Prof. Dr. Jens Honore Walther
Dr. Julija Zavadlav
ETH Zentrum, CLT
CH-8092 Zürich

Solution Set 2

Issued: 18.03.2021

Question 1: Linear least squares on 3D data

a) The problem is formulated in the following way: given a set of N data $\{x_i, y_i, z_i\}_{i=1}^N$, we wish to fit to a linear function of 3 parameters α, β and γ given by

$$f(x_i, y_i) = \alpha + \beta x_i + \gamma y_i \approx z_i,$$

which can be rewritten as

$$\begin{bmatrix} f(x_1,y_1) \\ f(x_2,y_2) \\ \vdots \\ f(x_N,y_N) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{bmatrix}}_{A} \cdot \underbrace{\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}}_{\mathbf{w}} \approx \underbrace{\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix}}_{\mathbf{z}}.$$

The squared L_2 error norm of such an approximation reads as follows:

$$\begin{aligned} E^2 &= \sum_{i=1}^N \big(f(x_i,y_i) - z_i\big)^2 = \sum_{i=1}^N (\alpha + \beta x_i + \gamma y_i - z_i)^2 \\ &= (A\mathbf{w} - \mathbf{z})^T (A\mathbf{w} - \mathbf{z}). \end{aligned}$$

α, β and γ yield least error when the derivatives of the error are zero:

$$A^T A \mathbf{w} = A^T \mathbf{z}$$

Let us calculate explicitly left-hand side matrix $A^T A$ and right-hand side vector $A^T \mathbf{z}$:

$$H = A^T A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \\ y_1 & y_2 & \dots & y_N \end{bmatrix} \times \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^N x_i & \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N y_i^2 \end{bmatrix}$$

$$Z = A^T \mathbf{z} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \\ y_1 & y_2 & \dots & y_N \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N z_i \\ \sum_{i=1}^N z_i x_i \\ \sum_{i=1}^N z_i y_i \end{bmatrix}$$

To solve this linear system, it suffices to perform left-matrix-multiplication of the inverse matrix with both sides of the equation:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \mathbf{w}^* = H^{-1}Z \tag{1}$$

An applied counterpart to this question can be found in Notebook 2.1.

Question 2: Dependency of the LSQ fit on the noise

Let's make some rough estimates of the behavior of the error of the parameters given N data points and noise ε distributed normally with zero mean and σ^2 variance: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Define $y_i = \alpha_0 + \beta_0 x_i$ as the unperturbed data, and noisy data as $y_i^* = y_i + \varepsilon_i$.

Let $f(x) = \alpha + \beta x$ be the linear model which we will fit to the given data. Since the data y_i were created using the linear equation $y_i = \alpha_0 + \beta_0 x_i$, it is easy to deduce that the coefficients of the LSQ fit of this data will be $\alpha = \alpha_0$ and $\beta = \beta_0$. Moreover, let α^* and β^* be the coefficients of the LSQ fit using the perturbed data y^*. We denote the model that corresponds to perturbed data by $f^*(x) = \alpha^* + \beta^* x$. We want to understand how the difference,

$$e(x) = f(x) - f^*(x) \ , \tag{2}$$

behaves as a function of N.

First we observe that $e(x) = \alpha_0 - \alpha^* + (\beta_0 - \beta^*)x$. Thus, we have to find how the quantities $\alpha_0 - \alpha^*$ and $\beta_0 - \beta^*$ behave as a function of N.

Using the notation of the previous section, the coefficients of the LSQ fit using the unperturbed data (x_i, y_i) are given by

$$\begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} = H^{-1} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{bmatrix} \ , \tag{3}$$

or, equivalently,

$$\alpha_0 = H_{11}^{-1} \sum_{i=1}^N y_i + H_{12}^{-1} \sum_{i=1}^N x_i y_i \ , \tag{4}$$

$$\beta_0 = H_{21}^{-1} \sum_{i=1}^N y_i + H_{22}^{-1} \sum_{i=1}^N x_i y_i \ , \tag{5}$$

where

$$H = \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix} \ . \tag{6}$$

and its inverse is given by

$$H^{-1} = \frac{1}{N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix} \ . \tag{7}$$

Similarly, the coefficients that correspond to the perturbed data are given by

$$\begin{aligned} \alpha^* &= H_{11}^{-1} \sum_{i=1}^{N} y_i^* + H_{12}^{-1} \sum_{i=1}^{N} x_i y_i^* \\ &= H_{11}^{-1} \sum_{i=1}^{N} (y_i + \varepsilon_i) + H_{12}^{-1} \sum_{i=1}^{N} x_i (y_i + \varepsilon_i) \\ &= \alpha_0 + H_{11}^{-1} \sum_{i=1}^{N} \varepsilon_i + H_{12}^{-1} \sum_{i=1}^{N} x_i \varepsilon_i \ , \end{aligned} \tag{8}$$

thus,

$$\alpha^* - \alpha_0 = H_{11}^{-1} \sum_{i=1}^{N} \varepsilon_i + H_{12}^{-1} \sum_{i=1}^{N} x_i \varepsilon_i \ . \tag{9}$$

Notice that $H_{ij} = c_{ij} N$ because

$$\sum_{i=1}^{N} x_i = N\bar{x} \ , \text{ and } \sum_{i=1}^{N} x_i^2 = N\overline{x^2} \ , \tag{10}$$

and noting that the average $\bar{x}$ and $\overline{x^2}$ is independent of N, we have that $H_{ij}^{-1} = \frac{1}{N}\tilde{c}_{ij}$, where c_{ij} and $\tilde{c}_{ij}$ is independent of N. Using this fact in eq. (9) we get,

$$\alpha^* - \alpha_0 \approx \frac{\tilde{c}_{11}}{N} \sum_{i=1}^{N} \varepsilon_i + \frac{\tilde{c}_{12}}{N} \sum_{i=1}^{N} x_i \varepsilon_i \tag{11}$$

where the constant c_{ij} are independent of N. Since $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and for the the variance we know that $\operatorname{var}(aX + bY) = a^2 \operatorname{var}[X] + b^2 \operatorname{var}[Y]$, we conclude from the Central Limit Theorem, that the difference $\alpha^* - \alpha_0$ will follow the distribution:

$$\begin{aligned} \alpha^* - \alpha_0 &\sim \tilde{c}_{11} \mathcal{N}(0, \frac{\sigma^2}{N}) + \tilde{c}_{12} \mathcal{N}(0, \frac{\sigma^2 \sum x_i^2}{N^2}) \\ &\sim \mathcal{N}(0, \frac{\tilde{c}_{11}^2 \sigma^2}{N} + \frac{\tilde{c}_{12}^2 \sigma^2 \sum x_i^2}{N^2}) \\ &\sim \mathcal{N}(0, \frac{c\sigma^2}{N}), \end{aligned} \tag{12}$$

where we again used that $\sum x_i^2 = N\overline{x^2}$.

Working in the same way we get a similar result for the difference $\beta_0 - \beta^*$. Finally, the difference (2) follows a normal distribution,

$$e(x) \sim c(x) \mathcal{N}(0, \frac{\sigma^2}{N}) \ , \tag{13}$$

where $c(x)$ is independent of N. Therefore the error between the LSQ fit to a noisy and a noiseless data set is a random variable that follows normal distribution with mean zero and variance $\approx \frac{\sigma^2}{N}$. We therefore conclude that the difference in the approximation using noisy versus exact data scales with the noise strength (σ) linearly and with inverse square root of the number of samples (N).