A Proof of the Oja-Depth Conjecture in the Plane

Nabil H. Mustafa
Dept. of Computer Science,
LUMS, Pakistan.
nabil@lums.edu.pk

Hans Raj Tiwary
Inst. of Maths,
EPFL, Switzerland.
hansraj.tiwary@epfl.ch

Daniel Werner *
Institut für Informatik,
Freie Uni. Berlin, Germany.
daniel.werner@fu-berlin.de

Abstract

Given a set P of n points in the plane, the *Oja depth* of a point $x \in \mathbb{R}^2$ is defined to be the sum of the areas of all triangles defined by x and two points from P, normalized by the area of convex-hull of P. The Oja-depth of P is the minimum Oja-depth of any point in $\mathbb{R}^2$. The Oja-depth conjecture states that any set P of n points in the plane has Oja-depth at most $n^2/9$ (this would be optimal as there are examples where it is not possible to do better). We present a proof of this conjecture. We also improve the previously best bounds for all $\mathbb{R}^d$, $d \geq 3$, via a different, more combinatorial technique.

1 Introduction

The general area of statistical data analysis involves designing measures to quantitatively capture the spread and variance of multivariate data. For example, for a set of points in $\mathbb{R}$, the notion of *mean* and *median* are two natural measures. In particular, when the data consists of a set of finite points in Euclidean space $\mathbb{R}^d$, several notions for data depth have been proposed over the years. With each such measure, there come two questions: i) proving the existence of a point which suitably captures, with some guaranteed bounds, the spread under that measure and ii) devising efficient algorithms to compute this point.

Given a set P of n points in $\mathbb{R}^d$, some examples of various measures are the following. *Location depth* of a point x is the minimum number of points of P lying in any halfspace containing x [Hod55, Tuk75, RRT99]. The centerpoint theorem [Eck93] asserts that there always exists a point of location depth at least $n/(d+1)$, and that that is the best possible. The point with the highest location depth w.r.t. to a pointset P is called the Tukey-median of P. The corresponding computational question of finding the Tukey-median of a pointset has also been studied extensively, and an optimal algorithm with running time $O(n \log n)$ is known in $\mathbb{R}^2$ [Cha04]. Another example of a statistical depth measure is *Simplicial depth* [Liu90], which for a point x is the number of simplices spanned by P that contain x. The First Selection Lemma [Mat02] asserts that there always exists a point with simplicial depth at least $c_d \cdot n^{d+1}$, where $c > 0$ is a constant depending only d. The optimal value of c_d is known only for $d = 2$, where

*This research was funded by Deutsche Forschungsgemeinschaft within the Research Training Group (Graduiertenkolleg) "Methods for Discrete Structures"

