

Estimation of Functions

We will consider a real, scalar-valued function over real, vector-valued arguments.

$$f : \mathbb{R}^d \mapsto \mathbb{R}$$

We may denote a function by a single letter, f , for example, or by the function notation, $f(\cdot)$ or $f(x)$. When $f(x)$ denotes a function, x is merely a placeholder. The notation $f(x)$, however, may also refer to the value of the function at the point x .

Estimation; Notation

Using the common “hat” notation for an estimator, we use \hat{f} or $\hat{f}(x)$ to denote the estimator of f or of $f(x)$.

We use the term “estimator” to denote a random variable, and “estimate” to denote a realization of the random variable.

The hat notation is also used to denote an estimate, so we must determine from the context whether \hat{f} or $\hat{f}(x)$ denotes a random variable or a realization of a random variable.

Points and Functions; Notation

The estimate or the estimator of the value of the function at the point x may also be denoted by $\hat{f}(x)$.

Sometimes, to emphasize that we are estimating the ordinate of the function rather than evaluating an estimate of the function, we use the notation $\widehat{f(x)}$.

In this case also, we often make no distinction in the notation between the realization (the estimate) and the random variable (the estimator). We must determine from the context whether $\hat{f}(x)$ or $\widehat{f(x)}$ denotes a random variable or a realization of a random variable.

Continuous Functions

The usual optimality properties that we use in developing a theory of estimation of a finite-dimensional parameter must be extended for estimation of a general function.

As we will see, two of the usual desirable properties of point estimators, namely unbiasedness and maximum likelihood, cannot be attained in general by estimators of functions.

In the estimation of functions, we must be concerned about the properties of the estimators at specific points and also about properties over the full domain. Global properties over the full domain are often defined in terms of integrals or in terms of suprema or infima.

Approximation and Estimation

There are many similarities in *estimation* of functions and *approximation* of functions, but we must be aware of the fundamental differences in the two problems.

Estimation of functions is similar to other estimation problems: we are given a sample of observations; we make certain assumptions about the probability distribution of the sample; and then we develop estimators. The estimators are random variables, and how useful they are depends on properties of their distribution, such as their expected values and their variances.

Approximation of functions is an important aspect of numerical analysis. Functions are often approximated to interpolate functional values between directly computed or known values.

Functions are also approximated as a prelude to quadrature. Methods for estimating functions often use methods for approximating functions.

General Approaches

- maximum likelihood
- represent the function as a linear combination of basis functions
- estimate the function value at a given point with a *filter* or *kernel*

How to Compare Two Functions

1. define an inner product
2. define a norm
3. define a metric

Inner Products

The *inner product* or *dot product* of the real functions f and g over the domain D , denoted by $\langle f, g \rangle_D$ or usually just by $\langle f, g \rangle$, is defined as

$$\langle f, g \rangle_D = \int_D f(x)g(x) \, dx \quad (1)$$

if the (Lebesgue) integral exists.

To avoid questions about integrability, we generally restrict attention to functions whose dot products with themselves exist; that is, to functions that are square Lebesgue integrable over the region of interest. The set of such square integrable functions is denoted $L^2(D)$. In many cases, the range of integration is the real line, and we may use the notation $L^2(\mathbb{R})$, or often just L^2 , to denote that set of functions and the associated inner product.

Inner Products

The Cauchy-Schwarz inequality holds for the inner products of functions, just as for vectors, that is,

$$\langle f, g \rangle \leq \langle f, f \rangle^{1/2} \langle g, g \rangle^{1/2}.$$

This is easy to see by first observing that for every real number t ,

$$\begin{aligned} 0 &\leq \langle (tf + g), (tf + g) \rangle \\ &= \langle f, f \rangle t^2 + 2\langle f, g \rangle t + \langle g, g \rangle \\ &= at^2 + bt + c, \end{aligned}$$

where the constants a , b , and c correspond to the inner products in the preceding equation.

Look at the discriminant.

