CHAPTER 2

The Linear Least Squares Problem

In this chapter we study the linear least squares problem introduced in (4). Since this is such an important topic, we only briefly touch on a few aspects of this problem. We begin by introducing a few of the applications of the linear least squares from current research areas.

1. Applications

1.1. Polynomial Fitting. In many data fitting application one assumes a functional relationship between a set of "inputs" and a set of "outputs". For example, a patient is injected with a drug and the the research wishes to understand the clearance of the drug as a function of time. One way to do this is to draw blood samples over time and to measure the concentration of the drug in the drawn serum. The goal is to then provide a functional description of the concentration at any point in time.

Suppose the observed data is $y_i \in \mathbb{R}$ for each time point $t_i,\ i = 1, 2, \ldots, N$, respectively. The underlying assumption it that there is some function of time $f : \mathbb{R} \to \mathbb{R}$ such that $y_i = f(t_i),\ i = 1, 2, \ldots, N$. The goal is to provide and estimate of the function f. One way to do this is to try to approximate f by a polynomial of a fixed degree, say n:

$$p(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_n t^n.$$

We now wish to determine the values of the coefficients that "best" fit the data.

If were possible to exactly fit the data, then there would exist a value for the coefficient, say $\overline{x} = (\overline{x}_0, \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$ such that

$$y_i = \overline{x}_0 + \overline{x}_1 t_i + \overline{x}_2 t_i^2 + \cdots + \overline{x}_n t_i^n,\ i = 1, 2, \ldots, N.$$

But if N is larger than n, then it is unlikely that such an $\overline{x}$ exists; while if N is less than n, then there are probably many choices for $\overline{x}$ for which we can achieve a perfect fit. We discuss these two scenarios and their consequences in more depth at a future dat, but, for the moment, we assume that N is larger than n. That is, we wish to approximate f with a low degree polynomial.

When $n << N$, we cannot expect to fit the data perfectly and so there will be errors. In this case, we must come up with a notion of what it means to "best" fit the data. In the context of least squares, "best" means that we wish to minimized the sum of the squares of the errors in the fit:

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}}\ \tfrac{1}{2} \sum_{i=1}^{N} (x_0 + x_1 t_i + x_2 t_i^2 + \cdots + x_n t_i^n \ - \ y_i)^2 \ . \tag{5}$$

The leading one half in the objective is used to simplify certain computations that occur in the analysis to come. This minimization problem has the form

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}}\ \tfrac{1}{2} \left\| Vx - y \right\|_2^2 ,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad V = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & \ldots & t_2^n \\ \vdots & & & & \\ 1 & t_N & t_N^2 & \ldots & t_N^n \end{bmatrix},$$

since

$$Vx = \begin{pmatrix} x_0 + x_1 t_1 + x_2 t_1^2 + \cdots + x_n t_1^n \\ x_0 + x_1 t_2 + x_2 t_2^2 + \cdots + x_n t_2^n \\ \vdots \\ x_0 + x_1 t_N + x_2 t_N^2 + \cdots + x_n t_N^n \end{pmatrix}.$$

That is, the polynomial fitting problem (5) is an example of a linear least squares problem (4). The matrix V is called the *Vandermonde matrix* associated with this problem.

This is neat way to approximate functions. However, polynomials are a very poor way to approximate the clearance data discussed in our motivation to this approach. The concentration of a drug in serum typically rises quickly after injection to a maximum concentration and falls off gradually decaying exponentially. There is only one place where such a function is zero, and this occurs at time zero. On the other hand, a polynomial of degree n has n zeros (counting multiplicity). Therefore, it would seem that exponential functions would provide a better basis for estimating clearance. This motivates our next application.

1.2. Function Approximation by Bases Functions. In this application we expand on the basic ideas behind polynomial fitting to allow other kinds of approximations, such as approximation by sums of exponential functions. In general, suppose we are given data points $(z_i, y_i) \in \mathbb{R}^2,\ i = 1, 2, \ldots, N$ where it is assumed that the observation y_i is a function of an unknown function $f : \mathbb{R} \to \mathbb{R}$ evaluated at the point z_i for each $i = 1, 2, \ldots, N$. Based on other aspects of the underlying setting from which this data arises may lead us to believe that f comes from a certain space $\mathcal{F}$ of functions, such as the space of continuous or differentiable functions on an interval. This space of functions may itself be a vector space in the sense that the zero function is in the space ($0 \in \mathcal{F}$), two function in the space can be added pointwise to obtain another function in the space ($\mathcal{F}$ is closed with respect to addition), and any real multiple of a function is the space is also in the space ($\mathcal{F}$ is closed with respect to scalar multiplication). In this case, we may select from X a finite subset of functions, say $\phi_1, \phi_2, \ldots, \phi_k$, and try to approximate f as a linear combination of these functions:

$$f(x) \sim x_1\phi_1(z) + x_2\phi_2(z) + \cdots + x_n\phi_k(z).$$

This is exactly what we did in the polynomial fitting application discussed above. There $\phi_i(z) = z^i$ but we started the indexing at $i = 0$. Therefore, this idea is essentially the same as the polynomial fitting case. But the functions z^i have an additional properties. First, they are linearly independent in the sense that the only linear combination that yields the zero function is the one where all of the coefficients are zero. In addition, any continuous function on and interval can be approximated "arbitrarily well" by a polynomial assuming that we allow the polynomials to be of arbitrarily high degree (think Taylor approximations). In this sense, polynomials form a basis for the continuous function on and interval. By analogy, we would like our functions ϕ_i to be linearly independent and to come from basis of functions. There are many possible choices of bases, but a discussion of these would take us too far afield from this course.

Let now suppose that the functions $\phi_1, \phi_2, \ldots, \phi_k$ are linearly independent and arise from a set of basis function that reflect a deeper intuition about the behavior of the function f, e.g. it is well approximated as a sum of exponentials (or trig functions). Then the task to to find those coefficient $x_1, x_2, \ldots, x_n$ that best fits the data in the least squares sense:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2} \sum_{i=1}^{N} (x_1\phi_1(z_i) + x_2\phi_2(z_i) + \cdots + x_n\phi_k(z_i) - y_i)^2.$$

This can be recast as the linear least squares problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2} \left\| Ax - y \right\|_2^2,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad A = \begin{bmatrix} \phi_1(z_1) & \phi_2(z_1) & \ldots & \phi_n(z_1) \\ \phi_1(z_2) & \phi_2(z_2) & \ldots & \phi_n(z_2) \\ \vdots & & & \\ \phi_1(z_N) & \phi_2(z_N) & \ldots & \phi_n(z_N) \end{bmatrix}.$$

May possible further generalizations of this basic idea are possible. For example, the data may be multi-dimensional: $(z_i, y_i) \in \mathbb{R}^s \times \mathbb{R}^t$. In addition, constraints may be added, e.g., the function must be monotone (either increasing of decreasing), it must be unimodal (one "bump"), etc. But the essential features are that we estimate

using linear combinations and errors are measured using sums of squares. In many cases, the sum of squares error metric is not a good choice. But is can be motivated by assuming that the error are distributed using the Gaussian, or normal, distribution.

