2

Orthonormal Series and Approximation

The orthonormal series approach is the primary mathematical tool for approximation, data compression, and presentation of curves used in all statistical applications studied in Chapters 3–7. The core topics are given in the first two sections. Section 2.1 considers series approximations via visualization, and Section 2.2 gives a plain introduction in how fast Fourier coefficients can decay. Among special topics, Section 2.3 is devoted to a more formal discussion of the mathematics of series approximation, and it is highly recommended for study or review. Reading other special sections is optional and can be postponed until they are referred to in the following chapters.

2.1 Introduction to Series Approximation

In this section three particular orthonormal systems are introduced and discussed via visualization of their approximations. The first one is the cosine system that will be the main tool in the following chapters. The second one is a polynomial system based on orthonormalization of the powers \( \{1, x, x^2, \ldots\} \); this system is an excellent tool for approximating polynomial curves. The third one is a Haar system, which is a good tool for approximation of discontinuous functions; this basis is also of special interest because it is the simplest example of wavelets, which are relative newcomers to the orthogonal series scene.
For the performance assessment, we choose a set of corner (test) functions. Corner functions should represent different functions of interest that are expected to occur in practice. In this book eight specific corner functions with some pronounced characteristics are used, and they are expected to be approximated quite well or quite poorly by different systems. The set is shown in Figure 2.1.

To make all statistical simulations as simple as possible, the corner functions are some specific probability densities supported on $[0, 1]$. They are defined via uniform and normal $d_{\mu, \sigma}(x) := \frac{2\pi\sigma^2}{\sqrt{-1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ densities or their mixture.

Below, each of the corner functions is briefly discussed. These functions are arranged in order of the decreasing smoothness of their 1-periodic continuations.

1. **Uniform.** This is a uniform density on $[0, 1]$, that is, $f_1(x) := 1$. The Uniform is the smoothest 1-periodic function in our set, and we shall see that despite its triviality, neither its approximation nor statistical estimation is elementary. Moreover, this function plays a central role in asymptotic theory, and it is an excellent tool for debugging different types of errors.

2. **Normal.** This is a normal density with mean 0.5 and standard deviation 0.15, that is, $f_2(x) := d_{0.5, 0.15}(x)$. The normal (bell-shaped) curve is the most widely recognized curve. Recall the rule of three standard deviations, which states that a normal density $d_{\mu, \sigma}(x)$ practically vanishes whenever
2.1 Introduction to Series Approximation

\[ |x - \mu| > 3\sigma. \] This rule helps us to understand the curve. It also explains why we do not divide \( f_2 \) by its integral over the unit interval, because this integral is very close to 1.

3. **Bimodal.** This is a mixture of two normal densities, \( f_3(x) := 0.5d_{0.4,0.12}(x) + 0.5d_{0.7,0.08}(x) \). The curve has two pronounced and closely located modes, which why the curve is included in the set.

4. **Strata.** This is a function supported over two separated subintervals. In the case of a density, this corresponds to two distinct strata in the population. This is what differentiates the Strata from the Bimodal. The curve is obtained by a mixture of two normal densities, namely, \( f_4(x) := 0.5d_{0.2,0.06}(x) + 0.5d_{0.7,0.08}(x) \). (Note how the rule of three standard deviations was used to choose the parameters of the normal densities in the mixture.)

5. **Delta.** The underlying idea of the next curve is to have an extremely spatially inhomogeneous curve that vanishes over the entire interval except for an extremely small region at the center \((x = 0.5)\) where the function is very large. Such a function resembles many practical situations where a short but abrupt deviation from a normal process occurs. The Delta mimics the theoretical delta function, which has zero width and is integrated to 1. The Delta is defined as a normal density with very small standard deviation, \( f_5(x) := d_{0.5,0.02}(x) \).

6. **Angle.** This is a function whose 1-periodic continuation is continuous and extremely smooth except of the points \( x = k \) and \( x = k + 0.5 \), \( k = 0, \pm 1, \ldots \), where the derivative changes sign. The Angle is \( f_6(x) := (1/0.16095)d_{1,0.7}(x) \) if \( 0 \leq x \leq 0.5 \) and \( f_6(x) := (1/0.16095)d_{0.0.7}(x) \) if \( 0.5 < x \leq 1 \).

7. **Monotone.** This function is smooth over the interval, but its 1-periodic continuation has a jump at all integers \( x \). We shall see that this makes approximation of such a function challenging due to boundary effects. This also explains why the Monotone is ranked number 7 among the suggested corner functions. The Monotone is defined by the formula \( f_7(x) := d_{2,0.8}(x)/\int_0^1 d_{2,0.8}(u)du \).