It is clear from this proof that equality holds only if $f = 0$ or if $g = rf$ for some scalar r .

Inner Products

Just as we sometimes define vector inner products and vector norms in terms of a weight vector or matrix, we likewise define function inner products with respect to a weight function, $w(x)$, or with respect to the measure μ , where $d\mu = w(x)dx$,

$$\langle f, g \rangle_{(\mu; D)} = \int_D f(x)\bar{g}(x)w(x) dx,$$

if the integral exists. Often, both the weight and the range are assumed to be fixed, and the simpler notation $\langle f, g \rangle$ is used.

Scalar multiplication and function addition distribute over an inner product; if a is a scalar and f , g , and h are functions,

$$\langle af + g, h \rangle = a\langle f, h \rangle + \langle g, h \rangle.$$

This *linearity* is an important property of an inner product.

Norms and Pseudonorms

The *norm of a function* f , denoted generically as $\|f\|$, is a mapping into the nonnegative reals such that

- if $f \neq 0$ over an interval in which $w > 0$, then $\|f\| > 0$ and $\|0\| = 0$;
- $\|af\| = |a|\|f\|$ for a real scalar a ; and
- $\|f + g\| \leq \|f\| + \|g\|$.

Because of the linearity of a norm, a space together with a norm is called a *normed linear space*.

A norm of a function $\|f\|$ is often defined as some nonnegative, strictly increasing function of the inner product of f with itself, $\langle f, f \rangle$.

Norms

The most common type of norm for a real-valued function is the L_p norm, denoted as $\|f\|_p$, which is defined similarly to the L_p vector norm as

$$\|f\|_p = \left(\int_D |f(x)|^p w(x) dx \right)^{1/p},$$

if the integral exists.

The set of functions for which these integrals exist is often denoted by $L^p_{(\mu;D)}$. It is clear that $\|f\|_p$ satisfies the properties that define a norm.

Often μ is taken as Lebesgue measure, and $w(x)dx$ becomes dx . This is a uniform weighting.

Norms

A common L_p function norm is the L_2 norm, which is often denoted simply by $\|f\|$. This norm is related to the inner product:

$$\|f\|_2 = \langle f, f \rangle^{1/2}.$$

The space consisting of the set of functions whose L_2 norms over \mathbb{R} exist together with this norm is denoted L^2 .

The L_2 norm arises from the inner product.

Another common L_p function norm is the L_∞ norm, especially as a measure of the difference between two functions. This norm, which is called the *Chebyshev norm* or the *uniform norm*, is the limit as $p \rightarrow \infty$. This norm has the simpler relationship

$$\|f\|_\infty = \sup |f(x)w(x)|.$$

Norms of Differences: Metrics

How well one function approximates another function is usually measured by a norm of the difference in the functions over the relevant range.

If g approximates f , $\|g - f\|_\infty$ is likely to be the norm of interest. This is the norm most often used in numerical analysis when the objective is interpolation or quadrature.

In problems with noisy data, or when g may be very different from f , $\|g - f\|_2$ may be the more appropriate norm. This is the norm most often used in estimating probability density functions.

Basis Sets in Function Spaces

If each function in a linear space can be expressed as a linear combination of the functions in a set G , then G is said to be a *generating set*, a *spanning set*, or a *basis set* for the linear space. (These three terms are synonymous.) The basis sets for finite-dimensional vector spaces are finite; for most function spaces of interest, the basis sets are infinite.

A set of functions $\{q_k\}$ is *orthogonal over the domain D with respect to the nonnegative weight function $w(x)$* if the inner product with respect to $w(x)$ of q_k and q_l , $\langle q_k, q_l \rangle$, is 0 if $k \neq l$; that is,

$$\int_D q_k(x) \bar{q}_l(x) w(x) dx = 0 \quad k \neq l.$$

In the following, we will be concerned with real functions of real arguments, so we can take $\bar{q}_k(x) = q_k(x)$.