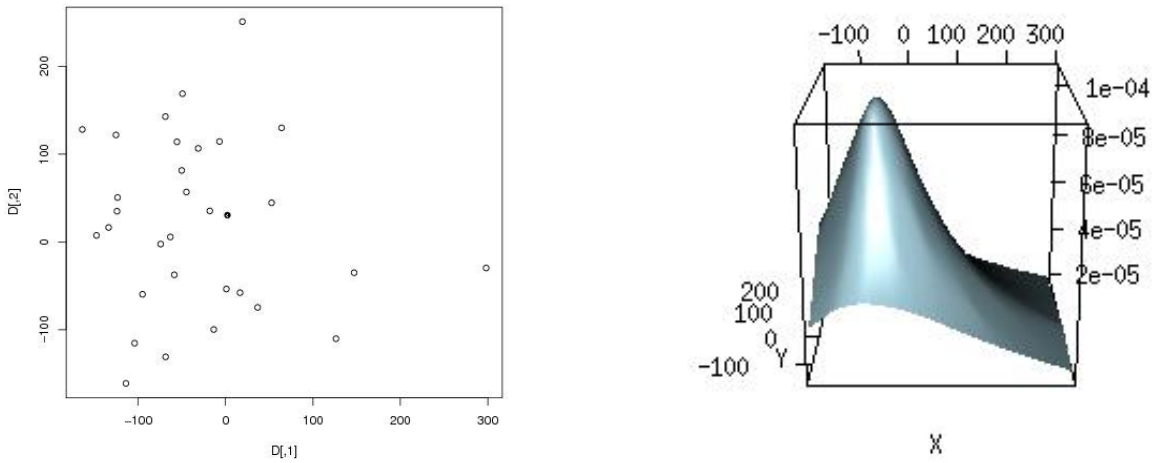


Figure 1: (a) Set of 30 points, together with a point (shaded black) with minimum Oja-depth, (b) Contours for a variant of Oja-depth for the pointset on the left implemented in the statistical package **R**.

$c_2 = 1/27$ [BF84]. Determining the exact value of c_3 is still open, though it has been the subject of a flurry of work recently [BMRR10, BMN10, Gro10]. The current-best algorithm computes the point with maximum simplicial depth in time $O(n^4 \log n)$ [Alo01].

Another well-studied measure, first proposed by Weber in 1909, is the so-called L_1 median, where the depth of a point $q \in \mathbb{R}^2$ is defined to be the sum of the distances of q to the n input points. Furthermore, it is known that the point with the lowest such depth is unique in $\mathbb{R}^2$.

In this paper, we study another well-known measure called the *Oja depth* of a pointset.

Oja depth. Given a set P of n points in $\mathbb{R}^d$, the *Oja* depth (first proposed by Oja [Oja83] in 1983) of a point $x \in \mathbb{R}^d$ w.r.t. P is defined to be the sum of the volumes of all $(d+1)$-simplices spanned by x and d other points of P. Formally, given a set $Q \subset \mathbb{R}^d$, let $conv(Q)$ denote the convex-hull of Q. And given an object $C \subset \mathbb{R}^d$, let vol(C) denote the d-dimensional volume of C. Then,

$$\text{Oja-depth}(x) = \sum_{y_1,\ldots,y_d \in \binom{P}{d}} \frac{\text{vol}(conv(x, y_1, \ldots, y_d))}{\text{vol}(conv(P))} \tag{1}$$

The Oja-depth of P is the minimum Oja depth over all $x \in \mathbb{R}^d$. From now onwards, w.l.o.g., assume that vol$(conv(P)) = 1$. See Figure 1(a) for an example of Oja-depth of a random pointset in the plane.

Known bounds. First note the following relation:

$$\left(\frac{n}{d+1}\right)^d \;\leq\; \text{Oja-depth}(P) \;\leq\; \binom{n}{d}$$

For the upper-bound, observe that any $(d+1)$-simplex spanned by points inside the convex-hull of P can have volume at most 1, and so a trivial upper-bound for Oja-depth of any $P \subset \mathbb{R}^d$ is $\binom{n}{d}$, achieved by picking any $x \in conv(P)$. For the lower-bound, construct P by placing $n/(d+1)$ points at each of the $(d+1)$ vertices of a unit-volume simplex in $\mathbb{R}^d$. It is easy to see that any point will have Oja-depth at least $(n/(d+1))^d$.

The conjecture [CDI$^+$10] is that the lower-bound given above is tight.

Oja-depth Conjecture. Oja-depth$(P) \leq (\frac{n}{d+1})^d$ for any $P \subset \mathbb{R}^d$ of n points.

The current-best upper-bound [CDI$^+$10] is that the Oja-depth of any set of n points is at most $\binom{n}{d}/(d+1)$. In particular, for $d = 2$, this gives $n^2/6$. This can be computed in linear time.

The Oja-depth conjecture states the existence of a low-depth point, but given P, computing the *lowest*-depth point is also an interesting problem. In $\mathbb{R}^2$, Rousseeuw and Ruts [RR96] presented a straightforward $O(n^5 \log n)$ time algorithm for computing the lowest-depth point, which was then improved to the current-best algorithm with running time $O(n \log^3 n)$ [Alo01]. An approximate algorithm utilizing fast rendering systems on current graphics hardware was presented in [KMV06, Mus04]. For general d, various heuristics for computing points with low Oja-depth were given by Ronkainen, Oja and Orponen [ROO03].

