Curve Fitting

Professor Henry Arguello

Universidad Industrial de Santander
Colombia

March 12, 2018

High Dimensional Signal Processing Group
www.hdspgroup.com
henarfu@uis.edu.co
LP 304

Least-squares Line

Least-squares Line

Given a set of data points (x_1, y_1), ..., (x_N, y_N), where the abscissas $\{x_k\}$ are distinct, one goal of numerical methods is to determine a formula $y = f(x)$ that relates these variables. This section emphasizes fitting the data to linear functions of the form,

$$(1) \qquad y = f(x) = Ax + B$$

Least-squares Line

If all the numerical values x_k, y_k are known to several significant digits of accuracy, then polynomial interpolation can be used successfully; otherwise it can not. Often there is an experimental error in the measurements, and although three digits are recorded for the values x_k and y_k, it is realized that the true value $f(x_k)$ satisfies

$$(2) \qquad f(x_k) = y_k + 2e_k,$$

Where e_k is the measurement error.

How do we find the best linear approximation of the form (1) that goes near (not always through) the points? To answer this question, we need to discuss the errors (also called derivations or residuals):

$$(3) \qquad e_k = f(x_k - y_k) \text{ for } 1 \leq k \leq N.$$

Least-squares Line

There are several norms that can be used with the residuals in (3) to measure how far the curve $y = f(x)$ lies from the data

Least-squares Line

There are several norms that can be used with the residuals in (3) to measure how far the curve $y = f(x)$ lies from the data

(4) Maximun error: $E\infty(f) = max|f(x_k - y_k)| \quad 1 <= k <= N,$

Least-squares Line

There are several norms that can be used with the residuals in (3) to measure how far the curve $y = f(x)$ lies from the data

(4) Maximun error: $E\infty(f) = max|f(x_k - y_k)| \quad 1 <= k <= N,$

(5) Average error: $E_1(f) = \frac{1}{N}\sum_{k=1}^{N} |f(x_k - y_k)|,$

Least-squares Line

There are several norms that can be used with the residuals in (3) to measure how far the curve $y = f(x)$ lies from the data

(4) Maximun error: $E\infty(f) = max|f(x_k - y_k)| \quad 1 <= k <= N,$

(5) Average error: $E_1(f) = \frac{1}{N}\sum_{k=1}^{N} |f(x_k - y_k)|,$

(6) Root-Mean-Square error: $E_2(f) = (\frac{1}{N}\sum_{k=1}^{N} |f(x_k - y_k)|^2)^{1/2}.$

Least-squares Line

Example Compare the maximum error, average error and RMS error for the linear approximation $y = f(x) = 8.61.6x$ to the data points (-1,10), (0,9), (1,7), (2,5), (3,4), (4,3), (5,0) and (6,-1).

The errors are found using the values for $f(x_k)$ and e_k given in table

x_k	y_k	$f(x_k) = 8.6 - 1.6x_k$	e_k	e_k^2
-1	10.0	10.2	0.2	0.04
0	9.0	8.6	0.4	0.16
1	7.0	7.0	0.0	0.00
2	5.0	5.4	0.4	0.16
3	4.0	3.8	0.2	0.04
4	3.0	2.2	0.8	0.64
6	-1.0	-1.0	0.0	0.00
			2.6	1.40

Least-squares Line

(7) $$E_\infty(f) = max\{0.2, 0.4, 0.0, 0.4, 0.2, 0.8, 0.6, 0.0\} = 0.8,$$

(8) $$E_1(f) = \frac{1}{8}(2.6) = 0.325,$$

(9) $$E_2(f) = (\frac{1}{N}(1.4)^{1/2}) \approx 0.41833.$$

We can see that the maximum error is largest, and if one point is badly in error, its value determines $E_1(f)$. The average error $E_1(f)$ simply averages the absolute value of the error at the various points. It is often used because it is easy to compute. The error $E_2(f)$ is often used when the statistical nature of the error is considered.