8. **Steps.** This is the least smooth function in our set. The function is challenging for smooth series like a trigonometric or polynomial one. Moreover, its approximation is not rosy even for wavelets. The name of the function is clear from the graph. The Steps is defined by \( f_8(x) := 0.6 \) for \( 0 \leq x < 0.3 \), \( f_8(x) := 0.9 \) for \( 0.3 \leq x < 0.5 \) and \( f_8(x) := \frac{204}{120} \) for \( 0.5 \leq x \leq 1 \).

Now, let us recall that a function \( f(x) \) defined on an interval (the domain) is a rule that assigns to each point \( x \) from the domain exactly one element from the range of the function. Three traditional methods to define a function are a table, a formula, and a graph. For instance, we used both formulae and graphs to define the corner functions.

The fourth (unconventional) method of describing a function \( f(x) \) is via a series expansion. Suppose that the domain is \([0, 1]\). Then
\[ f(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \quad \text{where} \quad \theta_j = \int_{0}^{1} f(x) \varphi_j(x) \, dx. \quad (2.1.1) \]

Here the functions \( \varphi_j(x) \) are known, fixed, and referred to as the orthonormal functions or elements of the orthonormal system \( \{ \varphi_0, \varphi_1, \ldots \} \), and the \( \theta_j \) are called the Fourier coefficients (for a specific system we may use the name of the system in place of "Fourier"; for instance, for a Haar system we may refer to \( \theta_j \) as Haar coefficients). A system of functions is called orthonormal if the integral \( \int_{0}^{1} \varphi_s(x) \varphi_j(x) \, dx = 0 \) for \( s \neq j \) and \( \int_{0}^{1} (\varphi_j(x))^2 \, dx = 1 \) for all \( j \). Examples will be given below.

Note that to describe a function via an infinite orthogonal series expansion (2.1.1) one needs to know the infinite number of Fourier coefficients. No one can store or deal with an infinite number of coefficients. Instead, a truncated (finite) orthonormal series (or so-called partial sum)

\[ f_J(x) := \sum_{j=0}^{J} \theta_j \varphi_j(x) \quad (2.1.2) \]

is used to approximate \( f \). The integer parameter \( J \) is called the cutoff.

The advantage of this approach is the possibility of an excellent compression of the data. In statistical applications this also leads to the estimation of a relatively small number of Fourier coefficients. Roughly speaking, the main statistical issue will be how to choose a cutoff \( J \) and estimate Fourier coefficients \( \theta_j \). Thus, the rest of this section is devoted to the issue of how a choice of \( J \) affects visualization of series approximations. This will give us a necessary understanding and experience in choosing reasonable cutoffs.

Below, several orthonormal systems are introduced and then analyzed via the visualization of partial sums.

**Cosine orthonormal system on \([0, 1]\).** The elements are

\[ \varphi_0(x) := 1 \quad \text{and} \quad \varphi_j(x) := \sqrt{2} \cos(\pi j x) \quad \text{for} \quad j = 1, 2, \ldots \quad (2.1.3) \]

\[ \text{FIGURE 2.2. The first four elements of the cosine system.} \{\text{Recall that any 4 (or fewer) elements may be visualized using the argument set.j} \} [\text{set.j} = c(0,1,2,3)] \]
The first four elements are shown in Figure 2.2. It is not easy to believe that such elements may be good building blocks for approximating different functions, but surprisingly, they do a good job in approximation of smooth functions.

To visualize partial sums, several particular cutoffs, namely \( J = 3 \), \( J = 5 \), and \( J = 10 \), are chosen. Then Fourier coefficients are calculated by (2.1.1), and the partial sums (2.1.2) are shown in Figure 2.3. (Note that here and in what follows an underlying corner function is always shown by the solid line. As a result, all other curves are “hidden” behind a solid line whenever they coincide.)

Consider the partial sums shown. The Uniform is clearly described by the single Fourier coefficient \( \theta_0 = 1 \), all other \( \theta_j \) being equal to zero because \( \int_0^1 \varphi_j(x)dx = 0 \) whenever \( j > 0 \) (recall that the antiderivative, see the definition below at (2.1.4), of \( \cos(\pi j x) \) is \((1/\pi j)\sin(\pi jx)\); thus \( \int_0^1 \sqrt{2}\cos(\pi j x)dx = \sqrt{2}(\pi j)^{-1}[\sin(\pi j1) - \sin(\pi j0)] = 0 \) for any positive integer \( j \)). Thus, there is no surprise that the Uniform is perfectly fitted by the cosine system—after all, the Uniform corner function is the first element of this system.

