◇ III.1

The Best Least-Squares Line Fit

David Alciatore
Mechanical Engineering Department
Colorado State University
Fort Collins, Colorado

Rick Miranda
Mathematics Department
Colorado State University
Fort Collins, Colorado

◇ Introduction ◇

Traditional approaches for fitting least-squares lines to a set of two-dimensional data points involve minimizing the sum of the squares of the minimum vertical distances between the data points and the fitted line. That is, the fit is against a set of independent observations in the range[1] y. This gem presents a numerically stable algorithm that fits a line to a set of ordered pairs (x, y) by minimizing its least-squared distance to each point without regard to orientation. This is a true 2D point-fitting method exhibiting rotational invariance.

◇ Background ◇

The classical formula for the univariate case based on vertical error measurement is

$$\begin{aligned} y &= m_y x + b_y, \\ m_y &= \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}, \\ b_y &= \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - (\sum x_i)^2}. \end{aligned} \tag{1}$$

Though well known, and presented in many numerical, statistical, and analytical texts (Charpra and Canale 1988, Chatfield 1970, Kryszig 1983), the method is not acceptable as a general line-fitting tool. Its frequent misapplication gives poor results when both coordinates are uncertain or when the line to be fit is near vertical ($m_y \to \infty$). Reversing the axes merely disguises the problem: The method still remains sensitive to the orientation of the coordinate system.

A least-squares line-fitting method that is insensitive to coordinate system orientation can be constructed by minimizing instead the sum of the squares of the perpendicular

[1] Horizontal distances can also be used by reversing the roles of the variables.

IBM ISBN 0-12-543455-3
Macintosh ISBN 0-12-543457-X

distances between the data points and their nearest points on the target line. (The perpendiculars are geometric features of the model independent of the coordinate system.) Such an algorithm has been presented in the literature (Ehrig 1985), but the algorithm is based on a slope–intercept form of the line resulting in solution degeneracy and numerical inaccuracies; as the line approaches vertical, the slope and intercept grow without bound. Also, the equations provided (*op. cit.*) have two solutions, and the user must perform a test to determine the correct one.

The algorithm presented in the next section uses a θ–ρ (line angle, distance from the origin) parameterization of the line that results in no degenerate cases and gives a unique solution. This parameterization has been used for statistical fitting of noisy data with outlying points as in image data (Weiss 1988, Rosenfeld and Sher 1986), but the parameterization has not been applied to a least-squares line fit.

The perpendicular error measurement least-squares technique is also readily applied to circular arc fitting. Several robust solutions to this problem have been presented in the literature (Karimaki 1992, Moura and Kitney 1991, Chernov and Ososkov 1984).

◇ Optimal Least-Squares Fit ◇

The problem may now be stated. Given an arbitrary line defined by parameters (θ, ρ) and the sum of the squares of the related perpendicular distances r_i between points (x_i, y_i) and their nearest points to this line (Figure 1), then find the values of θ and ρ that minimize this sum. That is, minimize the value

$$Z = \sum_{i=1}^{N} r_i^2(\rho, \theta); \tag{2}$$

where N is the number of data points to be fitted and r_i is a function of the chosen line. Locating the zeros of the derivative of this function forms the method of solution.

To simplify the analysis and to avoid degeneracies, the parameter ρ is chosen to be the length of a perpendicular erected between the line and the origin, and θ is chosen to be its orientation with respect to the x axis (Figure 1). From simple plane geometry, the parametric equation for the line is given by

$$xs_\theta \boxed{-} yc_\theta + \rho = 0, \tag{3}$$

where

$$c_\theta = \cos(\theta) \text{ and } s_\theta = \sin(\theta). \tag{4}$$

The perpendicular distance r_i is given by

$$r_i = y_i c_\theta - x_i s_\theta - \rho. \tag{5}$$

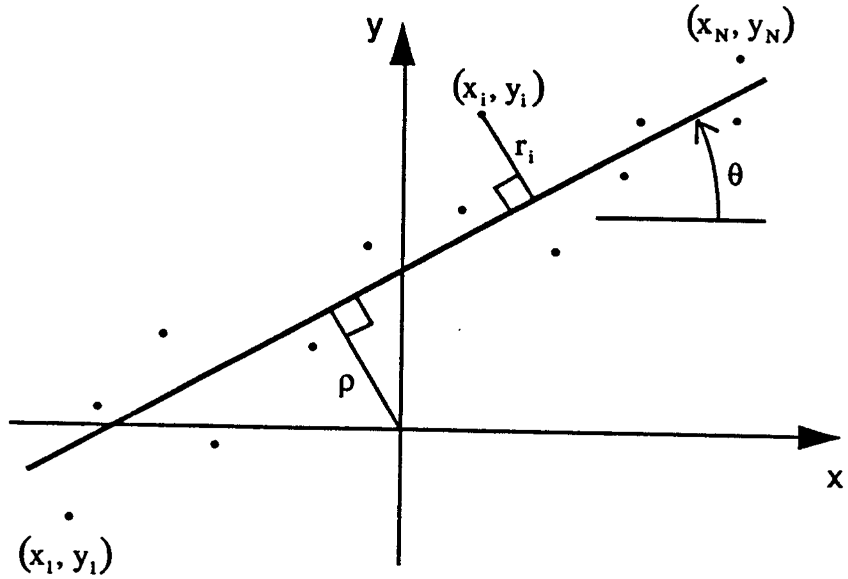

Figure 1. Least-squares line fit geometry.

