CHAPTER 1

INTRODUCTION TO OPERATORS, PROBABILITIES AND THE LINEAR MODEL

THIS CHAPTER has a little bit of everything: normed and Hilbert spaces, linear operators, probabilities, including conditional expectations and different modes of convergence, and matrix algebra. Introduction to the OLS method is given along with a discussion of methodological issues, such as the choice of the format of the convergence statement, choice of the conditions sufficient for convergence and the use of L_2-approximability. The exposition presumes that the reader is versed more in the theory of probabilities than in functional analysis.

1.1 LINEAR SPACES

In this book basic notions of functional analysis are used more frequently than in most other econometric books. Here I explain these notions the way I understand them—omitting some formalities and emphasizing the intuition.

1.1.1 Linear Spaces

The Euclidean space $\mathbb{R}^n$ is a good point of departure when introducing linear spaces. An element $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is called a *vector*. Two vectors x, y can be added coordinate by coordinate to obtain a new vector

$$x + y = (x_1 + y_1, \ldots, x_n + y_n). \tag{1.1}$$

A vector x can be multiplied by a number $a \in \mathbb{R}$, giving $ax = (ax_1, \ldots, ax_n)$. By combining these two operations we can form expressions like $ax + by$ or, more generally,

$$a_1 x^{(1)} + \cdots + a_m x^{(m)} \tag{1.2}$$

where $a_1, \ldots, a_n$ are numbers and $x^{(1)}, \ldots, x^{(m)}$ are vectors. Expression (1.2) is called a *linear combination of vectors* $x^{(1)}, \ldots, x^{(m)}$ with coefficients $a_1, \ldots, a_n$. Generally, multiplication of vectors is not defined.

Short-Memory Linear Processes and Econometric Applications. Kairat T. Mynbaev

Here we observe the major difference between $\mathbb{R}$ and $\mathbb{R}^n$. In $\mathbb{R}$ both summation $a + b$ and multiplication ab can be performed. In $\mathbb{R}^n$ we can add two vectors, but to multiply them we use elements of another set – the set of real numbers (or scalars) $\mathbb{R}$.

Generalizing upon this situation we obtain abstract *linear* (or *vector*) *spaces*. The elements x, y of a linear space L are called *vectors*. They can be added to give another vector $x + y$. Summation is defined axiomatically and, in general, there is no coordinate representation of type (1.1) for summation. A vector x can be multiplied by a scalar $a \in \mathbb{R}$. As in $\mathbb{R}^n$, we can form linear combinations [Eq. (1.2)].

The generalization is pretty straightforward, so what's the big deal? You see, in functional analysis complex objects, such as functions and operators, are considered vectors or points in some space. Here is an example. Denote $C[0, 1]$ the set of continuous functions on the segment [0, 1]. The sum of two functions F, $G \in C[0, 1]$ is defined as the function $F + G$ with values $(F + G)(t) = F(t) + G(t)$, $t \in [0, 1]$ [this is an analog of Eq. (1.1)]. Continuity of F, G implies continuity of their sum and of the product aF, for a a scalar, so $C[0, 1]$ is a linear space.

1.1.2 Subspaces of Linear Spaces

A subset L_1 of a linear space L is called its *linear subspace* (or just a subspace, for simplicity) if all linear combinations $ax + by$ of any elements $x, y \in L_1$ belong to L_1. Obviously, the set $\{0\}$ and L itself are subspaces of L, called trivial subspaces. For example, in $\mathbb{R}^n$ the set $L_1 = \{x : c_1x_1 + \cdots + c_nx_n = 0\}$ is a subspace because if $x, y \in L_1$, then $c_1(ax_1 + by_1) + \cdots + c_n(ax_n + by_n) = 0$. Thus, in $\mathbb{R}^3$ the usual straight lines and two-dimensional (2-D) planes containing the origin are subspaces. All intuition we get from our day-to-day experience with the space we live in applies to subspaces. Geometrically, summation $x + y$ is performed by the parallelogram rule. Multiplying x by a number $a \neq 0$ we obtain a vector ax of either the same ($a > 0$) or opposite ($a < 0$) direction. Multiplying x by all real numbers, we obtain a straight line $\{ax : a \in \mathbb{R}\}$ passing through the origin and parallel to x. This is a particular situation in which it may be convenient to call x a point rather than a vector. Then the previous sentence sounds like this: multiplying x by all real numbers we get a straight line passing through the origin and the given point x.

For a given $x_1, \ldots, x_n$ its *linear span* $\mathfrak{M}$ is, by definition, the least linear space of L containing those points. In the case $n = 2$ it can be constructed as follows. Draw a straight line $L_1 = \{ax_1 : a \in \mathbb{R}\}$ through the origin and x_1 and another straight line $L_2 = \{ax_2 : a \in \mathbb{R}\}$ through the origin and x_2. Then form $\mathfrak{M}$ by adding elements of L_1 and L_2 using the parallelogram rule: $\mathfrak{M} = \{x + y : x \in L_1, y \in L_2\}$.

1.1.3 Linear Independence

Vectors $x_1, \ldots, x_n$ are *linearly independent* if the linear combination $c_1x_1 + \cdots + c_nx_n$ can be null only when all coefficients are null.

EXAMPLE 1.1. Denote by $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ (unity in the jth place) the jth *unit vector* in $\mathbb{R}^n$. From the definition of vector operations in $\mathbb{R}^n$ we see that

$c_1e_1 + \cdots + c_ne_n = (c_1, \ldots, c_n)$. Hence, the equation $c_1e_1 + \cdots + c_ne_n = 0$ implies equality of all coefficients to zero and the unit vectors are linearly independent.

If in a linear space L there exist vectors $x_1, \ldots, x_n$ such that

1. $x_1, \ldots, x_n$ are linearly independent and
2. any other vector $x \in L$ is a linear combination of $x_1, \ldots, x_n$,

then L is called *n-dimensional* and the system $\{x_1, \ldots, x_n\}$ is called its *basis*. If, on the other hand, for any natural n, L contains n linearly independent vectors, then L is called *infinite-dimensional*.

EXAMPLE 1.2. The unit vectors in $\mathbb{R}^n$ form a basis because they are linearly independent and for any $x \in \mathbb{R}^n$ we can write $x = (x_1, \ldots, x_n) = x_1e_1 + \cdots + x_ne_n$.

EXAMPLE 1.3. $C[0, 1]$ is infinite-dimensional. Consider monomials $x_j(t) = t^j$, $j = 0, \ldots, n$. By the main theorem of algebra, the equation $c_0x_0(t) + \cdots + c_nx_n(t) = 0$ with nonzero coefficients can have at most n roots. Hence, if $c_0x_0(t) + \cdots + c_nx_n(t)$ is identically zero on [0, 1], the coefficients must be zero, so these monomials are linearly independent.

Functional analysis deals mainly with infinite-dimensional spaces. Together with the desire to do without coordinate representations of vectors this fact has led to the development of very powerful methods.

1.2 NORMED SPACES

1.2.1 Normed Spaces

The Pythagorean theorem gives rise to the Euclidean distance

$$\operatorname{dist}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{1.3}$$

between points $x, y \in \mathbb{R}^n$. In an abstract situation, we can first axiomatically define the distance $\operatorname{dist}(x, 0)$ from x to the origin and then the distance between any two points will be $\operatorname{dist}(x, y) = \operatorname{dist}(x - y, 0)$ (this looks like tautology, but programmers use such definitions all the time). $\operatorname{dist}(x, 0)$ is denoted $\|x\|$ and is called a norm.

Let X be a linear space. A real-valued function $\|\cdot\|$ defined on X is called a *norm* if

1. $\|x\| \geq 0$ (*nonnegativity*),
2. $\|ax\| = |a|\|x\|$ for all numbers a and vectors x (*homogeneity*),
3. $\|x + y\| \leq \|x\| + \|y\|$ (*triangle inequality*) and
4. $\|x\| = 0$ implies $x = 0$ (*nondegeneracy*).

By homogeneity the norm of the null vector is zero:

$$\left\| \underset{\text{(vector)}}{0} \right\| = \left\| \underset{\text{(number)}}{0} \cdot \underset{\text{(vector)}}{0} \right\| = |0|\|0\| = 0.$$

Nondegeneracy makes sure that the null vector is the only vector whose norm is zero. If we omit the nondegeneracy requirement, the result is the definition of a *seminorm*.

Distance measurement is another context in which points and vectors can be used interchangeably. $\|x\|$ is a length of the vector x and a distance from point x to the origin.

In this book, the way norms are used for bounding various quantities is clear from the next two definitions. Let $\{X_i\}$ be a nested sequence of normed spaces, $X_1 \subseteq X_2 \subseteq \dots$. Take one element from each of these spaces, $x_i \in X_i$. We say that $\{x_i\}$ is a *bounded* sequence if $\sup_i \|x_i\|_{X_i} < \infty$ and *vanishing* if $\|x_i\|_{X_i} \to 0$.

1.2.2 Convergence in Normed Spaces

A linear space X provided with a norm $\|\cdot\|$ is denoted $(X, \|\cdot\|)$. This is often simplified to X. We say that a sequence $\{x_n\}$ *converges* to x if $\|x_n - x\| \to 0$. In this case we write $\lim x_n = x$.

Lemma

(i) Vector operations are continuous: if $\lim x_n = x$, $\lim y_n = y$ *and* $\lim a_n = a$, *then* $\lim a_n x_n = ax$, $\lim(x_n + y_n) = \lim x_n + \lim y_n$.

(ii) If $\lim x_n = x$, *then* $\lim\|x_n\| = \|x\|$ *(a norm is continuous in the topology it induces).*

Proof.

(i) Applying the triangle inequality and homogeneity,

$$\begin{aligned}\|a_n x_n - ax\| &\le \|(a_n - a)x\| + \|a_n(x_n - x)\| \\ &= |a_n - a|\|x\| + \|a_n\|\|x_n - x\| \to 0.\end{aligned}$$

Here we remember that convergence of the sequence $\{a_n\}$ implies its boundedness: $\sup|a_n| < \infty$.

(ii) Let us prove that

$$\big|\|x\| - \|y\|\big| \le \|x - y\|. \tag{1.4}$$

The proof is modeled on a similar result for absolute values. By the triangle inequality, $\|x\| \le \|x - y\| + \|y\|$ and $\|x\| - \|y\| \le \|x - y\|$. Changing the places of x and y and using homogeneity we get $\|y\| - \|x\| \le \|y - x\| = \|x - y\|$. The latter two inequalities imply Eq. (1.4).

Equation (1.4) yields continuity of the norm: $|\|x_n\| - \|x\|| \le \|x_n - x\| \to 0$. ■

We say that $\{x_n\}$ is a *Cauchy sequence* if $\lim_{n,m\to\infty}(x_n - x_m) = 0$. If $\{x_n\}$ converges to x, then it is a Cauchy sequence: $\|x_n - x_m\| \le \|x_n - x\| + \|x - x_m\| \to 0$. If the converse is true (that is, every Cauchy sequence converges), then the space is called *complete*. All normed spaces considered in this book are complete, which ensures the existence of limits of Cauchy sequences.

1.2.3 Spaces l_p

A norm more general than (1.3) is obtained by replacing the index 2 by an arbitrary number $p \in [1, \infty)$. In other words, in $\mathbb{R}^n$ the function

$$\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p} \tag{1.5}$$

satisfies all axioms of a norm. For $p = \infty$, definition (1.5) is completed with

$$\|x\|_\infty = \sup_i |x_i| \tag{1.6}$$

because $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$. $\mathbb{R}^n$ provided with the norm $\|\cdot\|_p$ is denoted $\mathbb{R}^n_p$ $(1 \le p \le \infty)$.

The most immediate generalization of $\mathbb{R}^n_p$ is the *space* l_p of infinite sequences of numbers $x = (x_1, x_2, \dots)$ that have a finite norm $\|x\|_p$ [defined by Eqs. (1.5) or (1.6), where i runs over the set of naturals $\mathbb{N}$]. More generally, the set of indices $I = \{i\}$ in Eq. (1.5) or Eq. (1.6) may depend on the context. In addition to $\mathbb{R}^n_p$ we use $\mathbb{M}_p$ (the set of matrices of all sizes).

The jth *unit vector* in l_p is an infinite sequence $e_j = (0, \dots, 0, 1, 0, \dots)$ with unity in the jth place and 0 in all others. It is immediate that the unit vectors are linearly independent and l_p is infinite-dimensional.

1.2.4 Inequalities in l_p

The triangle inequality in l_p $\|x + y\|_p \le \|x\|_p + \|y\|_p$ is called the *Minkowski inequality*. Its proof can be found in many texts, which is not true with respect to another, less known, property that is natural to call *monotonicity* of l_p norms:

$$\|x\|_p \le \|x\|_q \quad \text{for all } 1 \le q \le p \le \infty. \tag{1.7}$$

If $x = 0$, there is nothing to prove. If $x \neq 0$, the general case can be reduced to the case $\|x\|_q = 1$ by considering the normalized vector $x/\|x\|_q$. $\|x\|_q = 1$ implies $|x_i| \le 1$ for all i. Hence, if $p < \infty$, we have

$$\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p} \le \left(\sum_i |x_i|^q\right)^{1/p} = \left(\sum_i |x_i|^q\right)^{1/q} = \|x\|_q .$$

If $p = \infty$, the inequality $\sup_i |x_i| \le \|x\|_q$ is obvious.

In l_p there is no general inequality opposite to Eq. (1.7). In $\mathbb{R}_p^n$ there is one. For example, in the case $n = 2$ we can write

$$\max\{|x_1|, |x_2|\} \le (|x_1|^p + |x_2|^p)^{1/p} \le 2^{1/p}\max\{|x_1|, |x_2|\}.$$

All such inequalities are easy to remember under the general heading of equivalent norms. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ defined on the same linear space X are called *equivalent* if there exist constants $0 < c_1 \le c_2 < \infty$ such that $c_1\|x\|_1 \le \|x\|_2 \le c_2\|x\|_1$ for all x.

Theorem. (Trenogin, 1980, Section 3.3) *In a finite-dimensional space any two norms are equivalent.*

1.3 LINEAR OPERATORS

1.3.1 Linear Operators

A linear operator is a generalization of the mapping $A : \mathbb{R}^m \to \mathbb{R}^n$ induced by an $n \times m$ matrix A according to $y = Ax$. Let L_1, L_2 be linear spaces. A mapping $A : L_1 \to L_2$ is called a *linear operator* if

$$A(ax + by) = aAx + bAy \tag{1.8}$$

for all vectors $x, y \in L_1$ and numbers a, b.