Approximation of the Normal is a great success story for the cosine system. Even the approximation based on the cutoff \( J = 3 \), where only 4 Fourier coefficients are used, gives us a fair visualization of the underlying function, and the cutoff \( J = 5 \) gives us an almost perfect fit. Just think about a possible compression of the data in a familiar table for a normal density into only several Fourier coefficients.

Now let us consider the approximations of the Bimodal and the Strata. Note that here partial sums with small cutoffs “hide” the modes. This is especially true for the Bimodal, whose modes are less pronounced and separated. In other words, approximations with small cutoffs oversmooth an underlying curve. Overall, about ten Fourier coefficients are necessary to get a fair approximation. On the other hand, even the cutoff \( J = 5 \) gives us a correct impression about a possibility of two modes for the Bimodal and clearly indicates two strata for the Strata. The cutoff \( J = 10 \) gives us a perfect visualization except for the extra mode between the two strata. This is how the cosine system approximates a constant part of a function. We shall see the same behavior in other examples as well.

The approximations of the Delta allow us to summarize the previous observations. The partial sum with \( J = 3 \) oversmooths the Delta. Approximations with larger cutoffs do a better job in the visualization of the peak, but the valley is approximated by confusing oscillations (“wiggles”). This corner function allows us to gain necessary experience in “reading” cosine approximations. Note that the wiggles are “suspiciously” symmetric about \( x = 0.5 \), which is the point of the pronounced mode. This will always be the case for approximating a function like the Delta. This is how a trigonometric approximation “tells” us about a spatially inhomogeneous underlying
22 2. Orthonormal Series and Approximation

FIGURE 2.3. Approximation of corner functions (solid lines) by cosine series: dotted, short-dashed, and long-dashed lines correspond to cutoffs $J = 3$, $J = 5$, and $J = 10$, respectively. The 6th function is custom-made. {The optional argument CFUN allows one to substitute a corner function by a custom-made corner function. For instance, the choice $CFUN = \text{list}(3, \"2 * x - 3 * \cos(x)\")$ implies that the third corner function (the Bimodal) is substituted by the positive part of $2x-3\cos(x)$ divided by its integral over $[0,1]$, i.e., the third corner function will be $(2x - 3\cos(x))_+ / \int_0^1 (2u - 3\cos(u))_+ du$. Any valid S–PLUS formula in $x$ (use only the lower case $x$) may be used to define a custom-made corner function. This option is available for all Figures where corner functions are used. Only for this figure to visualize approximations of the Angle set $CFUN=list(6,NA)$. The choice of cutoffs is controlled by the argument set.$J$. The smaller number of approximations may be used to make curves more recognizable. On the other hand, even 4 curves are well recognizable on a color monitor. Try > \ch2(f=0) to test colors. [set.$J = c(3,5,10)$, $CFUN = \text{list}(6, \"2 - 2 * x - \sin(8 * x)\")$]

function. Note that here even the cutoff $J = 10$ is not enough for a good representation of the Delta. Clearly the cosine system is not very good for approximation of this particular corner function. On the other hand, if it is known that an underlying function is nonnegative, then a projection onto the class of nonnegative functions creates a dramatically better visualization. This will be discussed in detail in Section 3.1.

The approximations of the custom-made function are fairly good even for $J = 3$ (of course, the representation of the tails needs more Fourier coefficients). Let us use this particular example to discuss the approximation of a function near the boundary points. As we see, the partial sums are flattened out near the edges. This is because derivatives of any partial
sum (2.1.2) are zeros at the boundary points (derivatives of \( \cos(\pi jx) \) are equal to \(-\pi j \sin(\pi jx)\)). In other words, the visualization of a cosine partial sum always reveals small flat plateaus near the edges (you could notice them in all previous approximations as well). Increasing the cutoff helps to decrease the length of the plateaus and improve the visualization. This is the boundary effect, and we shall discuss in Section 2.6 how to overcome it.

A similar situation occurs for the Monotone. Here the only reason to increase the cutoff is to diminish the boundary effect.

The approximations of the Steps are not aesthetically appealing, to say the least. On the other hand, it is the purpose of this corner function to “explain” to us how the cosine partial sums approximate a piecewise constant function. In particular, let us look at the long-dashed line, which exhibits overshoots of the steps in the underlying function. This is the famous Gibbs phenomenon, which has to do with how poorly a trigonometric series converges in the vicinity of a jump. The natural conjecture would be that the overshoots vanish as \( J \to \infty \), but surprisingly, this does not take place (actually, that overshoots are proportional to a jump).