A best-fitting line is found by minimizing one of the quantities in equations (4) through (6). Hence there are three best-fitting lines that we could find.

Finding the Least-Squares line

Let $\{(x_k, y_k)\}_{k=1}^{N}$ be a set of N points, where the abscissas $\{x_k\}$ are distinct. **The least-squares line** $y = f(x) = Ax + B$ is the line that minimizes the Root-Mean-Square error $E_2(f)$.

The quantity $E_2(f)$ will be a minimum if and only if the quantity

$$N(E_2(f))^2 = \sum_{k=1}^{N}(Ax_k + By_k)^2 \text{ is a minimum.}$$

The latter is visualized geometrically by minimizing the sum of the squares of the vertical distances from the points to the line.

Least-squares Line

Theorem: Least-Squares Line

Suppose that $\{(x_k, y_k)\}_{k=1}^{N}$ are N points, where the abscissas $\{x_k\}$ are distinct. The coefficients of the least-squares line

$$y = Ax + B$$

are the solution of the following linear system, known as the **normal equations:**

$$\left(\sum_{k=1}^{N} x_k^2\right) A + \left(\sum_{k=1}^{N} x_k\right) B = \sum_{k=1}^{N} x_k, y_k, \tag{10}$$

$$\left(\sum_{k=1}^{N} x_k\right) A + NB = \sum_{k=1}^{N} y_k.$$

Least-squares Line

Proof. Geometrically, we start with line $y = Ax + B$. The vertical distance d_k from the point (x_k, y_k) to the point $(x_k, Ax_k + B)$ on the line is $d_k = |Ax_k + By_k|$

Least-squares Line

(11) $\quad E(A,B) = \sum_{k=1}^{N}(Ax_k + B - y_k)^2 = \sum_{k=1}^{N} d_k^2$

The minimum value of $E(A,B)$ is determined by setting the partial derivatives $\partial E/\partial A$and $\partial E/\partial B$ equal to zero and solving these equations for A and B. Notice that $\{x_k\}$ and $\{y_k\}$ are constants in equation and that A and B are the variables! Hold B fixed, differentiate $E(A,B)$ with respect to A and get

(12) $\quad \frac{\partial_{E(A,B)}}{\partial_A} = \sum_{k=1}^{N} 2(Ax_k + B - y_k)(x_k) = 2\sum_{k=1}^{N}(Ax_k^2 + Bx_k - y_k x_k)$

Now hold A fixed and differentiate $E(A,B)$ with respect to B and get

(13) $\quad \frac{\partial_{E(A,B)}}{\partial_A} = \sum_{k=1}^{N} 2(Ax_k + B - y_k) = 2\sum_{k=1}^{N}(Ax_k + B - y_k)$

Setting the partial derivatives equal to zero in (12) and (13), use the distributive properties of summation to obtain

(14) $\quad 0 = \sum_{k=1}^{N}(Ax_k^2 + Bx_k y_k x_k) = A\sum_{k=1}^{N} x_k^2 + B\sum_{k=1}^{N} x_k - \sum_{k=1}^{N} y_k x_k$

(15) $\quad 0 = \sum_{k=1}^{N}(Ax_k + B - y_k) = A\sum_{k=1}^{N} x_k + NB - \sum_{k=1}^{N} y_k$

Equations (14) and (15) can be rearranged in the standard form for a system and result in the normal equations (10).

Example

Find the least-squares line for the data points given in the above example. The sums required for the normal equations (10) are easily obtained using the values in the table.

Least-squares Line

k	x_k	y_k	x_k^2	$x_k y_k$
0	-1	10	1	-10
1	0	9	0	0
2	1	7	1	7
3	2	5	4	10
4	3	4	9	12
5	4	3	16	12
6	5	0	25	0
7	6	-1	36	-6
$\sum$	20	37	92	25

The linear system involving A and B is

$$\begin{aligned} 92A + 20B &= 25 \\ 20A + 8B &= 37 \end{aligned}$$

The solution of the linear system is $A \approx -1.6071429$ and $B \approx 8.6428571$. Therefore, the least-squares line is (see figure)

$$y = -1.6071429x + 8.6428571$$

Least-squares Line

Power Fit $y = Ax^M$
Some situations involve $f(x) = Ax^M$, where M is a known constant. In these cases there is only one parameter A to be determined.

