Best Fit Straight Line

Author: Han de Bruijn
Date: 2006 July

Suppose we have a collection of points in the plane and we want to draw a straight line through these points, in such a way that the line is a "best fit".

Least Squares Method

An equation for a straight line can always be set up as follows:

$$\cos(\theta)(x-p)+\sin(\theta)(y-q)=0$$

Here θ is the angle of the line's normal with the x-axis and (p,q) is an arbitrary point on the line. The distance of an arbitrary point $\vec{r}=(x_k,y_k)$ to the line is given by the length of the projection of the point's vector $\vec{r}$ onto the normal $\vec{n}$ of the line. The latter is given by $\vec{n}=(\cos(\theta),\sin(\theta))$. Hence the length of the projection is:

$$\left|\frac{(\vec{r}\cdot\vec{n})}{(\vec{n}\cdot\vec{n})}\vec{n}\right| = |\cos(\theta)(x_k-p)+\sin(\theta)(y_k-q)|$$

For the straight line to be a "best fit", it will be required that the sum of the weighted squares of all distances shall be a minimum:

$$\sum_k w_k\left[\cos(\theta)(x_k-p)+\sin(\theta)(y_k-q)\right]^2 = \text{minimum}(p,q,\theta)$$

Here the weights w_k are chosen in such a way that $\sum_k w_k = 1$. Working out a bit:

$$\cos^2(\theta)\sum_k w_k(x_k-p)^2+\sin^2(\theta)\sum_k w_k(y_k-q)^2+$$

$$2\sin(\theta)\cos(\theta)\sum_k w_k(x_k-p)(y_k-q) = \text{minimum}(p,q,\theta)$$

Let us solve just one part of the puzzle, namely: how the points (p,q) must be selected in such a way that a minimum may be reached with respect to this choice. For certain parts of the above expression this would mean that:

$$\sum_k w_k(x_k-p)^2 = \text{minimum} \quad \text{and} \quad \sum_k w_k(y_k-q)^2 = \text{minimum}$$

It is sufficient to consider only the expression in x. Differentiate to p:

$$\frac{d}{dp}\left[\sum_k w_k x_k^2 - 2p\sum_k w_k x_k + p^2\sum_k w_k\right] = -2\sum_k w_k x_k + 2p\sum_k w_k = 0$$

$$\Longrightarrow \quad p = \sum_k w_k x_k$$

And in very much the same way:

$$q = \sum_k w_k y_k$$

Define second order momenta with respect to the midpoint as usual:

$$\sigma_{xx} = \sum_k w_k (x_k - p)^2 \quad \text{and} \quad \sigma_{yy} = \sum_k w_k (y_k - q)^2$$

$$\sigma_{xy} = \sum_k w_k (x_k - p)(y_k - q)$$

We can concentrate now on minimalization with respect to the angle θ:

$$\cos^2(\theta)\sigma_{xx} + \sin^2(\theta)\sigma_{yy} + 2\sin(\theta)\cos(\theta)\sigma_{xy} = \text{minimum}(\theta)$$

Extreme values may be found by differentiation to the independent variable:

$$-2\sin(\theta)\cos(\theta)\sigma_{xx} + 2\cos(\theta)\sin(\theta)\sigma_{yy} + 2\cos^2(\theta)\sigma_{xy} - 2\sin^2(\theta)\sigma_{xy} = 0$$

Which leads to the equation:

$$-\sin(2\theta)(\sigma_{xx} - \sigma_{yy}) + \cos(2\theta)2\sigma_{xy} = 0$$

A solution is:

$$\cos(2\theta) = \frac{\sigma_{xx} - \sigma_{yy}}{2}/N \quad \text{and} \quad \sin(2\theta) = \sigma_{xy}/N$$

Where N is a norming factor such that $\cos^2(2\theta) + \sin^2(2\theta) = 1$:

$$N = \sqrt{\left(\frac{\sigma_{xx} - \sigma_{yy}}{2}\right)^2 + \sigma_{xy}^2}$$

But instead of the double angle 2θ we rather need the single angle θ. Or the vector $(\cos(\theta), \sin(\theta))$ instead of $(\cos(2\theta), \sin(2\theta))$. By far the easiest way to obtain the former from the latter is by elementary Euclidian Geometry. Form the sum $(1, 0) + (\cos(2\theta), \sin(2\theta))$ by constructing an equilateral paralellogram and divide the sum vector by its length. For $\cos(2\theta) < 0$, the whole procedure becomes troublesome in the neighbourhood of $\theta = 90^o$. In this case, a fail-safe alternative is to consider instead the triangle $[(0, 0), (1, 0), -(\cos(2\theta), \sin(2\theta))]$. Do the same construction and finally draw the perpendicular to the end result. Something more can be said now about the expression (to be) minimized:

$$\cos^2(\theta)\sigma_{xx} + \sin^2(\theta)\sigma_{yy} + 2\sin(\theta)\cos(\theta)\sigma_{xy} = \text{minimum}(\theta)$$

Here:

$$\cos(2\theta) = 2\cos^2(\theta) - 1 = 1 - 2\sin^2(\theta) \quad \Longrightarrow$$

$$\cos^2(\theta) = \frac{1 + \cos(2\theta)}{2} \quad \text{and} \quad \sin^2(\theta) = \frac{1 - \cos(2\theta)}{2}$$

So:

$$\text{minimum}(\theta) = \frac{\sigma_{xx} + \sigma_{yy}}{}$$