Our results. We present progress on the Oja-depth conjecture. In Section 2, we present our main theorem, which completely resolves the planar case.

Theorem 1.1. *Every set P of n points in $\mathbb{R}^2$ has Oja-depth at most $\frac{n^2}{9}$. Furthermore, such a point can be computed in $O(n \log n)$ time.*

In Section 3, using completely different (and more combinatorial) techniques for higher dimensions, we also prove the following:

Theorem 1.2. *Every set P of n points in $\mathbb{R}^d$, $d \geq 3$, has Oja-depth at most $\frac{2n^d}{2^d d!} - \frac{2d}{(d+1)^2(d+1)!}\binom{n}{d} + O(n^{d-1})$.*

This improves the previously best bounds by an order of magnitude.

2 The optimal bound

We now come to prove the optimal bound for $\mathbb{R}^2$. First, let us give some basic definitions. The *center of mass* or *centroid* of a convex set X is defined as

$$c(X) = \frac{\int_{x \in X} x \, dx}{\text{area}\,(X)}.$$

For a discrete point set P, the center of mass is simply defined as the center of mass of the convex hull of P. When we talk about the *centroid of* P, we refer to the center of mass of the convex hull and hope the reader does not confuse this with the discrete centroid $\sum p/|P|$.

During this paper, we will bound the Oja-depth of the centroid of a set, and show that it is worst-case optimal. Our proof will rely on the following two Lemmas.

Lemma 2.1. *[Winternitz [Bla23]] Every line through the centroid of a convex object has at most $\frac{5}{9}$ of the total area on either side.*

Lemma 2.2. *[CDI+10] Let P be a convex object with unit area and let c be its center of mass. Then every simplex inside P which has c as a vertex has area at most $\frac{1}{3}$.*

To simplify matters, we will use the following proposition.

Proposition 1. *If we project an interior point $p \in P$ radially outwards from the centroid c to the boundary of the convex hull, the Oja-depth of the point c does not decrease.*

Proof. First, observe that the center of mass does not change. It suffices to show that every triangle that has p as one of its vertices increases its area. Let $T := \Delta(c, p, q)$ be any triangle. The area of T is $\frac{1}{2}\|c - p\| \cdot h$, where h is the height of T with respect to $p - c$. If we move p radially outwards to a point p', h does not change, but $\|c - p'\| > \|c - p\|$. See Fig. 2.

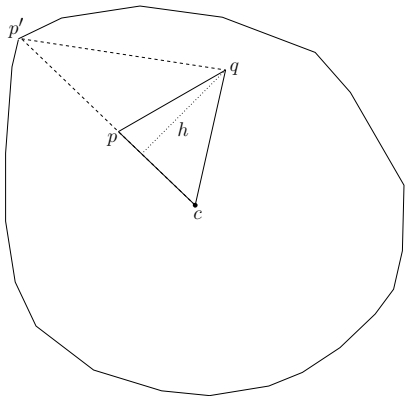


Figure 2: Moving points to the boundary increases the Oja-depth

□

This implies that in order to prove an upper bound, we can assume that all points lie on the convex hull.

Let us introduce some simple notation. From now on, let P be a set of points, and let $c := c(\text{ch}(P))$ denote its center of mass as defined above. Further, let $p_1, \ldots, p_n$ denote the points sorted clockwise by angle from c. We define the *distance* of two points p_i, p_j as

$$\text{dist}(p_i, p_j) := \min\{j - i \mod n, i - j \mod n\} \subseteq \{1, \ldots, \lfloor n/2 \rfloor\}.$$

A triangle that is formed by c and two points at distance i is called an *i-triangle*, or *triangle of type i*. Observe that for each i, $1 \leq i < \lfloor n/2 \rfloor$, there are exactly n triangles of type i. Further, if n is even, then there are $n/2$ triangles of type $\lfloor n/2 \rfloor$, otherwise there are n. These constitute all possible triangles.