Note that while cosine approximations are not perfect for some corner functions, understanding how these partial sums perform may help us to “read” messages of these approximations and guess about underlying functions. Overall, for the given set of corner functions, the cosine system does an impressive job in both representing the functions and the data compression.

**Polynomial orthonormal system on** \([0, 1]\). This is probably the most familiar system of functions \( \{ \varphi_j(x) = \sum_{l=0}^j a_{jl}x^l, j = 0, 1, 2, \ldots \} \). Here \( j \) is called the degree of the polynomial \( \varphi_j \), and the coefficients \( \{a_{jl}\} \) are chosen in such a way that the polynomials are orthonormal.

The underlying idea of this system is as follows. It is absolutely natural to approximate a function by a linear combination of the power functions \( 1, x, x^2, \ldots \) (this resembles the idea of a polynomial regression). Unfortunately, the power functions are not orthonormal. Indeed, recall that the antiderivative \( G(x) \) of \( x^k \) is equal to \( x^{k+1}/(k + 1) \), so \( \int_0^1 x^k dx = G(1) - G(0) = (k + 1)^{-1} \); see (2.1.4) below. On the other hand, the power functions may be used as building blocks for creating a polynomial orthonormal system using the Gram–Schmidt orthonormalization procedure discussed in detail in Section 2.3.

The Gram–Schmidt procedure is very simple and performs as follows. The first function is normalized and becomes the null element of the polynomial basis, namely, \( \varphi_0(x) := 1/\left[ \int_0^1 1^2 dx \right]^{1/2} = 1 \). The first element \( \varphi_1(x) \) is calculated using \( x \) and \( \varphi_0(x) \) by the formula

\[
\varphi_1(x) := \frac{x - \left( \int_0^1 \varphi_0(u)udu \right) \varphi_0(x)}{\left[ \int_0^1 (v - \left( \int_0^1 \varphi_0(u)udu \right) \varphi_0(v))^2 dv \right]^{1/2}}.
\]
A straightforward calculation shows that \( \varphi_1(x) = \sqrt{3}(2x - 1) \). Then any other element is defined by recursion. For instance, to find the element \( \varphi_j(x) \), all previous elements are to be calculated, and then \( \varphi_j \) is defined via the previous elements and \( x_j \) by

\[
\varphi_j(x) := \frac{x^j - \sum_{l=0}^{j-1} (\int_0^1 u^l \varphi_l(u) du) \varphi_l(x)}{[\int_0^1 (v^j - \sum_{l=0}^{j-1} (\int_0^1 u^l \varphi_l(u) du) \varphi_l(v))^2 dv]^{1/2}}.
\]

The first four elements of the polynomial orthonormal system are shown in Figure 2.4.

Note that the idea of the direct approximation of \( f \) by a power series \( \sum_{j=0}^{J} b_{JJ} x^j \) is so natural and so appealing that it is worthwhile to explain why in place of a power series the orthonormal series is recommended. The only (but absolutely crucial) reason is the simplicity in calculating polynomial coefficients \( \theta_j \). Indeed, we can always write \( f_J(x) = \sum_{j=0}^{J} \theta_j \varphi_j(x) = \sum_{j=0}^{J} b_{JJ} x^j \). The power series clearly looks simpler and more natural. On the other hand, its coefficients \( b_{JJ} \) should be calculated for every \( J \), and there is no simple formula for doing this. Actually, probably the best way to find \( b_{JJ} \) is first to calculate \( \theta_j \) (note that they do not depend on the cutoff \( J! \)) and then use them for calculating \( b_{JJ} \).

Prior to the discussion of the partial sums of the polynomial system, it is worthwhile to explain how Fourier coefficients \( \theta_j \) can be calculated for a particular underlying \( f \). The fundamental theorem of calculus states that for a function \( g(x) \) continuous on \([0, 1]\),

\[
\int_0^1 g(x) dx = G(1) - G(0),
\]

where \( G(x) \) is an antiderivative of \( g \), that is, \( dG(x)/dx = g(x) \), \( x \in [0, 1] \). Thus, if an antiderivative for \( f(x) \varphi_j(x) \) is known, then a calculation of the Fourier coefficient \( \theta_j = \int_0^1 f(x) \varphi_j(x) dx \) is elementary. Unfortunately, in many cases antiderivatives are unknown, and this natural approach can-
not be used. Also, we should always keep in mind statistical applications where an underlying function is unknown and, typically, only its noisy observations at some particular points are given.

