Robust Statistics and Arrangements

David Eppstein

Univ. of California, Irvine
School of Information and Computer Science

Two Worlds

Robust Statistics

Fit data points to model of data generation + errors

Error model allows arbitrary corruption (outliers)
Draw inferences about which points are corrupted
Fit of uncorrupt data should be unaffected by outliers

Fast algorithms needed to make methods practical

Computational Geometry

Fast algorithms for basic computational tasks on simple geometric objects, e.g. sets of data points

Powerful problem transformations
including projective duality: data points → arrangement

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

Slope selection

Regression depth

Multivariate regression

Outline

Projective duality and arrangements

Least median of squares

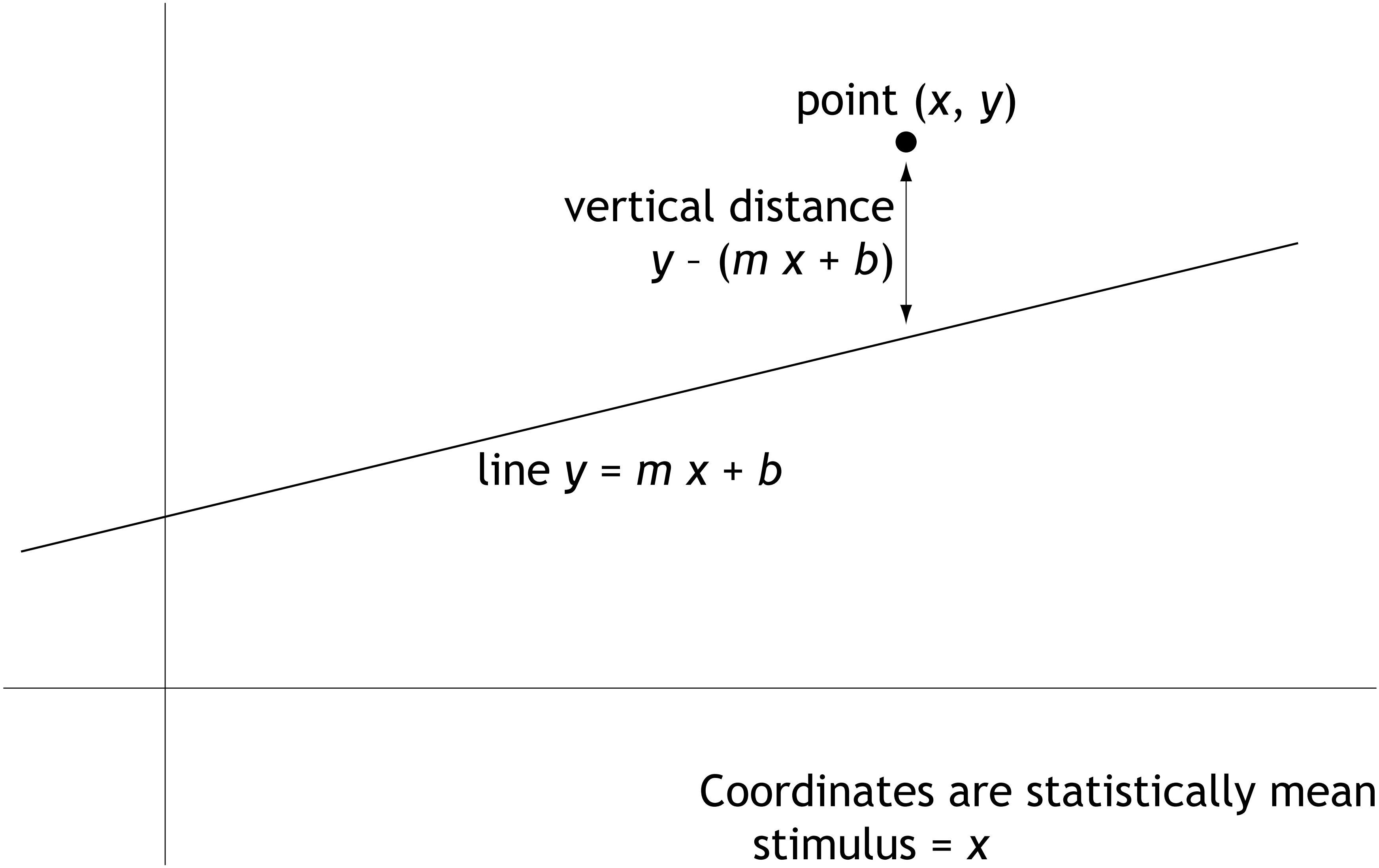
Least absolute deviation

Slope selection

Regression depth

Multivariate regression

Euclidean (Cartesian) Plane Geometry



point (x, y)