Let $C \subseteq P$, and let $\mathcal{C}$ be the unique convex polygon inscribed in P that consists of the points of C. This will be called a *cycle*, and the length of the cycle is simply the number of elements in C. A cycle $\mathcal{C}$ of

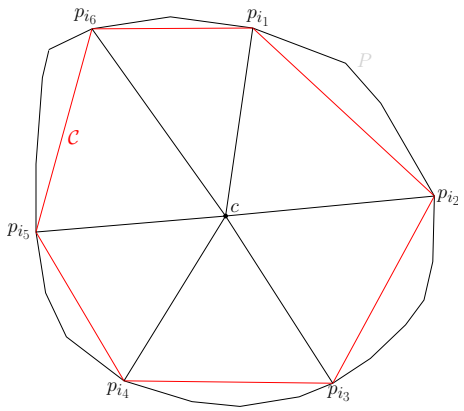


Figure 3: A cycle and its induced triangles

length k *induces* k *triangles* that arise by taking all the triangle formed by an edge in $\mathcal{C}$ and the center of mass c. The area induced by $\mathcal{C}$ is the sum of areas of these k triangles. See Fig. 3.

We now come to prove some basic facts about cycles. Because of Lemma 1, we get the following.

Corollary 2.1. *The total area of all triangles of type* 1 *is exactly* 1.

Note that this is the area of all triangles induced by the cycle arising from the entire set, P.

That we can bound the total area induced by *any* cycle is expressed in the next lemma.

Lemma 2.3. *Let $\mathcal{C}$ be a cycle. Then $\mathcal{C}$ induces a total area of at most* 1.

Proof. We distinguish two cases.

Case 1: The centroid lies in the convex hull of $\mathcal{C}$. In this case, all triangles are disjoint, so the area is at most one. See Fig. 4(a).

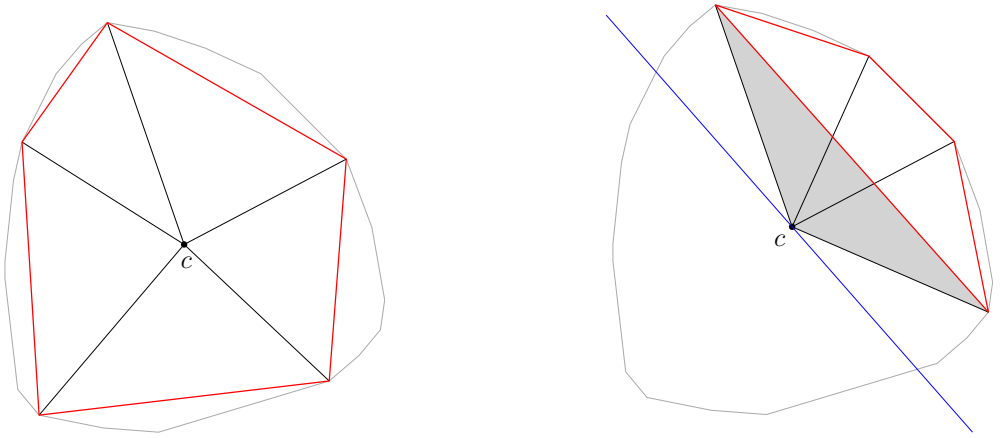
Case 2: The centroid does not lie in the convex hull of $\mathcal{C}$. By the Separation Theorem [Mat02], there is a line through c that contains all the triangles. Then we can remove one triangle to get a set of disjoint triangles, namely the one induces by the pair $\{p_{i_j}, p_{i_{j+1}}\}$ that has c on the left side. By Lemma 2.1, the area of the remaining triangles can thus be at most $5/9$. By Lemma 2.2, the removed triangle has an area of at most $1/3$. Thus, the total area is at most $8/9$. See Fig. 4(b). Here, the gray triangle can be removed to get a set of disjoint triangles.