1.3. Linear Regression and Maximum Likelihood. Suppose we are considering a new drug therapy for reducing inflammation in a targeted population, and we have a relatively precise way of measuring inflammation for each member of this population. We are trying to determine the dosing to achieve a target level of inflamation. Of course, the dose needs to be adjusted for each individual due to the great amount of variability from one individual to the next. One way to model this is to assume that the resultant level of inflamation is on average a linear function of the dose and other individual specific covariates such as sex, age, weight, body surface area, gender, race, blood iron levels, desease state, etc. We then sample a collection of N individuals from the target population, registar their dose z_{i0} and the values of their individual specific covariates $z_{i1}, z_{i2}, \ldots, z_{in}$, $i = 1, 2, \ldots, N$. After dosing we observe that the resultant inflammation for the ith subject to be $y_i,\ i = 1, 2, \ldots, N$. By saying that the "resultant level of inflamation is on average a linear function of the dose and other individual specific covariates ", we mean that there exist coefficients $x_0, x_1, x_2, \ldots, x_n$ such that

$$y_i = x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} + v_i,$$

where v_i is an instance of a random variable representing the individuals deviation from the linear model. Assume that the random variables v_i are independently identically distributed $N(0, \sigma^2)$ (norm with zero mean and variance σ^2). The probability density function for the the normal distribution $N(0, \sigma^2)$ is

$$\frac{1}{\sigma\sqrt{2\pi}} \mathrm{EXP}[-v^2/(2\sigma^2)] \ .$$

Given values for the coefficients x_i, the likelihood function for the sample $y_i,\ i = 1, 2, \ldots, N$ is the joint probability density function evaluated at this observation. The independence assumption tells us that this joint pdf is given by

$$L(x; y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \mathrm{EXP}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} - y_i\right)^2\right] \ .$$

We now wish to choose those values of the coefficients $x_0, x_2, \ldots, x_n$ that make the observation $y_1, y_2, \ldots, y_n$ most probable. One way to try to do this is to maximize the likelihood function $L(x; y)$ over all possible values of x. This is called *maximum likelihood estimation*:

$$\underset{x\in\mathbb{R}^{n+1}}{\text{maximize}}\, L(x; y) \ . \tag{6}$$

Since the natural logarithm is nondecreasing on the range of the likelihood function, the problem (6) is equivalent to the problem

$$\underset{x\in\mathbb{R}^{n+1}}{\text{maximize}} \ln(L(x; y)) \ ,$$

which in turn is equivalent to the minimization problem

$$\underset{x\in\mathbb{R}^{n+1}}{\text{minimize}} -\ln(L(x; y)) \ . \tag{7}$$

Finally, observe that

$$-\ln(L(x; y)) = K + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} - y_i\right)^2 ,$$

where $K = n\ln(\sigma\sqrt{2\pi})$ is constant. Hence the problem (7) is equivalent to the linear least squares problem

$$\underset{x\in\mathbb{R}^{n+1}}{\text{minimize}}\ \tfrac{1}{2}\left\|Ax - y\right\|_2^2 ,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad A = \begin{bmatrix} z_{10} & z_{11} & z_{12} & \cdots & z_{1n} \\ z_{20} & z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & & & & \\ z_{N0} & z_{N1} & z_{N2} & \cdots & z_{Nn} \end{bmatrix} .$$

This is the first step in trying to select an optimal dose for each individual across a target population. What is missing from this analysis is some estimation of the variability in inflammation response due to changes in the covariates. Understanding this sensitivity to variations in the covariates is an essential part of any regression analysis. However, a discussion of this step lies beyond the scope of this brief introduction to linear regression.

1.4. System Identification in Signal Processing. We consider a standard problem in signal processing concerning the behavior of a stable, causal, linear, continuous-time, time-invariant system with input signal $u(t)$ and output signal $y(t)$. Assume that these signals can be described by the convolution integral

$$y(t) = (g * u)(t) := \int_0^{+\infty} g(\tau)u(t-\tau)d\tau \ . \tag{8}$$

In applications, the goal is to obtain an estimate of g by observing outputs y from a variety of known input signals u. For example, returning to our drug dosing example, the function u may represent the input of a drug into the body through a drug pump any y represent the concentration of the drug in the body at any time t. The relationship between the two is clearly causal (and can be shown to be stable). The transfer function g represents what the body is doing to the drug. In the way, the model (8) is a common model used in pharmaco-kinetics.

The problem of estimating g in (8) is an infinite dimensional problem. Below we describe a way to approximate g using the the FIR, or *finite impulse response* filter. In this model we discretize time by choosing a fixed number N of time points t_i to observe y from a known input u, and a finite time horizon $n < N$ over which to approximate the integral in (8). To simplify matters we index time on the integers, that is, we equate t_i with the integer i. After selecting the data points and the time horizon, we obtain the FIR model

$$y(t) = \sum_{k=1}^{n} g(k)u(t-k), \tag{9}$$

where we try to find the "best" values for $g(k),\ k = 0, 1, 2, \ldots, n$ to fit the system

$$y(t) = \sum_{k=0}^{n} g(k)u(t-k), \quad t = 1, 2, \ldots, N.$$

Notice that this requires knowledge of the values $u(t-k)$ for $t = t = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, n$. One often assumes a observational error in this model that is $N(0, \sigma^2)$ for a given value of σ^2. In this case, the FIR model (9) becomes

$$y(t) = \sum_{i=1}^{n} g(k)u(t-k) + v(t), \tag{10}$$

where $v(t),\ t = 1, \ldots, N$ are iid $N(0, \sigma^2)$. In this case, the corresponding maximum likelihood estimation problem becomes the linear least squares problem

$$\underset{g \in \mathbb{R}^{n+1}}{\text{minimize}} \ \tfrac{1}{2} \left\| Hg - y \right\|_2^2 ,$$

where

$$y = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix}, \quad g = \begin{pmatrix} g(0) \\ g(1) \\ g(2) \\ \vdots \\ g(n) \end{pmatrix} \quad \text{and} \quad H = \begin{bmatrix} u(1) & u(0) & u(-1) & u(-2) & \ldots & u(1-n) \\ u(2) & u(1) & u(0) & u(-1) & \ldots & u(2-n) \\ u(3) & u(2) & u(1) & u(0) & \ldots & u(3-n) \\ \vdots & & & & & \\ u(N) & u(N-1) & u(N-2) & u(N-3) & \ldots & u(N-n) \end{bmatrix}.$$

Notice that the matrix H has constant "diagonals". Such matrices are called *Toeplitz matrices*.

1.5. Kalman Smoothing. Kalman smoothing is a fundamental topic in signal processing and control literature, with numerous applications in navigation, tracking, healthcare, finance, and weather. Contributions to theory and algorithms related to Kalman smoothing, and to dynamic system inference in general, have come from statistics, engineering, numerical analysis, and optimization. Here, the term 'Kalman smoother' includes any method of inference on any dynamical system fitting the graphical representation of Figure 1.

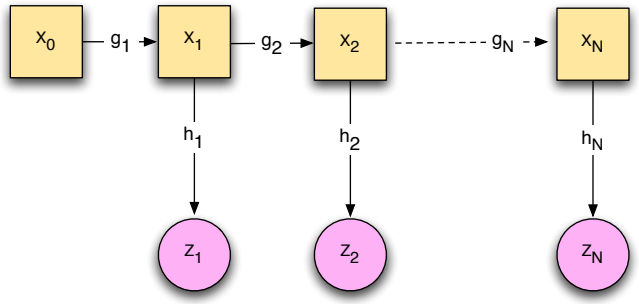

FIGURE 1. Dynamic systems amenable to Kalman smoothing methods.

The combined mathematical, statistical, and probablistic model corresponding to Figure 1 is specified as follows:

$$\begin{array}{rcll} \mathbf{x_1} & = & g_1(x_0) + \mathbf{w_1}, & \\ \mathbf{x_k} & = & g_k(\mathbf{x_{k-1}}) + \mathbf{w_k} & k = 2, \ldots, N, \\ \mathbf{z_k} & = & h_k(\mathbf{x_k}) + \mathbf{v_k} & k = 1, \ldots, N\ , \end{array} \tag{11}$$

where $\mathbf{w_k}$, $\mathbf{v_k}$ are mutually independent random variables with known positive definite covariance matrices Q_k and R_k, respectively. The vectors $\{\mathbf{x_k}\}$ are called the state sequence and the vectors $\{\mathbf{z_k}\}$ the observation sequence. Here, $\mathbf{w_k}$ often, but not always, arises from a probabilistic model (discretization of an underlying stochastic differential equation in the state x, from which the names 'smoother' is derived) and $\mathbf{v_k}$ comes from a statistical model for observations. We have $\mathbf{x_k}, \mathbf{w_k} \in \mathbb{R}^n$, and $\mathbf{z_k}, \mathbf{v_k} \in \mathbb{R}^{m(k)}$, so dimensions can vary between time points. The functions g_k and h_k as well as the matrices Q_k and R_k are known and given. In addition, the observation sequence $\{\mathbf{z_k}\}$ is also known. The goal is to estimate the unobserved state sequence $\{\mathbf{x_k}\}$. For example, in our drug dosing, the amount of the drug remaining in the body at time t is the unknown state sequence while the observation sequence is the observed concentration of the drug in each of our blood draws.