A linear operator is a function in the first place, and the general definition of an *image* applies to it:

$$\mathrm{Im}(A) = \{Ax : x \in L_1\} \subseteq L_2.$$

However, because of the linearity of A the image $\mathrm{Im}(A)$ is a linear subspace of L_2. Indeed, if we take two elements y_1, y_2 of the image, then there exist $x_1, x_2 \in L_1$ such that $Ax_i = y_i$. Hence, a linear combination

$$a_1y_1 + a_2y_2 = a_1Ax_1 + a_2Ax_2 = A(ax_1 + bx_2)$$

belongs to the image. With a linear operator A we can associate another linear subspace

$$N(A) = \{x \in L_1 : Ax = 0\} \subseteq L_1,$$

called a *null space* of A. Its linearity easily follows from that of A: if x, y belong to the null space of A, then their linear combination belongs to it too: $A(ax + by) = aAx + bAy = 0$.

The set of linear operators acting from L_1 to L_2 can be considered a linear space. A *linear combination* of operators $aA + bB$ of operators A, B is an operator defined by $(aA + bB)x = aAx + bBx$. It is easy to check linearity of $aA + bB$.

If A is a linear operator from L_1 to L_2 and B is a linear operator from L_2 to L_3, then we can also define a *product* of operators BA by $(BA)x = B(Ax)$. Applying Eq. (1.8) twice we see that BA is linear:

$$(BA)(ax + by) = B(aAx + bAy) = a(BA)x + b(BA)y.$$

1.3.2 Bounded Linear Operators

Let X_1, X_2 be normed spaces and let $A : X_1 \to X_2$ be a linear operator. We can relate $\|Ax\|_2$ to $\|x\|_1$ by composing the ratio $\|Ax\|_2/\|x\|_1$ if $x \neq 0$. A is called a *bounded* operator if all such ratios are uniformly bounded, and the *norm* of an operator A is defined as the supremum of those ratios:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_1}. \tag{1.9}$$

An immediate consequence of this definition is the bound $\|Ax\|_2 \leq \|A\|\|x\|_1$ for all $x \in X_1$, from which we see that the images Ax of elements of the unit ball $b_1 = \{x \in X_1 : \|x\|_1 \leq 1\}$ are uniformly bounded:

$$\|Ax\|_2 \leq \|A\| \quad \text{for all } x \in b_1. \tag{1.10}$$

To save a word, a bounded linear operator is called simply a bounded operator. Let $B(X_1, X_2)$ denote the set of bounded operators acting from X_1 to X_2.

Lemma. *$B(X_1, X_2)$ with the norm (1.9) is a normed space.*

Proof. We check the axioms from Section 1.2.1 one by one.

1. Nonnegativity is obvious.
2. Homogeneity of Eq. (1.9) follows from that of $\|\cdot\|_2$.
3. The inequality $\|(A + B)x\|_2 \leq \|Ax\|_2 + \|Bx\|_2$ implies

$$\|A + B\| = \sup_{x \neq 0} \frac{\|(A + B)x\|_2}{\|x\|_1} \leq \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_1} + \sup_{x \neq 0} \frac{\|Bx\|_2}{\|x\|_1} = \|A\| + \|B\|.$$

4. If $\|A\| = 0$, then $\|Ax\|_2 = 0$ for all x and, consequently, $A = 0$. ■

1.3.3 Isomorphism

Let X_1, X_2 be normed spaces. A linear operator $I : X_1 \to X_2$ is called an *isomorphism* if

1. $\|Ix\|_2 = \|x\|_1$ for all $x \in X_1$ (preservation of norms) and
2. $IX_1 = X_2$ (I is a surjection).

Item 1 implies that $\|I\| = 1$ and I is one-to-one (if $Ix_1 = Ix_2$, then $\|x_1 - x_2\|_1 = \|I(x_1 - x_2)\|_2 = 0$ and $x_1 = x_2$). Hence, the inverse of I exists and is an isomorphism from X_2 to X_1.

Normed spaces X_1 and X_2 are called *isomorphic* spaces if there exists an isomorphism $I : X_1 \to X_2$. Vector operations in X_1 are mirrored by those in X_2 and the norms are the same, so as normed spaces X_1 and X_2 are indistinguishable. However, a given operator in one of them may be easier to analyze than its isomorphic image in the other, because of special features. Let A be a bounded operator in X_1. It is easy to see that $\tilde{A} = IAI^{-1}$ is a linear operator in X_2. Moreover, the norms are preserved under this mapping:

$$\|\tilde{A}\| = \sup_{x \neq 0} \frac{\|IAI^{-1}x\|_2}{\|x\|_2} = \sup_{y \neq 0} \frac{\|IAy\|_2}{\|Iy\|_2} = \sup_{y \neq 0} \frac{\|Ay\|_1}{\|y\|_1} = \|A\|.$$

1.3.4 Convergence of Operators

Let $A, A_1, A_2, \ldots$ be bounded operators from a normed space X_1 to a normed space X_2. The sequence $\{A_n\}$ *converges to A uniformly* if $\|A_n - A\| \to 0$, where the norm is as defined in Eq. (1.9). This is convergence in a normed space $B(X_1, X_2)$. The word 'uniform' is pertinent because, as we can see from Eq. (1.10), when $\|A_n - A\| \to 0$, we also have the convergence $\|A_n x - Ax\|_2 \to 0$ uniformly in the unit ball b_1.

The sequence $\{A_n\}$ is said to *converge to A strongly*, or pointwise, if for each $x \in X_1$ we have $\|A_n x - Ax\|_2 \to 0$. Of course, uniform convergence implies strong convergence.

1.3.5 Projectors

Projectors are used (or implicitly present) in econometrics so often that it would be a sin to bypass them.

Let X be a normed space and let $P : X \to X$ be a bounded operator. P is called a *projector* if

$$P^2 = P. \tag{1.11}$$

Suppose y is a projection of x, $y = Px$. Then P doesn't change y: $Py = P^2x = Px = y$. This property is the key to the intuition behind projectors.

Consider on the plane two coordinate axes, X and Y, intersecting at a positive, not necessarily straight, angle. Projection of points on the plane onto the axis X parallel to the axis Y has the following geometrical properties:

1. The projection of the whole plane is X.
2. Points on X stay the same.
3. Points on Y are projected to the origin.
4. Any vector on the plane is uniquely represented as a sum of two vectors, one from X and another from Y.

All these properties can be deduced from linearity of P and Eq. (1.11).

Lemma. *Let P be a projector and denote $Q = I - P$, where I is the identity operator in X. Then*

(i) *Q is also a projector.*

(ii) $\mathrm{Im}(P)$ *coincides with the set of fixed points of* P: $\mathrm{Im}(P) = \{x : x = Px\}$.

(iii) $\mathrm{Im}(Q) = N(P)$, $\mathrm{Im}(P) = N(Q)$.

(iv) *Any $x \in X$ can be uniquely represented as $x = y + z$ with $y \in \mathrm{Im}(P)$, $z \in \mathrm{Im}(Q)$.*

Proof.

(i) $Q^2 = (I - P)^2 = I^2 - 2P + P^2 = I - P = Q$.

(ii) If $x \in \mathrm{Im}(P)$, then $x = Py$ for some $y \in X$ and $Px = P^2y = Py = x$, so that x is a fixed point of P. Conversely, if x is a fixed point of P, then $x = Px \in \mathrm{Im}(P)$.

(iii) The equation $Px = 0$ is equivalent to $Qx = (I - P)x = x$, and the equation $\mathrm{Im}(Q) = N(P)$ follows. $\mathrm{Im}(P) = N(Q)$ is obtained similarly.

(iv) The desired representation is obtained by writing $x = Px + (I - P)x = y + z$, where $y = Px \in \mathrm{Im}(P)$ and $z = (I - P)x = Qy \in \mathrm{Im}(Q)$. If $x = y_1 + z_1$ is another representation, then, subtracting one from another, we get $y - y_1 = -(z - z_1)$. Hence, $P(y - y_1) = -P(z - z_1)$. Here the right-hand side is null because $z, z_1 \in \mathrm{Im}(Q) = N(P)$. The left-hand side is $y - y_1$ because both y and y_1 are fixed points of P. Thus, $y = y_1$ and $z = z_1$.

■

1.4 HILBERT SPACES

1.4.1 Scalar Products

A Hilbert space is another infinite-dimensional generalization of $\mathbb{R}^n$. Everything starts with noticing how useful a *scalar product*

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \tag{1.12}$$

of two vectors $x, y \in \mathbb{R}^n$ is. In terms of it we can define the Euclidean norm, in $\mathbb{R}^n$:

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \langle x, x \rangle^{1/2} . \tag{1.13}$$

Most importantly, we can find the cosine of the angle between x, y by the formula

$$\cos(\widehat{x, y}) = \frac{\langle x, y \rangle}{\|x\|_2 \, \|y\|_2} . \tag{1.14}$$

To do without the coordinate representation, we observe algebraic properties of this scalar product. First of all, it is a *bilinear form*: it is linear with respect to one argument when the other is fixed:

$$\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle, \langle z, ax + by\rangle = a\langle z, x\rangle + b\langle z, y\rangle$$

for all vectors x, y, z and numbers a, b. Further, we notice that $\langle x, x\rangle$ is always nonnegative and $\langle x, x\rangle = 0$ is true only when $x = 0$.

Thus, on the abstract level, we start with the assumption that H is a linear space and $\langle x, y\rangle$ is a real function of arguments $x, y \in H$ having properties:

1. $\langle x, y\rangle$ is a bilinear form,

2. $\langle x, x\rangle \geq 0$ for all $x \in H$ and

3. $\langle x, x\rangle = 0$ implies $x = 0$.

4. $\langle x, y\rangle = \langle y, x\rangle$ for all x, y.

Such a function is called a *scalar product*. Put

$$\|x\| = \langle x, x\rangle^{1/2}. \tag{1.15}$$

Lemma. (*Cauchy–Schwarz inequality*) $|\langle x, y\rangle| \leq \|x\|\|y\|$.

Proof. The function $f(t) = \langle x + ty, x + ty\rangle$ of a real argument t is nonnegative by item 2. Using items 1 and 4 we see that it is a quadratic function:

$$f(t) = \langle x, x + ty\rangle + t\langle y, x + ty\rangle = \langle x, x\rangle + 2t\langle x, y\rangle + t^2\langle y, y\rangle.$$

Its nonnegativity implies that its discriminant $\langle x, y\rangle^2 - \langle x, x\rangle\langle y, y\rangle$ is nonpositive. ■

1.4.2 Continuity of Scalar Products

Notation (1.15) is justified by the following lemma.

Lemma

(i) Eq. (1.15) defines a norm on H and the associated convergence concept: $x_n \to x$ in H if $\|x_n - x\| \to 0$.

(ii) The scalar product is continuous: if $x_n \to x$, $y_n \to y$, then $\langle x_n, y_n\rangle \to \langle x, y\rangle$.

Proof.

(i) By the Cauchy–Schwarz inequality

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y\rangle = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2,\end{aligned}$$

which proves the triangle inequality in Section 1.2.1 (3). The other properties of a norm (nonnegativity, homogeneity and nondegeneracy) easily follow from the scalar product axioms.

(ii) Convergence $x_n \to x$ implies boundedness of the norms $\|x_n\|$. Therefore, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|\langle x_n, y_n\rangle - \langle x, y\rangle\| &\le \|\langle x_n, y_n - y\rangle\| + \|\langle x_n - x, y\rangle\| \\ &\le \|x_n\|\|y_n - y\| + \|x_n - x\|\|y\|.\end{aligned}$$

■

A linear space H that is endowed with a scalar product and is complete in the norm generated by that scalar product is called a *Hilbert space*.

1.4.3 Discrete Hölder's Inequality

An interesting generalization of the Cauchy–Schwarz inequality is in terms of the spaces l_p from Section 1.2.3. Let p be a number from $[1, \infty)$ or the symbol ∞. Its *conjugate* q is defined from $1/p + 1/q = 1$. Explicitly,

$$q = \begin{cases} p/(p-1) \in (1, \infty), & 1 < p < \infty; \\ 1, & p = \infty; \\ \infty, & p = 1. \end{cases}$$

Hölder's inequality states that

$$\sum_{i=1}^{\infty} |x_i y_i| \le \|x\|_p \, \|y\|_q \,. \tag{1.16}$$

A way to understand it is by considering the bilinear form $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$. It is defined on the Cartesian product $l_2 \times l_2$ and is continuous on it by Lemma 1.4.2 Hölder's inequality allows us to take arguments from different spaces: $\langle x, y\rangle$ is defined on $l_p \times l_q$ and is continuous on this product.

1.4.4 Symmetric Operators

Let A be a bounded operator in a Hilbert space H. Its *adjoint* is defined as the operator A^* that satisfies

$$\langle Ax, y\rangle = \langle x, A^* y\rangle \quad \text{for all } x, y \in H.$$

This definition arises from the property of the transpose matrix A',

$$\sum_{i=1}^{n} (Ax)_i y_i = \sum_{i=1}^{n} x_i (A' y)_i.$$

Existence of A^* is proved using the so-called Riesz theorem. We do not need the general proof of existence because in all the cases we need, the adjoint will be constructed explicitly. Boundedness of A^* will also be proved directly.

A is called *symmetric* if $A = A^*$. Symmetric operators stand out by having properties closest to those of real numbers.

1.4.5 Orthoprojectors

Cosines of angles between vectors from H can be defined using Eq. (1.14). We don't need this definition, but we do need its special case: vectors $x, y \in H$ are called *orthogonal* if $\langle x, y\rangle = 0$. For orthogonal vectors we have the Pythagorean theorem:

$$\|x + y\|^2 = \langle x + y, x + y\rangle = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 = \|x\|^2 + \|y\|^2.$$

Two subspaces $X, Y \subseteq H$ are called *orthogonal* if every element of X is orthogonal to every element of Y.

If a projector P in H ($P^2 = P$) is symmetric, $P = P^*$, then it is called an *orthoprojector*. In the situation described in Section 1.3.5, when points on the plane are projected onto one axis parallel to another, orthoprojectors correspond to the case when the axes are orthogonal.

Lemma. *Let P be an orthoprojector and let $Q = I - P$. Then*

(i) $\mathrm{Im}(P)$ *is orthogonal to* $\mathrm{Im}(Q)$.

(ii) For any $x \in H$, $\|Px\|$ is the distance from x to $\mathrm{Im}(Q)$.

Proof.

