Chapter 4

Least Squares Problems

Data fitting (or parameter estimation) is an important technique used for modeling in many areas of disciplines. The problem can be described as follows: Assuming a physical phenomenon is modeled by a relationship

$$y = f(z; x_1, \ldots, x_n) \tag{4.1}$$

where f is a prescribed function determined up to values of $x_1, \ldots, x_n$, z is the control variable and y is the expected response to z. After m experiments $(m \geq n)$, we have collected m observed quantities $(z_i, y_i), i = 1, \ldots, m$ Due to measurement errors (called noise), however, (z_i, y_i) may not satisfy (4.1) exactly. We, therefore, seek to adjust the parameters $x_1, \ldots, x_n$ so that the expression

$$g(x_1, \ldots, x_n) := \sum_{i=1}^{m} \|y_i - f(z_i; x_1, \ldots, x_n)\|_F^2 \tag{4.2}$$

is minimized.

As a necessary condition that $(x_1, \ldots, x_n)$ be a minimizer of (4.2), we need to solve the normal equation

$$\bigtriangledown g(x_1, \ldots, x_n) = 0 \tag{4.3}$$

A good reference on this topic can be found in the book, Solving Least Squares Problems, by Lawson and Hanson.

4.1 Linear Least Square Problems

Example. (Polynomial Fitting) Let (z_i, y_i), $i = 1, \ldots, m$ be the observed quantities. Suppose the function f in (4.1) is an $(n-1)$-th degree polynomial

$$f(z; x_1, \ldots, x_n) = x_1 z^{n-1} + \ldots + x_{n-1} z + x_n. \tag{4.4}$$

Then data fitting problem is to solve the system

$$\begin{bmatrix} z_1^{n-1} & z_1^{n-2} & \dots & z_1 & 1 \\ z_2^{n-1} & & & & \\ \vdots & & & & \\ & & & & \\ z_m^{n-1} & z_m^{n-1} & \dots & z_m & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_m \end{bmatrix} \tag{4.5}$$

for the coefficients $(x_1, \dots, x_n)$. Usually the system (4.5) is overdetermined, so we consider the least square problem

$$\min_{x \in R^n} \|Ax - b\|_2^2$$

where $A \in R^{m \times n}$, $b \in R^m$ are known quantities. Problem (4.1) where the function f is linear in the parameters $x_1, \dots, x_n$ is called a linear least square problem.

Definition 4.1.1 *Given $A \in R^{m \times n}$, we define*

$$\begin{aligned} R(A) : &= \{y \in R^m | y = Ax \text{ for some } x \in R^n\} && (4.6) \\ &= \text{ The range space of } A. \\ N(A) : &= \{x \in R^n | Ax = 0\} = \text{ The null space of } A. && (4.7) \end{aligned}$$

Theorem 4.1.1 *For every $z \in R^m$, there exist a unique $y \in R(A)$, and unique $w \in N(A^T)$ such that $z = y + w$. That is*

$$R^m = R(A) \oplus N(A^T).$$

(pf): It suffices to prove $R(A)^\perp = N(A^T)$. Now $w \in N(A^T) \Leftrightarrow A^T w = 0 \Leftrightarrow x^T A^T w = 0$ for all $x \in R^n \Leftrightarrow w \perp Ax$ for all $x \in R^n \Leftrightarrow R(A)$. $\oplus$

Theorem 4.1.2 *The linear least squares problem (4.1) has a solution for every b. The solution is unique if and only if $N(A) = \{0\}$.*

(pf): By (4.1.1, we may rewrite $b = b_1 + b_2$ with $b_1 \in R(A)$ and $b_2 \in N(A^T)$. Now $Ax - b = (Ax - b_1) - b_2$. Since $Ax - b_1 \in R(A)$, $Ax - b_2 \perp b_2$. So $\|Ax - b\|_2^2 = \|Ax - b_1\|_2^2\|_2^2$. Note that b_2 is fixed whenever b is given. Thus $\|Ax - b\|$ is minimized if and only if $Ax = b_1$. But $b_1 \in R(A)$. So there exists $x_0 \in R^n$ such that $Ax_0 = b_1$ and $\min \|Ax - b\| = \|b_2\|$. If $N(A) = (0)$, then the solution of $Ax = b_1$ is unique. The converse is true also. $\oplus$

Remark. In the proof of (4.5), we realize that the minimizer x_0 of (4.1) must satisfy the equation $Ax_0 - b = -b_2 \in N(A^T)$. It follows that x_0 must satisfy the equation (necessary condition)

$$A^T Ax = A^T b.$$

This linear system is called the normal equation. If, in particular, $A^T A$ is nonsingular (This is so if A is of full column rank), then the unique solution of (4.1) is given by $x = (A^T A)^{-1} Ab$.