Thus, instead of (2.1.4), a numerical integration based on values of a function at some points may be of a special interest. As an example, consider the widely used trapezoid rule for numerical integration. Let \( h = 1/N \) and \( x_k = kh \) for \( k = 0, 1, \ldots, N \). Assume that the second derivative \( \phi^{(2)} \) of the function \( \phi \) is continuous. Then it is possible to show that for some \( x^* \) in \([0, 1]\) the following formula holds:

\[
\int_0^1 \phi(x) \, dx = [(h/2)(\phi(x_0) + 2\phi(x_1) + 2\phi(x_2) + \cdots + 2\phi(x_{N-1}) + \phi(x_N))] \]

\[
- (1/(12N^2))\phi^{(2)}(x^*). \tag{2.1.5}
\]

The first term on the right side gives us the trapezoid rule, and the second is called the discretization (or numerical) error. Note that the discretization error decreases proportionally to \( 1/N^2 \).

As you see, to implement the trapezoid formula it is sufficient to know values of \( f \) at \( N + 1 \) equidistant points. Also, the formula is simple. Of course, a numerical error will be presented. On the other hand, for our purposes of understanding how partial sums perform, these errors can be considered as a positive phenomenon. Indeed, in all statistical applications Fourier coefficients are estimated with some stochastic errors. Here we do not have them, but the numerical errors can simulate for us the effect of stochastic ones. As a result, we shall gain experience in dealing with partial sums whose Fourier coefficients are contaminated by errors.

The trapezoid rule has been used to calculate the polynomial coefficients (with \( N = 300 \)) for the corner functions; these polynomial approximations are shown in Figure 2.5. The eye is drawn to the partial sums for the Uniform. The Uniform serves as an excellent test for debugging all possible errors because we know for sure that all partial sums are to be identical to the underlying function (indeed, the Uniform should be perfectly matched by \( \varphi_0(x) = 1 \)). But what we see is rather puzzling because only the partial sum with the cutoff \( J = 3 \) gives us a fair representation of the curve. Moreover, the approximations perform inversely to our expectations and previous experience, where larger cutoffs meant better approximation. The reason is the numerical errors, and we see how they affect the partial sums. Note that a larger cutoff implies a larger number of calculated coefficients and therefore a larger cumulative error. This is clearly seen in Figure 2.5.1. Thus, in the presence of errors, an optimal cutoff is not necessarily the largest, and a choice of an optimal cutoff is based on a compromise between a fair approximation and cumulative errors due to incorrectly calculated polynomial coefficients. We shall see that this is also the main issue for all statistical settings where stochastic errors are inevitable.
Among other approximations shown in Figure 2.5, it is worthwhile to mention the exceptionally good approximation of the Monotone. Here even $J = 3$ gives us a perfect approximation. Also, we see that the polynomial basis has its own boundary effects, and they can be pronounced.

**Haar orthonormal system on** $[0, 1]$. This system is of special interest because it is a good tool to approximate piecewise constant functions and it is the simplest example of wavelets. It is easier to draw elements of the Haar system than to define them by formulae; in Figure 2.6 the first four elements are shown.

The wavelet literature refers to the function $F(x)$ as the *scaling* or wavelet *father* function and to $M(x)$ as the *wavelet function* or wavelet *mother* function. Note that the mother function is integrated to zero, while the father function is integrated to one. The name *mother* is motivated by the fact that all other elements are generated by the mother function. For instance, the next two elements shown in Figure 2.6 are $\sqrt{2}M(2x)$ and $\sqrt{2}M(2x - 1)$. Already you have seen the two essential operations for creating the elements: *translation* and *dilation*. Translation is the step from $M(2x)$ to $M(2x - 1)$, while dilation is from $M(x)$ to $M(2x)$. Thus, starting from a single mother function, the graphs are shifted (translated) and compressed (dilated). The next *resolution level* (scale) contains functions $2M(2^2x)$, $2M(2^2x - 1)$, $2M(2^2x - 2)$, and $2M(2^2x - 3)$. Note that each of these four functions is supported on an interval of length $\frac{1}{4}$. This procedure
2.1 Introduction to Series Approximation

The first four functions of Haar system. In the title $a = 2^{1/2}$.

Elements of a Haar system are localized; for instance, $M_{jk}(x)$ is supported on $[2^{-j}k, 2^{-j}(k+1)]$. This is what makes them so special. In short, one can expect that Haar elements will be good building blocks for approximation of nonsmooth functions.