Basis Sets in Function Spaces

If also

$$\int_D q_k(x)q_k(x)w(x)dx = 1,$$

the functions are called *orthonormal*.

The weight function can also be incorporated into the individual functions to form a different set,

$$\tilde{q}_k(x) = q_k(x)w^{1/2}(x).$$

This set of functions also spans the same function space and is orthogonal over D with respect to a constant weight function.

Basis Sets in Function Spaces

Basis sets consisting of orthonormal functions are generally easier to work with and can be formed from any basis set. Given two nonnull, linearly independent functions, q_1 and q_2 , two orthonormal vectors, \tilde{q}_1 and \tilde{q}_2 , that span the same space can be formed as

$$\tilde{q}_1(\cdot) = \frac{1}{\|q_1\|} q_1(\cdot),$$

$$\tilde{q}_2(\cdot) = \frac{1}{\|q_2 - \langle \tilde{q}_1, q_2 \rangle \tilde{q}_1\|} (q_2(\cdot) - \langle \tilde{q}_1, q_2 \rangle \tilde{q}_1(\cdot)).$$

These are the Gram-Schmidt function transformations. They can easily be extended to more than two functions to form a set of orthonormal functions from any set of linearly independent functions.

Series Expansions in Basis Functions

Our objective is to represent a function of interest, $f(x)$, over some domain D , as a linear combination of “simpler” functions, $q_0(x), q_1(x), \dots$:

$$f(x) = \sum_{k=0}^{\infty} c_k q_k(x).$$

There are various ways of constructing the q_k functions. If they are developed through a linear operator on a function space, they are called *eigenfunctions*, and the corresponding c_k are called eigenvalues.

We choose a set $\{q_k\}$ that spans some class of functions over the given domain D . A set of orthogonal basis functions is often the best choice because they have nice properties that facilitate computations and a large body of theory about their properties is available.

Series Expansions in Basis Functions

If the function to be estimated, $f(x)$, is continuous and integrable over a domain D , the orthonormality property allows us to determine the coefficients c_k in the expansion:

$$c_k = \langle f, q_k \rangle.$$

The coefficients $\{c_k\}$ are called the *Fourier coefficients* of f with respect to the orthonormal functions $\{q_k\}$.

Approximation with a Series Expansion

In applications, we *approximate* the function using a truncated orthogonal series. The error due to finite truncation at j terms of the infinite series is the residual function $f - \sum_{k=1}^j c_k f_k$. The *mean squared error* over the domain D is the scaled, squared L_2 norm of the residual,

$$\frac{1}{d} \left\| f - \sum_{k=0}^j c_k q_k \right\|^2,$$

where d is some measure of the domain D . (If the domain is the interval $[a, b]$, for example, one choice is $d = b - a$.)

A very important property of Fourier coefficients is that they yield the minimum mean squared error for a given set of basis functions $\{q_i\}$; that is, for any other constants, $\{a_i\}$, and any j ,

$$\left\| f - \sum_{k=0}^j c_k q_k \right\|^2 \leq \left\| f - \sum_{k=0}^j a_k q_k \right\|^2$$

Approximation and Estimation

In applications of statistical data analysis, after forming the approximation, we then *estimate* the coefficients by identifying an appropriate probability density that is a factor of the function of interest, f .

Expected values can be estimated using observed or simulated values of the random variable and the approximation of the probability density function.

The basis functions are generally chosen to be easy to use in computations. Common examples include the Fourier trigonometric functions $\sin(kt)$ and $\cos(kt)$ for $k = 1, 2, \dots$, orthogonal polynomials such as Legendre, Hermite, and so on, splines, and wavelets.

Orthogonal Polynomials

The most useful type of basis function depends on the nature of the function being estimated. The orthogonal polynomials are useful for a very wide range of functions.

It is clear that for the k^{th} polynomial in the orthogonal sequence, we can choose an a_k that does not involve x , such that

$$q_k(x) - a_k x q_{k-1}(x)$$

is a polynomial of degree $k - 1$.

