Computing the Depth of a Flat

Marshall Bern* David Eppstein†

Abstract

We compute the regression depth of a k-flat in a set of n points in $\mathbb{R}^d$, in time $\mathcal{O}(n^{d-2} + n \log n)$ for $1 \leq k \leq d-2$. This contrasts with a bound of $\mathcal{O}(n^{d-1} + n \log n)$ when $k = 0$ or $k = d-1$.

1 Introduction

Regression depth was introduced by Hubert and Rousseeuw [7] as a distance-free quality measure for linear regression. The depth of a hyperplane is the minimum number of points crossed in any continuous motion taking the hyperplane to a vertical hyperplane (a "nonfit"). The deepest hyperplane provides a good fit even in the presence of skewed or data-dependent errors, and is robust against a constant fraction of arbitrary outliers.

Due to its combinatorial nature, regression depth leads to many interesting geometric and algorithmic problems. A simple construction called the *catline* provides a line of depth $\lceil n/3 \rceil$ for n points in the plane [3]. The catline's depth bound is best possible, and more generally in $\mathbb{R}^d$ the best depth bound is $\lceil n/(d+1) \rceil$ [1, 5]. On the algorithmic side, the deepest line in the plane can be found in time $\mathcal{O}(n \log n)$ [4]. In higher dimensions, the fastest known exact algorithm takes time $\mathcal{O}(n^d)$, and ϵ-cuttings can be used to obtain an $\mathcal{O}(n)$-time $(1+\epsilon)$-approximation to the maximum depth [9].

In previous work [2], we generalized depth to *multivariate regression*, fitting points in $\mathbb{R}^d$ by affine subspaces with dimension $k < d-1$ (*k-flats* for short). The generalization is most natural in a dual setting. The projective dual of a point set is a hyperplane arrangement, and linear regression dualizes to finding a central point in an arrangement. If the *depth* of point p is the minimum number of arrangement hyperplanes crossed by any line segment from p to the plane at infinity, then the regression depth of a hyperplane is exactly the depth of its dual point in the dual arrangement. Thus we defined the *crossing distance* between two flats in an arrangement to be the fewest crossings on any line segment having one endpoint on each flat, and defined the *regression depth of a k-flat* to be the crossing distance between its dual $(d-k-1)$-flat and a certain k-flat at *vertical infinity*. (This definition also subsumes the classical notion of *data depth* or *Tukey depth*.) We showed that deep flats always exist, meaning that for any point set in $\mathbb{R}^d$, there is always a k-flat of depth a constant fraction of n. Moreover, ϵ-cuttings can be used to obtain an $\mathcal{O}(n)$-time $(1+\epsilon)$-approximation for the deepest flat. The catline generalizes to give lines with depth $\lceil n/(2d-1) \rceil$, which is tight for $d \leq 3$ and would be tight for all d under a conjectured $\lceil n/((k+1)(d-k)+1) \rceil$ bound on maximum regression depth.

In this paper, we consider the problem of testing the depth of a given flat, or more generally the crossing depth of two flats. Rousseeuw and Struyf [8] studied similar problems for hyperplanes and points. The crossing distance between a point and a hyperplane can be found in time $\mathcal{O}(n^{d-1} + n \log n)$ by examining the arrangement's restriction to the hyperplane (as described later), and the same bound applies to testing the depth of a hyperplane or point. We show that, in contrast, the depth of a flat of any other dimension can be found in randomized time $\mathcal{O}(n^{d-2} + n \log n)$. More generally, the crossing distance between a j-flat and a k-flat can be found in time $\mathcal{O}(n^{j+k-1} + n \log n)$ when $1 \leq j, k \leq d-2$.

We omit many details in this extended abstract; see the longer version of this paper at http://arXiv.org/abs/cs.CG/0009024 for a more complete exposition.

