A

Appendix

A.1 Matrix algebra

A.1.1 Numbers associated with a matrix

Let $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ be an $m \times n$ matrix of real numbers. The spectral norm of A is defined as $\|A\| = \{\lambda_{\max}(A'A)\}^{1/2}$, and the 2-norm of A is defined as $\|A\|_2 = \{\mathrm{tr}(A'A)\}^{1/2}$, where $\lambda_{\max}$ denotes the largest eigenvalue (see below). The following inequalities hold (see Lemma 3.7):

$$\|A\| \le \|A\|_2 \le \sqrt{m \wedge n}\|A\|.$$

If $m = n$, A is called a square matrix. The trace of a square matrix A is defined as the sum of the diagonal elements of A; that is, $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii}$. The trace has the following properties:

(i) $\mathrm{tr}(A) = \mathrm{tr}(A')$, where A' denotes the transpose of A.

(ii) $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ for any square matrices A and B of the same dimension.

(iii) $\mathrm{tr}(cA) = c\,\mathrm{tr}(A)$ for any square matrix A and real number c.

(iv) $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrices A and B, provided that AB and BA are well-defined square matrices.

Let $\pi = (\pi_1, \ldots, \pi_n)$ denote an arbitrary permutation of $(1, \ldots, n)$. The number $\#(\pi)$ of inversions of π is the number of exchanges of pairs of integers π to bring them to the natural order $1, \ldots, n$. Let $A = (a_{ij})_{1\le i,j\le n}$ be a square matrix. The determinant of A is defined as

$$|A| = \sum_{\text{all } \pi} (-1)^{\#(\pi)} \prod_{i=1}^n a_{\pi_i, i}.$$

The determinant has the following properties:

(i) $|A| = |A'|$.

(ii) If the row (or column) of A is multiplied by a number c, $|A|$ is multiplies by c. It follows that $|cA| = c^n|A|$ for $n \times n$ matrix A.

J. Jiang, *Large Sample Techniques for Statistics,*
DOI 10.1007/978-1-4419-6827-2, © Springer Science+Business Media, LLC 2010

(iii) If two rows (or columns) of A are interchanged, the sign of $|A|$ changes. It follows that if two rows (or columns) of A are identical, then $|A| = 0$.

(iv) The value of $|A|$ is unchanged if to its ith row (column) is added c times the jth row (column), where c is a real number. Thus, in particular, if the rows (or columns) of A are not linearly independent, then $|A| = 0$.

(v) If $A = \text{diag}(a_1, \ldots, a_n)$, then $|A| = a_1 \cdots a_n$.

(vi) $|AB| = |A| \cdot |B|$ if the determinants are well defined.

(vii) $|AA'| \geq 0$ and $|A'A| \geq 0$ for any matrix $m \times n$ matrix A.

(viii) $\left| \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \right| = |A| \cdot |B|$.

(ix) $|I_m + AB| = |I_n + BA|$ for any $m \times n$ matrix A and $n \times m$ matrix B.

A.1.2 Inverse of a matrix

The inverse of a matrix A is defined as a matrix B such that $AB = BA = I$, the identity matrix. The inverse of A, if it exists, is unique and denoted by A^{-1}. The following are some basic properties of the inverse:

(i) A^{-1} exists if and only if $|A| \neq 0$.

(ii) $(A')^{-1} = (A^{-1})'$.

(iii) $(cA)^{-1} = c^{-1}A^{-1}$, where c is a nonzero real number.

(iv) $(AB)^{-1} = B^{-1}A^{-1}$, if A^{-1} and B^{-1} both exist.

(v) $\text{diag}(a_1, \ldots, a_n)^{-1} = \text{diag}(a_1^{-1}, \ldots, a_n^{-1})$. More generally, if A is a block-diagonal matrix, $A = \text{diag}(A_1, \ldots, A_k)$, where the diagonal blocks are nonsingular, which is equivalent to $|A_j| \neq 0, 1 \leq j \leq k$, then $A^{-1} = \text{diag}(A_1^{-1}, \ldots, A_k^{-1})$.

An inverse-matrix identity that is very useful is the following. For any $n \times n$ nonsingular matrix A, $n \times q$ matrix U, and $q \times n$ matrix V, we have

$$(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}. \tag{A.1}$$

One of the applications of identity (A.1) is the following. Denote the $n \times 1$ vector of 1's by 1_n (and recall that I_n is the $n \times n$ identity matrix). Let $J_n = 1_n 1_n'$. Then, by (A.1) it is easy to show that for any real numbers a and b such that $a \neq 0$ and $a + nb \neq 0$, we have

$$(aI_n + bJ_n)^{-1} = \frac{1}{a}\left(I_n - \frac{b}{a + nb}J_n\right);$$

we also have $|aI_n + bJ_n| = a^{n-1}(a + nb)$.

For any matrix A, whether it is nonsingular or not, there always exists a matrix A^- satisfying $AA^-A = A$. Such an A^- is called a generalized inverse of A. Note that here we use the term "a generalized inverse" instead of "the generalized inverse" because such an A^- may not be unique. Two special kinds of generalized inverse are often of interest.

Any matrix A^- satisfying

$$AA^{-}A = A \quad \text{and} \quad A^{-}AA^{-} = A^{-}$$

is called a reflexible generalized inverse of A. Given a generalized inverse A^{-} of A, one can produce a generalized inverse that is reflexible by $A_{\mathrm{r}}^{-} = A^{-}AA^{-}$.

If the generalized inverse is required to satisfy the conditions, known as the Penrose conditions, (i) $AA^{-}A = A$, (ii) $A^{-}AA^{-} = A^{-}$, (iii) AA^{-} is symmetric, and (iv) $A^{-}A$ is symmetric, it is called the Moore–Penrose inverse. In other words, a reflexible generalized inverse that satisfies the symmetry conditions (iii) and (iv) is the Moore–Penrose inverse. It can be shown that for any matrix A, its Moore–Penrose inverse exists and is unique.

A.1.3 Kronecker products

Let $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ be a matrix. Then for any matrix B, the Kronecker product $A \otimes B$ is defined as the partitioned matrix $(a_{ij}B)_{1\le i\le m, 1\le j\le n}$. For example, if $A = I_m$ and $B = 1_n$, then $A \otimes B = \mathrm{diag}(1_n, \ldots, 1_n)$. Below are some well-known and useful properties of the Kronecker products:

(i) $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B$.

(ii) $A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$.

(iii) $c \otimes A = A \otimes c = cA$, where c is a real number.

(iv) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.

(v) $(A \otimes B)' = A' \otimes B'$.

(vi) If A is partitioned as $A = [A_1\ A_2]$, then $[A_1\ A_2] \otimes B = [A_1 \otimes B\ A_2 \otimes B]$. However, if B is partitioned as $[B_1\ B_2]$, then $A \otimes [B_1\ B_2] \neq [A \otimes B_1\ A \otimes B_2]$.

(vii) $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1A_2) \otimes (B_2B_2)$.

(viii) If A and B are nonsingular, so is $A \otimes B$, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

(ix) $\mathrm{rank}(A \otimes B) = \mathrm{rank}(A)\mathrm{rank}(B)$.

(x) $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$.

(xi) If A is $m \times m$ and B is $k \times k$, then $|A \otimes B| = |A|^m |B|^k$.

(xii) The eigenvalues of $A \otimes B$ are all possible products of an eigenvalue of A and an eigenvalue of B.

A.1.4 Matrix differentiation

If A is a matrix whose elements are functions of θ, a real-valued variable, then $\partial A/\partial\theta$ represents the matrix whose elements are the derivatives of the corresponding elements of A with respect to θ. For example, if

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \text{then} \quad \frac{\partial A}{\partial\theta} = \begin{pmatrix} \partial a_{11}/\partial\theta & \partial a_{12}/\partial\theta \\ \partial a_{21}/\partial\theta & \partial a_{22}/\partial\theta \end{pmatrix}.$$

If $a = (a_i)_{1\le i\le k}$ is a vector whose components are functions of $\theta = (\theta_j)_{1\le j\le l}$, a vector-valued variable, then $\partial a/\partial\theta'$ is defined as the matrix $(\partial a_i/\partial\theta_j)_{1\le i\le k, 1\le j\le l}$. Similarly, $\partial a'/\partial\theta$ is defined as the matrix $(\partial a/\partial\theta')'$.

The following are some useful results.

(i) (Innerproduct) If a, b, and θ are vectors, then

$$\frac{\partial(a'b)}{\partial\theta} = \left(\frac{\partial a'}{\partial\theta}\right) b + \left(\frac{\partial b'}{\partial\theta}\right) a.$$

(ii) (Quadratic form) If x is a vector and A is a symmetric matrix, then

$$\frac{\partial}{\partial x} x'Ax = 2Ax.$$

(iii) (Inverse) If the matrix A depends on a vector θ and is nonsingular, then, for any component θ_i of θ,

$$\frac{\partial A^{-1}}{\partial\theta_i} = -A^{-1}\left(\frac{\partial A}{\partial\theta_i}\right)A^{-1}.$$

(iv) (Log-determinant) If the matrix A above is also positive definite, then, for any component θ_i of θ,

$$\frac{\partial}{\partial\theta_i}\log(|A|) = \operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial\theta_i}\right).$$

A.1.5 Projection

For any matrix X, the matrix $P_X = X(X'X)^{-1}X'$ is called the projection matrix to $\mathcal{L}(X)$, the linear space spanned by the columns of X. Here, it is assumed that $X'X$ is nonsingular; otherwise, $(X'X)^{-1}$ should be replaced by $(X'X)^-$, the generalized inverse (see Section A.1.2).

To see why P_X is given such a name, note that any vector in $\mathcal{L}(X)$ can be expressed as $v = Xb$, where b is a vector of the same dimension as the number of columns of X. Then we have $P_X v = X(X'X)^{-1}X'Xb = Xb = v$; that is, P_X keeps v unchanged.

The orthogonal projection to $\mathcal{L}(X)$ is defined as $P_{X^\perp} = I - P_X$, where I is the identity matrix. Then, for any $v \in \mathcal{L}(X)$, we have $P_{X^\perp}v = v - P_X v = v - v = 0$. In fact, $P_{X^\perp}$ is the projection matrix to the orthogonal space of X, denoted by $\mathcal{L}(X)^\perp$.

If we define the projection of any vector v to $\mathcal{L}(X)$ as $P_X v$, then if $v \in \mathcal{L}$, the projection of v is itself; if $v \in \mathcal{L}(X)^\perp$, the projection of v is zero (vector). In general, we have the orthogonal decomposition $v = v_1 + v_2$, where $v_1 = P_X v \in \mathcal{L}(X)$ and $v_2 = P_{X^\perp}v \in \mathcal{L}(X)^\perp$ such that $v_1'v_2 = v'P_X P_{X^\perp}v = 0$, because $P_X P_{X^\perp} = P_X(1 - P_X) = P_X - P_X^2 = 0$.

The last equation recalls an important property of a projection matrix; that is, any projection matrix is idempotent (i.e., $P_X^2 = P_X$). For example, if $X = 1_n$ (see Section A.1.2), then $P_X = 1_n(1_n'1_n)^{-1}1_n' = n^{-1}J_n = \bar{J}_n$. The orthogonal projection is thus $I_n - \bar{J}_n$. It is easy to verify that $\bar{J}_n^2 = \bar{J}_n$ and $(I_n - \bar{J}_n)^2 = I_n - \bar{J}_n$.

Another useful result involving projections is the following. Suppose that X is $n \times p$ such that $\text{rank}(X) = p$ and that V is $n \times n$ and positive definite. For any $n \times (n-p)$ matrix A such that $\text{rank}(A) = n - p$ and $A'X = 0$, we have

$$A(A'VA)^{-1}A' = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \tag{A.2}$$

Equation (A.2) may be expressed in a different way:

$$P_{V^{1/2}A} = I - P_{V^{-1/2}X},$$

where $V^{1/2}$ and $V^{-1/2}$ are the square root matrix of V and V^{-1}, respectively (see the next section). In particular, if $V = I$, we have $P_A = I - P_X = P_{X^\perp}$. If X is not of full rank, (A.2) still holds with $(X'V^{-1}X)^{-1}$ replaced by $(X'V^{-1}X)^-$ (see Section A.1.2).

A.1.6 Decompositions of matrices and eigenvalues

There are various decompositions of a matrix satisfying certain conditions. Two of them are most relevant to this book.

The first is Choleski's decomposition. Let A be a nonnegative definite matrix. Then there exists an upper-triangular matrix U such that $A = U'U$. An application of the Choleski decomposition is the following. For any $k \times 1$ vector μ and $k \times k$ covariance matrix V, one can generate a k-variate normal random vector with mean μ and covariance matrix V. Simply let $\xi = \mu + U'\eta$, where η is a $k \times 1$ vector whose components are independent $N(0,1)$ random variables and U is the upper-triangular matrix in the Choleski decomposition of V.

Another decomposition is the eigenvalue decomposition. For any $n \times n$ symmetric matrix A, there exists an orthogonal matrix T such that $A = TDT'$, where $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. In particular, if $A \geq 0$ (i.e., nonnegative definite, in which case the eigenvalues are nonnegative), we define $D^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ and $A^{1/2} = TD^{1/2}T'$, called the square root matrix of A. It follows that $(A^{1/2})^2 = A$. If A is positive definite, then we write $A^{-1/2} = (A^{1/2})^{-1}$, which is identical to $(A^{-1})^{1/2}$. Thus, for example, an alternative way of generating the k-variate normal random vector (see above) is to let $\xi = \mu + V^{1/2}\eta$. The definition of $A^{1/2}$ can be extended A^r for any $A \geq 0$ and $r \in [0,1]$; that is, $A^r = T\text{diag}(\lambda_1^r, \ldots, \lambda_n^r)$. For example, the Löwner–Heinz inequality states that for any matrices A and B such that $A \geq B \geq 0$ and $r \in [0,1]$, we have $A^r \geq B^r$.

The eigenvalue decomposition is one way of diagonalizing a matrix A such that $T'AT = D$, where T is orthogonal and D is diagonal. It follows that any symmetric matrix A can be diagonalized by an orthogonal matrix such that the diagonal elements are the eigenvalues of A. Furthermore, if symmetric matrices $A_1, \ldots, A_k$ are commuting—that is, if $A_jA_j = A_jA_i, 1 \leq i \neq j \leq k$—then they are simultanenously diagonalizable in the sense that there is an

orthogonal matrix T that $T'A_jT = D_j, 1 \leq j \leq k$, where D_j is a diagonal matrix whose diagonal elements are the eigenvalues of A_j, $1 \leq j \leq k$.

The largest and smallest eigenvalues of a symmetric matrix, A, are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. The following properties hold. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A.

(i) For any positive integer p, the eigenvalues of A^p are $\lambda_1^p, \ldots, \lambda_n^p$. Thus, if $A \geq 0$, then $\lambda_{\max}(A^p) = \{\lambda_{\max}(A)\}^p, \lambda_{\min}(A^p) = \{\lambda_{\min}(A)\}^p$.

(ii) $\mathrm{tr}(A) = \lambda_1 + \cdots + \lambda_n$.

(iii) $|A| = \lambda_1 \cdots \lambda_n$.

(iv) For any matrices A and B (not necessarily symmetric), the nonzero eigenvalues of AB are the same as the nonzero eigenvalues of BA. This implies, in particular, that $\lambda_{\max}(AB) = \lambda_{\max}(BA)$ and $\lambda_{\min}(AB) = \lambda_{\min}(BA)$, if the eigenvalues of AB, BA are all positive. Another consequence is that $\lambda_{\max}(A'A) = \lambda_{\max}(AA')$ for any matrix A.