It is customary to write the Haar partial sum as

$$f_J(x) := \theta_0 F(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} \theta_{jk} M_{jk}(x). \quad (2.1.6)$$

Here $J$ is the maximum multiresolution level (or the number of multiresolution components or scales), and thus $2^{J+1}$ Haar coefficients are used. In other words, a Haar partial sum may be based on 2, 4, 8, 16, etc. terms. Typically, only a small portion of Haar coefficients is significant, and all others are negligibly small. This implies good data compression.

The Haar system is so simple that we can even guess the Haar coefficients. For instance, let $f(x)$ be equal to 1 on the interval $[0, \frac{1}{4}]$ and vanish beyond the interval. Try to guess how to approximate the function by the elements shown in Figure 2.6. (This is a nice puzzle, and the answer is $f(x) = 0.25F(x) + 0.25M(x) + 0.5M(2x)$; of course, we can always solve such a puzzle using the formula (2.1.1).)

Figure 2.7 shows how a Haar system approximates the corner functions: the dotted line is based on 16 Haar coefficients ($J = 3$), and the short-dashed line on 64 Haar coefficients ($J = 5$). We see that the trapezoid rule of numerical integration gives relatively large numerical errors. Here we again do nothing to improve the numerical method of integration (later we shall use the toolkit S+WAVELETS for accurate calculation of these coefficients).

A promising case is the approximation of the Delta function. The localized Delta is almost perfectly (apart of its magnitude and smoothness) represented by the Haar partial sum with $J = 5$ due to the localized prop-
2. Orthonormal Series and Approximation

Figure 2.7. Approximation of the corner functions (solid lines) by the Haar basis: Dotted and short-dashed lines correspond to the number of multiresolution scales $J = 3$ and $J = 5$, respectively. $[set.J=c(3,5)]$

property of Haar elements. The Strata is also nicely approximated, and again this is due to the localized nature of Haar elements.

With all other corner functions the situation is not too rosy, and the main issue is not even the nonsmooth approximations but the large number of Haar coefficients necessary to get a fair approximation.

An interesting example is the case of the Steps. By all means, it should be the exhibition case for the Haar system. But we see that while the second jump is perfectly shown (let us ignore the numerical errors), this is not the case for the first jump. The issue is that the second jump is perfectly positioned at the point $x = \frac{3}{4}$, while the first jump is positioned at the point $x = \frac{1}{3}$, which cannot be matched by any dyadic Haar element. Thus, a Haar approximation is forced to use a sequence of elements to approximate the first jump. The important conclusion from the Steps is that even a piecewise constant function cannot be perfectly fitted by the Haar system whenever it has a jump at a point different from $2^{-l}$, $l = 0, 1, \ldots$.

To analyze Haar coefficients, the S+WAVELETS module of S–PLUS has two built-in functions: $dwt$ and $mra$. The former computes the discrete wavelet transform and allows us to visualize Haar coefficients at different resolution levels. The latter computes the multiresolution analysis and allows us to visualize a set of multiresolution approximations.

Figure 2.8 illustrates the analysis of the Normal and the Steps corner functions based on $64 = 2^6$ equidistant values of the corner functions.
2.1 Introduction to Series Approximation

FIGURE 2.8. Haar coefficients and multiresolution approximations of the Normal and the Delta functions by the Haar system. Approximations are based on $2^L$ equidistant values of the functions; the default is $L = 6$. {The set of approximated corner functions is controlled by the argument set.cf. Recall that before using any figure with wavelets, the S+WAVELETS module should be loaded using the command>
-module(wavelets) at the S–PLUS prompt.} [set.cf = c(2,5), $L=6$]

These plots are standard in S+WAVELETS, so let us explain how to read them. The first column of plots shows locations and magnitudes of Haar coefficients for the Normal. The top graph (idwt) shows the underlying curve; you can see that this is indeed the Normal. The bottom of the first column shows a magnitude of the Haar coefficient for the father function (see row s6). The Haar coefficient $\theta_{00}$ for the mother function should be shown in row d6, but the Normal function is symmetric about 0.5, so this coefficient is zero and thus not shown. In row d5 we see both approximate locations of elements and magnitudes of corresponding Haar coefficients for $M_{1,0}$ and $M_{1,1}$, etc.

The second column of plots illustrates multiresolution approximations. Row S6 shows the approximation by the father function. This approximation is often referred to as the low-frequency approximation. Because
\theta_{00} = 0$, the approximation $f_0(x)$, shown in row S5, is the same as the low-frequency approximation S6. The Haar partial sum with $J = 1$ is shown in row S4, with $J = 2$ in S3, with $J = 3$ in S2, and with $J = 4$ in S1. Finally, the approximated function is shown in the top row called Data. Note that the approximation with $J = 5$, which includes the finest elements with the Haar coefficients shown in d1, is not exhibited.

