# Computing the Depth of a Flat

Marshall Bern

Xerox PARC

and

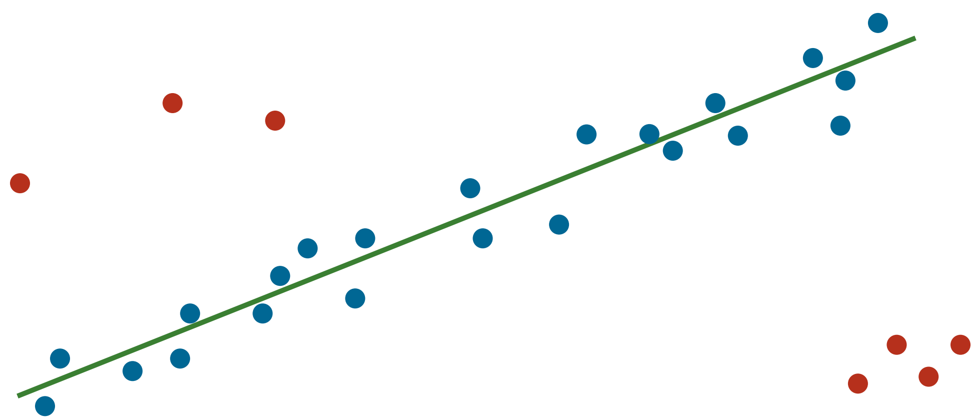
David Eppstein

UC Irvine

# Robust Regression

Given data with dependent and independent vars

Describe dependent vars as function of indep. ones



Should be robust against arbitrary outliers

Prefer distance-free methods for robustness against skewed and data-dependent noise

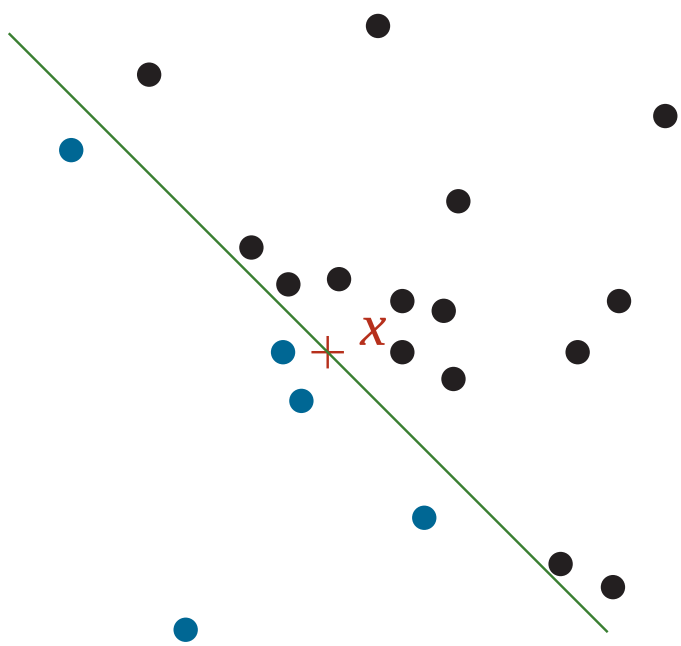
# Example: Data Depth

(no variables independent)

Fit a point to a cloud of data points

Depth of a fit $x$

= min # data points in halfspace containing $x$




Tukey median

= point with max possible depth

# Known Results for Data Depth

Tukey median has depth $\geq \left\lceil \frac{n}{d+1} \right\rceil$
[Radon 1946]

Deep (but not optimally deep) point can be found in time polynomial in $n$ and $d$
[Clarkson, Eppstein, Miller, Sturtivant, Teng 1996]

Deepest point can be found in time $O(n^d)$
(linear program with that many constraints)

Computing the depth of a point is
NP-complete for variable $d$ [Johnson & Preparata 1978]
$O(n^{d-1} + n \log n)$ for fixed $d$ [Rousseeuw & Struyf 1998]

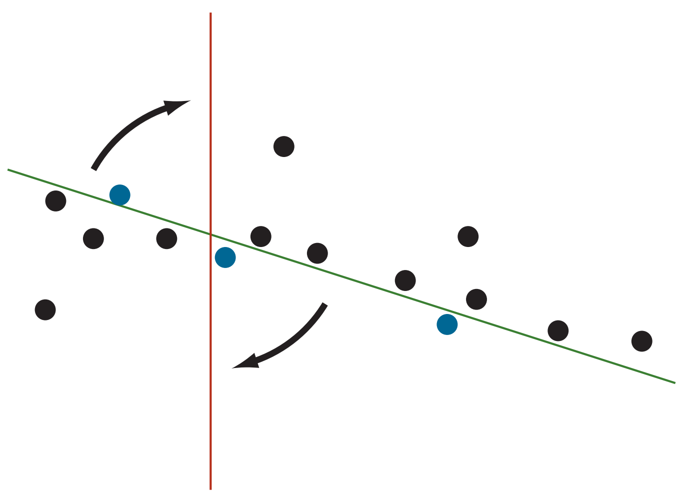
# Example: Regression Depth

(all but one variable independent)

