume that the 366 days of the year are chronologically numbered, that is, January 1 is day 1, and December 31 is day 366. Then the data can be represented in the form of a permutation of size 366, where $\pi(i)$ represents the ith day chosen in the lottery. The lottery was carried out in a somewhat flawed manner as follows:

366 capsules were made, one for each day of the year. The January capsules were put in a box first, and then mixed among themselves. The February capsules were then put in the box, and the capsules for the first two months were mixed. This was carried on until the December capsules were put in the box, and all the capsules were mixed. As a result of this mixing, the January capsules were mixed 12 times, the February capsules were mixed 11 times, and the December capsules were mixed just once. As a result, most of the capsules for the latter months stayed near the top, and ended up being drawn early in the lottery. The resulting permutation π thus seems to have a bias toward the permutation

$$(366, 365, \ldots, 1),$$

and so the permutation $\tau = 367 - \pi$ should be biased toward the identity.

Thus, the question of interest is to test whether the permutation τ is chosen uniformly at random from S_{366}, and the alternative hypothesis is that τ has a bias toward the identity permutation. For $\tau \in S_n$ with $n = 366$, one can construct the histogram of the points

$$\left\{\left(\frac{i}{n}, \frac{\tau(i)}{n}\right), 1 \le i \le n\right\}.$$

If τ is indeed drawn from the uniform distribution on S_n, then this histogram should be close to the uniform distribution on the unit square. The bivariate histogram is drawn with 10×10 bins in Figure 3(a). To compare this with the uniform distribution on S_n, a uniformly random permutation σ is chosen from S_n, and the

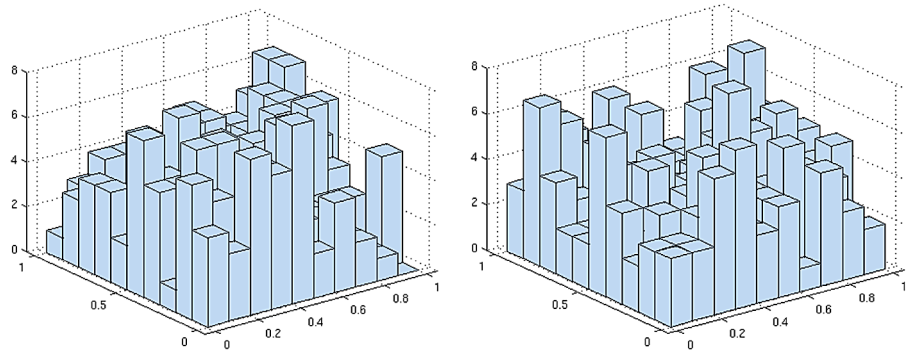


FIG. 3. *Bivariate histogram of the points* $\frac{1}{366}\{(i, \tau(i)), 1 \le i \le 366\}$ *with* 10×10 *bins where* τ *is* (a) *reverse permutation of draft lottery*, (b) *a random permutation chosen uniformly.*

corresponding histogram is drawn in Figure 3(b) with the same the number of bins as above. From Figure 3, it seems that the heights of the bins in the second picture are a bit more uniform than the first.

If τ is indeed uniform, then the statistic $\frac{1}{n^3}\sum_{i=1}^{n} i\tau(i)$ has a limiting normal distribution with mean $\frac{1}{4}(1+\frac{1}{n})^2 \approx 0.25$ and variance $\frac{1}{144n}(1-\frac{1}{n})(1+\frac{1}{n})^2 \approx \frac{1}{144n} \approx 1.89\times 10^{-5}$ ([13], page 116). The observed value of this statistic is 0.2702, which clearly falls outside a 99% acceptance region under the null hypothesis. Even if the normal approximation is not believed, by Chebyshev's inequality one has

$$\mathbb{P}_{\theta=0}\left(\frac{1}{366^3}\sum_{i=1}^{366} i\tau(i) \geq 0.2702\right) \leq \frac{1.89\times 10^{-5}}{0.0502^2} \approx 0.0075,$$

which suggests very strong evidence against the null hypothesis.