Similarly, the third and fourth columns show Haar coefficients and partial sums for the Delta.

These diagrams allow us to count the number of Haar coefficients needed for a “good” approximation of the curves. Let us begin with the Normal. Probably, the approximation S1 (which corresponds to $J = 4$) may be considered as a good one, and then $2^{J+1} = 32$ Haar coefficients should be calculated. However, we see that only 24 of them are significant (to get the number 24, just calculate the total number of coefficients shown in rows s6 and d6–d2). Thus, for the Normal curve the Haar system compresses the data essentially worse than the cosine or polynomial bases (just look again at Figure 2.3, where only 6 Fourier coefficients give us an almost perfect approximation and 4 Fourier coefficients give us a good visualization). Also, Haar approximation S3, based on 7 Haar coefficients, is a caricature of the Normal.

The outcome is quite the opposite for the Delta. Here just 9 Haar coefficients give us a fair visualization S1.

What is the conclusion? We see that there is no magical orthonormal system. Roughly speaking, smooth functions are better approximated by smooth elements, and thus cosine or polynomial systems can be recommended; nonsmooth functions may be better approximated by Haar or other wavelet systems. On the other hand, knowledge of how a particular system approximates a function allows us to recognize a pattern and then, if necessary, change the system. This is the reason why it is worthwhile to know both approximation properties of a particular orthonormal system and different orthonormal systems oriented on approximation of a specific type of function.

### 2.2 How Fast Fourier Coefficients May Decrease

The previous section introduced us to the world of orthonormal series approximations via visualization of partial sums for 8 corner functions. Another possible approach is a theoretical one that allows us to analyze simultaneously large classes of functions $f$ that are square integrable on $[0, 1]$, i.e., when $\int_0^1 f^2(x)dx < \infty$. This approach is based on the famous Parseval identity. For the cosine or polynomial orthonormal systems this
2.2 How Fast Fourier Coefficients May Decrease

identity is written as

$$\int_0^1 (f(x) - f_J(x))^2 dx = \sum_{j>J} \theta_j^2,$$  \hspace{1cm} (2.2.1)

where $f_J$ is the partial sum (2.1.2), and for the Haar system as

$$\int_0^1 (f(x) - f_J(x))^2 dx = \sum_{j>J} \sum_{k=0}^{2^j-1} \theta_{jk}^2,$$  \hspace{1cm} (2.2.2)

where here $f_J$ is the partial sum (2.1.6).

Thus, the faster Fourier coefficients decrease, the smaller cutoff $J$ is needed to get a good global approximation of $f$ by a partial sum $f_J(x)$ in terms of the integrated squared error (ISE). Note that in nonparametric statistics the ISE is customarily called the integrated squared bias (ISB).

The aim of this section is to explain the main characteristics of a function $f$ that influence the rate at which its Fourier coefficients decrease.

First, let us begin with the cosine system. We would like to understand what determines the rate at which Fourier coefficients $\theta_j = \int_0^1 \sqrt{2} \cos(\pi j x) f(x) dx$ of an integrable function $f$ decrease as $j \to \infty$.

To analyze $\theta_j$, let us recall the technique of integration by parts. If $u(x)$ and $v(x)$ are both differentiable functions, then the following equality, called integration by parts, holds:

$$\int_0^1 u(x) dv(x) = [u(1)v(1) - u(0)v(0)] - \int_0^1 v(x) du(x).$$ \hspace{1cm} (2.2.3)

Here $du(x) := u^{(1)}(x)dx$ is the differential of $u(x)$, and $u^{(k)}(x)$ denotes the $k$th derivative of $u(x)$.

Assume that $f(x)$ is differentiable. Using integration by parts and the relations

$$d \cos(\pi j x) = -\pi j \sin(\pi j x) dx, \quad d \sin(\pi j x) = \pi j \cos(\pi j x) dx,$$  \hspace{1cm} (2.2.4)

we may find $\theta_j$ for $j \geq 1$,

$$\theta_j = \sqrt{2} \int_0^1 \cos(\pi j x) f(x) dx = \sqrt{2} \pi j^{-1} \int_0^1 f(x) d\sin(\pi j x)$$

$$= \frac{\sqrt{2}}{(\pi j)} [f(1) \sin(\pi j) - f(0) \sin(0)] - \frac{\sqrt{2}}{(\pi j)} \int_0^1 \sin(\pi j x) f^{(1)}(x) dx.$$

