International Journal of Computational Geometry & Applications

Quantile Approximation for Robust Statistical Estimation and k-Enclosing Problems[*]

DAVID M. MOUNT[†]
Department of Computer Science, University of Maryland, College Park, Maryland 20742
E-mail: mount@cs.umd.edu

NATHAN S. NETANYAHU[‡]
Dept. of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel
and Ctr. for Automation Research, University of Maryland, College Park, Maryland 20742
E-mail: nathan@macs.biu.ac.il

CHRISTINE D. PIATKO
The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, 20723
E-mail: christine.piatko@jhuapl.edu

RUTH SILVERMAN
Center for Automation Research, University of Maryland, College Park, Maryland, 20742
E-mail: ruth@cfar.umd.edu

ANGELA Y. WU
Department of Computer Science and Information Systems, American University, Washington,
DC 20016 E-mail: awu@american.edu

Received received date
Revised revised date
Communicated by Editor's name

ABSTRACT

Given a set P of n points in R^d, a fundamental problem in computational geometry is concerned with finding the smallest shape of some type that encloses all the points of P. Well-known instances of this problem include finding the smallest enclosing box, minimum volume ball, and minimum volume annulus. In this paper we consider the following variant: Given a set of n points in R^d, find the smallest shape in question that contains at least k points or a certain quantile of the data. This type of problem is known as a k-enclosing problem. We present a simple algorithmic framework for computing quantile approximations for the minimum strip, ellipsoid, and annulus containing a given quantile of the points. The algorithms run in $O(n \log n)$ time.

Keywords: Robust estimation, LMS regression, minimum enclosing disk, minimum volume ball/ellipsoid/annulus estimator.

[*] A preliminary version of this paper appeared in *Proceedings of the Tenth Canadian Conference on Computational Geometry*, Montréal, Québec, Canada, August 9–12, 1998, pp. 18–19.

[†] This work was supported in part by National Science Foundation grant CCR–9712379.

[‡] This research was carried out in part while the author was also affiliated with the Center of Excellence in Space Data and Information Sciences (CESDIS), Code 930.5, NASA/GSFC.

1. Introduction

Given a set P of n points in R^d, a fundamental problem in computational geometry is concerned with finding the smallest enclosing "range" of P. Well-known instances of this problem include finding the minimal enclosing box,[23] smallest enclosing simplex,[24] minimum volume ball,[31,37] (2-D) smallest ellipsoid,[30,37,6] and minimum width/volume annulus.[1,19,13] It will be assumed throughout that the range in question is a *generalized simplex*, which is defined to be the intersection of a constant number of halfspaces, or can be transformed into such a range by a suitable lifting into a higher dimensional space. We also assume that the points are in general position. That is, we assume that no $d + 1$ points are cohyperplanar, no two distinct d-tuples of points define parallel hyperplanes, etc. These assumptions are not difficult to overcome by taking some care in the implementation of the algorithm, and can certainly be overcome formally through the use of standard methods of simulating nondegenerate configurations of points.[9] We consider here the following generic variant.

Problem definition: *Given a set P of n points in d-dimensional space, and a generalized simplex range, R, find the smallest instance of R that contains at least k points or a certain quantile of the data.*

This variant is known as a *k-enclosing problem.* Instances include enclosing k points by a circle[18,11] and finding the smallest axis-parallel rectangle enclosing k points.[29] Although it has been studied to some extent in computational geometry the algorithms proposed deal mainly with specific cases, and their running times are relatively high ($O(n^2 \log n)$ and $O(n^3)$, respectively, for the minimum enclosing circle and axis-parallel rectangle in the plane).

Our main motivation for studying this variant stems from the growing need for efficient data analysis techniques that are robust to outlying and noisy observations. In recent years, there has been a great deal of interest in robust statistical estimators, because of their lack of sensitivity to outliers. The basic measure of the robustness of an estimator is its *breakdown point*, which is defined to be the fraction (up to 50%) of outlying data points that can corrupt the estimator. As it turns out, a number of highly robust estimators are defined in terms equivalent to the above generic problem formulation. In this paper, we will consider the following robust estimators:

- Rousseeuw's least median-of-squares (LMS) regression estimator[25] is among the best known 50% breakdown-point estimators. As will be clarified below, it corresponds to finding the narrowest hyperstrip containing at least half of the data points.

- The minimum volume ball (MVB) estimator is an LMS-like location estimator, which corresponds to finding the ball of minimum volume enclosing at least half of the data,[25] and the minimum volume ellipsoid (MVE) is defined similarly for ellipses.[25,26,27]

- The minimum width/volume annulus (MWA/MVA) estimator corresponds to finding an annulus of smallest volume/width that contains at least half of the data.

In some applications the number of outliers may exceed 50%, so it is common to generalize the problem definitions with respect to a specific rank. That is, in addition to the point set P, an algorithm will be given a (*residual*) *rank* k, where $k \leq n$, and will return the smallest corresponding range instance that contains at least k of the points. It is often more natural to represent k as a given *quantile*, or fraction, of the size of the data set. That is, given the quantile q, where $0 < q < 1$, we set $k = \lceil nq \rceil$. We will assume this latter formulation henceforth.

We will see below that exact computation of these robust estimators involves a high computational cost. In this paper we derive efficient approximations for the estimators in question through the use of quantile approximation. Consider the LMS estimator for example. First observe that it is not really reasonable to approximate the coefficients of the hyperplane of fit, since the data may not lie on a hyperplane in the first place. One reasonable approach is to compute a *width approximation*. In addition to the point set P and the quantile q, the algorithm is given a real $\epsilon_w > 0$, and produces a hyperstrip containing at least a fraction q of the points, such that the width of the hyperstrip is at most $(1 + \epsilon_w)w_{\text{opt}}$, where w_{opt} is the width of the LMS hyperstrip. The approach that we will consider here is called a *quantile approximation*. Given the desired quantile q and a *quantile error bound* ϵ_q, the algorithm returns a hyperstrip that contains at least $(1 - \epsilon_q)qn$ points and is no wider than the LMS hyperstrip.