The classic case is obtained by making the following assumptions:

(1) x_0 is known, and g_k, h_k are known *linear* functions, which we denote by

$$g_k(x_{k-1}) = G_k x_{k-1} \quad h_k(x_k) \quad = H_k x_k \tag{12}$$

where $G_k \in \mathbb{R}^{n \times n}$ and $H_k \in \mathbb{R}^{m(k) \times n}$,

(2) $\mathbf{w_k}$, $\mathbf{v_k}$ are mutually independent *Gaussian* random variables.

In the classical setting, the connection to the linear least squares problem is obtained by formulating the maximum *a posteriori* (MAP) problem under linear and Gaussian assumptions. As in the linear regression and signal processing applications, this yields the following linear least squares problem:

$$\min_{\{x_k\}} f(\{x_k\}) := \sum_{k=1}^{N} \frac{1}{2}(z_k - H_k x_k)^T R_k^{-1}(z_k - H_k x_k) + \frac{1}{2}(x_k - G_k x_{k-1})^T Q_k^{-1}(x_k - G_k x_{k-1})\ . \tag{13}$$

To simplify this expression, we introduce data structures that capture the entire state sequence, measurement sequence, covariance matrices, and initial conditions. Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

$$\mathrm{vec}(\{u_k\}) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}\ ,\ \mathrm{diag}(\{T_k\}) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

We now make the following definitions:

$$
\begin{array}{lll}
R = \mathrm{diag}(\{R_k\}) & x = \mathrm{vec}(\{x_k\}) & \\
Q = \mathrm{diag}(\{Q_k\}) & w = \mathrm{vec}(\{g_0, 0, \ldots, 0\}) & G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}, \\
H = \mathrm{diag}(\{H_k\}) & z = \mathrm{vec}(\{z_1, z_2, \ldots, z_N\}) &
\end{array} \tag{14}
$$

where $g_0 := g_1(x_0) = G_1 x_0$. With definitions in (14), problem (13) can be written

$$
\min_x f(x) = \frac{1}{2}\|Hx - z\|^2_{R^{-1}} + \frac{1}{2}\|Gx - w\|^2_{Q^{-1}} \,, \tag{15}
$$

where $\|a\|^2_M = a^\top M a$.

Since the number of time steps N can be quite large, it is essential that the underlying tri-diagonal structure is exploited in any solution procedure. This is especially true when the state-space dimension n is also large which occurs when making PET scan movies of brain metabolics or reconstructing weather patterns on a global scale.

2. Optimality in the Linear Least Squares Problem

We now turn to a discussion of optimality in the least squares problem (4) which we restate here for ease of reference:

$$
\underset{x\in\mathbb{R}^n}{\text{minimize}} \; \tfrac{1}{2}\,\|Ax - b\|_2^2 \,, \tag{16}
$$

where

$$
A \in \mathbb{R}^{m\times n},\; b \in \mathbb{R}^m, \quad \text{and} \quad \|y\|_2^2 := y_1^2 + y_2^2 + \cdots + y_m^2 \,.
$$

In particular, we will address the question of when a solution to this problem exists and how they can be identified or characterized.

Suppose that $\overline{x}$ is a solution to (16), i.e.,

$$
\|A\overline{x} - b\|_2 \le \|Ax - b\|_2 \quad \forall\, x \in \mathbb{R}^n. \tag{17}
$$

Using this inequality, we derive necessary and sufficient conditions for the optimality of $\overline{x}$. A useful identity for our derivation is

$$
\|u + v\|_2^2 = (u+v)^T(u+v) = u^T u + 2u^T v + v^T v = \|u\|_2^2 + 2u^T v + \|v\|_2^2 \,. \tag{18}
$$

Let x be any other vector in $\mathbb{R}^n$. Then, using (18) with $u = A(\overline{x} - x)$ and $v = Ax - b$ we obtain

$$
\begin{aligned}
\|A\overline{x} - b\|_2^2 &= \|A(\overline{x} - x) + (Ax - b)\|_2^2 \\
&= \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) + \|Ax - b\|_2^2 \\
&\ge \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) + \|A\overline{x} - b\|_2^2 \qquad \text{(by (17))}.
\end{aligned} \tag{19}
$$

Therefore, by canceling $\|A\overline{x} - b\|_2^2$ from both sides, we know that, for all $x \in \mathbb{R}^n$,

$$
0 \ge \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) = 2(A(\overline{x} - x))^T(A\overline{x} - b) - \|A(\overline{x} - x)\|_2^2 \,.
$$

By setting $x = \overline{x} - tw$ for $t \in T$ and $w \in \mathbb{R}^n$, we find that

$$
\frac{t^2}{2}\,\|Aw\|_2^2 \ge tw^T A^T(A\overline{x} - b) \quad \forall\, t \in \mathbb{R} \quad \text{and} \quad w \in \mathbb{R}^n.
$$

Dividing by $t \neq 0$, we find that

$$
\frac{t}{2}\,\|Aw\|_2^2 \ge w^T A^T(A\overline{x} - b) \quad \forall\, t \in \mathbb{R}\setminus\{0\} \quad \text{and} \quad w \in \mathbb{R}^n,
$$

Sending t to zero gives

$$
0 \ge w^T A^T(A\overline{x} - b) \quad \forall\, w \in \mathbb{R}^n,
$$

which implies that $A^T(A\overline{x} - b) = 0$ (why?), or equivalently,

$$
A^T A\overline{x} = A^T b. \tag{20}
$$

The system of equations (20) is called the *normal equations* associated with the linear least squares problem (16). This derivation leads to the following theorem.

THEOREM 2.1. *[Linear Least Squares and the Normal Equations]*
The vector $\overline{x}$ solves the problem (16)*, i.e.,*

$$\|A\overline{x} - b\|_2 \leq \|Ax - b\|_2 \quad \forall\ x \in \mathbb{R}^n,$$

if and only if $A^TA\overline{x} = A^Tb$.

PROOF. We have just shown that if $\overline{x}$ is a solution to (16), then the normal equations are satisfied, so we need only establish the reverse implication. Assume that $A^TA\overline{x} = A^Tb$ or, equivalently, $A^T(A\overline{x} - b) = 0$. Then, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\|Ax - b\|_2^2 &= \|(Ax - A\overline{x}) + (A\overline{x} - b)\|_2^2 \\
&= \|A(x - \overline{x})\|_2^2 + 2(A(x - \overline{x}))^T(A\overline{x} - b) + \|A\overline{x} - b\|_2^2 \qquad \text{(by (18))} \\
&\geq 2(x - \overline{x})^TA^T(A\overline{x} - b) + \|A\overline{x} - b\|_2^2 \qquad \text{(since } \|A(x - \overline{x})\|_2^2 \geq 0) \\
&= \|A\overline{x} - b\|_2^2 \qquad \text{(since } A^T(A\overline{x} - b) = 0),
\end{aligned}$$

or equivalently, $\overline{x}$ solves (16). □

This theorem provides a nice characterization of solutions to (16), but it does not tell us if a solution exits. For this we use the following elementary result from linear algebra.

LEMMA 2.1. *For every matrix $A \in \mathbb{R}^{m\times n}$ we have*

$$\textit{Null}(A^TA) = \textit{Null}(A) \quad \textit{and} \quad \textit{Ran}(A^TA) = \textit{Ran}(A^T)\ .$$

PROOF. Note that if $x \in \text{Null}(A)$, then $Ax = 0$ and so $A^TAx = 0$, that is, $x \in \text{Null}(A^TA)$. Therefore, $\text{Null}(A) \subset \text{Null}(A^TA)$. Conversely, if $x \in \text{Null}(A^TA)$, then

$$A^TAx = 0 \implies x^TA^TAx = 0 \implies (Ax)^T(Ax) = 0 \implies \|Ax\|_2^2 = 0 \implies Ax = 0,$$

or equivalently, $x \in \text{Null}(A)$. Therefore, $\text{Null}(A^TA) \subset \text{Null}(A)$, and so $\text{Null}(A^TA) = \text{Null}(A)$.

Since $\text{Null}(A^TA) = \text{Null}(A)$, the Fundamental Theorem of the Alternative tells us that

$$\text{Ran}(A^TA) = \text{Ran}((A^TA)^T) = \text{Null}(A^TA)^\perp = \text{Null}(A)^\perp = \text{Ran}(A^T),$$

which proves the lemma. □

This lemma immediately gives us the following existence result.

