CS 163: Computational Geometry
Tufts University, Fall 2022

Professor Diane Souvaine
Scribe: Ziqi (Stephanie) Xu

Duality and Least Median of Squares Regression using Vertical Line Sweep

Oct 12, 2022

1. Overview

Last time, we talked about planer plan location using Kirkpatrick (1983) algorithm that involves hierarchical search.

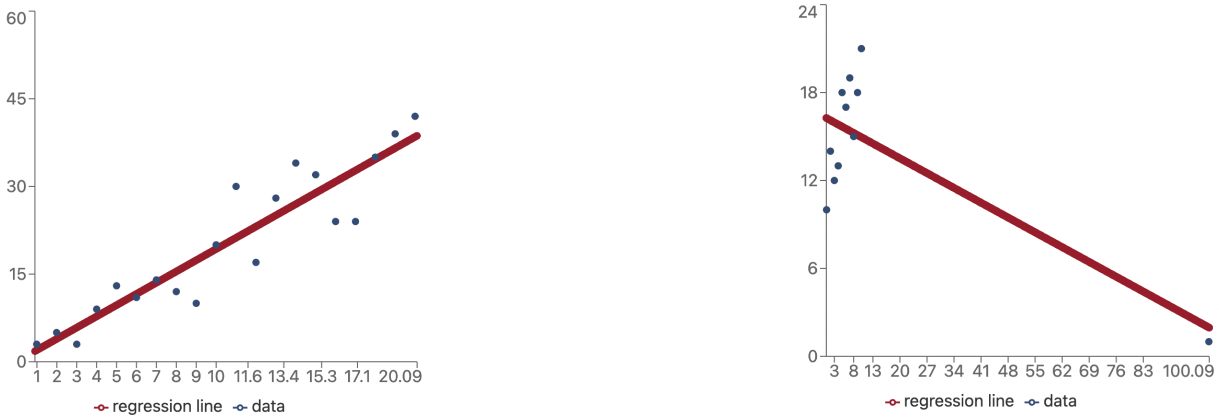
Today, the lecture is about computational statistics: finding the least median squares (LMS) regression using point-line duality and vertical line sweep.

The problem is to find a line that best describes a set of points on a plane.

2. Linear Regression: Ordinary Least Sum of Squares (OLS)

One classical approach to solve the problem is to find the line that minimizes the sum of the squares of the residuals, where a residual is the vertical distance from a data point to the line we find. (See Figure 1(a))[1]

However, the approach does not describe the major trend of the data set when there is one or some outliers in the data set. (See Figure 1(b))



(a) Good scenario

(b) Bad scenario

Figure 1: Finding regression using OLS

[1] Both pictures are generated using https://www.omnicalculator.com/math/least-squares-regression

3. Least Median of Squares (LMS) Regression

A more robust method (i.e. a method that's more tolerant of noises or outliers) is to find the line that minimizes the median of squares of the residuals instead.

3.1 Building up intuition to understand the algorithm

If we were given a line that claims to fit the data set (see Fig 2(a)), how can we adjust it so that we can reduce the median of squares of the residuals?

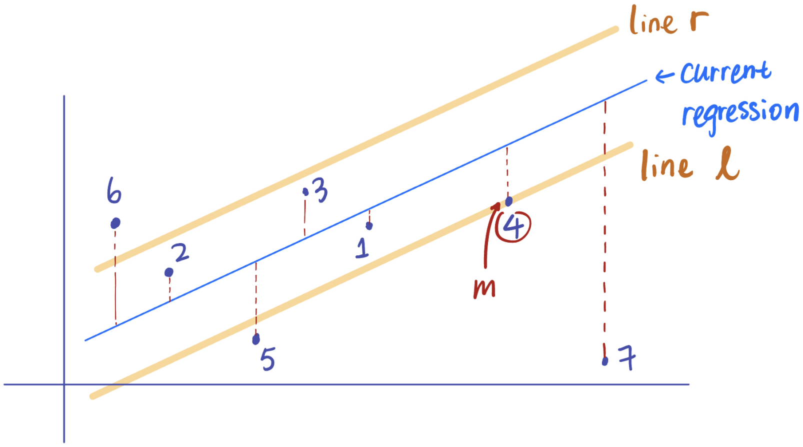


Figure 2: (a)

- I labelled the points with the sorted order of their vertical distances to the line. And point m is the point that has the median of squares of the residuals
- If we make a line parallel to the current regression line that passes through pt m, call it line l; and duplicate it to the other side with the same vertical distance between, call it line r; then there will be at least $\lceil \frac{n}{2} \rceil$ points in between the two lines, because half of the points that have residuals smaller than m would be locating between the slab, and the other half having larger residuals would be outside of the slab.
- The regression line is the bisector of this slab where the bottom binding line l goes through the "median" point. But currently the top binding line touches no point, so it gives us room to squish it down until it hits a point. In this way, we can decrease the vertical distance between the slab while still maintaining $\frac{n}{2}$ points between the slab.
- However, both line l and line r only go through 1 point each, so they are not stable, meaning we don't have a systemic way yet to control/update the slope of the regression.

