Discrete Mathematics 108 (1992) 37–51
North-Holland

Geometric medians

Joseph Gil
Computer Science, The Hebrew University, Jerusalem, Israel

William Steiger*
Department of Computer Science, Rutgers University, New Brunswick, NJ 08903, USA

Avi Wigderson
Computer Science, The Hebrew University, Jerusalem, Israel

Received 4 January 1991

Abstract

Gil, J., W. Steiger and A. Wigderson, Geometric medians, Discrete Mathematics 108 (1992) 37–51.

We discuss several generalizations of the notion of *median* to points in R^d. They arise in Computational Geometry and in Statistics. These notions are compared with respect to some of their mathematical properties. We also consider computational aspects. The issue of computational complexity raises several intriguing questions.

1. Introduction and summary

Suppose we are given a set $S = \{a_1, \ldots, a_n\}$ of reals, define the *rank* of a_i by $\rho(a_i) \equiv |\{a_j : a_j \leq a_i\}|$ and its *depth* by $\delta(a_i) \equiv \min(\rho(a_i), n+1-\rho(a_i))$. The ranking problem is to find ρ for a given $a_i \in S$ and the selection problem is to find an $a_i \in S$ with a given rank k. Sorting may be regarded as *complete* ranking or as complete selection; once S has been sorted we know the rank of each element, as well as an element of each rank. Finally we recall that a *median* of S is an element of rank $\lfloor (n+1)/2 \rfloor$ and note that it has maximal depth. We write $m(S)$ for the median and δ^* for its depth, and note that

$$1 \leq \delta^* \leq \left\lfloor \frac{n+1}{2} \right\rfloor. \tag{1}$$

* The author expresses gratitude to the NSF DIMACS Center at Rutgers. Research supported in part by NSF grant CCR-8902522 and CCR-9111491.

0012-365X/92/$05.00 © 1992 — Elsevier Science Publishers B.V. All rights reserved

The left-hand side is attained when the a_i have at most two different values. The right-hand side is attained when all elements are distinct, and this must be considered the general situation. Clearly the depth function is invariant under linear transformations.

In studying the complexity of these tasks it is usual to consider the number of comparisons needed for the worst case input. In this comparison model it is familiar that ranking has complexity $n-1$, that selection has complexity $\Theta(n)$, and sorting $\Theta(n \log n)$. Thus, it is not necessary to know the ranks of all elements in order to assert that a certain one, say a_i, has a certain rank, say k. Interestingly, this fact was only established in 1973 [2]; previously, it had not been known whether sorting provided the fastest way to find, e.g., the median.

In this paper we consider analogues of these comparison tasks in the case where the inputs are points in R^d. The greatest interest will be focused on selection and especially on analogues of the median. Such problems arise naturally in multivariate statistical analysis and in many problems in computational geometry. Each of the notions we will consider is based upon a different generalization of the idea of depth of a point. From now on, $S = \{a_1, a_2, \ldots, a_n\}$ denotes n points in R^d. We consider

(1) *Peel depth.* Let $C'(S) \subseteq S$ denote the subset of points which are vertices of $C(S)$, the convex hull of S. Define the sequence

$$S = S_1 \supset S_2 \supset \cdots \supset S_{t+1} = \emptyset,\ S_t \neq \emptyset \tag{2}$$

by $S_{i+1} = S_i \backslash C'(S_i)$. Points $a_j \in C'(S_i)$ are the points of *peel depth* i and we write $\pi(a_j) = i$. The points in S_t have maximal depth and form the *peel median* of S. We write m_π for the peel median and $\pi^* = t$ for its depth.

(2) *Tukey depth.* Given $x \in R^d$, $\|x\| = 1$, the directional depth of a_i in the direction x is defined by $\delta_x(a_i) = \delta(x \cdot a_i)$; this is the usual depth applied to the orthogonal projection of S onto the line tx, $t \in R$. The *Tukey depth* of a point is then defined to be

$$\tau(a_i) = \min[\delta_x(a_i) : x \in R^d,\ \|x\| = 1], \tag{3}$$

the minimum of its directional depths. Again, a median m_τ is a point of maximal depth, say k, and we write $\tau^* = k$ for the depth of the median. This depth was proposed by Tukey at the International Congress of Mathematicians held in Vancouver [19]. It was rediscovered independently by computational geometers, for example see [6].

(3) *Simplicial depth.* Let F be a probability distribution on R^d and let

$$p(x) = \text{Prob}[\{x\} \subset C(z_1, z_2, \ldots, z_{d+1})],$$

where C denotes the convex hull of the $d+1$ points, chosen independently according to F. A point $m \in R^d$ is a *simplicial median* of F if $p(m) \geq p(x)$ for all $x \in R^d$. If $S = \{a_1, \ldots, a_n\}$ is a sample of n points from F, the sample estimate of m is the point $a_i \in S$ which is strictly contained in the largest number of $d+1$

simplices. Specifically the *simplicial depth* of a_i is

$$\sigma(a_i) = 1 + \sum I[\{a_i\} \subset C(a_{j_1}, a_{j_2}, \ldots, a_{j_{d+1}})]; \tag{4}$$

the sum is over all subsets of S of size $d+1$ and I is the indicator function. A median is a point m_σ in S of maximal depth. This depth will be denoted by σ^*. This notion was recently proposed by Liu [14]. We mention a cruder version that arose in the study of σ, namely the *box depth* defined by

$$\beta(a_i) = 1 + \sum I[\{a_i\} \subset \mathrm{Box}(a_{j_1}, a_{j_2})]; \tag{5}$$

the sum is over all distinct pairs of points in S and 'Box(u, v)' denotes the set of points in R^d whose coordinates are *between* the corresponding coordinates of u and v. The box median is a point m_β in S of maximal depth. This depth will be denoted by β^*.