Teorem: Power Fit

Suppose that $\{x_k, y_k\}_{k=1}^{M}$ are N points, where the abscissas are distinct. The coefficient A of the least-squares power curve $y = Ax^M$ is given by

$$A = (\textstyle\sum_{k=1}^{N} x_k^M y_k) / \sum_{k=1}^{N} x_k^{2M}) \tag{16}$$

Using the least-squares technique, we seek a minimum of the function $E(A)$:

$$E(A) = \textstyle\sum_{k=1}^{N} (Ax_k^M y_k)^2 \tag{17}$$

Least-squares Line

In this case it will satisfy to solve $E'(A) = 0$. The derivative is

$$(18) \qquad E'(A) = 2\sum_{k=1}^{N}(Ax_k^M - y_k)(x_k^M) = 2\sum_{k=1}^{N}(Ax_k^{2M} - x_k^M y_k)$$

Hence the coefficient A is the solution of the equation

$$(19) \qquad 0 = A\sum_{k=1}^{N} x_k^{2M} - \sum_{k=1}^{N} x_k^M y_k,$$

which reduces to the formula in equation (16).

Example

Students collected the experimental data in table. The relation is $d = \frac{1}{2}gt^2$, where d is distance in meters and t is time in seconds. Find the gravitational constant g

Least-squares Line

Time, t_k	Distance, d_k	$d_k t_k^2$	t_k^4
0.200	0.1960	0.00784	0.0016
0.400	0.7850	0.12560	0.0256
0.600	1.7665	0.63594	0.1296
0.800	3.1405	2.00992	0.4096
1.000	4.9075	4.90750	1.0000
		7.68680	1.5664

The values in table are use to find the summations required in formula (16), where the power used is $M = 2$.

The coefficient $A = 7.68680/1.5664 = 4.9073$, and we get $d = 4.9073t^2$ and $g = 2A = 9.7146m/sec^2$.

Data Linearization Method for $y = Ce^{Ax}$

Suppose that we are given the points $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ and want to fit an exponential curve of the form

$$(1) \qquad y = Ce^{Ax}$$

The first step is to take the logarithm of both sides:

$$(2) \qquad In(y) = Ax + In(C).$$

Then introduce the change of variables:

(3) $\qquad Y = In(y), \qquad X = x, \qquad$ and $\qquad$ B=$ln(C)$.

This results in a linear relation between the new variables X and Y:

$$(4) \qquad Y = AX + B$$

Methods of Curve Fitting

The original points (x_k, y_k) in the xy-plane are transformed into the points $(X_k, Y_k) = (x_k, ln(y_k))$ in the XY-plane. This process is called data linearization. Then the least-squares line (4) is fit to the points $\{(X_k, Y_k)\}$. The normal equations for finding A and B are

$$\text{(5)} \qquad \left(\sum_{k=1}^{N} X_k^2\right) A + \left(\sum_{k=1}^{N} X_k\right) = \sum_{k=1}^{N} X_k Y_k,$$

$$\left(\sum_{k=1}^{N} X_k\right) A + NB = \sum_{k=1}^{N} Y_k.$$

After A and B have been found, the parameter C in equation (1) is computed:

$$\text{(6)} \qquad C = e^B.$$

Example

Use the data linearization method and find the exponential fit $y = Ce^{Ax}$ for the five data points (0, 1.5), (1, 2.5), (2, 3.5), (3, 5.0), and (4, 7.5). Apply the transformation (3) to the original points and obtain

x_k	y_k	X_k	$Y_k = In(y_k)$	X_k^2	$X_k Y_k$
0.0	1.5	0.0	0.405465	0.0	0.000000
1.0	2.5	1.0	0.916291	1.0	0.916291
2.0	3.5	2.0	1.252763	4.0	2.505526
3.0	5.0	3.0	1.609438	9.0	4.828314
4.0	7.5	4.0	2.014903	16.0	8.059612
		10.0	6.198860	30	16.309743