To minimize the sum of errors Z in (2), the following must hold:

$$\frac{\partial Z}{\partial \theta} = 0 \text{ and } \frac{\partial Z}{\partial \rho} = 0. \tag{6}$$

Taking derivatives of (2) using (5) results in the following expressions:

$$ac_\theta s_\theta + b(s_\theta^2 - c_\theta^2) + c\rho c_\theta + d\rho s_\theta = 0 \tag{7}$$

and

$$dc_\theta - cs_\theta = N\rho, \tag{8}$$

where

$$a = \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} y_i^2, b = \sum_{i=1}^{N} x_i y_i, c = \sum_{i=1}^{N} x_i, \text{ and } d = \sum_{i=1}^{N} y_i. \tag{9}$$

Equation (8) can be written as

$$\overline{x} s_\theta - \overline{y} c_\theta + \rho = 0, \tag{10}$$

where $(\overline{x}, \overline{y})$ is the centroid of the data set $\{(x_i, y_i)\}$. Since (10) appears in the form presented in (3), the fit necessarily passes through the centroid of the data.

Equation (7) can be simplified if the original data are translated so that the centroid is located at the origin, by setting

$$x_i' = x_i - \overline{x} \text{ and } y_i' = y_i - \overline{y}. \tag{11}$$

This translation results in

$$c' = d' = \rho' = 0, \tag{12}$$

and (7) reduces to

$$a' c_\theta s_\theta + b'(s_\theta^2 - c_\theta^2) = 0, \tag{13}$$

where

$$a' = \sum_{i=1}^{N} (x_i')^2 - \sum_{i=1}^{N} (y_i')^2 \text{ and } b' = \sum_{i=1}^{N} x_i' y_i'. \tag{14}$$

IF $b'=0$, THE BEST FIT LINE IS EXACTLY VERTICAL OR HORIZONTAL; OTHERWISE,

Equation (13) is a quadratic equation that can be solved for the ratio c_θ / s_θ, giving

$$\frac{c_\theta}{s_\theta} = \frac{\alpha \pm \gamma}{\beta}, \tag{15}$$

where

$$\alpha = a', \beta = 2b', \text{ and } \gamma = \sqrt{\alpha^2 + \beta^2}. \tag{16}$$

Equation (15) can be written as

$$c_\theta = t(\alpha \pm \gamma) \text{ and } s_\theta = t\beta, \tag{17}$$

where t is a constant satisfying the condition $s_\theta^2 + c_\theta^2 = 1$. One of these solutions is the minimum of (2) representing the best-fit line, and the other is a maximum representing the worst-fit line passing through the centroid of the data. It should be noted that this worst-fit line is always perpendicular to the best-fit line since the solutions of Equation (15) (which represent the line slopes) are negative reciprocals of each other. To determine which solution represents the best-fit line (other than by graphical inspection of the data), the second-derivative test can be employed. The following must hold:

$$\frac{\partial^2 Z}{\partial \theta^2} > 0. \tag{18}$$

The second derivative of the error function gives

$$\frac{\partial^2 Z}{\partial \theta^2} = 2\alpha(c_\theta^2 - s_\theta^2) + 4\beta c_\theta s_\theta. \tag{19}$$

After substituting (17) and simplifying, the second-derivative test (18) reduces to

$$t^2\gamma^2(\alpha \pm \gamma) > 0. \tag{20}$$

This forces $\alpha \pm \gamma > 0$, and since $\gamma > \alpha$, the $\alpha + \gamma$ solution represents the best-fit line. Therefore, the best-fit line [in the form of (3) and (17)] is defined by

$$\beta x - (\alpha + \gamma)y = -\rho/t = C, \tag{21}$$

where C is a constant that can be determined (10) by requiring that the line pass through the centroid:

$$C = \beta\overline{x} - (\alpha + \gamma)\overline{y}. \tag{22}$$

Therefore, from (16) and (21), the constants defining the best-fit line in standard form are

$$\begin{aligned} A &= 2b', \\ B &= -\left(a' + \sqrt{(a')^2 + 4(b')^2}\right), \\ C &= A\overline{x} + B\overline{y}. \end{aligned} \tag{23}$$

◇ Example ◇

The following data will be used to demonstrate the results of the method:

i	x_i	y_i
1	0.237	−1.000
2	0.191	−0.833
3	0.056	−0.667
4	0.000	−0.500
5	0.179	−0.333
6	0.127	−0.167
7	0.089	0.000
8	0.136	0.167
9	0.202	0.333
10	0.085	0.500
11	0.208	0.667
12	0.156	0.833
13	0.038	1.000