Finally, if A and B are symmetric matrices, whose eigenvalues, arranged in decreasing orders, are $\lambda_1 \geq \cdots \geq \lambda_k$ and $\mu_1 \geq \cdots \geq \mu_k$, respectively, then Weyl's perturbation theorem states that

$$\max_{1 \leq i \leq k} |\lambda_i - \mu_i| \leq \|A - B\|.$$

An application of Weyl's theorem is the following. If A_n is a sequence of symmetric matrices such that $\|A_n - A\| \to 0$ as $n \to \infty$, where A is a symmetric matrix, then the eigenvalues of A_n converge to those of A as $n \to \infty$.

A.2 Measure and probability

Let Ω denote the space of all elements of interest. In our case, Ω is typically the space of all possible outcomes so that the probability of Ω is equal to one. This said, the mathematical definition of probability has yet to be given, even although the concept might seem straightforward as a common sense.

A.2.1 Measures

First, we need to define what is a collection of "reasonable outcomes". Let $\mathcal{F}$ be a collection of subsets of Ω satisfying the following three properties:

(i) The empty set $\emptyset \in \mathcal{F}$.

(ii) $A \in \mathcal{F}$ implies the complement $A^c \in \mathcal{F}$.

(iii) $A_i \in \mathcal{F}, i \in I$, where I is a discrete set of indexes, implies $\cup_{i \in J} A_i \in \mathcal{F}$.

Then $\mathcal{F}$ is called a σ-field. The pair $(\Omega, \mathcal{F})$ is then called a measurable space. The elements of $\mathcal{F}$ are called measurable sets with respect to $\mathcal{F}$, or simply measurable sets, when the context is clear. Let $\mathcal{S}$ be a collection of subsets of Ω. The smallest σ-field that contains $\mathcal{S}$, denoted by $\sigma(\mathcal{S})$, is called the σ-field generated by $\mathcal{S}$. The smallest σ-field does exist. To see this, let $\mathcal{F}$ be the collection of all sets obtained by taking complements, union, or intersection

of elements of $\mathcal{S}$ in any order and possibly multiple times, plus the empty set (which may be understood as taking the union of no elements). It is easy to see that $\mathcal{F}$ is a σ-field and, in fact, $\mathcal{F} = \sigma(\mathcal{S})$. We consider some examples.

Example A.1. Let A be a nonempty proper subset of Ω (i.e., $A \neq \Omega$). Then $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$.

Example A.2 (Borel σ-field). Let $\mathcal{O}$ be the collection of all open sets on R, the real line. The σ-field $\mathcal{B} = \sigma(\mathcal{O})$ is called the Borel σ-field on R. It can be shown that $\mathcal{B} = \sigma(\mathcal{I})$, where $\mathcal{I}$ is the collection of all finite open intervals.

Let $(\Omega, \mathcal{F})$ be a measurable space. A function ν: $\mathcal{F} \mapsto [0, \infty]$ is called a *measure* if the following conditions are satisfied:

(i) $\nu(\emptyset) = 0$;

(ii) if $A_i \in \mathcal{F}, i = 1, 2, \ldots$ are disjoint (i.e., $A_i \cap A_j = \emptyset, i \neq j$), then $\nu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \nu(A_i)$.

In the special case, in which $\nu(\Omega) = 1$, ν is called a *probability measure*. The triple $(\Omega, \mathcal{F}, \nu)$ is then called a measure space; when $\nu(\Omega) = 1$, this is called a *probability space*. Although probability measures are what we are dealing with most of the time, the following nonprobability measures are often used in order to define a probability mass function, or a probability density fuction (see the next subsection).

Example A.3 (Counting measure). Let $\mathcal{F}$ be the collection of all subsets of Ω. It is easy to verify that $\mathcal{F}$ is a σ-field. Now, define, for any $A \in \mathcal{F}$, $\nu(A) =$ the number of elements in A [so $\nu(A) = \infty$ if A contains infinitely many elements]. It is easy to verify that ν is a measure on $(\Omega, \mathcal{F})$, known as the *counting measure*. In particular, if Ω is *countable* in the sense that there is a one-to-one correspondence between Ω and the set positive integers, we may let $\mathcal{F}$ be the collection of all subsets of Ω. This is a σ-field, known as the *trivial* σ-field. A counting measure is then defined on $(\Omega, \mathcal{F})$.

Example A.4 (Lebesgue measure). There is a unique measure ν on $(R, \mathcal{B})$ that satisfies $\nu([a, b]) = b - a$ for any finite interval $[a, b]$. This is called the *Lebesgue measure*. In particular, if $\Omega = [0, 1]$, the Lebesgue measure is a probability measure.

Some basic properties of a measure are the following. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and assume that all the sets considered are in $\mathcal{F}$.

(i) (Monotonicity) $A \subset B$ implies $\nu(A) \leq \nu(B)$.

(ii) (Subadditivity) For any collection of sets $A_i, i \in I$, in $\mathcal{F}$, we have $\nu\left(\bigcup_{i \in I} A_i\right) \leq \sum_{i \in I} \nu(A_i)$.

(iii) (Continuity) If $A_1 \subset A_2 \subset \cdots$, then

$$\nu(\lim_{n \to \infty} A_n) = \lim_{n \to \infty} \nu(A_n), \tag{A.3}$$

where $\lim_{n\to\infty} A_n = \cup_{i=1}^{\infty} A_i$. Similarly, if $A_1 \supset A_c \supset \cdots$ and $\nu(A_1) < \infty$, then (A.3) holds with $\lim_{n\to\infty} A_n = \cap_{i=1}^{\infty} A_i$.

A measure ν on $(\Omega, \mathcal{F})$ is called σ-finite if there is a countable collection of measurable sets $A_i, i \in I$, such that $\nu(A_i) < \infty, i \in I$, and $\Omega = \cup_{i\in I} A_i$.

Let P denote a probability measure on $(R, \mathcal{B})$. The *cumulative distribution function* (cdf) of P is defined as

$$F(x) = P((-\infty, x]), \quad x \in R. \tag{A.4}$$

The cdf has the following properties:

(i) $F(-\infty) \equiv \lim_{x\to-\infty} F(x) = 0$, $F(\infty) \equiv \lim_{x\to\infty} F(x) = 1$;

(ii) F is nondecreasing; that is, $F(x) \le F(y)$ if $x < y$;

(iii) F is right-continuous; that is, $\lim_{y\to x, y>x} F(y) = F(x)$.

It can be shown that for any real-valued function F that satisfies the above properties (i)–(iii), there is a unique probability measure P such that F can be expressed as (A.4).

The concept of cdf can be extended to multivariate case. Let $(\Omega_i, \mathcal{F}_i), i = 1, \ldots, k$, be measurable spaces. The product σ-field on the product space $\prod_{i=1}^k \Omega_i$ is defined as $\sigma(\prod_{i=1}^k \mathcal{F}_i)$, where $\prod_{i=1}^k \Omega_i = \{(x_1, \ldots, x_k) : x_i \in \Omega_i, 1 \le i \le k\}$ and $\prod_{i=1}^k \mathcal{F}_i = \{A_1 \times \cdots \times A_k : A_i \in \mathcal{F}_i, 1 \le i \le k\}$ with $A_1 \times \cdots \times A_k = \{(a_1, \ldots, a_k) : a_i \in A_i, 1 \le i \le k\}$. Note that $\prod_{i=1}^k \mathcal{F}_i$ is not necessarily a σ-field. If $(\Omega_i, \mathcal{F}_i, \nu_i), 1 \le i \le k$, are measure spaces, where $\nu_i, 1 \le i \le k$, are σ-finite, there is a unique σ-finite measure ν on $\{\prod_{i=1}^k \Omega_i, \sigma(\prod_{i=1}^k \mathcal{F}_i)\}$ such that for all $A_i \in \mathcal{F}_i, 1 \le i \le k$,

$$\nu(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu_k(A_k).$$

This is called the product measure, denoted by $\nu = \nu_1 \times \cdots \times \nu_k$. In particular, if $\Omega_i = R, \mathcal{F}_i = \mathcal{B}, 1 \le i \le k$, the corresponding product space and σ-field are denoted by R^k and $\mathcal{B}^k$, respectively. Let P be a probability measure on $(R^k, \mathcal{B}^k)$. The joint cdf of P is defined as

$$F(x_1, \ldots, x_k) = P\{(-\infty, x_1] \times \cdots \times (-\infty, x_k]\}, \tag{A.5}$$

$x_1, \ldots, x_k \in R$. In the special case where $P = P_1 \times P_k$, where P_i is a probability measure on $(\Omega_i, \mathcal{F}_i)$, $1 \le i \le k$, (A.5) becomes

$$F(x_1, \ldots, x_k) = F_1(x_1) \cdots F_k(x_k), \tag{A.6}$$

$x_1, \ldots, x_k \in R$, where F_i is the cdf of $P_i, 1 \le i \le k$.

A.2.2 Measurable functions

Let f be a map from Ω to Λ, an image space. Suppose that there is a σ-field $\mathcal{G}$ on Λ such that

$$f^{-1}(B) \equiv \{\omega \in \Omega : f(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{G}; \tag{A.7}$$

then f is said to be a measurable map from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. In particular, if $\Lambda = R^k$ for some k, f is called a measurable function. If, in addition, $\mathcal{G} = \mathcal{B}^k$, f is called *Borel measurable*.

Now, suppose that $(\Omega, \mathcal{F}, P)$ is a probability space. Any measurable function from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$ is called a *random variable*. Similarly, any measurable function from $(\Omega, \mathcal{F})$ to $(R^k, \mathcal{B}^k)$ $(k > 1)$ is called a random vector, or vector-valued random variable. Let X be a random variable. Define a probability measure on $(R, \mathcal{B})$ by

$$PX^{-1}(B) = P\{X^{-1}(B)\}, \quad B \in \mathcal{B}, \tag{A.8}$$

where $X^{-1}(B)$ is defined by (A.7) with f replaced by X. PX^{-1} is called the distribution of X. The cdf of X is defined as

$$F(x) = PX^{-1}\{(-\infty, x]\} = P(X \le x), \quad x \in R. \tag{A.9}$$

Note that (A.9) is the same as (A.4) with P replaced by PX^{-1}; in other words, the cdf of X is the same as the cdf of PX^{-1}. Note that a random variable, by definition, must be finite, whereas a measurable function may take infinite values, depending on the definition of Λ. We consider an example.

Example A.5. If X is a random variable, then for any $\epsilon > 0$, there is $b > 0$ such that $\mathrm{P}(|X| \le b) > 1 - \epsilon$. This is because, by continuity [see (A.3)],

$$\begin{aligned} 1 &= PX^{-1}\{(-\infty, \infty)\} \\ &= PX^{-1}\left(\lim_{n\to\infty}[-n, n]\right) \\ &= \lim_{n\to\infty} PX^{-1}([-n, n]) \\ &= \lim_{n\to\infty} P(|X| \le n). \end{aligned}$$

Therefore, there must be some $b = n$ such that $\mathrm{P}(|X| \le b) > 1 - \epsilon$.

The following are some basic facts related to Borel-measurable functions.

(i) f is Borel measurable if and only if $f^{-1}\{(-\infty, b)\} \in \mathcal{F}$ for any $b \in R$.

(ii) If f and g are Borel measurable, so are fg and $af + bg$ for any real numbers a and b; and f/g is Botel measurable if $g(\omega) \ne 0$ for any $\omega \in \Omega$.

(iii) If $f_1, f_2, \ldots$ are Borel measurable, so are $\sup_n f_n, \inf_n f_n, \limsup_n f_n$, and $\liminf_n f_n$.

(iv) If f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and g is measurable from $(\Lambda, \mathcal{G})$ to $(\Gamma, \mathcal{H})$, then the composite function, $g \circ f(\omega) = g\{f(\omega)\}, \omega \in \Omega$, is measurable from $(\Omega, \mathcal{F})$ to $(\Gamma, \mathcal{H})$. In particular, if $\Gamma = R$ and $\mathcal{H} = \mathcal{B}$, then $g \circ f$ is Borel measurable.

(iv) If Ω is a Borel subset of R^k, where $k \ge 1$, then any continuous function from Ω to R is Borel measurable.

A class of noncontinuous measurable functions plays an important role in the definition of integrals (see below). These are called *simple measurable functions*, or simply *simple functions*, defined as

$$f(\omega) = \sum_{i=1}^{s} a_i I_{A_i}(\omega), \tag{A.10}$$

where $a_1, \ldots, a_s$ are real numbers, $A_1, \ldots, A_s$ are measurable sets, and I_A is the indicator function defined as

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, \text{ otherwise} \end{cases}$$

A.2.3 Integration

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space. If f is a simple function defined by (A.10), its integral with respect to ν is defined as

$$\int f \, d\nu = \sum_{i=1}^{s} a_i \nu(A_i).$$

From the definition, it follows that $\int I_A \, d\nu = \nu(A)$. Thus, the integral may be regarded as an extension of the measure.

Next, if f is a nonnegative Borel-measurable function, let S_f denote the collection of all nonnegative simple functions g such that $g(\omega) \leq f(\omega)$ for all $\omega \in \Omega$. The integral of f with respect to ν is defined as

$$\int f \, d\nu = \sup_{g \in S_f} \int g \, d\nu.$$

Finally, any Borel-measurable function f can be expressed as $f = f^+ - f^-$, where $f^+(\omega) = f(\omega) \vee 0$ and $f^-(\omega) = -f(\omega) \wedge 0$. Because both f^+ and f^- are nonnegative Borel measurable (why?), the above definition of integral applies to both f^+ and f^-. If at least one of the integrals $\int f^+ \, d\nu$ and $\int f^- \, d\nu$ is finite, the integral of f with respect to ν is defined as

$$\int f \, d\nu = \int f^+ \, d\nu - \int f^- \, d\nu.$$

Example A.3 (continued). Let Ω be a countable set and ν be the counting measure. Then, for any Borel-measurable function f, we have

$$\int f \, d\nu = \sum_{\omega \in \Omega} f(\omega).$$

So, in this case, the integral is the summation.

Example A.4 (continued). If $\Omega = R$ and ν is the Lebesgue measure, the integral is called the *Lebesgue integral*, which is usually denoted by

$$\int f\, d\nu = \int_{-\infty}^{\infty} f(x)\, dx.$$

The definition extends to the multidimensional case in an obvious way. The connection between the Lebesgue integral and the Riemann integral (see §1.5.4.32) is that the two are equal when the latter is well defined. Below are some basic properties of the integrals and well-known results. Although some were already given in Section 1.5, they are listed again for completeness.

(i) For any Borel-measurable functions f and g, as long as the integrals involved exist, we have

$$\int (af + bg)\, d\nu = a \int f\, d\nu + b \int g\, d\nu$$

for any real numbers a, b.

(ii) For any Borel-measurable functions f and g such that $f \le g$ a.e. ν [which means that $\nu(\{\omega : f(\omega) > g(\omega)\}) = 0$], we have $\int f\, d\nu \le \int g\, d\nu$, provided that the integrals exist.

(iii) If f is Borel measurable and $f \ge 0$ a.e. ν, then $\int f d\nu = 0$ implies $f = 0$ a.e. ν.

(iv) (Fatou's lemma) Let $f_1, f_2, \ldots$ be a sequence of Borel-measurable functions such that $f_n \ge 0$ a.e. ν, $n \ge 1$; then

$$\int \left(\liminf_{n\to\infty} f_n \right) d\nu \le \liminf_{n\to\infty} \left(\int f_n\, d\nu \right).$$

(v) (Monotone convergence theorem) If $f_1, f_2, \ldots$ are Borel measurable such that $0 \le f_1 \le f_2 \le \cdots$ and $\lim_{n\to\infty} f_n = f$ a.e. ν, then

$$\int \left(\lim_{n\to\infty} f_n \right) d\nu = \lim_{n\to\infty} \left(\int f_n\, d\nu \right). \tag{A.11}$$

(vi) (Dominated convergence theorem) If $f_1, f_2, \ldots$ are Borel measurable such that $\lim_{n\to\infty} f_n = f$ a.e. ν and there is an integrable function g (i.e., $\int |g|\, d\nu < \infty$) such that $f_n \le g$ a.e. ν, $n \ge 1$, then (A.11) holds.