Simple examples show that these depth measures are quite different. We will briefly compare them in the next section, where we also study some other mathematical properties, like invariance. We also consider the *breakdown point* [7], an interesting property of a computational procedure. Specifically, let T be a mapping from sets of points in R^d to a point in R^d and let $P = \{p_1, \ldots, p_s\} \subset R^d$ be a 'polluting' set. We say 'T breaks down at S for pollution of size s' if

$$\sup(\|T(S) - T(S \cup P)\|) = \infty, \tag{6}$$

the sup taken over all polluting sets P of size s. Let s' be the smallest amount of pollution for which T breaks down at S; i.e.,

$$s' = \min(s: \sup(\|T(S) - T(S \cup P)\|) = \infty),$$

the sup again over all P of size s. The breakdown point of T at S is the fraction

$$\epsilon(T, S) = \frac{s'}{n + s'}. \tag{7}$$

The poorest behaviour is when the breakdown point is $1/(n+1)$, for example when T computes the arithmetic mean of $S \subseteq R$; i.e.,

$$T(S) = \frac{1}{n} \sum_{i=1}^{n} a_i.$$

The addition of only a single polluting point can cause arbitrarily large changes in $T(S)$. In contrast, the usual median has breakdown point $\frac{1}{2}$. In the next section we will study the breakdown point for the different generalizations of median under consideration.

Section 3 is devoted to computational questions. We will use the uniform cost RAM as the model of computation. Each arithmetic operation and comparison will be assigned the same unit cost. With all the generalizations of the median there is the interesting question regarding lower bounds on the computational

complexity. It is not known whether it is necesssary to find the depth of every point in order to assert that a certain point has maximal depth.

2. Comparisons and properties

We first remark that all four notions given meaningful generalizations of the median in the sense that each collapses to the usual median when $d = 1$: The peel depth is the usual linear depth because the min and max comprise $C'(S_1)$. In the case of Tukey depth, $\tau(a_i) = \delta(a_i)$ because there is only one direction in R. When $d = 1$ a simplex is a pair a_{j_1}, a_{j_2}, so $\sigma(a_i)$ counts the number of such pairs containing a_i, namely $\sigma(a_i) = (j-1)(n-j)$ when $\delta(a_i) = j$. This shows that δ and σ order the points in exactly the same way.

For $d > 1$ the depth measures may give very different orderings. It is straightforward to construct examples in which some point u has a small peel depth but a large simplicial depth while another point v has a large peel depth but a small simplical depth. Similar constructions reverse the depth orderings of peel and Tukey depths. Here is a simple example of n points in R^2 which has points u, v with $\sigma(u)$ much less than $\sigma(v)$ while $\tau(u)$ is much greater than $\tau(v)$. Point $u = (0, 0)$ and $v = (1, 1)$. Choose n^t points on the line $x = 1$ with y-coordinates at least 2, choose n^t points on the line $y = 1$ with x-coordinates at least 2, and choose n^a points on the line $y = x$ with x-coordinates at most -1, $a, t < 1$. The remaining $O(n)$ points are placed in the unit square, half above $y = x$ and half below. Clearly $\sigma(u) = n^{2+a}$, $\sigma(v) = n^{1+2t}$, $\tau(u) = n^a$, and $\tau(v) = n^t$. Therefore if we take $t/2 = a < \frac{1}{3}$,

$$\frac{\sigma(u)}{\sigma(v)} = n^{1-3a} \uparrow \infty$$

while

$$\frac{\tau(u)}{\tau(v)} = n^{-a} \uparrow 0.$$

The ordinary depth measure δ is clearly invariant under any linear transformation of the input data. It would be desirable to retain this property for multidimensional generalizations. Because convexity and simplicial containment are preserved under linear transformations, it is clear that both the peel and simplicial depths are invariant: if A is a d by d matrix of full rank and $b \in R^d$, the points in $AS + b$ have the same depths under σ and π as those in S. It is also clear that for a given direction $x \in R^d$, the directional depths can be altered by linear transformations of the points. This makes it easy to construct examples where τ is not invariant. The same is true for β which, because it depends on the coordinate system, is not invariant.

2.1. Medians' depths

Now, analogous to (1), we consider the range of variation of the depth of each of the medians. If the points of S are in convex position, each point will have depth one, in each of the depth measures except the box depth. On the other hand if S is $\lfloor n/(d+1) \rfloor$ nested simplices a median will have depth $\lceil n/(d+1) \rceil$ and so the inequality

$$1 \leq \pi^* \leq \left\lceil \frac{n}{d+1} \right\rceil \tag{8}$$

is sharp.

It is clear that

$$1 \leq \tau^* \leq \left\lceil \frac{n+1}{2} \right\rceil. \tag{9}$$

This is sharp in R^2. Just take $(0, 0)$ and $2k+1$ points evenly spaced on the unit circle and note that the origin has depth $k+1$. In general we can place $2k+d-1$ points on the unit sphere in R^d in such a way that every hyperplane containing the origin has at least k points in each open halfspace (Gale's theorem [10]). Again the origin has Tukey depth $k+1 = [n-(d-2)]/2$. It is also interesting to note that there is always a point x, not necessarily in S, which, if added to S would have $\tau(x) = O(n)$. Helly's theorem implies the existence of a *centerpoint* for S. This is a point x such that every hyperplane containing it, has at least $n/(d+1)$ points of S on each side (see, e.g. [8]) so $\tau(x) = \lfloor (n+1)/(d+1) \rfloor$.