[Hubert & Rousseeuw 1998]

Fit a hyperplane to a cloud of data points

Nonfit = vertical hyperplane
(doesn't predict dependent variable)



Depth of a fit = min # data points crossed
while moving to a nonfit

# Known Results for Regression Depth

Deepest hyperplane has depth $\geq \left\lceil \frac{n}{d+1} \right\rceil$

[Amenta, Bern, Eppstein, Teng 1998; Mizera 1998]

Deepest hyperplane can be found in time $O(n^d)$
(breadth first search in arrangement)

Planar deepest line can be found in $O(n \log n)$

[van Kreveld et al. 1999; Langerman & Steiger 2000]

Computing the depth of a hyperplane is
NP-complete for variable $d$ [Amenta et al. 1998]
$O(n^{d-1} + n \log n)$ for fixed $d$ [Rousseeuw & Struyf 1998]

# Multivariate Regression Depth

## (any number k of independent variables)

[Bern & Eppstein 2000]

Definition of depth for $k$-flat

Equals data depth for $k = 0$

Equals regression depth for $k = d - 1$

Deepest flat has depth $\Omega(n)$

Conjecture: depth $\geq \left\lceil \frac{n}{(k+1)(d-k)+1} \right\rceil$

true for $k = 0$, $k = 1$, $k = d - 1$

## New Results

Computing the depth of a $k$-flat is
$O(n^{d-2} + n \log n)$ when $0 < k < d - 1$

Saves a factor of $n$ compared to
similar results for regression depth, data depth

Deterministic $O(n \log n)$ for lines in space
$(k = 1, d = 3)$

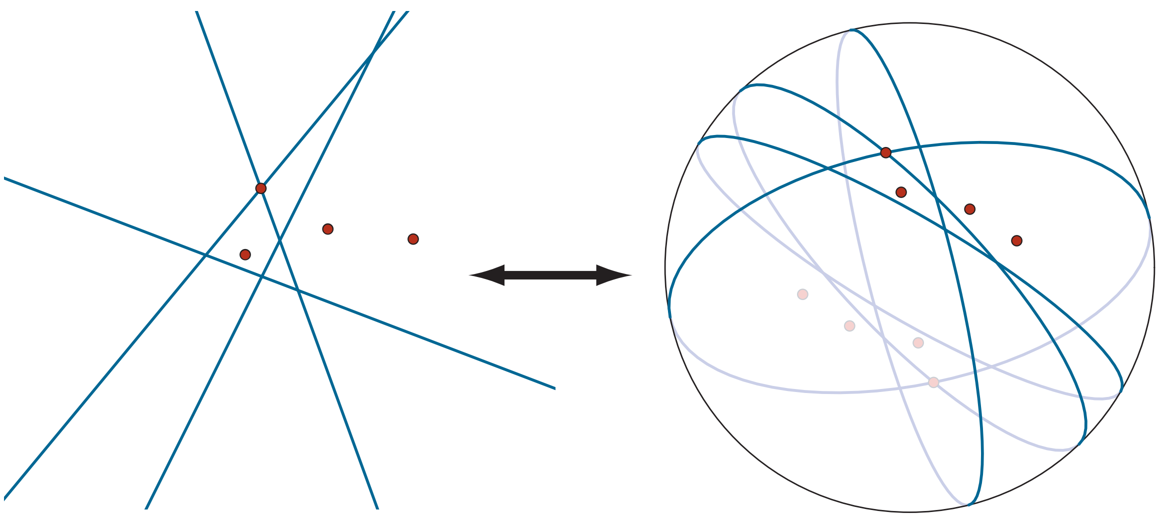
Randomized $O(n^{d-2})$ for all other cases

Likely can be derandomized using $\epsilon$-net techniques

# Projective Geometry

Augment Euclidean geom. by "points at infinity"
One infinite point per family of parallel lines
Set of infinite points forms "hyperplane at infinity"

Equivalently: view hyperplanes and points
as equators and pairs of poles on a sphere

