Geometry of Rank Tests

Jason Morton, Lior Pachter, Anne Shiu, Bernd Sturmfels, and Oliver Wienand
Department of Mathematics, UC Berkeley

Abstract

We study partitions of the symmetric group which have desirable geometric properties. The statistical tests defined by such partitions involve counting all permutations in the equivalence classes. These permutations are the linear extensions of partially ordered sets specified by the data. Our methods refine rank tests of non-parametric statistics, such as the sign test and the runs test, and are useful for the exploratory analysis of ordinal data. *Convex rank tests* correspond to probabilistic conditional independence structures known as semi-graphoids. *Submodular rank tests* are classified by the faces of the cone of submodular functions, or by Minkowski summands of the permutohedron. We enumerate all small instances of such rank tests. *Graphical tests* correspond to both graphical models and to graph associahedra, and they have excellent statistical and algorithmic properties.

1 Introduction

The non-parametric approach to statistics was introduced by (Pitman, 1937). The emergence of microarray data in molecular biology has led to a number of new tests for identifying significant patterns in gene expression time series; see e.g. (Willbrand, 2005). This application motivated us to develop a mathematical theory of rank tests. We propose that a *rank test* is a partition of S_n induced by a map $\tau : S_n \to T$ from the symmetric group of all permutations of $[n] = \{1, \ldots, n\}$ onto a set T of statistics. The statistic $\tau(\pi)$ is the *signature* of the permutation $\pi \in S_n$. Each rank test defines a partition of S_n into classes, where π and π' are in the same class if and only if $\tau(\pi) = \tau(\pi')$. We identify $T = \text{image}(\tau)$ with the set of all classes in this partition of S_n. Assuming the uniform distribution on S_n, the probability of seeing a particular signature $t \in T$ is $1/n!$ times $|\tau^{-1}(t)|$. The computation of a p-value for a given permutation $\pi \in S_n$ typically amounts to summing

$$\Pr(\pi') \quad = \quad \frac{1}{n!} \cdot |\, \tau^{-1}\big(\tau(\pi')\big)\, | \qquad (1)$$

over all permutations π' with $\Pr(\pi') < \Pr(\pi)$. In Section 2 we explain how existing rank tests can be understood from our point of view.

In Section 3 we describe the class of *convex rank tests* which captures properties of tests used in practice. We work in the language of algebraic combinatorics (Stanley, 1997). Convex rank tests are in bijection with polyhedral fans that coarsen the hyperplane arrangement of S_n, and with conditional independence structures known as semi-graphoids (Studený, 2005).

Section 4 is devoted to convex rank tests that are induced by submodular functions. These *submodular rank tests* are in bijection with Minkowski summands of the $(n-1)$-dimensional permutohedron and with structural imset models. Furthermore, these tests are at a suitable level of generality for the biological applications that motivated us. We make the connections to polytopes and independence models concrete by classifying all convex rank tests for $n \leq 5$.

In Section 5 we discuss the class of *graphical tests.* In mathematics, these correspond to graph associahedra, and in statistics to graphical models. The equivalence of these two structures is shown in Theorem 18. The implementation of convex rank tests requires the efficient enumeration of linear extensions of partially ordered sets (posets). A key ingredient is a highly optimized method for computing distributive lattices. Our software is discussed in Section 6.

2 Rank tests and posets

A permutation π in S_n is a total order on $[n] = \{1, \dots, n\}$. This means that π is a set of $\binom{n}{2}$ ordered pairs of elements in $[n]$. If π and π' are permutations then $\pi \cap \pi'$ is a partial order.

In the applications we have in mind, the data are vectors $u \in \mathbb{R}^n$ with distinct coordinates. The permutation associated with u is the total order $\pi = \{ (i,j) \in [n] \times [n] \,:\, u_i < u_j \}$. We shall employ two other ways of writing this permutation. The first is the *rank vector* $\rho = (\rho_1, \dots, \rho_n)$, whose defining properties are $\{\rho_1, \dots, \rho_n\} = [n]$ and $\rho_i < \rho_j$ if and only if $u_i < u_j$. That is, the coordinate of the rank vector with value i is at the same position as the ith smallest coordinate of u. The second is the *descent vector* $\delta = (\delta_1, \dots, \delta_n)$, defined by $u_{\delta_i} > u_{\delta_{i+1}}$. The ith coordinate of the descent vector is the position of the ith largest value of u. For example, if $u = (11, 7, 13)$ then its permutation is represented by $\pi = \{(2,1), (1,3), (2,3)\}$, by $\rho = (2,1,3)$, or by $\delta = (3,1,2)$.

A permutation π is a *linear extension* of a partial order P on $[n]$ if $P \subseteq \pi$. We write $\mathcal{L}(P) \subseteq S_n$ for the set of linear extensions of P. A partition τ of the symmetric group S_n is a *pre-convex rank test* if the following axiom holds:

$$(PC) \qquad \begin{array}{c} \text{If } \tau(\pi) = \tau(\pi') \text{ and } \pi'' \in \mathcal{L}(\pi \cap \pi') \\ \text{then } \tau(\pi) = \tau(\pi') = \tau(\pi''). \end{array}$$

Note that $\pi'' \in \mathcal{L}(\pi \cap \pi')$ means $\pi \cap \pi' \subseteq \pi''$. For $n = 3$ the number of all rank tests is the Bell number $B_6 = 203$. Of these 203 set partitions of S_3, only 40 satisfy the axiom (PC).