Obviously σ^* cannot exceed the number of distinct $d+1$-simplices in R^d. Boros and Füredi (and others, see e.g. [3]) showed in fact that

$$\sigma^* \leq \frac{1}{2^d} \binom{n}{d+1} + O(n^d)$$

and when $d=2$ the constant $\frac{1}{4}$ is best possible. For the planar case they also established the existence of a point x covered by $\frac{2}{9}$ of the triangles formed by the points of S and again the constant $\frac{2}{9}$ is best possible. Finally, a theorem of Bárány [1] generalizes the latter result by showing the existence of a point covered by a constant fraction of all $d+1$-simplices, namely

$$\sigma(x) \geq \frac{1}{(d+1)^{d+1}} \binom{n}{d+1}.$$

In all dimensions $\beta^* \leq n^2/2$, the number of boxes defined by the points of S. A distinctive property of the box median is that it always has quadratic depth.

Lemma 1. *There is a positive constant $c(d) \leq \frac{1}{2}$ such that for every set $S \subset R^d$ with n points, $\beta^* \geq c(d)n^2$.*

Proof. First we give a simple argument for $d = 2$, assuming the points are in general position. There are horizontal lines h_1, h_2 that separate the plane into three strips with at least $k = \lfloor n/3 \rfloor$ points of S in each. There are also two vertical lines v_1, v_2 with the same property and now we have nine regions R_{ij} where in each row *and* column there are at least k points of S. For each i, at least one R_{ij} must have a maximal number ($\geq n/9$) points of S and these cannot all occur in the same column, or that strip would have more than k points (the other possibility is that the maximal R_{ij} are not unique but in this case they may be taken in at least two columns). If the maximal R_{ij} line up along a diagonal we are finished. Otherwise repeat the same decomposition for the three maximal R_{ij}. It is easy to see that least three subregions, each of size at least $n/81$, are ordered up-right or up-left. This proves that at least $n/81$ points $a_i \in S$ are each in at least $(n/81)^2$ boxes so $c(2) \geq 1/3^8$. [Noga Alon (pers. com.) can show that $c(2) \leq \frac{1}{4}$; clearly there is a set where β^*/n^2 is about $\frac{1}{4}$].

Given a diagonal $\vec{e} = (1, e_2, \ldots, e_d)$, $e_i = \pm 1$, of the cube $K_d = \{(x_1, \ldots, x_n): |x_i| \leq 1\}$, $x, y \in R^d$ are *ordered along* $\vec{e}$ if $x - y$ has the same sign pattern as $\vec{e}$. We just proved the $d = 2$ case of the following statement: there is a constant $a(d) > 1$ and disjoint subsets $A, B, C \subset S \subset R^d$, each of size at least $n/a(d)$, so that for all triples $x \in A$, $y \in B$, $z \in C$, x, y and, y, z are ordered along one of the 2^{d-1} diagonals of the unit cube. To advance the induction from $d = t$ to $d = t + 1$, consider the first t coordinates of each point in $S \subset R^{t+1}$. We have a diagonal $\vec{e} = (1, e_2, \ldots, e_t)$ of K_t and subsets A, B, C of size at least $n/a(t)$, such that if $x \in A$, $y \in B$, $z \in C$, the first t coordinates of $x - y$ and $y - z$ have the same sign pattern as $\vec{e}$. Now apply the previous two dimensional argument to the points in A, B, C projected orthogonally onto the plane spanned by $\vec{e}$ and the $t + 1$st coordinate vector. This gives subsets A', B', C', of $A \cup B \cup C$ of size at least $n/(27a(d))$ whose elements are ordered like $\vec{e}' = (\vec{e}, e_{t+1}) \in R^{t+1}$. Finally we note that $a(d) \geq 3^{3d-2}$ and $c(d) \geq a^{-2}(d)$. □

If the points in S were generated independently, each according to the distribution F on R^d, the depth of the median is then a random variable and it is interesting to consider its expected value. Unfortunately very little is known. In the case of the peel median we need to know the expected number of peels. Although the expected size of $|C'(S_1)|$ has been studied in some detail ([17, 18]) it is not clear how to utilize this information because the successive peels are highly dependent. For example if F is the uniform distribution on the ball in R^d then the expected number of hull vertices is $O(n^{(d-1)/(d+1)})$ (see [17]). If the S_i in (1), $i \geq 2$ were also uniformly distributed in a ball, this observation could be repeated and would imply that $E(\pi^*) = O(n^{2/(d+1)} \log n)$. On the other hand it is not even known whether $E(\pi^*) = o(n)$ or if it is bounded. The situation may be simpler in the case of the other two medians. If F is uniform on the ball in R^d, $E(\tau^*) = n/2 + o(n)$ and $E(\sigma^*) = \Theta(n^{d+1})$.