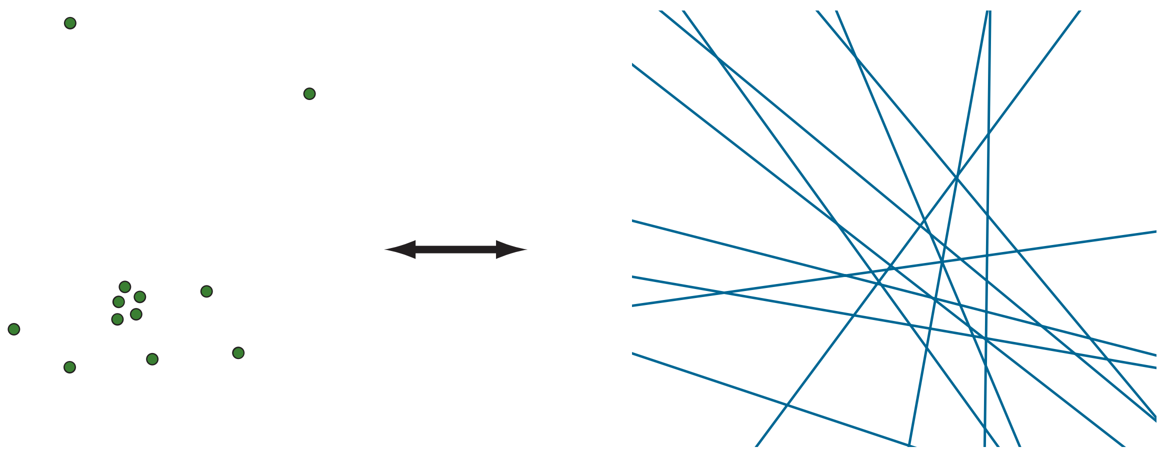


Nonfit = $k$-flat touching some
particular $(d - k - 1)$-flat at infinity

# Projective Duality

Incidence-preserving correspondence
between $k$-flats and $(d-k-1)$-flats

Cloud of data points becomes
arrangement of hyperplanes



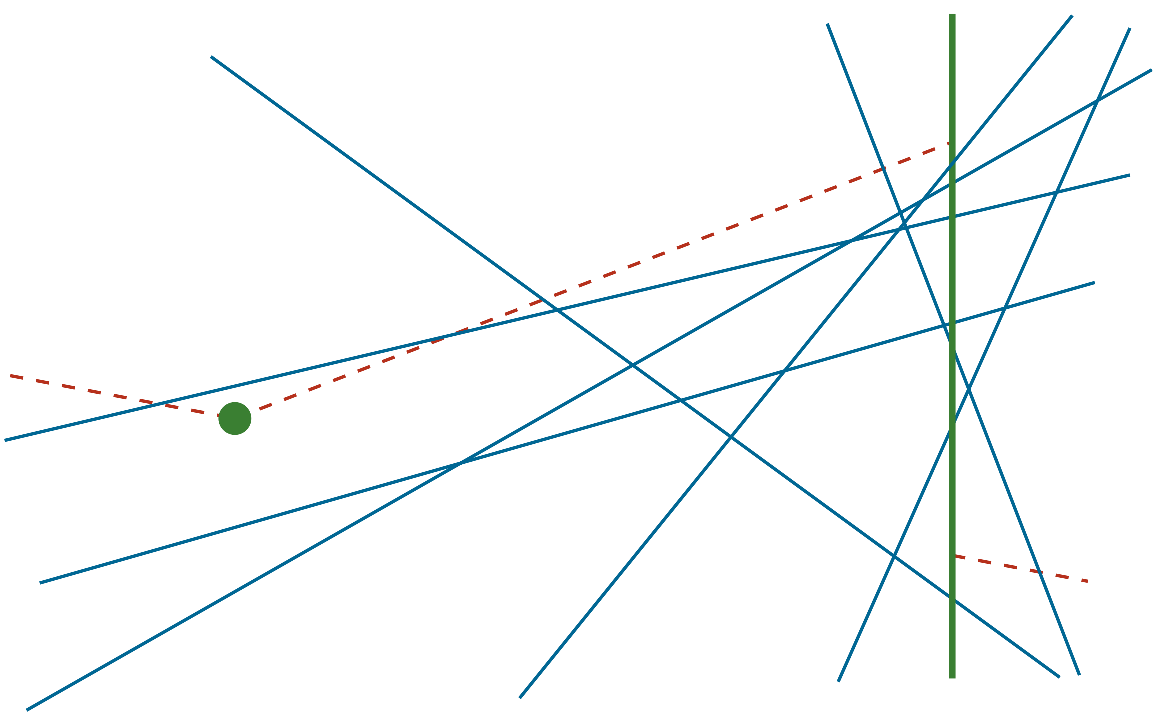
In coordinates (two dimensional case):

$$\begin{aligned} (a,b) &\mapsto y = ax + b \\ y = mx + c &\mapsto (-m, c) \end{aligned}$$