2 Reduction to Covering

As we now show, crossing distance can be reduced to finding a minimally covered point in a certain family of sets. Suppose we are given a hyperplane arrangement, a j-flat $\mathcal{F}_1$, and a k-flat $\mathcal{F}_2$ in $\mathbb{R}^d$. We wish to determine the line segment, having one endpoint on each flat, that crosses as few arrangement hyperplanes as possible. We first parametrize these line segments. Without loss of generality the two flats do not meet (else the crossing distance is zero) so any pair of points from $\mathcal{F}_1 \times \mathcal{F}_2$ determines a unique line. The pair divides the line into two line segments (one through infinity), so we need to augment each point of $\mathcal{F}_1 \times \mathcal{F}_2$ by an additional bit of information to specify each possible line segment. We do this topologically: $\mathcal{F}_1$ is a projective space, having as its double cover a j-sphere $\mathcal{S}_1$, and similarly the double cover of $\mathcal{F}_2$ is a k-sphere $\mathcal{S}_2$. The product $\mathcal{S}_1 \times \mathcal{S}_2$ supplies two extra bits of information per point, and there is a continuous two-to-one map from $\mathcal{S}_1 \times \mathcal{S}_2$ to the line segments connecting the two flats.

*Xerox PARC, 3333 Coyote Hill Rd., Palo Alto, CA 94304

†Univ. of California, Irvine, Dept. Inf. & Comp. Sci., Irvine, CA 92697.
Work done in part while visiting Xerox PARC and supported in part by NSF grant CCR-9912338.

Now consider subdividing $\mathcal{S}_1 \times \mathcal{S}_2$ according to whether the corresponding line segments cross or do not cross a hyperplane $\mathcal{H}$ of the arrangement. The boundary between crossing and non-crossing line segments is formed by the segments with an endpoint on a great sphere formed by intersecting $\mathcal{H}$ with $\mathcal{S}_1$ or $\mathcal{S}_2$. The line segments that cross $\mathcal{H}$ therefore correspond to a set $(\mathcal{H}_1 \times \overline{\mathcal{H}_2}) \cup (\overline{\mathcal{H}_1} \times \mathcal{H}_2)$, where $\mathcal{H}_i$ is a hemisphere bounded by the intersection of $\mathcal{H}$ with $\mathcal{S}_i$. The line segment crossing the fewest hyperplanes then simply corresponds to the point in the fewest such sets, and since the union in each such set is disjoint we have the following result.

LEMMA 2.1. *Computing crossing distance between flats $\mathcal{F}_1$ and $\mathcal{F}_2$ is equivalent to finding a point in $\mathcal{S}_1 \times \mathcal{S}_2$ covered by the fewest of a family of sets of the form $\mathcal{H}_1 \times \mathcal{H}_2$.*

As a special case, the crossing distance between a point and a hyperplane can be found as the point covered by the fewest hemispheres of a single $(d-1)$-sphere, justifying the $\mathcal{O}(n^{d-1} + n \log n)$ time bound claimed above.

LEMMA 2.2. *Given an arrangement of hyperplanes in $\mathbb{R}^d$, we can produce a recursive decomposition of $\mathbb{R}^d$, with high probability in time $\mathcal{O}(n^d + n \log n)$, such that any halfspace bounded by an arrangement hyperplane has (with high probability) a representation as a disjoint union of decomposition cells with $\mathcal{O}(n^{d-1} + \log n)$ ancestors.*

Proof. We apply a randomized incremental arrangement construction algorithm. Each cell in the recursive decomposition is an arrangement cell at some stage of the construction. The bound on the representation of a halfspace comes from applying the methods of [6, pp. 120–123] to the zone of the boundary hyperplane. □

The same method applies essentially without change to spheres and hemispheres, so we can apply it to the sets occurring in Lemma 2.1. Each product of hemispheres occurring in Lemma 2.1 can be represented as disjoint unions of $\mathcal{O}(n^{j+k-2})$ products of cells in the product of the two recursive decompositions formed by applying Lemma 2.2 to $\mathcal{S}_1$ and $\mathcal{S}_2$. Since there are $\mathcal{O}(n)$ products of hemispheres, we have overall $\mathcal{O}(n^{j+k-1})$ products of cells.