2.2. Breakdown points

We conclude this section by examining the breakdown point of the various medians. It is easy to break down the peel median when π^* is small. For example let S consist of $A=(0, 1)$, $B=(0, -1)$, $C=(1, 0)$, $O=(0, 0)$, and $n-4$ other points with negative x-coordinates, on the circle $x^2+y^2=1$. Clearly the origin is the median and $\pi^*=2$. Now add polluting points $D=(2+3t, 1)$ and $E=(2+2t, \frac{2}{3})$, $t>1$. This has points O, C, E with depth 2 and all others with depth 1. Finally add polluting point F in triangle $\triangle OCE$ and with x-coordinate $2+t$. This point is the new peel median. We have caused breakdown because, as in (6), $\|O-F\|\to\infty$ as $t\to\infty$ and $\epsilon(m_\pi, S)\leq 3/(n+3)$. A similar construction in R^d gives breakdown with $d+1$ polluting points. The peel median does not break down as easily when π^* is large. One can argue that $s'\geq\pi^*$ is necessary for breakdown and this implies that the breakdown point is at least $\pi^*/(2n)$. Even so, the median may be quite deep, say $\pi^*=n/\log n$ and still have an asymptotically zero breakdown point, in contrast with the usual median. The only way to avoid zero breakdown is when the peel median has linear depth. In view of the previous paragraph, this may be a most unlikely occurrence.

The situation with the Tukey median is similar. In the preceding example if we pollute with points $D=(-t, 0)$, $E=(-2t, 0)$, and $F=(-3t, 0)$ then D will have Tukey depth 3 so it must be the new median. Breakdown occurs when we let $t\to\infty$ so $\epsilon(m_\tau, S)\leq 3/(n+3)$. As before, $s'\geq\tau^*$ polluting points are necessary to cause breakdown. Therefore $\epsilon(m_\tau\cdot)\geq\tau^*/2n$. We should expect the Tukey median to be hard to break down. In a variety of random settings τ^* will be linear.

It would seem that the box median is hard to break down, since it always has quadratic depth. The argument after Lemma 1 implies that $s'>n/81$ in the plane and the breakdown point must be at least $\frac{1}{82}$.

Finally, let us consider the simplicial median in R^2. Suppose S consists of n points on the unit circle and $\arg(a_i)=\pi i/(4n)$, $i=1, \ldots, n$. Choose a point x on the line from the origin $O=(0, 0)$ to point a_2 which is also in triangles $\triangle a_1a_2a_3$, $\triangle a_1a_2a_4$, $\ldots$, $\triangle a_1a_2a_n$. x is the simplicial median and has depth $\sigma^*=n-2$. Consider the point $C=(2t, \pi+2\pi/(4n))$ (in polar coordinates). We pollute with points A and B in the triangle $\triangle Ca_2a_3$, both in the third quadrant, and A a distance $t/2$ from the origin, B a distance t. A creates one new triangle ($\triangle a_1a_2A$) containing x and so does B ($\triangle a_1a_2B$), so its depth is now n. However A is contained in $3(n-2)$ triangles and is therefore the new median. Breakdown occurs when $t\to\infty$ and $\epsilon(m_\sigma, S)\leq 2/(n+2)$. The simplicial median can be broken down with $o(n)$ polluting points even when it has quadratic depth.

3. Computational issues

There are some interesting aspects regarding the complexity of computing the four medians. We begin by mentioning previous work that relates to the peel,

Tukey, and box medians. Then we discuss the simplicial median in two and three dimensions (there is no fast algorithm for $d > 3$).

For $d = 2$ the computational issues related to the peel median are well understood. If $|C'(S_1)| = k$ the outer peel can be computed in $O(n \log k)$ time and this is optimal [13]. In addition Chazelle [5] has shown how to compute the entire sequence of peels in (2) in $O(n \log n)$ time which, in view of the foregoing result, is optimal. Since $\max[\pi(a_i)]$ may now be found in $\Theta(n)$ steps, $\Theta(n \log n)$ is the time complexity of the peel median if the depth of each point is to be computed (this is in fact required if the points are in convex position). On the other hand, if it were known that the points were not in convex position (the expected situation) a more efficient algorithm for the median may be possible. A clean question is: given $S \subset R^2$ with n points and $m_\pi(S) = k > 1$, what is the complexity of finding a point in S_k?

For $d = 3$ the $O(n \log n)$ algorithm of Preparata and Hong [15] computes $C'(S_1)$ optimally, though it is not sensitive to the size of the output. An exercise in [8] describes an $O(n^{3/2} \log n)$ algorithm to compute all the peels, but this must be far from optimal, even when there are $O(n)$ peels. Again, if the points were in convex position $\Theta(n \log n)$ is the cost of the peel median.

Finally, Raimund Seidel (see [8]) has devised an algorithm for $C'(S_1)$ that runs in time $O(n^{\lfloor (d+1)/2 \rfloor})$ and gives the whole combinatorial structure of the hull. It may be used to compute all peels in time $O(n^{\lfloor (d+3)/2 \rfloor})$, since there are at most $n/(d+1)$ peels. On the other hand we can compute $\pi(a_i)$ for each point using the linear-time linear programming algorithm and assuming d is fixed. This gives the current peel $C'(S_j)$ in quadratic time, and all depth in $O(n^3)$. There still remains the nice lower bound question for the peel median. Does there exist an algorithm that can compute m_π faster than $O(n)$ plus the time for an optimal algorithm to compute $\pi(a_i)$ for each point?