Orthogonal Polynomials

Because any polynomial of degree $k - 1$ can be represented by a linear combination of the first k members of any sequence of orthogonal polynomials, we can write

$$q_k(x) - a_k x q_{k-1}(x) = \sum_{i=0}^{k-1} c_i q_i(x).$$

Because of orthogonality, all c_i for $i < k - 2$ must be 0. Therefore, collecting terms, we have, for some constants a_k , b_k , and c_k , the three-term recursion that applies to any sequence of orthogonal polynomials:

$$q_k(x) = (a_k x + b_k) q_{k-1}(x) - c_k q_{k-2}(x), \quad \text{for } k = 2, 3, \dots$$

Orthogonal Polynomials

The recursion formula is often used in computing orthogonal polynomials. The coefficients in this recursion formula depend on the specific sequence of orthogonal polynomials, of course.

This three-term recursion formula can also be used to develop a formula for the sum of products of orthogonal polynomials $q_i(x)$ and $q_i(y)$:

$$\sum_{i=0}^k q_i(x)q_i(y) = \frac{1}{a_{k+1}} \frac{q_{k+1}(x)q_k(y) - q_k(x)q_{k+1}(y)}{x - y}.$$

This expression, which is called the Christoffel-Darboux formula, is useful in evaluating the product of arbitrary functions that have been approximated by finite series of orthogonal polynomials.

Standard Systems of Orthogonal Polynomials

The different systems are characterized by the one-dimensional intervals over which they are defined and by their weight functions. The Legendre, Chebyshev, and Jacobi polynomials are defined over $[-1, 1]$ and hence can be scaled into any finite interval. The weight function of the Jacobi polynomials is more general, so a finite sequence of them may fit a given function better, but the Legendre and Chebyshev polynomials are simpler and so are often used.

The Laguerre polynomials are defined over the half line $[0, \infty)$.

The Hermite polynomials are defined over the reals, $(-\infty, \infty)$.

Any of these systems of polynomials can be developed easily by beginning with the basis set $1, x, x^2, \dots$ and orthonormalizing them.

Table 6.1, p 136.

An Example

As an example of the use of orthogonal polynomials to approximate a given function, consider the expansion of $f(x) = e^{-x}$ over the interval $[-1, 1]$.

pp. 137 and 138.

Estimation of the Coefficients in an Orthogonal Expansion

We first decompose the function of interest to have a factor that is a probability density function, p , as in importance sampling in Monte Carlo:

$$f(x) = g(x)p(x).$$

We have

$$\begin{aligned} c_k &= \langle f, q_k \rangle \\ &= \int_D q_k(x)g(x)p(x)dx \\ &= E(q_k(X)g(X)), \end{aligned}$$

where X is a random variable whose probability density function is p .

Estimation of the Coefficients

If we can obtain a random sample, x_1, \dots, x_n , from the distribution with density p , the c_k can be unbiasedly estimated by

$$\hat{c}_k = \frac{1}{n} \sum_{i=1}^n q_k(x_i)g(x_i).$$

The series estimator of the function for all x therefore is

$$\hat{f}(x) = \frac{1}{n} \sum_{k=0}^j \sum_{i=1}^n q_k(x_i)g(x_i)q_k(x)$$

for some truncation point j .

The random sample, x_1, \dots, x_n , may be an observed dataset, or it may be the output of a random number generator.

Splines

An important type of basis function is a spline.

Designed to be 0 over certain intervals, and to have smooth joins at the interval endpoints.

Interpolating splines.

Smoothing splines.

Kernel Methods

Another approach to function estimation and approximation is to use a *filter* or *kernel* function to provide local weighting of the observed data. This approach ensures that at a given point the observations close to that point influence the estimate at the point more strongly than more distant observations.

A standard method in this approach is to convolve the observations with a unimodal function that decreases rapidly away from a central point.

This function is the filter or the kernel. A kernel function has two arguments representing the two points in the convolution, but we typically use a single argument that represents the distance between the two points.