These transformed points are shown in figure and exhibit a linearized form. The equation of the least-squares line $Y = AX + B$ for the points in the table is in the next figure

$$Y = 0.391202X + 0.457367 \tag{8}$$

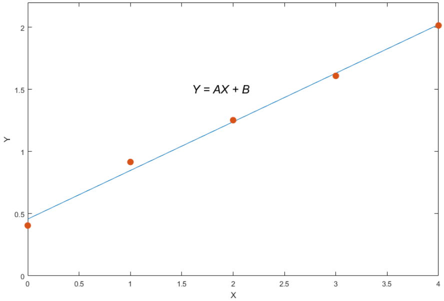

Figure: The transformed data points (X_k, Y_k)

Calculation of the coefficients for the normal equations in (5) is shown in table. The resulting linear system (5) for determining A and B is

$$
\begin{aligned}
30A + 10B &= 16.309742 \\
30A + 5B &= 6.198860
\end{aligned}
\tag{9}
$$

The solution is $A = 0.3912023$ and $B = 0.47367$. Then C is obtained wtih the calculation $C = e^{0.457367} = 1.579910$, and these values for A and C are substituted into equation (1) to obtain the exponential fit

(10) y = 1.579910e0.457367 (fit by data linearization).

Nonlinear Least-Squares Method for $y = Ce^{Ax}$ Suppose that we are given the points $(x1, y1), (x2, y2), ..., (xN, yN)$ and we want to fit an exponential curve:

$$(11) \qquad y = Ce^{Ax}.$$

The nonlinear least-squares procedure requires that we find a minimum of

$$(12) \qquad E(A, B) = \sum_{k=1}^{N} (Ce^{Ax_k} - y_k).$$

The partial derivatives of $E(A, B)$ with respect to A and C are

$$(13) \frac{\partial E}{\partial A} = 2 \sum_{k=1}^{N} (Ce^{Ax_k} - y_k)(Cx_k e^{Ax_k})$$

and

$$(14) \qquad \frac{\partial E}{\partial C} = 2 \sum_{k=1}^{N} (Ce^{Ax_k} - y_k)(x_k e^{Ax_k}).$$

When the partial derivatives in (13) and (14) are set equal to zero and then simplified, the resulting normal equations are

Transformations for Data Linearization

The technique of data linearization has been used by scientists to fit curves such as $y = Ce^{Ax}, y = Aln(x) + B$, and $y = A/x + B$. Once the curve has been chosen, a suitable transformation of the variables must be found so that a linear relation is obtained. For example, we can verify that $y = D/(x + C)$ is transformed into a linear problem $Y = AX + B$ by using the change of variables (and constants) $X = xy, Y = y, C = 1/A$, and $D = B/A$.

Linear Least Squares

The linear least-squares problem is stated as follows. Suppose that N data points $\{(X_k, Y_k)\}$ and a set of M linear independent functions $\{f_j(X)\}$ are given. We want to find M coefficients $\{c_j\}$ so that the function $f(x)$ given by the linear combination

$$f(x) = \textstyle\sum_{j=1}^{M} c_j f_j(x) \tag{16}$$

will minimize the sum of the squares of the errors:

$$E(c_1, c_2, \ldots, c_M) = \sum_{k=1}^{N} (f(x_k) - y_k)^2 = \sum_{k=1}^{N} \left(\left(\sum_{j=1}^{M} c_j f_j(x_k) \right) - y_k \right)^2. \tag{17}$$

Methods of Curve Fitting

For E to be minimized it is necessary that each partial derivative be zero $(i.e, \partial E/c_i = 0$ for $i = 1, 2, ..., M)$, and this results in the system of equations

$$(18) \qquad \sum_{k=1}^{N} \left(\left(\sum_{j=1}^{M} c_j f_j(x_k) \right) - y_k \right) (f_i(x_k)) = 0 \quad for \quad i = 1, 2, ..., M.$$