The same question pertians to the Tukey median. The brute-force algorithm would compute $\delta_x(a_i)$ for each x normal to a hyperplane containing d points of S. In this way we get each $\tau(a_i)$ in time $O(n^{d+1})$ and τ^* in $O(n^{d+2})$. Cole, Sharir, and Yap [6] outline an $O(n^d)$ algorithm to compute all of the $\tau(a_i)$, and now it is easy to compute the median and its depth in $O(n)$ additional steps.

The brute force algorithm for the box median would compute each $\beta(a_i)$ in time $O(n^2 d)$, so m_β may be obtained in time $O(n^3 d) + O(n)$. A better procedure uses a simple inductive algorithm, based on successive reduction of the dimension, a $\log n$ factor needed for each reduction. In this way we can get the box median in $O(n(\log n)^{d-1})$ time. For $d = 2$ this gives the optimal complexity to obtain the box depth of every point, by reduction to sorting. On the other hand, it may be possible to find the box median without computing all depths. We observe that the above algorithm has the same cost as a familiar one for the dominating pairs problem (see, e.g. [16]), to which the box median reduces.

3.1. Computing simplicial medians ($d = 2$)

A brute force algorithm for the simplicial median could check each possible simplex containment, for each point, in $O(n^{d+2})$. In the remainder of this section we discuss the complexity when $d \leq 3$.

All algorithms must be evaluated in comparison to the following result.

Lemma 2. *The cost of computing σ^* is $\Omega(n \log n)$.*

The argument is via reduction to element distinctness. Given $a_1, \ldots, a_n$ map a_i to the point $(a_i, a_i^2 + \epsilon^i)$. The n images will be in convex position if and only if the a_i are distinct, so $\sigma^* = 0$ is equivalent to distinctness. Still, it may be possible to compute the median in less time, although every algorithm that computes all $\sigma(a_i)$ must obey the lower bound of the lemma.

First, we give an $O(n^2)$ time algorithm to compute the simplicial median in the plane. It computes the depth of each point and then finds the maximum. The following two observations are basic to the algorithm.

Lemma 3. *Given points A, B, C and a reference point x, let A' be any point on the ray from x through A. Then $x \in \triangle ABC$ if and only if $x \in \triangle A'BC$.*

Lemma 4. *Given points A, B, C on the unit circle $\mathscr{C}$ centered at the origin, let A^* be antipodal to A. Then $\triangle ABC$ contains the origin if and only if A^* is on the short arc joining B and C.*

Let $a'_q = a_q - a_i$ have polar representation (r_q, θ_q). Lemma 3 says that $\sigma(a_i)$ may be computed by counting the number of triangles $\triangle \theta_j \theta_k \theta_m$ on the unit circle that contain the origin. Lemma 4 says we can do this by counting for each pair θ_j, θ_k the number of antipodal points θ_m^* that fall in the short arc between them, and summing over all such pairs. We abuse notation by saying θ_A when we mean the point A on $\mathscr{C}$ with polar angle θ_A. Here is a summary of an algorithm to count, for n points on $\mathscr{C}$, the number of triangles containing the center.

algorithm *Count_Triangles*$(\theta_j; n)$

1. Sort θ_j's anti-clockwise on $\mathscr{C}$.
   (a) For each θ_j, compute n_j, the number of θ_m^* in $[\theta_j, \theta_{j+1}]$, and $N_j = n_1 + \cdots + n_j$.
2. Pick the diameter D through θ_1 and divide $\mathscr{C}$ with it into upper half $(\theta_1, \ldots, \theta_t)$ and lower half $(\theta_{t+1}, \ldots, \theta_n)$ vertices.
3. Count all triangles with base in the upper half and having left endpoint θ_1.

4. **repeat**
   (a) Move D anti-clockwise to next θ_j and update upper half set to $(\theta_j, \ldots, \theta_{t+m})$ and lower half set to $(\theta_{t+m+1}, \ldots, \theta_{j-1})$.
   (b) Add to count the number of triangles with base in the new upper half and left endpoint θ_j.

   until $j = n$.
5. **return** the count divided by 3.

end *Count_Triangles*

Clearly Step 1 can be done in $O(n \log n)$ time; the sorting information allows all θ_m^* to be placed in the correct interval $[\theta_j, \theta_{j+1}]$ in linear time. Step 2 is linear.

We argue that Step 3 may be done in $O(n)$ time and thereafter, *all* the updates of Step 4 may also be done in linear time. By Lemma 4, n_1 is the number of triangles containing the origin and having base $\overline{\theta_1\theta_2}$. Similarly $n_1 + n_2$ is the number with base $\overline{\theta_1\theta_3}$, etc. The quantity evaluated in Step 3 is thus

$$T_1 = \sum_{i=1}^{t-1} (t-i)n_i. \tag{10}$$

It can be computed in $O(n)$ time.

When D is rotated to θ_2 suppose m new points $\theta_{t+1}, \ldots, \theta_{t+m}$ come into the upper half. The quantity computed in Step 4(b) is

$$T_2 = \sum_{i=2}^{t+m-1} (t+m-i)n_i. \tag{11}$$

We can compute it in time $O(m)$ by updating T_1. Subtract T_1 from T_2 to see

$$T_2 = T_1 + m(n_2 + \cdots + n_t) + [(m-1)n_{t+1} + \cdots + n_{t+m}] - (t-1)n_1.$$

The expression in parentheses is $N_t - N_1$ and takes $O(1)$ steps. The expression in square brackets requires $O(m)$ steps, but each n_j can only come into one such sum so that during the course of all the $n-1$ updates, the total cost of these steps is $O(n)$. This argument proves the following.