Each class C of a pre-convex rank test τ corresponds to a poset P on $[n]$; namely, P is the intersection of all total orders in that class: $P = \bigcap_{\pi \in C} \pi$. The axiom (PC) ensures that C coincides with the set $\mathcal{L}(P)$ of all linear extensions of P. The inclusion $C \subseteq \mathcal{L}(P)$ is clear. For the reverse inclusion, note that from any permutation π in $\mathcal{L}(P)$, we can obtain any other π' in $\mathcal{L}(P)$ by a sequence of reversals $(a,b) \mapsto (b,a)$, where each intermediate $\hat{\pi}$ is also in $\mathcal{L}(P)$. Assume $\pi_0 \in \mathcal{L}(P)$ and $\pi_1 \in C$ differ by one reversal $(a,b) \in \pi_0$, $(b,a) \in \pi_1$. Then $(b,a) \notin P$, so there is some $\pi_2 \in C$ such that $(a,b) \in \pi_2$; thus, $\pi_0 \in \mathcal{L}(\pi_1 \cap \pi_2)$ by (PC). This shows $\pi_0 \in C$.

A pre-convex rank test is thus an unordered collection of posets $P_1, , \dots, P_k$ on $[n]$ that satisfies the property that S_n is the disjoint union of the subsets $\mathcal{L}(P_1), \dots, \mathcal{L}(P_k)$. The posets P_i that represent the classes in a pre-convex rank test capture the shapes of data vectors.

Example 1 (The sign test for paired data)**.** The *sign test* is performed on data that are paired as two vectors $u = (u_1, u_2, \dots, u_m)$ and $v = (v_1, v_2, \dots, v_m)$. The null hypothesis is that the median of the differences $u_i - v_i$ is 0. The test statistic is the number of differences that are positive. This test is a rank test, because u and v can be transformed into the overall ranks of the $n = 2m$ values, and the rank vector entries can then be compared. This test coarsens the convex rank test which is the MSS of Section 4 with $\mathcal{K} = \{\{1, m+1\}, \{2, m+2\}, \dots\}$.

Example 2 (Runs tests)**.** A *runs test* can be used when there is a natural ordering on the data points, such as in a time series. The data are transformed into a sequence of 'pluses' and 'minuses,' and the null hypothesis is that the number of observed runs is no more than that expected by chance. A runs test is a coarsening of the convex rank test τ described in (Willbrand, 2005, Section 6.1.1) and in Example 4.

These two examples suggest that many tests from classical statistics have a natural refinement by a pre-convex rank test. The term "pre-convex" refers to the following interpretation of the axiom (PC). Consider any two vectors u and u' in $\mathbb{R}^n$, and a convex combination $u'' = \lambda u + (1 - \lambda) u'$, with $0 < \lambda < 1$. If π, π', π'' are the permutations of u, u', u'' then $\pi'' \in \mathcal{L}(\pi \cap \pi')$. Thus the regions in $\mathbb{R}^n$ specified by a pre-convex rank test are convex cones.

3 Convex rank tests

A *fan* in $\mathbb{R}^n$ is a finite collection $\mathcal{F}$ of polyhedral cones which satisfies the following properties: (i) if $C \in \mathcal{F}$ and C' is a face of C, then $C' \in \mathcal{F}$, (ii) If $C, C' \in \mathcal{F}$, then $C \cap C'$ is a face of C. Two vectors u and v in $\mathbb{R}^n$ are *permutation equivalent* when $u_i < u_j$ if and only if

$v_i < v_j$, and $u_i = u_j$ if and only if $v_i = v_j$ for all $i, j \in [n]$. The permutation equivalence classes (of which there are 13 for $n = 3$) induce a fan which we call the *S_n-fan*. The maximal cones in the S_n-fan, which are the closures of the permutation equivalence classes corresponding to total orders, are indexed by permutations δ in S_n. A *coarsening* of the S_n-fan is a fan $\mathcal{F}$ such that every permutation equivalence class of $\mathbb{R}^n$ is fully contained in a cone C of $\mathcal{F}$; $\mathcal{F}$ defines a partition of S_n because each maximal cone of the S_n-fan is contained in some cone $C \in \mathcal{F}$. We define a *convex rank test* to be a partition of S_n defined by a coarsening of the S_n-fan. We identify the fan with that test.