(vii) (Differentiation under the integral sign) Suppose that for each $\theta \in (a, b) \subset R$, $f(\cdot, \theta)$ is Borel measurable, $\partial f/\partial\theta$ exists, and $\sup_{\theta\in(a,b)} |\partial f/\partial\theta|$ is integrable. Then, for each $\theta \in (a, b)$, $\partial f/\partial\theta$ is integrable and

$$\frac{\partial}{\partial\theta} \int f(\omega, \theta)\, d\nu = \int \frac{\partial}{\partial\theta} f(\omega, \theta)\, d\nu.$$

(viii) (Change of variable) Let f be measurable from $(\Omega, \mathcal{F}, \nu)$ to $(\Lambda, \mathcal{G})$ and g be Borel measurable on $(\Lambda, \mathcal{G})$. Then we have

$$\int_\Omega (g \circ f)\, d\nu = \int_\Lambda g\, d(\nu \circ f^{-1}), \tag{A.12}$$

provided that either integral exists. Here, the measure $\nu \circ f^{-1}$ is defined similarly as (A.8); that is,

$$\nu \circ f^{-1}(B) = \nu\{f^{-1}(B)\}, \quad B \in \mathcal{G},$$

where $f^{-1}(B)$ is defined by (A.7).

(ix) (Fubini's theorem) Let ν_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i), i = 1, 2$, and f be Borel measurable on $\{\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2)\}$ whose integral with respect to $\nu_1 \times \nu_2$ exists. Then $\int_{\Omega_1} f(\omega_1, \omega_2)\, d\nu_1$ is Borel measurable on $(\Omega_2, \mathcal{F}_2)$ whose integral with respect to ν_2 exists, and

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, d\nu_1 \times \nu_2 = \int_{\Omega_2} \left\{ \int_{\Omega_1} f(\omega_1, \omega_2)\, d\nu_1 \right\} d\nu_2.$$

(x) (Radon-Nikodym derivative) Let μ and ν be two measures on $(\Omega, \mathcal{F})$ and μ be σ-finite. ν is said to be *absolutely continuous* with respect to μ, denoted by $\nu \ll \mu$, if for any $A \in \mathcal{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$. The Radon–Nikodym theorem states that if $\nu \ll \mu$, there exists a nonnegative Borel-measurable function f such that

$$\nu(A) = \int_A f\, d\mu, \quad \forall A \in \mathcal{F}. \tag{A.13}$$

The f is unique a.e. μ in the sense that if g is another Borel-measurable function that satisfies (A.13), then $f = g$ a.e. μ. The function f in (A.13) is called the Radon–Nikodym derivative, or density, of ν with respect to μ, denoted by $f = d\nu/d\mu$. The following result holds, which is, again, similar to the change-of-variables rule in calculus: If $\nu \ll \mu$, then for any nonnegative Borel-measurable function f, we have

$$\int f\, d\nu = \int f \left(\frac{d\nu}{d\mu} \right) d\mu. \tag{A.14}$$

A.2.4 Distributions and random variables

The pdf of a random variable X on $(\Omega, \mathcal{F}, P)$ is defined as $F(x) = P(X \leq x), x \in R$. Similarly, the joint pdf of random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ is defined as $F(x_1, \ldots, x_k) = P(X_1 \leq x_1, \ldots, X_k \leq x_k), x_1, \ldots, x_k \in R$. This definition is consistent with (A.9), whose multivariate version is

$$F(x_1, \ldots, x_k) = PX^{-1}\{(-\infty, x]\}, \quad x = (x_1, \ldots, x_k) \in R^k,$$

where $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_k]$.

The random variable X is said to be *discrete* if its possible values are a finite or countable subset of R. Let μ be the counting measure (see Example A.3) and $\nu = PX^{-1}$. It is clear that $\nu \ll \mu$ (why?). The Radon–Nikodym

derivative $f = d\nu/d\mu$ is called the *probability mass function*, or pmf, of X. The definition of pmf can be easily extended to the multivariate case.

A random variable X on on $(\Omega, \mathcal{F}, P)$ is said to be *continuous* if $\nu = PX^{-1} \ll \mu$, the Lebesgue measure. In such a case, the Radon–Nikodym derivative $f = d\nu/d\mu$ is called the *probability density function*, or pdf, of X. Again, the definition can be easily extended to the multivariate case.

A list of commonly used random variables and their pmf's or pdf's can be found, for example, in Casella and Berger (2002, pp. 621–626).

The mean, or expected value, of a random variable X on (Ω, F, P) is defined as $\mathrm{E}(X) = \int_\Omega X \, dP$. Usually, the expected value is calculated via a pmf or pdf. Let μ be a σ-finite measure on $(R, \mathcal{B})$ and assume that $PX^{-1} \ll \mu$. Let $f = dPX^{-1}/d\mu$. Then, by (A.12) and (A.14), we have

$$\begin{aligned}
\mathrm{E}(X) &= \int_\Omega x \circ X \, dP \\
&= \int_R x \, dPX^{-1} \\
&= \int_R x \left(\frac{dPX^{-1}}{d\mu} \right) d\mu \\
&= \int xf \, d\mu.
\end{aligned}$$

In particular, if μ is the counting measure, we have

$$\mathrm{E}(X) = \sum_{x \in S} xf(x) = \sum_{x \in S} xP(X = x),$$

where S is the set of possible values for X (note that in this case, X must be discrete—why?); if μ is the Lebesgue measure on (a, b), where a and b can be finite or infinite, we have

$$\mathrm{E}(X) = \int_a^b xf(x) \, dx.$$

The variance of X is defined as $\mathrm{var}(X) = \mathrm{E}(X - \mu_X)^2$, where $\mu_X = \mathrm{E}(X)$. Another expression of the variance is $\mathrm{var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$. The covariance and correlation between two random variables X and Y on the same probability space are defined as $\mathrm{cov}(X, Y) = \mathrm{E}(X - \mu_X)(Y - \mu_Y)$ and $\mathrm{cor}(X, Y) = \mathrm{cov}(X, Y)/\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}$, respectively. Similar to the variance, an alternative expression for the covariance is $\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$. If $X = (X_1, \ldots, X_k)'$ is a random vector, its expected value is defined as $\mathrm{E}(X) = [\mathrm{E}(X_1), \ldots, \mathrm{E}(X_k)]'$ and its covariance matrix is defined as $\mathrm{Var}(X) = [\mathrm{cov}(X_i, X_j)]_{1 \le i,j \le k}$. If $Y = (Y_1, \ldots, Y_l)'$ is another random vector, where l need not be the same as k, the covariance matrix between X and Y is defined as $\mathrm{Cov}(X, Y) = [\mathrm{cov}(X_i, Y_j)]_{1 \le i \le k, 1 \le j \le l}$. Some basic properties of the expected values and variances are the following, assuming their existence.

(i) $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ for any constants a and b.

(ii) $\mathrm{var}(X + Y) = \mathrm{var}(X) + 2\,\mathrm{cov}(X, Y) + \mathrm{var}(Y)$. In particular, if X and Y are *uncorrelated* in that $\mathrm{cov}(X, Y) = 0$, then $\mathrm{var}(X, Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

(iii) $\mathrm{Cov}(X, Y) = \mathrm{E}\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}'$, where the expected value of a random matrix, $\xi = (\xi_{ij})_{1\le i\le k, 1\le j\le l}$, is defined as $\mathrm{E}(\xi) = [\mathrm{E}(\xi_{ij})]_{1\le i\le k, 1\le j\le l}$. This gives an equivalent definition of the covariance matrix between X and Y, which is often more convenient in derivations. In particular, $\mathrm{Cov}(X, X) = \mathrm{Var}(X) = \mathrm{E}\{X - \mathrm{E}(X)\}\{X - \mathrm{E}(X)\}'$. Thus, we have $\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)'$, and, similar to (ii), $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Y)' + \mathrm{Var}(Y)$.

(iv) $\mathrm{var}(aX) = a^2\mathrm{var}(X)$ for any constant a; $\mathrm{Var}(AX) = A\ \mathrm{Var}(X)A'$ for any constant matrix A; and $\mathrm{Cov}(AX, BY) = A\ \mathrm{Cov}(X, Y)B'$ for any constant matrices A and B.

(v) $\mathrm{var}(X) \ge 0$ and $\mathrm{var}(X) = 0$ if and only if X is a.s. a constant; that is, $P(X = c) = 1$ for some constant c. Similarly, $\mathrm{Var}(X) \ge 0$ (i.e., positive definite); and $|\mathrm{Var}(X)| = 0$ if and only if there is a constant vector a and a constant c such that $a'X = c$ a.s.

(vi) (The covariance inequality) $|\mathrm{cov}(X, Y)| \le \sqrt{\mathrm{var}(X)\mathrm{var}(Y)}$. This implies, in particular, that the correlation between X and Y is always between -1 and 1; when $\mathrm{cor}(X, Y) = -1$ or 1, there are constants a, b, and c such that $aX + bY = c$ a.s.; in other words, one is (a.s.) a linear function of the other.

Other quantities associated with the expected values include the moments and central moments. The pth *moment* of a random variable X is defined as $\mathrm{E}(X^p)$; the pth *absolute moment* is $\mathrm{E}(|X|^p)$; and the pth *central moment* is $\gamma_p = \mathrm{E}\{(X - \mu_X)^p\}$, assuming existence, of course, in each case. The skewness and kurtosis of X are defined as $\kappa_3 = \gamma_3/\sigma^3$ and $\kappa_4 = (\gamma_4/\sigma^4) - 3$, respectively, where $\sigma = \sqrt{\mathrm{var}(X)}$, known as the *standard deviation* of X.

The random variables $X_1, \ldots, X_k$ are said to be independent if

$$P(X_1 \in B_1, \ldots, X_k \in B_k) = P(X_1 \in B_1) \cdots P(X_k \in B_k)$$

for any $B_1, \ldots, B_k \in \mathcal{B}$. Equivalently, $X_1, \ldots, X_k$ are independent if their joint pdf is the product of their individual pdf's; that is,

$$F(x_1, \ldots, x_k) = F_1(x_1) \cdots F_k(x_k)$$

for all $x_1, \ldots, x_k \in R$, where $F(x_1, \ldots, x_k) = P(X_1 \le x_1, \ldots, X_k \le x_k)$ and $F_j(x_j) = P(X_j \le x_j), 1 \le j \le k$. If the joint pdf f of $X_1, \ldots, X_k$ with respect to a product measure, $\mu = \mu_1 \times \cdots \times \mu_k$, exists, where the μ_j's are σ-finite, independence of $X_1, \ldots, X_k$ is also equivalent to

$$f(x_1, \ldots, x_k) = f_1(x_1) \cdots f_k(x_k)$$

for all $x_1, \ldots, x_k$, where f_j is the pdf of X_j with respect to μ_j, which can be derived by integrating the joint pdf; that is,

$$f_j(x_j) = \int f(x_1, \ldots, x_k) d\mu_1 \cdots d\mu_{j-1} d\mu_{j+1} \cdots d\mu_k.$$

The random variables $X_1, \ldots, X_k$ are said to be independent and identically distributed, or i.i.d., if they are independent and have the same distribution, that is, $F_1 = \cdots = F_k$.

A.2.5 Conditional expectations

An extension of the expected values is conditional expectations. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$. Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. The *conditional expectation* of X given $\mathcal{A}$, denoted by $\mathrm{E}(X|\mathcal{A})$, is the a.s. unique random variable that satisfies the following conditions:

(i) $\mathrm{E}(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(R, \mathcal{B})$.

(ii) $\int_A \mathrm{E}(X|\mathcal{A})\, dP = \int_A X\, dP$ for every $A \in \mathcal{A}$.

The *conditional probability* is a special case of the conditional expectation, with $X = 1_B$, the indicator function of $B \in \mathcal{F}$, denoted by $\mathrm{P}(B|\mathcal{A}) = \mathrm{E}(1_B|\mathcal{A})$.

The definition leads to the following result, which is useful in checking if a random variable is the conditional expectation of another random variable: Suppose that ξ is integrable and η is $\mathcal{A}$-measurable. Then $\eta = \mathrm{E}(\xi|\mathcal{A})$ if and only if $\mathrm{E}(\xi\zeta) = \mathrm{E}(\eta\zeta)$ for every ζ that is bounded and $\mathcal{A}$-measurable.

The conditional expectation has the same properties as the expected value—all one has to do is to keep the notation $|\mathcal{A}$ during the operations; however, there are also some important differences. A main difference is that, unlike the expected value which is a constant, the conditional expectation is a random variable, unless in some special cases, such as when $\mathcal{A} = \{\emptyset, \Omega\}$.

If Y is another random variable on $(\Omega, \mathcal{F}, P)$, the conditional expectation of X given Y is defined as $\mathrm{E}(X|Y) = \mathrm{E}\{X|\sigma(Y)\}$, where $\sigma(Y)$ is the σ-field generated by Y, defined as $\sigma(Y) = Y^{-1}(\mathcal{B}) = \{Y^{-1}(B) : B \in \mathcal{B}\}$, where $Y^{-1}(B)$ is defined by (A.7) with $f = Y$. Note that $\mathrm{E}(X|Y)$ is a function of Y so let $\mathrm{E}(X|Y) = h(Y)$, where h is a Borel-measurable function. Then the conditional expectation of X given $Y = y$ is defined as

$$\mathrm{E}(X|Y = y) = h(y).$$

Given the definitions of the conditional expectations, the conditional variances of X given $\mathcal{A}$, or X given Y, are defined as

$$\begin{aligned} \mathrm{var}(X|\mathcal{A}) &= \mathrm{E}[\{X - \mathrm{E}(X|\mathcal{A})\}^2|\mathcal{A}], \\ \mathrm{var}(X|Y) &= \mathrm{E}[\{X - \mathrm{E}(X|Y)\}^2|Y], \end{aligned}$$

respectively. Similar to the variance, the identities

$$\begin{aligned} \mathrm{var}(X|\mathcal{A}) &= \mathrm{E}(X^2|\mathcal{A}) - \{\mathrm{E}(X|\mathcal{A})\}^2, \\ \mathrm{var}(X|Y) &= \mathrm{E}(X^2|Y) - \{\mathrm{E}(X|Y)\}^2 \end{aligned}$$

hold. The latter gives an easier way to define $\mathrm{var}(X|Y = y)$ as

$$\mathrm{E}(X^2|Y=y) - \{\mathrm{E}(X|Y=y)\}^2.$$

Two of the most useful properties of the conditional expectations (variances) are the following, assuming the existence of those throughout.

(i) If $\mathcal{A}_1 \subset \mathcal{A}_2$, then

$$\mathrm{E}(X|\mathcal{A}_1) = \mathrm{E}\{\mathrm{E}(X|\mathcal{A}_2)|\mathcal{A}_1\} \quad \text{a.s.} \tag{A.15}$$

In particular, because the trivial σ-field, $\{\emptyset, \Omega\}$, is a sub-σ-field of any σ-field, we have, by letting $\mathcal{A}_1 = \{\emptyset, \Omega\}$ and $\mathcal{A}_2 = \mathcal{A}$ in (A.15),

$$\mathrm{E}(X) = \mathrm{E}\{\mathrm{E}(X|\mathcal{A})\}.$$

Similarly, if X, Y, and Z are random variables, we have

$$\begin{aligned} \mathrm{E}(X|Y) &= \mathrm{E}\{\mathrm{E}(X|Y,Z)|Y\} \quad \text{a.s.}, \\ \mathrm{E}(X) &= \mathrm{E}\{\mathrm{E}(X|Y)\}. \end{aligned}$$

Here, $\mathrm{E}(X|Y,Z) = \mathrm{E}\{X|\sigma(Y,Z)\}$ and $\sigma(Y,Z)$ is defined the same way as $\sigma(Y)$, treating (Y,Z) as a vector-valued random variable.