If τ is assumed to be generated from the model

$$\mathbb{Q}_{n,f,\theta}(\tau) = e^{(\theta/n^2)\sum_{i=1}^{n} i\tau(i) - Z_n(f,\theta)},$$

where $f(x,y)=xy$, the test used above is the most powerful test (in the sense of NP lemma) for testing $\theta=0$ versus $\theta>0$. Since the null is rejected, it might be of interest to see if there is another value of θ for which the model better fits the data. To investigate this, the value of θ is estimated using the estimators $\hat{\theta}_{\mathrm{LD}}$ of Corollary 1.7 and $\hat{\theta}_{\mathrm{PL}}$ of Theorem 1.11. By a direct computation, it turns out that $\hat{\theta}_{\mathrm{PL}} = 2.92$. To compute $\hat{\theta}_{\mathrm{LD}}$ requires estimating the limiting log normalizing constant, for which one needs to carry out the IPFP algorithm of Theorem 1.9. The grid size chosen for computing $\hat{\theta}_{\mathrm{LD}}$ is 1000×1000. It follows from the proof of Theorem 1.9 that the error in approximating the limiting log partition function $Z(f,\theta)$ by a k step approximation $W_k(f,\theta)$ is bounded by $|\theta|\varepsilon_k$, where

$$\varepsilon_k = \sup_{|x_1-x_2|\leq 1/k, |y_1-y_2|\leq 1/k} \bigl|f(x_1,y_1) - f(x_2,y_2)\bigr| \leq \frac{2}{k}.$$

Thus, a choice of $k=1000$ should ensure that the limiting log partition function is correct up to the first two decimal places, assuming the run time m is large. Larger values of k will increase accuracy of the estimate, at the cost of speed of computation. For each value of θ, the IPFP algorithm is run $m=200$ times. The estimate $\hat{\theta}_{\mathrm{LD}}$ turns out to be 2.96, which is close to the pseudo-likelihood estimate. To compare the relative performance of the two estimators $\hat{\theta}_{\mathrm{PL}}$ and $\hat{\theta}_{\mathrm{LD}}$, a sample of 1000 values is drawn from this model for $\theta=2.92$ and $\theta=2.96$, and the histogram of the statistic $n^{-3}\sum_{i=1}^{n} i\tau(i)$ is plotted side by side in Figure 4 with 25 bins. The observed value from the draft lottery data is 0.2702, represented by the green line. From Figure 4, it is clear that both estimates give a good fit to the observed data.

Finally, to test whether these values of θ gives a good fit to the given data, an independent random permutation $\hat{\tau}$ is drawn from this model with $\theta=2.92$. The

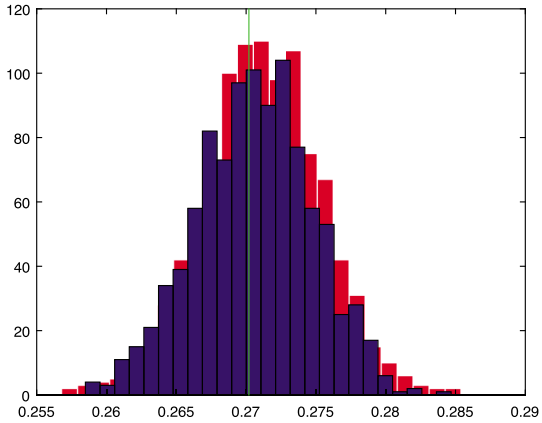
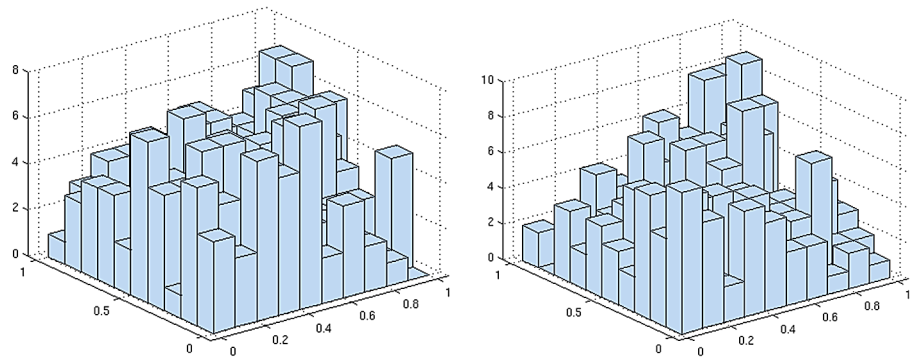


FIG. 4. *Histogram of the statistic* $366^{-3}\sum_{i=1}^{366} i\tau(i)$ *with* 1000 *independent draws grouped into* 25 *bins, where* τ *is a random permutation from Spearman's rank correlation model with* (a) $\theta = 2.92$ *in blue* (*pseudo-likelihood*), (b) $\theta = 2.96$ *in red* (*LD-MLE*). *The green line at* 0.2702 *is obtained when* τ *is the reverse permutation of draft lottery data.*

same auxiliary variable algorithm of Andersen–Diaconis from the previous section is used to draw the sample. The histogram of $\hat{\tau}$ is given in Figure 5(b) with 10×10 bins, along side the histogram for the observed permutation τ in 5(a).