(i) Let $x \in \mathrm{Im}(P)$ and $y \in \mathrm{Im}(Q)$. By Lemma 1.3.5(ii), $x = Px$, $y = Qy$. Hence, x and y are orthogonal:

$$\langle x, y\rangle = \langle Px, Qy\rangle = \langle x, P(I - P)y\rangle = \langle x, (P - P^2)y\rangle = 0.$$

(ii) For an arbitrary element $x \in H$ and a set $A \subseteq H$ the *distance* from x to A is defined by

$$\mathrm{dist}(x, A) = \inf_{y \in A} \|x - y\|.$$

Take any $y \in \mathrm{Im}(Q)$. In the equation

$$x - y = Px + Qx - Qy = Px + Q(x - y)$$

the two terms at the right are orthogonal, so by the Pythagorean theorem

$$\|x-y\|^2 = \|Px\|^2 + \|Q(x-y)\|^2 \geq \|Px\|^2,$$

which implies the lower bound for the distance $\text{dist}(x, \text{Im}(Q)) \geq \|Px\|$. This lower bound is attained on $y = Qx \in \text{Im}(Q)$: $\|x-y\| = \|Px + Qx - Qx\| = \|Px\|$. Hence, $\text{dist}(x, \text{Im}(Q)) = \|Px\|$. ■

1.5 L_p SPACES

1.5.1 σ-Fields

Let Ω be some set and let $\mathcal{F}$ be a nonempty family of its subsets. $\mathcal{F}$ is called a *σ-field* if

1. unions, intersections, differences and complements of any two elements of $\mathcal{F}$ belong to $\mathcal{F}$,
2. the union of any sequence $\{A_n : n = 1, 2, \ldots\}$ of elements of $\mathcal{F}$ belongs to $\mathcal{F}$ and
3. Ω belongs to $\mathcal{F}$.

This definition contains sufficiently many requirements to serve most purposes of analysis. In probabilities, σ-fields play the role of information sets. The precise meaning of this sentence at times can be pretty complex. The following existence statement is used very often.

Lemma. *For any system S of subsets of Ω there exists a σ-field $\mathcal{F}$ that contains S and is contained in any other σ-field containing S.*

Proof. The set of σ-fields containing S is not empty. For example, the set of all subsets of Ω is a σ-field and contains S. Let σ be the intersection of all σ-fields containing S. It obviously satisfies 1–3 and hence is the σ-field we are looking for. ■

The σ-field whose existence is affirmed in this lemma is called the *least σ-field generated by S* and denoted $\sigma(S)$.

1.5.2 Borel σ-field in $\mathbb{R}^n$

A ball in $\mathbb{R}^n$ centered at $x \in \mathbb{R}^n$ of radius $\varepsilon > 0$,

$$b_\varepsilon(x) = \{y \in \mathbb{R}^n : \|x-y\|_2 < \varepsilon\},$$

is called an *ε-neighborhood* of x. We say that the set $A \subseteq \mathbb{R}^n$ is an *open* set if each point x belongs to A with its neighborhood $b_\varepsilon(x)$ (where ε depends on x). The *Borel σ-field* $\mathcal{B}_n$ in $\mathbb{R}^n$ is defined as the smallest σ-field that contains all open subsets of $\mathbb{R}^n$. It exists by Lemma 1.5.1. In more general situations, when open subsets of Ω are not defined, σ-fields of Ω are introduced axiomatically.

1.5.3 σ-Additive Measures

A pair $(\Omega, \mathcal{F})$, where Ω is some set and $\mathcal{F}$ is a σ-field of its subsets, is called a *measurable space*. A set function μ defined on elements of $\mathcal{F}$ with values in the extended half-line $[0, \infty]$ is called a *σ-additive measure* if for any disjoint sets $A_1, A_2, \ldots \in \mathcal{F}$ one has

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

EXAMPLE 1.4. On a plane, for any rectangle A define $\mu(A)$ to be its area. The extension procedure from the measure theory then leads to the *Lebesgue measure* μ with $\Omega = \mathbb{R}^2$and $\mathcal{F} = \mathcal{B}_2$ (μ is defined on all Borel subsets of $\mathbb{R}^2$).

A *probabilistic measure* is a σ-additive measure that satisfies an additional requirement $\mu(\Omega) = 1$. In this case, following common practice, we write P instead of μ. Thus, a *probability space* (sometimes also called a *sample space*) is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set, $\mathcal{F}$ is a σ-field of its subsets and P is a σ-additive measure on $\mathcal{F}$ such that $P(\Omega) = 1$.

EXAMPLE 1.5. On a plane, take the square $[0, 1]^2$ as Ω and let P be the Lebesgue measure. Then $\mathcal{F}$ will be the set of Borel subsets of the square.

1.5.4 Measurable Functions

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is called *measurable* if $f^{-1}(A) \in \mathcal{F}_1$ for any $A \in \mathcal{F}_2$. More precisely, it is said to be *$(\mathcal{F}_1, \mathcal{F}_2)$-measurable*. In particular, when $(\Omega_1, \mathcal{F}_1) = (\mathbb{R}^n, \mathcal{B}_n)$ and $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}^m, \mathcal{B}_m)$, this definition gives the definition of *Borel-measurability*. Most of the time we deal with real-valued functions, when $\Omega_2 = \mathbb{R}$ and $\mathcal{F}_2 = \mathcal{B}_1$ is the Borel σ-field. In this case we simply say that f is $\mathcal{F}_1$-measurable. All analysis operations in the finite-dimensional case preserve measurability. The next theorem is often used implicitly.

Theorem. (Kolmogorov and Fomin, 1989, Chapter 5, Section 4)

1. *Let X, Y and Z be arbitrary sets with systems of subsets σ_X, σ_Y and σ_Z, respectively. Suppose the function $f : X \to Y$ is (σ_X, σ_Y)-measurable and $g : Y \to Z$ is (σ_Y, σ_Z)-measurable. Then the composition $z(x) = g(f(x))$ is (σ_X, σ_Z)-measurable.*
2. *Let f and g be defined on the same measurable space $(\Omega, \mathcal{F})$. Then a linear combination $af + bg$ and product fg are measurable. If g does not vanish, then the ratio f/g is also measurable.*

1.5.5 L_p Spaces

Let $(\Omega, \mathcal{F}, \mu)$ be any space with a σ-additive measure μ and let $1 \le p < \infty$. The set of measurable functions $f : \Omega \to \mathbb{R}$ provided with the norm

$$\|f\|_p = \left(\int_\Omega |f(x)|^p d\mu \right)^{1/p}, \; 1 \le p < \infty,$$

is denoted $L_p = L_p(\Omega)$. In the case $p = \infty$ this definition is completed with

$$\|f\|_\infty = \operatorname{ess\,sup}_{x\in\Omega} |f(x)| = \inf_{\mu(A)=0} \sup_{x\in\Omega\setminus A} |f(x)|.$$

The term in the middle is, by definition, the quantity at the right and is called *essential supremum*. These definitions mean that values taken by functions on sets of measure zero don't matter. An equality $f(t) = 0$ is accompanied by the caveat "almost everywhere" (a.e.) or "almost surely" (a.s.) in the probabilistic setup, meaning that there is a set of measure zero outside which $f(t) = 0$.

1.5.6 Inequalities in L_p

Apparently, L_p spaces should have a lot in common with l_p spaces. The triangle inequality in L_p $\|F + G\|_p \le \|F\|_p + \|G\|_p$ is called a *Minkowski inequality*.

Hölder's inequality looks like this:

$$\int_\Omega |f(x)g(x)| \, d\mu \le \|f\|_p \, \|g\|_q \,,$$

where q is the conjugate of p. When $\mu(\Omega) < \infty$, we can use this inequality to show that for $1 \le p_1 < p_2 \le \infty$, L_{p_2} is a subset of L_{p_1}:

$$\begin{aligned} \int_\Omega |f(x)|^{p_1} \, d\mu &\le \left(\int_\Omega |f(x)|^{p_1 p_2/p_1} \, d\mu \right)^{p_1/p_2} \left(\int_\Omega d\mu \right)^{1-p_1/p_2} \\ &= \|f\|_{p_2}^{p_1} \, [\mu(\Omega)]^{1-p_1/p_2} \,. \end{aligned}$$

In particular, when $(\Omega, \mathcal{F}, P)$ is a probability space, we get

$$\|f\|_{p_1} \le \|f\|_{p_2} \quad \text{if } 1 \le p_1 < p_2 \le \infty.$$

This is the opposite of the monotonicity relation (1.7).

1.5.7 Covariance as a Scalar Product

Real-valued measurable functions on a probability space $(\Omega, \mathcal{F}, P)$ are called *random variables*. Let X, Y be integrable random variables (integrability is necessary for their

means to exist). Denote $x = X - EX,\ \ y = Y - EY$. Then the *covariance* of X, Y is defined by

$$\operatorname{cov}(X, Y) = E(X - EX)(Y - EY) = Exy, \tag{1.17}$$

the *standard deviation* of X is, by definition,

$$\sigma(X) = \sqrt{\operatorname{cov}(X, X)} = \sqrt{Ex^2} = \sigma(x) \tag{1.18}$$

and the definition of *correlation* of X, Y is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma(X)\sigma(Y)} = \frac{Exy}{\sigma(x)\sigma(y)}. \tag{1.19}$$

Comparison of Eqs. (1.17), (1.18) and (1.19) with Eqs. (1.12), (1.13) and (1.14) from Section 1.4.1 makes clear that definitions (1.17), (1.18) and (1.19) originate in Euclidean geometry. In particular, $\sigma(X)$ is the distance from X to EX and from x to 0. While this idea has been very fruitful, I often find it more useful to estimate $(EX^2)^{1/2}$, which is the distance from X to 0.

1.5.8 Dense Sets in L_p, $p < \infty$

Let us fix some space with measure $(\Omega, \mathcal{F}, \mu)$. A set $M \subseteq L_p$ is said to be *dense* in L_p if any function $f \in L_p$ can be approximated by some sequence $\{f_n\} \subseteq M$: $\| f_n - f \|_p \to 0$. By 1_A we denote the *indicator* of a set A:

$$1_A = \begin{cases} 1, & x \in A; \\ 0, & x \notin A. \end{cases}$$

A finite linear combination $\sum_i c_i 1_{A_i}$ of indicators of measurable sets $A_i \in \mathcal{F}$ is called a *step function*. We say that the measure μ is a *σ-finite* measure if Ω can be represented as a union of disjoint sets Ω_i,

$$\Omega = \bigcup_i \Omega_i, \tag{1.20}$$

of finite measure $\mu(\Omega_i) < \infty$. For example, $\mathbb{R}^n$ is a union of rectangles of finite Lebesgue measure.

Lemma. *If $p < \infty$ and the measure μ is σ-finite, then the set M of step functions is dense in L_p.*

Proof.

Step 1. Let $f \in L_p$. First we show that the general case of Ω of infinite measure can be reduced to the case $\mu(\Omega) < \infty$. Since for the sets from Eq. (1.20) we have

$$\int_\Omega |f(x)|^p d\mu = \sum_l \int_{\Omega_l} |f(x)|^p d\mu < \infty,$$

for any $\varepsilon > 0$ there exists $L > 0$ such that $\sum_{l>L}\int_{\Omega_l}|f(x)|^p d\mu < \varepsilon$. Denote $\widetilde{\Omega} = \bigcup_{l=1}^{L}\Omega_l$. Whatever step function $\tilde{f}_\varepsilon$ we find to approximate f in $L_p(\widetilde{\Omega})$ in the sense that

$$\int_{\widetilde{\Omega}} |f(x) - \tilde{f}_\varepsilon(x)|^p d\mu < \varepsilon,$$

we can extend it by zero,

$$f_\varepsilon(x) = \begin{cases} \tilde{f}_\varepsilon(x), & x \in \widetilde{\Omega}; \\ 0, & x \in \Omega \backslash \widetilde{\Omega}, \end{cases}$$

to obtain an approximation to f in $L_p(\Omega)$:

$$\int_\Omega |f - f_\varepsilon|^p d\mu = \int_{\widetilde{\Omega}} |f - \tilde{f}_\varepsilon|^p d\mu + \int_{\Omega \backslash \widetilde{\Omega}} |f|^p d\mu < 2\varepsilon.$$

f_ε will be a step function and $\mu(\widetilde{\Omega}) < \infty$.

Step 2. Now we show that f can be considered bounded. From

$$\int_\Omega |f|^p d\mu = \sum_{l=1}^{\infty} \int_{\{l-1 \le |f(x)| < l\}} |f(x)|^p d\mu < \infty$$

we see that for any $\varepsilon > 0$, L can be chosen so that $\int_{\{L \le |f(x)|\}} |f(x)|^p d\mu < \varepsilon$. Then f is bounded on $\widetilde{\Omega} = |f(x)| \le L$ and, as above, we see that approximating f by a simple function on $\widetilde{\Omega}$ is enough.

Step 3. Now we can assume that $\mu(\Omega) < \infty$ and $|f(x)| \le L$. Take a large k and partition $[-L, L]$ into k nonoverlapping (closed, semiclosed or open, it does not matter) intervals $\Delta_1, \ldots, \Delta_k$ of length $2L/k$. Let $l_1, \ldots, l_k$ denote the left ends of those intervals and put $A_m = f^{-1}(\Delta_m)$, $m = 1, \ldots, k$. Then the sets A_m are disjoint,

$$|l_m - f(x)| \le \frac{2L}{k} \quad \text{for } x \in A_m \text{ and } \Omega = \bigcup_{m=1}^{k} A_m.$$

This implies

$$\int_\Omega \left| \sum_m l_m 1_{A_m}(x) - f(x) \right|^p d\mu = \sum_m \int_{A_m} |l_m - f(x)|^p d\mu$$
$$\le \left(\frac{2L}{k}\right)^p \mu(\Omega) \to 0, \, k \to \infty.$$

■

1.6 CONDITIONING ON σ-FIELDS

1.6.1 Absolute Continuity of Measures

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let f be an integrable function on Ω. Then the *σ-additivity* of Lebesgue integrals (Kolmogorov and Fomin, 1989, Chapter 5, Section 5.4)

$$\int_{\bigcup_{m=1}^{\infty} A_m} f(x)dP = \sum_{m=1}^{\infty} \int_{A_m} f(x)dP \text{ for disjoint measurable } A_m$$

means that

$$\nu(A) = \int_A f(x)dP \tag{1.21}$$

is a σ-additive set function with the same domain $\mathcal{F}$ as that of P. Another property of Lebesgue integrals (see the same source) states that ν is *absolutely continuous with respect to* P: $\nu(A) = 0$ for each measurable set A for which $P(A) = 0$. The Radon–Nikodym theorem affirms that the opposite is true: σ-additivity and absolute continuity are sufficient for a set function to be of form (1.21).