□

We now prove the general version of Corollary 2.1.

Lemma 2.4. *The total area of all triangles of type i is at most i.*

Proof. To prove this, we will proceed as follows. We will first create n cycles. Each cycle will consist of one triangle of type i, and $n - i$ triangles of type 1. We then determine the total area of these cycles and subtract the area of all 1-triangles. This will give the desired result.



(a) Case 1: c lies inside convex-hull of $\mathcal{C}$ (b) Case 2: c lies outside convex-hull of $\mathcal{C}$

Figure 4: The two cases

Let $p_1, \ldots, p_n$ be the points ordered by angles from the centroid c. Let $\mathcal{C}_j$ be the cycle consisting of the $n-i+1$ points $P - \{p_{i+1 \mod n}, \ldots, p_{i+j-1 \mod n}\}$. This is a cycle that consists of 1 triangle of type i, namely the one starting a p_j, and $n-i$ triangles of type 1.

By Lemma 2.3, every cycle $\mathcal{C}_j$ induces an area of at most 1. If we sum up the areas of all cycles $\mathcal{C}_j$, $1 \leq j \leq n$, we thus get an area of at most n.

We now determine how often we have counted each triangle. Each i-triangle is counted exactly once. Further, for every cycle we count $n-i$ triangles of type 1. For reasons of symmetry, each 1-triangle is counted equally often. Thus, each is counted *exactly* $n-i$ times over all the cycles. By Corollary 2.1, their area is *exactly* $n-i$, which we can subtract from n to get the total area of the i-triangles:

$$\sum_{i\text{-triangle } T} \text{area}(T) \leq n - (n-i) \cdot \left(\sum_{1\text{-triangle } T} \text{area}(T) \right) = n - (n-i) = i.$$

This completes the proof.

□

Theorem 2.5. *Let P be any set of points in the plane and c be the centroid of its convex hull. Then the Oja-depth of c is at most $\frac{n^2}{9}$.*

Proof. We will bound the area of the triangles depending on their type. For i-triangles with $1 \leq i \leq \lfloor n/3 \rfloor$, we will use Lemma 2.4. For i-triangles with $\lfloor n/3 \rfloor < i \leq \lfloor n/2 \rfloor$, this would give us a bound worse than $n/3$, so we will use Lemma 2.2 for each of these.

By Lemma 2.4, the sum of the areas of all triangles of type at most $\lfloor n/3 \rfloor$ is at most

$$\sum_{i=1}^{\lfloor n/3 \rfloor} i = \frac{\lfloor n/3 \rfloor \left(\lfloor n/3 \rfloor + 1 \right)}{2} \leq \frac{n^2}{18} + \frac{1}{2} \lfloor n/3 \rfloor .$$

There are $n\left(\lfloor n/2\rfloor - \lfloor n/3\rfloor\right)$ triangles remaining, n for each type j, $\lfloor n/2\rfloor < j \leq \lfloor n/3\rfloor$. For these we use Lemma 2.2 to bound the size of each by $1/3$. This gives an area of $\frac{n(\lfloor n/2\rfloor - \lfloor n/3\rfloor)}{3}$.

So the Oja-depth is at most

$$\frac{n^2}{18} + \frac{n\left(\lfloor n/2\rfloor - \lfloor n/3\rfloor\right)}{3} + \frac{n}{6}.$$

This completes the proof, almost. It looks like there is an additional lower order term $n/6$. To argue that this disappears, we have to distinguish two cases. If n is even, there are only $n/2$ triangles induced by $n/2$ points, but we have counted n. Thus, we can subtract $n/2 \cdot 1/3 = n/6$ from the total area. If n is odd, then $\lfloor n/2\rfloor = (n-1)/2$. In either case, the term $n/6$ cancels out.

Thus, the Oja-depth of the centroid is at most $n^2/9$.

□

This completes the proof of Theorem 1.1.

3 Higher Dimensions

We now present improved bounds for the Oja-depth problem in dimensions greater than two. Before the main theorem, we need the following two lemmas.

