CSC338 Winter 2022

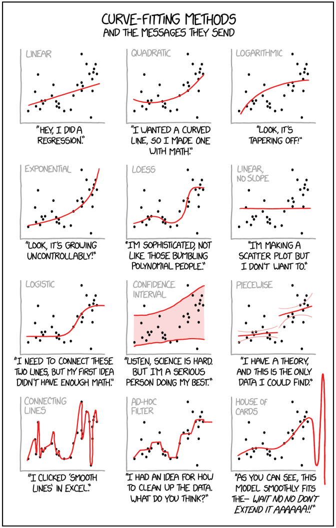
Week 5 - Linear Least Squares

Ilir Dema

University of Toronto

Feb 11, 2022

Curve Fitting



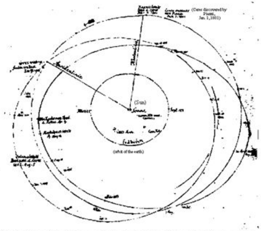
Source: `https://xkcd.com/2048/`

CURVE FITTING

- It's a known fact that $n+1$ points on the plane uniquely determine a polinomial of degree n.
- However, what if we are given the $n+1$ points, and from other considerations we know the shape of the curve, but we just need the parameters?
- A little history: Jan 1, 1801, Italian astronomer Giuzeppe Piazzi observed a new celestial object, whose motion he could trace for 42 days before the object disappeared.
- In Sep. 1801, the 24-years old Gauss, computed its orbit using curve fitting methods (regression). This is the first know attempt to use least squares method.
- Keep in mind, at that time, the theory of matrices was not yet developed in full.

Gauss

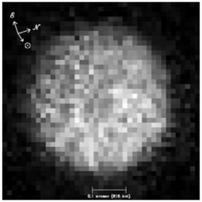
(Left) Carl Friedrich Gauss, considered one of the three greatest mathematicians of all time (along with Archimedes and Sir Isaac Newton). (Right) Gauss at 24, when he computed the orbit of Ceres.



Sketch of the orbits of Ceres and Pallas (nachlaß Gauß, Handb. 4). Courtesy of Universitätsbibliothek Göttingen.

(Left) Gauss' sketch of the orbit of Ceres.

(Right) Image of Ceres from the Hubble telescope.



CURVE FITTING

Example: Data Fitting

Fitting quadratic polynomial to five data points gives linear least squares problem

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cong \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \boldsymbol{b}$$

Matrix with columns (or rows) successive powers of independent variable called *Vandermonde* matrix

Curve Fitting

For data

t	-1.0	-0.5	0.0	0.5	1.0
y	1.0	0.5	0.0	0.5	2.0

overdetermined 5×3 linear system is

$$\boldsymbol{Ax} = \begin{bmatrix} 1 & -1.0 & 1.0 \\ 1 & -0.5 & 0.25 \\ 1 & 0.0 & 0.0 \\ 1 & 0.5 & 0.25 \\ 1 & 1.0 & 1.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cong \begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 0.5 \\ 2.0 \end{bmatrix} = \boldsymbol{b}$$