Two maximal cones of the S_n-fan share a *wall* if there exists an index k such that $\delta_k = \delta'_{k+1}$, $\delta_{k+1} = \delta'_k$ and $\delta_i = \delta'_i$ for $i \neq k, k+1$. That is, the corresponding permutations δ and δ' differ by an adjacent transposition. To such an unordered pair $\{\delta, \delta'\}$, we associate the following conditional independence (CI) statement:

$$\delta_k \perp\!\!\!\perp \delta_{k+1} \,|\, \{\delta_1, \ldots, \delta_{k-1}\}. \qquad (2)$$

This formula defines a map from the set of walls of the S_n-fan onto the set of all CI statements

$$\mathcal{T}_n \;=\; \big\{\, i \perp\!\!\!\perp j \,|\, K \;:\; K \subseteq [n] \backslash \{i, j\} \big\}.$$

The map from walls to CI statements is not injective; there are $(n-k-1)!(k-1)!$ walls which are labelled by the statement (2).

Any convex rank test $\mathcal{F}$ is characterized by the collection of walls $\{\delta, \delta'\}$ that are removed when passing from the S_n-fan to $\mathcal{F}$. So, from (2), any convex rank test $\mathcal{F}$ maps to a set $\mathcal{M}_{\mathcal{F}}$ of CI statements corresponding to missing walls. Recall from (Matúš, 2004) and (Studený, 2005) that a subset $\mathcal{M}$ of $\mathcal{T}_n$ is a *semi-graphoid* if the following axiom holds:

$$i \perp\!\!\!\perp j \,|\, K \cup \ell \in \mathcal{M} \text{ and } i \perp\!\!\!\perp \ell \,|\, K \in \mathcal{M}$$
$$\text{implies } i \perp\!\!\!\perp j \,|\, K \in \mathcal{M} \text{ and } i \perp\!\!\!\perp \ell \,|\, K \cup j \in \mathcal{M}.$$

Theorem 3. *The map $\mathcal{F} \mapsto \mathcal{M}_{\mathcal{F}}$ is a bijection between convex rank tests and semi-graphoids.*

Example 4 (Up-down analysis for $n = 3$)**.** The test in (Willbrand, 2005) is a convex rank test and is visualized in Figure 1. Permutations are in the same class if they are connected by a solid edge; there are four classes. In the S_3-fan, the two missing walls are labeled by conditional independence statements as defined in (2).

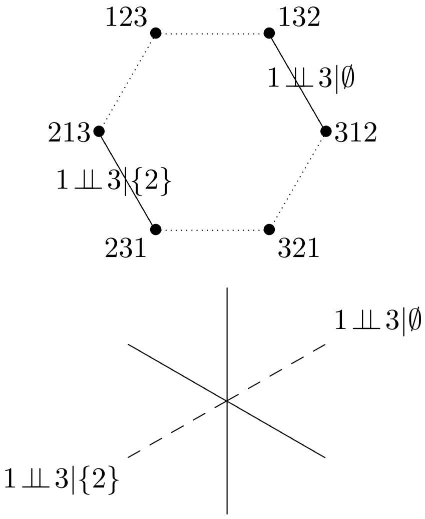
123
132
$1 \perp\!\!\!\perp 3|\emptyset$
213
312
$1 \perp\!\!\!\perp 3|\{2\}$
231
321
$1 \perp\!\!\!\perp 3|\emptyset$
$1 \perp\!\!\!\perp 3|\{2\}$

Figure 1: The permutohedron $\mathbf{P}_3$ and the S_3-fan projected to the plane. Each permutation is represented by its descent vector $\delta = \delta_1\delta_2\delta_3$.

Example 5 (Up-down analysis for $n = 4$)**.** The test $\mathcal{F}$ in (Willbrand, 2005) is shown in Figure 2. The double edges correspond to the 12 CI statements in $\mathcal{M}_{\mathcal{F}}$. There are 8 classes; e.g., the class $\{3412, 3142, 1342, 1324, 3124\}$ consists of the 5 permutations with up-down pattern $(-, +, -)$.

Our proof of Theorem 3 rests on translating the semi-graphoid axiom for a set of CI statements into geometric statements about the corresponding set of edges of the permutohedron.

The S_n-fan is the normal fan (Ziegler, 1995) of the *permutohedron* $\mathbf{P}_n$, which is the convex hull of the vectors $(\rho_1, \ldots, \rho_n) \in \mathbb{R}^n$, where ρ runs over all rank vectors of permutations in S_n. The edges of $\mathbf{P}_n$ correspond to walls and are thus labeled with CI statements. A collection of parallel edges of $\mathbf{P}_n$ perpendicular to a hyperplane $x_i = x_j$ corresponds to the set of CI statements $i \perp\!\!\!\perp j|K$, where K ranges over all subsets of $[n]\backslash\{i, j\}$. The two-dimensional faces of $\mathbf{P}_n$ are squares and regular hexagons, and two edges of $\mathbf{P}_n$ have the same label in $\mathcal{T}_n$