Kernels

Some examples of univariate kernel functions are

$$\begin{aligned} \text{uniform:} & \quad K_u(t) = 0.5, & \text{for } |t| \leq 1, \\ \text{quadratic:} & \quad K_q(t) = 0.75(1 - t^2), & \text{for } |t| \leq 1, \\ \text{normal:} & \quad K_n(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}, & \text{for all } t. \end{aligned}$$

Kernel Methods

In kernel methods, the locality of influence is controlled by a *window* around the point of interest. The choice of the size of the window is the most important issue in the use of kernel methods. In practice, for a given choice of the size of the window, the argument of the kernel function is transformed to reflect the size. The transformation is accomplished using a positive definite matrix, V , whose determinant measures the volume (size) of the window.

To estimate the function f at the point x , we first decompose f to have a factor that is a probability density function, p ,

$$f(x) = g(x)p(x).$$

For a given set of data, x_1, \dots, x_n , and a given scaling transformation matrix V , the kernel estimator of the function at the point x is

$$\widehat{f}(x) = (n|V|)^{-1} \sum_{i=1}^n g(x_i) K \left(V^{-1}(x - x_i) \right).$$

Kernel Methods

In the univariate case, the size of the window is just the width h . The argument of the kernel is transformed to s/h , so the function that is convolved with the function of interest is $K(s/h)/h$. The univariate kernel estimator is

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^n g(x) K\left(\frac{x - x_i}{h}\right).$$

Pointwise Properties of Function Estimators

The statistical properties of an estimator of a function at a given point are analogous to the usual statistical properties of an estimator of a scalar parameter.

The statistical properties involve expectations or other properties of random variables.

The expectations are usually taken with respect to the (unknown) distribution of the underlying random variable.

Occasionally, we may explicitly indicate the distribution by writing, for example, $E_p(\cdot)$, where p is the density of the random variable with respect to which the expectation is taken.

Bias

The bias of the estimator of a function value at the point x is

$$E(\hat{f}(x)) - f(x).$$

If this bias is zero, we would say that the estimator is unbiased at the point x .

If the estimator is unbiased at every point x in the domain of f , we say that the estimator is pointwise unbiased. Obviously, in order for $\hat{f}(\cdot)$ to be pointwise unbiased, it must be defined over the full domain of f .

Variance

The variance of the estimator at the point x is

$$V(\hat{f}(x)) = E \left((\hat{f}(x) - E(\hat{f}(x)))^2 \right).$$

Estimators with small variance are generally more desirable, and an optimal estimator is often taken as the one with smallest variance among a class of unbiased estimators.

Mean Squared Error

The mean squared error, MSE, at the point x is

$$\text{MSE}(\hat{f}(x)) = \text{E}((\hat{f}(x) - f(x))^2).$$

The mean squared error is the sum of the variance and the square of the bias:

$$\begin{aligned} \text{MSE}(\hat{f}(x)) &= \text{E}((\hat{f}(x))^2 - 2\hat{f}(x)f(x) + (f(x))^2) \\ &= \text{V}(\hat{f}(x)) + (\text{E}(\hat{f}(x)) - f(x))^2. \end{aligned}$$

Sometimes, the variance of an unbiased estimator is much greater than that of an estimator that is only slightly biased, so it is often appropriate to compare the mean squared error of the two estimators. In some cases, as we will see, unbiased estimators do not exist, so rather than seek an unbiased estimator with a small variance, we seek an estimator with a small MSE.

Mean Absolute Error

The mean absolute error, MAE, at the point x is similar to the MSE:

$$\text{MAE}(\hat{f}(x)) = E(|\hat{f}(x) - f(x)|).$$

It is more difficult to do mathematical analysis of the MAE than it is for the MSE. Furthermore, the MAE does not have a simple decomposition into other meaningful quantities similar to the MSE.