Solution, which we will see later how to compute, is

$$\boldsymbol{x} = [0.086 \quad 0.40 \quad 1.4]^T,$$

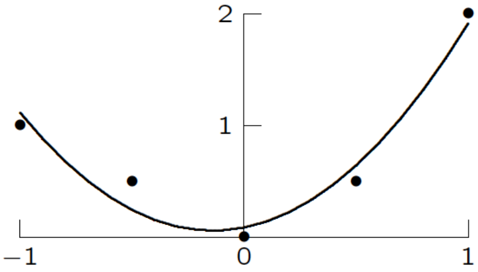
so approximating polynomial is

$$p(t) = 0.086 + 0.4t + 1.4t^2$$

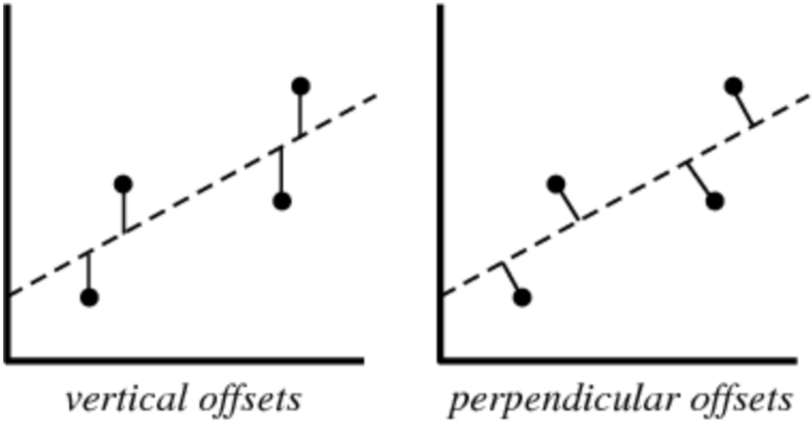
Curve fitting

Example: Data Fitting

Resulting curve and original data points shown in graph:



Linear case (credit: Wolfram)



Example

Example: Olympic winning times

To illustrate the computations, consider the following 20 data pairs, where x is the time in years since 1900 and y is the Olympic winning time in seconds for men in the final round of the 100-meter event [50, p. 248]:

x	0	4	8	12	20	24	28	32	36	48
y	10.8	11.0	10.8	10.8	10.8	10.6	10.8	10.3	10.3	10.3
x	52	56	60	64	68	72	76	80	84	88
y	10.4	10.5	10.2	10.0	9.95	10.14	10.06	10.25	9.99	9.92

The data set covers all Olympic events held between 1900 and 1988. (Olympic games were not held in 1916, 1940, and 1944.)

Credit: SIAM

Example (continued)

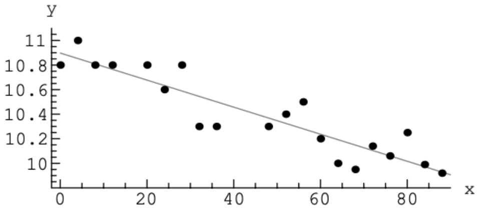


Figure 14.1. *Olympic winning time in seconds for men's 100-meter finals (vertical axis) versus year since* 1900 *(horizontal axis). The gray line is the linear least squares fit,* $y = 10.898 - 0.011x$.

Credit: SIAM

EXAMPLE (EXPLAINED)

- We want a simple linear model that fits the data and allows us to make predictions about future performance of athletes
- A line is determined by two points, so given we know their coordinates $(x_i, y_i), (x_j, y_j)$, using the general equation of the line $y = mx + k$ is not hard to write a system which allows us compute the parameters m, k:

$$\begin{bmatrix} x_i & 1 \\ x_j & 1 \end{bmatrix} \begin{bmatrix} m \\ k \end{bmatrix} = \begin{bmatrix} y_i \\ y_j \end{bmatrix}$$

- This way we get a mess of lines good for nothing (the problem is over determined).
- Therefore we seek the line that is closest to all data points simultaneously.

Example (explained)

$$\begin{bmatrix} 0 & 1 \\ 4 & 1 \\ 8 & 1 \\ 12 & 1 \\ 20 & 1 \\ 24 & 1 \\ 28 & 1 \\ 32 & 1 \\ 36 & 1 \\ 48 & 1 \\ \ldots & \\ 76 & 1 \\ 80 & 1 \\ 84 & 1 \\ 88 & 1 \end{bmatrix} \begin{bmatrix} m \\ k \end{bmatrix} \approx \begin{bmatrix} 10.8 \\ 11.0 \\ 10.8 \\ 10.8 \\ 10.8 \\ 10.6 \\ 10.8 \\ 10.3 \\ 10.3 \\ 10.3 \\ \ldots \\ 10.06 \\ 10.25 \\ 9.99 \\ 9.92 \end{bmatrix}$$

Example (explained)