Similarly to the LMS estimator, we can extend the notion of quantile approximation naturally to our other robust estimators. For example, the MVE estimator is a hyperellipsoid that contains at least $(1-\epsilon_q)qn$ points and is not larger in volume than $MVE_P(q)$, where $MVE_P(q)$ denotes the minimum volume ellipsoid containing $\lceil qn \rceil$ points of the data.

Our main result is that quantile approximations for all these estimators can be computed efficiently.

Theorem 1 *For a fixed quantile, ϵ_q-quantile approximations for the LMS, MVE, and MWA estimators in R^d for fixed d, as well for the MVA estimator in the plane can be computed in time*

$$O\left(n \log n + \left(\frac{1}{\epsilon_q}\right)^{O(d)}\right).$$

For fixed ϵ_q this is $O(n \log n)$.

This paper is organized as follows. In Section 2 we survey background and previous results on the estimators of interest. In Section 3 we draw on the notion of ϵ-approximation[17] to present a generic algorithmic framework for quantile approximation. In Section 4 we show that this framework leads to efficient quantile approximation algorithms for the robust estimators in question. In addition, we will show that the theoretical framework derived in this paper applies to any k-enclosing problem satisfying certain conditions. One such example is finding the minimal axis-parallel hyper-rectangle that encloses k of the points. Section 5 contains a proof of one of the technical lemmas needed for our results. Section 6 provides concluding remarks.

2. Background and previous results

While outlying and noisy data can be handled successfully by the use of robust statistical estimators, the efficient computation of these estimators continues to pose a formidable algorithmic challenge. In this subsection we provide additional background on the computation of robust estimators and other k-enclosing problems.

2.1. Least median of squares regression estimator

The LMS hyperplane estimator is defined formally as follows. Let $P = \{p_i = (x_{i1}, \ldots, x_{id}), i = 1, \ldots, n\}$, be a set of n points in R^d. Given a hyperplane $H : x_d = \theta_1 x_1 + \ldots + \theta_{d-1} x_{d-1} + \theta_d$, the *residual* of a point $(\xi_1, \xi_2, \ldots, \xi_d)$ with respect to H is $\xi_d - (\theta_1 \xi_1 + \ldots + \theta_{d-1} \xi_{d-1} + \theta_d)$. The *least median-of-squares (LMS) regression hyperplane* for P is the hyperplane H that minimizes the median of the squared residuals of P with respect to H. Put symbolically, this is the hyperplane H which minimizes $(\text{med}_i\, {r_i}^2)$, where r_i denotes the residual of point p_i with respect to H. In contrast, the ordinary least squares (OLS) estimator minimizes the sum of the squared residuals. Intuitively, if less than half of the points are outliers, then these points cannot adversely affect the median squared residual, which explains why LMS is a 50% breakdown point estimator. Observe that the median of the squared residuals is the same as the square of the median of the absolute values of the residuals. Henceforth we dispense with the squaring operation, and just describe the computation in terms of the absolute values of the residuals.

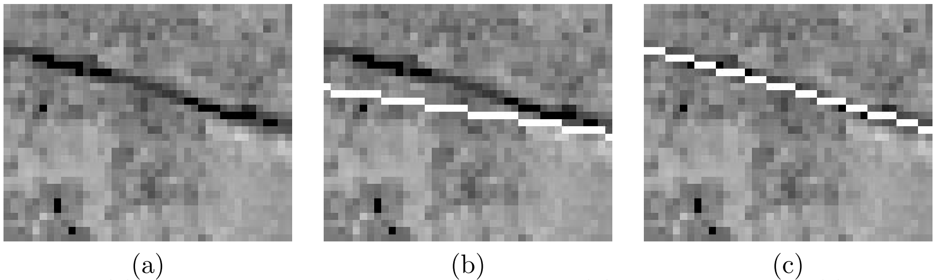

Figure 1: (a) A road segment with outlying pixels. (b) Its OLS line fit. (c) Its LMS line fit.

To view the LMS estimator as a k-enclosing problem, define a *hyperstrip* $\sigma = (H_1, H_2)$ to be the closed region in R^d lying between two nonvertical parallel hyperplanes H_1 and H_2. The *vertical width* of a hyperstrip, width(σ), is the length of its intersection with the x_d axis. Define $LMS_P(q)$ to be the hyperstrip of minimum vertical width that encloses at least $\lceil qn \rceil$ points from the set P. It is easy to see that the center hyperplane of $LMS_P(1/2)$ is the desired LMS regression hyperplane, and that the width of the hyperstrip is twice the magnitude of the median residual. We call $LMS_P(1/2)$ the *LMS hyperstrip*.