Consistency

Consistency of an estimator refers to the convergence of the expected value of the estimator to what is being estimated as the sample size increases without bound. A point estimator T_n , based on a sample of size n , is consistent for θ if

$$E(T_n) \rightarrow \theta \quad \text{as } n \rightarrow \infty.$$

The convergence is stochastic, of course, so there are various types of convergence that can be required for consistency.

The most common kind of convergence considered is weak convergence, or convergence in probability.

Consistency

In addition to the type of stochastic convergence, we may consider the convergence of various measures of the estimator.

In general, if m is a function (usually a vector-valued function that is an elementwise norm), we may define consistency of an estimator T_n in terms of m if

$$E(m(T_n - \theta)) \rightarrow 0.$$

Consistency: Type of Convergence

For an estimator, we are often interested in *weak convergence in mean square* or *weak convergence in quadratic mean*, so the common definition of consistency of T_n is

$$E((T_n - \theta)^T (T_n - \theta)) \rightarrow 0,$$

where the type of convergence is convergence in probability. Consistency defined by convergence in mean square is also called L_2 consistency.

Rate of Convergence

If convergence does occur, we are interested in the rate of convergence. We define rate of convergence in terms of a function of n , say $r(n)$, such that

$$E(m(T_n - \theta)) = O(r(n)).$$

A common form of $r(n)$ is n^α , where $\alpha < 0$.

For example, in the simple case of a univariate population with a finite mean μ and finite second moment, use of the sample mean \bar{x} as the estimator T_n , and use of $m(z) = z^2$, we have

$$\begin{aligned} E(m(\bar{x} - \mu)) &= E((\bar{x} - \mu)^2) \\ &= \text{MSE}(\bar{x}) \\ &= O(n^{-1}). \end{aligned}$$

Consistency at a Point

In the estimation of a function, we say that the estimator \hat{f} of the function f is *pointwise consistent* if

$$E(\hat{f}(x)) \rightarrow f(x)$$

for every x the domain of f . Just as in the estimation of a parameter, there are various kinds of pointwise consistency in the estimation of a function.

If the convergence is in probability, for example, we say that the estimator is weakly pointwise consistent. We could also define other kinds of pointwise consistency in function estimation along the lines of other types of consistency.

Global Properties of Estimators of Functions

Often we are interested in some measure of the statistical properties of an estimator of a function over the full domain of the function.

The obvious way of defining statistical properties of an estimator of a function is to integrate the pointwise properties discussed in the previous section.

Statistical properties of a function, such as the bias of the function, are often defined in terms of a norm of the function.

Global Properties of Estimators of Functions

For comparing $\hat{f}(x)$ and $f(x)$, the L_p norm of the error is

$$\left(\int_D |\hat{f}(x) - f(x)|^p dx \right)^{1/p},$$

where D is the domain of f . The integral may not exist, of course. Clearly, the estimator \hat{f} must also be defined over the same domain.

Global Properties of Estimators of Functions

Three useful measures are the L_1 norm, also called the *integrated absolute error*, or IAE,

$$\text{IAE}(\hat{f}) = \int_D |\hat{f}(x) - f(x)| \, dx,$$

the square of the L_2 norm, also called the *integrated squared error*, or ISE,

$$\text{ISE}(\hat{f}) = \int_D (\hat{f}(x) - f(x))^2 \, dx,$$

and the L_∞ norm, the *sup absolute error*, or SAE,

$$\text{SAE}(\hat{f}) = \sup |\hat{f}(x) - f(x)|.$$

Global Properties of Estimators of Functions

The L_1 measure is invariant under monotone transformations of the coordinate axes, but the measure based on the L_2 norm is not.

The L_∞ norm, or SAE, is the most often used measure in general function approximation.

In statistical applications, this measure applied to two cumulative distribution functions is the *Kolmogorov distance*.

The measure is not so useful in comparing densities and is not often used in density estimation.

