Quantile Approximation for Robust Statistical Estimation

David M. Mount* Nathan S. Netanyahu† Christine Piatko‡ Ruth Silverman§
Angela Y. Wu¶

Abstract

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, e.g., the smallest enclosing box [10], smallest enclosing simplex [11], minimum volume ball [17, 20], (2-D) smallest enclosing ellipsoid [16, 20, 2, 14], and minimum volume annulus [1, 8, 5].

In this paper, we consider the following generic variant: Given a set of n points in R^d, find the smallest range in question that contains (at least) a certain quantile (up to 50%) of the data. Although this variant has been studied to some extent in computational geometry — instances include enclosing k points by a circle [7, 4] and finding the smallest axis parallel rectangle enclosing k points [15] — 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).

The motivation for studying the above variant stems from the growing need for efficient data analysis techniques that are robust to outlying/noisy observations. While such data can be handled successfully due to robust statistical estimators, the efficient computation of these estimators continues to pose a formidable algorithmic challenge.

The basic measure of the robustness of an estimator is its breakdown point, that is, the fraction (up to 50%) of outlying data that can corrupt the estimator (see, e.g., [13] for an exact definition). For example, Rousseeuw's least median-of-squares (LMS) regression estimator [12] (which is equivalent to finding the narrowest hyperstrip contanining at least 50% of the data) is among the best known 50% breakdown-point estimators. The best exact algorithm known for this problem runs in $O(n^{d+2} \log n)$ time [19]. Because of this high running time, a simple $O(n \log n)$ Monte Carlo algorithm (for fixed d) is commonly used [13]. However, the latter provides no guarantee of accuracy (even probabilistic) unless the data set satisfies certain assumptions.[1] Other robust estimators considered in this paper are the minimum volume disk (e.g., the minimum volume ellipsoid (MVE)) and the minimum volume annulus (MVA) enclosing (at least) half of the points in P. As in the case of LMS, algorithms known for these estimators have either high complexity or they provide no guarantee of accuracy due to their Monte Carlo nature [13].

In an attempt to provide efficient algorithms for the above mentioned estimators which do guarantee the accuracy of the results, we pursue in this paper an approximation-based approach. Specifically, we compute *quantile approximations* of these estimators. That is to say, given two integers $k^- \leq k^+$, the algorithm returns a hyperstrip/disk/annulus that contains at least k^- points and is no wider/larger than the smallest hyperstrip/ellipsoid/annulus enclosing (at least) k^+ of the points in P. We present a con-

*Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland. Email: mount@cs.umd.edu. The support of the National Science Foundation under grant CCR–9310705 is gratefully acknowledged.

†Center for Automation Research, University of Maryland, College Park, Maryland, and Space Data and Computing Division, NASA Goddard Space Flight Center, Greenbelt, Maryland. Email: nathan@cfar.umd.edu.

‡The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland. Email: christine.piatko@jhuapl.edu.

§Department of Computer Science, University of the District of Columbia, Washington, DC, and Center for Automation Research, University of Maryland, College Park, Maryland. Email: ruth@cfar.umd.edu. The support of the National Science Foundation under ROA supplement CCR–9310705 is gratefully acknowledged.

¶Department of Computer Science and Information Systems, The American University, Washington, DC. Email: awu@american.edu.

[1] In contrast, see, e.g., [18, 3, 9] for efficient LMS computation in the plane.

ceptually simple algorithmic framework to compute the above described quantile approximation. Given the point set and the quantile error, we compute an epsilon-approximation for an appropriately defined range and an appropriately chosen value of epsilon. We then apply a simple, brute-force algorithm to solve the problem on the approximation. Drawing on a result by Matoušek [6], we show that an exact solution to the epsilon-approximation provides a quantile approximation to the original problem.

Our main result states that (quantile) approximations for the LMS and MVE estimators (for fixed d), as well as a (quantile) approximation for the MVA estimator in the plane, can be computed in linear time.

Key words: Robust estimation, least median-of-squares regression, minimum enclosing disk, minimum volume ball/ellipsoid/annulus estimator, epsilon-approximation, approximation algorithms.

References

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

[2] B. Chazelle and J. Matoušek. On linear-time deterministic algorithms for optimization problems in fixed Dimension. In *Proc. Fourth ACM-SIAM Symp. Disc. Alg.*, Austin, Texas, January 1993, 281–289.

[3] H. Edelsbrunner and D. L. Souvaine. Computing median-of-squares regression lines and guided topological sweep. *J. Amer. Statist. Assoc.*, 85:115–119, 1990.

[4] A. Efrat, M. Sharir, and A. Ziv. Computing the smallest k-enclosing circle and related problems. *Comput. Geom. Theory Appl.*, 4:119–136, 1994.

[5] J. García and P. A. Ramos. Fitting a set of points by a circle. In *Proc. Thirteenth ACM Symp. Comp. Geom.*, Nice, France, June 1997, 139–146.

[6] J. Matoušek. Efficient partition trees. *Discrete Comput. Geom.*, 8:315–334, 1992.

[7] J. Matoušek. On enclosing k points by a circle. *Inf. Proc. Lett.*, 53:217–221, 1995.

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

[9] D. M. Mount, N. S. Netanyahu, K. Romanik, R. Silverman, and A. Wu. A practical approximation algorithm for the LMS line estimator. In *Proc. Eighth ACM-SIAM Symp. Disc. Alg.*, New Orleans, Louisiana, January 1997, 473–482.

[10] J. O'Rourke. Finding minimal enclosing boxes. *Int'l. J. Comp. Inf. Sci.*, 14:183–199.

[11] J. O'Rourke, A. Aggarwal, S. Maddila, and M. Baldwin. An optimal algorithm for finding minimal enclosing traingles. *J. Algs.*, 7:258–269, 1986.

[12] P. J. Rousseeuw. Least median-of-squares regression. *J. Amer. Statist. Assoc.*, 79:871–880, 1984.

[13] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[14] S. Schonherr. *Berechnung kleinster ellipsoide um punktemengen.* Diploma thesis, Free University Berlin, 1994.

[15] M. Segal and K. Kedem. Enclosing k points in the smallest axix parallel rectangle. *Inf. Proc. Lett.*, 65:95–99, 1998.

[16] B. W. Silverman and D. M. Titterington. Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.*, 1:401–409, 1980.

[17] S. Skyum. A simple algorithm for computing the smallest enclosing circle. Inf. Proc. Lett., 37:121–125, 1991.

[18] D. L. Souvaine and J. M. Steele. Time- and space- efficient algorithms for least median of squares regression. *J. Amer. Statist. Assoc.*, 82:794–801, 1987.

[19] A. J. Stromberg. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Sci. Comp.*, 14:1289–1299, 1993.

[20] 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.