THEOREM 2.2. *[Existence and Uniqueness for the Linear Least Squares Problem]*
Consider the linear least squares problem (16)*.*

(1) A solution to the normal equations (20) *always exists.*
(2) A solution to the linear least squares problem (16) *always exists.*
(3) The linear least squares problem (16) *has a unique solution if and only if Null$(A) = \{0\}$ in which case $(A^TA)^{-1}$ exists and the unique solution is given by $\overline{x} = (A^TA)^{-1}A^Tb$.*
(4) If Ran$(A) = \mathbb{R}^m$, then $(AA^T)^{-1}$ exists and $\overline{x} = A^T(AA^T)^{-1}b$ solves (16)*, indeed, $A\overline{x} = b$.*

PROOF. (1) Lemma 2.1 tells us that $\text{Ran}(A^TA) = \text{Ran}(A^T)$; hence, a solution to $A^TAx = A^Tb$ must exist.
(2) This follows from Part (1) and Theorem 2.1.
(3) By Theorem 2.1, $\overline{x}$ solves the linear least squares problem if and only if $\overline{x}$ solves the normal equations. Hence, the linear least squares problem has a uniques solution if and only if the normal equations have a unique solution. Since $A^TA \in \mathbb{R}^{n\times n}$ is a square matrix, this is equivalent to saying that A^TA is invertible, or equivalently, $\text{Null}(A^TA) = \{0\}$. However, by Lemma 2.1, $\text{Null}(A) = \text{Null}(A^TA)$. Therefore, the linear least squares problem has a uniques solution if and only if $\text{Null}(A) = \{0\}$ in which case A^TA is invertible and the unique solution is given by $\overline{x} = (A^TA)^{-1}A^Tb$.
(4) By the hypotheses, Lemma 2.1, and the Fundamental Theorem of the Alternative, $\{0\} = (\mathbb{R}^m)^\perp = (\text{Ran}(A))^\perp = \text{Null}(A^T) = \text{Null}(AA^T)$; hence, $AA^T \in \mathbb{R}^{m\times m}$ is invertible. Consequently, $\overline{x} = A^T(AA^T)^{-1}b$ is well-defined and satisfies $A\overline{x} = b$ □

Theorem 2.2 establishes the existence of solutions to the linear least squares problem as well as necessary conditions for optimality and uniqueness. When the solution is unique, it also provides a formula for this solution. However, these results do not provide a numerical mechanism for computing a solution even in the case when the solution is unique. Here the dimension of the problem, or the problem size, plays a key role. In addition, the

level of accuracy in the solution as well as the greatest accuracy possible are also issues of concern. Linear least squares problems range in size from just a few variables and equations to millions. Some are so large that all of the computing resources at our disposal today are insufficient to solve them, and in many cases the matrix A is not even available in the sense that it is not stored on a computer. However, in this latter case, it is often possible to either compute or approximate the vector Ax for a given vector x. Therefore, great care and inventiveness is required in the numerical solution of these problems. Research into how to solve this class of problems remains an important area of research to this day.

In our study of numerical solution techniques we present a few classical methods. But before doing so, we study other aspects of the problem in order to gain further insight into its geometric structure.

3. Orthogonal Projection onto a Subspace

In this section we view the linear least squares problem from the perspective of a least distance problem to a subspace, or equivalently, as the problem of projecting onto a subspace. Suppose $S \subset \mathbb{R}^m$ is a given subspace and $b \notin S$. The least distance problem for S and b is to find that element of S that is as close to b as possible. That is we wish to solve the problem

$$\min_{z \in S} \tfrac{1}{2} \|z - y\|_2^2 \ , \tag{21}$$

or equivalently, we wish to find the point $\overline{z} \in S$ such that

$$\|\overline{z} - b\|_2 \leq \|z - b\|_2 \qquad \forall\ z \in S.$$

If we take the subspace to be the range of A, $S = \mathrm{Ran}(A)$, then the problem (21) is closely related to the problem (16) since

(22) $\overline{z} \in \mathbb{R}^m$ solves (21) if and only if there is an $\overline{x} \in \mathbb{R}^n$ with $\overline{z} = A\overline{x}$ such that $\overline{x}$ solves (16). (why?)

Below we discuss this connection and its relationship to the notion of an orthogonal projection onto a subspace.

A matrix $P \in \mathbb{R}^{m \times m}$ is said to be a *projection* if and only if $P^2 = P$. In this case we say that P is a projection onto the subspace $S = \mathrm{Ran}(P)$, the range of P. Note that if $x \in \mathrm{Ran}(P)$, then there is a $w \in \mathbb{R}^m$ such that $x = Pw$, therefore, $Px = P(Pw) = P^2 w = Pw = x$. That is, P leaves all elements of $\mathrm{Ran}(P)$ fixed. Also, note that, if P is a projection, then

$$(I - P)^2 = I - P - P + P^2 = I - P,$$

and so $(I - P)$ is also a projection. Since for all $w \in \mathbb{R}^m$,

$$w = Pw + (I - P)w,$$

we have

$$\mathbb{R}^m = \mathrm{Ran}(P) + \mathrm{Ran}(I - P).$$

In this case we say that the subspaces $\mathrm{Ran}(P)$ and $\mathrm{Ran}(I - P)$ are *complementary subspaces* since their sum is the whole space and their intersection is the origin, i.e., $\mathrm{Ran}(P) \cap \mathrm{Ran}(I - P) = \{0\}$ (why?).

Conversely, given any two subspaces S_1 and S_2 that are complementary, that is, $S_1 \cap S_2 = \{0\}$ and $S_1 + S_2 = \mathbb{R}^m$, there is a projection P such that $S_1 = \mathrm{Ran}(P)$ and $S_2 = \mathrm{Ran}(I - P)$. We do not show how to construct these projections here, but simply note that they can be constructed with the aid of bases for S_1 and S_2.

The relationship between projections and complementary subspaces allows us to define a notion of *orthogonal projection*. Recall that for every subspace $S \subset \mathbb{R}^m$, the subspace orthogonal to S is given by

$$S^\perp := \left\{ x \,\middle|\, x^T y = 0 \ \forall\ y \in S \right\}.$$

We say that S and $S^\perp$ are *orthogal subspaces*. Clearly, S and $S^\perp$ are complementary:

$$S \cap S^\perp = \{0\} \qquad \text{and} \qquad S + S^\perp = \mathbb{R}^m. \qquad \text{(why?)}$$

Therefore, there is a projection P such that $\mathrm{Ran}(P) = S$ and $\mathrm{Ran}(I - P) = S^\perp$, or equivalently,

$$((I - P)y)^T (Pw) = 0 \qquad \forall\ y, w \in \mathbb{R}^m. \tag{23}$$

The orthogonal projection plays a very special role among all possible projections onto a subspace. For this reason, we denote the orthogonal projection onto the subspace S by P_S.

We now use the condition (23) to derive a simple test of whether a linear transformation is an orthogonal projection. For brevity, we write $P := P_S$ and set $M = (I - P)^T P$. Then, by (23),

$$0 = e_i^T M e_j = M_{ij} \qquad \forall\ i, j = 1, \ldots, n,$$

i.e., M is the zero matrix. But then, since $0 = (I-P)^T P = P - P^T P$, we have

$$P = P^T P = (P^T P)^T = P^T.$$

Conversely, if $P = P^T$ and $P^2 = P$, then $(I-P)^T P = 0$. Therefore, a matrix P is an orthogonal projection if and only if $P^2 = P$ and $P = P^T$.

An orthogonal projection for a given subspace S can be constructed from any orthonormal basis for that subspace. Indeed, if the columns of the matrix Q form an orthonormal basis for S, then the matrix $P = QQ^T$ satisfies

$$P^2 = QQ^TQQ^T \overset{\text{why?}}{=} QI_kQ^T = QQ^T = P \qquad \text{and} \qquad P^T = (QQ^T)^T = QQ^T = P,$$

where $k = \dim(S)$, and so P is the orthogonal projection onto S since, by construction, $\mathrm{Ran}(QQ^T) = \mathrm{Ran}(Q) = S$. We catalogue these observations in the following lemma.

LEMMA 3.1. *[Orthogonal Projections]*

(1) The matrix $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection if and only if $P = P^2$ and $P = P^T$.