In addition to having a high breakdown-point, the LMS estimator is regression-, scale-, and affine-equivariant, which means that the estimate transforms "properly" under these types of transformations (see Ref. [26], pp. 116–117, for exact definitions). The LMS estimator may be used in its own right or as an initial step in a more complex estimation scheme.[38] It has been widely used in numerous applications of science and technology, and is considered a standard technique for robust data analysis. Our main motivation for studying the LMS estimator stems from its usage in computer vision (see, e.g., Refs. [20,33,34,22]). For example, Figure 1 demonstrates the enhanced performance obtained by LMS versus OLS (for $d = 2$) in detecting straight road segments in a noisy aerial image.[22]

An exact algorithm for computing the LMS hyperstrip is due to Stromberg.[35] Define an *elemental subset* for the LMS problem to be a subset of $d+1$ points. Each such subset defines a constant number (depending on d but not n) of hyperstrips. Stromberg shows that the optimal solution is realized by one of these hyperstrips. He presents an algorithm with worst-case complexity $O(n^{d+2} \log n)$. For our purposes, it suffices to use a simpler $O(n^{d+2})$ time algorithm due to Agulló.[4] His algorithm considers all $\binom{n}{d+1}$ elemental subsets and computes the number of points within each resulting hyperstrip. Among all hyperstrips containing the desired number of points, the strip of minimum width is returned. Obviously, these algorithms are impractical for large values of n. For $d = 2$, the best algorithm known for finding the LMS strip is the topological plane-sweep algorithm due to Edelsbrunner and Souvaine[10,32]. It runs in $O(n^2)$ time and requires $O(n)$ space. However, even quadratic running time is unacceptably high for many applications involving large data sets.

For this reason, what is often used in practice is a simple Monte Carlo approximation algorithm which runs in $O(n \log n)$ time for fixed d.[26] This algorithm randomly samples some constant number of d-tuples of points depending on the expected fraction of outliers and the user's desired confidence in the final result, but not on n. For each sampled tuple the coefficients of the hyperplane passing through these points are computed, and in $O(n \log n)$ time it is possible to approximate the hyperplane having those coefficients that minimize the k-th smallest squared residual. The intercept is computed by a reduction to a 1-dimensional LMS problem. Rousseeuw[26] shows that if a constant fraction of points does indeed lie on or very close to a hyperplane, then this sampling procedure will return the correct result with the desired confidence. However, if the data fail to satisfy this assumption, then there are no guarantees (even probabilistic) on the results of this algorithm. Likewise, the *feasible set algorithm* due to Hawkins,[15] an alternative Monte Carlo approximation, does not guarantee an error bound on the computed estimate either. Hence, neither of these Monte Carlo algorithms provides a completely satisfactory solution to the problem.

2.2. Minimum volume disk estimators

In addition to regression analysis/function fitting, the LMS principle can be extended to location and covariance estimation of mutivariate data. For example, the *minimum volume ball* (MVB) estimator can be defined formally as follows. Given $P = \{p_i = (x_{i1}, \ldots, x_{id})\}$, $i = 1, \ldots, n$, find the minimum volume sphere that covers at least half of the data. This corresponds to locating a point in R^d such that its median distance (with respect to each of the points in P) is minimized. As previously mentioned, algorithms have been proposed in the planar case,[18,11] but their complexity is rather high in the worst case, i.e., when $k = n/2$. This definition can be generalized as follows. Given a multivariate data set in R^d, find the *minimum volume ellipsoid* (MVE) that covers at least half of the data points. The MVB/MVE estimators also have a 50% breakdown point and they are translation-, scale-, and rotation-equivariant. Furthermore, the MVE estimator is affine-equivariant for any nonsingular affine transformation (see Ref. [26], p. 258).

Studying these estimators is of interest due to their usage in data clustering. Figure 2 illustrates, for example, the advantage gained by using an MVE estimator (as opposed to a standard maximum likelihood estimator (MLE)) in the presence of outliers. (The ellipse depicted in the figure is not the actual MVE, but is due to postprocessing applied to the actual MVE.)

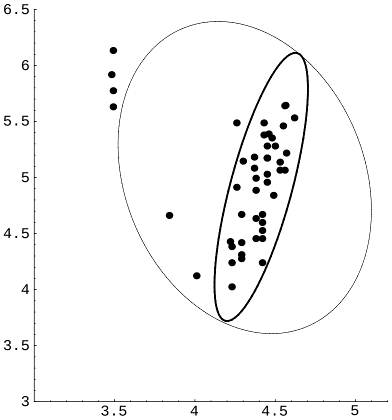

Figure 2: Cluster estimates for a contaminated set of data; MVE (bold) vs. MLE. Source: Rousseeuw and Leroy, p. 261.

An exact algorithm for the MVE estimator is due to Cook, Hawkins, and Weisberg.[5] The algorithm considers all subsets of the data of size $n/2$ and runs essentially in time $O\binom{n}{n/2}$. A recent algorithm by Agulló[3] appears to be more efficient in practice.

More commonly used algorithms are the Monte-Carlo approximations due to Rousseeuw (Ref. [26], pp. 258–262) and Hawkins.[16] However, as was noted in the previous subsection, such algorithms provide no guarantee on the accuracy of the result.

2.3. Minimum width/volume annulus

Similarly to a minimum volume disk, we may define a *minimum width/volume annulus* (MWA/MVA) as an annulus with the smallest width/volume that covers at least half of the data. (See Fig. 3 below.) We know of no efficient algorithm for solving this variant of the problem.

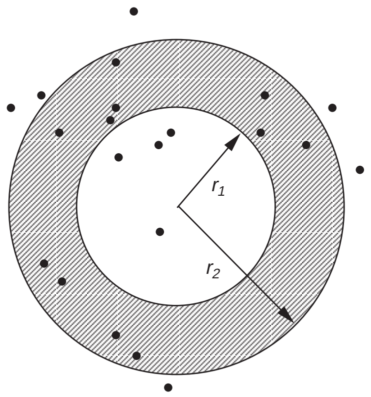

Figure 3: An illustration of a minimum width/volume annulus containing at least half of the data.