3.2 Stable slab needs a 3rd point

Since two points can't define a "slab" (see Figure 3a), we need a third point (see Figure 3b).

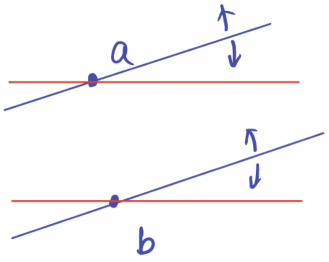


Fig 3(a)

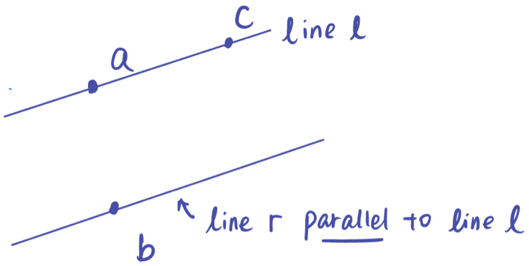


Fig 3(b)

We know from 3.1 that a point having the median of squares of the residuals will lie on one binding line of the slab, and that the slab contains $\lceil \frac{n}{2} \rceil$ points in between (including points on the boundary).

So instead of finding the line that minimizes the median of squares of the residuals, we can transform the problem into finding a triplet of points a, b, c that defines a slab by line $\overrightarrow{ac}$ and a line parallel to $\overrightarrow{ac}$ that passes through b, such that it contains $\lceil \frac{n}{2} \rceil$ points in between and that the two binding lines has the smallest possible vertical distance.

Now we have all the tools to design a naive algorithm.

3.3 Naive Algorithm for finding LMS regression

Algorithm

```
1) Set min-width to MAXINT
2) For every triplet of points $p_i, p_j, p_k$ in the set $S$:
    • Put a line through $p_i, p_j$
    • Put a parallel line through $p_k$
    • Count the number of points in the slab
    • If the number $\geq \lceil \frac{n}{2} \rceil$ and the width $<$ min-width, then reset the min-width
      and remember $i, j, k$
3) Return the bisector of the min-width slab that includes at least $\frac{n}{2}$ points in it.
```

Justification

Assuming general positions of points, we consider all possible candidates and choose the optimal one (i.e. the line associated with the min-width slab). A min-width slab also means that the distance from the bisector of the slab to any points on the binding lines of the slab is minimum, which is what we want: the least median of squares of the residuals.

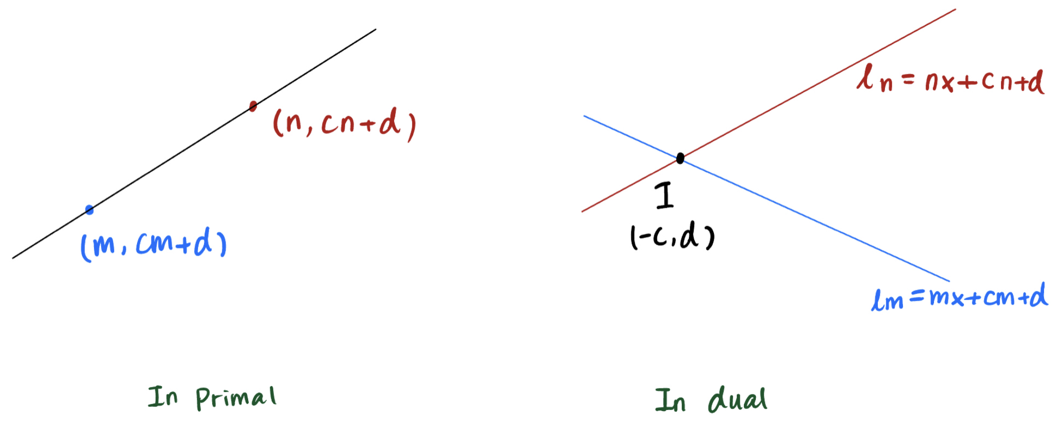
Runtime and Space

Runtime is $O(n^4)$ since there are $O(n^3)$ triplets of points in a set of size n, and for each triplet we are doing linear-time computation to count the number of points within the slab. Space complexity is constant.