Lemma 5. *The number of triangles containing a_i may be counted in time $O(n)$, once the $\theta_j = \arg(a_i - a_j)$ have been sorted.*

Finally (see [8]) we can obtain the radial order of the other $n-1$ points about a_i, for *all* the $a_i \in S$, in $O(n^2)$ time using duality. We map $a_i = (x_i, y_i)$ to the line $v = x_i u + y_i$ with slope x_i and intercept y_i and we map a line with equation $y = mx + b$ to the point $(-m, b)$. In the dual of S we have the set $\mathcal{L}$ of n lines which decompose the plane into *cells* bounded by *edges* which intersect in *vertices*. This dissection is the *arrangement* $\mathcal{A}(\mathcal{L})$ of the lines and may be represented by the *incidence graph* $\mathcal{I}(\mathcal{L})$. Edelsbrunner, O'Rourke, and Seidel [9] show how to construct $\mathcal{I}(\mathcal{L})$ in $O(n^2)$ time. Once constructed, we can traverse

part of the graph in linear time to obtain the vertices—in order of increasing x-coordinate—formed by the intersections of line i and the other $n-1$ lines. Transforming this information back to the primal gives the radial order of the other points about a_i. Therefore we have the following theorem.

Theorem 1 (Gill, Steiger, Wigderson [11]). *Given $S \subset R^2$ with n points, $m_\sigma(S)$ may be computed in* $O(n^2)$ steps.

This result was established independently by Khuller and Mitchell [12].

Lemma 2 also gives a lower bound on the complexity of simplicial medians if the depth of every point is computed. We conjecture that Theorem 1 gives the best possible upper bound if all simplicial depths are computed. The difference between the upper and lower bounds for simplicial medians is intriguing. These bounds match those for the general position question: given n point in R^2, is it true that no three are colinear?

3.2. *Computing simplicial medians* ($d = 3$)

The three-dimensional generalization is interesting. Here are some of the basic ideas. In dimension $d = 3$, we want to count $\sigma(a_i)$, the number of tetrahedra that contain a_i. Take the unit sphere $\mathcal{B}(a_i)$ centred at a_i, and write θ_j for the intersection point of $\mathcal{B}(a_i)$ and the ray from a_i through a_{i_j}, $i_j \neq i$. The obvious analogue of Lemma 3 shows that we need only count tetrahedra $\triangle' \theta_j \theta_k \theta_m \theta_r$ which contain the center. The analogue of Lemma 4 says that we may do this by counting how many spherical trianges $\triangle_s \theta_j \theta_k \theta_m$ (the sides are short arcs on great circles) contain how many antipidal points θ_r^*. Each such triangular containment is a 'good' tetrahedron. These triangular containments are counted via an algorithm that generalizes the foregoing one, in which triangle containments from points in a hemisphere are counted and then the plane defining that hemisphere is advanced. Here is a brief description of the counting of tetrahedra containing the origin O given n points θ_j on the unit sphere $\mathcal{B}(O)$.

algorithm *Count_Tetrahedra*$(\theta_j; n)$

1. Pick a point $x \in \mathcal{B}$, $x \neq \theta_j$, $j = 1, \ldots, n$ and define the plane Π_1 through O, θ_1, and x.
2. Renumber the θ_j, $j > 1$ by rotation of Π_1 about $\overrightarrow{Ox}$.
   Upper hemisphere points are $\mathcal{U}_1 = (\theta_1, \ldots, \theta_t; \theta_{t+1}^*, \ldots, \theta_n^*)$;
3. Centrally project $\mathcal{U}_1$ up onto a plane Λ_1 parallel to Π_1.
   Compute the arrangement for the dual of $\mathcal{U}_1$ in Λ_1.
4. Count $\sigma(\theta_r^*)$ in the projection, for each $\theta_r^* \in \mathcal{U}_1$ and save as SUM.

5. **repeat**
   (a) Rotate Π_j about $\overrightarrow{Ox}$ from θ_j to θ_{j+1} and update $\mathcal{U}_j$ to $\mathcal{U}_{j+1} = (\theta_{j+1}, \ldots, \theta_{t+m}; \theta^*_{t+m+1}, \ldots, \theta^*_j)$.
   (b) Radially project onto Λ_{j+1}, update arrangement, and update SUM.
   until $j = n$.
6. **return** SUM divided by 2.

end *Count_Tetrahedra*

Clearly Step 2 may be done in time $O(n \log n)$ by projecting the θ_j and O orthogonally onto a plane with normal vector $\overrightarrow{Ox}$ and then clockwise sorting the images about the image of O. If we have chosen x correctly the images of the θ_j and of O will be $n+1$ distinct points in general position. A random choice will certainly be good, or we could construct one in quadratic time.

The central projection from O in Step 3 preserves triangle containment: a great circle through $\overline{\theta_j\theta_k}$ projects to a straight line on Λ_1; spherical triangle $\triangle_s\theta_j\theta_k\theta_m$ containing θ^*_r on $\mathcal{B}(O)$ projects to a triangle in Λ_1 containing the image of θ^*_r. The arrangement of the points projected into Λ_1 may be computed in time $O(n^2)$.