(2) If the columns of the matrix $Q \in \mathbb{R}^{n\times k}$ form an orthonormal basis for the subspace $S \subset \mathbb{R}^n$, then $P := QQ^T$ is the orthogonal projection onto S.

Let us now apply these projection ideas to the problem (21). Let $P := P_S$ be the orthogonal projection onto the subspace S, and let $\overline{z} = Pb$. Then, for every $z \in S$,

$$\begin{aligned}
\|z-b\|_2^2 &= \|Pz - Pb - (I-P)b\|_2^2 && \text{(since } z \in S\text{)}\\
&= \|P(z-b) + (I-P)b\|_2^2 \\
&= \|P(z-b)\|_2^2 + 2(z-b)^T P^T (I-P)b + \|(I-P)b\|_2^2 \\
&= \|P(z-b)\|_2^2 + \|(I-P)b\|_2^2 && \text{(since } P = P^T \text{ and } P = P^2\text{)}\\
&\geq \|(P-I)b\|_2^2 && \text{(since } \|P(z-b)\|_2^2 >\geq 0\text{)}\\
&= \|\overline{z} - b\|_2^2 .
\end{aligned}$$

Consequently, $\|\overline{z}-b\|_2 \leq \|z-b\|_2$ for all $z \in S$, that is, $\overline{z} = Pb$ solves (21). This observation yield the following theorem as an elementary consequence of the *parallelogram law*:

$$2\|u\|_2^2 + 2\|v\|_2^2 = \|u+v\|_2^2 + \|u-v\|_2^2 \qquad \forall\, u, v \in \mathbb{R}^n.$$

THEOREM 3.1. *[Subspace Projection Theorem]*
Let $S \subset \mathbb{R}^m$ be a subspace and let $b \in \mathbb{R}^m \setminus S$. Then the unique solution to the least distance problem

$$\underset{z\in S}{\text{minimize}} \|z-b\|_2$$

is $\overline{z} := P_S b$, where P_S is the orthogonal projector onto S.

PROOF. Everything but the uniqueness of the solution has been established in the discussion preceeding the theorem. To show uniqueness, apply the parallelogram law to obtain

$$\|(1-t)u + tv\|_2^2 = (1-t)\|u\|_2^2 + t\|v\|_2^2 - t(1-t)\|u-v\|_2^2 \quad \forall\ 0 \leq t \leq 1 \quad \text{and} \quad u, v \in \mathbb{R}^m.$$

Let $z^1, z^2 \in \mathbb{R}^m$ be two points that solve the minimum distance problem. Then, $\|z^1 - b\|_2 = \|z^2 - b\|_2 =: \eta > 0$, and so by the identity given above,

$$\begin{aligned}
\left\|\tfrac{1}{2}(z^1+z^2) - b\right\|_2^2 &= \left\|\tfrac{1}{2}(z^1-b) + \tfrac{1}{2}(z^2-b)\right\|_2^2 \\
&= \tfrac{1}{2}\|z^1-b\|_2^2 + \tfrac{1}{2}\|z^2-b\|_2^2 - \frac{1}{4}\|z^1-z^2\|_2^2 \\
&= \eta^2 - \frac{1}{4}\|z^1-z^2\|_2^2 .
\end{aligned}$$

Since $\eta = \inf\{\|z-b\|_2 \,|\, z \in S\}$, we must have $z^1 = z^2$. □

Let us now reconsider the linear least-squares problem (16) as it relates to our new found knowledge about orthogonal projections and their relationship to least distance problems for subspaces. Consider the case where

$m >> n$ and $\text{Null}(A) = \{0\}$. In this case, Theorem 2.2 tells us that $\overline{x} = (A^TA)^{-1}A^Tb$ solves (16), and $\overline{z} = P_Sb$ solves (23) where P_S is the orthogonal projector onto $S = \text{Ran}(A)$. Hence, by (22),

$$P_Sb = \overline{z} = A\overline{x} = A(A^TA)^{-1}A^Tb.$$

Since this is true for all possible choices of the vector b, we have

$$P_S = P_{\text{Ran}(A)} = A(A^TA)^{-1}A^T \ ! \tag{24}$$

That is, the matrix $A(A^TA)^{-1}A^T$ is the orthogonal projector onto the range of A. One can also check this directly by showing that the matrix $M = A(A^TA)^{-1}A^T$ satisfies $M^2 = M$, $M^T = M$, and $\text{Ran}(M) = \text{Ran}(A)$.

PROPOSITION 3.1. *Let $A \in \mathbb{R}^{m\times n}$ with $m \leq n$ and Null$(A) = \{0\}$. Then*

$$P_{Ran(A)} = A(A^TA)^{-1}A^T.$$

4. Minimal Norm Solutions to $Ax = b$

Again let $A \in \mathbb{R}^{m\times n}$, but now we suppose that $m << n$. In this case A is short and fat so the matrix A most likely has rank m, or equivalently,

$$\text{Ran}(A) = \mathbb{R}^m \ . \tag{25}$$

But regardless of the range of A and the choice of the vector $b \in \mathbb{R}^m$, the set of solutions to $Ax = b$ will be infinite if a solution exists since the nullity of A is $n - m$. Indeed, if x^0 is any particular solution to $Ax = b$, then the set of solutions is given by $x^0 + \text{Null}(A) := \left\{x^0 + z \,\middle|\, z \in \text{Null}(A)\right\}$. In this setting, one might prefer the solution to the system having least norm. This solution is found by solving the problem

$$\min_{z\in\text{Null}(A)} \tfrac{1}{2}\left\|z + x^0\right\|_2^2 \ . \tag{26}$$

This problem is of the form (21). Consequently, the solution is given by $\overline{z} = -P_Sx^0$ where P_S is the orthogonal projection onto $S := \text{Null}(A)$. In particular, this implies that the least norm solution to the system $Ax = b$ is uniquely given by the orthogonal projection of x^0 onto the range of A^T since $S^\perp = \text{Null}(A)^\perp = \text{Ran}\left(A^T\right)$ and

$$x^0 + \overline{z} = x^0 - P_{\text{Null}(A)}x^0 = (I - P_{\text{Null}(A)})x^0 = P_{\text{Null}(A)^\perp}x^0 = P_{\text{Ran}\left(A^T\right)}x^0 \ . \tag{27}$$

Recall that the formula (24) shows that if $M \in \mathbb{R}^{k\times s}$ is such that $\text{Null}(M) = \{0\}$, then the orthogonal projector onto $\text{Ran}(M)$ is given by

$$P_{\text{Ran}(M)} = M(M^TM)^{-1}M^T \ . \tag{28}$$

In our case, $M = A^T$ and $M^TM = AA^T$. Thus, if we assume that (25) holds, then

$$\text{Null}(M) = \text{Null}(A^T) = \text{Ran}(A)^\perp = (\mathbb{R}^m)^\perp = \{0\}$$

and consequently, by (28), the orthogonal projector onto $\text{Ran}(A^T)$ is given by

$$P_{\text{Ran}(A^T)} = A^T(AA^T)^{-1}A \ .$$

Therefore, when (25) holds, the least norm solution to $Ax = b$ is uniquely given by

$$\overline{x} = A^T(AA^T)^{-1}Ax^0,$$

where x^0 is any particular solution to $Ax = b$. These observations establish the following theorem.

THEOREM 4.1. *[Least Norm Solution to Linear Systems] Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and let x^0 be any solution to the system $Ax = b$. Then the least norm solution to the system $Ax = b$ is given by the orthogonal projection of x^0 onto the range of A^T. If it is further assumed that Ran$(A) = \mathbb{R}^m$, then the following hold.*

1. *The matrix AA^T is invertible.*
2. *The orthogonal projection onto* $\text{Ran}\left(A^T\right)$ *and Null(A) are given by*

$$P_{\text{Ran}\left(A^T\right)} = A^T(AA^T)^{-1}A \quad \text{and} \quad P_{Null(A)} = I - A^T(AA^T)^{-1}A \ .$$

3. *For every $b \in \mathbb{R}^m$, the system $Ax = b$ is consistent, and the least norm solution to this system is uniquely given by*

$$\overline{x} = A^T(AA^T)^{-1}Ax^0 \ ,$$

where x^0 is any particular solution to the system $Ax = b$.

5. Gram-Schmidt Orthogonalization, the QR Factorization, and Solving the Normal Equations

5.1. Gram-Schmidt Orthogonalization. In the previous sections we learned the significance of orthogonal projections for the linear least squares problem. In addition, we found that if the columns of the matrix U form an orthonormal basis for the subspace S, then the matrix UU^T is the orthogonal projection onto S. Hence, one way to obtain an orthogonal projection onto a subspace S is to compute an orthogonal basis for S. This is precisely what the *Gram-Schmidt orthogonalization* process does.

Let us recall the Gram-Schmidt orthogonalization process for a sequence of linearly independent vectors $a_1, \ldots, a_n \in \mathbb{R}^m$ (note that this implies that $n \leq m$ (why?)). In this process we define vectors $q_1, \ldots, q_n$ inductively, as follows: set

$$p_1 = a_1, \qquad q_1 = p_1/\|p_1\|,$$

$$p_j = a_j - \sum_{i=1}^{j-1} \langle a_j,\, q_j\rangle\, q_i \quad \text{and} \quad q_j = p_j/\|p_j\| \quad \text{for} \quad 2 \leq j \leq n.$$

For $1 \leq j \leq n$, $q_j \in \text{Span}\{a_1, \ldots, a_j\}$, so $p_j \neq 0$ by the linear independence of $a_1, \ldots, a_j$. An elementary induction argument shows that the q_j's form an orthonormal basis for span $(a_1, \ldots, a_n)$.

If we now define

$$r_{jj} = \|p_j\| \neq 0 \quad \text{and} \quad r_{ij} = \langle q_i,\, a_j\rangle \quad \text{for} \quad 1 \leq i < j \leq n,$$

then

$$\begin{aligned} a_1 &= r_{11}\, q_1, \\ a_2 &= r_{12}\, q_1 + r_{22}\, q_2, \\ a_3 &= r_{13}\, q_1 + r_{23}\, q_2 + r_{33}\, q_3, \\ &\vdots \\ a_n &= \sum_{i=1}^{n} r_{in}\, q_i. \end{aligned}$$

Set

$$A := [a_1\ a_2\ \ldots\ a_n] \in \mathbb{R}^{m\times n}, \quad R := [r_{ij}] \in \mathbb{R}^{n\times n}, \quad \text{and} \quad Q := [q_1\ q_2\ \ldots\ q_n] \in \mathbb{R}^{m\times n}\ ,$$

where $r_{ij} = 0,\ i > j$. Then

$$A = QR\ ,$$

where Q is unitary and R is an upper triangular $n \times n$ matrix. In addition, R is invertible since the diagonal entries r_{jj} are non-zero. This is called the *QR factorization* of the matrix A.

REMARK 5.1. *If the a_j's for $j = 1, \ldots, n$ are linearly dependent, then, for at least one value of j,*

$$a_j \in \text{Span}\{a_1, \ldots, a_{j-1}\}, \quad \textit{and so} \quad p_j = 0.$$

The process can be modified by setting $r_{jj} = 0$, not defining a new q_j for this iteration, but continuing to define $r_{ij} = \langle a_j,\, q_i\rangle \quad$ for $\quad 1 \leq i < j$, and proceeding. We still obtain with orthogonormal vectors $\{q_1, q_2, \ldots, q_k\}$, but now $k < n$. In general, after n iterations, there will be $1 \leq k \leq n$ vectors $\{q_1, \ldots, q_k\}$ that form an orthonormal basis for $\text{Span}\{a_1, \ldots, a_n\}$*, where $n - k$ is the number of diagonal entries r_{jj} that take the value zero. Again we obtain $A = QR$, but now Q may not be square and the matrix R may have zero diagonal entries in which case it is not invertible.*

REMARK 5.2. *The classical Gram-Schmidt algorithm as described above can have poor computational behavior due to the accumulation of round-off error. In particular, the computed vectors q_j's are not orthogonal: $\langle q_j,\, q_k\rangle$ is small for $j \neq k$ with j near k, but not so small for $j \ll k$ or $j \gg k$.*

An alternate version, "Modified Gram-Schmidt," is equivalent in exact arithmetic, but behaves better numerically. In the following "pseudo-codes," p denotes a temporary storage vector used to accumulate the sums defining the p_j's.

Classic Gram-Schmidt

For $j = 1, \cdots, n$ *do*
- $p := a_j$
- *For* $i = 1, \ldots, j-1$ *do*
  - $r_{ij} = \langle a_j,\, q_i \rangle$
  - $p := p - r_{ij} q_i$
- $r_{jj} := \|p\|$
- $q_j := p / r_{jj}$

Modified Gram-Schmidt

For $j = 1, \ldots, n$ *do*
- $p := a_j$
- *For* $i = 1, \ldots, j-1$ *do*
  - $r_{ij} = \langle p,\, q_i \rangle$
  - $p := p - r_{ij} q_i$
- $r_{jj} = \|p\|$
- $q_j := p / r_{jj}$

The only difference is in the computation of r_{ij}*: in Modified Gram-Schmidt, we orthogonalize the accumulated partial sum for* p_j *against each* q_i *successively.*

THEOREM 5.1. *[The Full QR Factorization] Suppose* $A \in \mathbb{R}^{m\times n}$ *with* $m \geq n$*. Then there exists a permutation matrix* $P \in \mathbb{R}^{n\times n}$*, a unitary matrix* $Q \in \mathbb{R}^{m\times m}$*, and an upper triangular matrix* $R \in \mathbb{R}^{m\times n}$ *such that* $AP = QR$*. Let* $Q_1 \in \mathbb{R}^{m\times n}$ *denote the first* n *columns of* Q*,* Q_2 *the remaining* $(m-n)$ *columns of* Q*, and* $R_1 \in \mathbb{R}^{n\times n}$ *the first* n *rows of* R*, then*

$$AP = QR = [Q_1\ Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1. \tag{29}$$

Moreover, we have the following:

(a) We may choose R *to have nonnegative diagonal entries.*

(b) If A *is of full rank, then we can choose* R *with positive diagonal entries, in which case we obtain the condensed factorization* $A = Q_1 R_1$*, where* $R_1 \in \mathbb{R}^{n\times n}$ *invertible and the columns of* Q_1 *forming an orthonormal basis for the range of* A*.*

(c) If $\operatorname{rank}(A) = k < n$*, then*

$$R_1 = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

where R_{11} *is a* $k \times k$ *invertible upper triangular matrix and* $R_{12} \in \mathbb{R}^{k\times(n-k)}$*. In particular, this implies that* $AP = Q_{11}[R_{11}\ R_{12}]$*, where* Q_{11} *are the first* k *columns of* Q*. In this case, the columns of* Q_{11} *form an orthonormal basis for the range of* A*.*

REMARK 5.3. *We call the factorization* $AP = Q_{11}[R_{11}\ R_{12}]$ *in Part (c) above the condensed QR Factorization. Note that if* P *is a permutation matrix, then so is* P^T *with* $P^{-1} = P^T$ *(i.e. permutation matrices are unitary). The role of the permutation matrix is to make the first* $k = \operatorname{rank}(A)$ *columns of* AP *linearly independent.*

To distinguish the condensed QR Factorization from the factorization in (29) *with* Q *an* $m \times m$ *unitary matrix, we will refer the factorization where* Q *is unitary as the full QR factorization.*

PROOF. If necessary, permute the columns of A so that the first $k = \operatorname{rank}(A)$ columns of A are linearly independent and let P denote the permutation matrix that accomplishes this task so the the first k columns of AP are linearly independent. Apply the Gram-Schmidt orthogonalization process to obtain the matrix

$$Q_1 = [q_1, \ldots, q_k] \in \mathbb{R}^{m\times k} \quad \text{and the upper triangular matrix} \quad \widetilde{R}_{11} = [r_{ij}] \in \mathbb{R}^{k\times k}$$

so that $Q_1 R_1$ gives the first k columns of A. The write the remaining columns of A as linear combinations of the columns of Q_1 to obtain the coefficient matrix $R_{12} \in \mathbb{R}^{k\times(n-k)}$ yielding $AP = Q_1[R_{11}\ R_{12}]$. Finally, extend $\{q_1, \ldots, q_k\}$ to an orthonormal basis $\{q_1, \ldots, q_m\}$ of $\mathbb{R}^m$, and set

$$Q = [q_1, \ldots, q_m] \quad \text{and} \quad R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \in \mathbb{R}^{m\times n}, \text{ so } AP = QR.$$

As $r_{jj} > 0$ in the Gram-Schmidt process, we have (b). □

REMARK 5.4. *There are more efficient and better computationally behaved ways of calculating the* Q *and* R *factors. The idea is to create zeros below the diagonal (successively in columns* $1, 2, \ldots$*) as in Gaussian Elimination, except instead of doing this by successive left multiplication by Gaussian elimination matrices, we left multiply by*

unitary matrices. Below, we show how this can be done with Householder transformations. But another popular approach is to use Givens rotations.

