Normal Best Fit 2D Line to a Set of Points

By *Dick Kostelnicek* 01-24-2021

Find m and b for the line

$$y = mx + b \tag{1a}$$

in two dimensional x, y space that best fits, in the normal sense, the set of points

$$x_i \,,\ y_i \,;\ \ i = 1 \dots N \tag{2}$$

The perpendicular distance from a point (2) to the line (1) is

$$d_i = \frac{\mathrm{m}\, x_i - y_i + b}{\sqrt{m^2+1}} \tag{3}$$

The distance may be positive or negative, depending on which side of the line the point resides.

One measure E of how close the set of points (2) cluster about the line (1) is given by the average of all the point's squared distances (3)

$$E = \frac{1}{N}\sum_{i=1}^{N} {d_i}^2 = \frac{\sum_{i=1}^{N}[m x_i - (y_i - b)]^2}{N(m^2+1)} \tag{4a}$$

The extremal of E for a variation of b is

$$\frac{dE}{db} = 0 = \frac{2\,[\mathrm{m}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) + b]}{(m^2+1)} \tag{5}$$

The center of mass of the points (2) is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,,\ \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{6}$$

Inserting (6) into (5) yields

$$b = -m\bar{x} + \bar{y} \tag{7}$$

The y-intercept b will be determined from (7) after m is found.

Inserting (7) into (4a) yields

$$E = \frac{\frac{1}{N}\sum_{i=1}^{N}[m(x_i - \bar{x}) - (y_i - \bar{y})]^2}{(m^2+1)} \tag{4b}$$

Transform coordinates to u, v space so that the new origin is at the center of mass (6).

$$u = x - \bar{x} \,,\ v = y - \bar{y} \tag{8}$$

The line (1a) now passes through the origin of coordinates and becomes

$$v = mu \tag{1b}$$

Introduce the notation

$$S_{uu} = \frac{1}{n}\sum_{i=1}^{N}(u_i{}^2), \quad S_{vv} = \frac{1}{n}\sum_{i=1}^{N}(v_i{}^2), \quad S_{uv} = S_{vu} = \frac{1}{n}\sum_{i=1}^{N}(u_i v_i) \tag{10}$$

Equation (4b) becomes

$$E = \frac{m^2 S_{uu} - 2mS_{uv} + S_{vv}}{m^2+1} \tag{4c}$$

The extremal of (4c) for the variation of m is

$$\frac{dE}{dm} = 0 = \frac{2}{m^2+1}\left[-mE + m\frac{1}{n}S_{uu} - S_{uv}\right] \tag{9}$$

Eqation (9) yields

$$E = S_{uu} - \frac{S_{uv}}{m} \tag{4d}$$

Solving (4c) and (4d) for E and m yields two solutions

$$E_1 = \left(\frac{S_{vv}+S_{uu}}{2}\right) - \sqrt{\left(\frac{S_{vv}-S_{uu}}{2}\right)^2 + S_{uv}{}^2}, \quad m_1 = \frac{\left(\frac{S_{vv}-S_{uu}}{2}\right) - \sqrt{\left(\frac{S_{vv}-S_{uu}}{2}\right)^2 + S_{uv}{}^2}}{S_{uv}} \tag{11}$$

$$E_2 = \left(\frac{S_{vv}+S_{uu}}{2}\right) + \sqrt{\left(\frac{S_{vv}-S_{uu}}{2}\right)^2 + S_{uv}{}^2}, \quad m_2 = \frac{\left(\frac{S_{vv}-S_{uu}}{2}\right) + \sqrt{\left(\frac{S_{vvu}-uu}{2}\right)^2 + S_{uv}{}^2}}{S_{uv}} \tag{12}$$

Since by (4a) $E \geq 0$, the slope m_1 provides the least mean square error E_1. It represents the best fitting line (1a) to the set of points (2). Recall, b in (1a) is given in terms of m by (7).

Note that

$$m_1 = -\frac{1}{m_2} \tag{13}$$

This indicates that the best fitting line is perpendicular to the worst.

Since $E \geq 0$, for both lines then from (11) or (12)

$$S_{vv}S_{uu} \geq S_{uv}{}^2 \tag{14}$$

The solutions (11) and (12) can be cast as a matrix eiegn value problem:

$$\begin{bmatrix} S_{vv} & -S_{sv} \\ -S_{uv} & S_{uu} \end{bmatrix}\begin{bmatrix} 1 \\ m \end{bmatrix} = E\begin{bmatrix} 1 \\ m \end{bmatrix} \tag{15}$$

The square matrix in (15) is symmetric and (14) shows that it is positive definite. Hence, the two eigen values E_1 and E_2 are real and not negative. Furthermore, the eigen vectors $\begin{bmatrix} 1 \\ m_1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ m_2 \end{bmatrix}$ are orthogonal.

If all of the points (2) lie exactly on a single line, the square matrix in (15) is singular and the smllest eigen value $E_1 = 0$. The solution then can be determined by taking any two different points, inserting each into (1a) and solving the resulting two simultaneous equations for b and m.