Theorem. (*Radon–Nikodym*) (Kolmogorov and Fomin, 1989, Chapter 6, Section 5.3) *If $(\Omega, \mathcal{F}, P)$ is a probability space and ν is a set function defined on $\mathcal{F}$ that is σ-additive and absolutely continuous with respect to P, then there exists an integrable function f on Ω such that Eq. (1.21) is true. If g is another such function, then $f = g$ a.s.*

1.6.2 Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space, X an integrable random variable and $\mathcal{G}$ a σ-field contained in $\mathcal{F}$. The *conditional expectation* $E(X|\mathcal{G})$ is defined as a $\mathcal{G}$-measurable function Y such that

$$\int_A YdP = \int_A XdP \quad \text{for all } A \in \mathcal{G}. \tag{1.22}$$

EXAMPLE 1.6. Let $\mathcal{G} = \{\emptyset, \Omega\}$ be the smallest σ-field. In the case $A = \emptyset$ (or, more generally, $P(A) = 0$) Eq. (1.22) turns into an equality of two zeros. In the case $A = \Omega$ we see that the means of Y and X should be the same. Since a constant is the only $\mathcal{G}$-measurable random variable, it follows that $E(X|\mathcal{G}) = EX$.

EXAMPLE 1.7. Let $\mathcal{G} = \mathcal{F}$ be the largest σ-field contained in $\mathcal{F}$. Since X is $\mathcal{G}$-measurable, $Y = X$ satisfies Eq. (1.22). Hence, $E(X|\mathcal{G}) = X$ by a.s. uniqueness.

$Y = X$ is an incorrect answer for Example 1.6 because inverse images $X^{-1}(B)$ of some Borel sets would not belong to $\{\emptyset, \Omega\}$ unless $\mathcal{F} = \{\emptyset, \Omega\}$. $Y = E(X|\mathcal{G})$ contains precisely as much information about X as is necessary to calculate the integrals in (1.22).

1.6.3 Conditioning as a Projector

Lemma. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G}$ be a σ-field contained in $\mathcal{F}$.*

(i) For any integrable X, $E(X|\mathcal{G})$ exists. Denote $P_{\mathcal{G}}X = E(X \mid \mathcal{G})$ for $X \in L_1(\Omega)$.

(ii) $P_{\mathcal{G}}$ is linear, $P_{\mathcal{G}}(aX + bY) = aP_{\mathcal{G}}X + bP_{\mathcal{G}}Y$, and bounded, $\|P_{\mathcal{G}}X\|_1 \leq \|X\|_1$.

(iii) $P_{\mathcal{G}}$ is a projector.

Proof.

(i) $\nu(A) = \int_A X dP$ defines a σ-additive set function on $\mathcal{G}$ that is absolutely continuous with respect to P. By the Radon–Nikodym theorem there exists a $\mathcal{G}$-measurable function Y such that Eq. (1.22) is true. This proves the existence of $Y = E(X \mid \mathcal{G})$.

(ii) We can use Eq. (1.22) repeatedly to obtain

$$\begin{aligned}\int_A P_{\mathcal{G}}(aX + bY)\, dP &= \int_A (aX + bY)\, dP = a\int_A X dP + b\int_A Y\, dP \\ &= a\int_A P_{\mathcal{G}}X\, dP + b\int_A P_{\mathcal{G}}Y\, dP \\ &= \int_A (aP_{\mathcal{G}}X + bP_{\mathcal{G}}Y)\, dP, \quad A \in \mathcal{G}.\end{aligned}$$

Since $aP_{\mathcal{G}}X + bP_{\mathcal{G}}Y$ is $\mathcal{G}$-measurable, it must coincide with $P_{\mathcal{G}}(aX + bY)$.

For any real-valued function f define its *positive part* by $f_+ = \max\{f, 0\}$ and *negative part* by $f_- = -\min\{f, 0\}$. Then it is geometrically obvious that $f = f_+ - f_-$ and $|f| = f_+ + f_-$. Decomposing $P_{\mathcal{G}}X$ into its positive and negative parts, $P_{\mathcal{G}}X = (P_{\mathcal{G}}X)_+ - (P_{\mathcal{G}}X)_-$, and remembering that both sets $\{P_{\mathcal{G}}X > 0\}$ and $\{P_{\mathcal{G}}X < 0\}$ are $\mathcal{G}$-measurable we have

$$\begin{aligned}\int_\Omega |P_{\mathcal{G}}X|\, dP &= \int_\Omega [(P_{\mathcal{G}}X)_+ + (P_{\mathcal{G}}X)_-]\, dP \\ &= \int_{P_{\mathcal{G}}X > 0} P_{\mathcal{G}}X\, dP + \int_{P_{\mathcal{G}}X < 0} P_{\mathcal{G}}X\, dP \\ &= \int_{P_{\mathcal{G}}X > 0} X\, dP + \int_{P_{\mathcal{G}}X < 0} X\, dP \leq \int_\Omega |X|\, dP.\end{aligned}$$

This proves that $\| P_{\mathcal{G}} \| \leq 1$.

(iii) $P_{\mathcal{G}}^2 X$ is defined as a $\mathcal{G}$-measurable function Y such that $\int_A Y\, dP = \int_A P_{\mathcal{G}}X\, dP$ for all $A \in \mathcal{G}$. Since $P_{\mathcal{G}}X$ itself is $\mathcal{G}$-measurable, we have $Y = P_{\mathcal{G}}X$ a.s.

■

1.6.4 The Law of Iterated Expectations

In a 3-D space, projecting first to a plane and then to a straight line in that plane gives the same result as projecting directly to the straight line. This is also true of conditioning (and projectors in general).

Lemma. *Let $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ be nested σ-fields and denote $P_{\mathcal{H}}$ and $P_{\mathcal{G}}$ as the conditioning projectors on $\mathcal{H}$ and $\mathcal{G}$, respectively. Then $P_{\mathcal{H}}P_{\mathcal{G}} = P_{\mathcal{G}}P_{\mathcal{H}} = P_{\mathcal{H}}$. Using the conditional expectation notation, this is the same as*

$$E[E(X|\mathcal{G})|\mathcal{H}] = E[E(X|\mathcal{H})|\mathcal{G}] = E(X|\mathcal{H}). \tag{1.23}$$

In particular, when $\mathcal{H} = \{\emptyset, \Omega\}$ is the least σ-field, we get $E[E(X|\mathcal{G})] = EX$ for all integrable X.

Proof. $\mathcal{H}$-measurability of $P_{\mathcal{H}}X$ implies its $\mathcal{G}$-measurability. Hence, by Lemma 1.6.3 (iii) $P_{\mathcal{G}}$ doesn't change it. This proves that $P_{\mathcal{G}}P_{\mathcal{H}} = P_{\mathcal{H}}$.

$P_{\mathcal{G}}X$ is $\mathcal{G}$-measurable and satisfies $\int_A P_{\mathcal{G}}X dP = \int_A X\, dP$ for all $A \in \mathcal{G}$. In particular, this is true for $A \in \mathcal{H}$: $\int_A P_{\mathcal{G}}X dP = \int_A X dP$, $A \in \mathcal{H}$. Confronting this with the definition of $P_{\mathcal{H}}P_{\mathcal{G}}X$,

$$\int_A P_{\mathcal{H}}P_{\mathcal{G}}X dP = \int_A P_{\mathcal{G}}X dP, \quad A \in \mathcal{H},$$

we see that $\int_A P_{\mathcal{H}}P_{\mathcal{G}}X dP = \int_A X dP$, $A \in \mathcal{H}$. But $P_{\mathcal{H}}$ satisfies the same equation with $P_{\mathcal{H}}X$ in place of $P_{\mathcal{H}}P_{\mathcal{G}}X$ and both are $\mathcal{H}$-measurable. Hence, $P_{\mathcal{H}}P_{\mathcal{G}}X = P_{\mathcal{H}}X$ a.s. ■

1.6.5 Extended Homogeneity

In the usual homogeneity, $P_{\mathcal{G}}(aX) = aP_{\mathcal{G}}X$, a is a number. In the conditioning context, a can be any $\mathcal{G}$-measurable function, according to the next theorem. I call this property *extended homogeneity*.

Theorem. *If the variables X and XY are integrable and Y is $\mathcal{G}$-measurable, then $P_{\mathcal{G}}(XY) = YP_{\mathcal{G}}X$.*

The proof can be found, for example, in (Davidson 1994, Section 10.4).

1.6.6 Independence

σ-fields $\mathcal{H}$ and $\mathcal{G}$ are called *independent* σ-fields if any event $A \in \mathcal{H}$ is independent of any event $B \in \mathcal{G}$: $P(A \cap B) = P(A)P(B)$. Random variables X and Y are said to be *independent* if σ-fields $\sigma(X)$ and $\sigma(Y)$ are independent. Moreover, a family $\{X_i : i \in I\}$ of random variables is called *independent* if, for any two disjoint sets of indices J, K, σ-fields $\sigma(X_i : i \in J)$ and $\sigma(X_i : i \in K)$ are independent.

Theorem. (Davidson 1994, Section 10.5) *Suppose X is integrable and $\mathcal{H}$-measurable. If $\mathcal{G}$ is independent of $\mathcal{H}$, then conditioning X on $\mathcal{G}$ provides minimum information: $E(X|\mathcal{G}) = EX$.*

1.7 MATRIX ALGEBRA

Everywhere we follow the matrix algebra convention: all matrices and vectors in the same formula are compatible. All matrices in this section are assumed to be of size $n \times n$. The determinant of A is denoted as $\det A$ or $|\mathrm{A}|$.

1.7.1 Orthogonal Matrices

A matrix T is called *orthogonal* if

$$T'T = I. \tag{1.24}$$

Since both $T'T$ and TT' have generic elements $\sum_l t_{il}t_{li}$, Eq. (1.24) is equivalent to $TT' = I$. Equation (1.24) means, by definition of the inverse, that $T^{-1} = T'$.

Geometrically, the mapping $y = Tx$ is rotation in $\mathbb{R}^n$. This is proved by noting that T preserves scalar products: $\langle Tx, Ty\rangle = \langle x, T'Ty\rangle = \langle x, y\rangle$. Hence, it preserves vector lengths and angles between vectors, see Equations (1.13) and (1.14) in Section 1.4.1. Rotation around the origin is the only mapping that has these properties.

1.7.2 Diagonalization of Symmetric Matrices

A number $\lambda \in \mathbb{R}$ is called an *eigenvalue* of a matrix A if there exists a nonzero vector x that satisfies $Ax = \lambda x$. Such a vector x is named an *eigenvector* corresponding to λ. From this definition it follows that A reduces to multiplication by λ along the straight line $\{ax : a \in \mathbb{R}\}$.

The set L of all eigenvectors corresponding to λ, completed with the null vector, is a subspace of $\mathbb{R}^n$, because $Ax = \lambda x$ and $Ay = \lambda y$ imply $A(ax + by) = \lambda(ax + by)$. This subspace is called a *characteristic subspace* of A corresponding to λ. The dimension of the characteristic subspace (see Section 1.1.3) is called *multiplicity* of λ. A reduces to multiplication by λ in L.

We say that a system of vectors $x_1, \ldots, x_k$ is *orthonormal* if

$$\langle x_i, x_j\rangle = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

The system of unit vectors in $\mathbb{R}^n$ is an example of an orthonormal system. An orthonormal system is necessarily linearly independent because scalar multiplication of the equation $a_1x_1 + \cdots + a_kx_k = 0$ by vectors $x_1, \ldots, x_k$ yields $a_1 = \cdots = a_k = 0$.

Theorem. (Diagonalization theorem) (*Bellman 1995, Chapter 4, Section 7). If A is symmetric of size* $n \times n$, *then it has n real eigenvalues* $\lambda_1, \ldots, \lambda_n$, *repeated with their multiplicities. Further, there is an orthogonal matrix T such that*

$$A = T'\Lambda T, \tag{1.25}$$

where Λ *is a diagonal matrix* $\Lambda = \text{diag}[\lambda_1, \ldots, \lambda_n]$. *Finally, the eigenvectors* $x_1, \ldots, x_n$ *that correspond to* $\lambda_1, \ldots, \lambda_n$ *can be chosen orthonormal.*

Equation (1.25) embodies the following geometry. In the original coordinate system with the unit vectors e_j (see Section 1.1.3) the matrix A has generic elements a_{ij}. The first transformation T in Eq. (1.25) rotates the coordinate system to a new position in which A is of simple diagonal form, the new axes being eigenvectors along which applying A amounts to multiplication by numbers. The final transformation by $T' = T^{-1}$ rotates the picture to the original position.

1.7.3 Finding and Applying Eigenvalues

Eigenvalues are the roots of the equation $\det(A - \lambda I) = 0$. Application of this matrix algebra rule is complicated as the left side of the equation is a polynomial of order n. Often it is possible to exploit the analytical structure of A to find its eigenvalues using the next lemma. A subspace L of $\mathbb{R}^n$ is called an *invariant subspace* of a matrix A if $AL \subseteq L$.

Lemma

(i) λ *is an eigenvalue of A if and only if* $\lambda - c$ *is an eigenvalue of* $A - cI$.

(ii) Let L be an invariant subspace of a symmetric matrix A. Denote P an orthoprojector onto L, $Q = I - P$ *and* $M = \text{Im}(Q)$. *Then M is an invariant subspace of A and the analysis of A reduces to the analysis of its restrictions* $A|_L$ *and* $A|_M$.

Proof. Statement

(i) is obvious because the equation $Ax = \lambda x$ is equivalent to $(A - cI)x = (\lambda - c)x$.

(ii) For any $x, y \in \mathbb{R}^n$ by symmetry of A, P,

$$\langle PAQx, y\rangle = \langle AQx, Py\rangle = \langle Qx, APy\rangle = 0.$$

The last equality follows from the facts that $Py \in L = \text{Im}(P)$, $APy \in L$ and $\text{Im}(P)$ is orthogonal to $\text{Im}(Q)$ [see Lemma 1.4.5(i)]. Plugging in $y = PAQx$ we get $\|PAQx\| = 0$ and $PAQx = 0$. Since Qx runs over M when x runs over $\mathbb{R}^n$, we obtain $PAM = \{0\}$ or, by Lemma 1.4.5(ii), $AM \subseteq M$ and M is invariant with respect to A.