The bivariate histogram of the points $(i/n, \tau(i)/n)_{i=1}^{n}$ for the observed permutation τ and the points $(i/n, \hat{\tau}(i)/n)_{i=1}^{n}$ for the simulated permutation $\hat{\tau}$ is drawn in Figure 5. This seems to be a better match than the histograms for τ and σ in Figure 5, where σ was a permutation drawn uniformly at random. This agrees with the observation made in [17] that the observed permutation does not seem uniformly random.



(a) (b)

FIG. 5. *Bivariate histogram of the points* $\frac{1}{366}\{(i, \tau(i)), 1 \le i \le 366\}$ *with* 10×10 *bins where* τ *is* (a) *reverse permutation of draft lottery,* (b) *a random permutation chosen from Spearman's model with* $\theta = 2.92$.

Acknowledgement. This paper benefited from helpful discussions with Persi Diaconis, Amir Dembo, Sourav Chatterjee, Susan Holmes, Bhaswar Bhattacharya and Austen Head. I would like to thank Maxwell Grazier G'Sell for help with acquiring the draft lottery data. I would also like to thank the Editor, Associate Editor and two referees, whose recommendations improved significantly the content and presentation of this paper.

The contents of this paper also appear in the author's Ph.D. thesis advised by Persi Diaconis.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation in exponential families on permutations" (DOI: 10.1214/15-AOS1389SUPP; .pdf). The supplementary material contain the proofs of all theorems, corollaries, propositions and supporting lemmas. It also states Proposition 2.2, which deals with the joint limiting distribution of $\{\pi(1), \ldots, \pi(n)\}$ for π from either the model $\mathbb{Q}_{n,f,\theta}$ of (1.1) or from the Mallows model with Kendall's tau of Proposition 1.12. A short proof of this proposition is included using the more recent results of [6].

REFERENCES

[1] AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884

[2] ANDERSEN, H. C. and DIACONIS, P. (2007). Hit and run as a unifying device. *J. Soc. Fr. Stat. & Rev. Stat. Appl.* **148** 5–28. MR2502361

[3] AWASTHI, P., BLUM, A., SHEFFET, O. and VIJAYARAGHAVAN, A. (2014). Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems* **27** 2609–2617. Curran Associates, Inc. Montreal.

[4] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

[5] BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

[6] BHATTACHARYA, B. and MUKHERJEE, S. (2015). Degree Sequence of Random Permutation Graphs. Preprint. Available at arXiv:1503.03582.

[7] BRIGO, D., PALLAVICINI, A. and TORRESETTI, R. (2010). *Credit Models and the Crisis: A Journey Into CDOs, Copulas, Correlations and Dynamic Models*. Wiley, New York.

[8] CHEN, H., BRANAVAN, S. R. K., BARZILAY, R. and KARGER, D. R. (2009). Content modeling using latent permutations. *J. Artificial Intelligence Res.* **36** 129–163. MR2602357

[9] CRITCHLOW, D. E. (1985). *Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics* **34**. Springer, Berlin. MR0818986

[10] CRITCHLOW, D. E., FLIGNER, M. A. and VERDUCCI, J. S. (1991). Probability models on rankings. *J. Math. Psych.* **35** 294–318. MR1128236

[11] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798

[12] DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11** 427–444. MR0003527

[13] DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069

[14] DIACONIS, P., GRAHAM, R. and HOLMES, S. P. (2001). Statistical problems involving permutations with restricted positions. In *State of the Art in Probability and Statistics* (*Leiden*, 1999). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 195–222. IMS, Beachwood, OH. MR1836562

[15] DIACONIS, P. and RAM, A. (2000). Analysis of systematic scan Metropolis algorithms using Iwahori–Hecke algebra techniques. *Michigan Math. J.* **48** 157–190. MR1786485

[16] FEIGIN, P. and COHEN, A. (1978). On a model of concordance between judges. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **40** 203–213.

[17] FIENBERG, S. (1971). Randomization and social affairs, the 1970 draft lottery. *Science* **171** 255–261.

[18] FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48** 359–369. MR0876847

[19] FLIGNER, M. A. and VERDUCCI, J. S. (1988). Multistage ranking models. *J. Amer. Statist. Assoc.* **83** 892–901. MR0963820

[20] GENEST, C. and MACKAY, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *Amer. Statist.* **40** 280–283. MR0866908

[21] HOEFFDING, W. (1951). A combinatorial central limit theorem. *Ann. Math. Stat.* **22** 558–566. MR0044058

[22] HOPPEN, C., KOHAYAKAWA, Y., MOREIRA, C. G., RÁTH, B. and MENEZES SAMPAIO, R. (2013). Limits of permutation sequences. *J. Combin. Theory Ser. B* **103** 93–113. MR2995721

[23] HUANG, J., GUESTRIN, C. and GUIBAS, L. (2009). Fourier theoretic probabilistic inference over permutations. *J. Mach. Learn. Res.* **10** 997–1070. MR2520800

[24] IRUROZKI, E., CALVO, B. and LOZANO, A. (2014). Sampling and learning the Mallows and Generalized Mallows models under the Cayley distance. Technical report. Available at https://addi.ehu.es/handle/10810/11239.

