Numerical Analysis: Least Squares Approximation – Part 3

Computer Science, Ben-Gurion University

2018/2019, Fall Semester
1 Deriving the Solution for Linear Least Squares

2 Nonlinear Least Squares
Deriving the Least-squares Estimate in a Linear Model

\[
\begin{align*}
H & \triangleq \begin{bmatrix} H_1^T & \cdots & H_N^T \end{bmatrix}^T \in \mathbb{R}^{(Nd) \times k} \\
y & \triangleq \begin{bmatrix} y_1^T & \cdots & y_N^T \end{bmatrix}^T \in \mathbb{R}^{(Nd) \times 1} \\
r & \triangleq \begin{bmatrix} r_1^T & \cdots & r_N^T \end{bmatrix}^T = H\theta - y \in \mathbb{R}^{(Nd) \times 1}
\end{align*}
\]

\[
\hat{\theta}_{LS} \triangleq \arg \min_{\theta \in \mathbb{R}^k} \sqrt{\sum_{i=1}^{N} \|r_i\|^2} = \arg \min_{\theta \in \mathbb{R}^k} \sum_{i=1}^{N} \|r_i\|^2
\]

\[
\begin{align*}
H^T H \hat{\theta}_{LS} &= H^T y
\end{align*}
\]

We want to show that a minimizer satisfies

\[
\begin{align*}
H^T H \hat{\theta}_{LS} &= H^T y
\end{align*}
\]

and is unique if \( \text{rank}(H) = k; \) i.e., if \( H^T H \) is invertible.
Fact

Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^k$. Suppose that these two column vectors are possibly dependent on a scalar $t$. Then, similarly to the scalar case,

$$\frac{d}{dt} (\mathbf{a}^T \mathbf{b}) = \frac{d\mathbf{a}^T}{dt} \mathbf{b} + \mathbf{a}^T \frac{d\mathbf{b}}{dt} = \mathbf{b}^T \frac{d\mathbf{a}}{dt} + \frac{d\mathbf{b}^T}{dt} \mathbf{a}$$
Proof.

\[
\frac{d}{dt} (\mathbf{a}^T \mathbf{b}) = \frac{d}{dt} \left( \sum_{i=1}^{k} a_i b_i \right) = \sum_{i=1}^{k} \frac{d}{dt} (a_i b_i) = \sum_{i=1}^{k} (\frac{d}{dt} a_i) b_i + a_i \left( \frac{d}{dt} b_i \right) \\
= \frac{d\mathbf{a}^T}{dt} \mathbf{b} + \mathbf{a}^T \frac{d\mathbf{b}}{dt}
\]
Corollary

\[
\frac{d}{dt} \left( \| \mathbf{a} \|^2 \right) = \frac{d}{dt} (\mathbf{a}^T \mathbf{a}) = \frac{d\mathbf{a}^T}{dt} \mathbf{a} + \mathbf{a}^T \frac{d\mathbf{a}}{dt} = 2 \frac{d\mathbf{a}^T}{dt} \mathbf{a} = 2 \mathbf{a}^T \frac{d\mathbf{a}}{dt}
\]
Corollary

Let \( A \) and \( B \) be two matrices such that \( AB \) is well defined. Suppose that \( A \) and \( B \) are possibly dependent on a scalar \( t \). Then, similarly to the scalar case,

\[
\frac{d}{dt} (A^T B) = \frac{dA^T}{dt} B + A^T \frac{dB}{dt}
\]
Let \( \theta = [\theta_1 \ldots \theta_k]^T \) and define

\[
f : \mathbb{R}^k \rightarrow \mathbb{R}_{\geq 0} \quad f : \theta \mapsto ||y - H\theta||^2
\]

Now:

\[
f(\theta) = ||y - H\theta||^2 = (y - H\theta)^T(y - H\theta)
\]

\[
= ||y||^2 - \theta^T H^T y - y^T H\theta + \theta^T H^T H\theta
\]

\[
= ||y||^2 - \theta^T H^T y - \underbrace{\theta^T H^T y}_{\text{transposing a scalar}} + \theta^T H^T H\theta
\]

\[
= ||y||^2 - 2\theta^T H^T y + \theta^T H^T H\theta
\]

\[
\text{const w.r.t. } \theta
\]

\[
\text{arg min } f(\theta) = \text{arg min } \theta^T H^T H\theta - 2\theta^T H^T y
\]

We need to compute the \( k \) partial derivatives and set them to 0:

\[
\frac{\partial}{\partial \theta_i} (\theta^T H^T H\theta - 2\theta^T H^T y) = 0 \quad i = 1, \ldots, k
\]
Let us derive each of these two terms.