Another interesting k-enclosing problem is that of finding the smallest axis-parallel hyper-rectangle that encloses k points of the data. There are a number of algorithms for solving the problem in the plane. They can be computed in $O(k^2 n \log n)$ time and $O(kn)$ space,[2] $O(n \log n + k^2 n)$ time and $O(kn)$ space,[12] and $O(n \log n + k^2 n)$ time and $O(n)$ space.[8] Recently, Segal and Kedem presented a more general algorithm in d-dimensional space. Their algorithm requires $O(dn + dk(n-k)^{2(d-1)})$ time and $O(dn)$ space.[29] If k is roughly $n/2$ then, even in the plane, all of the above algorithms require $\Omega(n^3)$ time.

3. Generic Approach to Quantile Approximation.

We define our algorithmic framework in the abstract setting of set systems (also called *range spaces*). First, define a *set system* to be a pair (P, Σ), where P is a set and Σ is a collection of subsets of P. For the applications of interest, the elements of Σ can each be thought of as the subset of P contained within some shape, e.g., strip, ball, ellipse, or box. We will use the terms shape and range rather than set when referring to these geometric objects.

Given $\epsilon > 0$, a subset $A \subseteq P$ is an *ϵ-approximation* for P relative to the set system if for every range $\sigma \in \Sigma$,

$$\left| \frac{|A \cap \sigma|}{|A|} - \frac{|P \cap \sigma|}{|P|} \right| < \epsilon.$$

That is, the fraction of A that lies in any range and the fraction of the entire set that lies in the same range differ by at most ϵ. We will make use of the following result on the construction of ϵ-approximations for generalized simplices. It involves a straightforward generalization of a construction due to Matoušek.[17] The proof is presented in Section 5.

Lemma 1 *Given an n-point set P in Euclidean d-space and a parameter $\delta > 0$, for $t \geq 1$, an (unweighted) $(1/t)$-approximation of size $O(t^{d+1+\delta})$ for P with respect to generalized simplices can be computed in $O(n \log \min(t, n))$ time.*

Now we present our general algorithm for computing a quantile approximation for a generic k-enclosing problem. We assume that the ranges are described in the form of a set system (P, Σ), where the points P lie in d-dimensional Euclidean space, and the ranges Σ are generalized simplices in this space or can be mapped to generalized simplices in a higher-dimensional space through standard linearization, e.g., by lifting to a paraboloid. We also assume that we have access to an exact version of some exact k-enclosing problem. In particular, we assume that there exists an algorithm which, given any set of m points and an integer $h \leq m$, computes the smallest subset in the set system that contains at least h points. Moreover, we assume that this algorithm runs in $O(m^c)$ time for some constant c. Our analysis is based on the assumption that q, ϵ_q and dimension d are constants, independent of n.

The approximation algorithm is rather simple. Given the point set P, the quantile q, and the quantile error bound ϵ_q, it returns a minimum sized shape that contains at least a fraction $q(1 - \epsilon_q)$ of the points.

(1) Set $t = \max(n^{1/(c(d+1+\delta))}, \frac{2}{q\epsilon_q})$, where δ is the constant from Lemma 1.

(2) Compute a $(1/t)$-approximation $A \subseteq P$ of size $m = O(\min(n, t^{d+1+\delta}))$.

(3) Invoke the exact version of the algorithm on the above $(1/t)$-approximation to compute the smallest shape σ enclosing at least a fraction $q\left(1 - \frac{\epsilon_q}{2}\right)$ of A. This shape is the final quantile approximation.

Lemma 2 *The shape computed by the above algorithm is an ϵ_q quantile approximation.*

Proof. Let σ^* denote the minimum sized shape that contains a fraction q of the points of P. Let σ denote the shape returned by the approximation algorithm. Let size(σ) denote the associated geometric size measure of the shape, e.g., its width or volume. To establish that σ is a quantile approximation we need to establish the following two things.

(a) The size of σ is not greater than the optimum, that is, size$(\sigma) \leq$ size(σ^*). Suppose to the contrary that size$(\sigma) >$ size(σ^*). Since the exact algorithm did not select σ^* when run on the $(1/t)$-approximation, it follows that σ^* encloses less than a fraction

$$q\left(1 - \frac{\epsilon_q}{2}\right)$$

of the set A. Since σ^* encloses at least a fraction q of the points in P, we have

$$\left|\frac{|A \cap \sigma^*|}{|A|} - \frac{|P \cap \sigma^*|}{|P|}\right| \;>\; \left|q - q\left(1 - \frac{\epsilon_q}{2}\right)\right| \;=\; \frac{q\epsilon_q}{2} \;\geq\; \frac{1}{t},$$

contradicting the fact that A is a $(1/t)$-approximation.

(b) The shape σ encloses at least a fraction $q(1-\epsilon_q)$ of P. Suppose to the contrary that it enclosed a smaller fraction. Then we would have

$$\left|\frac{|A\cap\sigma|}{|A|} - \frac{|P\cap\sigma|}{|P|}\right| \;>\; \left|q\left(1-\frac{\epsilon_q}{2}\right) - q\left(1-\epsilon_q\right)\right| \;=\; \frac{q\epsilon_q}{2} \;\geq\; \frac{1}{t}.$$

Again, this is a contradiction.

□

Lemma 3 *Given fixed q and d and access to an exact polynomial time k-enclosing algorithm, for any $0<\epsilon_q<1$ the generic quantile approximation algorithm runs in time*

$$O\left(n\log n + \left(\frac{1}{\epsilon_q}\right)^{O(d)}\right).$$

Under our assumption that ϵ_q is a constant, the running time is $O(n\log n)$.