Interchanging the order of the summations in (18) will produce an $M \times M$ system of linear equations where the unknowns are the coefficients $\{c_j\}$. They are called the normal equations:

$$(19) \qquad \sum_{j=1}^{M} \left(\sum_{k=1}^{N} f_i(x_k) f_i(x_k) \right) c_j = \sum_{k=1}^{N} f_i(x_k) y_k \quad for \quad i = 1, 2, ..., M.$$

Matrix Formulation

Al though (19) is easily recognized as a system of M linear equations in M unknowns, one must be clever so that wasted computations are not performed when writing the system in matrix notation. The key is to write dawn the matrices ***F*** and ***F'*** as follows:

$$\boldsymbol{F} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_M(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_M(x_N) \end{bmatrix}, \boldsymbol{F'} = \begin{bmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_N) \\ \vdots & \vdots & & \vdots \\ f_M(x_1) & f_M(x_2) & \cdots & f_M(x_N) \end{bmatrix}.$$

Methods of Curve Fitting

Consider the product of ***F*** and the column matrix ***Y***:

$$(20) \qquad \boldsymbol{F'Y} = \begin{bmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_N) \\ \vdots & \vdots & & \vdots \\ f_M(x_1) & f_M(x_2) & \cdots & f_M(x_N) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

The element in the ith row of the product ***F'Y*** in (20) is the same as the ith element in the column matrix in equation (19); that is,

$$(21) \qquad \sum_{k=1}^{N} f_i(xk)y_k = row_i - \boldsymbol{F'} \, . \, \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}'$$

Now consider the product ***F'F***, which is an $M \times M$ matrix. The element in the ith row and jth column of F?F is the coefficient of cj in the ith row in equation (19); that is,

$$(22) \qquad \sum_{k=1}^{N} f_i(x_k)f_j(x_k) = f_i(x_1)f_j(x_1) + f_i(x_2)f_j(x_2) + \cdots + f_i(x_N)f_j(x_N).$$

When M is small, a computationally efficient way to calculate the linear least-squares coefficients for (16) is to store the matrix F,compute $F'F$, and $F'Y$ and then solve the linear system

(23) ***F'FC*** = ***F'Y*** for the coefficient matrix C

Polynomial Fitting

When the foregoing method is adapted to using the functions $\{f_i(x) = x^{j-1}\}$ and the index of summation ranges from $j = 1$ to $j = M + 1$, the function $f(x)$ will be a polynomial of degree M:

$$f(x) = c_1 + c_2 x + c_3 x^2 + ... + c_M + 1x^M \tag{24}$$

Least-Squares Parabola

Suppose that $\{(Xk, Yk)\}_{k=1}^{N}$ are N points, where the abscissas are distinct. The coefficients of the least-squares parabola

$$y = f(x) = Ax^2 + Bx + C \tag{25}$$

are the solution values A, B and C of the linear system

$$
\begin{aligned}
&\left(\sum_{k=1}^{N} x_k^4\right) A + \left(\sum_{k=1}^{N} x_k^3\right) B + \left(\sum_{k=1}^{N} x_k^2\right) C = \sum_{k=1}^{N} y_k x_k^2,\\
(26)\qquad &\left(\sum_{k=1}^{N} x_k^3\right) A + \left(\sum_{k=1}^{N} x_k^2\right) B + \left(\sum_{k=1}^{N} x_k\right) C = \sum_{k=1}^{N} y_k x_k,\\
&\left(\sum_{k=1}^{N} x_k^2\right) A + \left(\sum_{k=1}^{N} x_k\right) B + NC = \sum_{k=1}^{N} y_k
\end{aligned}
$$