(ii) The following identity holds for the conditional variance (see above):

$$\mathrm{var}(X) = \mathrm{E}\{\mathrm{var}(X|Y)\} + \mathrm{var}\{\mathrm{E}(X|Y)\}.$$

In addition, the conditional expectation $\mathrm{E}(X|Y)$ is the minimizer of

$$\mathrm{E}\{X - g(Y)\}^2$$

over all Borel-measurable functions g such that $\mathrm{E}\{g^2(Y)\} < \infty$. Here is another useful result: Let X be a random variable such that $\mathrm{E}(|X|) < \infty$, and let Y and Z be random vectors. If (X,Y) and Z are independent, then

$$\mathrm{E}(X|Y,Z) = \mathrm{E}(X|Y) \quad \text{a.s.} \tag{A.16}$$

We say that, given Y, X and Z are *conditionally independent* if

$$\mathrm{P}(A|Y,Z) = \mathrm{P}(A|Y) \quad \text{a.s. } \forall A \in \sigma(X), \tag{A.17}$$

where the conditional probability $\mathrm{P}(A|\xi)$ is defined as $\mathrm{E}\{1_A|\sigma(\xi)\}$ for any random vector ξ. Given the definition, a similar result to (A.16) is the following. If (X,Y) and Z are independent, then given Y, X and Z are conditionally independent. The conclusion may not be true if Z is independent of X but not of (X,Y).

A.2.6 Conditional distributions

Let X be a random vector on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. There exists a function $P(\cdot,\cdot)$ defined on $\mathcal{B}^k \times \Omega$, where k is the dimension of X, such that the following hold:

(i) $P(B,\omega) = P\{X^{-1}(B)|\mathcal{A}\}$ a.s. for any fixed $B \in \mathcal{B}^k$ [see (A.7) for the definition of $X^{-1}(B)$].

(ii) $P(\cdot,\omega)$ is a probability measure on $(R^k,\mathcal{B}^k)$ for any fixed $\omega \in \Omega$.

If Y is measurable from $(\Omega,\mathcal{F})$ to $(\Lambda,\mathcal{G})$. Then there exists a function defined on $\mathcal{B}^k \times \Lambda$, denoted by $P_{X|Y}(\cdot|\cdot)$, such that the following hold:

(i) $P_{X|Y}(B|y) = P\{X^{-1}(B)|Y=y\} \equiv \mathrm{E}\{1_{X^{-1}(B)}|Y=y\}$ (see above) a.s. PY^{-1} [see (A.8)] for any fixed $B \in \mathcal{B}^k$.

(ii) $P_{X|Y}(\cdot|y)$ is a probability measure on $(R^k,\mathcal{B}^k)$ for any fixed $y \in \Lambda$. The following holds. If $g(\cdot,\cdot)$ is a Borel function such that $\mathrm{E}|g(X,Y)| < \infty$, then

$$\begin{aligned}\mathrm{E}\{g(X,Y)|Y=y\} &= \mathrm{E}\{g(X,y)|Y=y\} \\ &= \int_{R^k} g(x,y)\, dP_{X|Y}(x|y) \ \text{ a.s. } PY^{-1}.\end{aligned}$$

Let $(\Lambda,\mathcal{G},P_1)$ be a probability space. Suppose that P_2 is a function from $\mathcal{B}^k \times \Lambda$ to R such that

(i) $P_2(B,\cdot)$ is Borel measurable for any $B \in \mathcal{B}^k$; and

(ii) $P_2(\cdot|y)$ is a probability measure on $(R^k,\mathcal{B}^k)$ for any $y \in \Lambda$.

Then there is a unique probability measure on $\{R^k \times \Lambda.$
$\sigma(\mathcal{B}^k \times \mathcal{G})\}$ such that

$$P(B \times C) = \int_C P_2(B,y)\, dP_1(y) \tag{A.18}$$

for any $B \in \mathcal{B}^k$ and $C \in \mathcal{G}$. In particular, if $(\Lambda,\mathcal{G}) = (R^l,\mathcal{B}^l)$ and define $X(x,y) = x$ and $Y(x,y) = y$ for $(x,y) \in R^k \times R^l$, then $P_1 = PY^{-1}$ and $P_2(\cdot,y) = P_{X|Y}(\cdot|y)$ (see above), and the probability measure (A.18) is the joint distribution of (X,Y) that has the joint cdf

$$F(x,y) = \int_{(-\infty,y]} P_{X|Y}\{(-\infty,x]|v\}\, dPY^{-1}(v), \quad x \in R^k, y \in R^l,$$

where $(-\infty,a]$ is defined above (see the beginning of Section A.2.4). $P_{X|Y}(\cdot|y)$, denoted by $P_{X|Y=y}$, is called the conditional distribution of X given $Y=y$. If $P_{X|Y=y}$ has a pdf with respect to ν, a σ-finite measure on $(R^k,\mathcal{B}^k)$, the pdf is denoted by $f_{X|Y}(\cdot|y)$, known as the conditional pdf of X given $Y=y$.

A.3 Some results in statistics

A.3.1 The multivariate normal distribution

A random vector ξ is said to have a k-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ, or $\xi \sim N(\mu,\Sigma)$, if the joint pdf of ξ is given by

$$f(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \qquad x \in R^k.$$

When $k = 1$, the pdf reduces to that of $N(\mu, \sigma^2)$ with $\sigma^2 = \Sigma$. Below are some useful results associated with the multivariate normal distribution. Here, we assume that all of the matrix products are well defined.

1. If $\xi \sim N(\mu, \Sigma)$, then for any constant matrix A, $A\xi \sim N(A\mu, A\Sigma A')$.

2. If $\xi \sim N(\mu, \Sigma)$, then for any constant matrices A and B, $A\xi$ and $B\xi$ are independent if and only if $A\Sigma B' = 0$. If ξ is multivariate normal, the components of ξ are independent if and only if they are uncorrelated; that is, $\mathrm{cov}(\xi_i, \xi_j) = 0$, $i \neq j$, where ξ_i is the ith component of ξ.

3. If $\xi \sim N(\mu, \Sigma)$ and ξ, μ and Σ are partitioned accordingly as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then the conditional distribution of ξ_1 given ξ_2 is multivariate normal with mean vector $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\xi_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Note that $\Sigma_{21} = \Sigma_{12}'$.

4. Let ξ be a random vector such that $\mathrm{E}(\xi) = \mu$ and $\mathrm{Var}(\xi) = \Sigma$. Then, for any constant symmetric matrix A, we have

$$\mathrm{E}(\xi' A \xi) = \mu' A \mu + \mathrm{tr}(A\Sigma).$$

In particular, if $\xi \sim N(0, \Sigma)$, then $\xi' A \xi$ is distributed as χ_r^2 if and only if $A\Sigma$ is idempotent (see Section A.1.5) and $r = \mathrm{rank}(A)$. In particular, if $\xi = (\xi_i)_{1 \le i \le k} \sim N(0, I_k)$, then $|\xi|^2 = \xi_1^2 + \cdots + \xi_k^2 \sim \chi_k^2$.

5. If $\xi \sim N(\mu, \Sigma)$, a is a constant vector, and A and B are constant symmetric matrices, then $a'\xi$ and $\xi' A \xi$ are independent if and only if $b'\Sigma A = 0$; $\xi' A \xi$ and $\xi' B \xi$ are independent if and only if $A\Sigma B = 0$. Also, we have

$$\begin{aligned} \mathrm{cov}(\xi' A \xi, b'\xi) &= 2b'\Sigma A \mu, \\ \mathrm{cov}(\xi' A \xi, \xi' B \xi) &= 4\mu' A \Sigma B \mu + 2\,\mathrm{tr}(A\Sigma B \Sigma). \end{aligned}$$

6. If $\xi \sim N(0, 1)$, $\eta \sim \chi_d^2$, and ξ and η are independent, then

$$t = \frac{\xi}{\sqrt{\eta/d}} \sim t_d,$$

the t-distribution with d degrees of freedom, which has the pdf

$$f(x) = \frac{\Gamma\{(d+1)/2\}}{\sqrt{d\pi}\,\Gamma(d/2)} \left(1 + \frac{x^2}{d}\right)^{-(d+1)/2}, \quad -\infty < x < \infty.$$

An extension of the t-distribution is the multivariate t-distribution. A k-dimensional multivariate t-distribution with mean vector μ, covariance matrix Σ, and degrees of freedom d has the joint pdf

$$\frac{\Gamma\{(d+k)/2\}}{(d\pi)^{k/2}\Gamma(d/2)}|\Sigma|^{-1/2}\left\{1+\frac{1}{d}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}^{-(d+k)/2}, \quad x \in R^d.$$

7. If $\xi_j \sim \chi^2_{d_j}$, $j = 1, 2$, and ξ_1 and ξ_2 are independent, then

$$F = \frac{\xi_1/d_1}{\xi_2/d_2} \sim F_{d_1,d_2},$$

the F-distribution with d_1 and d_2 degrees of freedom, which has the pdf

$$f(x) = \frac{\Gamma\{(d_1+d_2)/2\}}{\Gamma(d_1/2)\Gamma(d_2/2)}\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}\left\{1+\left(\frac{d_1}{d_2}\right)x\right\}^{-(d_1+d_2)/2},$$

$-\infty < x < \infty$.

A.3.2 Maximum likelihood

Let X be a vector of observations and let $f(\cdot|\theta)$ the pdf of X, with respect to a σ-finite measure μ—that is dependent on a vector of parameters, θ. Let x denote the observed value of X. The notation $f(x|\theta)$ can be viewed in two ways. For a fixed θ, it is the pdf of X when considered as a function of x; for the fixed x (as the observed X), it is viewed as a function of θ, known as the *likelihood function*. In the latter case, a different notation is often used, $L(\theta|x) = f(x|\theta)$. The *log-likelihood* is the logarithm of the likelihood function, denoted by $l(\theta|x) = \log\{L(\theta|x)\}$.

A widely used method of estimation is the *maximum likelihood*; namely, the parameter vector θ is estimated by the maximizer of the likelihood function. More precisely, let Θ be the *parameter space*—that is, the space of possible values of θ. Suppose that there is $\hat{\theta} \in \Theta$ such that

$$L(\hat{\theta}|x) = \sup_{\theta\in\Theta} L(\theta|x);$$

then $\hat{\theta}$ is called the maximum likelihood estimator, or MLE, of θ.

Under widely existing regularity conditions, the maximum likelihood is carried out by differentiating the log-likelihood with respect to θ and solving the equations that equate the derivatives to zero; that is,

$$\frac{\partial}{\partial\theta} l(\theta|x) = 0. \tag{A.19}$$

Equation (A.19) is known as the ML equation. It should be noted that a solution to the ML equation is not necessarily the MLE. However, under some more restrictive conditions, the solution indeed coincides with the MLE. For example, if (A.19) has a unique solution and it can be made sure that the maximum of $L(\theta|x)$ does not occur on the boundary of Θ, then the MLE is identical to the solution of the ML equation.

Associated with the log-likelihood function is the *information matrix*, or Fisher information matrix, defined as

$$I(\theta) = \mathrm{E}_\theta \left(\frac{\partial l}{\partial \theta} \frac{\partial l}{\partial \theta'} \right), \tag{A.20}$$

where $\partial l/\partial\theta = (\partial/\partial\theta) l(\theta|X)$ and E_θ stands for expectation with θ being the true parameter vector. It should be noted—and this is important—that the θ in E_θ must be the same as the θ in $\partial l/\partial\theta$ on the right side of (A.20).

Under some regularity conditions, the following nice properties hold for the log-likelihood function.

(i) [An integrated version of (A.19)]

$$\mathrm{E}_\theta \left\{ \frac{\partial}{\partial \theta} l(\theta|X) \right\} = 0.$$

(ii) [Another expression of (A.20)]

$$I(\theta) = -\mathrm{E}_\theta \left\{ \frac{\partial^2 l}{\partial\theta\partial\theta'} \right\},$$

where $\partial^2 l/\partial\theta\partial\theta' = (\partial^2/\partial\theta\partial\theta') l(\theta|X)$. Properties (i) and (ii) lead to a third expression for $I(\theta)$:

$$I(\theta) = \mathrm{Var}_\theta \left(\frac{\partial l}{\partial \theta} \right),$$

where $\partial l/\partial\theta = (\partial/\partial\theta) l(\theta|X)$ and Var_θ stands for covariance matrix with θ being the true parameter vector.

A well-known result involving the Fisher information matrix is the Cramér–Rao lower bound. For simplicity, let θ be a scalar parameter. Let $\hat{\delta}$ be an unbiased estimator of $\delta = g(\theta)$; that is, $\mathrm{E}_\theta(\hat{\delta}) = g(\theta)$ for all θ, where g is a differentiable function. Under regularity conditions, we have

$$\mathrm{var}_\theta(\hat{\delta}) \geq \frac{\{g'(\theta)\}^2}{I(\theta)}.$$

For a multivariate version, see, for example, Shao (2003, p. 169).

An important and well-known property of the MLE is its asymptotic efficiency in the sense that, under regularity conditions, the asymptotic covariance matrix of the MLE is equal to the Cramér–Rao lower bound. For example, in the i.i.d. case, we have, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} N\{0, I^{-1}(\theta)\},$$

where the right side is the multivariate normal distribution with mean vector 0 and covariance matrix $I^{-1}(\theta) = I(\theta)^{-1}$ (see Section A.3.1).

Many testing problems involve the *likelihood ratio*. This is defined as

$$\frac{\sup_{\theta\in\Theta_0} L(\theta|x)}{\sup_{\theta\in\Theta} L(\theta|x)}, \tag{A.21}$$

where Θ_0 is the parameter space under the null hypothesis, H_0, and Θ is the parameter space without assuming H_0.

A.3.3 Exponential family and generalized linear models

The concept of generalized linear models, or GLMs, is closely related to that of the exponential family. The distribution of a random variable Y is a member of the exponential family if its pdf or pmf can be expressed as

$$f(y;\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{A.22}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are known functions, θ is an unknown parameter, and ϕ is an additional dispersion parameter, which may or may not be known. Many of the well-known distributions are members of the exponential family. These include normal, Gamma, binomial, and Poisson distributions.

An important fact regarding the exponential family is the following relationship between the mean of Y and θ:

$$\mu = \mathrm{E}(Y) = b'(\theta).$$

In many cases, this establishes an 1–1 correspondence between μ and θ. Another relationship among θ, ϕ, and the variance of Y is

$$\mathrm{var}(Y) = b''(\theta)a(\phi).$$

The following is an example.

Example A.6. Suppose that $Y \sim \text{Binomial}(n,p)$. Then the pmf of Y can be expressed as (A.22) with

$$\theta = \log\left(\frac{p}{1-p}\right), \quad b(\theta) = n\log(1+e^{\theta}), \quad \text{and} \quad a(\phi) = \log\binom{n}{y}.$$

Note that in this case, $\phi = 1$. It follows that $b'(\theta) = ne^{\theta}/(1+e^{\theta}) = np = \mathrm{E}(Y)$ and $b''(\theta) = ne^{\theta}/(1+e^{\theta})^2 = np(1-p) = \mathrm{var}(Y)$.

McCullagh and Nelder (1989) introduced the GLM as an extension of the classical linear models. Suppose the following:

(i) The observations $y_1,\ldots,y_n$ are independent.