Now premultiply by A the identity $I = P + Q$ to get

$$A = AP + AQ = A|_L P + A|_M Q. \qquad \blacksquare$$

The second part of this lemma leads to the following practical rule. If you have managed to find the first eigenvalue λ and the corresponding characteristic subspace L of A, then consider the restriction $A|_M$ to find the rest of the eigenvalues. This process of "chipping off" characteristic subspaces can be repeated. While you do that, construct the orthonormal systems of eigenvectors until their total number reaches n.

Denoting $y = Tx$, from Theorem 1.7.2 we have

$$\langle Ax, x\rangle = \langle T'\Lambda Tx, x\rangle = \langle \Lambda Tx, Tx\rangle = \langle \Lambda y, y\rangle = \sum_{i=1}^{n} \lambda_i y_i^2.$$

Hence, A is nonnegative and $\langle Ax, x\rangle \geq 0$ for all x if and only if all eigenvalues of A are nonnegative. Therefore we can define the *square root* of a nonnegative symmetric matrix by

$$A^{1/2} = T'\mathrm{diag}[\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}]T.$$

1.7.4 Gram Matrices

In a Hilbert space H consider vectors $x_1, \ldots, x_k$. Their *Gram matrix* is defined by

$$G = \begin{pmatrix} \langle x_1, x_1\rangle & \ldots & \langle x_1, x_k\rangle \\ \ldots & \ldots & \ldots \\ \langle x_k, x_1\rangle & \ldots & \langle x_k, x_k\rangle \end{pmatrix}.$$

Theorem. (Gantmacher 1959, Chapter IX, Section 5) *Vectors $x_1, \ldots, x_k$ are linearly independent if and only if* $\det G > 0$.

1.7.5 Positive Definiteness of Gram Matrices

Lemma. *If vectors $x_1, \ldots, x_k \in \mathbb{R}^n$ are linearly independent, then G is positive definite: $\langle Gx, x\rangle > 0$ for all $x \neq 0$.*

Proof. According to the Silvester criterion (Bellman 1995, Chapter 5, Section 3), G is positive definite if and only if all determinants

$$\langle x_1, x_1\rangle, \det\begin{pmatrix} \langle x_1, x_1\rangle & \langle x_1, x_2\rangle \\ \langle x_2, x_1\rangle & \langle x_2, x_2\rangle \end{pmatrix}, \ldots, \det G \tag{1.26}$$

are positive. Linear independence of the system $\{x_1, \ldots, x_k\}$ implies that of all its subsystems $\{x_1\}, \{x_1, x_2\}, \ldots$. Thus all determinants are positive by Theorem 1.7.4. $\blacksquare$

1.7.6 Partitioned Matrices: Determinant and Inverse

Lemma. (Lütkepohl 1991, Section A.10). *Let matrix A be partitioned as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are square. Then

(i) If A_{11} is nonsingular, $|A| = |A_{11}| \cdot |A_{22} - A_{21}A_{11}^{-1}A_{12}|$.

(ii) If A_{11} and A_{22} are nonsingular,

$$\begin{aligned} A^{-1} &= \begin{pmatrix} D & -DA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}D & A_{22}^{-1} + A_{22}^{-1}A_{21}DA_{12}A_{22}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}GA_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}G \\ -GA_{21}A_{11}^{-1} & G \end{pmatrix}, \end{aligned}$$

where $D = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$ and $G = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$.

1.8 CONVERGENCE OF RANDOM VARIABLES

A *random variable* is nothing but a $(\mathcal{F}, \mathcal{B})$-measurable function $X : \Omega \to \mathbb{R}$ where $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathcal{B}$ is the Borel σ-field of $\mathbb{R}$. In the case of a *random vector* it suffices to replace $\mathbb{R}$ by $\mathbb{R}^n$ and $\mathcal{B}$ by $\mathcal{B}_n$, the Borel σ-field of $\mathbb{R}^n$.

1.8.1 Convergence in Probability

Let $X, X_1, X_2, \ldots$ be random vectors defined on the same probability space and with values in the same space $\mathbb{R}^n$. If

$$\lim_{n\to\infty} P(\|X_n - X\|_2 > \varepsilon) = 0 \text{ for any } \varepsilon > 0,$$

then $\{X_n\}$ is said to *converge in probability* to X. Convergence in probability is commonly denoted $X_n \xrightarrow{p} X$ or $\text{plim} X_n = X$. From the equivalent definition

$$\lim_{n\to\infty} P(\|X_n - X\|_2 \leq \varepsilon) = 1 \text{ for any } \varepsilon > 0$$

it may be easier to see that this notion is a natural generalization of convergence of numbers. A nice feature of convergence in probability is that it is preserved under arithmetic operations.

Lemma. *Let* $\{X_i\}$ *and* $\{Y_i\}$ *be sequences of* $n \times 1$ *random vectors and let* $\{A_i\}$ *be a sequence of random matrices such that* $\operatorname{plim} X_i$, $\operatorname{plim} Y_i$ *and* $\operatorname{plim} A_i$ *exist. Then*

(i) $\operatorname{plim}(X_i \pm Y_i) = \operatorname{plim} X_i \pm \operatorname{plim} Y_i$.

(ii) $\operatorname{plim} A_i X_i = \operatorname{plim} A_i \operatorname{plim} X_i$.

(iii) Let $g : \mathbb{R}^n \to \mathbb{R}$ *be a Borel-measurable function such that* $X = \operatorname{plim} X_i$ *takes values in the continuity set* C_g *of* g *with probability 1,* $P(X \in C_g) = 1$. *Then* $\operatorname{plim} g(X_i) = g(X)$.

(iv) If $\operatorname{plim} A_i = A$ *and* $P(\det A \neq 0) = 1$, *then* $\operatorname{plim} A_n^{-1} = A^{-1}$.

Proof. Statements (i) and (ii) are from (Lütkepohl 1991, Section C.1). (iii) is proved in (Davidson 1994, Theorem 18.8).

(iv) The real-valued function $1/\det A$ of a square matrix A of order n is continuous everywhere in the space $\mathbb{R}^{n^2}$ of its elements except for the set $\det A = 0$. Elements of A^{-1} are cofactors of elements of A divided by $\det A$. Hence, they are also continuous where $\det A \neq 0$. The statement follows on applying (iii) element by element. ■

Part (iv) of this lemma does not imply invertibility of A_n a.e. It merely implies that the set on which A_n is not invertible has probability approaching zero.

1.8.2 Distribution Function of a Random Vector

Let X be a random vector with values in $\mathbb{R}^k$. Its *distribution function* is defined by

$$F_X(x) = P(X_1 \leq x_1, \ldots, X_k \leq x_k) = P\left(X^{-1}\left(\prod_{n=1}^{k} (-\infty, x_n]\right)\right), x \in \mathbb{R}^k.$$

It is proved that F_X induces a probability measure on $\mathbb{R}^k$, also denoted by F_X. We say that X has *density* p_X if F_X is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^k$, that is if

$$F_X(A) = \int_A p_X(t)\, dt$$

for any Borel set A. Random vectors X, Y are said to be *identically distributed* if their distribution functions are identical: $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}^k$. The original pair consisting of the vector X and probability space $(\Omega, \mathcal{F}, P)$ is distributed identically with the pair consisting of the identity mapping $X(t) = t$ on $\mathbb{R}^k$ and probability space $(\mathbb{R}^k, \mathcal{B}_k, F_X)$ where $\mathcal{B}_k$ is the Borel field of subsets of $\mathbb{R}^k$. Identically distributed vectors have equal moments. In particular, there are two different formulas for

$$EX = \int_\Omega X(\omega)\, dP(\omega) = \int_{\mathbb{R}^k} t dF_X(t)$$

(see Davidson 1994, Section 9.1).

1.8.3 Convergence in Distribution

We say that a sequence of random vectors $\{X_i\}$ *converges in distribution* to X if $F_{X_i}(t) \to F_X(t)$ at all continuity points t of the limit distribution F_X. For convergence in distribution we use the notation $X_i \xrightarrow{d} X$ or $\text{dlim} X_i = X$.

In econometrics, we are interested in convergence in distribution because confidence intervals for X in the one-dimensional (1-D) case can be expressed in terms of $F_X : P(a < X \leq b) = F_X(b) - F_X(a)$. Here the right-hand side can be approximated by $F_{X_i}(b) - F_{X_i}(a)$ if $\text{dlim} X_i = X$ and a and b are continuity points of F_X (which is always the case if X is normal).

Convergence in distribution is so weak that it is not preserved under arithmetic operations. In expressions like $X_i + Y_i$ or $A_i X_i$ we can pass to the limit in distribution if one sequence converges in distribution and the other *in probability to a constant*.

Lemma. *Let* $\{X_i\}$ *and* $\{Y_i\}$ *be sequences of* $n \times 1$ *random vectors and let* $\{A_i\}$ *be a sequence of random matrices such that* $\text{dlim} X_i$, $\text{plim} Y_i$ *and* $\text{plim} A_i$ *exist*.

(i) *If* $c = \text{plim} Y_i$ *is a constant, then* $\text{dlim}(X_i + Y_i) = \text{dlim} X_i + c$.

(ii) *If* $A = \text{plim} A_i$ *is constant, then* $\text{dlim} A_i X_i = A\,\text{dlim} X_i$.

(iii) $\text{plim} X_i = X$ *implies* $\text{dlim} X_i = X$. If X *is a constant, then the converse is true:* $\text{dlim} X_i = c$ *implies* $\text{plim} X_i = c$.

(iv) *(Dominance of convergence in probability to zero) If* $\text{plim} A_i = 0$*, then the same is true for the product:* $\text{plim} A_i X_i = 0$.

(v) *Suppose* $X_n \xrightarrow{d} X$ *where all random vectors take values in* $\mathbb{R}^k$. *Let* $h : \mathbb{R}^k \to \mathbb{R}^m$ *be measurable and denote* D_h *the set of discontinuities of* h. *If* $F_X(D_h) = 0$, *then* $h(X_n) \xrightarrow{d} h(X)$.

Proof. For (i) and (ii) see (Davidson 1994, Theorem 22.14) (1-D case). The proof of (iii) can be found in (Davidson 1994, Theorems 22.4 and 22.5).

Statement (iv) is proved like this. If $\text{plim} A_i = 0$, then $\text{dlim} A_i X_i = 0$ by (ii), which implies $\text{plim} A_i X_i = 0$ by (iii).

The proof of (v) is contained in (Billingsley 1968, Chapter 1, Section 5). ■

The case $c = 0$ of statement (i) is a perturbation result: adding to $\{X_i\}$ a sequence $\{Y_i\}$ such that $\text{plim} Y_i = 0$ does not change $\text{dlim} X_i$. A continuous h (for which D_h is empty) is a very special case of (v). This case is called a *continuous mapping theorem* (CMT). For (ii) "$\text{plim} A_i$" is not constant, the way around is to prove convergence in distribution of the pair $\{A_i, X_i\}$. Then CMT applied to $h(A_i, X_i) = A_i X_i$ does the job.

1.8.4 Boundedness in Probability

Let $\{X_n\}$ be a sequence of random variables. We know that a (proper) random variable X satisfies $P(|X| > M) \to 0$ as $M \to \infty$. Requiring this property to hold uniformly in n gives us the definition of *boundedness in probability*: $\sup_n P(|X_n| > M) \to 0$

as $M \to \infty$. We write $X_n = O_p(1)$ when $\{X_n\}$ is bounded in probability. This notation is justified by item (i) of the next lemma.

Lemma

(i) If $X_n = x_n = constant$, then $x_n = O(1)$ is equivalent to $X_n = O_p(1)$.
(ii) If $X_n = O_p(1)$ and $Y_n = O_p(1)$, then $X_n + Y_n = O_p(1)$ and $X_n Y_n = O_p(1)$.

Proof.

(i) It is easy to see that

$$\sup_n P(|x_n| > M) = \sup_n 1_{\{|x_n|>M\}} = 1_{\{\sup_n |x_n|>M\}}. \tag{1.27}$$

This implies that $\sup_n P(|X_n| > M) \to 0$ if and only if $\sup_n |x_n| \le M$.

(ii) Let us show that

$$\{|X_n + Y_n| > M\} \subseteq \{|X_n| > M/2\} \cup \{|Y_n| > M/2\}. \tag{1.28}$$

Suppose the opposite is true. Then there exists $\omega \in \Omega$ such that

$$M < |X_n(\omega) + Y_n(\omega)| \le |X_n(\omega)| + |Y_n(\omega)| \le M,$$

which is nonsense. Equation (1.28) implies

$$\begin{aligned}\sup_n P(|X_n + Y_n| > M) &\le \sup_n P\left(|X_n| > \frac{M}{2}\right) + \sup_n P\left(|Y_n| > \frac{M}{2}\right) \\ &\to 0, M \to \infty,\end{aligned}$$

that is, $X_n + Y_n = O_p(1)$. Further, along with Eq. (1.28), we can prove

$$\{|X_n Y_n| > M\} \subseteq \left\{|X_n| > \sqrt{M}\right\} \cup \left\{|Y_n| > \sqrt{M}\right\}$$

and therefore

$$\begin{aligned}\sup_n P(|X_n Y_n| > M) &\le \sup_n P\left(|X_n| > \sqrt{M}\right) + \sup_n P\left(|Y_n| > \sqrt{M}\right) \\ &\to 0, M \to \infty,\end{aligned} \tag{1.29}$$

which proves that $X_n Y_n = O_p(1)$.

■

1.8.5 Convergence in Probability to Zero

The definition of Section 1.8.1 in the special case when $\{X_n\}$ is a sequence of random variables gives the definition of *convergence in probability to zero*: $\lim_{n\to\infty} P(|X_n| > \varepsilon) = 0$ for any ε. In this case, instead of $X_n \xrightarrow{p} 0$ people often write $X_n = o_p(1)$.

Lemma

(i) *If $X_n = x_n =$ constant, then $x_n = o(1)$ is equivalent to $X_n = o_p(1)$.*

(ii) *$X_n = o_p(1)$ implies $X_n = O_p(1)$.*

(iii) *If $X_n = o_p(1)$ and $Y_n = o_p(1)$, then $X_n \pm Y_n = o_p(1)$.*

(iv) *Suppose $X_n = o_p(1)$ or $X_n = O_p(1)$ and $Y_n = o_p(1)$. Then $X_nY_n = o_p(1)$.*

(v) *If $X_n \xrightarrow{d} X$ and $Y_n = o_p(1)$, then $X_nY_n = o_p(1)$.*

Proof.