Global Properties of Estimators of Functions

Other measures of the difference in \hat{f} and f over the full range of x are the Kullback-Leibler measure,

$$\int_D \hat{f}(x) \log \left(\frac{\hat{f}(x)}{f(x)} \right) dx,$$

and the Hellinger distance,

$$\left(\int_D \left(\hat{f}^{1/p}(x) - f^{1/p}(x) \right)^p dx \right)^{1/p}.$$

For $p = 2$, the Hellinger distance is also called the Matusita distance.

Integrated Bias and Variance

We now want to develop global concepts of bias and variance for estimators of functions.

Bias and variance are statistical properties that involve expectations of random variables. The obvious global measures of bias and variance are just the pointwise measures integrated over the domain.

In the case of the bias, of course, we must integrate the absolute value, otherwise points of negative bias could cancel out points of positive bias.

The estimator \hat{f} is pointwise unbiased if

$$E(\hat{f}(x)) = f(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Integrated Bias

Because we are interested in the bias over the domain of the function, we define the *integrated absolute bias* as

$$\text{IAB}(\hat{f}) = \int_D |\mathbb{E}(\hat{f}(x)) - f(x)| dx$$

and the *integrated squared bias* as

$$\text{ISB}(\hat{f}) = \int_D (\mathbb{E}(\hat{f}(x)) - f(x))^2 dx.$$

If the estimator is unbiased, both the integrated absolute bias and integrated squared bias are 0. This, of course, would mean that the estimator is pointwise unbiased almost everywhere. Although it is not uncommon to have unbiased estimators of scalar parameters or even of vector parameters with a countable number of elements, it is not likely that an estimator of a function could be unbiased at almost all points in a dense domain. (“Almost” means all except possibly a set with a probability measure of 0.)

Integrated Variance

The *integrated variance* is defined in a similar manner:

$$\begin{aligned}\text{IV}(\hat{f}) &= \int_D \text{V}(\hat{f}(x)) \, dx \\ &= \int_D \text{E}((\hat{f}(x) - \text{E}(\hat{f}(x)))^2) \, dx.\end{aligned}$$

Integrated Mean Squared Error

Global unbiasedness is generally not to be expected.

An important measure for comparing estimators of functions is, therefore, based on the mean squared error.

The *integrated mean squared error* is

$$\begin{aligned}\text{IMSE}(\hat{f}) &= \int_D \text{E}((\hat{f}(x) - f(x))^2) dx \\ &= \text{IV}(\hat{f}) + \text{ISB}(\hat{f}).\end{aligned}$$

Mean Integrated Squared Error

If the expectation integration can be interchanged with the outer integration in the expression above, we have

$$\begin{aligned}\text{IMSE}(\hat{f}) &= \mathbb{E} \left(\int_D (\hat{f}(x) - f(x))^2 dx \right) \\ &= \text{MISE}(\hat{f}),\end{aligned}$$

the *mean integrated squared error*. We will assume that this interchange leaves the integrals unchanged, so we will use MISE and IMSE interchangeably.

Integrated Mean Absolute Error

Similarly, for the *integrated mean absolute error*, we have

$$\begin{aligned}\text{IMAE}(\hat{f}) &= \int_D \mathbb{E}(|\hat{f}(x) - f(x)|) \, dx \\ &= \mathbb{E} \left(\int_D |\hat{f}(x) - f(x)| \, dx \right) \\ &= \text{MIAE}(\hat{f}),\end{aligned}$$

the *mean integrated absolute error*.

Mean SAE

The *mean sup absolute error*, or MSAE, is

$$\text{MSAE}(\hat{f}) = \int_D \text{E}(\sup|\hat{f}(x) - f(x)|) dx.$$

This measure is not very useful unless the variation in the function f is relatively small. For example, if f is a density function, \hat{f} can be a “good” estimator, yet the MSAE may be quite large. On the other hand, if f is a cumulative distribution function (monotonically ranging from 0 to 1), the MSAE may be a good measure of how well the estimator performs. As mentioned earlier, the SAE is the *Kolmogorov distance*.