(ii) The distribution of y_i is a member of the exponential family, which can be expressed as

$$f_i(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y,\phi)\right\}.$$

(iii) The mean of y_i, μ_i, is associated with a linear predictor $\eta_i = x_i'\beta$ through a link function; that is,

$$\eta_i = g(\mu_i),$$

where x_i is a vector of known covariates, β is a vector of unknown regression coefficients, and $g(\cdot)$ is a link function.

Assumptions (i)–(iii) define a GLM. By the properties of the exponential family mentioned above, θ_i is associated with η_i. In particular, if

$$\theta_i = \eta_i,$$

the link function $g(\cdot)$ is called *canonical*.

The function $a_i(\phi)$ typically takes the form $a_i(\phi) = \phi/w_i$, where w_i is a weight. For example, if the observation y_i is the average of k_i observations (e.g., a binomial proportion, where k_i is the number of Bernoulli trials), then $w_i = k_i$; if the observation is the sum of k_i observations (e.g., a binomial or sum of Bernoulli observations), then $w_i = 1/k_i$.

A.3.4 Bayesian inference

Suppose that Y is a vector of observations and θ is a vector of parameters that are not observable. Let $f(y|\theta)$ represent the probability density function (pdf) of Y given θ and let $\pi(\theta)$ represent a *prior* pdf for θ. Then the *posterior* pdf of θ is given by

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)\, d\theta}. \tag{A.23}$$

Obtaining the posterior is often the goal of Bayesian inference. In particular, some numerical summaries may be obtained from the posterior. For example, a Bayesian point estimator of θ is often obtained as the *posterior mean*:

$$\begin{aligned} \mathrm{E}(\theta|y) &= \int \theta p(\theta|y)\, d\theta \\ &= \frac{\int \theta f(y|\theta)\pi(\theta)\, d\theta}{\int f(y|\theta)\pi(\theta)\, d\theta}; \end{aligned}$$

the *posterior variance*, $\mathrm{var}(\theta|y)$, on the other hand, is often used as a Bayesian measure of uncertainty. The notation $d\theta$ in the above, which corresponds to the Lebesgue measure, can be replaced by $d\mu$, where μ is a σ-finite measure with respect to which $\pi(\cdot)$ is defined.

A discrete probabilistic version of (A.23) is called the *Bayes rule*, which is often useful in computing the conditional. Suppose that there are a number of events, $A_1, \ldots, A_k$, such that $A_i \cap A_j = \emptyset, i \neq j$, and $A_1 \cup \cdots \cup A_k = \Omega$, the sample space. Then, for any event B, we have

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_{j=1}^{k} \mathrm{P}(B|A_j)\mathrm{P}(A_j)}, \tag{A.24}$$

$1 \le i \le k$. To see the connection between (A.23) and (A.24), suppose that π is a discrete distribution over $\theta_1, \ldots, \theta_k$, and the distribution of Y is also discrete. Then, for any possible value y of Y, we have, by (A.24) with $A_i = \{\theta = \theta_i\}, 1 \le i \le k$, and $B = \{Y = y\}$,

$$\mathrm{P}(\theta = \theta_i|Y = y) = \frac{\mathrm{P}(Y = y|\theta = \theta_i)\mathrm{P}(\theta = \theta_i)}{\sum_{j=1}^{k} \mathrm{P}(Y = y|\theta = \theta_j)\mathrm{P}(\theta = \theta_j)},$$

which is the discrete version of (A.23).

The posterior can be used to obtain a *posterior predictive distribution* of a future observation, $\tilde{Y}$. Suppose that Y and $\tilde{Y}$ are conditionally independent given θ (see Section A.2.5); then, by (A.17), we have $f(\tilde{y}|\theta, y) = f(\tilde{y}|\theta)$. Therefore, the posterior predictive pdf is given by

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)\, d\theta \\ &= \int f(\tilde{y}|\theta, y)p(\theta|y)\, d\theta \\ &= \int f(\tilde{y}|\theta)p(\theta|y)\, d\theta \end{aligned}$$

(here, as usual, f and p denote the pdf's, and the rule of notation is that p is used whenever the conditioning involves y only; otherwise, f is used).

Similar to Section A.3.2, the pdf $f(y|\theta)$, considered as a function of θ, is called the *marginal likelihood*, or simply likelihood. The ratio of the posterior $p(\theta|y)$ evaluated at the points θ_1 and θ_2 under a given model is called the *posterior odds* for θ_1 compared to θ_2—namely,

$$\begin{aligned} \frac{p(\theta_1|y)}{p(\theta_2|y)} &= \frac{\pi(\theta_1)f(y|\theta_1)/f(y)}{\pi(\theta_2)f(y|\theta_2)/f(y)} \\ &= \frac{\pi(\theta_1)}{\pi(\theta_2)} \cdot \frac{f(y|\theta_1)}{f(y|\theta_2)}, \end{aligned}$$

according to the Bayes rule (A.23). In other words, the posterior odds is simply the prior odds multiplied by the *likelihood ratio*, $f(y|\theta_1)/f(y|\theta_2)$ [see (A.21)]. The concept of (posterior) odds is most familiar when θ takes two possible values, with θ_2 being the complement of θ_1.

A similar concept is the *Bayesian factor*. This is used, for example, when a discrete set of competing models is proposed for model selection. The Bayesian factor is the ratio of the marginal likelihood under one model to that under another model. If we label the two competing models by M_1 and M_2, respectively, then the ratio of their posterior probabilities is

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{\pi(M_1)}{\pi(M_2)} \times \text{Bayesian factor}(M_1, M_2),$$

which defines the Bayesian factor—namely,

$$\begin{aligned}\text{Bayesian factor}(M_1, M_2) &= \frac{f(y|M_1)}{f(y|M_2)} \\ &= \frac{\int \pi(\theta_1|M_1) f(y|\theta_1, M_1)\, d\theta_1}{\int \pi(\theta_2|M_2) f(y|\theta_2, M_2)\, d\theta_2}.\end{aligned}$$

A.3.5 Stationary processes

Many important processes have the *stationarity* properties, in one way or the other. For simplicity, consider a process $X(t), t \geq 0$, taking values in R.

The process is said to be *strongly stationary* if $[X(t_1), \ldots, X(t_n)]$ and $[X(t_1 + h), \ldots, X(t_n + h)]$ have the same joint distribution for all $t_1, \ldots, t_n$ and $h > 0$. Note that if $X(t), t \geq 0$, is strongly stationary, then, in particular, the $X(t)$'s are identically distributed. However, strong stationarity is a much stronger property than identical distribution. The process is said to be *weakly* (or *second-order*) stationary if $\mathrm{E}\{X(t_1)\} = \mathrm{E}\{X(t_2)\}$ and $\mathrm{cov}\{X(t_1), X(t_2)\} = \mathrm{cov}\{X(t_1 + h), X(t_2 + h)\}$ for all t_1, t_2 and $h > 0$.

The terms used here might suggest that a strongly stationary process must be weakly stationary. However, this is not implied by the definition, unless the second moment of $X(t)$ is finite for all t (in which case, the claim is true). On the other hand, a weakly stationary process may not be strongly stationary, of course, unless the process is Gaussian, as in the first example below.

Example A.7 (Gaussian process). A real-valued process $X(t), t \geq 0$, is said to be *Gaussian* if each finite-dimensional vector $[X(t_1), \ldots, X(t_n)]'$ has a multivariate normal distribution. Now, suppose that the Gaussian process is weakly stationary. Then the vectors $U = [X(t_1), \ldots, X(t_n)]'$ and $V = [X(t_1 + h), \ldots, X(t_n + h)]'$ have the same mean vector [which is $(\mu, \ldots, \mu)$, where $\mu = \mathrm{E}\{X(0)\}$]. Furthermore, the weak stationarity property implies that $\mathrm{Var}(U) = \mathrm{Var}(V)$ (see Section A.2.4). Thus, by the properties of multivariate normal distribution (see Section A.3.1), U and V have the same joint distribution. In other words, the Gaussian process is strongly stationary.

Example A.8 (Markov chains). Let $X(t), t \geq 0$, be an irreducible Markov chain taking values in a countable subset S of R and with a unique stationary distribution π (see Section 10.2). The finite-dimensional distributions of the process depend on the initial distribution p_0 of $X(0)$, and it is not generally true that $X(t), t \geq 0$, is stationary in either sense. However, if $p_0 = \pi$—that is, the initial distribution is the same as the stationary distribution—then the distribution p_t of $X(t)$ satisfies

$$\begin{aligned} p_t(j) &= \mathrm{P}\{X(t) = j) \\ &= \sum_{i \in S} \mathrm{P}\{X(0) = i\} \mathrm{P}\{X(t) = j | X(0) = i\} \\ &= \sum_{i \in S} p_0(i) p^{(t)}(i, j) \\ &= \sum_{i \in S} \pi(i) p^{(t)}(i, j) \quad = \quad \pi(j), \quad j \in S, \end{aligned}$$

the last identity implied by (10.7) and (10.17). In other words, the distribution of $X(t)$ does not depend on t. By a similar argument, it can be shown that the joint distribution of $X(t_1 + h), \ldots, X(t_n + h)$ does not depend on h for every $n > 1$. Thus, the process $X(t), t \geq 0$, is strongly stationary.

A fundamental theory associated with weak stationary processes is called the spectral theorem, or spectral representation. Define the *autocovariance function* of a weakly stationary process $X(t), -\infty < t < \infty$, by

$$c(t) = \mathrm{cov}\{X(s), X(s + t)\}$$

for any $s, t \in R$. Thus, in particular, $c(0) = \mathrm{var}\{X(t)\}$. The *autocorrelation function* is defined as $\rho(t) = c(t)/c(0), t \in R$. The spectral theorem for autocorrelation functions states that if $c(0) > 0$ and $\rho(t)$ is continuous at $t = 0$, then $\rho(t)$ is the cf (see Section 2.4) of some distribution F; that is,

$$\rho(t) = \int_{-\infty}^{\infty} e^{it\lambda} \, dF(\lambda), \quad t \in R.$$

The distribution F is called the *spectral distribution function* of the process. If $X(n), n = 0, \pm 1, \ldots$, is a discrete-time process such that $\sum_{n=-\infty}^{\infty} |\rho(n)| < \infty$, then F has a density f, called the *spectral density function*, given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \rho(n), \quad \lambda \in [-\pi, \pi].$$

Not only does the autocorrelation function of a weakly stationary process have the spectral representation, but the process itself also enjoys a nice spectral representation. Suppose that $X(t), -\infty < t < \infty$, is weakly stationary with $\mathrm{E}\{X(t)\} = 0$ and that $\rho(t)$ is continuous. Then there exists a complex-valued process $S(\lambda), -\infty < \lambda < \infty$, such that

$$X(t) = \int_{-\infty}^{\infty} e^{it\lambda} \, dS(\lambda), \quad -\infty < t < \infty$$

(see Section 10.6 for the definition of a stochastic integral). The process S is called the *spectral process* of X.

As for strongly stationary processes, a well-known result is the ergodic theorem. This may be regarded as an extension of the SLLN. The theorem is usually stated for a discrete-time process, $X_n, n = 1, 2, \ldots$. If the latter is strongly stationary such that $\mathrm{E}(|X_1|) < \infty$, then there is a random variable, Y, with the same mean as the X's such that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} Y$$

and $\mathrm{E}(X_n) \to \mathrm{E}(Y)$ as $n \to \infty$. The random variable Y can be expressed as a conditional expectation. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A measurable map $T\colon \Omega \to \Omega$ is said to be measure-preserving if $P(T^{-1}A) = P(A)$ for $A \in \mathcal{F}$, where $T^{-1}A = \{\omega \in \Omega, T(\omega) \in A\}$. Any stationary process $X_n, n = 0, 1, \ldots$, can be thought of as being generated by a measure-preserving transformation T in the sense that there exists a random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$ and a map $T\colon \Omega \to \Omega$ such that the process $XT^n, n \geq 0$, has the same joint distribution as $X_n, n \geq 0$, where $XT^n(\omega) = X\{T^n(\omega)\}$, $\omega \in \Omega$, and $XT^0 = X$. The process $X_n, n \geq 0$, is said to be *ergodic* if the transformation T satisfies the following: For any $A \in \mathcal{F}$, $T^{-1}(A) = A$ implies $P(A) = 0$ or 1. The ergodic theorem can now be restated as that if T is measure-preserving and $\mathrm{E}(|X|) < \infty$, then

$$\frac{1}{n}\sum_{i=0}^{n-1} XT^i \xrightarrow{\text{a.s.}} \mathrm{E}(X|\mathcal{I}),$$

where $\mathcal{I}$ is the invariant σ-field defined as $\mathcal{I} = \{A \in \mathcal{F} : T^{-1}A = A\}$. In particular, if the process $X_n, n \geq 0$, is ergodic, then $Y = \mathrm{E}(X|\mathcal{I}) = \mathrm{E}(X)$.

A.4 List of notation and abbreviations

The list is in alphabetical order, although the actual letters that appear in different places in the text may be different:

$a \wedge b$: $= \min(a, b)$.

$a \vee b$: $= \max(a, b)$.

a.s.: almost surely.

a': the transpose of vector a.

$\dim(a)$: the dimension of vector a.

$A \leq B$, where A and B are symmetric matrices: This means $B - A$ is nonnegative definite.

$A < B$, where A and B are symmetric matrices: This means $B - A$ is positive definite.

A^c: the complement of set A.

$|A|$: the determinant of matrix A.

A': the transpose of matrix A.

$\lambda_{\min}(A)$: the smallest eigenvalue of matrix A.

$\lambda_{\max}(A)$: the largest eigenvalue of matrix A.

$\mathrm{tr}(A)$: the trace of matrix A.

$\|A\|$: the spectral norm of matrix A defined as $\|A\| = \{\lambda_{\max}(A'A)\}^{1/2}$.

$\|A\|_2$: the 2-norm of matrix A defined as $\|A\|_2 = \{\mathrm{tr}(A'A)\}^{1/2}$.

$\mathrm{rank}(A)$: the (column) rank of matrix A.

$A^{1/2}$: the square root of a nonnegative definite matrix A (see Section A.1.6).

If A is a set, $|A|$ represents the cardinality of A.

ACR: autocorrelation.

ACV: autocovariance.

AIC: Akaike's information criterion.

ANOVA: analysis of variance.

AR: autoregressive process.

ARMA: autoregressive moving average process.

ARE: asymptotic relative efficiency.

$a_n = O(b_n)$: This means that the sequence $a_n/b_n, n = 1, 2, \ldots$, is bounded.

$a_n = o(b_n)$: This means that the sequence $a_n/b_n \to 0$ as $n \to \infty$.

$a_n \sim b_n$, where both a_n and b_n are sequences of real numbers: This means $a_n/b_n \to 1$ as $n \to \infty$.

AU: asymptotically unbiased.

$\mathcal{B}$: the Borel σ-field.

BIC: Bayesian information criterion.

BLUE: best linear unbiased estimator.

BLUP: best linear unbiased predictor.

BP: best predictor.

$\mathcal{C}$: the space of continuous functions with the uniform metric.

cdf: cumulative distribution function.

cf: characteristic function.

CLT: central limit theorem.

C_k^m: the binomial coefficient equal to the number of ways of choosing k items from m items without considering the order; this is also denoted by $\binom{m}{k}$.

$\mathrm{Cov}(\xi, \eta)$: the covariance matrix between random vectors ξ and η (see Section A.2.4).

$\xrightarrow{\text{a.s.}}$: almost sure convergence.

$\xrightarrow{\text{d}}$: convergence in distribution.

$\xrightarrow{\text{P}}$: convergence in probability.

$\xrightarrow{L^p}$: convergence in L^p.

c.u.: continuous uniformly.

$\mathcal{D}$: the space of functions that are right continuous and possess left-limit at each point, with the uniform metric.