Proof. By Lemma 1, Step (2) of the algorithm can be performed in time $O(n\log\min(t,n)) \leq O(n\log n)$ time. The size of the approximation is

$$m \;=\; O(t^{d+1+\delta}) \;=\; O\left(\max\left(n^{1/c}, \left(\frac{2}{q\epsilon_q}\right)^{d+1+\delta}\right)\right) \;=\; O\left(n^{1/c} + \left(\frac{1}{\epsilon_q}\right)^{O(d)}\right),$$

given that q is fixed. Step (3) can be accomplished in $O(m^c) = O(n + (1/\epsilon_q)^{O(d)})$ time. The overall running time is dominated by the sum of times for Steps (2) and (3). If ϵ_q is fixed, then this is $O(n\log n)$. □

Remark: Note from the proof of Lemma 2, the choice $t = 2/(q\epsilon_q)$ in the algorithm would have been sufficient to guarantee the approximation bounds. The actual choice made in the algorithm allows the algorithm to produce even more accurate results, subject to the restriction that the running time of the exact k-enclosing algorithm on the $(1/t)$-approximation is linear in n. This feature is often useful in practice. The price we pay is the additional $O(\log n)$ factor in the running time, which would be replaced by $O(\log t) = O(\log 1/(q\epsilon_q))$ otherwise.

4. Applications

Based on the definitions and discussion in the introduction, and in view of the generic algorithmic framework provided in the previous section, we now demonstrate how quantile approximations can be computed in $O(n\log n)$ time for the estimators in question.

4.1. LMS.

The shape in this case is simply a hyperstrip (in d-space), which is a generalized simplex, and its size is its vertical width. Once an ϵ-approximation A is found, we set $c = d+2+\delta'$ (for any $\delta' > 0$) and invoke an exact LMS algorithm (as in Ref. [35] or Ref. [3]) with respect to A. Let $LMS_A(q(1-\epsilon_q/2))$ denote the resulting hyperstrip.

According to Lemma 2, the computed hyperstrip is a quantile approximation of $LMS_P(q)$. The overall time complexity is determined by the computation time of the ϵ-approximation. According to Lemma 1, this requires $O(n \log n)$ time.

4.2. MVE.

This is perhaps the most interesting case considered. Note that the shape is an ellipsoid, i.e., it is not a generalized simplex. However, we may still draw on Lemma 1 by mapping the given set of points, through linearization, to a higher-dimensional space as follows. A point $p = (x_1, \ldots, x_d)$ in R^d is transformed to a point p' in R^D, where $D = d(d+3)/2$. The first d coordinates of p' are $x_1, \ldots, x_d$, and the remaining coordinates are the products $x_i x_j$ for $1 \le i \le j \le d$. That is

$$p' = (x_1, \ldots, x_d, {x_1}^2, x_1x_2, \ldots x_1x_d, {x_2}^2, x_2x_3, \ldots, {x_d}^2).$$

In doing so, every ellipsoid in R^d corresponds to a hyperplane in D-dimensional space and vice versa. It is easy to verify that the corresponding range is thus a halfspace in D-dimensional space, which is a generalized simplex. An ϵ-approximation A of the transformed points in R^D is computed with respect to this halfspace. Once A is found, we map its m points to the original d space. This yields an ϵ-approximation with respect to ellipsoids (in the original space) due to the correspondence between ellipsoids and halfspaces.

We now invoke an exact algorithm on the m points in R^d to obtain the desired MVE. To comply with the algorithmic framework of Section 3, we need to show that there exists an exact polynomial time algorithm to compute an MVE that contains $h \le m$ of the points. Let E denote the desired MVE, and let H denote the set of h points contained in the MVE. Also, let E_H denote the smallest ellipse containing (all of the points in) H. By an easy contradiction argument it follows that $E = E_H$.

It is well known that the number of points on the surface of the smallest ellipse that contains a set of h points in d-space is between $d+1$ and $d(d+3)/2$ (see Ref. [36]). Let $h^{'}$ denote this number, and let S denote the set of $h^{'}$ points lying on the surface of this ellipse. S is called the *support set* of the ellipse.

Claim 1 *$E_H = E_S$, i.e., the smallest ellipse containing H is identical to the smallest ellipse containing the support set of H.*

Proof. This follows from well-known uniqueness properties of the John-Löwner ellipsoid.[7] See also Ref. [14] Proposition 2.1(iii) and Ref. [28] for a more formal proof. □

The above discussion suggests that to compute the desired MVE that contains h points, we may consider all $h^{'}$-tuples, where $h^{'} = d+1, \ldots, D$, and where each tuple is a support-set candidate. For each $h^{'}$-tuple we then compute the smallest ellipsoid containing this tuple and report that ellipsoid for which a minimum is attained.

Claim 2 *In any fixed-dimensional space, the above algorithm runs in polynomial time in h.*

Proof. Based on Chazelle and Matoušek[6] and Matoušek, Sharir, and Welzl,[19] computing the smallest ellipsoid containing $h^{'}$ points in R^d can be done in $O(h^{'})$

time. Specifically, it was shown that such an ellipsoid can be computed deterministically in $D^{O(D)}h'$ time[6]. Hence computing the smallest ellipsoid for each h'-tuple will require

$$\sum_{h'=d+1}^{D} D^{O(D)}h' \binom{h}{h'},$$

or $O(D^{O(D)}h^D) = O(m^c)$ time, where $c = D$. □

Setting $c = D$ and applying the general framework results in an $O(n \log n)$-time quantile approximation algorithm for the MVE.

4.3. MWA/MVA.

As with the minimum volume ellipsoid, since an annulus is not a generalized simplex, we need to map the data points to a higher dimensional space and compute an ϵ-approximation on the transformed data set with respect to a corresponding generalized simplex. This can be done through linearization, where each point $(x_1, \ldots, x_d)$ in d-space is mapped to $(x_1, \ldots, x_d, \sum_{i=1}^{d} {x_i}^2)$ in R^{d+1}. It is easy to see that this transformation establishes a one-to-one correspondence between a d-dimensional annulus and a hyperstrip in R^{d+1}, the latter being a generalized simplex. Thus an ϵ-approximation A is computed in $d+1$-space with respect to hyperstrips. The points of A can then be transformed back to R^d.