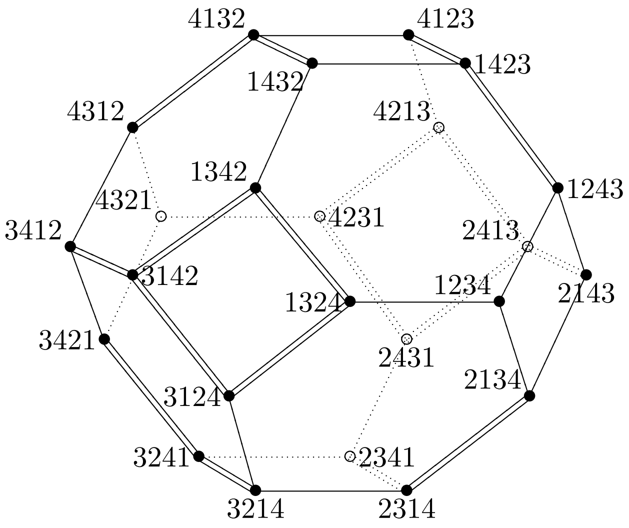

Figure 2: The permutohedron $\mathbf{P}_4$ with vertices marked by descent vectors δ. The test "up-down analysis" is indicated by the double edges.

if, but not only if, they are opposite edges of a square. A semi-graphoid $\mathcal{M}$ can be identified with the set $\mathbf{M}$ of edges with labels from $\mathcal{M}$. The semi-graphoid axiom translates into a geometric condition on the hexagonal faces of $\mathbf{P}_n$.

Observation 6. *A set* $\mathbf{M}$ *of edges of the permutohedron* $\mathbf{P}_n$ *is a semi-graphoid if and only if* $\mathbf{M}$ *satisfies the following two axioms:*
Square axiom: *Whenever an edge of a square is in* $\mathbf{M}$*, then the opposite edge is also in* $\mathbf{M}$.
Hexagon axiom: *Whenever two adjacent edges of a hexagon are in* $\mathbf{M}$*, then the two opposite edges of that hexagon are also in* $\mathbf{M}$.

Let $\mathbf{M}$ be the subgraph of the edge graph of $\mathbf{P}_n$ defined by the statements in $\mathcal{M}$. Then the classes of the rank test defined by $\mathcal{M}$ are given by the permutations in the path-connected components of $\mathbf{M}$. We regard a path from δ to δ' on $\mathbf{P}_n$ as a word $\sigma^{(1)} \cdots \sigma^{(l)}$ in the free associative algebra $\mathcal{A}$ generated by the adjacent transpositions of $[n]$. For example, the word $\sigma_{23} := (23)$ gives the path from δ to $\delta' = \sigma_{23}\delta = \delta_1\delta_3\delta_2\delta_4 \dots \delta_n$. The following relations in $\mathcal{A}$ define a presentation of the group algebra of S_n:

(BS) $\sigma_{i,i+1}\sigma_{i+k+1,i+k+2} - \sigma_{i+k+1,i+k+2}\sigma_{i,i+1}$,
(BH) $\sigma_{i,i+1}\sigma_{i+1,i+2}\sigma_{i,i+1} - \sigma_{i+1,i+2}\sigma_{i,i+1}\sigma_{i+1,i+2}$,
(BN) $\sigma_{i,i+1}^2 - 1$,

where suitable i and k vary over $[n]$. The first two are the *braid relations*, and the last represents the idempotency of each transposition.

Now, we regard these relations as properties of a set of edges of $\mathbf{P}_n$, by identifying a word and a permutation δ with the set of edges that comprise the corresponding path in $\mathbf{P}_n$. For example, a set satisfying (BS) is one such that, starting from any δ, the edges of the path $\sigma_{i,i+1}\sigma_{i+k+1,i+k+2}$ are in the set if and only if the edges of the path $\sigma_{i+k+1,i+k+2}\sigma_{i,i+1}$ are in the set. Note then, that (BS) is the square axiom, and (BH) is a weakening of the hexagon axiom of semi-graphoids. That is, implications in either direction hold in a semi-graphoid. However, (BN) holds only directionally in a semi-graphoid: if an edge lies in the semi-graphoid, then its two vertices are in the same class; but the empty path at some vertex δ certainly does not imply the presence of all incident edges in the semi-graphoid. Thus, for a semi-graphoid, we have (BS) and (BH), but must replace (BN) with the directional version

$$(BN') \qquad \sigma_{i,i+1}^2 \to 1.$$

Consider a path p from δ to δ' in a semi-graphoid. A result of (Tits, 1968) gives the following lemma; see also (Brown, 1989, p. 49-51).

Lemma 7. *If* $\mathcal{M}$ *is a semi-graphoid, then if* δ *and* δ' *lie in the same class of* $\mathcal{M}$*, then so do all shortest paths on* $\mathbf{P}_n$ *between them.*

We are now equipped to prove Theorem 3. Note that we have demonstrated that semi-graphoids and convex rank tests can be regarded as sets of edges of $\mathbf{P}_n$, so we will show that their axiom systems are equivalent. We first show that a semi-graphoid satisfies (PC). Consider δ, δ' in the same class C of a semi-graphoid, and let $\delta'' \in \mathcal{L}(\delta, \delta')$. Further, let p be a shortest path from δ to δ'' (so, $p\delta = \delta''$), and let q be a shortest path from δ'' to δ'. We claim that qp is a shortest path from δ to δ', and thus $\delta'' \in C$ by Lemma 7. Suppose qp is not a shortest path. Then, we can obtain a shorter path in the semi-graphoid by some sequence of substitutions according to (BS), (BH), and (BN'). Only (BN') decreases the length of a path, so the sequence must involve (BN'). Therefore, there is some i, j in $[n]$, such that their positions relative to each other are reversed twice in qp. But p and q are shortest paths, hence one reversal occurs in each p and q. Then δ and δ' agree on whether $i > j$ or $j > i$, but the reverse holds in δ'', contradicting $\delta'' \in \mathcal{L}(\delta, \delta')$. Thus every semi-graphoid is a pre-convex rank test.

