Math 19b: Linear Algebra with Probability **Oliver Knill, Spring 2011**

Lecture 20: More data fitting

Last time, we saw how the geometric formula $P = A(A^TA)^{-1}A^T$ for the projection on the image of a matrix A allows us to fit data. Given a fitting problem, we write it as a system of linear equations

$$Ax = b\ .$$

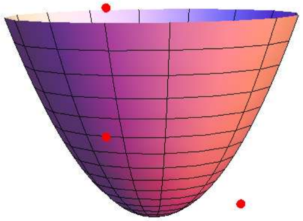
While this system is not solvable in general, we can look for the point on the image of A which is closest to b. This is the "best possible choice" of a solution" and called the **least square solution**:

The vector $x = (A^TA)^{-1}A^Tb$ is the **least square solution** of the system $Ax = b$.

The most popular example of a data fitting problem is **linear regression**. Here we have data points (x_i, y_i) and want to find the best line $y = ax + b$ which fits these data. But data fitting can be done with any finite set of functions. Data fitting can be done in higher dimensions too. We can for example look for the best surface fit through a given set of points (x_i, y_i, z_i) in space. Also here, we find the least square solution of the corresponding system $Ax = b$ which is obtained by assuming all points to be on the surface.

1 Which paraboloid $ax^2 + by^2 = z$ best fits the data

x	y	z
0	1	2
-1	0	4
1	-1	3



In other words, find the least square solution for the system of equations for the unknowns a, b which aims to have all data points on the paraboloid.

Solution: We have to find the least square solution to the system of equations

$$\begin{aligned} a0 + b\,1 &= 2 \\ a1 + b\,0 &= 4 \\ a1 + b\,1 &= 3\ . \end{aligned}$$

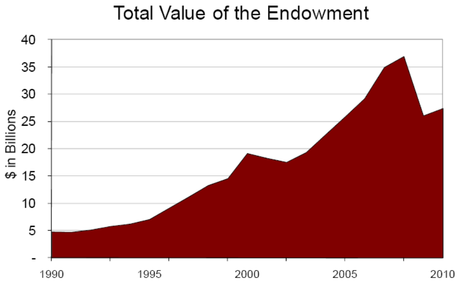
In matrix form this can be written as $A\vec{x} = \vec{b}$ with

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \vec{b} = \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}.$$

We have $A^TA = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $A^Tb = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$. We get the least square solution with the formula

$$x = (A^TA)^{-1}A^Tb = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

The best fit is the function $\boxed{f(x,y) = 3x^2 + y^2}$ which produces an elliptic paraboloid.



2 A graphic from the Harvard Management Company Endowment Report of October 2010 is shown to the left. Assume we want to fit the growth using functions $1, x, x^2$ and assume the years are numbered starting with 1990. What is the best parabola $a + bx + cx^2 = y$ which fits these data?

quintenium	endowment in billions
1	5
2	7
3	18
4	25
5	27

We solved this example in class with linear regression. We saw that the best fit. With a quadratic fit, we get the system $A\vec{x} = \vec{b}$ with

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix},$$

$$\vec{b} = \begin{bmatrix} 5 \\ 7 \\ 18 \\ 25 \\ 27 \end{bmatrix}.$$

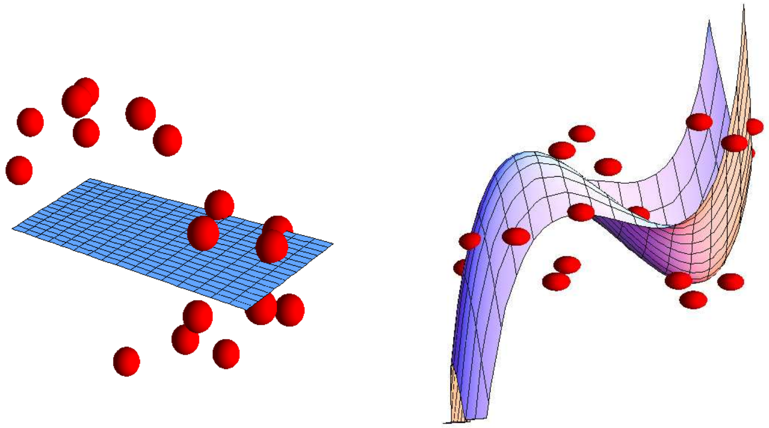
The solution vector $\vec{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -21/5 \\ 277/35 \\ -2/7 \end{bmatrix}$ which indicates strong linear growth but some slow down.

3 Here is a problem on data analysis from a website. We collect some data from users but not everybody fills in all the data

Person 1	3	5	-	3	9	-	-	-	2	9
Person 2	4	-	-	8	-	5	6	2	-	9
Person 3	-	4	2	5	7	-	1	9	8	-
Person 4	1	-	-	-	-	-	-	-	-	-

It is difficult to do statistic with this. One possibility is to filter out all data from people who do not fulfill a minimal requirement. Person 4 for example did not do the survey seriously enough. We would throw this data away. Now, one could sort the data according to some important row. Arter tha one could fit the data with a function $f(x, y)$ of two variables. This function could be used to fill in the missing data. After that, we would go and seek correlations between different rows.

Whenever doing datareduction like this, one must always compare different scenarios and investigate how much the outcome changes when changing the data.



The left picture shows a linear fit of the above data. The second picture shows a fit with cubic functions.

Homework due March 23, 2011

1 Here is an example of a fitting problem, where the solution is not unique:

x	y
0	1
0	2
0	3

Write down the corresponding fitting problem for linear functions $f(x) = ax + b = y$.What is going wrong?

2 If we fit data with a polynomial of the form $y = a_0 + a_1x + a_2x^2 + a_3x^3 + ...a + yx^7$. How many data points $(x_1, y_1), \dots, (x_m, y_m)$ do you expect to fit exactly if the points $x_1, x_2, ..., x_m$ are all different?

3 The first 6 prime numbers $2, 3, 5, 7, 11$ define the data points $(1, 2), (2, 3), (3, 5), (5, 7), (6, 11)$ in the plane. Find the best parabola of the form $y = ax^2 + c$ which fits these data.