We now need to show that there exists an exact, polynomial time algorithm that computes the MWA/MVA containing h of the m points of A. We first consider finding the desired minimum-width annulus. Analogously to the MVE derivation, let W denote the desired width annulus, and let H denote the set of h points contained in W. Also, let W_H denote the minimum width annulus containing H. One can easily show that $W = W_H$. Furthermore, the number of points on the boundaries of W_H is at least $d+2$. (This is implied by the discussion in García, et al.[13] following their Theorem 1.) Note that $d+2$ points uniquely determine a d-dimensional annulus. (To see this observe that lifting to a paraboloid yields a linear system of $d+2$ equations with $d+2$ unknowns from which the center's d coordinates and the two radii of the annulus can be found.) Thus, we need only consider all $d+2$-tuples (or support-set candidates) and record that support set for which the corresponding annulus attains minimum width. Since the number of candidates is $2^{d+2}\binom{h}{d+2}$ (as there are 2^{d+2} ways of distributing $d+2$ points between the inner and outer disks of an annulus), and since for each candidate the corresponding annulus can be computed in $O(d^3)$ time, finding the desired minimum-width annulus requires $O(2^d d^3 h^{d+2}) = O(m^c)$ time, where $c = d+2$ and d is fixed. Setting $c = d+2$ leads to an $O(n \log n)$-time quantile approximation algorithm for the MWA.

We now consider finding the minimum-volume annulus containing a certain fraction of the points. While an ϵ-approximation in R^{d+1} can be found as before, we know of no analogous result (to that of García and Ramos) that implies that the optimal MVA can also be determined by an elemental subset. We show how to solve the problem for $d = 2$. By projecting the points to a paraboloid, it can be shown that an annulus $A(r_1, r_2)$ (in any d-space) maps to a hyperstrip (in R^{d+1}) whose vertical width is ${r_2}^2 - {r_1}^2$. Since the area of annulus in the plane is proportional to the difference of its squared radii, the MVA problem for $d = 2$ is reduced to finding an LMS estimator in 3-space. Setting $c = 5 + \delta$ for some small δ, leads to an $O(n \log n)$-time quantile approximation algorithm in the plane.

4.4. Generalizations.

As previously noted, the algorithmic framework presented in Section 3 applies to any k-enclosing problem, provided that the associated range is either a generalized simplex (or it can be mapped to one) and that there exists an exact polynomial time algorithm to compute a desired k-enclosing range.

As an example, reconsider the problem of finding the smallest axis-parallel hyper-rectangle that encloses k points. Obviously the range here is a generalized simplex, where the number of halfspaces is $2d$. In addition, based on the result by Segal and Kedem,[29] the smallest k-enclosing hyper-rectangle can be computed in $O(n^{2d-1})$ time (for $k \approx n/2$). Thus, according to the algorithmic framework, setting $c = 2d-1$ in this case would lead to an $O(n \log n)$-time quantile approximation algorithm.

5. Proof of Lemma 1

In this section we present a proof of Lemma 1 on the existence and construction of ϵ-approximations. Our approach is to generalize a result of Matoušek on the existence of weighted ϵ-approximations for simplices to unweighted ϵ-approximations for generalized simplices. Given a set system (P, Σ) and $\epsilon > 0$, a *weighted ϵ-approximation* is a subset $A \subseteq P$ and associated positive integer weights $w(a)$ for each $a \in A$, such that for all ranges $\sigma \in \Sigma$,

$$\left| \frac{w(A \cap \sigma)}{w(A)} - \frac{|P \cap \sigma|}{|P|} \right| < \epsilon,$$

where $w(X)$ denotes the total weight of set X.

For our purposes, define a *simplex* in d-space to be the intersection of $d + 1$ halfspaces. Note that this definition allows for unbounded simplices. Define a *generalized simplex* to be the intersection of at most g halfspaces, for some constant g. Matoušek has shown that given an n-point set P in d-space, and a set system consisting of P and ranges consisting of (standard) simplices, a weighted $(1/t)$-approximation for P of size $O(t^{d+\delta})$ can be computed in $O(n \log t)$ time.[17] In order to present our generalization, we first review the concept of simplicial partitions. Given an n-point set P, define a *simplicial partition* to be a collection

$$\Pi = \{(P_1, \Delta_1), \ldots, (P_\ell, \Delta_\ell)\},$$

where the P_i's form a partition of P and each Δ_i is a relatively open simplex containing P_i. A simplicial partition Π has *crossing number κ* if no hyperplane intersects more than κ simplices of Π. We use the following result, which was proved by Matoušek.[17]

Lemma 4 (Construction of Simplicial Partitions) *Given $\delta > 0$, an n-point set $P \subseteq R^d$, and an integer parameter s, $2 \leq s < n$, a simplicial partition for P satisfying $s \leq |P_i| < 2s$ for every class P_i and crossing number $O((n/s)^{1-(1/d)+\delta})$ can be constructed in time $O(n \log(n/s))$.*

From the class size restriction, the partition is of size $O(n/s)$. Our result differs from the one presented in Ref. [17] in that it avoids the need to attach weights to the members of the approximation, but we pay a price in having a slightly larger size.