[25] IRUROZKI, E., CALVO, B. and LOZANO, A. (2014). Sampling and learning the Mallows model under the Ulam distance. Technical report. Available at https://addi.ehu.es/handle/10810/11241.

[26] IRUROZKI, E., CALVO, B. and LOZANO, A. (2014). Sampling and learning the Mallows and Weighted Mallows models under the Hamming distance. Technical report. Available at https://addi.ehu.es/handle/10810/11240.

[27] JAWORSKI, P., DURANTE, F., HÄRDLE, W. and RYCHLIK, T. (2010). Copula theory and its applications. In *Proceedings of the Workshop Held at the University of Warsaw, Warsaw, September* 25–26, 2009. *Lecture Notes in Statistics—Proceedings* **198**. Springer, Heidelberg. MR3075361

[28] KONDOR, R., HOWARD, A. and JEBARA, T. (2007). Multi-object tracking with representations of the symmetric group. In *AISTATS* **2** 211–218.

[29] KULLBACK, S. (1968). Probability densities with given marginals. *Ann. Math. Stat.* **39** 1236–1243. MR0229330

[30] LEBANON, G. and LAFFERTY, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the* 19*th International Conference on Machine Learning* 363–370. Morgan Kaufmann, San Francisco, CA.

[31] LEBANON, G. and MAO, Y. (2008). Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.* **9** 2401–2429. MR2452632

[32] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035

[33] MAI, J.-F. and SCHERER, M. (2012). *Simulating Copulas*: *Stochastic Models*, *Sampling Algorithms*, *and Applications*. *Series in Quantitative Finance* **4**. Imperial College Press, London. MR2906392

[34] MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267

[35] MARDEN, J. I. (1995). *Analyzing and Modeling Rank Data*. *Monographs on Statistics and Applied Probability* **64**. Chapman & Hall, London. MR1346107

[36] MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management*: *Concepts*, *Techniques and Tools*. Princeton Univ. Press, Princeton, NJ. MR2175089

[37] MCNEIL, A. J. and NEŠLEHOVÁ, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *Ann*. *Statist*. **37** 3059–3097. MR2541455

[38] MEILA, M. and BAO, L. (2008). Estimation and clustering with infinite rankings. In *Proceedings of the* 24*th Conference in Uncertainty in Artificial Intelligence* 393–402. Helsinki.

[39] MEILĂ, M. and BAO, L. (2010). An exponential model for infinite rankings. *J*. *Mach*. *Learn*. *Res*. **11** 3481–3518. MR2756191

[40] MEILA, M., PHADNIS, K., PATTERSON, A. and BLIMES, J. (2007). Consensus ranking under the exponential model, Technical Report 515, Dept. Statistics, Univ. Washington, Seattle, WA.

[41] MEUCCI, A. (2011). A new breed of copulas for risk and portfolio management. *Risk* **24** 122–126.

[42] MUKHERJEE, S. (2015). Supplement to "Estimation in exponential families on permutations." DOI:10.1214/15-AOS1389SUPP.

[43] NELSEN, R. B. (1999). *An Introduction to Copulas*. *Lecture Notes in Statistics* **139**. Springer, New York. MR1653203

[44] RÜSCHENDORF, L. (1995). Convergence of the iterative proportional fitting procedure. *Ann*. *Statist*. **23** 1160–1174. MR1353500

[45] RUSCHENDORF, L., SCHWEIZER, B. and TAYLOR, M. (1997). *Distributions with Fixed Marginals & Related Topics*. *Lecture Notes—Monograph Series* **28**. IMS, Hayward, CA. MR1485518

[46] SCHWEIZER, B. and WOLFF, E. F. (1981). On nonparametric measures of dependence for random variables. *Ann*. *Statist*. **9** 879–885. MR0619291

[47] SINKHORN, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann*. *Math*. *Stat*. **35** 876–879. MR0161868

[48] SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ*. *Inst*. *Statist*. *Univ*. *Paris* **8** 229–231. MR0125600

[49] STARR, S. (2009). Thermodynamic limit for the Mallows model on S_n. *J*. *Math*. *Phys*. **50** 095208, 15. MR2566888

[50] TRASHORRAS, J. (2008). Large deviations for symmetrised empirical measures. *J*. *Theoret*. *Probab*. **21** 397–412. MR2391251

[51] WHITT, W. (1976). Bivariate distributions with given marginals. *Ann*. *Statist*. **4** 1280–1289. MR0426099

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: sm3949@columbia.edu