(i) From an equation similar to Eq. (1.27):

$$\limsup_{n\to\infty} P(|x_n| > \varepsilon) = \limsup_{n\to\infty} 1_{\{|x_n|>\varepsilon\}} = 1_{\{\limsup_{n\to\infty} |x_n|>\varepsilon\}},$$

we see that $\lim_{n\to\infty} P(|X_n| > \varepsilon) = 0$ is equivalent to $\limsup_{n\to\infty} |x_n| \le \varepsilon$ and $X_n = o_p(1)$ is equivalent to $x_n = o(1)$.

(ii) If $X_n = o_p(1)$, then, for any given $\delta > 0$, there exists n_0 such that $P(|X_n| > M) \le \delta$, $n \ge n_0$. Increasing M, if necessary, we can make sure that $P(|X_n| > M) \le \delta$, $n < n_0$. Thus, $\sup_n P(|X_n| > M) \le \delta$. Since $\delta > 0$ is arbitrary, this proves $X_n = O_p(1)$.

(iii) This statement follows from Lemma 1.8.1(i).

(iv) By (ii) $X_n = O_p(1)$, modify Eq. (1.29) to get

$$\sup_{n\ge n_0} P(|X_nY_n| > \varepsilon M) \le \sup_n P(|X_n| > M) + \sup_{n\ge n_0} P(|Y_n| > \varepsilon).$$

Taking an arbitrary $\delta > 0$, choose a sufficiently large M, define $\varepsilon = \delta/M$ and then select a sufficiently large n_0. The right-hand side will be small, which proves $X_nY_n = o_p(1)$.

(v) This is just a different way of stating Lemma 1.8.3(iv).

■

1.8.6 Criterion of Convergence in Distribution of Normal Vectors

A normal vector is defined using its density. We don't need the formula for the density here. It suffices to know that the density of a normal vector e is completely determined by its *first moment* $Ee = \int_{\mathbb{R}^n} t dF_e(t)$ and *second moments* $Ee_ie_j = \int_{\mathbb{R}^n} t_it_j dF_e(t)$.

Lemma. *Convergence in distribution of a sequence* $\{X_k\}$ *of normal vectors takes place if and only if the limits* $\lim EX_k$ *and* $\lim V(X_k)$ *exist where* $V(X) = E(X - EX)(X - EX)'$.

Proof. This statement is obtained by combining two facts. The *characteristic function* ϕ_X of a random vector X is defined by

$$\phi_X(t) = Ee^{i\langle t,X\rangle}, \quad t \in \mathbb{R}^n.$$

Here $i = \sqrt{-1}$. The first fact is that convergence in distribution $\operatorname{dlim} X_k = X$ is equivalent to the pointwise convergence

$$\lim \phi_{X_k}(t) = \phi_X(t) \quad \text{for all } t \in \mathbb{R}^n$$

(see Billingsley 1995, Theorem 26.3). The second fact is that the characteristic function of a normal vector X depends only on two parameters: its mean EX and variance $V(X)$ see (Rao 1965, Section 8a.2). ■

1.9 THE LINEAR MODEL

1.9.1 The Classical Linear Model

The usual assumptions about the linear regression

$$y = X\beta + e \tag{1.30}$$

are the following:

1. y is an observed n-dimensional random vector,
2. the matrix of regressors (or independent variables) X of size $n \times k$ is assumed known,
3. $\beta \in \mathbb{R}^k$ is the parameter vector to be estimated from data (y and X),
4. e is an unobserved n-dimensional error vector with mean zero and
5. $n > k$ and $\det X'X \neq 0$.

The matrix X is assumed constant (deterministic). In dynamic models, with lags of the dependent variable at the right side, those lags are listed separately. I am in favor of separating deterministic regressors from stochastic ones from the very beginning, rather than piling them up together and later trying to specify the assumptions by sorting out the exogenous regressors.

1.9.2 Ordinary Least Squares Estimator

The least squares procedure first gives rise to the *normal equation*

$$X'X\hat{\beta} = X'y$$

for the OLS estimator $\hat{\beta}$ of β and then, subject to the condition $\det X'X \neq 0$, to the formula of the estimator

$$\hat{\beta} = (X'X)^{-1}X'y.$$

This formula and model (1.30) itself lead to the representation

$$\hat{\beta} - \beta = (X'X)^{-1}X'e \tag{1.31}$$

used to study the properties of $\hat{\beta}$. In particular, the assumption $Ee = 0$ implies that $\hat{\beta}$ is unbiased, $E\hat{\beta} = \beta$ and that its distribution is centered on β.

1.9.3 Normal Errors

$N(\mu, \Sigma)$ denotes the *class of normal vectors* with mean μ and variance Σ (which in general may be singular). Errors distributed as $N(0, \sigma^2 I)$ are assumed as the first approximation to reality. Components $e_1, \ldots, e_n$ of such errors satisfy

$$\operatorname{cov}(e_i, e_j) = 0, \quad i \neq j, \quad Ee_i = 0, \quad Ee_i^2 = \sigma^2. \tag{1.32}$$

The first equation here says that $e_1, \ldots, e_n$ are *uncorrelated*.

Lemma. *If $e \sim N(0, \sigma^2 I)$, then the components of e are independent identically distributed.*

Proof. By the theorem from (Rao 1965, Section 8a.2) uncorrelatedness of the components of e plus normality of e imply independence of the components. By Eq. (1.32) the first and second moments of the components coincide, therefore their densities and distribution functions coincide. ■

1.9.4 Independent Identically Distributed Errors

We write $e \sim \mathrm{IID}(0, \sigma^2 I)$ to mean that the components of e are independent identically distributed (i.i.d.), have mean zero and covariance $\sigma^2 I$. Lemma 1.9.3 means that $N(0, \sigma^2 I) \subseteq \mathrm{IID}(0, \sigma^2 I)$.

Lemma. *Suppose $e \sim \mathrm{IID}(0, \sigma^2 I)$ and put $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_t = \sigma(e_j : j \leq t)$, $t = 1, 2, \ldots$ Then e_t is $\mathcal{F}_t$-measurable, $E(e_t|\mathcal{F}_{t-1}) = 0$, $E(e_t^2|\mathcal{F}_{t-1}) = \sigma^2$, $t = 1, \ldots, n$.*

Proof. For $t = 1$, $E(e_1|\mathcal{F}_0) = Ee_1 = 0$ (see Example 1.6 in Section 1.6.2). Let $t > 1$. By definition, $\mathcal{F}_{t-1} = \sigma(e_j : j \leq t-1)$ and $\sigma(e_t)$ are independent.

By Theorem 1.6.6, $E(e_t|\mathcal{F}_{t-1}) = Ee_t = 0$. Similarly, $E(e_t^2|\mathcal{F}_{t-1}) = Ee_t^2 = \sigma^2$ (see Theorem 1.5.4(i) about nonlinear transformations of measurable functions). ■

1.9.5 Martingale Differences

Let $\{\mathcal{F}_t : t = 1, 2, \ldots\}$ be an increasing sequence of σ-fields contained in $\mathcal{F}$: $\mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n \subseteq \ldots \subset \mathcal{F}$. A sequence of random variables $\{e_t : t = 1, 2, \ldots\}$ is called *adapted* to $\{\mathcal{F}_t\}$ if e_t is $\mathcal{F}_t$-measurable for $t = 1, 2, \ldots$ If a sequence of integrable variables $\{e_t\}$ satisfies

1. $\{e_t\}$ is adapted to $\{\mathcal{F}_t\}$ and
2. $E(e_t|\mathcal{F}_{t-1}) = 0$ for $t = 1, 2, \ldots,$ where $\mathcal{F}_0 = \{\emptyset, \Omega\}$,

then we say that $\{e_t, \mathcal{F}_t\}$ or, shorter, $\{e_t\}$ is a *martingale difference* (m.d.) sequence.

Lemma. *Square-integrable m.d. sequences are uncorrelated and have mean zero.*

Proof. By the law of iterated expectations (LIE) [Eq. (1.23)] and the m.d. property item 2 the means are zero:

$$Ee_t = E[E(e_t|\mathcal{F}_{t-1})|\mathcal{F}_0] = 0, \; t = 1, 2, \ldots$$

Let $s < t$. Since e_s is $\mathcal{F}_s$-measurable, it is $\mathcal{F}_{t-1}$-measurable. By extended homogeneity (Section 1.6.5) and the LIE

$$Ee_se_t = E[E(e_se_t|\mathcal{F}_{t-1})] = E[e_sE(e_t|\mathcal{F}_{t-1})] = 0. \quad ■$$

The generality of the m.d. assumption is often reduced by the necessity to restrict the behavior of the second-order conditional moments by the condition

$$E(e_t^2|\mathcal{F}_{t-1}) = \sigma^2, \; t = 1, 2, \ldots \tag{1.33}$$

Owing to the LIE this condition implies $Ee_t^2 = \sigma^2, \; t = 1, 2, \ldots$ We denote by MD(0, σ^2) the square-integrable m.d.'s that satisfy Eq. (1.33). By Lemma 1.9.4, IID(0, $\sigma^2 I$) $\subseteq$ MD(0, σ^2) if we put $\mathcal{F}_t = \sigma(e_j : j \leq t)$.

1.9.6 The Hierarchy of Errors

We have proved that

$$N(0, \sigma^2 I) \subseteq \text{IID}(0, \sigma^2 I) \subseteq \text{MD}(0, \sigma^2). \tag{1.34}$$

Members of any of these three classes have a mean of zero and are uncorrelated. Normal errors are in the core of all error classes considered in this book. This means that any asymptotic results should hold for normal errors and the class of normal errors can be used as litmus paper for tentative assumptions and proofs. The

criterion of convergence in the distribution of normal vectors (Section 1.8.6) facilitates verifying convergence in this class.

Some results will be proved for linear processes as errors. Let $\{\psi_j : j \in \mathbb{Z}\}$ be a double-infinite summable sequence of numbers, $\sum_{j\in\mathbb{Z}}|\psi_j| < \infty$, and let $\{e_j : j \in \mathbb{Z}\}$ be a sequence of integrable zero-mean random variables, called *innovations*. A *linear process* is a sequence $\{v_j : j \in \mathbb{Z}\}$ defined by the convolution

$$v_t = \sum_{j\in\mathbb{Z}} \psi_j e_{t-j}, \quad t \in \mathbb{Z}. \tag{1.35}$$

Members of any of the above three classes may serve as the innovations. If $\psi_0 = 1$ and $\psi_j = 0$ for any $j \neq 0$, we get $v_t = e_t$, which shows that the class of linear processes includes any of the three classes of Eq. (1.34).

Linear processes with summable $\{\psi_j\}$ are called *short-memory processes*. If $\sup_j E|e_j| < \infty$ and $\sum_j|\psi_j| < \infty$, then v_t have uniformly bounded L_1-norms, $E|v_t| \leq \sup_j E|e_j| \sum_j|\psi_j| < \infty$, and zero means. More general processes with square-summable $\{\psi_j\}$, $\sum_{j\in\mathbb{Z}}\psi_j^2 < \infty$, are called *long-memory processes*. In this case, if the innovations are uncorrelated and have uniformly bounded L_2-norms, then v_t exist in the sense of L_2: $Ev_t^2 \leq \sup_j Ee_j^2 \sum_j\psi_j^2 < \infty$. There are also mixing processes, see (Davidson, 1994), which are more useful in nonlinear problems. Long-memory and mixing processes are not considered here. Long-memory processes do not fit Theorem 3.5.2, as discussed in Section 3. Conditions in terms of mixing processes do not look nice, perhaps because they are inherently complex or the theory is underdeveloped.

1.10 NORMALIZATION OF REGRESSORS

1.10.1 Normal Errors as the Touchstone of the Asymptotic Theory

Suppose we have a series of regressions $y = X\beta + e$ with the same β and n going to infinity (dependence of y, X and e on n is not reflected in the notation). We would like to know if the sequence of corresponding OLS estimators $\hat{\beta}$ converges in distribution to a normal vector. We shall see that, as a preliminary step, $\hat{\beta}$ should be centered on β and properly scaled, so that convergence takes place for $D_n(\hat{\beta} - \beta)$, where D_n is some matrix function of the regressors. The factor D_n is called a *normalizer* (it normalizes variances of components of the transformed errors in the OLS estimator formula to a constant). The choice of the normalizer is of crucial importance as it affects the conditions imposed later on X and e.

The classes of regressors and errors should be as wide as possible. The search for these classes is complicated if both regressors and errors are allowed to vary. However, under the hierarchy of errors described above the normal errors are the core of the theory. The implication is that, whatever the conditions imposed on X, they should work for the class of normal errors. The OLS estimator, being a linear transformation

of e, is normal when e is normal. Therefore from the criterion of convergence in distribution of normal vectors (Section 1.8.6) we conclude that the choice of the normalizer and the class of regressors should satisfy the conditions

1. lim $ED_n(\hat{\beta} - \beta)$ exists and
2. lim $V(D_n(\hat{\beta} - \beta))$ exists

when $e \sim N(0, \sigma^2 I)$. For deterministic X, it is natural to stick to deterministic D_n, so condition 1 trivially holds because of unbiasedness of $\hat{\beta}$. The second condition can be called a *variance stabilization condition*.

1.10.2 Where Does the Square Root Come From?

Consider n independent observations on a normal variable with mean β and standard deviation σ. In terms of regression, we are dealing with $X = (1, \ldots, 1)'$ (n unities) and $e \sim N(0, \sigma^2 I)$. From the representation of the OLS estimator (1.31) $\hat{\beta} - \beta = (e_1 + \cdots + e_n)/n$. By independence of the components of e this implies

$$V(\hat{\beta} - \beta) = \frac{1}{n^2}[V(e_1) + \cdots + V(e_n)] = \frac{\sigma^2}{n}.$$

Now it is easy to see that with $D_n = \sqrt{n}$ the variance stabilization condition is satisfied and the criterion of convergence of normal variables gives $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2)$. The square root also works for stable autoregressive models (Hamilton, 1994).

1.10.3 One Nontrivial Regressor and Normal Errors

Consider a slightly more general case $y = x\beta + e$ with $x \in \mathbb{R}^n$ and a scalar β. The representation of the OLS estimator reduces to $\hat{\beta} - \beta = x'e/\|x\|_2^2$ and we easily find that

$$V(\|x\|_2 (\hat{\beta} - \beta)) = \frac{1}{\|x\|_2^2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \sigma^2$$

under the same assumption $e \sim N(0, \sigma^2 I)$. It follows that

$$\|x\|_2 (\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2) \tag{1.36}$$

and $D_n = \|x\|_2$ is the right normalizer.