$\mathrm{diag}(A)$: for A being a square matrix, this is the vector of diagonal elements of A.

diag$(A_1, \ldots, A_k)$: the block-diagonal matrix with $A_1, \ldots, A_k$ on its diagonal; the definition also includes the diagonal matrix, when $A_1, \ldots, A_k$ are numbers.

Distributions: Binomial(n, p) — binomial distribution with n independent trials and probability p of success for each trial; Cauchy(μ, σ) — Cauchy distribution with pdf $f(x|\mu, \sigma) = (\pi\sigma[1+\{(x-\mu)/\sigma\}^2])^{-1}$, $-\infty < x < \infty$; DE(μ, σ) — double exponential distribution with pdf $f(x|\mu, \sigma) = (2\sigma)^{-1} \exp(-|x - \mu|/\sigma)$, $-\infty < x < \infty$; χ^2_ν — χ^2-distribution with ν degrees of freedom; Exponential(λ) — exponential distribution with mean λ; $N(\mu, \sigma^2)$ — normal distribution with mean μ and variance σ^2, or $N(\mu, \Sigma)$ — multivariate normal distribution with mean vector μ and covariance matrix Σ; NM(p, τ) — normal mixture distribution with cdf $(1 - p)\phi(x) + p\Phi(x/\tau)$, where Φ is the cdf of $N(0, 1)$; Poisson(λ) — Poisson distribution with mean λ; t_ν — t-distribution with ν degrees of freedom; Uniform$[a, b]$ — uniform distribution over $[a, b]$.

∇: the gradient operator.

$\delta_x(y)$: the Dirac (or point) mass at x, which $= 1$ if $y = x$ and 0 otherwise.

E_θ: This notation is often used for expectation under the distribution with θ being the true parameter (vector).

E_M: This notation is sometimes used for model-based expectation; or expectation under model M.

E_d: This notation is sometimes used for design-based expectation.

$\mathrm{E}(\xi|\eta)$: conditional expectation of ξ given η.

EBLUE: empirical best linear unbiased estimator.

EBLUP: empirical best linear unbiased predictor.

EBP: empirical best predictor.

EM: Expectation–Maximization (algorithm).

$\emptyset$: empty set.

E_θ: expectation when θ is the true parameter (vector).

$f \circ g$: $f \circ g(x) = f(g(x))$ for functions f, g.

$F^{-1}(t)$: If F is a cdf, this is defined as $\inf\{x : F(x) \geq t\}$.

$f(x) = O\{g(x)\}$: This means $f(x)/g(x)$ is bounded for all x.

$f(x) = o\{g(x)\}$: This means $f(x)/g(x) \to 0$ as $x \to \infty$ (or $x \to 0$).

$f(x) \sim g(x)$: This means $f(x)/g(x) \to 1$ as $x \to \infty$ (or $x \to 0$).

$f(x|y)$: the conditional density function.

$\mathcal{F}$: This notation is usually used for a σ-field.

$\mathcal{F}_n$: a sequence of σ-fields such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}, n = 1, 2, \ldots$.

F_n: This notation is often (but not always) used for the empirical distribution of observations $X_1, \ldots, X_n$.

$F'_-(x)$ ($F'_+(x)$): the left (right) derivative of F at x.

$\Gamma(\cdot)$: the gamma function.

GLM: generalized linear model.

GLMM: generalized linear mixed model.

HQ: Hannan–Quinn criterion.

i: in the definition of cf (see above), for example, this represents $\sqrt{-1}$.

iff: if and only if.
i.i.d.: independent and identically distributed.
inf: infimum.
I_n: the n-dimensional identity matrix.
$I(\theta)$ (or $\mathcal{I}(\theta)$): the Fisher information (matrix).
J_n: the $n \times n$ matrix of 1's, or $J_n = 1_n 1_n'$ (see below).
$\bar{J}_n$: $= n^{-1} J_n$.
κ_3: the skewness (parameter).
κ_4: the kurtosis (parameter).
$\mathcal{L}(A)$: the linear space spanned by the columns of matrix A.
LIL: law of the iterated logarithm.
log: logarithm of base e, or natural logarithm.
logit: the logit function defined as $\text{logit}(p) = \log\{p/(1-p)\}$, $0 < p < 1$.
L^p: the L^p space of functions or random variables.
$\liminf x_n$: the smallest limit point of the sequence x_n.
$\limsup x_n$: the largest limit point of the sequence x_n.
$\limsup A_n$, where $A_1, A_2, \ldots$ is a sequence of events: This is defined as $\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_n$.
LLN: law of large numbers.
LSE: least squares estimator.
MA: moving average process.
MC: Markov chain.
MCEM: Monte Carlo EM (algorithm).
MCMC: Markov-chain Monte Carlo.
M_{f}: This notation is often used to denote a full model.
mgf: moment generating function.
MINQUE: minimum norm quadratic unbiased estimation.
ML: maximum likelihood.
MLE: maximum likelihood estimator.
MM: method of moments.
M_{opt}: This notation is often used to denote an optimal model.
MSA: for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, $\text{MSA} = \text{SSA}/(k-1)$.
MSE: mean squared error (or, see below).
MSE: for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, $\text{MSE} = \text{SSE}/m(k-1)$.
MSM: method of simulated moments.
MSPE: mean squared prediction error.
$\#A$, where A is a set: This represents the cardinality of set A.
$N(\mu, \Sigma)$: The multivariate normal distribution with mean vector μ and covariance matrix Σ.
OLS: ordinary least squares.
1_A, where A is an event: This represents the indicator of event A.
1_n, where n is a positive integer: the n-dimensional vector of 1s.
1_n^0: $= I_n$.
1_n^1: $= 1_n$.
Ω: This usually represents a probability space.

$O_{\mathrm{P}}, o_{\mathrm{P}}$: big O and small o in probability (see Section 3.4).

$\otimes$: Kronecker product.

$\mathrm{P}(A|B)$: conditional probability of A given B.

P_A: the projection matrix to $\mathcal{L}(A)$ defined as $P_A = A(A'A)^- A'$, where A^- is the generalized inverse of A (see §A.1.2).

$P_{A^\perp}$: the projection matrix with respect to the linear space orthogonal to $\mathcal{L}(A)$, defined as $P_A = I - P_A$, where I is the identity matrix.

∂A, the boundary of set A.

$\partial\xi/\partial\eta'$: When $\xi = (\xi_i)_{1\le i\le a}$, $\eta = (\eta_j)_{1\le j\le b}$, this notation means the matrix $(\partial\xi_i/\partial\eta_j)_{1\le i\le a, 1\le j\le b}$.

$\partial^2\xi/\partial\eta\partial\eta'$: When ξ is a scalar, $\eta = (\eta_j)_{1\le j\le b}$, this notation means the matrix $(\partial^2\xi/\partial\eta_j\partial\eta_k)_{1\le j,k\le b}$.

pdf: probability density function.

pmf: probability mass function.

PQL: penalized quasi-likelihood.

$\propto$: proportional to.

r.c.: relatively compact.

R^d: the d-dimensional Euclidean space; in particular, $R^1 = R$ represents the real line.

REML: restricted maximum likelihood.

RSS: residual sum of squares.

s.d.: standard deviation.

SDE: stochastic differential equation.

$S_\delta(a)$: the δ-neighborhood of a; that is, $\{x : |x - a| < \delta\}$.

SLLN: strong law of large numbers.

SSA: for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, $\mathrm{SSA} = k\sum_{i=1}^m (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

SSE: for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, $\mathrm{SSE} = \sum_{i=1}^m \sum_{j=1}^k (y_{ij} - y_{i\cdot})^2$.

sup: supremum.

TMD: two-parameter martingale differences.

$\mathrm{Var}(\xi)$: covariance matrix of the random vector ξ.

var_θ: variance when θ is the true parameter (vector).

WLLN: weak law of large numbers.

WLS: weighted least squares.

WN: white noise process.

w.r.t.: with respect to.

$[x]$ for real number x: This is the largest integer less than or equal to x.

$|x|$ for $x \in R^d$: This is defined as $(\sum_{i=1}^d x_i^2)^{1/2}$, where x_i is the ith component of x, $1 \le i \le d$.

x^+: defined as x is $x > 0$ and 0 otherwise.

x^-: defined as $-x$ if $x < 0$ and 0 otherwise.

$\bar{X}$: the sample mean of $X_1, \ldots, X_n$.

$X_{(i)}$: the ith order statistic of $X_1, \ldots, X_n$ such that $X_{(1)} \le \cdots \le X_{(n)}$.

$\mathrm{Var}(\xi)$: the covariance matrix of random vector ξ (see §A.2.4).

$(X_i)_{1\le i\le m}$: When $X_1,\ldots,X_m$ are matrices with the same number of columns, this is the matrix that combines the rows of $X_1,\ldots,X_m$, one after the other.

$(y_i)_{1\le i\le m}$: When $y_1,\ldots,y_m$ are column vectors, this notation means the column vector $(y_1',\ldots,y_m')'$.

$(y_{ij})_{1\le i\le m,1\le j\le n_i}$: In the case of clustered data, where y_{ij}, $j=1,\ldots,n_i$, denote the observations from the ith cluster, this notation represents the vector $(y_{11},\ldots,y_{1n_1},y_{21},\ldots,y_{2n_2},\ldots,y_{m1},\ldots,y_{mn_m})'$.

$y_{i\cdot}$, $\bar{y}_{i\cdot}$, $y_{\cdot j}$, $\bar{y}_{\cdot j}$, $y_{\cdot\cdot}$ and $\bar{y}_{\cdot\cdot}$: In the case of clustered data y_{ij}, $i=1,\ldots,m$, $j=1,\ldots,n_i$, $y_{i\cdot}=\sum_{j=1}^{n_i}y_{ij}$, $\bar{y}_{i\cdot}=n_i^{-1}y_{i\cdot}$, $y_{\cdot\cdot}=\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{ij}$, $\bar{y}_{\cdot\cdot}=(\sum_{i=1}^{m}n_i)^{-1}y_{\cdot\cdot}$; in the case of balanced data y_{ij}, $1\le i\le a$, $j=1,\ldots,b$, $y_{i\cdot}=\sum_{j=1}^{b}y_{ij}$, $\bar{y}_{i\cdot}=b^{-1}y_{i\cdot}$, $y_{\cdot j}=\sum_{i=1}^{a}y_{ij}$, $\bar{y}_{\cdot j}=a^{-1}y_{\cdot j}$, $y_{\cdot\cdot}=\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij}$, $\bar{y}_{\cdot\cdot}=(ab)^{-1}y_{\cdot\cdot}$.

$y|\eta\sim$: the distribution of y given η is ...; note that here η may represent a vector of parameters or random variables, or a combination of both.

Y-W: Yule–Walker.

References

Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki eds.), Akademiai Kiado, Budapest, 267–281.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automatic Control* 19, 716–723.

An Hong-Zhi, Chen Zhao-Guo, and Hannan, E. J. (1982), Autocorrelation, autoregression and autoregressive approximation, *Ann. Statist.* 10, 926–936.

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.

Anderson, T. W., and Darling, D. A. (1954), A test of goodness of fit, *J. Amer. Statist. Assoc.* 49, 765–769.

Arnold, L. (1974), *Stochastic Differential Equations: Theory and Applications*, Wiley, New York.

Atwood, L. D., Wilson, A. F., Bailey-Wilson, J. E., Carruth, J. N., and Elston, R. C. (1996), On the distribution of the likelihood ratio test statistic for a mixture of two normal distributions, *Commun. Statist. - Simulation* 25, 733–740.

Bachman, G., Narici, L., and Beckenstein, E. (2000), *Fourier and Wavelet Analysis*, Springer, New York.

Bamber, D. (1975), The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *J. Math. Psych.* 12, 387–415.

Barndorff-Nielsen, O. (1983), On a formula for the distribution of the maximum likelihood estimator, *Biometrika* 70, 343–365.

Barndorff-Nielsen, O. E., and Cox, D. R. (1989), *Asymptotic Techniques for Use in Statistics*, Chapman & Hall, London.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data, *J. Amer. Statist. Assoc.* 80, 28–36.

Beran, R. (1984), Bootstrap methods in statistics, *Jber. Dt. Math. Verein.* 86, 24–30.

Bernstein, S. N. (1937), On several modifications of Chebyshev's inequality, *Doklady Akad. Nauk SSSR* 17, 275–277.

Berry, A. C. (1941), The accuracy of the Gaussian approximation to the sum of independent variates, *Trans. Amer. Math. Soc.* 49, 122–136.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statist. Soc. B* 36, 192–236.

Bickel, P. J. (1966), Some contributions to the theory of order statistics, *Proc. 5th Berkeley Symp. Math. Statist. Probab.* 1, 575–592.

Bickel, P. J. (1967), Some contributions to the theory of order statistics, *Proc. 5th Berkeley Symp. Math. Statist. Probab.* 1, 575–591.

Bickel, P. J., and Bühlmann, P. (1997), Closure of linear processes, *J. Theoret. Probab.* 10, 445–479.

Bickel, P. J., and Freedman, D. A. (1981), Some asymptotic theory for the bootstrap, *Ann. Statist.* 9, 1196–1217.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

Billingsley, P. (1968), *Convergence of Probability Measure*, Wiley, New York.

Billingsley, P. (1995), *Probability and Measure*, 3nd ed., Wiley, New York.

Birnbaum, Z. W. and McCarty, R. C. (1958), A distribution-free upper confidence bound for $\mathrm{P}(Y < X)$ based on independent samples of X and Y, *Ann. Math. Statist.* 29, 558–562.

Black, F., and Sholes, M. (1973), The pricing of options and corporate liabilities, *J. Politi. Econ.* 81, 637–659.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *J. Econometrics* 31, 307–327.

Booth, J. G., and Hobert, J. P. (1999), Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm, *J. Roy. Statist. Soc. B* 61, 265–285.

Bozdogan, H. (1994), Editor's general preface, in *Engineering and Scientific Applications*, Vol. 3 (H. Bozdogan ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, ix–xii, Kluwer Academic Publishers, Dordrecht.

Brackstone, G. J. (1987), Small area data: policy issues and technical challenges, in *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh eds.) 3–20, Wiley, New York.

Breslow, N. E., and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9–25.

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, 2nd ed., Springer, New York.

Brown, L. D., Wang, Y., and Zhao, L. H. (2003), Statistical equivalence at suitable frequencies of GARCH and stochastic volatility models with the corresponding diffusion model, *Statist. Sinica* 13, 993–1013.

Bühlmann, P. (1997), Sieve bootstrap for time series, *Bernoulli* 3, 123–148.

Bühlmann, P. (2002), Bootstraps for time series, *Statist. Sci.* 17, 52–72.

Burkholder, D. L. (1966), Martingale transforms, *Ann. Math. Statist.* 37, 1494–1504.

Calvin, J. A., and Sedransk, J. (1991), Bayesian and frequentist predictive inference for the patterns of care studies, *J. Amer. Statist. Assoc.* 86, 36–48.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, 2nd ed., Duxbury, Thomson Learning, Pacific Grove, CA.

Casella, G., and George, E. I. (1992), Explaining the Gibbs sampler, *Amer. Statist.* 46, 167–174.

Chan, N. N., and Kwong, M. K. (1985), Hermitian matrix inequalities and a conjecture, *Amer. Math. Monthly* 92, 533–541.

Chatterjee, S., Lahiri, P., and Li, H. (2008), Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models, *Ann. Statist.* 36, 1221–1245.

Chernoff, H. and Lehmann, E. L. (1954), The use of maximum-likelihood estimates in χ^2 tests for goodness of fit, *Ann. Math. Statist.* 25, 579–586.

Chiang, T.-P. (1987), On Markov models of random fields, *Acta Math. Appl. Sinica* (English) 3, 328–341.

Chiang, T.-P. (1991), Stationary random field: Prediction theory, Markov models, limit theorems, *Contemp. Math.* 118, 79–101.