Consider an n-point set $P \subseteq R^d$. Let $\Pi = \{(P_1, \Delta_1), \ldots, (P_\ell, \Delta_\ell)\}$ be a simplicial partition for P as given in Lemma 4, and let κ denote its crossing number. If $n = O(t^{d+\delta})$, we may just take P to be the approximation. Otherwise, let $s = \beta n / t^{d+\delta}$, for a constant $\beta < 1$ to be determined later. Observe that s grows with n but is smaller than n, and hence satisfies the conditions of Lemma 4. We may assume that $s \geq 4t$, for otherwise, $n = O(t^{d+1+\delta})$ and again, we may simply take P to be the approximation.

Let $\alpha = 4t/s$. Since $s \geq 4t$, it follows that $\alpha \leq 1$. For each class P_i, let A_i denote any subset of P_i of size $\lceil \alpha |P_i| \, \rceil$, and let A denote the union of these (disjoint) subsets. We assert that A is the desired $(1/t)$-approximation. Clearly $|A| \geq \alpha |P| = \alpha n$. Also, since the partition has at most n/s classes, we have

$$|A| \;=\; \sum_i \lceil \alpha |P_i| \, \rceil \;\leq\; \alpha \sum_i |P_i| + \frac{n}{s} \;\leq\; \alpha n + \frac{\alpha n}{4t} \;=\; \left(1 + \frac{1}{4t}\right) \alpha n.$$

It follows that $|A| = O(\alpha n) = O(t^{d+1+\delta})$, and by Lemma 4, A can be computed in $O(n \log(n/s)) = O(n \log \min(t, n))$ time.

All that remains to be shown is that A is a $(1/t)$-approximation for generalized simplices. Let σ be a generalized simplex with g sides. First, consider any class P_i whose corresponding simplex Δ_i is crossed by one of the g hyperplanes of σ. Because $|A| \geq \alpha n$ and $|A_i| \leq \alpha |P_i| + 1$ we have

$$\frac{|A_i \cap \sigma|}{|A|} \;\leq\; \frac{|A_i|}{\alpha n} \;\leq\; \frac{|P_i| + (1/\alpha)}{n}.$$

Since $t \geq 1$ we have $1/\alpha = s/(4t) \leq s$. Combining this with the fact that $|P_i| < 2s$ we have

$$\left| \frac{|A_i \cap \sigma|}{|A|} - \frac{|P_i \cap \sigma|}{|P|} \right| \;\leq\; \frac{|P_i| + (1/\alpha)}{n} \;\leq\; \frac{3s}{n}.$$

Since none of the g sides of σ can cross more than κ hyperplanes of the simplicial partition, if we sum this quantity over all crossed simplices, the total is at most $3sg\kappa/n$.

Let $\delta' = (1/d) - 1/(d+\delta)$, implying that $(d+\delta)((1/d) - \delta') = 1$. By applying Lemma 4 (using δ' in place of δ) and using the above definition of s, we have

$$\begin{aligned} \frac{3sg\kappa}{n} \;&=\; 3g(s/n) O\left((s/n)^{-1+(1/d)-\delta'}\right) \;=\; 3gO\left((s/n)^{(1/d)-\delta'}\right) \\ &=\; 3g\beta^{1/(d+\delta)} O\left((1/t)^{(d+\delta)((1/d)-\delta')}\right) \;=\; 3g\beta^{1/(d+\delta)} O(1/t) \;\leq\; \frac{1}{2t}, \end{aligned}$$

for a suitable choice of $\beta < 1$.

Next, we consider simplices that are not crossed by any of the sides of σ. Any such simplex lies entirely inside or entirely outside of σ. It contributes to the approximation error only if it lies inside of σ, that is if $|A_i \cap \sigma| = |A_i|$ and $|P_i \cap \sigma| = |P_i|$. Thus for these simplices, it suffices to bound the absolute value of

$$x = \frac{|A_i|}{|A|} - \frac{|P_i|}{|P|}.$$

Based on our previous bounds on $|A|$ we have

$$x \;\le\; \frac{|A_i|}{\alpha n} - \frac{|P_i|}{n} \;\le\; \frac{1}{\alpha n}(\lceil \alpha |P_i| \rceil - \alpha |P_i|) \;\le\; \frac{1}{\alpha n}.$$

Similarly,

$$\begin{aligned} -x \;&\le\; \frac{|P_i|}{n} - \frac{|A_i|}{(1+1/4t)\alpha n} \;\le\; \frac{1}{n}\left(|P_i| - \frac{|P_i|}{(1+1/4t)}\right) \\ &=\; \frac{|P_i|}{n(4t+1)} \;\le\; \frac{2s}{4nt} \;=\; \frac{2}{n\alpha}. \end{aligned}$$

So $|x| \le 2/(n\alpha)$. Since there are at most n/s simplices in the simplicial partition, by summing $|x|$ over all of the simplices that are not crossed by σ, the total is at most $2/(s\alpha) = 1/(2t)$.

Combining the bounds from the two cases of simplices (crossed and not crossed), it follows that the total absolute difference is no larger than $1/(2t) + 1/(2t) = 1/t$, which completes the proof.

6. Discussion

In this paper we have presented an algorithmic methodology based on the framework of ϵ-approximations to compute quantile approximations efficiently for a number of robust estimators and other k-enclosing problems. Specifically, we showed that a quantile approximation for the problems considered can be computed in $O(n \log n)$ time for fixed q, ϵ_q, and d. The following questions arise in the context of future research:

1. Can the result for the minimum volume annulus be extended to $d > 2$?

2. Given that the algorithms described rely on the efficient computation of ϵ-approximations whose practical implementation has yet to be demonstrated, it remains to be seen whether (quantile) approximations can be further pursued to yield practical algorithms in higher dimensions for the problems considered.

References

1. P. K. Agarwal, M. Sharir, and S. Toledo. Applications of parametric searching in geometric optimization. *Journal of Algorithms*, 17:292–318, 1994.