Now, we show that a semi-graphoid corresponds to a fan. Consider the cone corresponding to a class C. We need only show that it meets any other cone in a shared face. Since C is a cone of a coarsening of the S_n-fan, each nonmaximal face of C lies in a hyperplane $H = \{x_i = x_j\}$. Suppose a face of C coincides with the hyperplane H and that $i > j$ in C. A vertex δ borders H if i and j are adjacent in δ. We will show that if $\delta, \delta' \in C$ border H, then their reflections $\hat{\delta} = \delta_1 \ldots ji \ldots \delta_n$ and $\hat{\delta}' = \delta'_1 \ldots ji \ldots \delta'_n$ both lie in some class C'. Consider a 'great circle' path between δ and δ' which stays closest to H: all vertices in the path have i and j separated by at most one position, and no two consecutive vertices have i and j nonadjacent. This is a shortest path, so it lies in C, by Lemma 7. Using the square and hexagon axioms (Observation 6), we see that the reflection of the path across H is a path in the semi-graphoid that connects $\hat{\delta}$ to $\hat{\delta}'$ (Figure 3). Thus a semigraphoid is a convex rank test.

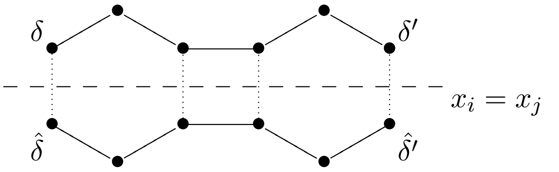

Figure 3: Reflecting a path across a hyperplane.

Finally, if $\mathbf{M}$ is a set of edges of $\mathbf{P}_n$, representing a convex rank test, then it is easy to show that $\mathbf{M}$ satisfies the square and hexagon axioms. This completes the proof of Theorem 3.

Remark 8. For $n = 3$ there are 40 pre-convex rank tests, but only 22 of them are convex rank tests. The corresponding CI models are shown in Figure 5.6 on page 108 in (Studený, 2005).

4 The submodular cone

In this section we examine a subclass of the convex rank tests. Let $2^{[n]}$ denote the collection of all subsets of $[n] = \{1, 2, \ldots, n\}$. Any real-valued function $w : 2^{[n]} \to \mathbb{R}$ defines a convex polytope Q_w of dimension $\leq n-1$ as follows:

$$\begin{aligned} Q_w \;\; := \; \big\{ \, x \in \mathbb{R}^n \, : \, x_1 + x_2 + \cdots + x_n = w([n]) \\ \text{and } \textstyle\sum_{i \in I} x_i \leq w(I) \text{ for all } \emptyset \neq I \subseteq [n] \, \big\}. \end{aligned}$$

A function $w : 2^{[n]} \to \mathbb{R}$ is called *submodular* if $w(I) + w(J) \geq w(I \cap J) + w(I \cup J)$ for $I, J \subseteq [n]$.

Proposition 9. *A function $w : 2^{[n]} \to \mathbb{R}$ is submodular if and only if the normal fan of the polyhedron Q_w is a coarsening of the S_n-fan.*

This follows from greedy maximization as in (Lovász, 1983). Note that the function w is submodular if and only if the optimal solution of

$$\text{maximize } u \cdot x \text{ subject to } x \in Q_w$$

depends only on the permutation equivalence class of u. Thus, solving this linear programming problem constitutes a convex rank test. Any such test is called a *submodular rank test.*

A convex polytope is a *(Minkowski) summand* of another polytope if the normal fan of the latter refines the normal fan of the former. The polytope Q_w that represents a submodular rank test is a summand of the permutohedron $\mathbf{P}_n$.

Theorem 10. *The following combinatorial objects are equivalent for any positive integer n:*
1. *submodular rank tests,*
2. *summands of the permutohedron $\mathbf{P}_n$,*
3. *structural conditional independence models,*
4. *faces of the submodular cone $\mathbf{C}_n$ in $\mathbb{R}^{2^n}$.*

We have $1 \Longleftrightarrow 2$ from Proposition 9, and $1 \Longleftrightarrow 3$ follows from (Studený, 2005). Further $3 \Longleftrightarrow 4$ holds by definition.

The *submodular cone* is the cone $\mathbf{C}_n$ of all submodular functions $w : 2^{[n]} \to \mathbb{R}$. Working modulo its lineality space $\mathbf{C}_n \cap (-\mathbf{C}_n)$, we regard $\mathbf{C}_n$ as a pointed cone of dimension $2^n - n - 1$.