vertical distance

$y - (m\, x + b)$

line $y = m\, x + b$

Coordinates are statistically meaningful

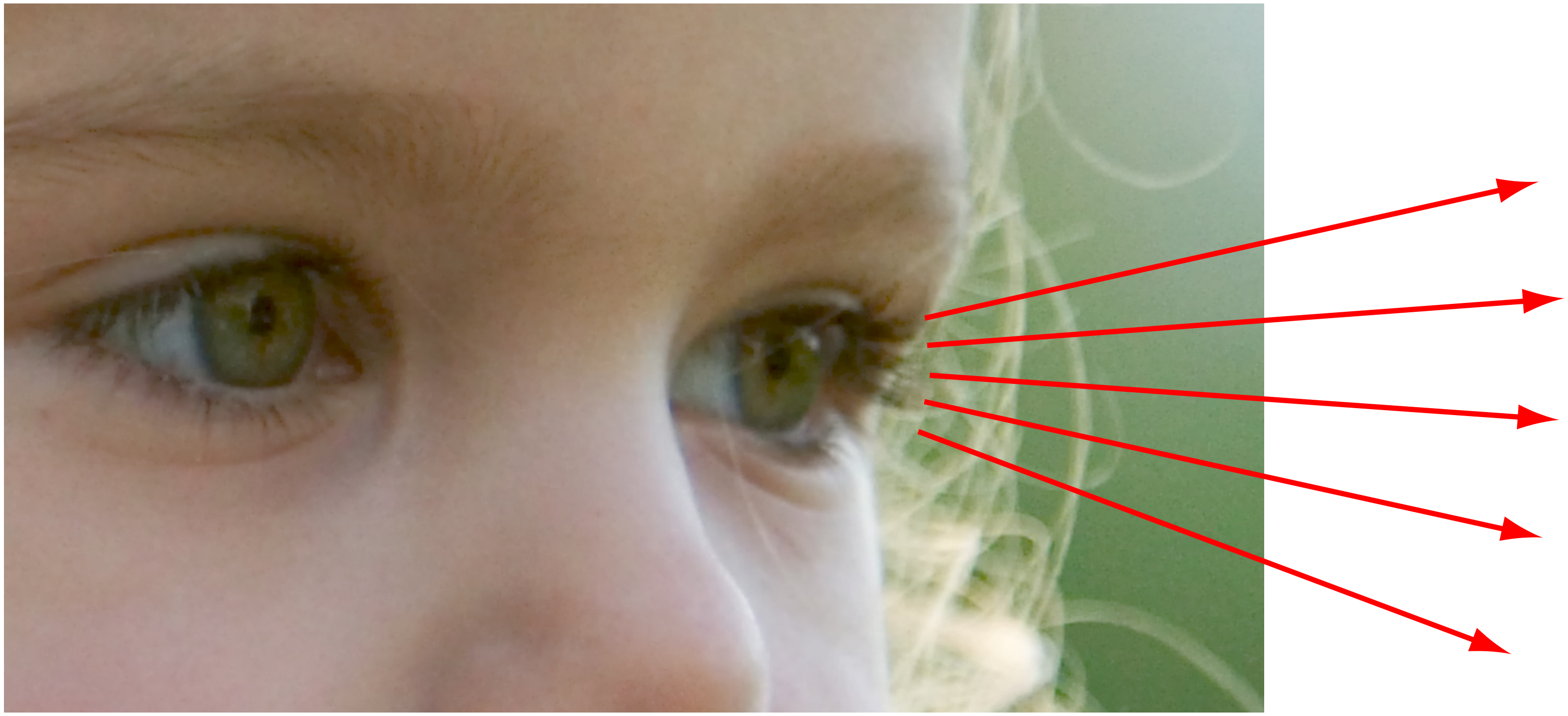
stimulus = x

response = y

residual deviation = vertical distance

Projective Plane Geometry

Mathematical model of human vision and visual perspective
(Brunelleschi, 1425, and many others since)



"projective point" = line in 3d space that passes through focal point

"projective line" = set of projective points touching a line in 3d
= plane in 3d space that passes through focal point

Projective plane coordinates

Place focal point at origin of 3d coordinate system
specify line through origin: triple of real numbers (x, y, z)
representing any Euclidean point on the line

A triple represents a projective point iff not all coordinates zero

Any projective point has many different representations:
(x, y, z) and (kx, ky, kz) represent same point if $k \neq 0$

Projective line = points satisfying linear equation $a\,x + b\,y + c\,z = 0$
vector dot product: $(a, b, c)\bullet(x, y, z) = 0$

So defined by triple of real numbers (a, b, c)
(a, b, c) and (ka, kb, kc) represent same line if $k \neq 0$

Conversions between Euclidean and Projective Geometry

Intuitively: what do you see when you look at a Euclidean plane?
intersect lines through origin with plane z=1

Euclidean to Projective:

$(x, y) \rightarrow (x, y, 1)$

$(m, b) \rightarrow (m, -1, \mathrm{b})$

Preserves point-line incidences:

$y = m\, x + b$ iff
$(m, -1, \mathrm{b}) \bullet (x, y, 1) = 0$

$(a, b, c) \bullet (x, y, z) = 0$ iff
$y/z = (-a/b)(x/z) + (-c/b)$

Projective to Euclidean:

$(x, y, z) \rightarrow (x/z, y/z)$

$(a, b, c) \rightarrow (-a/b, -c/b)$

Except, get division by zero for:

projective points with z = 0
"points at infinity"

projective lines with b = 0, a ≠ 0
"vertical lines"

projective line (0, 0, 1)
"the line at infinity"

Projective Duality (basic version)

Point-line incidence formula $(a, b, c) \bullet (x, y, z)$ is symmetric
so, changing which triple is a point and which is a line doesn't change incidence

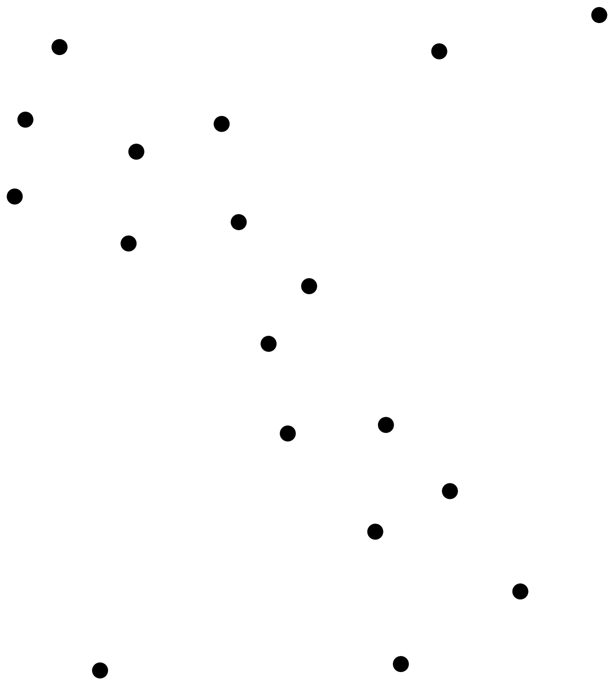
Switching points for lines and vice versa
preserves truth or correctness of any theorem or algorithm

Replacing problem by its dual
changes our visual intuition of the problem
without changing its underlying mathematics

Example:
algorithm for computing coordinates of line incident to two points
is identical to
algorithm for computing for coordinates of point incident to two lines

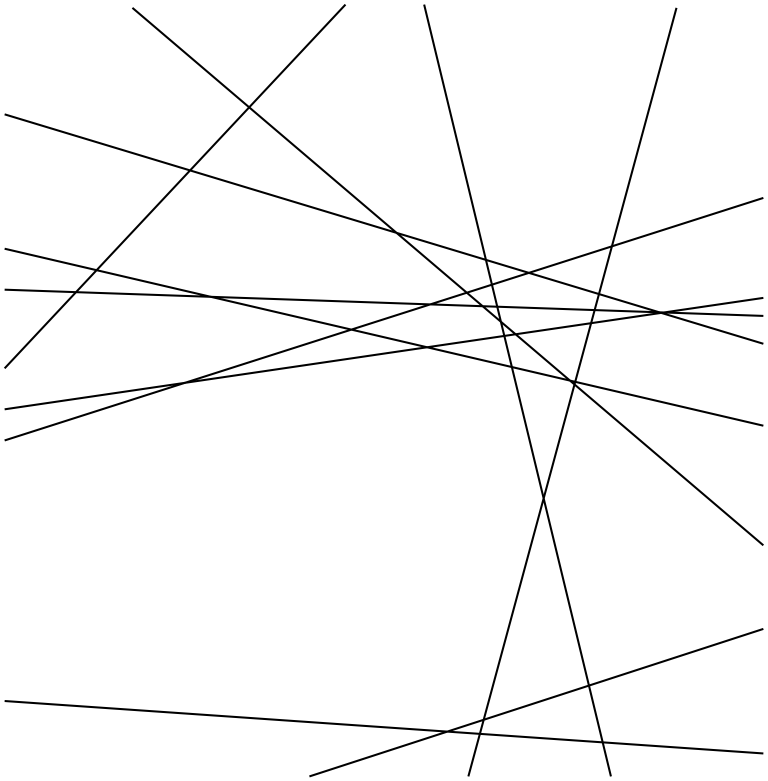
Projective duality (statistical version)