3 The Algorithm

Our crossing distance algorithm performs a depth-first traversal of the recursive decomposition for $\mathcal{S}_1$, while maintaining a number for each cell of the decomposition of $\mathcal{S}_2$. This number measures the fewest $\mathcal{H}_2$ hemispheres covering some point in that cell, where the $\mathcal{H}_2$ hemispheres come from pairs $\mathcal{H}_1 \times \mathcal{H}_2$ for which $\mathcal{H}_1$ covers the current cell in the traversal of $\mathcal{S}_1$. These numbers are computed by taking the minimum number for the cell's two children and adding the number of hemispheres whose decomposition uses that cell directly. When the traversal visits a cell in $\mathcal{S}_1$, we determine the set of hemispheres whose decomposition uses that cell, and update the numbers for the ancestors of cells covering the corresponding hemispheres in $\mathcal{S}_2$. Each hemisphere product leads to $\mathcal{O}(n^{j+k-2})$ update steps, so the total time for this traversal is $\mathcal{O}(n^{j+k-1})$. We also maintain the overall minimum covering seen so far, and take the minimum with the number at the root of the decomposition of $\mathcal{S}_2$ whenever the traversal reaches a leaf in the decomposition of $\mathcal{S}_1$.

When one flat (say $\mathcal{F}_1$) is a line, this method's time includes an unwanted logarithmic factor. To avoid this, we traverse $\mathcal{S}_1$ directly instead of its hierarchical decomposition. We use the same data structure for $\mathcal{S}_2$, and when the traversal reaches an endpoint of an interval $\mathcal{H}_1$, we update the cells for the corresponding hemisphere $\mathcal{H}_2$.

We summarize our results.

THEOREM 3.1. *The crossing distance between a j-flat and a k-flat can be found in time $\mathcal{O}(n^{j+k-1} + n \log n)$ with high probability for $1 \le j, k$. The depth of a single k-flat in $\mathbb{R}^d$ can be found in time $\mathcal{O}(n^{d-2} + n \log n)$ with high probability for $1 \le k \le d-2$.*

It is likely that ϵ-cuttings can be used to derandomize this result. If both flats are lines, we can substitute segment trees for the randomized hierarchical decomposition.

References

[1] N. Amenta, M. Bern, D. Eppstein, and S.-H. Teng. Regression depth and center points. *Discrete & Computational Geometry* 23(3):305–323, 2000, cs.CG/9809037.

[2] M. Bern and D. Eppstein. Multivariate regression depth. *Proc. 16th Symp. Computational Geometry*, pp. 315–321. ACM, June 2000, cs.CG/9912013.

[3] 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.

[4] S. Langerman and W. Steiger. An $\mathcal{O}(n \log n)$ algorithm for the hyperplane median in $\mathbb{R}^2$. *Proc. 11th Symp. Discrete Algorithms*, pp. 54–59. ACM and SIAM, January 2000.

[5] I. Mizera. On depth and deep points: a calculus. *Inst. Mathematical Statistics Bull.* 27(4), 1998, http://www.dcs.fmph.uniba.sk/∼mizera/PS/depthps.ps. Full version to appear in *Annals of Statistics*.

[6] K. Mulmuley. *Computational Geometry: An Introduction Through Randomized Algorithms*. Prentice-Hall, 1994.

[7] P. J. Rousseeuw and M. Hubert. Regression depth. *J. Amer. Statistical Assoc.* 94(446):388–402, June 1999, http://win-www.uia.ac.be/u/statis/publicat/rdepth_abstr.html.

[8] P. J. Rousseeuw and A. Struyf. Computing location depth and regression depth in higher dimensions. *Statistics and Computing* 8(3):193–203, August 1998, http://win-www.uia.ac.be/u/statis/publicat/compdepth_abstr.html.

[9] 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.gz.