3.4 Point-Line Duality

A more efficient algorithm requires the concept of point-line duality: we want a set of points S_p in the primal to have symmetrical roles as a set of lines L_d in the dual; and a set of points S_d in the dual to have symmetrical roles as a set of lines L_p in the primal.

Define the transformation T_1 as: $(a, b) \mapsto y = ax + b$ from the primal space to the dual space.



As shown in the figure above,

- Starting with two points in the primal: $p_1 = (m, cm + d)$ and $p_2 = (n, cn + d)$.
- In the dual: $T(p_1) = l_m = mx + cm + d$, $T(p_2) = l_n = nx + cn + d$.
- The intersection point I of l_n and l_m can be derived from solving the equation $mx + cm + d = nx + cn + d$, which gives us $I = (-c, d)$.

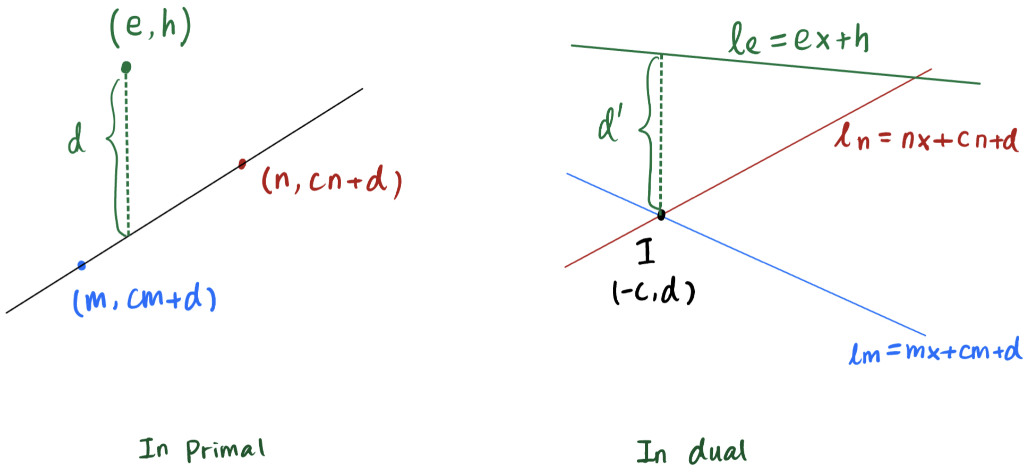
- I is represented by the line $y = cx + d$ in the primal, i.e. the line that points p_1 and p_2 define.

Therefore, we know that the transformation from a point in the dual to a line in the primal can be defined as: $T_2 : (-c, d) \mapsto y = cx + d$.

The point-line duality preserves the distance between a point and a line.

Lemma: The distance from a point p to a line l in the primal is the same as the distance from a point $T(l)$ to a line $T(p)$ in the dual.

As shown in the figure below, the new point (e, h) in the primal is presented by a line l_e in the dual. And the distance between (e, h) and the line $y = cx + d$ in the prmal is equal to the distance between l_e and point I in the dual.



3.5 Vertical Line Sweep for finding LMS regression

Given a set of points S, we could use the below algorithm to find the line that has the least median of squares of residuals.

Algorithm

1) Transform all the points in S to a set of lines L in a dual, using $(a, b) \mapsto y = ax + b$.

2) Apply a vertical sweep line algorithm that would report all intersections of lines in L in the dual, with a stopping point min-heap and a STATUS array:

- At each intersection point p, count down or up $\lceil \frac{n}{2} \rceil$ lines in the STATUS array to find a line l.
- Calculate the distance between p and l.
- Maintain a running min to store the p_m and l_m that has the minimum distance through out the process.

3) Let p_m and l_m define a slab which contains $\lceil \frac{n}{2} \rceil$ points in between. Return the bisector of the slab as the LMS regression.

Justification

The algorithm works because the point-line transformation preserves the distance between a point and a line. We consider all possible lines that could act as a binding line of the slab in step 2) by traversing through all intersection points in the dual. And for each binding line, we jump up or down $\lceil \frac{n}{2} \rceil$ indices in the STATUS array to find the other binding point of a slab. There will be $\lceil \frac{n}{2} \rceil$ points within the slab, because all the lines in between represent points in the primal that have smaller distance to the binding line. So at the end we are able to compare and choose the min-width slab that contains half points in between to make the LMS regression.

Runtime and Space

Runtime is $O(n^2 \log n)$ because transforming the points to lines takes linear time, the vertical sweep line algorithm in step 2) takes $O(n^2 \log n)$, and step 3) takes constant time,

Space complexity is $O(n)$ to store the lines in the dual and to store additional space created by the vertical sweep line algorithm in step 2).