2. A. Aggarwal, H. Imai, N. Katoh, and S. Suri. Finding k points with minimum diameter and related problems. *Journal of Algorithms*, 12:38–56, 1991.

3. J. Agulló. Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. *Proceedings of the Twelfth COMPSTAT Symposium on Computational Statistics*, Barcelona, Spain, August 1996. A. Prat, ed., Physica-Verlag, 175–180.

4. J. Agulló. Exact algorithms for computing the least median of squares estimate in multiple linear regression algorithm. Presented at the Third International Conference on Statistical Data Analysis based on the L1-Norm and Related Methods, Neuchatel, Switzerland, August, 1997.

5. R. D. Cook, D. M. Hawkins, and S. Weisberg. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics and Probability Letters*, 16:213–218, 1993.

6. B. Chazelle and J. Matoušek. On linear-time deterministic algorithms for optimization problems in fixed dimension. *Journal of Algorithms*, 21:579–597, 1996.

7. L. Danzer, D. Laugwitz, and H. Lenz. Über das Löwnersche Ellipsoid und sein Analogon unter sen einem Eikörper eingeschriebenen Ellipsoiden. *Arch. Math.*, 8:214–219, 1957.

8. A. Datta, H. P. Lenhof, C. Schwarz, M. Smid. Static and dynamic algorithms for k-point clustering problems. *Lecture Notes in Computer Science*, vol. 70, Springer, Berlin, 1993, 265–276.

9. H. Edelsbrunner and E. P. Mücke. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics*, 9:66–104, 1990.

10. H. Edelsbrunner and D. L. Souvaine. Computing median-of-squares regression lines and guided topological sweep. *Journal of the American Statistical Association*, 85:115–119, 1990.

11. A. Efrat, M. Sharir, and A. Ziv. Computing the smallest k-enclosing circle and related problems. *Computational Geometry Theory and Applications*, 4:119–136, 1994.

12. D. Eppstein and J. Erickson. Iterated nearest neighbors and finding minimal polytopes. *Discrete & Computational Geometry*, 11: 321–350, 1994.

13. J. García-López, P. Ramos and J. Snoeyink. Fitting a set of points by a circle. *Discrete & Computational Geometry*, 20: 389–402, 1998.

14. B. Gärtner and S. Schönherr. Smallest enclosing ellipses—Fast and exact. Technical report B 97-03, Freie University, Berlin, 1997.

15. D. M. Hawkins. The feasible set algorithm for least median of squares regression. *Computational Statistics and Data Analysis*, 16:81–101, 1993.

16. D. M. Hawkins. The feasible set algorithm for the minimum volume ellipsoid estimator in multivariate data. *Computational Statistics*, 8:95–107, 1993.

17. J. Matoušek. Efficient partition trees. *Discrete & Computational Geometry*, 8:315–334, 1992.

18. J. Matoušek. On enclosing k points by a circle. *Information Processing Letters*, 53:217–221, 1995.

19. J. Matoušek, M. Sharir, and E. Welzl. A subexponential bound for linear programming. *Algorithmica*, 16:498–516.

20. P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for

computer vision—A review. *International Journal of Computer Vision*, 6:59–70, 1991.

21. D. M. Mount, N. S. Netanyahu, K. Romanik, R. Silverman, and A. Wu. A practical approximation algorithm for the LMS line estimator. *Proceedings of the Eighth ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, Louisiana, January 1997, 473–482.
22. N. S. Netanyahu, V. Philomin, A. Rosenfeld, and A. J. Stromberg. Robust detection of road segments in noisy aerial images. *Pattern Recognition*, 30:1673–1686, 1997.
23. J. O'Rourke. Finding minimal enclosing boxes. *International Journal of Computer and Information Sciences*, 14:183–199, 1985.
24. J. O'Rourke, A. Aggarwal, S. Maddila, and M. Baldwin. An optimal algorithm for finding minimal enclosing triangles. *Journal of Algorithms*, 7:258–269, 1986.
25. P. J. Rousseeuw. Least median-of-squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
26. P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.
27. P. J. Rousseeuw and B. L. van Zomeren. Unmasking multiple outliers and leverage points (with comments and rejoinder). *Journal of the American Statistical Association*, 85:633–651, 1990.
28. S. Schönherr. *Berechnung kleinster Ellipsoide um Punktemengen.* Diploma thesis, Freie University Berlin, 1994.
29. M. Segal and K. Kedem. Enclosing k points in the smallest axis parallel rectangle. *Information Processing Letters*, 65:95–99, 1998.
30. B. W. Silverman and D. M. Titterington. Minimum covering ellipses. *SIAM Journal on Scientific and Statistical Computing*, 1:401–409, 1980.
31. S. Skyum. A simple algorithm for computing the smallest enclosing circle. *Information Processing Letters*, 37:121–125, 1991.
32. D. L. Souvaine and J. M. Steele. Time- and space- efficient algorithms for least median of squares regression. *Journal of the American Statistical Association*, 82:794–801, 1987.
33. A. Stein and M. Werman. Robust statistics in shape fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Champaign, Illinois, June 1992, 540–546.
34. C. V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:925–938, 1996.
35. A. J. Stromberg. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal on Scientific Computing*, 14:1289–1299, 1993.
36. D. M. Titterington. Optimal design: Some geometrical aspects of D-optimality. *Biometrika*, 62:313–320, 1975.
37. E. Welzl. Smallest enclosing disks (balls and ellipses). In H. Maurer, ed., *LNCS 555 (New Results and New Trends in Computer Science)*, 359–370. Springer Verlag, 1991.
38. V. J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656, 1987.