Map point (x, y, z) to line $(x, -z, -y)$
Map line (a, b, c) to point $(a, -c, -b)$



set of data points

becomes



arrangement of lines

Preserves vertical distances (deviation)
so points near a line become lines near a point

Higher dimensional projective geometry

k-flats in *d*-dimensional projective geometry
= (*k* + 1)-flats through origin in (*d* + 1)-dimensional Euclidean geometry

Euclidean-projective coordinate transformation formulas
are similar to planar formulas

Duality maps points to hyperplanes
and *k*-flats to (*d* – *k* – 1)-flats

Still possible to dualize preserving point-hyperplane vertical distance

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

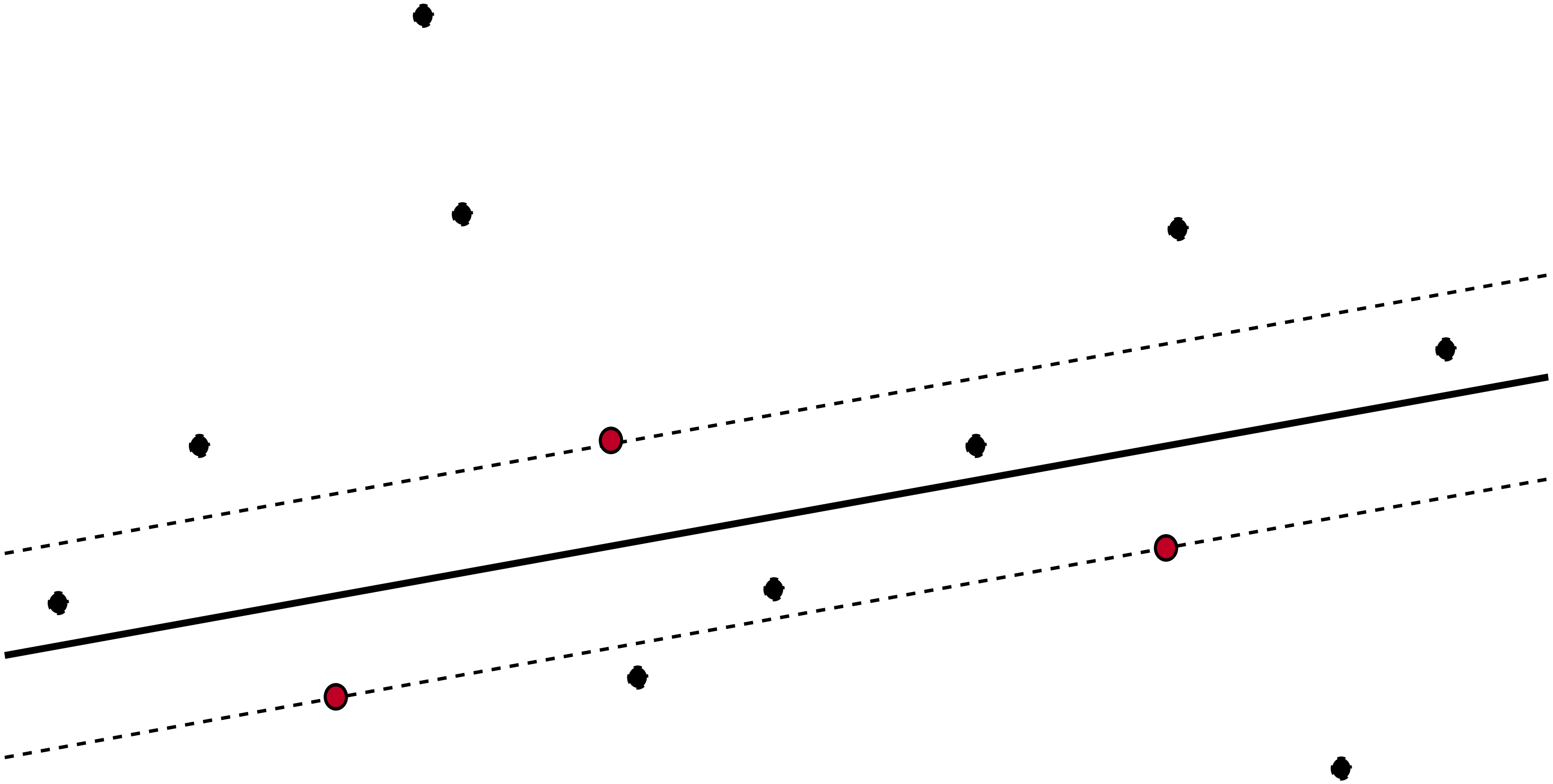
Slope selection

Regression depth

Multivariate regression

Least Median of Squares (LMS) Regression

[Rousseeuw, JASA 1984]