Remark 11. All 22 convex rank tests for $n = 3$ are submodular. The submodular cone $\mathbf{C}_3$ is a 4-dimensional cone whose base is a bipyramid. The polytopes Q_w, as w ranges over the faces of $\mathbf{C}_3$, are all the Minkowski summands of $\mathbf{P}_3$.

Proposition 12. *For $n \geq 4$, there exist convex rank tests that are not submodular rank tests. Equivalently, there are fans that coarsen the S_n-fan but are not the normal fan of any polytope.*

This result is stated in Section 2.2.4 of (Studený, 2005) in the following form: "There exist semi-graphoids that are not structural."

We answered Question 4.5 posed in (Postnikov, 2006) by finding a non-submodular convex rank test in which all the posets P_i are trees:

$$\begin{aligned} \mathcal{M} \;\; = \;\; \{ & 2 \perp\!\!\!\perp 3 | \{1,4\}, \, 1 \perp\!\!\!\perp 4 | \{2,3\}, \\ & 1 \perp\!\!\!\perp 2 | \emptyset, \, 3 \perp\!\!\!\perp 4 | \emptyset \, \}. \end{aligned}$$

Remark 13. For $n = 4$ there are 22108 submodular rank tests, one for each face of the 11-dimensional cone $\mathbf{C}_4$. The base of this submodular cone is a polytope with f-vector $(1, 37, 356, 1596, 3985, 5980, 5560, 3212, 1128, 228, 24, 1)$.

Remark 14. For $n = 5$ there are 117978 coarsest submodular rank tests, in 1319 symmetry classes. We confirmed this result of (Studený, 2000) with `POLYMAKE` (Gawrilow, 2000).

We now define a class of submodular rank tests, which we call *Minkowski sum of simplices (MSS) tests*. Note that each subset K of $[n]$ defines a submodular function w_K by setting $w_K(I) = 1$ if $K \cap I$ is non-empty and $w_K(I) = 0$ if $K \cap I$ is empty. The corresponding polytope Q_{w_K} is the simplex $\Delta_K = \text{conv}\{e_k : k \in K\}$.

Now consider an arbitrary subset $\mathcal{K} = \{K_1, K_2, \ldots, K_r\}$ of $2^{[n]}$. It defines the submodular function $w_{\mathcal{K}} = w_{K_1} + w_{K_2} + \cdots + w_{K_r}$. The corresponding polytope is the Minkowski sum

$$\Delta_{\mathcal{K}} \quad = \quad \Delta_{K_1} + \Delta_{K_2} + \cdots + \Delta_{K_r}.$$

The associated MSS test $\tau_{\mathcal{K}}$ is defined as follows. Given $\rho \in S_n$, we compute the number of indices $j \in [r]$ such that $\max\{\rho_k : k \in K_j\} = \rho_i$, for each $i \in [n]$. The signature $\tau_{\mathcal{K}}(\rho)$ is the vector in $\mathbb{N}^n$ whose ith coordinate is that number. Few submodular rank tests are MSS tests:

Remark 15. For $n = 3$, among the 22 submodular rank tests, only 15 are MSS tests. For $n = 4$, among the 22108, only 1218 are MSS.

5 Graphical tests

Graphical models are fundamental in statistics, and they also lead to a useful class of rank tests. First we show how to associate a semi-graphoid to a family $\mathcal{K}$. Let $\mathcal{F}_{w_{\mathcal{K}}}$ be the normal fan of $Q_{w_{\mathcal{K}}}$. We write $\mathcal{M}_{\mathcal{K}}$ for the CI model derived from $\mathcal{F}_{w_{\mathcal{K}}}$ using the bijection in Theorem 3.

Proposition 16. *The semi-graphoid $\mathcal{M}_{\mathcal{K}}$ is the set of CI statements $i \perp\!\!\!\perp j \,|\, K$ which satisfy the following property: all sets containing $\{i,j\}$ and contained in $\{i,j\} \cup [n] \backslash K$ are not in $\mathcal{K}$.*

Let G be a graph with vertex set $[n]$. We define $\mathcal{K}(G)$ to be the collection of all subsets K of $[n]$ such that the induced subgraph of $G|_K$ is connected. Recall that the *undirected graphical model* (or *Markov random field*) derived from the graph G is the set $\mathcal{M}^G$ of CI statements:

$$\begin{aligned} \mathcal{M}^G \; = \; \big\{ \, & i \perp\!\!\!\perp j \,|\, C \; : \; \text{the restriction of } G \text{ to} \\ & [n] \backslash C \;\; \text{contains no path from } i \text{ to } j \big\}. \end{aligned}$$

The polytope $\Delta_G = \Delta_{\mathcal{K}(G)}$ is the *graph associahedron*, which is a well-studied object in combinatorics (Carr, 2004; Postnikov, 2005). The next theorem is derived from Proposition 16.