Chow, Y. S. (1960), A martingale inequality and the law of large numbers, *Proc. Amer. Math. Soc.* 11, 107–111.

Chow, Y. S. (1965), Local convergence of martingales and the law of large numbers, *Ann. Math. Statist.* 36, 552–558.

Chow, Y. S., and Teicher, H. (1988), *Probability Theory*, Springer, New York.

Chung, K.-L. (1948), On the maximum partial sums of sequence of independent random variables, *Trans. Amer. Math. Soc.* 64, 205–233.

Chung, K.-L. (1949), An estimate concerning the Kolmogoroff limit distribution, *Trans. Amer. Math. Soc.* 67, 36–50.

Claeskens, G., and Hart, J. D. (2009), Goodness-of-fit tests in mixed models (with discussion), *TEST* 18, 213–239.

Cox, D. R. (1975), Prediction intervals and empirical Bayes confidence intervals, in *Perspectives in Probability and Statistics, Papers in Honor of M. S. Bartlett*

(J. Gani, ed.), Applied Probability Trust, University of Sheffield, Sheffield, 47–55.

Cramér, H. (1936), Über eine Eigenschaft der normalen Verteilungsfunction, *Math. Zeitschr.* 41, 405–414.

Cramér, H. (1946), *Mathematical Mathods of Statistics*, Princeton University Press, Princeton, N.J.

Csörgő, M., and Révész, R. (1975), A new method to prove Strassen-type laws of invariance principle, I and II, *Z. Wahrsch. verw. Geb.* 31, 255–269.

Das, K., Jiang, J., and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818–840.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York.

Datta, G. S., and Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statist. Sinica* 10, 613–627.

Datta, G. S., and Lahiri, P. (2001), Discussions on a paper by Efron and Gous, *Model Selection*, IMS Lecture Notes/Monograph 38, P. Lahiri ed., Institute of Mathematical Statistics, Beachwood, OH.

Datta, G. S., Kubokawa, T., Rao, J. N. K., and Molina, I. (2009), Estimation of mean squared error of model-based small area estimators, *TEST*, to appear.

Datta, G. S., Rao, J. N. K., and Smith, D. D. (2005), On measuring the variability of small area estimators under a basic area level model, *Biometrika* 92, 183–196.

David, H. A. and Nagaraja, H. N. (2003), *Order Statistics*, 3rd ed., Wiley, New York.

De Bruijn, N. G. (1961), *Asymptotic Methods in Analysis*, North-Holland, Amsterdam.

Dehling, H., Mikosch, T., and Sørensen, M. (2002), *Empirical Process Techniques for Dependent Data*, Birkhäuser, Boston.

Dehling, H., and Philipp, W. (2002), Empirical process techniques for dependent data, in *Empirical Process Techniques for Dependent Data* (H. Dehling, T. Mikosch and M. Sørensen eds.), Birkhäuser, Boston.

de Leeuw, J. (1992), Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle, in *Breakthroughs in Statistics* (S. Kotz and N. L. Johnson eds.), Springer, London, Vol. 1, 599–609.

Dempster, A., Laird, N., and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. B* 39, 1–38.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1996). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Donsker, M. (1951), An invariance principle for certain probability limit theorems, *Mem. Amer. Math. Soc.* 6, .

Donsker, M. (1952), Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems, *Ann. Math. Statist.* 23, 277-i-281.

Doob, J. L. (1949), Heuristic approach to the Kolmogorov-Smirnov theorems, *Ann. Math. Statist.* 20, 393–403.

Doob, J. L. (1953), *Stochastic Processes*, Wiley, New York.

Doob, J. L. (1960), Notes on martingale theory, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 2, 95–102.

Durbin, J. (1960), The fitting of time series models, *Int. Statist. Rev.* 28, 233–244.

Durrett, R. (1991), *Probability: Theory and Examples*, Wadsworth, Pacific Grove, CA.

Durrett, R. (1996), *Stochastic Calculus: A Practical Introduction*, CRC Press, Boca Raton, FL.

Dvoretzky, A., Erdös, P., and Kakutani, S. (1961), Nonincrease everywhere of the Brownian motion process, *Proc. Fourth Berkeley Symposium* II, 103–116.

Dvoretzky, A., Keifer, J., and Wolfowitz, J. (1956), Asymptotic minimax character of the sample distribution functions and of the classical multinomial estimator, *Ann. Math. Statist.* 27, 642–669.

Dynkin, E. B. (1957), Inhomogeneous strong Markov processes, *Dokl. Akad. Nauk SSSR* 113, 261–263.

Engle, R. F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* 50, 987–1007.

Efron, B. (1979), Bootstrap method: Another look at the jackknife, *Ann. Statist.* 7, 1–26.

Efron, B., and Morris, C. (1973), Stein's estimation rule and its competitors: An empirical Bayes approach, *J. Amer. Statist. Assoc.* 68, 117–130.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York.

Esseen, C. G. (1942), On the Liapounoff limit of error in the theory of probability, *Ark. Math. Astr. och Fysik* 28A, 1–19.

Faber, G. (1910), Über stetige Funktionen II, *Math. Ann.* 69, 372–443.

Fan, J., and Yao, Q. (2003), *Nonlinear Time Series*, Springer, New York.

Fay, R. E., and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269–277.

Federá Serio, G., Manara, A., and Sicoli, P. (2002), Giuseppe Piazzi and the Discovery of Ceres, in W. F. Bottke Jr., A. Cellino, P. Paolicchi, and R. P.

Binzel, *Asteroids III*, Tucson, Arizona: Univ. of Arizona Press, pp. 17–24, Retrieved 2009–06–25.

Feller, W. (1968), *An introduction to Probability Theory and Its Applications*, Wiley, New York, Vol. I.

Feller, W. (1971), *An introduction to Probability Theory and Its Applications*, Wiley, New York, Vol. II.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, London.

Finkelstein, H. (1971), The law of iterated logarithm for empirical distributions, *Ann. Math. Statist.* 42, 607–615.

Fisher, R. A. (1922a), On the interpretation of chi-square from contingency tables, and the calculation of P, *J. Roy. Statist. Soc.* 85, 87–94.

Fisher, R. A. (1922b), On the mathematical foundations of the theoretical statistics, *Phil. Trans. R. Soc.* 222, 309–368.

Foderà Serio, G., Manara, A., and Sicoli, P. (2002), Giuseppe Piazzi and the discovery of Ceres, in *Asteroids* III (W. F. Bottke Jr., A. Cellino, P. Paolicchi and R. P. Binzel, eds.), University of Arizona Press, Tucson, 17–24.

Forrester, J., and Ury, H. K. (1969), The signed-rank (Wilcoxon) test in the rapid analysis of biological data, *Lancet* 1, 239–241.

Fox, R., and Taqqu, M. S. (1985), Noncentral limit theorems for quadractic forms in random variables having long-range dependence, *Ann. Probab.* 13, 428–446.

Freedman, D. (1984), On bootstrapping two-stage least squares estimates in stationary linear models, *Ann. Statist.* 12, 827–842.

Fuller, W. A. (1989), Prediction of true values for the measurement error model, in *Conference on Statistical Analysis of Measurement Error Models and Applications*, Humbolt State University, Arcata, CA.

Gelfand, A. E., and Smith, A. F. M. (1990), Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.* 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman & Hall London.

Geman, S., and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Patterns. Anal. Mach. Intell.* 6, 721–741.

Ghosh, M., and Rao, J. N. K. (1994), Small area estimation: An appraisal (with discussion), *Statist. Sci.* 9, 55–93.

Goethe, J. W. von (1808), *Faust*, Part One.

Grenander, U. (1981), *Abstract Inference*, Wiley, New York.

Green, P. J. (1987), Penalized likelihood for general semi-parametric regression

models, *Int. Statist. Rew.* 55, 245–259.

Guttorp, P., and Lockhart, R. A. (1988), On the asymptotic distribution of quadratic forms in uniform order statistics, *Ann. Statist.* 16, 433–449.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer, New York.

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and Its Application*, Academic Press, New York.

Hall, P., and Maiti, T. (2006), Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Ann. Statist.* 34, 1733–1750.

Hannan, E. J. (1970), *Multiple Time Series*, Wiley, New York.

Hannan, E. J. (1980), The estimation of the order of an ARMA process, *Ann. Statist.* 8, 1071–1081.

Hannan, E. J., and Heyde, C. C. (1972), On limit theorems for quadratic functions of discrete time series, *Ann. Math. Statist.* 43, 2058–2066.

Hannan, E. J., and Quinn, B. G. (1979), The determination of the order of an autoregression, *J. Roy. Statist. Soc. B* 41, 190–195.

Hanna, E. J., and Rissanen, J. (1982), Recursive estimation of mixed autoregressive-moving average order, *Biometrika* 69, 81–94.

Hardy, G., Littlewood, J. E. and Pólya, G. (1934), *Inequalities*, Cambridge Univ. Press.

Hartigan, J. A. (1985), A failure of likelihood asymptotics for normal mixtures, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen eds.), Vol. II, 807–810, Wadsworth, Belmont, CA.

Hartley, H. O., and Rao, J. N. K. (1967), Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika* 54, 93–108.

Hartman, P., and Wintner, A. (1941), On the law of the iterated logarithm, *Amer. J. Math.* 63, 169–176.

Harville, D. A. (1974), Bayesian inference for variance components using only error contrasts, *Biometrika* 61, 383–385.

Harville, D. A. (1977), maximum likelihood approaches to variance components estimation and related problems, *J. Amer. Statist. Assoc.* 72, 320–340.

Harville, D. A. (1985), Decomposition of prediction error, *J. Amer. Statist. Assoc.* 80, 132–138.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* 57, 97–109.

Henderson, C. R. (1953), Estimation of variance and covariance components, *Biometrics* 9, 226–252.

Heyde, C. C. (1994), A quasi-likelihood approach to the REML estimating equations, *Statist. Probab. Lett.* 21, 381–384.

Heyde, C. C. (1997), *Quasi-Likelihood and Its Application*, Springer, New York.

Heyde, C. C., and Scott, D. J. (1973), Invariance principle for the law of the iterated logarithm for martingales and processes with stationary increments, *Ann. Probab.* 1, 428–436.

Hinkley, D. V. (1977), Jackknifing in unbalanced situations, *Technometrics* 19, 285–292.

Hjort, N. L., and Jones, M. C. (1996), Locally parametric nonparametric density estimation, *Ann. Statist.* 24, 1619–1647.

Hoeffding, W. (1956), On the distribution of the number of successes in independent trials, *Ann. Math. Statist.* 27, 713–721.

Hoeffding, W. (1961), The strong law of large numbers for U-statistics, *Inst. Mimeo Ser.* 302, 1–10.

Hsieh, F., and Turnbull, B. W. (1996), Non-parametric and semi-parametric estimation of the receiver operating characteristic curve, *Ann. Statist.* 24, 25–40.

Hu, I. (1985), A uniform bound for the tail probability of Kolmogorov-Smirnov statistics, *Ann. Statist.* 13, 821–826.

Huang, D. (1988a), Convergence rate of sample autocorrelations and autocovariances for stationary time series, *Sci. Sinica* XXXI, 406–424.

Huang, D. (1988b), Recursive method for ARMA model estimation (I), *Acta Math. Appl. Sinica* 4, 169–192.

Huang, D. (1989), Recursive method for ARMA model estimation (II), *Acta Math. Appl. Sinica* 5, 332–354.

Huang, D. (1992), Central limit theorem for two-parameter martingale differences with application to stationary random fields, *Sci. Sinica* XXXV, 413–425.

Hunt, G. (1956), Some theorems concerning Brownian motion, *Trans. Amer. Math. Soc.* 81, 294–319.

Hurvich, C. M., and Tsai, C. L. (1989), Regression and time series model selection in small samples, *Biometrika* 76, 297–307.

Ibragimov, I. A., and Linnik, Y. V. (1971), *Independent and Stationary Sequence of Random Variables*, Wolters-Noordhoff, Groningen.

James, F. (2006), *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific Singapore.

Jiang, J. (1989), Uniform convergence rate of sample autocovariances and autocorrelations for linear spatial series, *Adv. Math. (Chinese)* 18, 497–499.

Jiang, J. (1991a), Uniform convergence rate of sample ACV and ACR for linear spatial series under more general martingale condition, *Adv. Math. (Chinese)* 20, 39–50.

Jiang, J. (1991b), Parameter estimation of spatial AR model, *Chinese Ann.*

Math. 12B, 432–444.

Jiang, J. (1993), Estimation of spatial AR models, *Acta Math. Appl. Sinica* 9, 174–187.

Jiang, J. (1996), REML estimation: Asymptotic behavior and related topics, *Ann. Statist.* 24, 255–286.

Jiang, J. (1997a), Wald consistency and the method of sieves in REML estimation, *Ann. Statist.* 25, 1781–1803.

Jiang, J. (1997b), Sharp upper and lower bounds for asymptotic levels of some statistical tests, *Statist. Probab. Lett.* 35, 395–400.

Jiang, J. (1998a), Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.* 93, 720–729.

Jiang, J. (1998b), On unbiasedness of the empirical BLUE and BLUP, *Statist. Probab. Lett.* 41, 19–24.

Jiang, J. (1998c), Asymptotic properties of the empirical BLUP and BLUE in mixed linear models, *Statist. Sinica* 8, 861–885.

Jiang, J. (1999a), Some laws of the iterated logarithm for two parameter martingales, *J. Theoret. Probab.* 12, 49–74.

Jiang, J. (1999b), On maximum hierarchical likelihood estimators, *Commun. Statist.: Theory Methods* 28, 1769–1776.

Jiang, J. (2000a), A matrix inequality and its statistical application, *Linear Algebra Applic.* 307, 131–144.

Jiang, J. (2000b), A nonlinear Gauss-Seidel algorithm for inference about GLMM, *Comput. Statist.* 15, 229–241.

Jiang, J. (2001), On actual significance levels of combined tests, *Nonparametric Statist.* 13, 763–774.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

Jiang, J., Jia, H., and Chen, H. (2001), Maximum posterior estimation of random effects in generalized linear mixed models, *Statist. Sinica* 11, 97–120.

Jiang, J., and Lahiri, P. (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.* 53, 217–243.

Jiang, J., and Lahiri, P. (2006), Mixed model prediction and small area estimation (with discussion), *TEST* 15, 1–96.

Jiang, J., Lahiri, P., and Wan, S. (2002a), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782–1810.

Jiang, J., Lahiri, P., and Wan, S. (2002b), Jackknifing the mean squared error of empirical best predictor: A theoretical synthesis, Tech. Report, Dept. Statistics, University of California, Davis, CA.

Jiang, J., Lahiri, P., and Wu, C.-H. (2001), A generalization of the Pearson's

χ^2 goodness-of-fit test with estimated cell frequencies, *Sankhyā, Ser. A* 63, 260–276.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008), Fence methods for mixed model selection, *Ann. Statist.* 36, 1669–1692.

Jiang, J., Nguyen, T., and Rao, J. S. (2009), A simplified adaptive fence procedure, *Statist. Probab. Lett.* 79, 625–629.

Jiang, J., and Zhang, W. (2001), Robust estimation in generalized linear mixed models, *Biometrika* 88, 753–765.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.* 91, 401–407.

Kackar, R. N., and Harville, D. A. (1981), Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Commun. Statist.: Theory Methods* 10, 1249–1261.

Kackar, R. N., and Harville, D. A. (1984), Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *J. Amer. Statist. Assoc.* 79, 853–862.

Karim, M. R., and Zeger, S. L. (1992), Generalized linear models with random effects: Salamander mating revisited, *Biometrics* 48, 631–644.

Katz, M. (1968), A note on the weak law of large numbers, *Ann. Math. Statist.* 39, 1348– 1349.