Proof. The coefficients A, B, and C will minimize the quantity:

$$
(27)\qquad E(A,B,C) \sum_{k=1}^{N} (Ax_k^2 + Bx_k + C - y_k)^2.
$$

The partial derivatives $\partial E/\partial A, \partial E/\partial B$, and $\partial E/\partial C$ must be zero. This results in (28)

$$0 = \partial E(A,B,C)/\partial A = 2\sum_{k=1}^{N}(Ax_k^2 + Bx_k * C - y_k)(x_k^2),$$

$$0 = \partial E(A,B,C)/\partial B = 2\sum_{k=1}^{N}(Ax_k^2 + Bx_k * C - y_k)(x_k),$$

$$0 = \partial E(A,B,C)/\partial B = 2\sum_{k=1}^{N}(Ax_k^2 + Bx_k * C - y_k)(1).$$

Using the distributive property of addition, we can move the values A, B, and C outside the summations in (28) to obtain the normal equations that are given in (28).

Polynomial Wiggle

It is tempting to used a least-squares polynomial to fit data that are non-lineal. But if the data do not exhibit a polynomial nature, the resulting curve may exhibit large oscillations. This phenomenum, called polynomial wiggle, becomes more pronounced with higher-degree polynomials. For this reason we seldom use a polynomial of degree 6 or above unless it is known that the true function we are working with is a polynomial.

For example let $f(x) = 1.44/x^2 + 0.24x$ be used to generate the six data points (0.25,23.1), (1.0,1.68), (1.5,1.0), (2.0,0.84), (2.4,0.826), and (5.0,1.2576). The result of curve fitting with the least-square polynomials

$$
\begin{aligned}
P_2(x) &= 22.93 - 16.96x + 2.553x^2, \\
P_3(x) &= 33.04 - 46.51x + 19.51x^2 - 2.296^3, \\
P_4(x) &= 39.92 - 80.93x + 58.39x^2 - 17.15x^3 + 1.680x^4, \\
P_5(x) &= 46.02 - 118.1x + 119.4x^2 - 57.51x^3 + 13.03x^4 - 1.085x^5
\end{aligned}
$$

is shown in Figure through (d). Notice that $P_3(x), P_4(x)$, and $P_5(x)$, exhibit a large wiggle in the interval [2,5]. Even though $P_5(x)$ goes through the 6 points, it produces the worst fit. If we must fit a polynomial to this data, $P_2(x)$ should be the choice.

Methods of Curve Fitting

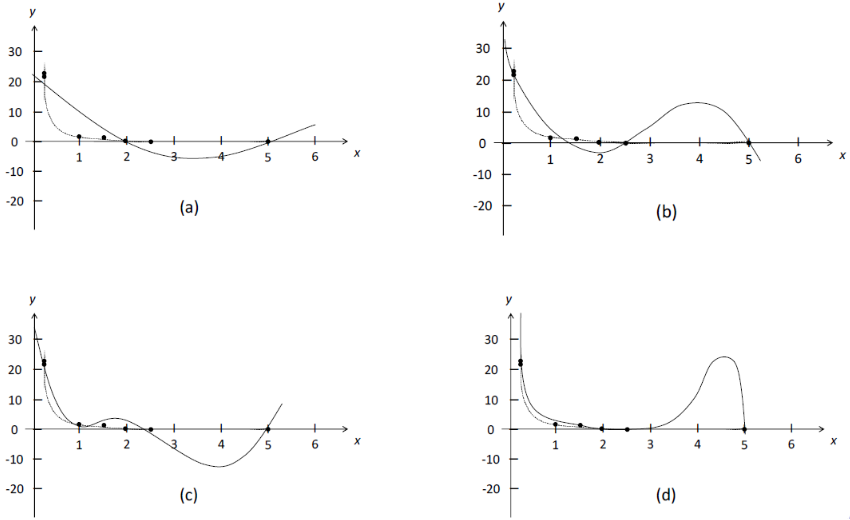

Figure: (a) using $P_2(x)$ to fit data. (b) using $P_3(x)$ to fit data. (c) using $P_4(x)$ to fit data. (d) using $P_5(x)$ to fit data