What if instead of D_n we use $\sqrt{n}$? Then $\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}x'e/\|x\|_2^2$ and the variance stabilization condition leads to

$$\frac{n}{\|x\|_2^2} \to \textit{constant}.$$

This means that the $\sqrt{n}$-rule separates a narrow class of regressors for which $\|x\|_2$ is of order $\sqrt{n}$ for large n. In general, any function of n tending to ∞ as $n \to \infty$ can be used as a normalizer for some class of regressors, and there are as many classes as there are functions with different behavior at infinity.

The normalizer $D_n = \|x\|_2$ is better because it *adapts* to the regressor instead of separating some class. For example, for $x = (1, \ldots, 1)'$ (n unities) it gives the classical square root and for a *linear trend* $x_1 = 1, x_2 = 2, \ldots, x_n = n$ it grows as $n^{3/2}$. As D_n is self-adjusting, you don't need to know the rate of growth of $\|x\|_2$. This is especially important in applications where regressors don't have any particular analytical pattern. The decisive argument is that D_n is in some sense unique (see Section 1.11.3).

1.10.4 The Errors Contribution Negligibility Condition

Let us look again at $y = x\beta + e$ where $e_1, \ldots, e_n$ are now IID(0, $\sigma^2 I$) and not necessarily normal. Having made up our mind regarding the normalizer we need to prove convergence in distribution of

$$\|x\|_2 (\hat{\beta} - \beta) = \frac{x_1}{\|x\|_2} e_1 + \cdots + \frac{x_n}{\|x\|_2} e_n.$$

Here is where CLTs step in. The CLTs we need affirm the asymptotic normality of weighted sums

$$\sum_{t=1}^{n} w_{nt} e_t$$

of random variables $e_1, \ldots, e_n$, which are not necessarily normal. Convergence in distribution of such sums is possible under two types of restrictions.

The first type limits dependence among the random variables and is satisfied in the case under consideration because we assume independence. The second type requires contribution of each term in the sum to vanish asymptotically where

$$\text{contribution} = \frac{\text{variance of a term}}{\text{variance of the sum}}.$$

Under our assumptions this type boils down to the condition

$$\lim_{n \to \infty} \max_{1 \le t \le n} \frac{|x_t|}{\|x\|_2} = 0, \tag{1.37}$$

often called an *errors contribution negligibility condition*. This condition in combination with $e \sim$ IID(0, $\sigma^2 I$) is sufficient to prove Eq. (1.36).

1.11 GENERAL FRAMEWORK IN THE CASE OF K REGRESSORS

1.11.1 The Conventional Scheme

Now in the model $y = X\beta + e$ we allow X to have more than one column and assume $\det X'X \neq 0$, $e \sim \text{IID}(0, \sigma^2 I)$.

The rough approach consists in generalizing upon Section 1.10.2 (with a constant regressor) by relying on the identity

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{X'X}{n}\right)^{-1} \frac{X'e}{\sqrt{n}}. \tag{1.38}$$

Suppose that here

$$\text{limit } A = \lim_{n\to\infty} \frac{X'X}{n} \text{ exists and is nonsingular} \tag{1.39}$$

and that

$$\frac{X'e}{\sqrt{n}} \xrightarrow{d} N(0, B). \tag{1.40}$$

Then, by continuity of matrix inversion $(X'X/n)^{-1} \to A^{-1}$ and the rule for convergence in distribution [Lemma 1.8.3(ii)] implies

$$\left(\frac{X'X}{n}\right)^{-1} \frac{X'e}{\sqrt{n}} \xrightarrow{d} A^{-1}u,\ u \sim N(0, B).$$

As a result,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1}). \tag{1.41}$$

As in case $k = 1$, the rough approach separates a narrow class of regressor matrices by virtue of conditions (1.39) and (1.40).

The refined approach is based on the variance stabilization idea.

Partitioning X into columns, $X = (X_1, \ldots, X_k)$, we see that the vector $u = X'e$ has components $u_j = X_j'e$ with variances $V(u_j) = \sigma^2 \|X_j\|_2^2$. Since $X'X$ is the Gram matrix of the system $\{X_1, \ldots, X_k\}$, the condition $\det X'X \neq 0$ is equivalent to linear independence of the columns (Section 1.7.4) and implies $\|X_j\|_2 \neq 0$ for all j and large n. If we define the normalizer by

$$D_n = \text{diag}\big[\, \|X_1\|_2,\ \ldots,\ \|X_k\|_2 \,\big], \tag{1.42}$$

then the matrix

$$H = XD_n^{-1} = \left(\frac{X_1}{\|X_1\|_2}, \ \ldots, \ \frac{X_k}{\|X_k\|_2} \right) = (H_1, \ \ldots, H_k)$$

has normalized columns, $\|H_j\|_2 = 1$. This construction is simple yet so important that I would love to name it after the discoverer. Unfortunately, the historical evidence is not clear-cut, as is shown in Section 1.11.2. For this reason I call D_n a *variance-stabilizing* (*VS*) normalizer.

The analog of Eq. (1.38) is [see Eq. (1.31)]

$$\begin{aligned} D_n(\hat{\beta} - \beta) &= D_n(X'X)^{-1}X'e \\ &= (D_n^{-1}X'XD_n^{-1})^{-1}D_n^{-1}X'e = (H'H)^{-1}H'e. \end{aligned} \tag{1.43}$$

Naturally, the place of Eqs. (1.39) and (1.40) is taken by

$$\text{limit } A = \lim_{n\to\infty} H'H \text{ exists and is nonsingular} \tag{1.44}$$

and

$$H'e \ \xrightarrow{d} \ N(0, B). \tag{1.45}$$

We call both the combinations of Eqs. (1.38) + (1.39) + (1.40) and Eqs. (1.43) + (1.44) + (1.45) a *conventional scheme* of derivation of the OLS asymptotics.

The result in Section 1.11.3 implies that, if we want to use Eq. (1.43), condition (1.44) is unavoidable. If Eq. (1.44) is not satisfied with any normalization, the conventional scheme itself should be modified (see in Chapter 4, how P.C.B. Phillips handles this issue).

1.11.2 History

The probabilists became aware of the variance stabilization principle a long time ago. It is realized in one or another form in all CLTs. It took some time for the idea to penetrate econometrics.

Eicker (1963) introduced the normalizer D_n, but considered convergence of components of the OLS estimator instead of convergence of the estimator in joint distribution. Anderson (1971) proved convergence in joint distribution using D_n and mentioned that the result "in a slightly different form was given by Eicker". Schmidt (1976), without reference to either Eicker or Anderson, established a result similar to Anderson's. None of these three authors compare D_n to the classical normalizer. Moreover, Schmidt's comments imply that he thinks of D_n as complementary to the square root.

Amemiya (1985) proved Anderson's result, without referring to the three authors just cited. Evidently, he was the first to show that D_n is superior to $\sqrt{n}$ in

the sense that Eq. (1.44) is more general than Eq. (1.39). He also noticed that D_n-type normalization is applicable to maximum likelihood estimators.

Finally, Mynbaev and Castelar (2001) established that D_n is more general than *any other* normalizer, as long as the conventional scheme is employed. This result is the subject of Section 1.11.3.

1.11.3 Universality of D_n

Definition. A diagonal matrix (actually, a sequence of matrices) D_n is called a *conventional-scheme-compliant* (*CSC*) normalizer if $H = XD_n^{-1}$ satisfies Eqs. (1.44) and (1.45) for all errors $e \sim \text{IID}(0, \sigma^2 I)$.
If $\{M_n\}$ is any sequence of nonstochastic diagonal matrices satisfying the condition

$$\text{limit } M = \lim M_n \text{ exists and is nonsingular} \tag{1.46}$$

and D_n is a CSC normalizer, then it is easily checked that $\widetilde{D}_n = M_n D_n$ is also a CSC normalizer with

$$\widetilde{H} = HM_n^{-1}, \tilde{A} = \lim \widetilde{H}' \widetilde{H} = M^{-1}AM^{-1}, \tilde{B} = M^{-1}BM^{-1}.$$

Theorem. (Mynbaev and Castelar 2001) *The VS normalizer (1.42) is unique in the class of CSC normalizers up to a factor satisfying Eq. (1.46). It follows that if with some normalizer the conventional scheme works, then D_n can also be used, while the converse may not be true.*

Proof. Let $\overline{D}_n = \text{diag}[\,\bar{d}_{n1}, \ldots, \bar{d}_{nk}\,]$ be some CSC normalizer, $\overline{H} = X\overline{D}_n^{-1}$, and let $\bar{A}$ and $\bar{B}$ be the corresponding elements of the conventional scheme. The diagonal of the limit relation $\overline{H}'\overline{H} \to \bar{A}$ gives

$$\overline{H}_j'\overline{H}_j = \|X_j\|_2^2/\bar{d}_{nj}^2 \longrightarrow \bar{a}_{jj}, \quad j = 1, \ldots, k, \tag{1.47}$$

where $\overline{H}_j$ denote the columns of $\overline{H}$, X_j the columns of X and $\bar{a}_{ij}$ the elements of $\bar{A}$. Recalling that D_n has $d_{nj} = \|X_j\|_2$ on its diagonal we deduce from Eq. (1.47) that

$$d_{nj}/\bar{d}_{nj} \longrightarrow \bar{a}_{jj}^{1/2}, \quad j = 1, \ldots, k. \tag{1.48}$$

By the Cauchy–Schwarz inequality the elements of $\overline{H}'\overline{H}$ satisfy the inequality $|\overline{H}_i'\overline{H}_j| \le \|\overline{H}_i\|_2 \|\overline{H}_j\|_2$. Letting $n \to \infty$ here and using Eq. (1.47) we get $|\bar{a}_{ij}| \le (\bar{a}_{ii}\,\bar{a}_{jj})^{1/2}$. This tells us that none of the diagonal elements can be zero because otherwise a whole cross in $\bar{A}$ would consist of zeros and $\bar{A}$ would be singular.

Now from Eq. (1.48) we see that $M_n = D_n\overline{D}_n^{-1}$ satisfies Eq. (1.46) and $D_n = M_n\overline{D}_n$ differs from $\overline{D}_n$ by an asymptotically constant diagonal factor. It follows that D_n is CSC with $A = M^{-1}\overline{A}M^{-1}$ and $B = M^{-1}\overline{B}M^{-1}$.

The square root is an example of a normalizer that has a narrower area of applicability than D_n. ■

1.11.4 The Moore–Penrose Inverse

Suppose A is a singular square matrix. According to (Rao 1965, Section 1b.5) the *Moore–Penrose inverse* A^+ of a matrix A is uniquely defined by the properties

$$AA^+A = A, \tag{1.49}$$

$$A^+AA^+ = A^+, \tag{1.50}$$

$$AA^+ \quad \text{and} \quad A^+A \text{ are symmetric.} \tag{1.51}$$

When A is symmetric, A^+ can be constructed explicitly using its diagonal representation. Let A be of order n and diagonalized as $A = P\Lambda P'$ where P is orthogonal, $P'P = I$ and Λ is a diagonal of eigenvalues of A (see Theorem 1.7.2). Denote

$$\left(\frac{1}{\lambda}\right)^+ = \begin{cases} \frac{1}{\lambda}, & \lambda \neq 0; \\ 0, & \lambda = 0. \end{cases} \quad (\Lambda^{-1})^+ = \operatorname{diag}\left[\left(\frac{1}{\lambda_1}\right)^+, \ \ldots, \ \left(\frac{1}{\lambda_n}\right)^+\right],$$

$$A^+ = P(\Lambda^{-1})^+P'.$$

Lemma. *A^+ is the Moore–Penrose inverse of A. It is symmetric and the matrix $Q = A^+A$ is an orthoprojector: $Q' = Q$, $Q^2 = Q$.*

Proof. A^+ is symmetric by construction. It is easy to see that the product $\Delta = (\Lambda^{-1})^+\Lambda$ has zeros where Λ has zeros and unities where Λ has nonzero eigenvalues. Therefore $\Lambda\Delta = \Lambda$ and $\Delta\Lambda^+ = \Lambda^+$, so that Eqs. (1.49) and (1.50) are true:

$$AA^+A = P\Lambda\Delta P' = A, \ A^+AA^+ = P\Delta\Lambda^+P' = A^+.$$

Besides, the matrices $AA^+ = P\Lambda\Lambda^+P'$ and $A^+A = P\Lambda^+\Lambda P' = P\Delta P'$ are symmetric. By the uniqueness of the Moore–Penrose inverse, A^+ is that inverse.

The symmetry of $Q = A^+A$ has just been shown. Q is idempotent: $Q^2 = (A^+A)^2 = P\Delta^2P' = Q$. ■

Note that A^+ is not a continuous function of A. For example,

$$A_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix}$$

converges to

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = A^+$$

but

$$A_n^+ = \begin{pmatrix} 1 & 0 \\ 0 & n \end{pmatrix}$$

does not converge to A^+.

1.11.5 What if the Limit of the Denominator Matrix is Singular?

Can the Moore–Penrose inverse save the situation? It is important to realize that convergence in distribution of $D_n(\hat{\beta} - \beta)$ in the conventional scheme is obtained as a consequence of Equations (1.43)–(1.45) from Section 1.11.1. Since the Moore–Penrose inversion is not continuous, the scheme does not work when the limit of the denominator matrix is singular. The next proposition shows that the Moore–Penrose inverse can be applied if outside (independent of the conventional scheme) information is available in the form

$$\text{limit } v = \text{dlim} D_n(\hat{\beta} - \beta) \text{ exists.} \tag{1.52}$$

Lemma. *If instead of Eq. (1.44) we assume that*

$$\textit{limit } A = \lim_{n \to \infty} H'H \textit{ exists and is singular} \tag{1.53}$$

and if two pieces of information about convergence in distribution are available in the form of Eqs. (1.45) and (1.52), then

$$Qv \sim N(0, A^+BA^+)$$

where $Q = A^+A$ is an orthoprojector.

Proof. The normal equation $X'X(\hat{\beta} - \beta) = X'e$ can be rewritten as

$$H'HD_n(\hat{\beta} - \beta) = H'e.$$

Denoting u the limit of the numerator and using Eqs. (1.53), (1.45) and (1.52) we get $Av = u$. Premultiply this by A^+ to obtain $Qv = A^+u$. Now the statement follows from Eq. (1.45). ■

Thus, under the additional condition (1.52) some projection of v is normally distributed, with a degenerate variance A^+BA^+.