Lemma 3.1. *Given any set P of n points in $\mathbb{R}^d$ and any point $q \in \mathbb{R}^d$, any line through q intersects at most $f(n,d)$ $(d-1)$-simplices spanned by P, where $f(n,d) = \frac{2n^d}{2^d d!} + O(n^{d-1})$.*

Proof. Given P, and the point q, let l be any line through q. Project P onto the hyperplane H orthogonal to l to get the pointset P' in $\mathbb{R}^{d-1}$. The line l becomes a point on H, say point l^*. Then l intersects the $(d-1)$-simplex spanned by $\{p_1, \ldots, p_d\}$ if and only if the corresponding points in P' contain the point l^*. By Barany [Bar82], given n points in $\mathbb{R}^d$, any point in $\mathbb{R}^d$ is contained in at most these many d-simplices:

$$\frac{2n}{n+d+1} \cdot \binom{(n+d+1)/2}{d+1} \text{ if } n-d \text{ is odd} \qquad \frac{2(n-d)}{n+d+2} \cdot \binom{(n+d+2)/2}{d+1} \text{ if } n-d \text{ is even}$$

Note that both the bounds above are equal within additive factor of $O(n^d)$, and simplifying the first, we get:

$$\frac{2n}{n+d+1} \cdot \binom{(n+d+1)/2}{d+1} \leq 2 \cdot \binom{(n+d+1)/2}{d+1} \leq \frac{2(n+d+1)^{d+1}}{2^{d+1}(d+1)!} \leq \frac{2n^{d+1}}{2^{d+1}(d+1)!} + O(n^d)$$

We apply this to P' in $\mathbb{R}^{d-1}$ to get the desired result. □

Lemma 3.2. *Given any set P of n points in $\mathbb{R}^d$, there exists a point q such that any half-infinite ray from q intersects at least $\frac{2d}{(d+1)^2(d+1)!}\binom{n}{d}$ $(d-1)$-simplices spanned by P.*

Proof. This follows directly from a recent result of Gromov [Gro10], who showed that given any set P, there exists a point q contained in at least $\frac{2d}{(d+1)(d+1)!}\binom{n}{d+1}$ simplices spanned by P. Now note that any half-infinite ray from q must intersect exactly one $(d-1)$-dimensional face of each simplex containing q and each such $(d-1)$-simplex can be counted at most $n-d$ times. Simplifying, we get the desired result. □

Given a set P and a point q, call a simplex a q-simplex if it is spanned by q and d other points of P.

Theorem 3.3. *Given any set P of n points in $\mathbb{R}^d$, there exists a point q with Oja-depth at most $\frac{2n^d}{2^d d!} - \frac{2d}{(d+1)^2(d+1)!}\binom{n}{d} + O(n^{d-1})$.*

Proof. Let q be the point from Lemma 3.2. Assign a weight function, $w(r)$, to each point $r \in conv(P)$, where $w(r)$ is the number of q-simplices spanned by P and q that contain r. Then note that if r is contained in a q-simplex, say spanned by $\{q, p_{i_1}, \ldots, p_{i_d}\}$, then the half-infinite ray $\overrightarrow{qr}$ intersects the $(d-1)$-simplex spanned by $\{p_{i_1}, \ldots, p_{i_d}\}$. Therefore $w(r)$ is upper-bounded by the number of $(d-1)$-simplices intersected by the ray $\overrightarrow{qr}$. To upper-bound this, note that the ray starting from q but in the opposite direction to the ray $\overrightarrow{qr}$, intersects at least $\frac{2d}{(d+1)^2(d+1)!}\binom{n}{d}$ $(d-1)$-simplices (by Lemma 3.2). On the other hand, by Lemma 3.1, the entire line passing through q and r intersects at most $\frac{2n^d}{2^d d!} + O(n^{d-1})$ $(d-1)$-simplices spanned by P. These two together imply that the ray $\overrightarrow{qr}$ intersects at most $\frac{2n^d}{2^d d!} - \frac{2d}{(d+1)^2(d+1)!}\binom{n}{d} + O(n^{d-1})$ $(d-1)$-simplices spanned by P, and this is also an upper-bound on $w(r)$. Finally, we have