- The previous system looks like $\boldsymbol{Ax} = \boldsymbol{B}$ where $\boldsymbol{A}$ is a 20×2 matrix and $\boldsymbol{b}$ is a 20 dimensional vector, whereas $\boldsymbol{x}$ is a 2 dimensional vector.
- The system is overdetermined, (almost) always impossible to satify precisely, so we instead seek to minimize $\|\boldsymbol{r}\|_{\mathbf{2}}$ where $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}$ is the residual.
- For $\|\boldsymbol{r}\|_{\mathbf{2}}^{\mathbf{2}} = \phi(x, y) = \|\boldsymbol{b} - \boldsymbol{Ax}\|_2^2$ to attain minimal value, we need $\nabla\phi = (\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}) = (0, 0)$ and Hessian

$$H(x,y) = \begin{bmatrix} \phi_{xx} & \phi_{xy} \\ \phi_{yx} & \phi_{yy} \end{bmatrix}$$

be positive definite.

Formalization of the linear problem

Linear Least Squares

For linear problems, obtain *overdetermined* linear system $\boldsymbol{Ax} = \boldsymbol{b}$, with $m \times n$ matrix $\boldsymbol{A}$, $m > n$

Better written $\boldsymbol{Ax} \cong \boldsymbol{b}$, since equality usually not exactly satisfiable when $m > n$

Least squares solution $\boldsymbol{x}$ minimizes squared Euclidean norm of residual vector $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}$,

$$\min_{\boldsymbol{x}} \|\boldsymbol{r}\|_2^2 = \min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{Ax}\|_2^2$$

Let's do the math

- $\phi(\boldsymbol{x}) = (\boldsymbol{b} - \boldsymbol{Ax})^{\boldsymbol{T}}(\boldsymbol{b} - \boldsymbol{Ax})$
- Require $\nabla\phi = 0$.
- $\nabla\phi = 2(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Ax} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}) = 0$.
- $H(x) = 2\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$.
- Therefore we need two conditions:
  - $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Ax} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}$ (normal equations)
  - $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$ be positive definite.
  - The last condition can be guaranteed if $\boldsymbol{A}$ has full rank (that is n, assuming $\boldsymbol{A}$ is $m \times n, m > n$.
- Since $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$ is positive definite, using Cholesky factorization, $\boldsymbol{L}^{\boldsymbol{T}}\boldsymbol{Lx} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}$ we can solve the system.
- If $rank(\boldsymbol{A}) < n$ the solution may not be unique.

Geometrical viewpoint

Vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are *orthogonal* if their inner product is zero, $\boldsymbol{v}_1^T \boldsymbol{v}_2 = 0$

Space spanned by columns of $m \times n$ matrix $\boldsymbol{A}$, $\text{span}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\}$, is of dimension at most n

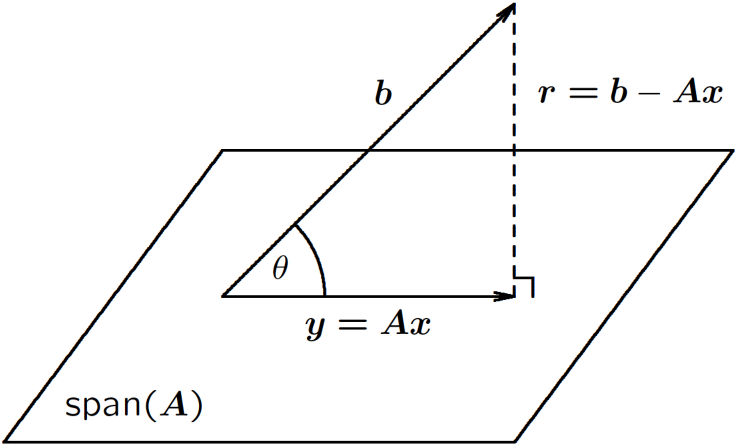
If $m > n$, $\boldsymbol{b}$ generally does not lie in $\text{span}(\boldsymbol{A})$, so no exact solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

Vector $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ in $\text{span}(\boldsymbol{A})$ closest to $\boldsymbol{b}$ in 2-norm occurs when residual $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ orthogonal to $\text{span}(\boldsymbol{A})$

Thus,

$$\boldsymbol{o} = \boldsymbol{A}^T \boldsymbol{r} = \boldsymbol{A}^T (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}),$$

Geometrical viewpoint



References

Michael T. Heath
Scientific Computing (Revised Second Edition)
SIAM