1.12 INTRODUCTION TO L_2-APPROXIMABILITY

1.12.1 Asymptotic Linear Independence

By Theorem 1.7.4 the Gram matrix

$$G = H'H = \begin{pmatrix} H_1'H_1 & \dots & H_1'H_k \\ \dots & \dots & \dots \\ H_k'H_1 & \dots & H_k'H_k \end{pmatrix}$$

is nonsingular if and only if the columns $H_1, \dots, H_k$ of H are linearly independent. Therefore condition (1.44) is termed the *asymptotic linear independence condition*. The question is: can the word "asymptotic" be removed from this name, that is, are there any vectors for which nonsingularity of the limit $A = \lim_{n\to\infty} H'H$ would mean simply linear independence? Imagine that for each j we have convergence of columns $H_j \to M_j$, as $n \to \infty$, in such a way that $H_k'H_l \to M_k'M_k$. Then existence of the limit $A = \lim_{n\to\infty} H'H$ would be guaranteed and $\det A \neq 0$ would mean linear independence of $M_1, \dots, M_k$.

Unfortunately, the sequences $\{H_j : n > k\}$ do not converge. Their elements belong to $\mathbb{R}_2^n$, which can be embedded naturally into $l_2(\mathbb{N})$. A necessary condition for convergence $x^{(n)} \to x$ in $l_2(\mathbb{N})$ is the coordinate-wise convergence $x_i^{(n)} \to x_i$, $n \to \infty$, for all $i = 1, 2, \dots$. But for Eq. (1.45) to be true we have to require the errors contribution negligibility condition (1.37) which in terms of the elements of H looks like this:

$$\lim_{n\to\infty} \max_{i,j} |h_{ij}| = 0.$$

Thus, convergence $H_j \to M_j$, as $n \to \infty$, implies $M_j = 0$, but this is impossible because $\|H_j\|_2 = 1$ for all n because of normalization.

1.12.2 Discretization

The general idea is to approximate sequences of vectors (functions of a discrete argument) with functions of a continuous argument.

For any natural n a function $f \in C[0, 1]$ generates a vector with coordinates $f(i/n)$, $i = 1, \dots, n$. A sequence of vectors $\{x^{(n)}\}$, with $x^{(n)} \in \mathbb{R}^n$ for all n, can be considered close to f if

$$\max_{1\le i\le n} \left| x_i^{(n)} - f\left(\frac{i}{n}\right) \right| \to 0, \quad n \to \infty.$$

This kind of approximation was used by Nabeya and Tanaka (1988), see also, (Tanaka, 1996). A better idea is to use the class $L_2(0, 1)$, which is wider than $C[0, 1]$. However, the members of $L_2(0, 1)$ are defined only up to sets of Lebesgue measure 0, and it doesn't make sense to talk about values $f(i/n)$ for $f \in L_2(0, 1)$. Instead of values we can use integrals $\int_{(i-1)/n}^{i/n} f(t)\,dt$, $i = 1, \dots, n$. For convenience,

the vector of integrals is multiplied by $\sqrt{n}$, which gives the definition of the *discretization operator* δ_n

$$(\delta_n f)_i = \sqrt{n} \int_{(i-1)/n}^{i/n} f(t)\, dt, \quad i = 1, \ldots, n. \tag{1.54}$$

The sequence $\{\delta_n f : n \in \mathbb{N}\}$ is called *L_2-generated by f*. L_2-generated sequences were introduced by Moussatat (1976).

With the volatility of economic data, in econometrics it is unacceptable to require regressors to be L_2-generated or, in other words, to be exact images of some $f \in L_2(0, 1)$ under the mapping δ_n. To allow some deviation from exact images, in a conference presentation (Mynbaev 1997) I defined an *L_2-approximable* sequence as a sequence $\{x^{(n)}\}$ for which there is a function $f \in L_2(0, 1)$ satisfying

$$\|x^{(n)} - \delta_n f\|_2 \to 0.$$

If this is true, we also say that $\{x^{(n)}\}$ is *L_2-close* to $f \in L_2(0, 1)$.

It is worth emphasizing that the OLS estimator asymptotics can be proved without this condition. When the errors are independent, the asymptotic linear independence and errors contribution negligibility condition are sufficient for this purpose, see (Anderson 1971; Amemiya 1985). In 1997 I needed this notion to find the asymptotic behavior of the fitted value, which is a more advanced problem. Note also that (Pötscher and Prucha 1997) and Davidson (1994) used the term L_p-approximability in a different context.

L_2-approximable sequences and, more generally, L_p-approximable sequences defined in (Mynbaev 2000) possess some continuity properties when $p < \infty$. This is their main advantage over general sequences.

1.12.3 Ordinary Least Squares Asymptotics

Theorem. *Consider a linear model $y = X\beta + u$ where*

(i) the errors $u_1, \ldots, u_n$ are defined by Eq. (1.35), the innovations $\{e_j : j \in \mathbb{Z}\}$ are IID$(0, \sigma^2 I)$, $\sum_{j\in\mathbb{Z}} |\psi_j| < \infty$ and e_j^2 are uniformly integrable;

(ii) for each $j = 1, \ldots, k$, the sequence of columns $\{H_j : n > k\}$ of the normalized regressor matrix $H = XD_n^{-1}$ is L_2-close to $M_j \in L_2(0, 1)$;

(iii) the functions $M_1, \ldots, M_k$ are linearly independent.

Then the denominator matrix $H'H$ converges to the Gram matrix G of the system $M_1, \ldots, M_k$ and

$$D_n(\hat{\beta} - \beta) \xrightarrow{d} N(0, (\sigma\beta_\psi)^2 G^{-1}). \tag{1.55}$$

Proof. By Theorem 2.5.3 $\lim_{n\to\infty} H_i'H_j = \int_0^1 M_i M_j\, dt$ and, in consequence, $\lim H'H = G$. By Theorem 3.5.2 $H'u \xrightarrow{d} N(0, (\sigma\beta_\psi)^2 G)$ (this includes the case when $H'u$ converges in distribution and in probability to a null vector). Equation (1.55) follows from the conventional scheme. ■

In similar results with VS normalization (Anderson, 1971, Theorem 2.6.1; Schmidt, 1976, Section 2.7; Amemiya, 1985, Theorem 3.5.4) the errors are assumed independent. Assumptions on H vary from source to source. In Theorems 2.5.3 and 3.5.2 the necessary properties of H are derived from the L_2-approximability assumption. Instead, we could require them directly. When the errors are independent, these properties are: existence of the limit $A = \lim_{n\to\infty} H'H$, asymptotic linear independence $\det A \neq 0$ and the errors contribution negligibility condition $\lim_{n\to\infty} \max_{i,j}|h_{ij}| = 0$. Thus, as far as the OLS asymptotics for the classical model is concerned, the L_2-approximability condition is stronger than the minimum required. It becomes indispensable when deeper properties are needed, like convergence of the fitted value considered next.

1.12.4 Convergence of the Fitted Value

The *fitted value* is defined by $\hat{y} = X\hat{\beta}$. The need for its asymptotics may arise in the following way. Suppose we have to estimate stock $q(t)$ based on its known initial value $q(t_0)$ and flow (rate of change) $q'(t)$. By the Newton–Leibniz formula, $q(t) - q(t_0) = \int_{t_0}^{t} q'(s)\,ds$. If $q'(t)$ is measured at discrete points and regressed on, say, a polynomial of time, the interpolated fitted value approximates q' on the whole interval $[t_0, t]$ and integrating it gives an estimate of $q(t) - q(t_0)$.

As is the case with the OLS estimator, the fitted value has to be transformed to achieve convergence in distribution. Centering on $X\beta$ results in

$$\hat{y} - X\beta = X(\hat{\beta} - \beta) = XD_n^{-1}D_n(\hat{\beta} - \beta) = HD_n(\hat{\beta} - \beta). \tag{1.56}$$

Convergence of $D_n(\hat{\beta} - \beta)$ is available from Theorem 1.12.3, but H does not converge, as explained in Section 1.12.1. It happens, though, that interpolating H leads to a convergent sequence in $L_2(0, 1)$.

A vector x with n values is interpolated by constants to obtain a step function $\Delta_n x = \sum_{t=1}^{n} x_t 1_{it}$. The *interpolation operator* Δ_n is applied to columns of H. From Eq. (1.56) we get

$$\Delta_n(\hat{y} - X\beta) = \Delta_n \sum_{l=1}^{k} H_l[D_n(\hat{\beta} - \beta)]_l = \sum_{l=1}^{k} (\Delta_n H_l)[D_n(\hat{\beta} - \beta)]_l.$$

Theorem. *Under the assumptions of Theorem 1.12.3 the fitted value converges in distribution to a linear combination of the functions $M_1, \ldots, M_k$,*

$$\Delta_n(\hat{y} - X\beta) \xrightarrow{d} \sum_{l=1}^{k} M_l c_t,$$

where the random vector $c = (c, \ldots, c_k)'$ is distributed as $N(0, (\sigma\beta_\psi)^2 G^{-1})$.

Proof. By Lemma 2.5.1 the L_2-approximability condition $\|H_l - \delta_n M_l\| \to 0$ is equivalent to $\|\Delta_n H_l - M_l\| \to 0$. Convergence of $\{\Delta_n H_l\}$ to M_l in L_2 implies convergence in distribution of $\{\Delta_n H\}$ to the vector $M = (M_1, \ldots, M_k)'$. In the expression $\Delta_n(\hat{y} - X\beta) = [\Delta_n H'][D_n(\hat{\beta} - \beta)]$ both factors in brackets at the right

converge in distribution. Since their limits M and $u = \text{dlim} D_n(\hat{\beta} - \beta)$ are independent and, for each n, $\Delta_n H$ and $D_n(\hat{\beta} - \beta)$ are independent, the relations $\Delta_n H \xrightarrow{d} M$ and $D_n(\hat{\beta} - \beta) \xrightarrow{d} u$ imply convergence of the pair $(\Delta_n H, D_n(\hat{\beta} - \beta)) \xrightarrow{d} (M, u)$ see (Billingsley 1968, pp. 26–27). By the continuous mapping theorem then $\Delta_n(\hat{y} - X\beta) \xrightarrow{d} M'u$. ■

1.12.5 Convictions and Preconceptions

In econometrics too much depends on the views of the researcher. Apparently, a set of real-world data can be looked at from different angles. Unfortunately, theoretical studies also suffer from the subjectivity of their authors. Two different sets of assumptions for the same model may lead to quite different conclusions. The choice of the assumptions depends on the previous experience of the researcher, the method employed and the desired result. Assumptions made for and views drawn from a simple model are often taken to a higher level where they can be called convictions if justified or preconceptions if questionable.

A practitioner usually worries only about the qualitative side of the result. A highly technical paper about estimator asymptotics in his/her interpretation boils down to "under some regularity conditions the estimator is asymptotically normal". Hypotheses tests are conducted accordingly, the result is cited without proofs in expository monographs for applied specialists and, with time, becomes a part of folklore. The probability of a critical revision of the original paper declines exponentially.

Imagine that you are a security agent entrusted with the task of capturing an alien that is killing humans. If you presuppose that the beast is disguised like a human your course of actions will be quite different from what it would be if you were looking for a giant cockroach.

When you see a new estimator, its asymptotics is that alien. The best of all is not to presume that it is of a particular type. Make simplified assumptions and look at the finite-sample distributions in the case of normal disturbances. If they are normal, perhaps the asymptotics is also normal. If they are not, a suitable transformation of the estimator, such as centering and scaling, may result in normal asymptotics. Alternatively, you may have to apply a CLT in conjunction with the CMT to obtain nonnormal asymptotics. All these possibilities are illustrated in the book.

By choosing the format of the result you make a commitment. Normal asymptotics is usually proved using a CLT. Let us say it comes with conditions (A), (B) and (C). To satisfy them, you impose in terms of your model conditions (A′), (B′) and (C′), respectively. These conditions determine the class of processes your result is applicable to. By selecting a different format you are bound to use different techniques and obtain a different class.

In the case of the conventional scheme an easy way to go is simply assume that X and e are such that either Eqs. (1.39) + (1.40) or Eqs. (1.44) + (1.45) are satisfied. I call such a "theorem" a *pig-in-a-poke* result. While this approach serves illustrative purposes in a university course well, its value in a research paper or monograph is doubtful. Eicker (1963) mentions that conditions should be imposed separately on the errors and regressors.

In this relation it is useful to distinguish between *low-level conditions*, stated directly in terms of the primary elements of the model, such as Eq. (1.44), and *high-level conditions*, expressed in terms of some complex combinations of the basic elements, such as Eq. (1.45). Of course, this distinction is relative. For instance, the L_2-approximability assumption about deterministic regressors made in the most part of this book is of a lower level than Eq. (1.44).

The *parsimony principle* in econometric modeling states that a model should contain as few parameters as possible or be simple otherwise and still describe the process in question well. A similar principle applies to the choice of conditions. If you have imposed several of them and are about to require a new one, make sure that it is not implied or contradicted by the previous conditions. My major professor, M. Otelbaev, used to say, "If I am allowed to impose many conditions, I can prove anything".

Transparency, simplicity and beauty are other subjective measures of the assumptions quality. A good taste is acquired by reading and comparing many sources. It is not a good idea to have a prospective user of your result prove a whole theorem to check whether your assumptions are satisfied. Nontransparent conditions appealing to *existence* of objects with certain properties are especially dangerous. It is quite possible to use the right theorems and comply with all the rules of formal logic and get a bad statement because the set of objects it applies to will be empty if the conditions are contradictory or existence requirements are infeasible. Contradictions are easy to avoid by using conditions with nonoverlapping responsibilities. In other words, beware of two different conditions governing the behavior of the same object.

Generalizations do not always work, as we have seen when going from constant to variable regressors. However, when studying a dynamic model, such as the mixed spatial model $Y = X\beta + \rho WY + e$ in Chapter 5, I choose the conditions and methods that work for its two submodels, $Y = X\beta + e$ and $Y = \rho WY + e$. In this sense, this book is not free from subjectivity.

Generalizations based on the conventional scheme can be as harmful as any others. The study of the purely spatial model in Chapter 5 shows that the said model violates the habitual notions in several ways:

1. the OLS asymptotics is not normal,

2. the limit of the numerator vector is not normal,

3. the limit of the denominator matrix is not constant,

4. the normalizer is identically 1 (that is, no scaling is necessary) and

5. there is no consistency.

These days requirements to econometric papers are very high. If you suggest a new model, you have to defend it by showing its theoretical advantages and testing its practical performance, preferably in the same paper. The author of a new model can be excused if he/she studies the model under simplified assumptions and leaves the generalizations and refinements to the followers. The way of modeling deterministic regressors advocated here allows us to combine simple assumptions with rigorous proofs.