We discussed in Section 2.5 that for any matrix $A \in R^{m\times n}$, there exist an orthogonal matrix $Q \in R^{m\times m}$ and $R = \begin{bmatrix} R_1 \\ \\ 0 \end{bmatrix} \in R^{m\times n}$ with R_1 upper triangular such that

$$Q^T A = R. \tag{4.8}$$

Note that orthogonal transformations leave the norm $\|x\|_2$ of a vector x invariant ($\|Vx\|_2 = \sqrt{x^T V^T V x} = \|x\|_2$). Denoting $Q^T b := [h_1^T, h_2^T]^T$, we see that

$$\|Ax-b\|_2^2 = \|Q^T(Ax-b)\|_2^2 = \| \begin{bmatrix} R_1 \\ \\ 0 \end{bmatrix} x - \begin{bmatrix} h_1 \\ \\ h_2 \end{bmatrix} \|_2^2 = \|R_1 x - h_1\|_2^2 + \|h_2\|_2^2.$$

Hence $\|Ax-b\|_2$ is minimized if x is chosen so that

$$R_1 x = h_1. \tag{4.9}$$

If we assume that A is of full column ran, then R_1 is nonsingular and (4.9) has a unique solution.

4.2 Singular Value Decomposition

Given any $A \in R^{m\times n}$, $A^T A \in R^{n\times n}$ is symmetric and positive semi-definite. So $A^T A$ has a complete set of orthogonal eigenvectors and all eigenvalues of $A^T A$ are non-negative. We have similar situation for the matrix $AA^T \in R^{m\times m}$. Let the positive eigenvalues of $A^T A$ be denoted as $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2 > 0$.

Lemma 4.2.1 *The two matrices $A^T A$ and AA^T have the same positive eigenvalues. More precisely, let $u_j \in R^n$ be the normalized eigenvector of $A^T A$ associated with the eigenvalue σ_j^2. Then $Au_j \in R^m$ is an eigenvector of AA^T whose corresponding eigenvalue is also σ_j^2. If we take $v_j := Au_j/\sigma_j$, then v_j is normalized.*

(pf): If $A^T A u_j = \sigma_j^2 u_j$, then $(AA^T)Au_j = \sigma_j^2(Au_j)$. Also, since u_j is normalized, $u_j^T A^T A u_j = \|Au_j\|_2^2 = \sigma_j^2$.

Definition 4.2.1 *The numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq, \sigma_r > \sigma_{r+1} = \ldots = 0$ are called the singular values of A.*

Definition 4.2.2 *The normalized eigenvectors $u_1, \ldots, u_n$ (or, $v_1, \ldots, v_m$) of $A^T A$ (or, AA^T) are called the right (or, left) singular vectors of A.*

Remark. Let $U := [u_1, \ldots, u_n] \in R^{n\times n}$ where columns are orthonomal eigenvectors of $A^T A$. Define $V := [v_1, \ldots, v_m] \in R^{m\times m}$ where

(1) For $j = 1, \ldots, r, v_j = Au_j/\sigma_j$, and

(2) $\{v_{r+1}, \ldots, v_m\}$ are orthonormal eigenvectors corresponding to the zero eigenvalue of AA^T.

Then both U and V are orthogonal matrices. Furthermore,

Theorem 4.2.1 *With U and V given above, we have*

$$A = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T \tag{4.10}$$

where $\Sigma :=$ diag $\{\sigma_1, \ldots, \sigma_r\}$.

(pf): Write $U = [U_1, U_2],\ V = [V_1, V_2]$. Then

$$V^T AU = \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} A[U_1, U_2] = \begin{bmatrix} V_1^T AU_1 & V_1 AU_2 \\ V_2^T AU_1 & V_2^T AU_2 \end{bmatrix}.$$

Note that $AU_2 = 0,\ V_2^T AU_1 = V_2^T V_1 \Sigma = 0$ and $V_1^T AU_1 = \Sigma$ by the choice of V.

Definition 4.2.3 *The composition in (4.10) is called the singular value decomposition of A.*

Theorem 4.2.2 *Let $A \in R^{m\times n}$ and have singular value decomposition given by (4.10). Then the vector $\tilde{x}$ given by*

$$\tilde{x} := U \begin{bmatrix} \Sigma^{-1}, 0 \\ 0, 0 \end{bmatrix} V^T b \tag{4.11}$$

minimizes $\|Ax - b\|_2$ among all $x \in R^n$. Moreover, if $\overline{x}$ is another least square solution of (4.1), then $\|\tilde{x}\|_2 \le \|\overline{x}\|_2$.