The count in Step 4 is based on the previous algorithm. For each $\theta^*_r \in \mathcal{U}_1$, $\sigma(\theta^*_r)$ counts the number of triangles $\triangle\theta_j\theta_k\theta_m$ from $\mathcal{U}_1$ that contain it. By Theorem 1 the quantity

$$\text{SUM} = \sum_{\theta^*_r \in \mathcal{U}} \sigma(\theta^*_r) \tag{12}$$

may be obtained in $O(n^2)$ time.

Step 5 is less straightforward. Rotate Π_1 from θ_1 to θ_2 and then centrally project the points in $\mathcal{U}_2$ onto Λ_2. We need to count triangle containments that were not present in Λ_1. There are two new features; θ^*_1 and $\theta_{t+1}, \ldots, \theta_{t+m}$ have entered and θ_1 and $\theta^*_{t+1}, \ldots, \theta^*_{t+m}$ have left. For the leaving θ^*_j, there is nothing to do. But to efficiently account for the other changes, we need to use the dual arrangement of the points that are projected into Λ_2. The naive approach would compute this arrangement from scratch in $O(n^2)$ time. We can get it in amortized linear time, using the following observation.

Lemma 6. *Suppose θ^*_r, and $\theta_{j_1}, \ldots, \theta_{j_q}$ are in successive upper hemispheres $\mathcal{U}_m$, $\mathcal{U}_{m+1}$. The rotational order of the images of $\theta_{j_1}, \ldots, \theta_{j_q}$ about the image of θ^*_r is the same in Λ_m and Λ_{m+1}.*

The proof is straightforward because the great circle through θ^*_r and θ_i projects to a straight line in Λ_m and in Λ_{m+1} and these lines are both in the plane defined by the origin, θ^*_r, and θ_i. The meaning of Lemma 6 is that although lines corresponding to $\theta^*_r, \theta_{j_1}, \ldots, \theta_{j_q}$ may all change their positions as Λ_m is rotated to Λ_{m+1}, their combinatorial structure remains fixed. Therefore the arrangement of lines dual to the points projected into Λ_2 may be obtained by simply adding θ^*_1

and $\theta^*_{t+1}, \ldots, \theta_{t+m}$ and deleting θ_1 and $\theta^*_{t+1}, \ldots, \theta^*_{t+m}$ from the arrangement for Λ_1. The cost in Step 5(b) is $O(n)$ for each line added to, or deleted from, the arrangement. Since each point leaves and enters once, these updates to the line arrangements use a total of $O(n^2)$ time. Now that the arrangement describes the current points in Λ_2, we can use the previous algorithm to compute $\sigma(\theta^*_1)$ in linear time and add it to SUM.

To complete the update of SUM in Step 5(b), we need only count the new triangle containments of points $\theta^*_j \in \mathcal{U}_2$, $j > 1$, caused by the new points $\theta_k \in \mathcal{U}_2$, and add them to SUM. The complexity is $O(n^2)$ because of the following.

Lemma 7. *Suppose θ_k is a given new point in Λ_{m+1}. The number of new triangle containments $\theta^*_r \in \triangle\theta_i\theta_j\theta_k$, θ_i, θ_j in Λ_{m+1}, may be counted in linear time.*

Proof. As in Lemma 4 we consider points in Λ_{m+1} projected onto the unit circle $\mathcal{C}(\theta^*_r)$ centred at θ^*_r. Let θ'_k denote the point on $\mathcal{C}(\theta^*_r)$ which is antipodal to the given point, θ_k. By Lemma 4 we need to count the number of pairs θ_i, θ_j which have θ'_k on the short arc between them. Now, using duality in Λ_{m+1}, let ℓ be the new line (dual to θ'_k) that we are accounting for. Let $c_1, \ldots, c_p$ denote the duals of the θ_j, and $c^*_1, \ldots, c^*_q$ the duals of the θ^*_r. We must count the number of times lines c^*_r intersect triangles bounded by ℓ, c_i, and c_j.

Consider a particular c^*_r, and suppose the rank of $c^*_r \cap \ell$ is kth among the $p+1$ x-coordinates of the intersections $c_i \cap c^*_r$. Then c^*_r intersects $k(p-k)$ triangles bounded by ℓ, c_i, and c_j. If we add this quantity to SUM for each of the c^*_r, we will have counted all the new triangle containments involving the new line ℓ. The time taken by these updates is also $O(n)$ because the rank of $\ell \cap c^*_r$ may be obtained from the incidence graph in constant time. □

Each tetrahedron containing O has been counted exactly twice. The line $\overrightarrow{Ox}$ about which the planes are rotated is an axis of $\mathcal{B}(O)$ and meets exactly two faces of every tetrahedron containing the origin. Each of the other two faces lies in at least one of our upper hemispheres, and will be counted exactly once as a triangle containing the fourth, antipodal point. This explains Step 6 and concludes the proof of Theorem 2.

Theorem 2. *Given n points in R^3, the simplicial depth of any point may be counted in $O(n^2)$ time and $m_\sigma(S)$ may be found in $O(n^3)$ time.*

There doesn't seem to be any fundamental obstacle to generalizing this approach to higher dimensions, but we have not really considered the details.

4. Concluding remarks

In this paper we have considered analogues of ranking, selection, and sorting problems for points in R^d. The analogues are based on four different notions of the depth of a point. In studying properties of these measures, and algorithms to compute them, we have raised many questions. Perhaps the most interesting is whether sorting (ranking every point) is the most efficient way to perform selection (finding e.g., a median). Here are some of the other interesting problems:

(1) What is the expectation of π^*, the number of peels, under various distributions for n points in R^d?

(2) What is the value of $c(d) = \inf[\sigma(m_\sigma(S)): S \subset R^d, |S| = n]/n^2$ from Lemma 1?

(3) What is the breakdown point of the simplicial median when its depth is greater than n^d?

(4) What is the cost of computing all peels if $d > 2$?

(5) The way box medians are defined suggests a notion of median for any partial order $\prec$. Let n_i be the number of pairs (a_i, a_k) satisfying $a_j \prec a_i \prec a_k$, and the median, the element with maximum n_i. If all relations of the partial order were explicitly given, a brute force algorithm would solve this problem in $O(n^3)$ time. A partial order Q is d-dimensional if it is the intersection of d total orders. If these orders were explicitly given, the box median algorithm would apply, and would have the same time bound. The complexity for arbitrary partial orders is not known to us.

(6) It is interesting to seek a median analogue that is easy to compute, affine invariant, and has high breakdown point. The box median fails with respect to invariance. The others are hard to compute or easy to break down. Here are two alternatives. First, define a score function by

$$f(a_i) = \sum_{j \neq i} \|a_i - a_j\|;$$

$\|\cdot\|$ the Euclidean norm for R^d. A median is a point which minimizes f. This agrees with the usual median in R. Its advantage is $O(dn^2)$ cost.

Another interesting notion is the superposition of unit vectors from a_i in the direction of each a_j, i.e.,

$$v(a_i) = \sum_{j \neq i} \frac{a_i \vec{a}_j}{\|a_i \vec{a}_j\|}.$$

A median would be an a_i with $\|v(a_i)R\| \leq 1$. This would also agree with the usual median in R. J.E. Goodman (pers. com.) showed that such a median is unique. It could also be computed in quadratic time in all dimensions.

Notes added in proof. (1) J. Matoušek has shown that sorting is not necessary for the Tukey median by giving an $O(n(\log n)^5)$ algorithm [J. Matoušek, 'Computing the Center of Planar Point Sets', in *Discrete and Computational Geometry: Papers from the DIMACS Special Year*, J.E. Goodman, R. Pollack and W. Steiger, eds., American Math. Soc., 1991, pps. 221–230.]

(2) Luc Devroye can show (pers. com.) that $E(\pi^*) = \Theta(n^{2/3})$ if the points are a random sample of size n from a uniform distribution on a convex body $K \subset R^2$. Imre Bárány can show (pers. com.) that if S is a sample of n points from a uniform distribution on a convex body $K \subset R^d$, $E(\pi^*)$ is $n^{2/(d+1)}$, up to poly-logarithmic factors.

References

[1] I. Bárány, A generalization of Carathéodory's Theorem, Discrete Math. 40 (1982) 141–150.
[2] M. Blum, R. Floyd, V. Pratt, R. Rivest and R.E. Tarjan, Time bounds for selection, J. Comput. System Sci. 7 (1973) 448–461.
[3] E. Boros and Z. Füredi, The maximal number of covers by the triangles of a given vertex set on the plane, Geom. Dedicata 17 (1984) 69–77.
[4] B. Chazelle, New techniques for computing order statistics in Euclidean space, in: Proc. ACM Symp. on Comp. Geom. 1 (1985) 125–134.
[5] B. Chazelle, On the convex layers of a convex set, IEEE Trans. Inform. Theory 31 (1985) 509–517.
[6] R. Cole, M. Sharir and C. Yap, On k-hulls and related topics, SIAM J. Comput. 16 (1987) 61–77.
[7] D. Donoho, Breakdown properties of multivariate location estimators, Ph.D Qualifying Paper, Dept. Statistics, Harvard Univ., 1982.
[8] H. Edelsbrunner, Algorithms in Combinatorial Geometry (Springer, Berlin, 1987).
[9] H. Edelsbrunner, J. O'Rourke and R. Seidel, Constructing arrangements of lines and hyperplanes with applications, SIAM J. Comput. 15 (1986) 341–363.
[10] D. Gale, Neighboring vertices on a convex polyhedron, in: H. Kuhn and A. Tucker, eds., Linear Inequalities and Related Systems (Princeton University Press, Princeton, NJ, 1956) 255–263.
[11] J. Gill, W. Steiger and A. Wigderson, Computing certain medians, Tech. Report, Dept. of Comp. Sci., Rutgers Univ., 1988).
[12] S. Khuller and J. Mitchell, On a triangle counting problem, Inform. Proc. Letters 33 (1989) 319–321.
[13] D. Kirkpatrick and R. Seidel, The ultimate planar convex hull algorithm?, SIAM J. Comput. 15 (1986) 287–299.
[14] Regina Liu, A notion of data depth based on random simplicies, Ann. Statist. 18 (1989) 405–414.
[15] F. Preparata and S. Hong, Convex hulls of finite sets of points in two and three dimensions, Comm. ACM 2 (1977) 87–93.
[16] F. Preparata and M.I. Shamos, Computational Geometry (Springer, New York, 1985).
[17] H. Raynaud, Sur l'enveloppe convexe des nuages des points aléatoires dans R^n, I, J. Appl. Prob. 7 (1970) 35–48.
[18] A. Rényi and R. Sulanke, Ueber die konvexe Hulle von n zufallig gewahlen Punkten, I, Z. Wahrschein. 2 (1963) 75–84.
[19] J. Tukey, Mathematics and the picturing of data, Int. Conf. of Math. Vancouver, 1971.