Khan, L. A., and Thaheem, A. B. (2000), On the equivalence of the Heine-Borel and the Bolzano-Weierstrass theorems, *Int. J. Math. Edu. Sci. Technol.* 31, 620–622.

Khintchine, A. (1924), Über einen Satz der Wahrscheinlichkeitsrechnung, *Fund. Math.* 6, 9–20.

Kolmogorov, A. (1929), Über das Gesetz des iterierten Logarithmus, *Math. Ann.* 101, 126–135.

Koroljuk, V. S., and Borovskich, Yu. V. (1994), *Theory of U-Statistics*, Kluwer, Dordrecht, The Netherlands.

Kosorok, M. P. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer, New York.

Künsch, H. R. (1989), The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* 17, 1217–1241.

Kutoyants, Y. A. (2004), *Statistical Inference for Ergodic Diffusion Processes*, Springer, London.

Łagodowski, Z. A., and Rychlik, Z. (1986), Rate of convergence in the strong law of large numbers for martingales, *Probab. Theory Rel. Fields* 71, 467–476.

Lai, T. L., Robbins, H., and Wei, C. Z. (1979), Strong consistency of least squares estimates in multiple regression II, *J. Multivariate Anal.* 9, 343–361.

Lai, T. L., and Wei, C. Z. (1982), A law of the iterated logarithm for double arrays of independent random variables with applications to regression and time series models, *Ann. Probab.* 10, 320–335.

Lai, T. L., and Wei, C. Z. (1984), Moment inequalities with applications to regression and time series models, *Inequalities in Statistics and Probability*, IMS Lecture Notes - Monograph Ser. 5, Institute of Mathematical Statistics, Hayward, CA, 165–172.

Laird, N. M., and Ware, J. M. (1982), Random effects models for longitudinal data, *Biometrics* 38, 963–974.

Lange, N., and Ryan, L. (1989), Assessing normality in random effects models, *Ann. Statist.* 17, 624–642.

Langyintuo, A., and Mekuria, M. (2008), Assessing the influence of neighborhood effects on the adoption of improved agricultural technologies in developing agriculture, *African J. Agric. Resource Econ.* 2, 151–169.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer, Berlin.

Le Cam, L., and Yang, G. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer, New York.

Lee, A. J. (1990), *U-Statistics*, Marcel Dekker, New York.

Lehmann, E. L. (1975), *Nonparametrics*, Holden-Day, San Francisco.

Lehmann, E. L. (1983), *Theory of Point Estimation*, Wiley, New York.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* 2nd. ed., Chapman & Hall, London.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer, New York.

Lichstein, J. W., Simons, T. R., Shriner, S. A., and Franzreb, K. E. (2002), Spatial autocorrelation and autoregressive models in ecology, *Ecol. Monogr.* 72, 445–463.

Lin, X. (1997), Variance components testing in generalized linear models with random effects, *Biometrika* 84, 309–326.

Lin, X., and Breslow, N. E. (1996), Bias correction in generalized linear mixed models with multiple components of dispersion, *J. Amer. Statist. Assoc.* 91, 1007–1016.

Liu, J. S. (1994), The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem, *J. Amer. Statist. Assoc.* 89, 958–966.

Liu, J. S., Wong, W. H., and Kong, A. (1994), Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes, *Biometrika* 81, 27–40.

Liu, J. S., Wong, W. H., and Kong, A. (1995), Covariance structure and convergence rate of the Gibbs sampler with various scans, *J. R. Statist. Soc. B* 57, 157–169.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), Small area inference for binary variables in the National Health Interview Survey, *J. Amer. Statist. Assoc.* 92, 815–826.

Mallows, C. L. (1972), A note on asymptotic joint normality, *Ann. Math. Statist.* 43, 508–515.

Manski, C. F., and McFadden, D. (1981), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA.

Marcinkiewicz, J., and Zygmund, A. (1937a), Sur les fonctions indepdendantes, *Fundam. Math.* 29, 60–90.

Marcinkiewicz, J., and Zygmund, A. (1937b), Remarque sur la loi du logarithme itéré, *Fundam. Math.* 29, 215–222.

Marron, J. S., and Nolan, D. (1988), Canonical kernels for density estimation, *Statist. Probab. Lett.* 7, 195–199.

Marshall, R. J. (1991), A review of methods for the statistical analysis of spatial patterns of disease, *J. Roy. Statist. Soc. A* 154, 421–441.

Mason, D. M., Shorack, G. R., and Wellner, J. A. (1983), Strong limit theorems for oscillation moduli of the uniform empirical process, *Z. Wahrsch. verw. Geb.* 65, 83–97.

Massart, P. (1990), The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality, *Ann. Probab.* 18, 1269–1283.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

McCulloch, C. E. (1994), Maximum likelihood variance components estimation for binary data, *J. Amer. Statist. Assoc.* 89, 330–335.

McCulloch, C. E. (1997), Maximum likelihood algorithms for generalized linear mixed models, *J. Amer. Statist. Assoc.* 92, 162–170.

McFadden, D. (1989), A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrika* 57, 995–1026.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), Equations of state calculations by fast computing machines, *J. Chem. Phys.* 21, 1087–1092.

Meza, J., and Lahiri, P. (2005), A note on the Cp statistic under the nested error regression model, *Survey Methodology* 31, 105–109.

Meza, J., Chen, S., and Lahiri, P. (2003), Estimation of lifetime alcohol abuse for Nebraska counties, unpublished manuscript.

Miller, J. J. (1977), Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance, *Ann. Statist.* 5, 746–762.

Miller, R. G. (1974), An unbalanced jackknife, *Ann. Statist.* 2, 880–891.

Mises, R. von (1947), On the asymptotic distribution of differentiable statistical

functions, *Ann. Math. Statist.* 18, 309–348.

Moeanaddin, R., and Tong, H. (1990), Numerical evaluation of distributions in nonlinear autoregression, *J. Time Series Anal.* 11, 33–48.

Moore, D. S. (1978), Chi-square tests, in *Studies in Statistics* (R. V. Hogg, ed.), Mathematical Society of America, Providence, RI.

Móricz, F. (1976), Moment inequalities and the strong laws of large numbers, *Z. Wahrsch. verw. Geb.* 35, 299–314.

Morris, C. N. (1983), Parametric empirical Bayes inference: theory and applications, *J. Amer. Statist. Assoc.* 78, 47–59.

Murray, G. D. (1977), Comment on "Maximum likelihood from incomplete data via the EM algorithm" by A. P. Dempster, N. Laird, and D. B. Rubin, *J. Roy. Statist. Soc. B* 39, 27–28.

Nelson, D. (1990), ARCH models as diffusion approximations, *J. Econometrics* 45, 7–38.

Neyman, J., and Scott, E. (1948), Consistent estimates based on partially consistent observations, *Econometrika* 16, 1–32.

Nishii, R. (1984), Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.* 12, 758–765.

Nummelin, E. (1984), *General Irreducible Markov Chains and Non-negatice Operators*, Cambridge University Press, Cambridge.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008), Non-parametric small area estimation using penalized spline regression, *J. R. Statist. Soc. B* 70, 265–286.

Owen, D. B. (1962), *Handbook of Statistical Tables*, Addison-Wesley, Reading, MA.

Paley, R. E. A. C., Wiener, N., and Zygmund, A. (1933), Notes on random functions, *Math. Z.* 37, 647–668.

Patterson, H. D., and Thompson, R. (1971), Recovery of interblock information when block sizes are unequal, *Biometrika* 58, 545–554.

Pearson, K. (1900), On a criterion that a given system of deviations from the probable in the case of a corrected system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philos. Mag. 5th Series* 50, 157–175.

Peng, L., and Zhou, X.-H. (2004), Local linear smoothing of receiver operating characteristic (ROC) curves, *J. Statist. Planning Inference* 118, 129–143.

Petrov, V. V. (1975), *Sums of Independent Random Variables*, Springer, Berlin.

Prasad, N. G. N., and Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163–171.

Quenouille, M. (1949), Approximation tests of correlation in time series, *J. Roy.*

Statist. Soc., Ser. B 11, 18–84.

Rao, C. R. (1972), Estimation of variance and covariance components in linear models, *J. Amer. Statist. Assoc.* 67, 112–115.

Rao, C. R., and Kleffe, J. (1988), *Estimation of Variance Components and Applications*, North-Holland, Amsterdam.

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley, Hoboken, NJ.

Rice, J. A. (1995), *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press, Belmont, CA.

Richardson, A. M., and Welsh, A. H. (1994), Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models, *Austral. J. Statist.* 36, 31–43.

Rivest, L.-P., and Belmonte, E. (2000), A conditional mean squared error of small area estimators, *Survey Methodology* 26, 67–78.

Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods*, 2nd ed., Springer, New York.

Roberts, G. O., and Smith, A. F. M. (1994), Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms, *Stoch. Proc. Appl.* 49, 207–216.

Roberts, G. O., and Tweedie, R. L. (1996), Geometric convergence and central limit theorem for multidimensional Hastings and Metropolic algorithms, *Biometrika* 83, 95–110.

Robinson, G. K. (1991), That BLUP is a good thing: The estimation of random effects (with discussion), *Statist. Sci.* 6, 15–51.

Rosenblatt, M. (1952), Limit theorems associated with variants of the von Mises statistic, *Ann. Math. Statist.* 23, 617–623.

Rosenthal, H. P. (1970), On the subspaces L^p $(p > 2)$ spanned by sequences of independent random variables, *Israel J. Math.* 8, 273–303.

Ross, S. M. (1983), *Stochastic Processes*, Wiley, New York.

Samuels, M. L., and Witmer, J. A. (2003), *Statistics for the Life Sciences*, 3rd ed., Pearson Education, Upper Saddle River, NJ.

Schmidt, W. H., and Thrum, R. (1981), Contributions to asymptotic theory in regression models with linear covariance structure, *Math. Operationsforsch. Statist. Ser. Statist.* 12, 243–269.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461–464.

Scott, D. W. (1992), *Multivariate Density Estimation*, Wiley, New York.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

Sen, A., and Srivastava, M. (1990), *Regression Analysis*, Springer, New York.

Serfling, R. J. (1980), *Approximation Theorems of Statistics*, Wiley, New York.

Shao, J. (2003), *Mathematical Statistics*, Springer, New York.

Shao, J., and Wu, C. F. J. (1987), Heteroscedasticity-robustness of jackknife variance estimators in linear models, *Ann. Statist.* 15, 1563–1579.

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann. Statist.* 8, 147–164.

Shibata, R. (1984), Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika* 71, 43–49.

Shorack, G. R., and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

Singh, K. (1981), On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.* 9, 1187–1195.

Slepian, D. (1962), The one-sided barrier problem for Gaussian noise, *Bell System Tech. J.* 41, 463–501.

Smirnov, N. V. (1936), Sur la distribution de ω^2 (Critérium de M. R. v. Mises), *C. R. Acad. Sci. Paris* 202, 449–452.

Smirnov, N. V. (1944), Approximating the distribution of random variables by empirical data (in Russian), *Usp. Mat. Nauk* 10, 179–206.

Smythe, R. T. (1973), Strong laws of large numbers for r-dimensional arrays of random variables, *Ann. Probab.* 1, 164–170.

Stout, W. F. (1974), *Almost Sure Convergence*, Academic Press, New York.

Strassen, V. (1964), An invariance principle for the law of the iterated logarithm, *Z. Warsch. verw. Geb.* 3, 211–226.

Strassen, V. (1966), A converse to the law of the iterated logarithm, *Z. Warsch. verw. Geb.* 4, 265–268.

Stroock, D. W., and Varadhan, S. R. S. (1979), *Multidimensional Diffusion Processes*, Springer, Berlin.

Swets, J. A., and Pickett, R. M. (1982), *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, New York.

Tanner, M. A., and Wong, W. H. (1987), The calculation of posterior distribution by data augmentation (with discussion), *J. Amer. Statist. Assoc.* 82, 528–550.

Thompson, W. A., Jr. (1962), The problem of negative estimates of variance components, *Ann. Math. Statist.* 33, 273–289.

Tierney, L. (1991), Exploring posterior distributions using Markov chains, in *Computer Science and Statistics: Proc. 23rd Symp. Interface* (E. M. Keramidas ed.), Interface Foundation, Fairfax Station, VA, 563–570.

Tierney, L. (1994), Markov chains for exploring posterior distributions (with discussion), *Ann. Statist.* 22, 1701–1786.

Tjøstheim, D. (1978), Statistical spatial series modelling, *Adv. Appl. Probab.* 10, 130–154.

Tjøstheim, D. (1983), Statistical spatial series modelling II: Some further results on unilateral lattice processes, *Adv. Appl. Probab.* 15, 562–584.

Tong, Y. L. (1980), *Probability Inequalities in Multivariate Distributions*, Academic Press, New York.

Tukey, J. (1958), Bias and confidence in not quite large samples, *Ann. Math. Statist.* 29, 614.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York.

Varadhan, S. R. S. (1966), Asymptotic probabilities and differential equations, *Commun. Pure Appl. Math.* 19, 261–286.

Verbeke, G., and Lesaffre, E. (1996), A linear mixed-effects model with heterogeneity in the random-effects population, *J. Amer. Statist. Assoc.* 91, 217–221.

Verbyla, A. P. (1990), A conditional derivation of residual maximum likelihood, *Austral. J. Statist.* 32, 227–230.

Ville, J. (1939), *Etude Critique de la Notion de Collectif*, Gauthier-Villars, Paris.

Wald, A. (1949), Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* 20, 595–601.

Wand, M. (2003), Smoothing and mixed models, *Comput. Statist.* 18, 223–249.

Wang, Y. (2002), Asymptotic nonequivalence of GARCH models and diffusions, *Ann. Statist.* 30, 754–783.

Weiss, L. (1955), The stochastic coverage of a function of sample successive differences, *Ann. Math. Statist.* 26, 532–535.

Weiss, L. (1971), Asymptotic properties of maximum likelihood estimators in some nonstandard cases, *J. Amer. Statist. Assoc.* 66, 345–350.

Wichura, M. J. (1973), Some Strassen-type laws of the iterated logarithm for multiparameter stochastic processes with independent increments, *Ann. Probab.* 1, 272–296.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989), A family of nonparametric statistics for comparing diagnostic markers with paired and unpaired data, *Biometrika* 76, 585–592.

Wiener, N. (1923), Differential spaces, *J. Math. Phys.* 2, 131–174.

Wittmann, R. (1985), A general law of iterated logarithm, *Z. Wahrsch. verw. Geb.* 68, 521–543.

Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*, Almqvist and Wiksell, Uppsala.

Wolfe, J. H. (1971), A Monte Carlo study of the sampling distribution of the likelihood ratio for mixture of multinomial distributions, *Tech. Bull. STB* 72-2, Naval Research and Training Laboratory, San Diego, CA.

Wolfowitz, J. (1949), On Wald's proof of the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* 20, 601–602.

Wood, C. L., and Altavela, M. M. (1978), Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions, *Biometrika* 65, 239–240.

Wu, C. F. J. (1983), On the convergence properties of the EM algorithm, *Ann. Statist.* 11, 95–103.

Wu, C. F. J. (1986), Jackknife, bootstrap and other resampling methods in regression analysis (with discussion), *Ann. Statist.* 14, 1261–1350.

Wu, W. B. (2007), Strong invariance principles for dependent random variables, *Ann. Probab.* 35, 2294–2320.

Yaglom, A. M. (1957), Some classes of random fields in n-dimensional space, related to stationary processes, *Theory Probab. Appl.* 2, 273–320.

Young, D. (1971), *Iterative Solutions of Large Linear Systems*, Academic Press, New York.

Zhan, X. (2002), *Matrix inequalities*, Lecture Notes in Mathematics No. 1790, Springer, New York.

Zygmund, A. (1951), An individual ergodic theorem for non-commutative transformations, *Acta Math. Szeged* 14, 103–110.

Index