Quality of line = median deviation of data points

Choose line that minimizes median deviation

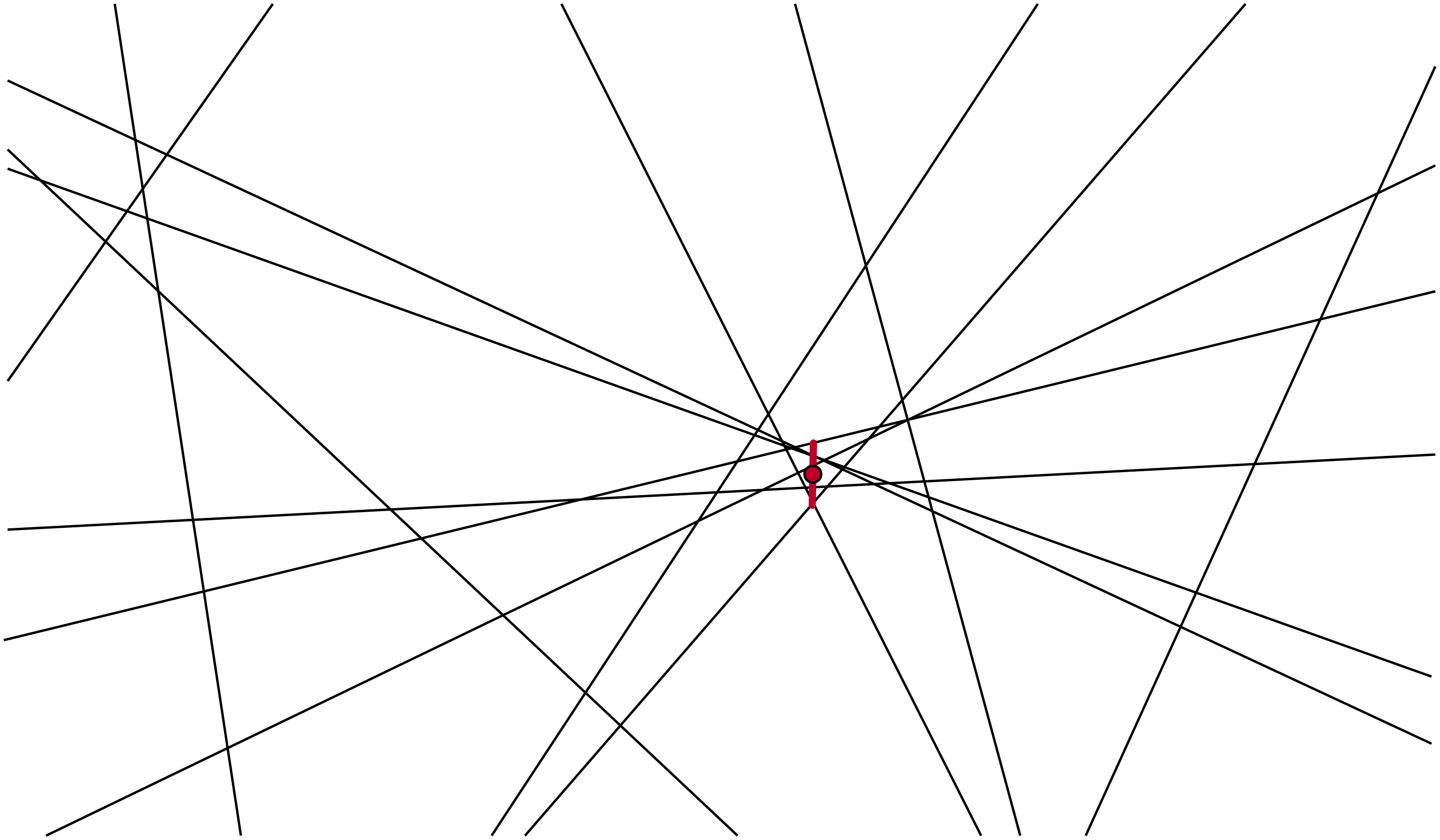
Equivalently, find minimum height strip containing ≥ $n/2$ points

If #outliers < (n – 1)/2, wrong lines can’t have small deviation

So breakdown point = 50%

LMS and Duality

Statement of problem involves only points, lines, and deviations
so: can be restated in projective dual



In line arrangement, find point minimizing median vertical distance
equivalently, find min-length vertical line segment touching ≥ n/2 lines

LMS Algorithm via Duality

Sweep vertical line left-right across arrangement
(plane sweep technique)

Maintain data structures:

Binary search tree of points where arrangement lines cross the sweep line, sorted by *y* coordinate

Priority queue of intersections between lines with adjacent crossings, sorted by *x* coordinate

Repeat:

1. Use priority queue to find next intersection point *p*
2. Use search tree to find line *L*, $n/2$ positions away from *p*
3. Test vertical line segment *p*-*L* as candidate solution
4. Update data structures

With further refinements (topological sweeping)
time is $O(n^2)$ [Edelsbrunner & Souvaine, JASA 1990]

Duality also leads to fast approx [Mount et al., SODA '90]

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

Slope selection

Regression depth

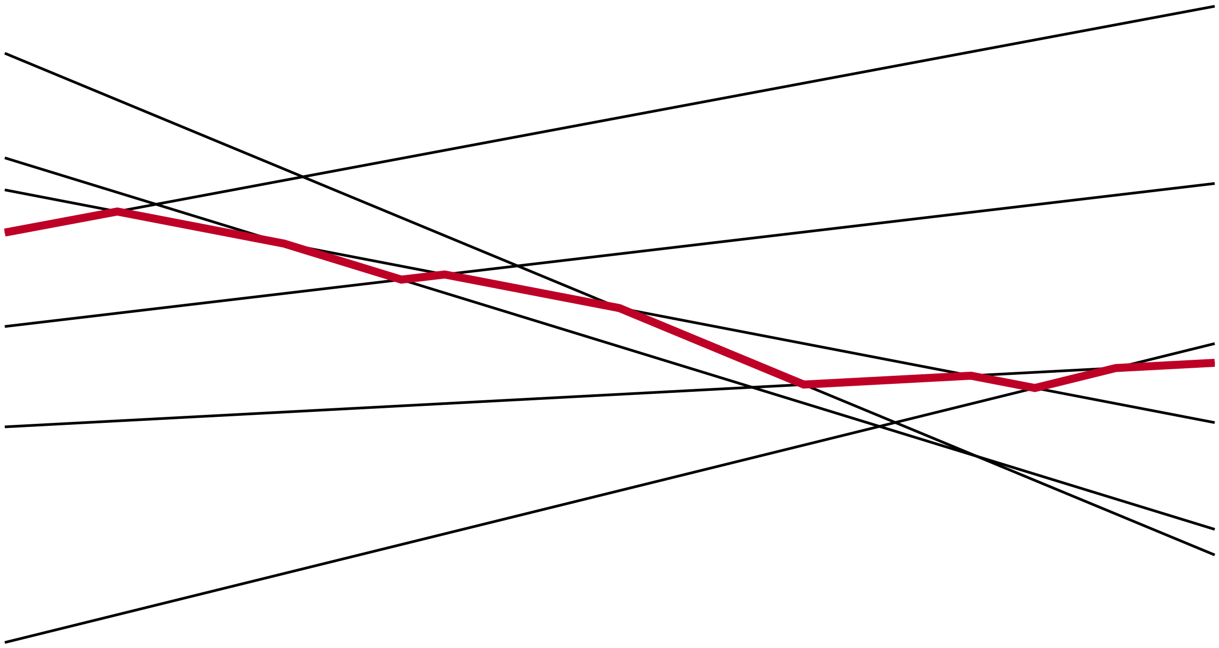
Multivariate regression

Least absolute deviation

Fit line or hyperplane minimizing mean absolute value (L_1 norm) of deviations

Dually, find point in line or hyperplane arrangement
minimizing sum of vertical distances to all hyperplanes

