Regression Depth and Center Points[1]

Nina Amenta[2] Marshall Bern[3]
David Eppstein[4] Shang-Hua Teng[5]

Robust statistics [1,6] has attracted much attention recently within the computational geometry community due to the natural geometric formulation of many of its problems. In contrast to least-squares regression, in which measurement error is assumed to be normally distributed, robust estimators allow some of the data to be affected by completely arbitrary errors; the *breakdown value* of an algorithm measures how many erroneous data points can be tolerated.

Recently, Rousseeuw and Hubert [2, 4, 5] introduced *regression depth* as a quality measure for robust linear regression: in statistical terminology, the *regression depth* of a hyperplane H is the smallest number of residuals that need to change sign to make H a nonfit. This definition has convenient statistical properties such as invariance under affine transformations; hyperplanes with high regression depth behave well in general error models, including skewed or heteroskedastic error distributions.

Geometrically, the regression depth of a hyperplane is the minimum number of points intersected by the hyperplane as it undergoes any continuous motion taking it from its initial position to vertical. The *contractible hull* of a data set is the set of hyperplanes having nonzero regression depth. In the dual setting of hyperplane arrangements, the *directional depth* of a ray is the number of planes that are touched by or parallel to the ray; the *hyperplane depth* of a point is the minimum directional depth of any ray containing the point, and the contractible hull of an arrangement is the set of points not in the interior of an infinite cell. Standard techniques of projective duality transform any statement about regression depth to a mathematically equivalent statement about hyperplane depth and vice versa.

Rousseeuw and Hubert [4, 5] showed that the regression depth of n points in d dimensions is upper bounded by $\lceil n/(d+1)\rceil$; i.e., there exist point sets for which no hyperplane has regression depth larger than this bound. For bivariate data, they found a simple linear-time construction, the *catline*, which achieves the optimal $\lceil n/3\rceil$ bound. These facts, together with an analogy to *center points* (points such that any halfspace containing them also contains many data points), led to the following conjectures:

Conjecture 1 (Rousseeuw and Hubert)**.** *For any d-dimensional set of n points there exists a hyperplane having regression depth $\lceil n/(d+1)\rceil$.*

Conjecture 2 (Rousseeuw and Hubert)**.** *Any point set can be partitioned into $\lceil n/(d+1)\rceil$ subsets such that the intersection of the contractible hulls of the subsets is nonempty.*

Steiger and Wenger [8] made some progress on Conjectures 1 and 2: they found a constant c_d (depending on the dimension d) such that any point set can be partitioned into $c_d n$ subsets such that the subsets' contractible hulls have nonempty common intersection. Note that any hyperplane in this intersection must

[1] The full version of this paper is on the xxx.lanl.gov e-print server, cs.CG/9809037.

[2] Univ. of Texas, Austin, Dept. of Comp. Sci., amenta@cs.utexas.edu, http://www.cs.utexas.edu/users/amenta/
[3] Xerox Palo Alto Research Ctr., bern@parc.xerox.com, http://www.parc.xerox.com/csl/members/bern/
[4] Univ. of California, Irvine, Dept. of Inf. and Comp. Sci., eppstein@ics.uci.edu, http://www.ics.uci.edu/~eppstein/
[5] Univ. of Illinois, Urbana-Champaign, Dept. of Comp. Sci., steng@cs.uiuc.edu, http://www-sal.cs.uiuc.edu/~steng/

have regression depth at least $c_d n$. Their value c_d is not stated explicitly, however it appears to be quite small: roughly $1/(6^{d^2}(d+1))$.

Questions of computational efficiency of problems related to regression depth have also been studied. Rousseeuw and Struyf [7] described algorithms for testing the regression depth of a given hyperplane. The same paper also considers algorithms for testing the *location depth* of a point (its quality as a center point). One can find the hyperplane of greatest regression depth for a given point set in time $O(n^d)$ by a breadth first search of the dual hyperplane arrangement; standard ϵ-cutting methods [3] can be used to develop a linear-time approximation algorithm that finds a hyperplane with regression depth within a factor $(1-\epsilon)$ of the optimum in any fixed dimension. For bivariate data, van Kreveld, Mitchell, Rousseeuw, Sharir, Snoeyink, and Speckmann have recently found an algorithm for finding the optimum regression line in time $O(n \log^2 n)$ (Jack Snoeyink, personal communication).

Our main result is to prove the truth of Conjecture 1. We do this by finding a common generalization of location depth and regression depth that formalizes the analogy between these two concepts: the *crossing distance* between a point and a plane is the smallest number of sites crossed by the plane in any continuous motion from its initial location to a location incident to the point, the location depth of a point is just its crossing distance from the plane at infinity, and the regression depth of a plane is just its crossing distance from the point at vertical infinity. We then prove the conjecture by using Brouwer's fixed point theorem to find a projective transformation that maps the point at vertical infinity to a center point of the transformed sites; the inverse transformation maps the plane at infinity to a deep plane.

We also improve the partial result of Steiger and Wenger on Conjecture 2: we show that one can always partition a data set into $\lceil n/d(d+1) \rceil$ subsets with nonintersecting contractible hulls; we further improve this to $\lfloor (n+1)/6 \rfloor$ for $d = 3$. Our technique of projective transformation also sheds some light on issues of computational complexity: the two problems of testing regression depth and location depth considered by Rousseeuw and Struyf are in fact computationally equivalent. Known NP-hardness results for center points then lead to the observation that testing regression depth is NP-hard for data sets of unbounded dimension.

References

[1] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: the Approach Based on Influence Functions*. Series in Probability and Mathematical Statistics. Wiley Interscience, 1986.

[2] M. Hubert and P. J. Rousseeuw. The catline for deep regression. *J. Multivariate Analysis* 66:270–296, 1998, http://win-www.uia.ac.be/u/statis/publicat/catline_abstr.html.

[3] K. Mulmuley and O. Schwarzkopf. Randomized algorithms. *Handbook of Discrete and Computational Geometry*, chapter 34, pp. 633–652. CRC Press, 1997.

[4] P. J. Rousseeuw and M. Hubert. Depth in an arrangement of hyperplanes. To appear in *Discrete & Computational Geometry*, http://win-www.uia.ac.be/u/statis/publicat/arrang_abstr.html.

[5] P. J. Rousseeuw and M. Hubert. Regression depth. To appear in *J. Amer. Stastical Assn.*, http://win-www.uia.ac.be/u/statis/publicat/rdepth_abstr.html.

[6] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Series in Applied Probability and Statistics. Wiley Interscience, 1987.

[7] P. J. Rousseeuw and A. Struyf. Computing location depth and regression depth in higher dimensions. To appear in *Statistics and Computing*, http://win-www.uia.ac.be/u/statis/publicat/compdepth_abstr.html.

[8] W. Steiger and R. Wenger. Hyperplane depth and nested simplices. *Proc. 10th Canad. Conf. Computational Geometry*. McGill Univ., 1998, http://cgm.cs.mcgill.ca/cccg98/proceedings/cccg98-steiger-hyperplane.ps.