$$\begin{aligned}
\sum_{|P'|=d} \text{vol}(conv(\{q\} \cup P')) &= \int_{x \in conv(P)} w(x)\, dx \\
&\leq \left(\frac{2n^d}{2^d d!} - \frac{2d}{(d+1)^2(d+1)!}\binom{n}{d} + O(n^{d-1})\right) \int_{x \in conv(P)} dx \\
&= \frac{2n^d}{2^d d!} - \frac{2d}{(d+1)^2(d+1)!}\binom{n}{d} + O(n^{d-1}),
\end{aligned}$$

finishing the proof. □

References

[Alo01] Greg Aloupis. On computing geometric estimators of location, 2001. M.Sc. Thesis, McGill University.

[Bar82] I. Barany. A generalization of caratheodory's theorem. *Discrete Mathematics*, 40:141–152, 1982.

[BF84] E. Boros and Z. Furedi. The maximal number of covers by the triangles of a given vertex set on the plane. *Geom. Dedicata*, 17:69–77, 1984.

[Bla23] W. Blaschke. *Vorlesungen uber Differentialgeometrie. II, Affine Differentialgeometrie.* Springer, 1923.

[BMN10] Boris Bukh, Jirí Matousek, and Gabriel Nivasch. Stabbing simplices by points and flats. *Discrete & Computational Geometry*, 43(2):321–338, 2010.

[BMRR10] Abdul Basit, Nabil H. Mustafa, Saurabh Ray, and Sarfraz Raza. Hitting simplices with points in 3d. *Discrete & Computational Geometry*, 44(3):637–644, 2010.

[CDI+10] Dan Chen, Olivier Devillers, John Iacono, Stefan Langerman, and Pat Morin. Oja Medians and Centers of Mass. In *Proceedings of the 22nd Canadian Conference on Computational Geometry (CCCG2010)*, pages 147–150, 2010.

[Cha04] Timothy M. Chan. An optimal randomized algorithm for maximum tukey depth. In *SODA*, pages 430–436, 2004.

[Eck93] J. Eckhoff. Helly, Radon and Carathéodory type theorems. In *Handbook of Convex Geometry*, pages 389–448. North-Holland, 1993.

[Gro10] M. Gromov. Singularities, expanders and topology of maps. part 2: From combinatorics to topology via algebraic isoperimetry. *Geometric and Functional Analysis*, 20:416–526, 2010.

[Hod55] J. Hodges. A bivariate sign test. *Annals of Mathematical Statistics*, 26:523–527, 1955.

[KMV06] Shankar Krishnan, Nabil Mustafa, and Suresh Venkatasubramanian. Statistical data depth and the graphics hardware. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, pages 223–250. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 2006.

[Liu90] R. Liu. A notion of data depth based upon random simplices. *The Annals of Statistics*, 18:405–414, 1990.

[Mat02] J. Matoušek. *Lectures in Discrete Geometry*. Springer-Verlag, New York, NY, 2002.

[Mus04] Nabil H. Mustafa. *Simplification, Estimation and Classification of Geometric Objects*. PhD thesis, Duke University, Durham, North Carolina, 2004.

[Oja83] H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1:327–332, 1983.

[ROO03] T. Ronkainen, H. Oja, and P. Orponen. Computation of the multivariate oja median. In *R. Dutter and P. Filzmoser*. International Conference on Robust Statistics, 2003.

[RR96] P. Rousseeuw and I. Ruts. Bivariate location depth. *Applied Statistics*, 45:516–526, 1996.

[RRT99] P. J. Rousseeuw, I. Ruts, and J. W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

[Tuk75] J. Tukey. Mathematics and the picturing of data. In *Proc. of the international congress of mathematicians*, pages 523–531, 1975.