For 1d point set, optimum is median, so
in higher dimensions, lies on median level of arrangement

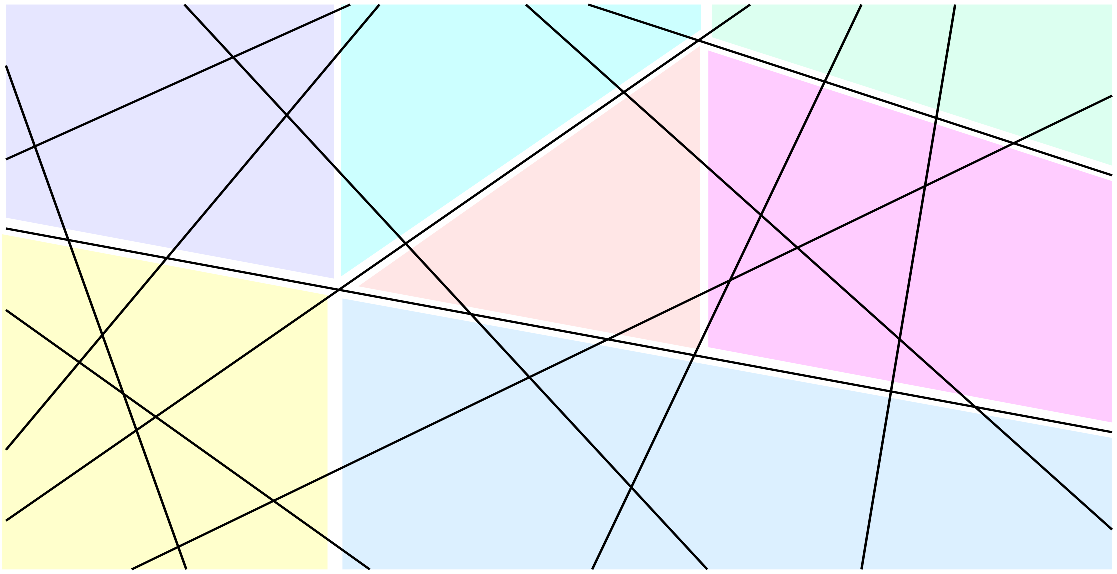


Not robust against outliers, but less sensitive than normal least squares

LAD Algorithm via Duality

Sum of vertical distances is piecewise-linear and convex

Partition arrangement into cells crossed by few hyperplanes
(epsilon-cutting technique)



Recurse on dimension to find best side of each cell wall
locate single cell containing the optimal point

Solve subproblem with constant fraction of original size

Linear time when dimension is constant

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

Slope selection

Regression depth

Multivariate regression

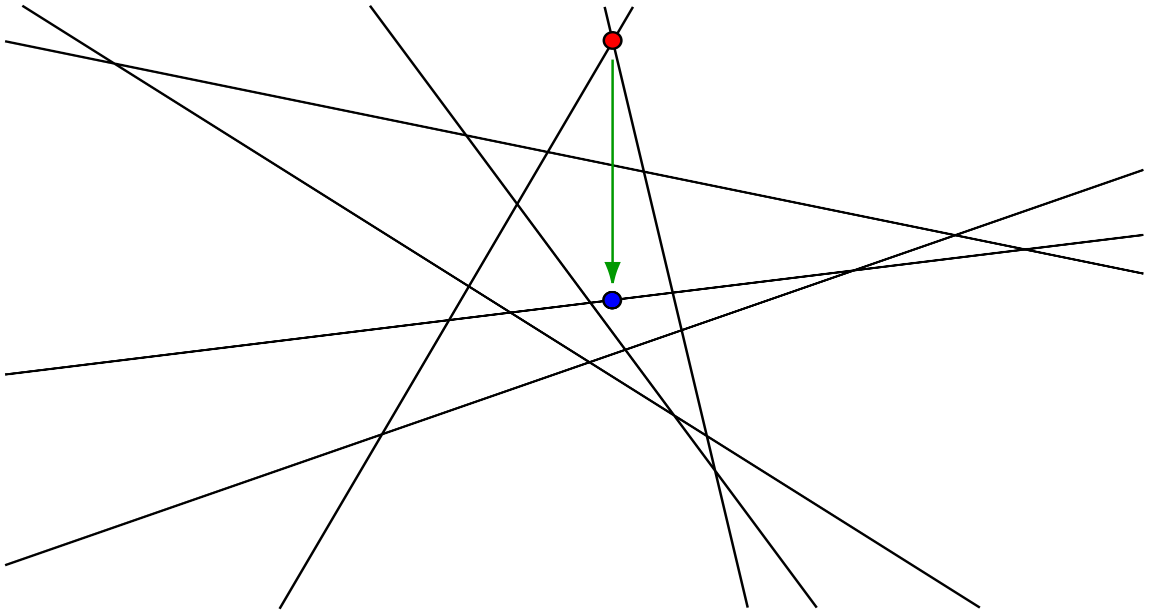
Slope Selection (Thiel-Sen Estimator)