# Crossing Distance

Crossing distance between a $j$-flat and a $k$-flat in a hyperplane arrangement

= minimum number of hyperplanes crossed by any line segment connecting the two flats



(incl. line segments “through infinity”)

# Definition of Depth

Depth of a k-flat $F$

= crossing distance between dual($F$)
and dual(($d - k - 1$)-flat at infinity)

In primal space, minimum # data points
in double wedge bounded by $F$
and by (($d - k - 1$)-flat at infinity

Nonfit always has depth zero
(zero-length line seg, empty wedge)

## Parametrizing Line Segments

Let $F_1$, $F_2$ be flats (unoriented projective spaces)

If $F_1 \cap F_2 = \emptyset$, any pair $(p_1 \in F_1, p_2 \in F_2)$
determines unique line through them

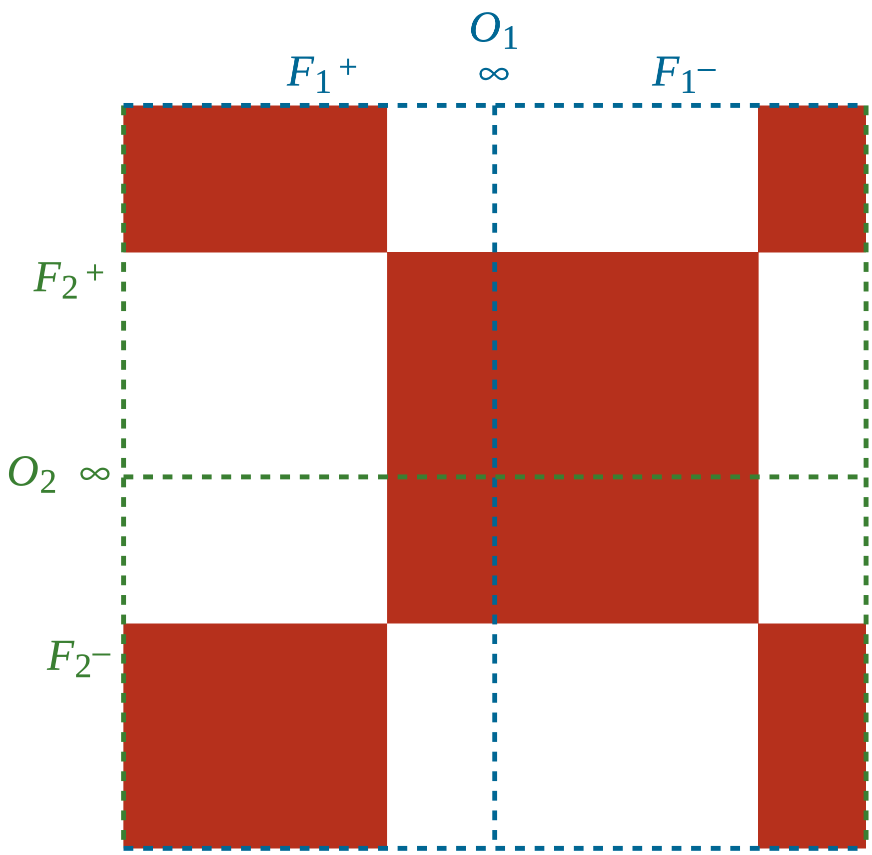
Need one more bit of information
to specify which of two line segments:
double cover (oriented proj. spaces) $O_1$, $O_2$

Two-to-one correspondence
$O_1 \times O_2 \mapsto$ line segments

# When does a segment cross a hyperplane?

Set of line segments crossing hyperplane $H$
is $h_1 \oplus h_2$ where $h_i$ are halfspaces in $O_i$
with boundary$(h_i) = H \cap O_i$

Or more simply, disjoint union of two sets
halfspace $\times$ halfspace




Line seg w/fewest crossings
= point covered fewest times by such sets

## Algorithm for $k = 1$, $d = 3$:

Want point in torus $O_1 \times O_2$
covered by fewest rectangles $h_1 \times h_2$

Sweep left-right (i.e., by $O_1$-coordinate),
use segment tree to keep track of
shallowest point in sweep line

Time: $O(n \log n)$

## Algorithm for Higher Dimensions:

Replace segment tree by history tree of
randomized incremental arrangement

Replace sweep by traversal of history tree

$O(n^{j+k-1})$ for crossing distance between
$j$-flat and $k$-flat $\Rightarrow$ $O(n^{d-2})$ for flat depth

# Conclusions

Presented efficient algorithm for testing depth

Many remaining open problems in algorithms, combinatorics, & statistics

- How to find deepest flat efficiently?
- What is its depth?
- Can we find deep flats efficiently when $d$ is variable?
- Do local optimization heuristics work?
- Are similar ideas of depth useful for nonlinear regression?