(pf): We observe $\|Ax - b\|_2^2 = \|V^T(Ax - b)\|_2^2 = \|V^T AUU^T x - V^T b\|_2^2 = \| \begin{bmatrix} \Sigma, 0 \\ 0, 0 \end{bmatrix} z - c\|_2^2$ with $z := U^T x$ and $c := V^T b$. Write $z := \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ and $c := \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$ with $z_1, c_1 \in R^r$. Then $\|Ax - b\|_2^2 = \| \begin{bmatrix} \Sigma z_1 - c_1 \\ c_2 \end{bmatrix} \|_2^2 = \|\Sigma z_1 - c_1\|_2^2 + \|c_2\|_2^2$. Obviously, $\|Ax - b\|$ is minimized if and only if $z_1 = \Sigma^{-1} c_1$. Note that z_2 can be arbitrary. Suppose we choose $\tilde{z} = \begin{bmatrix} \Sigma^{-1} c_1 \\ 0 \end{bmatrix}$. Then $\tilde{x} = U\tilde{z} = U \begin{bmatrix} \Sigma^{-1} c_1 \\ 0 \end{bmatrix} = U \begin{bmatrix} \Sigma^{-1}, 0 \\ 0, 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = U \begin{bmatrix} \Sigma^{-1}, 0 \\ 0, 0 \end{bmatrix} V^T b$. For any other least squares solution $\overline{x}$, the corresponding $\overline{z}$ must be of the form $\begin{bmatrix} \Sigma^{-1} c_1 \\ z_2 \end{bmatrix}$. Thus $\|\overline{x}\|_2^2 = \|U\overline{z}\|_2^2 = \|\overline{z}\|_2^2 = \|\Sigma^{-1} c_1\|_2^2 + \|z_2\|_2^2 \ge \|\tilde{z}\|_2^2 = \|\tilde{x}\|_2^2$.

4.3 Nonlinear Least Squares Problems

Nonlinear least squares problems (4.2) may be solved by general unconstrained minimization techniques. (cf: Practical Optimization by Gill, Murray and Wright). The special form of (4.2), however, make it worthwhile to use methods designed specifically for the least squares problems.

For demonstrate, we shall consider the following problem

$$\text{Minimize } F(x) = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2 = \frac{1}{2}\|f(x)\|_2^2 \tag{4.12}$$

where $x \in R^n$, $f_i : R^n \to R$ and $f(x) = (f_1(x), \ldots, f_m(x))^T$ are smooth functions of x. A necessary condition for $\overline{x}$ of being a critical point is

$$g(x) := \bigtriangledown F(\overline{x}) = 0 \tag{4.13}$$

Toward this, we calculate

$$g(x) := \bigtriangledown F(x) = J(x)^T f(x) \text{ (Justify this!)} \tag{4.14}$$

where

$$J(x) := \frac{\partial f}{\partial x} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}, \frac{\partial f_1}{\partial x_2}, \ldots, \frac{\partial f_1}{\partial x_n} \\ \\ \frac{\partial f_m}{\partial x_1}, \frac{\partial f_m}{\partial x_2}, \ldots, \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{4.15}$$

is the $m \times n$ Jacobian matrix of f. Note that $g : R^n \to R^n$. So we may apply the Newton-Ralphson method to solve $g(x) = 0$, i.e., let $x^{(k)}$ denote the current estimate of $\overline{x}$, we calculate a Newton step p_k by solving

$$g'(x^{(k)})p_k = -g(x^{(k)}) \tag{4.16}$$

and then update

$$x^{(k+1)} = x^{(k)} + \tau_k p_k, \tag{4.17}$$

where the parameter τ_k is chosen to guarantee the sequence $\{F(x^{(k+1)}\}$ is strictly monotone decreasing. The matrix $g'(x)$ (the Hessian of F) is calculated as

$$g'(x) = J_f(x)^T J_f(x) + \sum_{i=1}^{m} H_i(x) f_i(s) \text{ (Justify this !)} \tag{4.18}$$

where

$$H_i(x) := [\frac{\partial^2 f_i}{\partial x_s \partial x_t}] \tag{4.19}$$

is the Hessian matrix of $f_i(x)$.

Generally, if $\|f_i\|$ tends to zero as $x^{(k+1)}$ approaches the solution, the second matrix in (4.18) also tends to zero. Thus the Newton direction is approximated by the solution of the equation

$$J(x^{(k)})^T J(x^{(k)}) P_k = -J(x^{(k)})^T f(x^{(k)}). \tag{4.20}$$

We note that the solution of (4.20) is the solution of the linear least squares problem

$$\text{Minimize } \frac{1}{2}\|J(x^{(k)})p + f(x^{(k)})\|_2^2, \tag{4.21}$$

and is unique if $J(x^{(k)})$ has full column ran. We note also that the Taylor series expansion

$$f(\overline{x}) \approx f(x) + J(\overline{x})(\overline{x} - x) \approx f(x) + J(x)(\overline{x} - x), \tag{4.22}$$

if x is close to $\overline{x}$. Thus

$$\text{Minimize } \frac{1}{2}\|f(z)\|_2^2 \approx \frac{1}{2}\|J(x)(z - x) + f(x)\|_2^2. \tag{4.23}$$

In other words, the linear least squares problem (4.21) may be regarded as a linear approximation to the nonlinear problem (4.12). The vector that solves (4.21) is called the Gauss-Newton direction. The method in which this vector is used as a search direction is known as the Gauss-Newton method.