Find median slope *m* among *n*(*n* – 1)/2 lines determined by pairs of points

Choose line with slope *m* splitting points evenly

Dually: find median *x*-coordinate of line arrangement intersections

Choose *y* so that point (*x*,*y*) is above and below equal numbers of lines



If #outliers < 0.29*n*, fewer than half the slopes determined by outliers

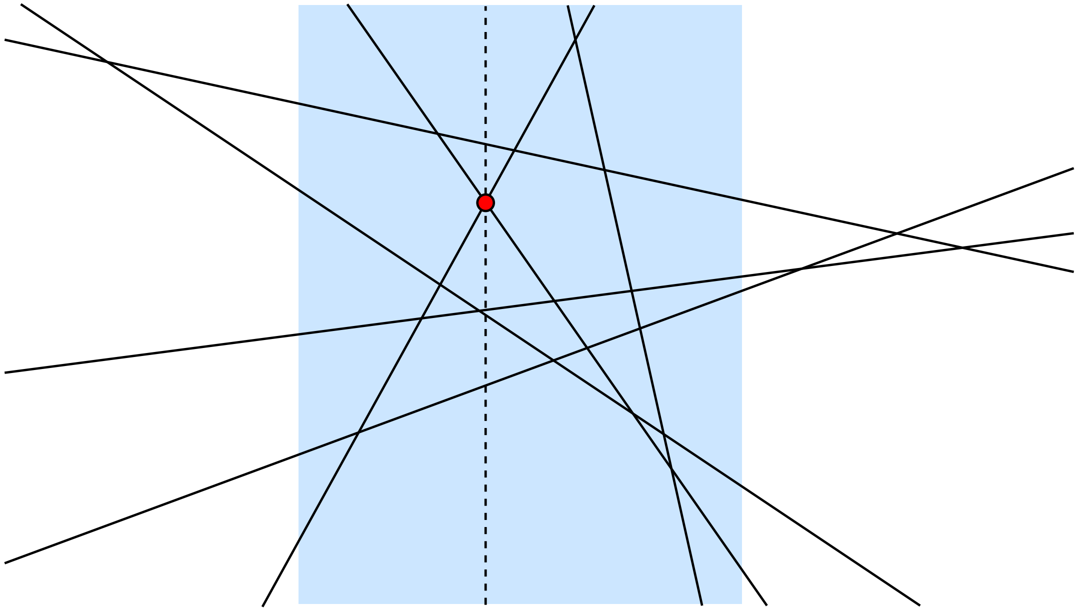
So breakdown point = 29%

Not as robust as LMS, but easier to compute

Slope Selection Algorithm, I

Restrict location of median arrangement intersection to vertical slab (initially, whole plane)

Choose pivot point within slab



Count intersections to left of pivot
(= count inversions in permutation, solve by sorting)

Split slab into two smaller pieces,
continue recursively in only one of the pieces

Slope Selection Algorithm, II

How to choose pivot?

- Simulate parallel sorting algorithm (parametric search)
  [Cole, Salowe, Steiger, and Szemerédi, SICOMP 1989]

- Randomly chosen arrangement intersection
  [Matousek, IPL 1991]
  [Dillencourt, Mount, and Netanyahu, IJCGA 1992]

- Expander graph
  [Katz and Sharir, IPL 1993]

- Epsilon-cutting
  [Brönnimann and Chazelle, CGTA 1998]

...also need some care to avoid doing O(log n) independent O(n log n) time sorting algorithms...

Result: O(n log n) time

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

Slope selection

Regression depth

Multivariate regression

Depth-based statistic

Idea: start with a combinatorial definition of a nonfit

To measure the quality of a fit F:
Remove minimum set S of points to make F a nonfit
breakdown point = depth of F = cardinality of S

Other good statistical properties (e.g. affine invariance)

Goals:

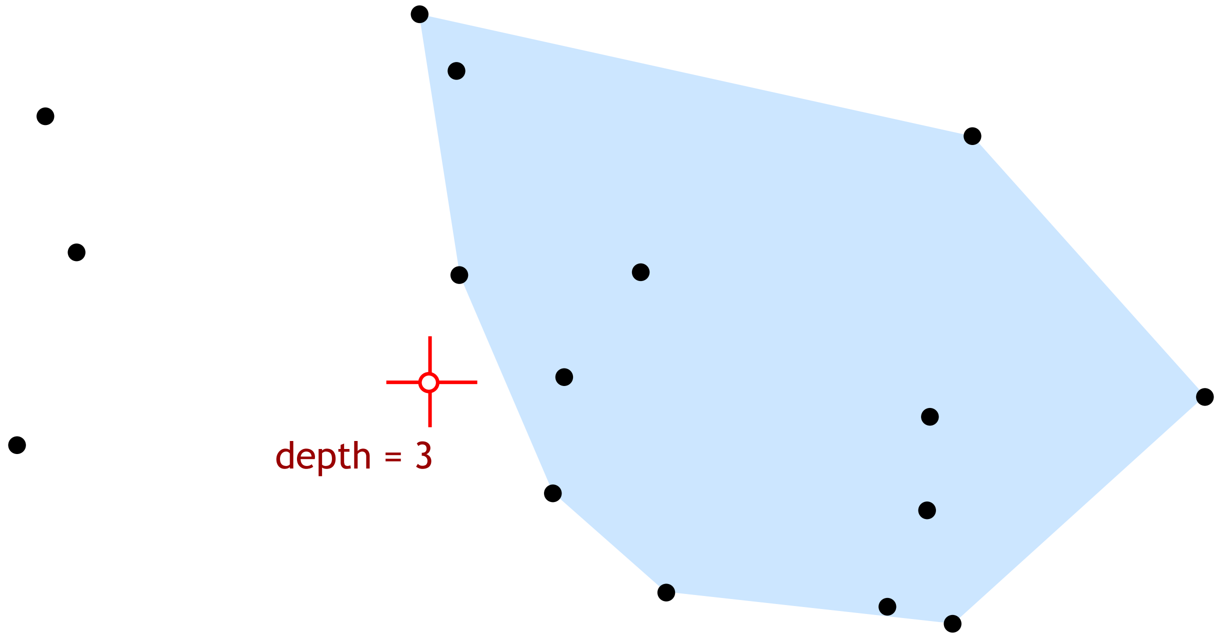
- Algorithm for finding deepest fit or approximating it
- Algorithm for computing depth of given fit
- Visualization of depth contours
- Proof that deepest fit has high depth
- Proof that some efficiently computable heuristic guarantees high depth

Classical example: Tukey depth

(for estimating single point given samples of that point + errors)

Nonfit = point outside convex hull of data

Depth(*p*) = min *k* s.t. *p* is outside hull of some set of *n* – *k* points



What's known about Tukey depth?

For any point set, maximum depth ≥ $n/(d+1)$

Can find point of depth ≥ $n/3$ in the plane
in time $O(n)$ [Jadhav and Mukhopadhay, SoCG 1993]

Find 2d deepest point in time $O(n \log^3 n)$ [Langerman & Steiger, STACS 2003]
3d deepest point in time $O(n^2 \log^7 n)$ [Matousek, DIMACS 1991]

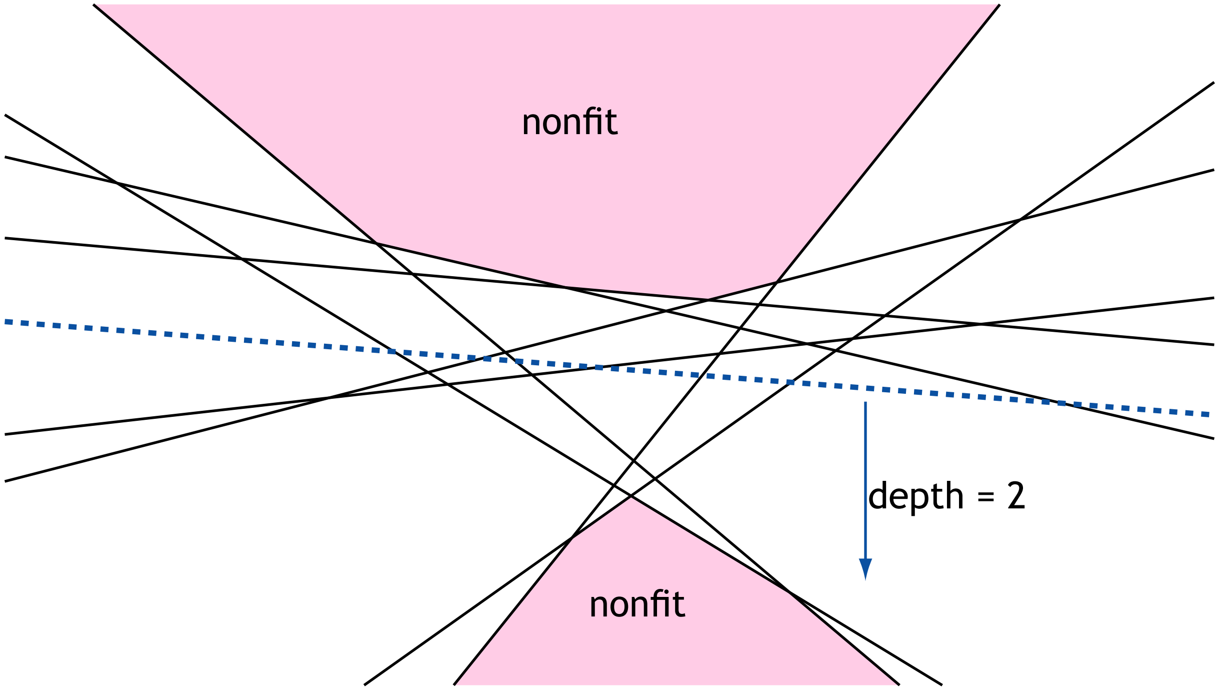
In any dimension d, can find point with depth
a constant fraction of n in time poly(n,d)
[Clarkson, E, Miller, Sturtivant, Teng, IJCGA 1996]

Many other results e.g. on depth contours

Tukey depth and projective duality

Point *p* is a nonfit iff
some line through *p* passes entirely above all data points
or some line through *p* passes entirely below all data points

In dual arrangement, line *L* = dual(*p*) is a nonfit iff
some point on *L* lies entirely above all arrangement lines
or some point on *L* lies entirely below all arrangement lines

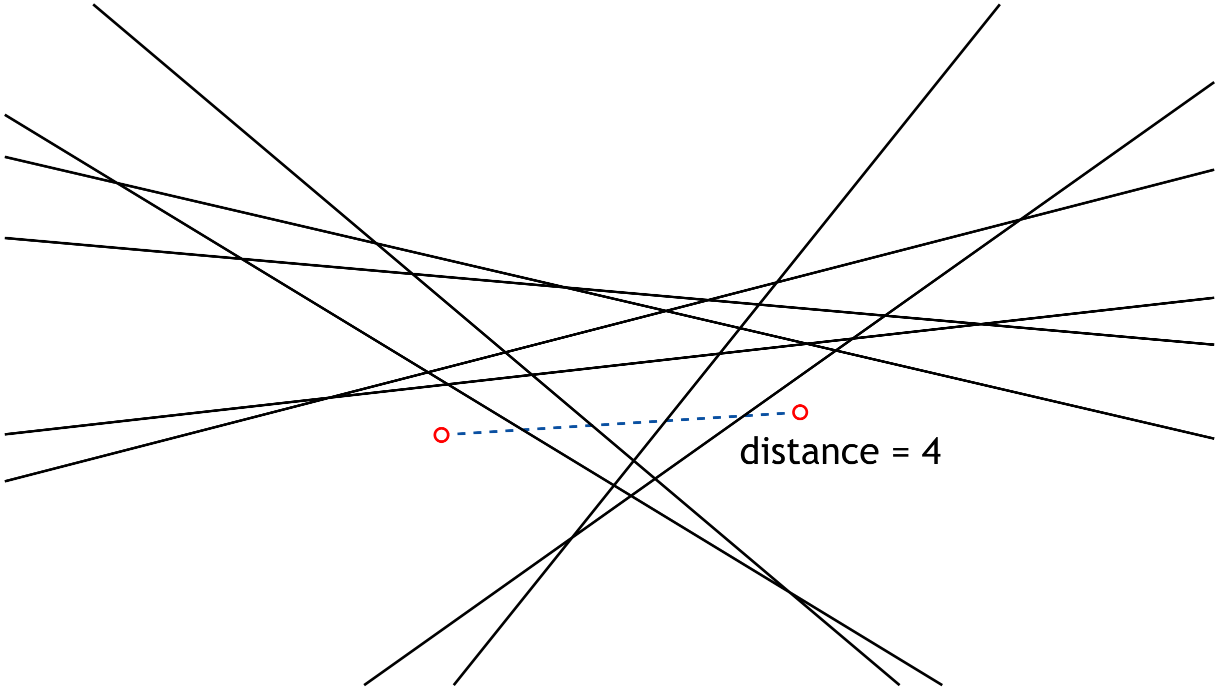


So depth(*L*) = min # lines separating it from vertical infinity

Distance in arrangements

distance(*p*,*q*) = # lines separating *p* from *q*
= min (# cells in path from *p* to *q*) – 1

If *S* and *T* are sets of points,
distance(*S*,*T*) = min(*s*,*t*) for *s* in *S* and *t* in *T*



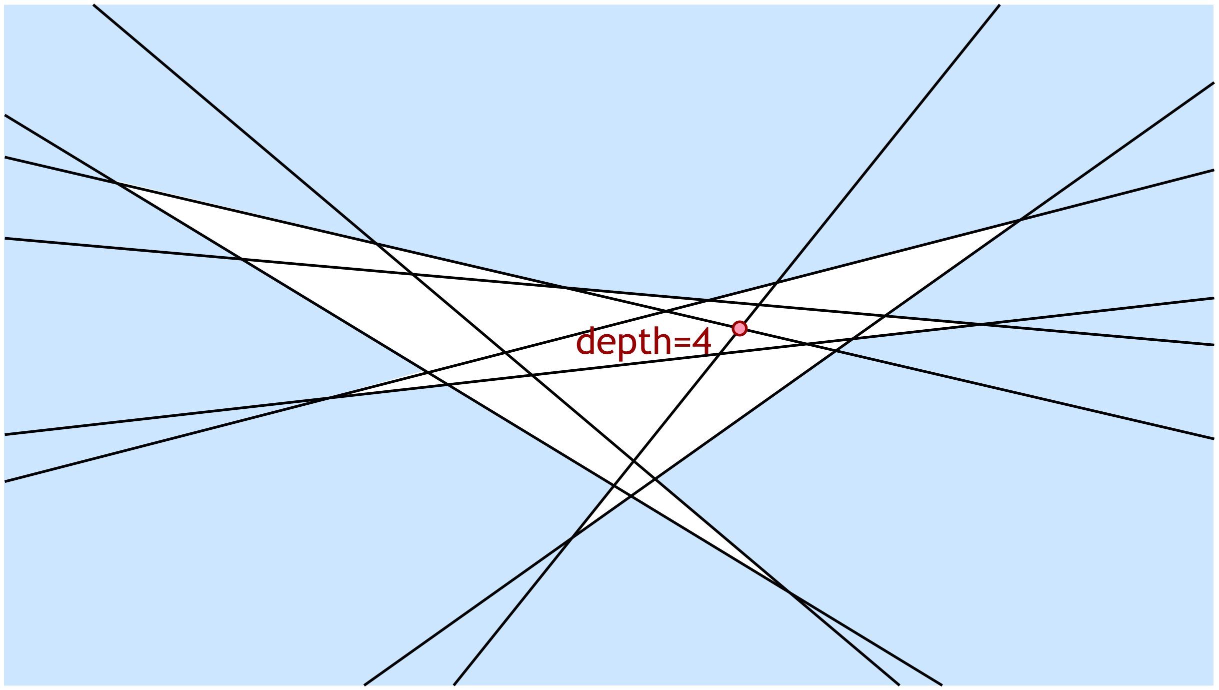
Tukey depth of *L* = distance from point at vertical infinity

Regression Depth

[Rousseeuw and Hubert, JASA 1999]

Nonfit = line combinatorially equivalent to a vertical line
(splits the data points into the same two subsets)

Duals of vertical lines = points at infinity
So dual nonfit = point in infinite cell of arrangement



Dual depth = arrangement distance from line at infinity

Regression depth and Tukey depth

Not quite projective duals of each other, but very closely related

Both seek object at maximum arrangement distance from another object

Regression depth: seek farthest point from line at infinity

Tukey depth: seek farthest line from point at vertical infinity

Which properties and algorithms for Tukey depth
can be carried over to regression depth?

E.g. conjecture [Rousseeuw]: maximum regression depth ≥ $n/(d+1)$

Proof that depth ≥ *n*/(*d* + 1): key idea

[Amenta, Bern, E, Teng, DCG 2000]

We want to find a point far from line at infinity

Tukey depth gives a point (vertical infinity)
far from a different line (dual to the deep point)

Find a distance-preserving transformation of projective space
that maps original line at infinity to transformed (dual) Tukey median

Then:
vertical infinity is far from the Tukey median
inverse image of vertical infinity is far from inverse image of Tukey median
inverse image of vertical infinity is far from line at infinity, QED

Proof that depth ≥ $n/(d + 1)$: distance preserving transformations

Tukey median is Euclidean not projective concept
(depends on location of line at infinity)

Standard embedding Euclidean into projective:
Form 3d Euclidean plane z=1 tangent to unit sphere at (0,0,1)
Line through origin intersects plane in single Euclidean point

Other embeddings:
Form 3d Euclidean plane tangent to unit sphere at another point
Line through origin intersects plane in single Euclidean point

Transformation: extend arrangement to projective space,
pick different tangent point, project back to Euclidean space

Proof that depth ≥ $n/(d + 1)$: the rest of the proof

For each choice of Euclidean embedding (tangent point on unit sphere)
we have a different Tukey median

Define function T:
T(tangent point) = point on unit sphere
on line from origin to Tukey median

Modify definition of depth (smear out points to continuous distribution)
so that T is a continuous function

Opposite tangencies give same Euclidean embedding
so T is symmetric: $T(-x) = -T(x)$

Borsuk-Ulam theorem: symmetric function on sphere covers all points
so some x has T(x) = vertical infinity, QED

Other results on regression depth

Can test depth of a hyperplane in time O(n^{d-1} + n log n), space O(n)
(transform to Tukey depth, view samples from viewpoint of test point
as oriented-projective (d – 1)-space, sweep dual arrangement)
[O(n^{d-1} log n): Rousseeuw and Struyf, Statistics & Comput. 1998;
O(n^d): Rosta, this workshop]

Planar point of depth ≥ $n/3$ can be found in time O(n)
("catline", ham sandwich cut technique)
[Hubert and Rousseeuw, J.Multivariate Anal. 1998]

Deepest point can be found in time O(n log n)
(similar approach to slope selection)
[Langerman and Steiger, SODA 2000]

Outline

Projective duality and arrangements

Least median of squares

Least absolute deviation

Slope selection

Regression depth

Multivariate regression

What is multivariate regression?

More than one response variable
fewer than d – 1 stimulus variables

Example: one stimulus, two responses:
fitting a line to three-dimensional data
more generally, fit k-flat, $0 < k < d - 1$

Sometimes (e.g. ordinary least squares)
works well to treat each response variable independently

But for other optimization criteria (e.g., depth),
treating variables independently may lead to suboptimal fit

Multivariate regression depth

[Bern and E, SoCG 2000]

Nonfits should include any *k*-flat that does not predict response variables
e.g. for lines in space, lines parallel to *yz* plane

These flats all meet a (d – k – 1)-flat V at vertical infinity

Define regression depth = arrangement distance from V

Matches Tukey depth when k = 0
Matches regression depth when k = d – 1

Results for multivariate depth

[Bern and E, SoCG 2000 and SODA 2001]

For any d and k, depth is a constant fraction of n
(so deepest flat has constant breakdown point)
Conjecture: depth $\geq n/((k + 1)(d - k) - 1)$

Conjecture is true for $k = 1$
e.g. depth for lines in space $\geq n/5$
(generalization of catline)
Or for any k, n if samples are in general position [Mizera]

Can find deepest flat in time $O(n^{O(1)})$,
approximately deepest flat in time $O(n)$

Can test depth of a flat in time $O(n^{d-2} + n \log n)$
(order of magnitude faster than Tukey depth or regression depth)

Conclusions

Duality helps us understand many robust regression methods

Statistics can benefit from fast geometric algorithms

Computational geometry can benefit from interesting statistical problems

Open question:
What about data that is not dual to hyperplane arrangements?
E.g. data consisting of lines in space