The Kolmogorov distance (and, hence, the SAE and the MSAE) does poorly in measuring differences in the tails of the distribution.

Large-Sample Statistical Properties

The pointwise consistency properties are extended to the full function in the obvious way.

Consistency of the function estimator is defined in terms of

$$\int_D \mathbb{E}(m(\hat{f}(x) - f(x))) dx \rightarrow 0.$$

Large-Sample Statistical Properties

The estimator of the function is said to be *mean square consistent* or *L_2 consistent* if the MISE converges to 0; that is,

$$\int_D \mathbb{E}((\hat{f}(x) - f(x))^2) dx \rightarrow 0.$$

If the convergence is weak, that is, if it is convergence in probability, we say that the function estimator is weakly consistent; if the convergence is strong, that is, if it is convergence almost surely or with probability 1, we say the function estimator is strongly consistent.

Large-Sample Statistical Properties

The estimator of the function is said to be L_1 consistent if the mean integrated absolute error (MIAE) converges to 0; that is,

$$\int_D \mathbf{E}(|\hat{f}(x) - f(x)|) dx \rightarrow 0.$$

As with the other kinds of consistency, the nature of the convergence in the definition may be expressed in the qualifiers “weak” or “strong”.

As we have mentioned above, the integrated absolute error is invariant under monotone transformations of the coordinate axes, but the L_2 measures are not. As with most work in L_1 , however, derivation of various properties of IAE or MIAE is more difficult than for analogous properties with respect to L_2 criteria.

Convergence

If the MISE converges to 0, we are interested in the rate of convergence. To determine this, we seek an expression of MISE as a function of n . We do this by a Taylor series expansion.

In general, if $\hat{\theta}$ is an estimator of θ , the Taylor series for $\text{ISE}(\hat{\theta})$ about the true value is

$$\text{ISE}(\hat{\theta}) = \sum_{k=0}^{\infty} \frac{1}{k!} (\hat{\theta} - \theta)^k \text{ISE}^{k'}(\theta),$$

where $\text{ISE}^{k'}(\theta)$ represents the k^{th} derivative of ISE evaluated at θ .

Taking the expectation yields the MISE. The limit of the MISE as $n \rightarrow \infty$ is the *asymptotic mean integrated squared error*, AMISE. One of the most important properties of an estimator is the order of the AMISE.

Large-Sample Statistical Properties

In the case of an unbiased estimator, the first two terms in the Taylor series expansion are zero, and the AMISE is

$$V(\hat{\theta}) ISE''(\theta)$$

to terms of second order.

Other Global Properties of Estimators of Functions

There are often other properties that we would like an estimator of a function to possess.

Weights or Densities

We may want the estimator to weight given functions in some particular way. For example, if we know how the function to be estimated, f , weights a given function r , we may require that the estimate \hat{f} weight the function r in the same way; that is,

$$\int_D r(x)\hat{f}(x)dx = \int_D r(x)f(x)dx.$$

Range of the Function

We may want to restrict the minimum and maximum values of the estimator.

For example, because many functions of interest are nonnegative, we may want to require that the estimator be nonnegative.

Smoothness

We may want to restrict the variation in the function. This can be thought of as the “roughness” of the function. A reasonable measure of the variation is

$$\int_D \left(f(x) - \int_D f(x) dx \right)^2 dx.$$

If the integral $\int_D f(x) dx$ is constrained to be some constant (such as 1 in the case that $f(x)$ is a probability density), then the variation can be measured by the square of the L_2 norm,

$$\mathcal{S}(f) = \int_D (f(x))^2 dx.$$

Smoothness

We may want to restrict the derivatives of the estimator or the smoothness of the estimator. Another intuitive measure of the roughness of a twice-differentiable and integrable univariate function f is the integral of the square of the second derivative:

$$\mathcal{R}(f) = \int_D (f''(x))^2 dx.$$

Often, in function estimation, we may seek an estimator \hat{f} such that its roughness (by some definition) is small.