\[ \frac{\partial}{\partial \theta_i} (\theta^T H^T H \theta - 2 \theta^T H^T y) = \frac{\partial}{\partial \theta_i} (\theta^T H^T H \theta) - 2 \frac{\partial}{\partial \theta_i} (\theta^T H^T y) \]
\[
\frac{\partial}{\partial \theta_i} (\theta^T H^T H \theta) = \left( \frac{\partial}{\partial \theta_i} (\theta^T H^T H) \right) \theta + \theta^T H^T H \frac{\partial}{\partial \theta_i} \theta
\]
\[
= \left( \frac{\partial}{\partial \theta_i} \theta^T \right) H^T H \theta + \theta^T H^T H \frac{\partial}{\partial \theta_i} \theta = (H^T H)_{:,i} \theta + \theta^T (H^T H)_{:,i}
\]
\[
H^T H \text{ is sym} \quad (H^T H)_{:,i}^T \theta + \theta^T (H^T H)_{:,i} = 2 \theta^T (H^T H)_{:,i}
\]
Deriving the Solution for Linear Least Squares

\[-2 \frac{\partial}{\partial \theta_i} (\theta^T H^T y) = -2 \left( \frac{\partial}{\partial \theta_i} \theta^T \right) H^T y = -2 \left( H^T \right)_{i,:} \, y\]

\[= -2 \begin{pmatrix} H_{:,i}^T \end{pmatrix}_{\text{col } i} \, y = -2 y^T H_{:,i}^T \]
Deriving the Solution for Linear Least Squares

\[
\frac{\partial}{\partial \theta_i} (\theta^T H^T H \theta - 2\theta^T H^T y) = \frac{\partial}{\partial \theta_i} (\theta^T H^T H \theta) - 2 \frac{\partial}{\partial \theta_i} (\theta^T H^T y)
\]

\[
= 2\theta^T (H^T H)_{:,i} - 2y^T H_{:,i}
\]

Setting to zero all these \(k\) partial derivatives:

\[
2\theta^T (H^T H)_{:,i} = 2y^T H_{:,i} \quad i = 1, \ldots, k
\]

\[
\theta^T (H^T H)_{:,i} = y^T H_{:,i} \quad i = 1, \ldots, k
\]

Thus,

\[
\theta^T H^T H = y^T H
\]

or, equivalently:

\[
H^T H \theta = H^T y
\]

where we used the symmetry of \(H^T H\).
The Normal Equations

- The equations captured by the matrix equation
  \[ H^T H \theta = H^T y \]

  are called the normal equations.
- A positive-definite matrix is always invertible.

**Proof.**

For every square matrix \( A \),

\[ Ax = 0 \Rightarrow x^T Ax = 0. \]

Thus, if the matrix is PD then \( Ax = 0 \) implies \( x = 0 \). But a square matrix \( A \) is invertible \( \iff (Ax = 0 \Rightarrow x = 0) \). Thus, a PD matrix is invertible.

- The symmetric matrix \( H^T H \) is always positive semi-definite. It is positive definite (and thus invertible) iff \( H \) is full rank.
Nonlinear Least Squares

In linear least squares, the model was

\[ y = h(x)\theta = \sum_{j=1}^{k} h_i(x)\theta_j. \]

More generally, we may have a nonlinear model, i.e., one that cannot be written that way.

Example
Consider \( f : \mathbb{R} \to \mathbb{R} \) where \( f(x) = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x) \).

Example
Consider \( f : \mathbb{R} \to \mathbb{R} \) where \( f(x) = \theta_1 \sin(x + \theta_2) + \theta_3 \cos(x + \theta_4). \)
Nonlinear least-squares problems are usually much harder than the linear case. In certain such problems, however, we can apply some transformation that makes the problem linear.
Example

\( y = \theta_2 \exp(\theta_1 x) \) can be written as

\[
\tilde{y} = \log y = \theta_1 x + \log \theta_2 = \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \log \theta_2 \end{bmatrix} = \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \tilde{\theta}_2 \end{bmatrix}
\]

where \( \tilde{y} = \log y \) and \( \tilde{\theta}_2 = \log \theta_2 \). So use linear least squares to solve for the optimal \( \begin{bmatrix} \theta_1 & \tilde{\theta}_2 \end{bmatrix}^T \) and then set \( \theta_2 = \exp(\tilde{\theta}_2) \) (no need to change \( \theta_1 \)).
Example

$y = \theta_1 \log(x) + \theta_2$ is already linear in $\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$:

$$y = \begin{bmatrix} \log x & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

so there is no need to perform a transformation.
Example

$y = 1/(\theta_1 x + \theta_2)$ can be written as

$$\tilde{y} = \frac{1}{y} = \theta_1 x + \theta_2 = \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

(so it enough to transform the $y_i$'s).
Example

\[ y = \frac{x}{(\theta_1 x + \theta_2)} \text{ can be written as } \]

\[ \tilde{y} = \frac{1}{y} = \frac{\theta_1 x + \theta_2}{x} = \begin{bmatrix} 1 & \frac{1}{x} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \]