Chapter 1
Preliminaries

1.1. Introduction

We give some necessary background. The first section gives the notational conventions to be used in the book. In the second section, we give some elementary results and definitions from linear algebra and probability. A third section is on matrix and vector norms, and stresses the importance of the two-norm. The results are presented without proof. For such proofs, see some of the references given in §1.5.

1.2. Notation for Vectors, Matrices, and Algorithms

Vectors and scalars are denoted by lower case Roman or Greek letters. Scalars are written in ordinary type, whereas vectors are written in boldface type. For instance, we write a real \( n \)-vector \( \mathbf{y} \) componentwise in a column as

\[
\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}
\]

where \( y_1, y_2, \ldots, y_n \) are real scalars. We can also have a \( p \)-vector \( \mathbf{\beta} \) given by

\[
\mathbf{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}
\]

Usually instead of saying “\( \mathbf{y} \) is an \( n \)-vector”, we will say “\( \mathbf{y} \in \mathbb{R}^n \)” where \( \mathbb{R}^n \) is the vector space of \( n \)-tuples of real numbers.

Matrices are denoted by capital Roman or Greek letters. Thus the \( m \times n \)
real matrix $X$ and the $p \times q$ real matrix $\Psi$ are given by

$$X = \begin{pmatrix}
x_{11} & x_{12} & \ldots & x_{1n} \\
x_{21} & x_{22} & \ldots & x_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
x_{m-1,1} & x_{m-1,2} & \ldots & x_{m-1,n} \\
x_{m1} & x_{m2} & \ldots & x_{mn}
\end{pmatrix},$$

$$\Psi = \begin{pmatrix}
\psi_{11} & \psi_{12} & \ldots & \psi_{1q} \\
\psi_{21} & \psi_{22} & \ldots & \psi_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
\psi_{p-1,1} & \psi_{p-1,2} & \ldots & \psi_{p-1,q} \\
\psi_{p1} & \psi_{p2} & \ldots & \psi_{pq}
\end{pmatrix}.$$

Corresponding to our designation for vectors, instead of saying “$X$ is a real $m \times n$ matrix”, we will say “$X \in \mathbb{R}^{m \times n}$” where $\mathbb{R}^{m \times n}$ is the vector space of $m \times n$ matrices. We will often write simply

$$X = (x_{ij}) \in \mathbb{R}^{m \times n}.$$

We denote the transpose of a matrix $X \in \mathbb{R}^{m \times n}$ by the matrix $X^T \in \mathbb{R}^{n \times m}$ that is given by

$$X^T = \begin{pmatrix}
x_{11} & x_{21} & \ldots & x_{m1} \\
x_{12} & x_{22} & \ldots & x_{m2} \\
\vdots & \vdots & \ddots & \vdots \\
x_{1,n-1} & x_{2,n-1} & \ldots & x_{m,n-1} \\
x_{1n} & x_{2n} & \ldots & x_{mn}
\end{pmatrix}.$$

We write the matrix $X$ in terms of its column vectors $x_1, \ldots, x_n \in \mathbb{R}^m$ as

$$X = (x_1, \ldots, x_n).$$

If $v \in \mathbb{R}^{1 \times n}$ is a row vector, we always write it as $v^T$ where $v = r^T$ and $r \in \mathbb{R}^n$. Likewise a matrix $X \in \mathbb{R}^{m \times n}$ is written in terms of its row vectors as

$$X = \begin{pmatrix}
r_1^T \\
r_2^T \\
\vdots \\
r_m^T
\end{pmatrix}$$

where $r_1, \ldots, r_m \in \mathbb{R}^n$. Thus vectors are always column vectors.
A vector $\mathbf{y} \in \mathbb{R}^n$ that is the concatenation of vectors $\mathbf{y}_1 \in \mathbb{R}^{n_1}, \ldots, \mathbf{y}_s \in \mathbb{R}^{n_s}$ is written
\[
\mathbf{y} = \begin{pmatrix}
\mathbf{y}_1 \\
\mathbf{y}_2 \\
\vdots \\
\mathbf{y}_s
\end{pmatrix},
\]
or in the more compact form
\[
\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_s^T)^T.
\]

It is often appropriate discuss partitioned matrices. That is, let $X \in \mathbb{R}^{m \times n}$ have the form
\[
X = \begin{pmatrix}
X_{11} & X_{12} & \ldots & X_{1s} \\
X_{21} & X_{22} & \ldots & X_{2s} \\
\vdots & \vdots & \ddots & \vdots \\
X_{r-1,1} & X_{r-1,2} & \ldots & X_{r-1,s} \\
X_{r1} & X_{r2} & \ldots & X_{rs}
\end{pmatrix}
\]
where $X_{ij} \in \mathbb{R}^{m_i \times n_j}$, and $m_1 + m_2 + \ldots + m_r = m$ and $n_1 + n_2 + \ldots + n_s = n$. If any $m_i$ or $n_j$ is zero, the corresponding row or column is absent from the matrix. An upper case $X_{ij}$ will always be used to denote a submatrix whereas the lower case $x_{ij}$ will denote a scalar. A partitioned matrix may also be written in blocks like so
\[
\Psi = \begin{pmatrix}
p_1 & p_2 & p_3 \\
\psi_{11} & \psi_{12} & \psi_{13} \\
n_1 & \psi_{21} & \psi_{22} \\
n_2 & \psi_{31} & \psi_{32} \\
n_3 & \psi_{33}
\end{pmatrix}.
\]

If we wish to discuss a particular submatrix matrix of $X$ or subvector of $\mathbf{y}$, we use the MATLAB convention of writing $X(i:j,k:\ell)$ and $\mathbf{y}(i:j)$ to mean
\[
X(i:j,k:\ell) = \begin{pmatrix}
x_{i,k} & x_{i,k+1} & \ldots & x_{i,\ell} \\
x_{i+1,k} & x_{i+1,k+1} & \ldots & x_{i+1,\ell} \\
\vdots & \vdots & \ddots & \vdots \\
x_{j,k} & x_{j,k+1} & \ldots & x_{j,\ell}
\end{pmatrix}, \quad \mathbf{y}(i:j) = \begin{pmatrix}
y_i \\
y_{i+1} \\
\vdots \\
y_j
\end{pmatrix}.
\]

We shall also use $X(:,k:\ell)$ to denote all of columns $k$ through $\ell$ and $X(i:,;)$ to denote all of rows $i$ through $j$.

For a vector $\mathbf{y} \in \mathbb{R}^n$ or a matrix $X = (x_{ij}) \in \mathbb{R}^{m \times n}$ their componentwise absolute values $|\mathbf{y}| \in \mathbb{R}^n$ and $|X| \in \mathbb{R}^{m \times n}$ are given by
\[
|\mathbf{y}| = (|y_1|, |y_2|, \ldots, |y_n|)^T, \quad |X| = (|x_{ij}|).
\]
Likewise the relations $=, <, >, \leq, \geq$ applied to vectors and matrices are valid if and only if they are valid for every component. That is for $X = (x_{ij}), Y = (y_{ij}) \in \mathbb{R}^{m \times n},$

\[ X \leq Y \]

is the same as the statement

\[ x_{ij} \leq y_{ij}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n. \]

Clearly the relation $\neq$ is valid for vectors and matrices if and only if it is valid for any component. A matrix $A \in \mathbb{R}^{m \times n}$ is diagonal if

\[ \lambda_{ij} = 0, \quad i \neq j. \]

We will use a special notation for diagonal matrices:

\[ A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \quad \lambda_{ij} = \begin{cases} 0, & i \neq j \\ \lambda_i, & i = j. \end{cases} \]

We will also use $A = \text{diag}(\mathbf{\lambda})$ where $\mathbf{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T.$

Algorithms will be written either using an English step-by-step description, a stylized version of MATLAB [?], or some mixture of the two. We give two examples of how we write functions below. The first example is the computation of the absolute value function. A sophisticated knowledge of MATLAB will not be necessary to understand our algorithms. With rare exceptions, our descriptions will be based upon basic scalar, vector, and matrix operations. Two such exceptions are the MATLAB functions \texttt{size} and \texttt{length}. For $X \in \mathbb{R}^{m \times n},$

\[ [m, n] = \text{size}(X); \quad (1.1) \]

where $[m, n]$ denotes the ordered pair of output variables $m$ and $n$. The function \texttt{length} yields

\[ p = \text{length}(X) = \max\{m, n\} \]

where $m$ and $n$ are the values generated by \texttt{size} in (1.1). For a vector $x \in \mathbb{R}^n, \quad n = \text{length}(x)$ and $[n, 1] = \text{size}(x).$
1.2. NOTATION FOR VECTORS, MATRICES, AND ALGORITHMS

Function 1.1 (Absolute Value Function)

\[
\text{function } absx = abs(x) \\
\text{if } x \geq 0 \\
\quad absx = x; \\
\text{else} \\
\quad absx = -x; \\
\text{end}; \\
\text{end abs}
\]

The second example call \textit{infnorm} takes a vector \( \mathbf{x} \in \mathbb{R}^n \) and produces its maximum component in absolute value called \( x_{\text{max}} \) and the index of the first component with that value call \( j_{\text{max}} \). As will be described in §7, this is the infinity norm which we write

\[
x_{\text{max}} = |x_{j_{\text{max}}}| = \|\mathbf{x}\|_{\infty} = \max_{1 \leq j \leq n} |x_j|.
\]

Function 1.2 (Infinity Norm)

\[
\text{function } [x_{\text{max}}, j_{\text{max}}] = \text{infnorm}(\mathbf{x}) \\
\text{if } n > 0 \\
\quad x_{\text{max}} = |x_1|; j_{\text{max}} = 1; \\
\quad \text{for } j = 2:n \\
\quad \quad absx_j = |x_j|; \\
\quad \quad \text{if } absx_j > x_{\text{max}} \\
\quad \quad \quad x_{\text{max}} = absx_j; j_{\text{max}} = j; \\
\quad \text{end; } \\
\quad \text{end; } \\
\text{else} \\
\quad x_{\text{max}} = 0; j_{\text{max}} = 0; \\
\text{end } \\
\text{end infnorm}
\]

The general form of a function statement is

\[
\text{function } [b_1, \ldots, b_m] = \text{fun}(a_1, \ldots, a_n)
\]
where \( a_1, \ldots, a_n \) are the formal input parameters and \( b_1, \ldots, b_m \) are the formal output parameters. Functions are called by a statement of the form

\[ \text{fun}(c_1, \ldots, c_n) = [d_1, \ldots, d_m] \]

where \( d_1, \ldots, d_m \) and \( c_1, \ldots, c_n \) are the actual parameters. The parameters may be scalars, vectors, or matrices. All functions are called by value and information is transmitted from the function only through the output parameters.

Very often we will use vector and matrix operations in our description. For example, a dot product may be written

\[ \text{dot} = x^T y, \]

the above gaxpy operation may be written,

\[ y = y + X \beta, \]

and a matrix-matrix product may just be written

\[ C = AB. \]

### 1.3. Elementary Definitions and Results

#### 1.3.1. Linear Algebra Basics

Below are some definitions and elementary results from linear algebra that are used in this book.

A *subspace* \( S \) of \( \mathbb{R}^m \) is a subset such for any real numbers \( \alpha \) and \( \beta \),

\[ x, y \in S, \]

implies that

\[ \alpha x + \beta y \in S. \]

The *dimension* of a subspace \( k \), is the smallest number of vectors \( x_1, \ldots, x_k \in S \) such that for any \( y \in S \), there are scalars \( \beta_1, \ldots, \beta_k \) such that

\[ y = \sum_{i=1}^{k} \beta_i x_i. \]

The *range* of a matrix \( X \in \mathbb{R}^{m \times n} \) is the set

\[ \text{range}(X) = \{ Xy : y \in \mathbb{R}^n \}. \]

The range is a subspace of \( \mathbb{R}^m \). The *rank* of \( X \) is the dimension of its range space and is written \( \text{rank}(X) \). Clearly, \( \text{rank}(X) \leq \min\{m, n \} \). The *span* of a set of vectors \( x_1, \ldots, x_n \) is given by

\[ \text{span}\{ x_1, \ldots, x_n \} = \text{range}(X) \]
1.3. ELEMENTARY DEFINITIONS AND RESULTS

where \( X = (x_1, \ldots, x_n) \). If \( \text{rank}(X) = n \), then \( x_1, \ldots, x_n \) is a basis for \( \text{range}(X) \) and \( X \) is a basis matrix. A matrix \( X \) is rank deficient if \( \text{rank}(X) < \min\{m, n\} \), is said to have full column rank if \( \text{rank}(X) = n \), and is said to have full row rank if \( \text{rank}(X) = m \).

The null space of \( X \) is the linear subspace of \( \mathbb{R}^n \) given by

\[
\text{null}(X) = \{ y \in \mathbb{R}^n : Xy = 0 \}.
\]

The matrix \( I_n = \text{diag}(1, 1, \ldots, 1) \) is the identity matrix for \( \mathbb{R}^{n \times n} \). The \( j \)th column of the identity matrix will be denoted \( e_j \).

A square matrix \( X \in \mathbb{R}^{n \times n} \) is nonsingular if it has full column rank (or full row rank), and said to be singular if it is rank deficient. If \( X \) is nonsingular then there exists an inverse matrix, denoted \( X^{-1} \), such that

\[
XX^{-1} = X^{-1}X = I_n.
\]

Furthermore, for any \( b \in \mathbb{R}^n \), there exists a vector \( y \in \mathbb{R}^n \) such that

\[
Xy = b
\]

given by \( y = X^{-1}b \).

For a square matrix \( X \) we also define a real-valued function called the determinant denoted \( \det(\cdot) \). It is defined recursively as follows. If \( n = 1 \) then \( \det(X) = x \) where \( X = (x) \). For larger values of \( n \), let \( X_{ij} \) be the \((n-1) \times (n-1)\) submatrix resulting from deleting row \( i \) and column \( j \). Then

\[
\det(X) = \sum_{i=1}^{n} (-1)^{i+1} x_{ii} \det(X_{ii}).
\]

The value \( \det(X_{ij}) \) is called the \((i, j)\) minor of \( X \). The following facts about determinants are useful:

1. \( \det(X) = \det(X^T) \).
2. \( \det(XY) = \det(X) \det(Y) \).
3. \( \det(X) = 0 \) if and only if \( X \) is singular.
4. \( \det(X^{-1}) = 1 / \det(X) \).
5. If \( X \) has the form

\[
X = \begin{pmatrix} k & n-k \\ n-k & X_{11} & X_{12} \\ \end{pmatrix}
\]

then

\[
\det(X) = \det(X_{11}) \det(X_{22}).
\]
A square matrix $X \in \mathbb{R}^{n \times n}$ is \textit{symmetric} if $X = X^T$. A symmetric matrix $X$ is \textit{positive semidefinite} if for all $y \in \mathbb{R}^n$ we have
\[ y^T X y \geq 0. \]
We say that $X$ is \textit{positive definite} if
\[ y^T X y > 0, \quad \text{if } y \neq 0. \]
A symmetric matrix $X$ is \textit{indefinite} if neither $X$ nor $-X$ is positive semidefinite.

\textbf{1.3.2. Probability Basics}

Some definitions and elementary theorems from probability are necessary in explaining the context of linear least squares problems. These definitions correspond to those from Rao [3, 1973].

Let \( \Pr\{A\} \) denote the probability that the event $A$ occurs. A real valued \textit{random variable} $x$ is a variable associated with a distribution function $F(\cdot) : \mathbb{R} \to [0,1]$ such that
\[ F(a) = \Pr\{x < a\} \]
for all $a \in \mathbb{R}$. The distribution function $F(\cdot)$ has the properties
\[
\lim_{a \to -\infty} F(a) = 0, \quad \lim_{a \to \infty} F(a) = 1,
\]
\[ F(a_1) \leq F(a_2) \text{ if } a_1 \leq a_2, \]
\[ \lim_{a \to b^-} F(a) = F(b). \]
The last property says that $F(\cdot)$ is continuous from the left. (Some books define distribution functions as continuous from the right.)

The distributions referred to in this book are continuous and differentiable. Thus they will have associated with them a \textit{probability density function} given by
\[ f(a) = F'(a). \quad (1.2) \]

The expectation of function $g(\cdot)$ of a random variable $x$ is defined by
\[ \mathbb{E}(g(x)) = \int_{-\infty}^{\infty} g(a)f(a)da \]
where $f(\cdot)$ is defined in (1.2). Important expectations include the \textit{mean} defined by
\[ \mu = \text{Mean}(x) = \mathbb{E}(x) = \int_{-\infty}^{\infty} a f(a) da \]
and variance given by

\[ \sigma^2 = \text{Var}(x) = \mathbb{E}((x - \mu)^2) = \int_{-\infty}^{\infty} (a - \mu)^2 f(a) \, da. \]

The most used density function is the normal or Gaussian density function. It is parameterized by its mean \( \mu \) and its variance \( \sigma^2 \) and is given by

\[ f(a) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(a - \mu)^2}{2\sigma^2}\right). \]  

(1.3)

The associated distribution function \( F(a) \) does not have a closed form expression. The distribution is referred to as \( \text{N}(\mu, \sigma^2) \).

Likewise, we can define a random vector \( \mathbf{x} \in \mathbb{R}^n \) as a vector associated with a distribution function \( F : \mathbb{R}^n \to [0, 1] \) such that

\[ F(\mathbf{a}) = \Pr\{ \mathbf{x} < \mathbf{a} \}. \]

Again, we are concerned only with distributions that are continuous and differentiable. If \( \mathbf{a} = (a_1, \ldots, a_n)^T \), the density function \( f(\cdot) \) is defined by

\[ f(\mathbf{a}) = \frac{\partial^n F(\mathbf{a})}{\partial a_1 \cdots \partial a_n}. \]

The definition of expectation extends in an obvious manner to vectors in \( \mathbb{R}^n \). If \( g(\cdot) \) is a function mapping \( \mathbb{R}^n \) to \( \mathbb{R}^m \) then

\[ \mathbb{E}(g(\mathbf{x})) = \int_{\mathbb{R}^n} g(\mathbf{a}) f(\mathbf{a}) \, d\mathbf{a}. \]

Expectation is a linear function. That is, for any \( A \in \mathbb{R}^{m \times n} \), we have

\[ \mathbb{E}(Ag(\mathbf{x})) = A \mathbb{E}(g(\mathbf{x})). \]

The mean is then given by

\[ \mathbf{\mu} = \mathbb{E}(\mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{a} f(\mathbf{a}) \, d\mathbf{a}. \]  

(1.4)

and the variance-covariance matrix is given by

\[ \Sigma = \text{Cov}(\mathbf{x}) = \mathbb{E}((\mathbf{x} - \mathbf{\mu})(\mathbf{x} - \mathbf{\mu})^T) = \int_{\mathbb{R}^n} (\mathbf{a} - \mathbf{\mu})(\mathbf{a} - \mathbf{\mu})^T f(\mathbf{a}) \, d\mathbf{a}. \]

There is also a nice matrix relation for variance-covariance matrices. For any \( A \in \mathbb{R}^{m \times n} \), it is easily verified that

\[ \text{Cov}(A\mathbf{x}) = A \text{Cov}(\mathbf{x}) A^T. \]
CHAPTER 1. PRELIMINARIES

The covariance of any two linear functions of $\mathbf{x}$, say, $\mathbf{b}^T \mathbf{x}$ and $\mathbf{c}^T \mathbf{x}$ is defined by

$$
\text{Cov}(\mathbf{b}^T \mathbf{x}, \mathbf{c}^T \mathbf{x}) = \mathbf{b}^T \text{Cov}(\mathbf{x}) \mathbf{c}.
$$

The matrix $\Sigma = \text{Cov}(\mathbf{x})$ is always symmetric and positive semidefinite. It is positive definite if and only if

$$
\text{Pr}\{\mathbf{c}^T (\mathbf{x} - \mu) = 0\} \neq 1
$$

for all nonzero vectors $\mathbf{c} \in \mathbb{R}^n$.

Mostly we are interested in the multivariate normal distribution. That distribution is parameterized by its mean vector $\mu$ and its variance-covariance matrix $\Sigma$. Its density function is

$$
f(\mathbf{a}) = c_n \exp(- (\mathbf{a} - \mu)^T \Sigma^{-1} (\mathbf{a} - \mu)/2) \tag{1.5}
$$

where $c_n$ is a constant given by

$$
c_n = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}}.
$$

This distribution is denoted $\mathcal{N}(\mu, \Sigma)$. The normal distribution can be defined for a singular $\Sigma$, but no density function exists for it on all of $\mathbb{R}^n$. However, the density function does exist on a subspace, see Rao [3, 1973, pp.527-528].

A vector $\mathbf{x}$ of random variables is uncorrelated if the matrix $\text{Cov}(\mathbf{x})$ is diagonal, otherwise $\mathbf{x}$ is said to be correlated. It is equivalent to say that the components of $\mathbf{x}$, $x_1, \ldots, x_n$, are a set of uncorrelated (correlated) random variables.

Let $\text{Pr}\{\mathcal{A}|\mathcal{B}\}$ denote the probability that event $\mathcal{A}$ will occur given that $\mathcal{B}$ is known to have occurred. The vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is a vector of independent random variables $x_1, \ldots, x_n$ if for any two sets $D_1 \subseteq \mathbb{R}$ and $D_2 \subseteq \mathbb{R}^{n-1}$ and any integer $i$,

$$
\text{Pr}\{x_i \in D_1 | (x_1, \ldots, x_{i-1}, x_i, \ldots, x_n)^T \in D_2\} = \text{Pr}\{x_i \in D_1\}.
$$

We may also say that the components of $\mathbf{x}$, $x_1, \ldots, x_n$ form an independent set. Any set of independent random variables is also uncorrelated, but not all uncorrelated sets of random variables are independent. However, if $\mathbf{x}$ has a multivariate normal distribution, $\mathbf{x}$ is uncorrelated if and only if $\mathbf{x}$ is independent.

An estimate $\hat{\beta}$ of a parameter $\beta$ is a function of a random vector $\mathbf{x}$. It is unbiased if

$$
\text{E}(\hat{\beta}) = \beta,
$$

and biased otherwise.
1.4. Inner Products, Vector Norms, and Matrix Norms

Let $f(x; \theta)$ be a family of density functions parameterized by a vector of parameters $\theta = (\theta_1, \ldots, \theta_p)^T$ belonging to a set $\Theta \subseteq \mathbb{R}^p$. The likelihood function $\ell(\theta; x)$ is a function of $\theta$ parameterized by $x$ which satisfies

$$\ell(\theta; x) = cf(x; \theta).$$

where $c$ is a constant that can be ignored. A maximum likelihood estimate (MLE) estimate $\hat{\theta}$ of $\theta$ is an estimate such that

$$\ell(\hat{\theta}; x) = \sup_{\theta \in \Theta} \ell(\theta; x).$$

For instance, consider $x$ from a family of normal distributions with $N(\mu c, \sigma^2 I)$ where $c = (1, 1, \ldots, 1)^T$ and $\mu$ and $\sigma^2$ are unknown parameters. It is desired to estimate $\mu$ from the sampled random variables in the vector $x$. Thus we have the likelihood function

$$\ell(\mu, \sigma^2; x) = \sigma^{-n} \exp\left(-\frac{(x - \mu c)^T(x - \mu c)}{2\sigma^2}\right).$$

The constant factor $(2\pi)^{-n/2}$ can be ignored in maximizing the likelihood function. It can be shown that the maximum likelihood estimates of $\mu$ and $\sigma^2$ are

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.$$

The estimate $\hat{\mu}$ is unbiased, but $E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2$ and is thus biased. An unbiased estimate is

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.$$

In $\S$ we show that the linear least squares problem arises out of a maximum likelihood estimation problem.

1.4. Inner Products, Vector Norms, and Matrix Norms

We briefly review basic results for inner products, vector norms, and matrix norms. In $\S 1.4.1$ basic properties of norms and inner products are given. In $\S 1.4.2$, we discuss the properties of the norms that are based on the Euclidean norm.
1.4.1. Basic Definitions and Lemmas

First, we give the definition of an inner product.

**Definition 1.1** An inner product in \( \mathbb{R}^n \) is a function \( (\cdot,\cdot) \) mapping \( \mathbb{R}^n \times \mathbb{R}^n \) into \( \mathbb{R} \) that satisfies the four axioms

1. \( (x,x) \geq 0; \) 
   \( (x,x) = 0, \) if and only if \( x = 0, \) \( x \in \mathbb{R}^n, \)
2. \( (ax,y) = a(x,y), \) \( x,y \in \mathbb{R}^n, a \in \mathbb{R}, \)
3. \( (x,y) = (y,x), \) \( x,y \in \mathbb{R}^n, \)
4. \( (x+z,y) = (x,y) + (z,y), \) \( x,y,z \in \mathbb{R}^n. \)

We note that if we define an inner product for complex vectors, the third axiom becomes

\[
(x,y) = \overline{(y,x)} \quad x,y \in \mathbb{C}^n
\]

where \( \overline{\cdot} \) denotes the complex conjugate of \( \cdot \).

The inner product that is most often is the Euclidean dot product

\[
(x,y) = x^T y = \sum_{i=1}^{n} x_i y_i .
\]

Others take the form

\[
(x,y) = x^T A y
\]

for an \( A \in \mathbb{R}^{n \times n} \) that is symmetric and positive definite.

The **Cauchy-Schwarz** inequality given below is quite useful for all inner products.

**Lemma 1.3 (Cauchy-Schwarz inequality)** Let \( (\cdot,\cdot) \) be an inner product in \( \mathbb{R}^n \). Then for all \( x,y \in \mathbb{R}^n \) we have

\[
| (x,y) | \leq (x,x)^{1/2} (y,y)^{1/2} .
\]

Moreover, equality in (1.7) holds if and only if \( x = \alpha y \) for some \( \alpha \in \mathbb{R} \).

This inequality leads to the following definition of the angle between two vectors relative to an inner product.

**Definition 1.2** The angle \( \theta \) between the two nonzero vectors \( x,y \in \mathbb{R}^n \) with respect to an inner product \( (\cdot,\cdot) \) is given by

\[
\cos \theta = \frac{(x,y)}{(x,x)^{1/2} (y,y)^{1/2}}.
\]
1.4. INNER PRODUCTS, VECTOR NORMS, AND MATRIX NORMS

The next important definition is that of a vector norm.

**Definition 1.3** A norm in $\mathbb{R}^n$ is a function $\| \cdot \|$ mapping $\mathbb{R}^n$ into $\mathbb{R}$ satisfying the following three axioms

1. $\|x\| \geq 0$;
   $\|x\| = 0$ if and only if $x = 0$, $x \in \mathbb{R}^n$.
2. $\|\alpha x\| = |\alpha| \|x\|$, $x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$, $x, y \in \mathbb{R}^n$.

Definition 1.3 and the Cauchy-Schwarz inequality give us that for any inner product $(\cdot, \cdot)_\alpha$ we can define a norm $\| \cdot \|_\alpha$ by

$$\|x\|_\alpha = (x, x)^{1/2}. \quad (1.9)$$

The most well known (and arguably the most important) of these norms is the Euclidean norm given by

$$\|x\|_2 = (x^T x)^{1/2} = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}. \quad (1.10)$$

It is also referred to as the two-norm. The subscript “2” is explained below.

The two-norm is one of the class of $p$-norms. These are given by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad p \geq 1. \quad (1.11)$$

Clearly, $p = 2$ leads to the definition (1.10). Except for the two-norm, the only norms we will use from this class are the one-norm given by

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad (1.12)$$

and the $\infty$-norm given by

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| = \lim_{p \to \infty} \|x\|_p. \quad (1.13)$$

The next lemma, the Hölder inequality, relates the $p$-norms to the Euclidean inner product.

**Lemma 1.4 (Hölder inequality)** Let $\| \cdot \|_p$ and $\| \cdot \|_q$ be norms in $\mathbb{R}^n$ from the class (1.11) such that $p^{-1} + q^{-1} = 1$. Then for all $x, y \in \mathbb{R}^n$ we have

$$|x^T y| \leq \|x\|_p \|y\|_q. \quad (1.14)$$
The interesting cases for us are \( p = q = 2, p = 1, q = \infty, \) and \( p = \infty, q = 1. \) For \( p = q = 2, \) the Hölder inequality is just the Cauchy-Schwarz inequality.

The next inequality states that if we can bound \( \|x\|_\alpha, \) for a given \( x \in \mathbb{R}^n \) in any norm \( \|\cdot\|_\alpha, \) the quantity \( \|x\|_\beta \) can be bounded in any other norm \( \|\cdot\|_\beta. \)

**Lemma 1.5** All norms on \( \mathbb{R}^n \) are uniformly equivalent, meaning that for any two norms \( \|\cdot\|_\alpha \) and \( \|\cdot\|_\beta \) there are constants \( c_1 \) and \( c_2 \) such that

\[
c_1 \|x\|_\beta \leq \|x\|_\alpha \leq c_2 \|x\|_\beta, \quad c_1, c_2 > 0
\]

for all \( x \in \mathbb{R}^n. \)

For the two-norm, the one-norm, and the \( \infty-\)norm the uniform equivalence relations are summarized by

\[
\frac{1}{\sqrt{n}} \|x\|_1 \leq \|x\|_2 \leq \|x\|_1, \quad (1.16)
\]

\[
\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty, \quad (1.17)
\]

\[
\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty. \quad (1.18)
\]

We will also need norms for matrices.

**Definition 1.4** A norm in \( \mathbb{R}^{m \times n} \) is a function \( \|\cdot\| \) mapping \( \mathbb{R}^{m \times n} \) into \( \mathbb{R} \) satisfying the following three axioms

1. \( \|X\| \geq 0; \)
   \( \|X\| = 0 \) if and only if \( X = 0, \quad X \in \mathbb{R}^{m \times n} \)

2. \( \|\alpha X\| = |\alpha| \|X\|, \quad X \in \mathbb{R}^{m \times n}, \alpha \in \mathbb{R} \)

3. \( \|X + Y\| \leq \|X\| + \|Y\|, \quad X, Y \in \mathbb{R}^{m \times n}. \)

This definition is isomorphic to the definition of a vector norm on \( \mathbb{R}^m. \)

For example, the Frobenius norm defined by

\[
\|X\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2 \right)^{1/2}
\]

is isomorphic to the two-norm on \( \mathbb{R}^{mn}. \)

For semantic reasons, we define a family of norms.
1.4. INNER PRODUCTS, VECTOR NORMS, AND MATRIX NORMS

**Definition 1.5** Let \( \| \cdot \|_{\alpha,m,n} \) be a norm on \( \mathbb{R}^{m \times n} \) and let it be well-defined for every positive integer \( m \) and \( n \). Then the set

\[ \mathcal{N}_\alpha = \{ \| \cdot \|_{\alpha,m,n} : m, n \text{ positive integers} \} \]

is called a family of norms. For any positive integers \( m \) and \( n \), and any \( X \in \mathbb{R}^{m \times n} \), we denote \( \| X \|_{\alpha,m,n} \) by \( \| X \|_\alpha \). That is, for any matrix \( X \), the quantity \( \| X \|_\alpha \) is the appropriate member of \( \mathcal{N}_\alpha \) applied to \( X \).

The set \( \mathcal{V}_\alpha \subseteq \mathcal{N}_\alpha \) given by

\[ \mathcal{V}_\alpha = \{ \| \cdot \|_{\alpha,m,1} : m \text{ positive integer} \} \]

is the associated family of vector norms. For any \( x \in \mathbb{R}^m \), the quantity \( \| x \|_\alpha = \| (x) \|_\alpha \) is the appropriate member of \( \mathcal{V}_\alpha \) applied to \( x \).

Since \( X \) represents a linear operator from \( \mathbb{R}^n \) to \( \mathbb{R}^m \), it is appropriate to define the induced norm \( \| \cdot \|_\alpha \) on \( \mathbb{R}^{m \times n} \) by

\[ \| X \|_\alpha = \sup_{y \neq 0} \frac{\| Xy \|_\alpha}{\| y \|_\alpha}. \tag{1.20} \]

It is a simple matter to show that

\[ \| X \|_\alpha = \max_{\| y \|_\alpha = 1} \| Xy \|_\alpha. \tag{1.21} \]

Note that the maximum is taken over a closed, bounded set, thus we have that

\[ \| X \|_\alpha = \| Xy^* \|_\alpha \tag{1.22} \]

for some \( y^* \) such that \( \| y^* \|_\alpha = 1 \). The above definition leads to the very useful bound

\[ \| Xy \|_\alpha \leq \| X \|_\alpha \| y \|_\alpha \tag{1.23} \]

where equality occurs for every vector of the form \( \gamma y^* \), \( \gamma \in \mathbb{R} \).

If \( X \in \mathbb{R}^{m \times 1} \) then (1.23) appears to give a second definition for \( \| X \|_\alpha \). However, it is easily verified that if \( X = (x) \), for some \( x \in \mathbb{R}^m \), then \( \| x \|_\alpha \), defined from the matrix norm (1.23), and \( \| x \|_\alpha \), defined from the vector norm, are the same value. Thus the notation \( \| \cdot \|_\alpha \) is unambiguous for all \( m \) and \( n \).

For any induced norm \( \| \cdot \|_\alpha \), the identity matrix \( I_n \) for \( \mathbb{R}^{n \times n} \) satisfies

\[ \| I_n \|_\alpha = 1. \tag{1.24} \]

However, for the Frobenius norm

\[ \| I_n \|_F = \sqrt{n}, \]

thus it is not an induced norm for any vector norm.
For the one-norm and the \( \infty \)-norm there are formulas for the corresponding matrix norms and for a vector \( \mathbf{y}^* \) satisfying (1.22). The one-norm formula is
\[
\|X\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |x_{ij}|. \tag{1.25}
\]
If \( j_{\max} \) is the index of a column such that
\[
\|X\|_1 = \sum_{i=1}^m |x_{i,j_{\max}}|
\]
then \( \mathbf{y}^* = e_{j_{\max}} \), the corresponding column of the identity matrix.

The \( \infty \)-norm formula is
\[
\|X\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |x_{ij}|. \tag{1.26}
\]
If \( i_{\max} \) is the index of a row such that
\[
\|X\|_\infty = \sum_{j=1}^n |x_{i_{\max},j}|
\]
then the vector \( \mathbf{y}^* = (y_1^*, \ldots, y_n^*)^T \) with components
\[
y_j^* = \text{sign}(x_{i_{\max}j})
\]
satisfies (1.22). Note that \( \|X\|_\infty = \|X^T\|_1 \).

The matrix two-norm does not have a formula like (1.25) or (1.26) and all other formulations are really equivalent to (1.20). Moreover, computing the vector \( \mathbf{y}^* \) in (1.22) is a nontrivial task that we will discuss in Chapters ?? and ??.

The induced norms have a convenient property that is important in understanding matrix computations. For \( X \in \mathbb{R}^{m \times n} \) and \( Y \in \mathbb{R}^{n \times r} \) consider \( \|XY\|_\alpha \). We have that
\[
\|XY\|_\alpha = \max_{\|z\|_\alpha = 1} \|XYz\|_\alpha \leq \max_{\|z\|_\alpha = 1} \|X\|_\alpha \|YZ\|_\alpha
\]
\[
= \|X\|_\alpha \max_{\|z\|_\alpha = 1} \|YZ\|_\alpha = \|X\|_\alpha \|Y\|_\alpha.
\]
Thus
\[
\|XY\|_\alpha \leq \|X\|_\alpha \|Y\|_\alpha. \tag{1.27}
\]
A norm \( \| \cdot \|_\alpha \) (or really family of norms) that satisfies the property (1.27) is said to be consistent. Since they are induced norms the two-norm, one-norm, and the \( \infty \)-norm are all consistent. The Frobenius norm also satisfies (1.27). An example of a matrix norm that is not consistent is given below.
1.4. INNER PRODUCTS, VECTOR NORMS, AND MATRIX NORMS

Example 1.1 Consider the norm $\| \cdot \|_\beta$ on $\mathbb{R}^{m \times n}$ given by

$$
\|X\|_\beta = \max_{(i,j)} |x_{ij}|
$$

This is simply the $\infty$-norm applied to $X$ written out as vector in $\mathbb{R}^{mn}$. For $m = n = 2$, consider

$$
X = Y = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}
$$

Note that

$$
XY = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}
$$

and thus $\|XY\|_\beta = 2 > \|X\|_\beta \|Y\|_\beta = 1$. Clearly, $\| \cdot \|_\beta$ norm is not consistent.

Henceforth, we use only consistent families of norms.

The above defined matrix norms satisfy the following useful inequalities for all $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{n \times s}$:

$$
(mn)^{-\frac{1}{4}} (\|X\|_1 \|X\|_\infty)^{1/2} \leq \|X\|_2 \leq (\|X\|_1 \|X\|_\infty)^{1/2},
$$

(1.28)

$$
\max \left\{ \frac{1}{\sqrt{m}}, \frac{1}{\sqrt{n}} \right\} \|X\|_F \leq \|X\|_2 \leq \|X\|_F,
$$

(1.29)

$$
\|XY\|_F \leq \|X\|_2 \|Y\|_F,
$$

(1.30)

$$
\|XY\|_F \leq \|X\|_F \|Y\|_2.
$$

(1.31)

For $X \neq 0$, the lower bound in (1.29) can be tightened into

$$
\frac{1}{\sqrt{\text{rank}(X)}} \|X\|_F \leq \|X\|_2.
$$

(1.32)

Since rank($X$) is often more expensive to compute than a good approximation to $\|X\|_2$, the practical use of (1.32) is limited.

For any diagonal matrix $A = \text{diag}(\lambda)$ and any $p$-norm $\| \cdot \|_p$

$$
\|A\|_p = \|\lambda\|_\infty = \max_{1 \leq i \leq n} |\lambda_i| \quad 1 \leq p \leq \infty.
$$

For any matrix $X \in \mathbb{R}^{m \times n}$ we have the following relations for the matrix $|X|$:

$$
\|X\|_2 \leq \| |X| \|_2 \leq \sqrt{\text{rank}(X)} \| |X| \|_2 \leq \min \{ \sqrt{m}, \sqrt{n} \} \|X\|_2,
$$

$$
\|X\|_1 = \| |X| \|_1, \quad \|X\|_\infty = \| |X| \|_\infty,
$$

$$
\|X\|_F = \| |X| \|_F.
$$

The following theorem is a simplification of result due to Robert [?].
Theorem 1.6 For any \( p \geq 1 \) let \( X \in \mathbb{R}^{m \times n} \) be partitioned

\[
X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1s} \\
X_{21} & X_{22} & \cdots & X_{2s} \\
\vdots & \vdots & \ddots & \vdots \\
X_{r-1,1} & X_{r-1,2} & \cdots & X_{r-1,p} \\
X_{r1} & X_{r2} & \cdots & X_{rs}
\end{pmatrix}.
\]

Then

\[
\|X\|_p \leq \left( \begin{array}{ccccc}
\|X_{11}\|_p & \|X_{12}\|_p & \cdots & \|X_{1s}\|_p \\
\|X_{21}\|_p & \|X_{22}\|_p & \cdots & \|X_{2s}\|_p \\
\vdots & \vdots & \ddots & \vdots \\
\|X_{r-1,1}\|_p & \|X_{r-1,2}\|_p & \cdots & \|X_{r-1,p}\|_p \\
\|X_{r1}\|_p & \|X_{r2}\|_p & \cdots & \|X_{rs}\|_p
\end{array} \right)_p
\]

In the context of linear least squares, our interest will be in the two-norm or in the Frobenius norm. In this book, Theorem 1.6 is used only for \( p = 2 \). The other norms will be used to bound them. Some special properties of these norms are given in the next section.

1.4.2. The Two-Norm, the Frobenius Norm, and Orthogonality

We begin by defining orthogonality. We then relate orthogonality to the matrix two-norm and Frobenius norm.

Definition 1.6 Two vectors \( \mathbf{u}, \mathbf{v} \in \mathbb{R}^n \) are orthogonal with respect to an inner product \((\cdot, \cdot)\) if \((\mathbf{u}, \mathbf{v}) = 0\). The set \( S_1 \subseteq \mathbb{R}^n \) is orthogonal to the set \( S_2 \subseteq \mathbb{R}^n \), if for each \( \mathbf{u} \in S_1 \) and \( \mathbf{v} \in S_2 \), we have \((\mathbf{u}, \mathbf{v}) = 0\).

We write

\[
\mathbf{u} \perp \mathbf{v}, \quad S_1 \perp S_2
\]

to mean “\( \mathbf{u} \) is orthogonal to \( \mathbf{v} \)” and “\( S_1 \) is orthogonal to \( S_2 \).”

The zero vector is orthogonal to any vector and the set \{0\} is orthogonal to any set.

The following lemma gives two well known results from Euclidean geometry.

Lemma 1.7 For all \( \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \) we have

1. (Law of Cosines) If \( \mathbf{x} \) and \( \mathbf{y} \) are nonzero then

\[
\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| \cos \theta
\]

where \( \cos \theta \) is defined by (1.8) for the Euclidean inner product.
2. (Pythagorean Theorem) If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal then
\[
\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.
\]

An orthogonal set of vectors is defined as follows.

**Definition 1.7** A set of $k$ vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, where each $\mathbf{u}_i \in \mathbb{R}^n$, is said to be an orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle$ if $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for $i \neq j$. The set is said to be orthonormal if it is orthogonal and $\langle \mathbf{u}_i, \mathbf{u}_i \rangle = 1$ for $i = 1, 2, \ldots, k$.

The definition of an orthogonal matrix is related to the definition for vectors.

**Definition 1.8** The matrix $U = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k) \in \mathbb{R}^{n \times k}$ whose columns form an orthonormal set is said to be left orthogonal. If $k = n$, that is, $U$ is square, then $U$ is said to be an orthogonal matrix.

Note that the columns of (left) orthogonal matrices are orthonormal, not merely orthogonal. Square complex matrices whose columns form an orthonormal set are called unitary.

If $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, then an orthogonal matrix $U$ satisfies
\[
U^T U = U U^T = I_n
\]
where $I_n$ is the identity matrix of order $n$. Thus the inverse of an orthogonal matrix is its transpose. For a left orthogonal matrix, we merely have
\[
U^T U = I_n.
\]

Unless stated otherwise, orthogonality and orthonormality will be with respect to the Euclidean inner product.

The decomposition of $\mathbb{R}^n$ into two complementary subspaces is fundamental to least squares computation.

**Lemma 1.8** Let $\mathcal{S} \subset \mathbb{R}^n$ be a subspace. There exists a unique subspace $\mathcal{S}^\perp$, called the orthogonal complement of $\mathcal{S}$ with the properties

- $\mathcal{S} \perp \mathcal{S}^\perp$
- For every $\mathbf{z} \in \mathbb{R}^n$ there exists a unique $\mathbf{x} \in \mathcal{S}$ and a unique $\mathbf{y} \in \mathcal{S}^\perp$ such that $\mathbf{z} = \mathbf{x} + \mathbf{y}$.

The following lemma gives some important facts about orthogonal sets.
Lemma 1.9 Let \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \) be an orthogonal (orthonormal) set where each \( \mathbf{u}_i \in \mathbb{R}^n \). Then

1. The integer \( k \), the cardinality of the set, satisfies \( k \leq n \).

2. If \( k < n \), then there exist vectors \( \mathbf{u}_{k+1}, \ldots, \mathbf{u}_n \) such that the set \( \{\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_n\} \) is orthogonal (orthonormal).

3. If \( U = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbb{R}^{n \times k} \) is left orthogonal and \( k < n \), then there is an orthogonal matrix \( \tilde{U} \) such that

\[
\tilde{U} = \begin{pmatrix} \hat{U} & U^\perp \end{pmatrix},
\]

where \( \text{range}(U^\perp) = \text{range}(U)^\perp \) and \( U^TU^\perp = 0 \).

The third statement in Lemma 1.9 is just a matrix version of the second. The next lemma gives some simple facts about orthogonal matrices.

Lemma 1.10 Let \( U \in \mathbb{R}^{n \times n} \) and \( V \in \mathbb{R}^{n \times n} \) be orthogonal matrices and let \( Z \in \mathbb{R}^{m \times n} \) be a left orthogonal matrix.

1. The matrices \( U^T \) and \( U^\top \) are orthogonal.

2. The matrix \( UV \) is orthogonal and the matrix \( ZU \) is left orthogonal.

The final lemma gives significance to orthogonal matrices, the two-norm, and the Frobenius norm.

Lemma 1.11 Let \( X \in \mathbb{R}^{m \times n} \), \( Y \in \mathbb{R}^{n \times k} \) and \( y \in \mathbb{R}^n \). Let \( U \in \mathbb{R}^{m \times m} \) and \( V \in \mathbb{R}^{n \times n} \) be orthogonal matrices, and let \( Z \in \mathbb{R}^{m \times n} \) be a left orthogonal matrix. Then

\[
\|Y\|_2 = \|y\|_2, \quad \|Zy\|_2 = \|y\|_2, \quad \|U X V\|_2 = \|X\|_F, \quad \|UXV\|_F = \|X\|_F, \quad \|Z\|_F = \|Y\|_F,
\]

\[
\|U\|_2 = \|V\|_2 = \|Z\|_2 = 1, \quad \|U\|_F = \sqrt{m}, \quad \|V\|_F = \|Z\|_F = \sqrt{n}.
\]

The properties (1.34)–(1.36) are called orthogonal invariance.

Orthogonal invariance leads directly to the singular value decomposition. The singular value decomposition and a template for other orthogonal decompositions in this book are given in the next section.
Bibliography


1.5. Notes

For a more thorough discussion of the results from linear algebra see, for instance, the book by Strang [?, 2003]. The first chapter of Stewart’s book [4, 1973] gives a summary of the important theorems for applications to least squares. The section on probability, §1.3.2, was deliberately nonrigorous. For a thorough introduction read Chung [1, 1974]. Much of the material in §1.3.2 was adapted from Rao [3, 1973].

A thorough treatment of norms is available from a number of books on numerical linear algebra. Householder [2, 1964, Chapter 2] is still an excellent source. A recent treatment of norms with special attention to orthogonally invariant norms is in Stewart and Sun [5, 1990, part II].

21