Theorem 17. *The CI model induced by the graph associahedron coincides with the graphical model $\mathcal{M}^G$, i.e., $\mathcal{M}_{\mathcal{K}(G)} = \mathcal{M}^G$.*

There is a natural involution $*$ on the set of all CI statements which is defined as follows:

$$(i \perp\!\!\!\perp j \,|\, C)^* \quad := \quad i \perp\!\!\!\perp j \,|\, [n] \backslash (C \cup \{i,j\}).$$

If $\mathcal{M}$ is any CI model, then the CI model $\mathcal{M}^*$ is obtained by applying the involution $*$ to all the CI statements in the model $\mathcal{M}$. Note that this involution was called duality in (Matúš, 1992).

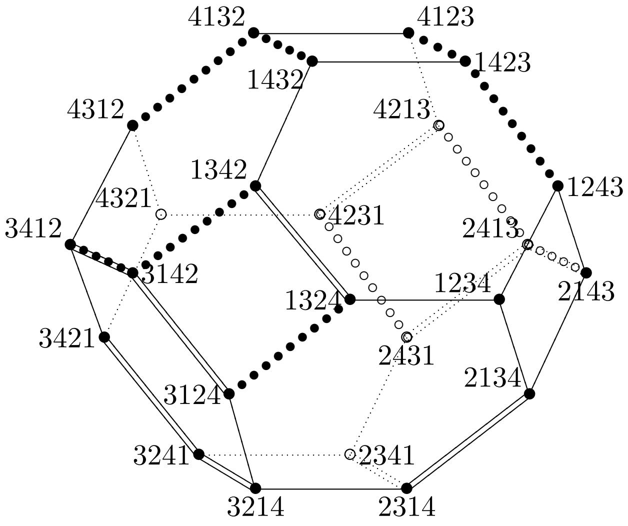

Figure 4: The permutohedron $\mathbf{P}_4$. Double edges indicate the test $\tau_{\mathcal{K}(G)}$ when G is the path. Edges with large dots indicate the test $\tau^*_{\mathcal{K}(G)}$.

The *graphical tubing rank test* $\tau^*_{\mathcal{K}(G)}$ is the test associated with $\mathcal{M}^*_{\mathcal{K}(G)}$. It can be obtained by a construction similar to the MSS test $\tau_{\mathcal{K}}$, with the function $w_{\mathcal{K}}$ defined differently and supermodular. The *graphical model rank test* $\tau_{\mathcal{K}(G)}$ is the MSS test of the set family $\mathcal{K}(G)$.

We next relate $\tau_{\mathcal{K}(G)}$ and $\tau^*_{\mathcal{K}(G)}$ to a known combinatorial characterization of the graph associahedron Δ_G. Two subsets $A, B \subset [n]$ are *compatible* for the graph G if one of the following conditions holds: $A \subset B$, $B \subset A$, or $A \cap B = \emptyset$, and there is no edge between any node in A and B. A *tubing* of the graph G is a subset $\mathbf{T}$ of $2^{[n]}$ such that any two elements of $\mathbf{T}$ are compatible. Carr and Devadoss (2005) showed that Δ_G is a simple polytope whose faces are in bijection with the tubings.

Theorem 18. *The following four combinatorial objects are isomorphic for any graph G on $[n]$:*

- *the graphical model rank test $\tau_{\mathcal{K}(G)}$,*
- *the graphical tubing rank test $\tau^*_{\mathcal{K}(G)}$,*
- *the fan of the graph associahedron Δ_G,*
- *the simplicial complex of all tubings on G.*

The maximal tubings of G correspond to vertices of the graph associahedron Δ_G. When G is the path of length n, then Δ_G is the *associahedron*, and when it is a cycle, Δ_G is the *cyclohedron.* The number of classes in the tubing test $\tau^*_{\mathcal{K}(G)}$ is the G-Catalan number of (Postnikov, 2005). This number is $\frac{1}{n+1}\binom{2n}{n}$ for the *associahedron test* and $\binom{2n-2}{n-1}$ for the *cyclohedron test.*

6 Enumerating linear extensions

In this paper we introduced a hierarchy of rank tests, ranging from pre-convex to graphical. Rank tests are applied to data vectors $u \in \mathbb{R}^n$, or permutations $\pi \in S_n$, and locate their cones. In order to determine the significance of a data vector, one needs to compute the quantity $|\,\tau^{-1}\big(\tau(\pi)\big)\,|$, and possibly the probabilities of other maximal cones. These cones are indexed by posets $P_1, P_2, \ldots, P_k$ on $[n]$, and the probability computations are equivalent to finding the cardinality of some of the sets $\mathcal{L}(P_i)$.

We now present our methods for computing linear extensions. If the rank test is a tubing test then this computation is done as follows. From the given permutation, we identify its signature (image under τ), which we may assume is its G-tree $\mathbf{T}$ (Postnikov, 2005). Suppose the root of the tree $\mathbf{T}$ has k children, each of which is a root of a subtree $\mathbf{T}^i$ for $i = 1, \ldots, k$. Writing $|\mathbf{T}^i|$ for the number of nodes in $\mathbf{T}^i$, we have

$$|\,\tau^{-1}(\mathbf{T})\,| = \binom{\sum_{i=1}^{k} |\mathbf{T}^i|}{|\mathbf{T}^1|, \ldots, |\mathbf{T}^k|} \left(\prod_{i=1}^{k} |\tau^{-1}(\mathbf{T}^i)| \right).$$

This recursive formula can be translated into an efficient iterative algorithm. In (Willbrand, 2005) the analogous problem is raised for the test in Example 3. A determinantal formula for (1) appears in (Stanley, 1997, page 69).

