Master of Information and Data Science
DATASCI 281: Computer Vision
Spring 2022

Professor Hany Farid
University of California, Berkeley

Least Squares

You are given a collection of data points (x_i, y_i), for $i \in [1, m]$, and asked to fit a line to these data. The model of a line is parametrized by two parameters the slope a and intercept b:

$$y_i = ax_i + b, \tag{1}$$

Each data point (x_i, y_i) provides one constraint on each model parameter. A total of m such constraints provides an over-constrained system that can be expressed in matrix form as:

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$X\vec{u} = \vec{y}. \tag{2}$$

This linear system of equations will have a solution only when the vector $\vec{y}$ is contained in the column space of matrix X: that is, when $\vec{y}$ can be expressed as a linear combination of the columns of the matrix X. From a geometric perspective, this will occur when all points (x_i, y_i) lie precisely on a line. If, however, the points deviate even slightly from a perfect line, then $\vec{y}$ will not be in the column space of X, and there will be no solution to the above linear

system. It is in these situations that we seek a solution that minimizes some measure of goodness of fit of a line to the points.

The six data points in Figure 1(a), for example, lie close, but not exactly along a line. The solid line minimizes the overall vertical displacement of the points from the line. Minimizing this vertical distance, $ax + b - y$, lends itself to a particularly straightforward optimization, termed least-squares. In matrix form, the vertical distance between points and a line with slope a and intercept b is given by:

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$
$$X\vec{u} - \vec{y}. \tag{3}$$

We seek to minimize the sum of these squared distances:

$$\sum_{i=1}^{m} (ax_i + b - y_i)^2 \;=\; (X\vec{u} - \vec{y})^T (X\vec{u} - \vec{y}) \;=\; \|X\vec{u} - \vec{y}\|^2, \tag{4}$$

where $\|\cdot\|$ denotes vector norm. The sum of squared distances is minizmied by first establishing a quadratic error function in the parameters of the line:

$$E(\vec{u}) \;=\; \|X\vec{u} - \vec{y}\|^2. \tag{5}$$

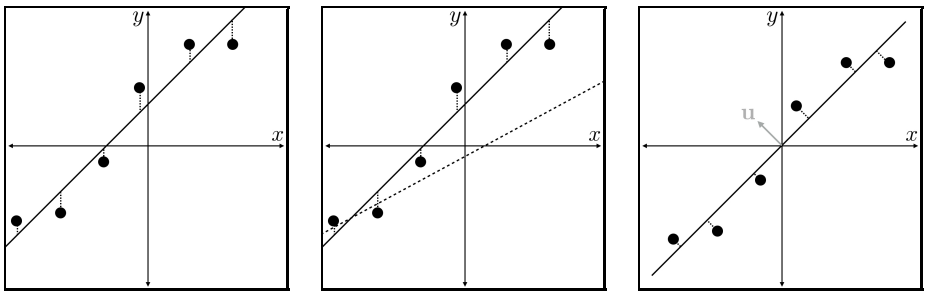
This quadratic error is minimized by differentiating:

$$\frac{dE}{d\vec{u}} \;=\; 2X^T(X\vec{u} - \vec{y}), \tag{6}$$

setting the result equal to zero, and solving for $\vec{u}$:

$$\begin{aligned} 2X^T(X\vec{u} - \vec{y}) &= 0 \\ X^T X\vec{u} &= X^T\vec{y} \\ \vec{u} &= (X^T X)^{-1} X^T \vec{y}, \end{aligned} \tag{7}$$

yielding the least-squares estimation of the line parameters. The matrix $X^T X$ will be singular and hence not invertible if the rows of the matrix X are linearly dependent. Geometrically, this corresponds to one of two



(a) (b) (c)

Figure 1: (a) Least squares; (b) weighted least-squares; and (c) total least-squares.

situations: (1) the m points are identical in which case there is no unique solution to fitting a line to a single point; or (2) the m points lie on a vertical line ($x_1 = x_2 = \ldots = x_m$) in which case $a = \infty$.

This basic framework is, of course, applicable to estimating any model that is linear in their unknown parameters. For example, a parabola $y = ax^2 + bx + c$ can be fit to m points by first constructing the following linear system:

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_m^2 & x_m & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$X\vec{u} = \vec{y}, \tag{8}$$

and then solving using the least-squares solution in Equation (7).

Weighted least-squares: In the standard least-squares formulation, each data point contributes equally to the overall error being minimized. Weighted least-squares allows for a non-uniform treatment of the contribution of each individual point to the overall error. The error function of Equation (5) takes the form:

$$E(\vec{u}) = \|W(X\vec{u} - \vec{y})\|^2, \tag{9}$$

where W is a diagonal $m \times m$ weighting matrix with diagonal elements w_i corresponding to the weight associated with the i^{th} point. A larger weight

places more emphasis on minimizing that point's deviation from the model. This error function is minimized by differentiating:

$$\frac{dE}{d\vec{u}} = 2X^T W^T (WX\vec{u} - W\vec{y}), \tag{10}$$

setting the result equal to zero and solving for $\vec{u}$ to yield the weighted least-squares solution:

$$\begin{aligned} 2X^T W^T (WX\vec{u} - W\vec{y}) &= 0 \\ X^T W^T WX\vec{u} &= X^T W^T W\vec{y} \\ \vec{u} &= (X^T W^T WX)^{-1} X^T W^T W\vec{y} \\ \vec{u} &= (X^T W_2 X)^{-1} X^T W_2 \vec{y}, \end{aligned} \tag{11}$$

where the diagonal elements of the matrix W_2 contain the square of the individual weights w_i. Notice that if W is the identity matrix, then this solution reverts back to the least-squares solution of Equation (5).