In practice, every $A \in \mathbb{R}^{m\times n}$ has a QR-factorization, even when $m < n$. This follows immediately from Part (c) Theorem 5.1.

COROLLARY 5.1.1. *[The General Condensed QR Factorization] Let $A \in \mathbb{R}^{m\times n}$ have rank $k \leq \min\{m, n\}$. Then there exist*

$$\begin{aligned} Q &\in \mathbb{R}^{m\times k} \quad \textit{with orthonormal columns,} \\ R &\in \mathbb{R}^{k\times n} \quad \textit{full rank upper triangular, and} \\ P &\in \mathbb{R}^{n\times n} \quad \textit{a permutation matrix} \end{aligned}$$

such that

$$AP = QR.$$

In particular, the columns of the matrix Q form a basis for the range of A. Moreover, the matrix R can be written in the form

$$R = [R_1 \ R_2],$$

where $R_1 \in \mathbb{R}^{k\times k}$ is nonsingular.

REMARK 5.5. *The permutation P in the corollary above can be taken to be any permutation that re-orders the columns of A so that the first k columns of A are linearly independent, where k is the rank of A (similarly for $\widetilde{P}$ in permuting the columns of A^T).*

COROLLARY 5.1.2. *[Orthogonal Projections onto the Four Fundamental Subspaces] Let $A \in \mathbb{R}^{m\times n}$ have rank $k \leq \min\{m, n\}$. Let A and A^T have generalized QR factorizations*

$$AP = Q[R_1 \ R_2] \quad \text{and} \quad A^T\widetilde{P} = \widetilde{Q}[\widetilde{R}_1 \ \widetilde{R}_2].$$

Since row rank equals column rank, $P \in \mathbb{R}^{n\times n}$ is a permutation matrix, $\widetilde{P} \in \mathbb{R}^{m\times m}$ is a permutation matrix, $Q \in \mathbb{R}^{m\times k}$ and $\widetilde{Q} \in \mathbb{R}^{n\times k}$ have orthonormal columns, $R_1, \widetilde{R}_1 \in \mathbb{R}^{k\times k}$ are both upper triangular nonsingular matrices, $R_2 \in \mathbb{R}^{k\times(n-k)}$, and $\widetilde{R}_2 \in \mathbb{R}^{k\times(m-k)}$. Moreover,

$$\begin{aligned} QQ^T &\quad \textit{is the orthogonal projection onto } \textit{Ran}(A), \\ I - QQ^T &\quad \textit{is the orthogonal projection onto } \textit{Null}(A^T), \\ \widetilde{Q}\widetilde{Q}^T &\quad \textit{is the orthogonal projection onto } \textit{Ran}(A^T), \textit{ and} \\ I - \widetilde{Q}\widetilde{Q}^T &\quad \textit{is the orthogonal projection onto } \textit{Null}(A)^\perp. \end{aligned}$$

PROOF. The result follows immediately from Corollary 5.1.1 and the Fundamental Theorem of the Alternative. □

EXERCISE 5.1. *Verify the representations of the orthogonal projections onto $\textit{Ran}(A)$ and $\textit{Null}(A)$ given in Corollary 5.1.2 correspond to those given in Proposition 3.1 and Theorem 4.1.*

5.2. Solving the Normal Equations with the QR Factorization. Let's now reconsider the linear least squares problem (16) and how the QR factorization can be used in its solution. Specifically, we examine how is can be used to solve the normal equations $A^TAx = A^Tb$. Let A and b be as in (16), and let

$$AP = Q[R_1 \ R_2]$$

be the general condensedQR factorization of A, where $P \in \mathbb{R}^{n\times n}$ is a permutation matrix, $Q \in \mathbb{R}^{m\times k}$ has orthonormal columns, $R_1 \in \mathbb{R}^{k\times k}$ is nonsingular and upper triangular, and $R_2 \in \mathbb{R}^{k\times(n-k)}$ with $k = \text{rank}\,(A) \leq \min\{n, m\}$. Replacing A by $A = Q[R_1 \ R_2]P^T$ in the normal equations gives the following equivalent system:

$$P^T\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^TQ \begin{bmatrix} R_1 & R_2 \end{bmatrix} Px = P^T\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} \begin{bmatrix} R_1 & R_2 \end{bmatrix} Px = A^Tb = P^T\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^Tb,$$

since $Q^TQ = I_k$ the $k \times k$ identity matrix. By multiplying on the left by P, replacing b by $\hat{b} := Q^Tb \in \mathbb{R}^k$ and x by

$$z := \begin{bmatrix} R_1 & R_2 \end{bmatrix} Px, \tag{30}$$

we obtain

$$\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} z = \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} \hat{b}.$$

Let us see if we can reconstruct a solution to the normal equations by choosing the most obvious solution to the this system, namely, $\overline{z} := \hat{b}$. If this is to yield a solution to the normal equations, then, by (30), we need to solve the system

$$\begin{bmatrix} R_1 & R_2 \end{bmatrix} Px = \hat{b} \ .$$

Set

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} := Px \ ,$$

where $w_1 \in \mathbb{R}^k$ and $w_2 \in \mathbb{R}^{(n-k)}$, and consider the system

$$R_1 w_1 = \hat{b} \ \in \mathbb{R}^k.$$

Since $R_1 \in \mathbb{R}^{k\times k}$ is invertible, this system has a unique solution $\overline{w}_1 := R_1^{-1}\hat{b}$. Indeed, this system is very easy to solve using *back substitution* since R_1 is upper triangular. Next set $\overline{w}_2 = 0 \in \mathbb{R}^{(n-k)}$ and

$$\overline{x} := P^T \overline{w} = P^T \begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix}.$$

Then

$$\begin{aligned}
A^T A\overline{x} &= A^T A P^T \begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix} \\
&= A^T Q \begin{bmatrix} R_1 & R_2 \end{bmatrix} P P^T \begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix} \\
&= A^T Q R_1 R_1^{-1}\hat{b} && \text{(since } PP^T = I\text{)} \\
&= A^T Q \hat{b} \\
&= A^T Q Q^T b \\
&= P^T \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T Q Q^T b && \text{(since } A^T = P^T \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T\text{)} \\
&= P^T \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T b && \text{(since } Q^T Q = I\text{)} \\
&= A^T b,
\end{aligned}$$

that is, $\overline{x}$ solves the normal equations!

Let us now consider the computational cost of obtaining the solution to the linear least squares problem in this way. The key steps is this computation are as follows:

$AP = Q[R_1 \ R_2]$	the general condensed QR factorization	$o(m^2 n)$
$\hat{b} = Q^T b$	a matrix-vector product	$o(km)$
$\overline{w}_1 = R_1^{-1}\hat{b}$	a back solve	$o(k^2)$
$\overline{x} = P^T \begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix}$	a matrix-vector product	$o(kn)$.

Therefore, the majority of the numerical effort is in the computation of the QR factorization.

5.3. Computing the Full QR Factorization using Householder Reflections. In subsection 5.1 we showed how to compute the QR factorization using the Gram-Schmidt orthogonalization procedure. We also indicated that due to numerical round-off error this procedure has difficulty in preserving the orthogonality of the columns of the matrix Q. To address this problem we presented the mathematically equivalent *modified* Gram-Schmidt process which has improved performance. We now present a very different method for obtaining the full QR factorization. The approach we describe is very much like Gauss-Jordan Elimination to obtain reduced echelon form. However, now we successively multiply A on the left by unitary matrices, rather than Gauss-Jordan elimination matrices, which eventually put A into upper triangular form. The matrices we multiply by are the *Householder reflection matrices*.