For an arbitrary convex rank test we proceed as follows. The test is specified (implicitly or explicitly) by a collection of posets $P_1, \ldots, P_k$ on $[n]$. From the given permutation, we first identify the unique poset P_i of which that permutation is a linear extension. We next construct the *distributive lattice* $L(P_i)$ of all order ideals of P_i. Recall that an *order ideal* is a subset O of $[n]$ such that if $l \in O$ and $(k, l) \in P_i$ then $k \in O$. The set of all order ideals is a lattice with meet and join operations given by set intersection $O \cap O'$ and set union $O \cup O'$. Knowledge of this distributive lattice $L(P_i)$ solves our problem because the linear extensions of P_i are precisely the maximal chains of $L(P_i)$. Computing the number of linear extensions is #P-complete (Brightwell, 1991). Therefore we developed efficient heuristics to build $L(P_i)$.

The key algorithmic task is the following: given a poset P_i on $[n]$, compute an efficient representation of the distributive lattice $L(P_i)$. Our program for performing rank tests works as follows. The input is a permutation π and a rank test τ. The test τ can be specified either

- by a list of posets $P_1, \ldots, P_k$ (pre-convex),
- or by a semigraphoid $\mathcal{M}$ (convex rank test),
- or by a submodular function $w : 2^{[n]} \to \mathbb{R}$,
- or by a collection $\mathcal{K}$ of subsets of $[n]$ (MSS),
- or by a graph G on $[n]$ (graphical test).

The output of our program has two parts. First, it gives the number $|\mathcal{L}(P_i)|$ of linear extensions, where the poset P_i represents the equivalence class of S_n specified by the data π. It also gives a representation of the distributive lattice $L(P_i)$, in a format that can be read by the **maple** package **posets** (Stembridge, 2004). Our software for the above rank tests is available at `www.bio.math.berkeley.edu/ranktests/`.

Acknowledgments

This paper originated in discussions with Olivier Pourquié and Mary-Lee Dequéant in the DARPA Fundamental Laws of Biology Program, which supported Jason Morton, Lior Pachter, and Bernd Sturmfels. Anne Shiu was supported by a Lucent Technologies Bell Labs Graduate Research Fellowship. Oliver Wienand was supported by the Wipprecht foundation.

References

G Brightwell and P Winkler. Counting linear extensions. *Order*, 8(3):225-242, 1991.

K Brown. *Buildings.* Springer, New York, 1989.

M Carr and S Devadoss. Coxeter complexes and graph associahedra. 2004. Available from `http://arxiv.org/abs/math.QA/0407229`.

E Gawrilow and M Joswig. Polymake: a framework for analyzing convex polytopes, in *Polytopes – Combinatorics and Computation*, eds. G Kalai and G M Ziegler, Birkhäuser, 2000, 43-74.

L Lovász. Submodular functions and convexity, in *Math Programming: The State of the Art*, eds. A Bachem, M Groetschel, and B Korte, Springer, 1983, 235-257.

F Matúš. Ascending and descending conditional independence relations, in *Proceedings of the Eleventh Prague Conference on Inform. Theory, Stat. Dec. Functions and Random Proc.*, Academia, B, 1992, 189-200.

F Matúš. Towards classification of semigraphoids. *Discrete Mathematics*, 277, 115-145, 2004.

EJG Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

A Postnikov. Permutohedra, associahedra, and beyond. 2005. Available from `http://arxiv.org/abs/math/0507163`.

A Postnikov, V Reiner, L Williams. Faces of Simple Generalized Permutohedra. Preprint, 2006.

RP Stanley. *Enumerative Combinatorics* Volume I, Cambridge University Press, Cambridge, 1997.

J Stembridge. Maple packages for symmetric functions, posets, root systems, and finite Coxeter groups. Available from `www.math.lsa.umich.edu/~jrs/maple.html`.

M Studený. *Probablistic conditional independence structures.* Springer Series in Information Science and Statistics, Springer-Verlag, London, 2005.

M Studený, RR Bouckaert, and T Kocka. Extreme supermodular set functions over five variables. *Institute of Information Theory and Automation*, Research report n. 1977, Prague, 2000.

J Tits. *Le problème des mots dans les groupes de Coxeter.* Symposia Math., 1:175-185, 1968.

K Willbrand, F Radvanyi, JP Nadal, JP Thiery, and T Fink. Identifying genes from up-down properties of microarray expression series. *Bioinformatics*, 21(20):3859–3864, 2005.

G Ziegler. *Lectures on polytopes.* Vol. 152 of Graduate Texts in Mathematics. Springer-Verlag, 1995.