The six data points in Figure 1(b) are fit with a line using least-squares (solid line) and weighted least squares (dashed line). In the weighted least-squares fit, the two bottom left most points were weighted disproportionately relative to the other points. Notice how the line minimizes the error for these points at the price of significantly higher error for the remaining points.

Total Least-Squares: In the above line fitting examples it is the *vertical* distances between each data point and the line that are minimized, Figure 1(a). This gives special distinction to the y-component of the data. In some cases, it is preferable to make no such distinction and to instead minimize the *perpendicular* distance between each data point and the line, Figure 1(c). Such a minimization is termed total least-squares.

For ease of exposition, but without loss of generality, assume that the data points (x_i, y_i) are centered about the origin. In this case, the line can be parameterized by a unit vector $\vec{u}$. With this parameterization, the perpendicular distance between each data point and a line is the inner product $\begin{pmatrix} x_i & y_i \end{pmatrix} \cdot \vec{u}$. In matrix form, the perpendicular distances between each of m points and the line is given by:

$$X\vec{u} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_m & y_m \end{pmatrix} \begin{pmatrix} u_x \\ u_y \end{pmatrix}. \tag{12}$$

We seek to minimize the sum of these squared distances:

$$\sum_{i=1}^{m}(x_i u_x + y_i u_y)^2 \;=\; (X\vec{u})^T(X\vec{u}) \;=\; \|X\vec{u}\|^2. \tag{13}$$

The sum of squared distances is minizmied by establishing a quadratic error function in the parameters of the line:

$$E(\vec{u}) \;=\; \|X\vec{u}\|^2, \tag{14}$$

which we seek to minimize subject to the constraint that $\vec{u}$ is unit length, $\|\vec{u}\| = 1$. Unlike least-squares minimization, total least-squares requires a *constrained* minimization to ensure that the estimated solution $\vec{u}$ is unit-length.[1]

The constrained minimization can be framed using a Lagrange multiplier[2] or equivalently as an unconstrained minimization of the Rayleigh quotient:

$$R(\vec{u}) \;=\; \frac{\|X\vec{u}\|^2}{\|\vec{u}\|^2} \;=\; \frac{\vec{u}^T X^T X \vec{u}}{\vec{u}^T\vec{u}}. \tag{15}$$

This error function is minimized by differentiating:

$$\frac{dR}{d\vec{u}} \;=\; \frac{(2X^TX\vec{u})(\vec{u}^T\vec{u}) - (\vec{u}^TX^TX\vec{u}^T)(2\vec{u})}{(\vec{u}^T\vec{u})^2}, \tag{16}$$

and setting the result equal to zero:

$$\begin{aligned}
\frac{(2X^TX\vec{u})(\vec{u}^T\vec{u}) - (\vec{u}^TX^TX\vec{u}^T)(2\vec{u})}{(\vec{u}^T\vec{u})^2} &= 0 \\
(X^TX\vec{u})(\vec{u}^T\vec{u}) &= (\vec{u}^TX^TX\vec{u}^T)\vec{u} \\
X^TX\vec{u} &= \frac{(\vec{u}^TX^TX\vec{u}^T)}{(\vec{u}^T\vec{u})}\vec{u}.
\end{aligned} \tag{17}$$

The scalar-valued term on the right-hand side that multiplies $\vec{u}$ is itself the Rayleigh quotient being minimized:

$$X^TX\vec{u} \;=\; R(\vec{u})\vec{u}. \tag{18}$$

[1] Without the constraint that $\|\vec{u}\| = 1$, the zero vector $\vec{u} = \vec{0}$ would trivially minimize the error function $E(\vec{u}) = \|X\vec{u}\|^2$.

[2] The constrained minimization can be solved by minimizing the error function: $L(u) = \|X\vec{u}\|^2 + \lambda(\|\vec{u}\|^2 - 1)$, where λ is the Lagrange multiplier.

This constraint on the solution $\vec{u}$ tells us two things: (1) the vector $\vec{u}$ that minimizes (or maximizes) $R(\vec{u})$ is an eigenvector[3] of the matrix $X^T X$; and (2) since the eigenvalue for the eigenvector $\vec{u}$ is $R(\vec{u})$, the vector $\vec{u}$ that *minimizes* $R(\vec{u})$ must be the *minimial* eigenvalue eigenvector. The total least-squares estimate is therefore the minimial eigenvalue eigenvector of the matrix $X^T X$.

[3]The non-zero eigenvector $\vec{e}$ of a square matrix M satisfies $M\vec{e} = \lambda\vec{e}$ where λ is the eigenvalue. That is, the product of a matrix and its eigenvector $\vec{e}$ yields a scaled version of $\vec{e}$, by an amount λ.