Given $w \in \mathbb{R}^n$ we can associate the matrix

$$U = I - 2\frac{ww^T}{w^T w}$$

which reflects $\mathbb{R}^n$ across the hyperplane $\text{Span}\{w\}^\perp$. The matrix U is call the Householder reflection across this hyperplane.

Given a pair of vectors x and y with

$$\|x\|_2 = \|y\|_2, \quad \text{and} \quad x \neq y,$$

the Householder reflection

$$U = I - 2\frac{(x-y)(x-y)^T}{(x-y)^T(x-y)}$$

is such that $y = Ux$, since

$$\begin{aligned} Ux &= x - 2(x-y)\frac{\|x\|^2 - y^T x}{\|x\|^2 - 2y^T x + \|y\|^2} \\ &= x - 2(x-y)\frac{\|x\|^2 - y^T x}{2(\|x\|^2 - y^T x)} \qquad (\text{since } \|x\| = \|y\|) \\ &= y\ . \end{aligned}$$

We now show how Householder reflections can be used to obtain the QR factorization. Let $\mu := \min\{n, m\}$. The procedure described below terminates in at most $\kappa \leq \tau$ steps. The approach is based on a numerical linear algebra procedure call *deflation* where the dimension of the problem is reduced at each iteration. Here we describe the basic idea of a deflation step in the QR-factorization of the matrix $A_0 \in \mathbb{R}^{m\times n} \setminus \{0\}$. Begin by block decomposing A_0 as

$$A_0 = \begin{bmatrix} \alpha_0 & a_0^T \\ b_0 & \tilde{A}_0 \end{bmatrix}, \text{ with } \tilde{A}_0 \in \mathbb{R}^{(m-1)\times(n-1)},$$

and set

$$\nu_0 = \left\| \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \right\|_2 .$$

If $\nu_0 = 0$, then multiply A_0 on the left by a permutation matrix P_0 to bring a non-zero (largest magnitude) column in A_0 into the first column and the zero column to the last column. Then block decompose A_0P_0 as above with

$$A_0P_0 = \begin{bmatrix} \alpha_0 & a_0^T \\ b_0 & \tilde{A}_0 \end{bmatrix}, \text{ with } \tilde{A}_0 \in \mathbb{R}^{(m-1)\times(n-1)},$$

and set

$$\nu_0 = \left\| \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \right\|_2 \neq 0.$$

Let H_0 be the Householder transformation that maps

$$\begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \mapsto \nu_0\, e_1 \quad :$$

$$H_0 = I - 2\frac{ww^T}{w^T w} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} - \nu_0 e_1 = \begin{pmatrix} \alpha_0 - \nu_0 \\ b_0 \end{pmatrix}.$$

Then, there is a matrix $A_1 \in \mathbb{R}^{(m-1)\times(n-1)}$ and a vector $a_1 \in \mathbb{R}^{n-1}$ such that

$$H_0A_0P_0 = \begin{bmatrix} \nu_0 & a_1^T \\ 0 & A_1 \end{bmatrix}$$

If $\tau = 1$ or $A_1 = 0$, we are done; otherwise, repeat the process on the matrix A_1. Decompose A_1 as

$$A_1 = \begin{bmatrix} \alpha_1 & a_1^T \\ b_1 & \tilde{A}_1 \end{bmatrix}, \text{ with } \tilde{A}_1 \in \mathbb{R}^{(m-2)\times(n-2)},$$

and set

$$\nu_1 = \left\| \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \right\|_2 .$$

If $\nu_1 = 0$, then multiply A_1 on the left by a permutation matrix P_1 to bring a non-zero (largest magnitude) column in A_1 into the first column and the zero column to the last column. Then block decompose $A_1 P_1$ as above with

$$A_1 P_1 = \begin{bmatrix} \alpha_1 & a_1^T \\ b_1 & \tilde{A}_1 \end{bmatrix}, \text{ with } \tilde{A}_1 \in \mathbb{R}^{(m-2)\times(n-2)},$$

and set

$$\nu_1 = \left\| \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \right\|_2 \neq 0.$$

Let H_1 be the Householder transformation that maps

$$\begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \mapsto \nu_1 \, e_1 \quad :$$

$$H_1 = I - 2\frac{ww^T}{w^T w} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} - \nu_1 e_1 = \begin{pmatrix} \alpha_1 - \nu_1 \\ b_1 \end{pmatrix}.$$

Then, there is a matrix $A_2 \in \mathbb{R}^{(m-2)\times(n-2)}$ and a vector $a_2 \in \mathbb{R}^{n-1}$ such that

$$H_1 A_1 P_1 = \begin{bmatrix} \nu_2 & a_2^T \\ 0 & A_2 \end{bmatrix}.$$

Consequently,

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ 0 & H_1 \end{bmatrix} H_0 A_0 P_0 \begin{bmatrix} 1 & 0 \\ 0 & P_1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & H_1 \end{bmatrix} \begin{bmatrix} \nu_0 & a_1^T \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & P_1 \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_1^T \\ 0 & H_1 A_1 P_1 \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_1^T \\ 0 & \begin{bmatrix} \nu_2 & a_2^T \\ 0 & A_2 \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_{12} & \tilde{a}_1^T \\ 0 & \nu_2 & a_2^T \\ 0 & 0 & A_2 \end{bmatrix}. \end{aligned}$$

If $\tau = 2$ or $A_2 = 0$, we are done; otherwise repeat as above on the matrix A_2. This process terminates after $\kappa \leq \tau$ iterations with an upper triangular factorization of the form

$$\widetilde{H}_\kappa \widetilde{H}_{\kappa-1} \dots \widetilde{H}_0 A_0 P_0 \widetilde{P}_1 \dots \widetilde{P}_\kappa = \begin{bmatrix} \nu_1 & a_{12} & a_{13} & a_{14} & \cdots & a_{1\kappa} & \cdots & a_{1n} \\ 0 & \nu_2 & a_{23} & a_{24} & \cdots & a_{2\kappa} & \cdots & a_{2n} \\ 0 & 0 & \nu_3 & a_{34} & \cdots & a_{3\kappa} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \nu_\kappa & \cdots & a_{\kappa n} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} = R,$$

where the zeros below the κ row are absent if $m = \kappa$. Then $P := P_0 \widetilde{P}_1 \dots \widetilde{P}_\kappa$ is a permutation matrix and $Q := \widetilde{H}_\kappa \widetilde{H}_{\kappa-1} \dots \widetilde{H}_0$ is unitary with $A_0 P = QR$. The matrix A_0 is surjective if and only if $\kappa = m$ in which case $P = I$. On the other hand, A_0 is injective if and only if $\kappa = n$ and again $P = I$.

If the above method is implemented by always permuting the column of greatest magnitude into the current pivot column, then

$$AP = QR$$

gives a QR-factorization with the diagonal entries of R nonnegative and listed in the order of descending magnitude, i.e. $\nu_1 \geq \nu_2 \cdots \geq \nu_\kappa > 0$. Since Q is unitary, this is the full QR factorization in (29).

The numerical stability of the procedure can be improved with a slight change to the Householder transformations at each step. Let H_s be the Householder transformation used at iteration s. Redefine H_s so that it maps

$$\begin{pmatrix} \alpha_s \\ b_s \end{pmatrix} \mapsto -\text{sign}(\alpha_s)\nu_s \; e_1 \quad :$$

$$H_s = I - 2\frac{ww^T}{w^T w} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_s \\ b_s \end{pmatrix} + \text{sign}(\alpha_s)\nu_1 e_1 = \begin{pmatrix} \alpha_s + \text{sign}(\alpha_s)\nu_s \\ b_s \end{pmatrix},$$

where

$$\text{sign}(\alpha) := \begin{cases} 1 & , \; \alpha \geq 0 \\ -1 & , \; \alpha < 0 \\ 0. & \end{cases}$$

Note that this avoids the possibility of subtraction by nearly like terms and increases the magnitude of the vector w. The monotonicity and non-negativity of the ν_s's can be recovered on termination by redefining $Q := QD$ and $R := DR$, where $D := \text{diag}(\text{sign}(\nu_1), \text{sign}(\nu_2), \cdots, \text{sign}(\nu_\kappa), 1, \cdots, 1) \in \mathbb{R}^{m \times m}$ is a diagonal unitary matrix since $D^T D = I$.