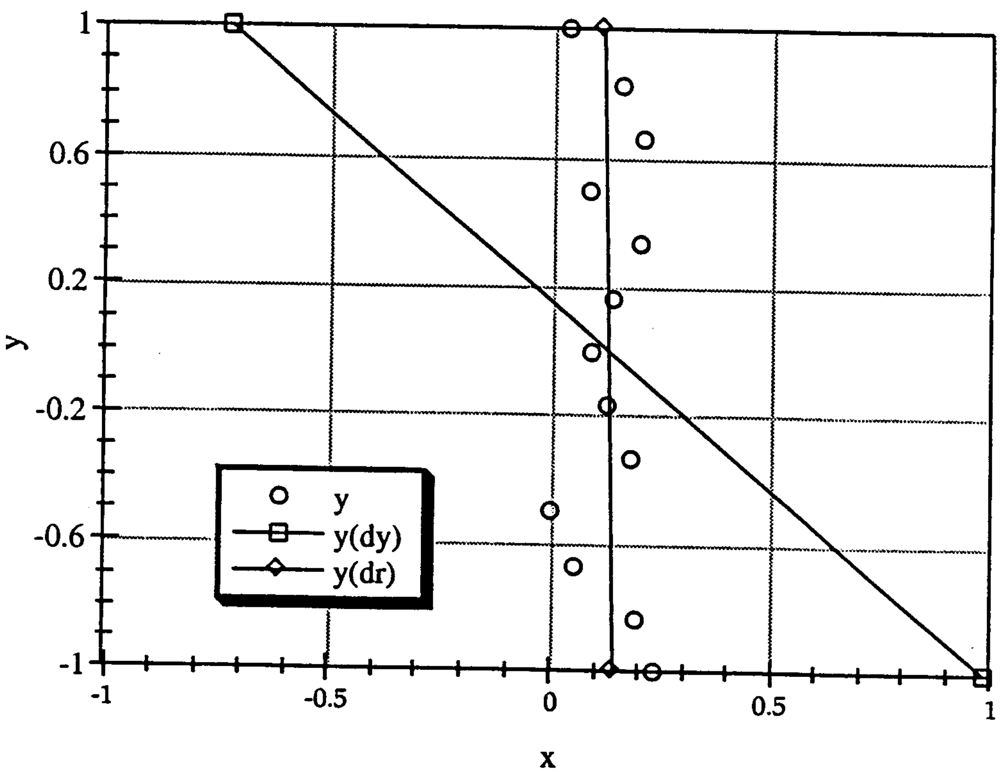

Figure 2. Example line fit.

The centroid of this data is located at

$$\overline{x} = 0.131, \overline{y} = 0.000.$$

Expressed in tems of (14), this gives

$$a' = -4.992 \quad \text{and} \quad b' = -0.075,$$

and so from (23) the final solution is

$$A = -0.149, B = -0.002, \text{ and } \quad C = -0.020.$$

This line $(Ax + By = C)$ is plotted in Figure 2 along with the results from Equation (1) for purposes of comparison. The original $y = m_y x + b_y$ fit afforded by (1) is extremely poor since the data lie near a vertical line.

◊ Conclusions ◊

The method for determining the line passing through a two-dimensional data set and having best least-squares fit was derived. This line's orientation minimizes the sum of the squares of the perpendicular distances between the data and the line. A ρ–θ parameterization of the line resulted in a fairly straightforward analysis. The results, which were expressed in standard ($Ax + By = C$) form, provide a unique, general, and robust solution that is free of degenerate cases. The only possible indeterminacy occurs when $a' = b' = 0$. However, this case can occur only when the data exhibit a perfect circular symmetry (isomorphism under arbitrary rotation). In this case, there is no line of "best fit" because all lines passing through the centroid have a fit that is equally good or bad.

◊ Bibliography ◊

(Charpra and Canale 1988) S. Charpra and R. Canale. *Numerical Methods for Engineers*, 2nd ed. McGraw-Hill, 1988.

(Chatfield 1970) C. Chatfield. *Statistics for Technology*. Penguin Books, 1970.

(Chernov and Ososkov 1984) N. Chernov and G. Ososkov. Effective algorithm for circle fitting. *Computer Physics Communications*, 33:329–333, 1984.

(Ehrig 1985) H. Ehrig. 45th annual meeting of the ACMS. Technical report, ACMS, 1985.

(Karimaki 1992) V. Karimaki. Fast code to fit circular arcs. *Computer Physics Communications*, 69:133–141, 1992.

(Kryszig 1983) E. Kryszig. *Advanced Engineering Mathematics*, 5th ed. John Wiley and Sons, 1983.

(Moura and Kitney 1991) L. Moura and R. Kitney. A direct method for least-squares circle fitting. *Computer Physics Communications*, 64:57–63, 1991.

(Rosenfeld and Sher 1986) A. Rosenfeld and A. Sher. Direction weighted line fitting to edge data. Technical Report CAR-TR-189, Computer Vision Laboratory, University of Maryland, 1986.

(Weiss 1988) I. Weiss. Straight line fitting in a noisy image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 647–652, 1988.