Recall that $\sin(\pi j) = 0$ for all integers $j$, so we obtain

$$\theta_j = -\sqrt{2} \pi j^{-1} \int_0^1 \sin(\pi j x) f^{(1)}(x) dx.$$ \hspace{1cm} (2.2.5)

Note that $| \int_0^1 \sin(\pi j x) f^{(1)}(x) dx | \leq \int_0^1 | f^{(1)}(x) | dx$, and thus we may conclude the following. If a function $f(x)$ is differentiable, then for the cosine
32  2. Orthonormal Series and Approximation

We established the first rule (regardless of a particular \( f \)) about the rate at which the Fourier coefficients decrease. Namely, if \( f \) is differentiable and \( \int_0^1 |f^{(1)}(x)|dx < \infty \), then \( |\theta_j| \) decrease with rate at least \( j^{-1} \).

Let us continue the calculation. Assume that \( f \) is twice differentiable. Then using the method of integration by parts on the right-hand side of (2.2.5), we get

\[
\theta_j = -\frac{\sqrt{2}}{\pi j} \int_0^1 \sin(\pi j x) f^{(1)}(x)dx = \frac{\sqrt{2}}{(\pi j)^2} \int_0^1 f^{(1)}(x)d\cos(\pi j x)
\]

\[
= \frac{\sqrt{2}}{(\pi j)^2} [f^{(1)}(1) \cos(\pi j) - f^{(1)}(0) \cos(0)] - \frac{\sqrt{2}}{(\pi j)^2} \int_0^1 \cos(\pi j x) f^{(2)}(x)dx.
\]

(2.2.7)

We conclude that if \( f(x) \) is twice differentiable then for some finite constant \( c \),

\[
|\theta_j| \leq cj^{-2} \int_0^1 |f^{(2)}(x)|dx, \quad j \geq 1.
\]

(2.2.8)

Thus, the Fourier coefficients \( \theta_j \) of smooth (twice differentiable) functions decrease with rate not slower than \( j^{-2} \).

So far, boundary conditions (i.e., values of \( f(x) \) near boundaries of the unit interval \([0, 1]\)) have not affected the rate. The situation changes if \( f \) is smoother, for instance, it has three derivatives. In this case integration by parts can be used again. However, now the decrease of \( \theta_j \) may be defined by boundary conditions, namely by the term \([f^{(1)}(1) \cos(\pi j) - f^{(1)}(0) \cos(0)]\) on the right-hand side of (2.2.7). Note that \( \cos(\pi j) = (-1)^j \), so all these terms are equal to zero only if \( f^{(1)}(1) = f^{(1)}(0) = 0 \). This is the boundary condition that allows \( \theta_j \) to decrease faster than \( j^{-2} \). Otherwise, if the boundary condition does not hold, then \( \theta_j \) cannot decrease faster than \( j^{-2} \) regardless of how smooth the underlying function \( f \) is.

Now we know two main factors that define the decay of Fourier coefficients of the cosine system and therefore the performance of an orthonormal approximation: smoothness and boundary conditions.

A customary rule of thumb, used by many statisticians, is that an underlying function is twice differentiable. As we have seen, for twice differentiable functions the cosine system yields the optimal decrease of \( \theta_j \) regardless of the boundary conditions. Thus, the cosine system may be a good tool in statistical applications.

The case of the polynomial basis is discussed in Section 2.6.
Now let us consider a similar problem for the Haar basis. Using the specific shape of the mother function $M(x)$ (see Figure 2.6) we write,

$$|\theta_{jk}| = \int_{k2^{-j}}^{(k+1)2^{-j}} f(x)M_{jk}(x)dx$$

$$\leq 2^{-j}(\max |M_{jk}(x)|)(\max f(x) - \min f(x))/2,$$

where both the maximum and the minimum are taken over $x \in [k2^{-j}, (k+1)2^{-j}]$. Because $\max |M_{jk}(x)| = 2^j/2$, we get

$$\sum_{k=0}^{2^j-1} |\theta_{jk}| \leq 2^{-1-j/2} \sup\sum_{k=1}^{2^j+1} |f(t_k) - f(t_{k-1})|,$$  \hspace{1cm} (2.2.9)

where the supremum (see definition of the supremum below line (A.45) in Appendix A) is taken over all possible partitions of the unit interval $0 \leq t_0 < t_1 < \cdots < t_{2^j+1} \leq 1$.

The quantity $TV(f) := \lim_{m \to \infty} \sup \sum_{k=1}^{m} |f(t_k) - f(t_{k-1})|$, where the supremum is taken over all possible partitions $0 \leq t_0 < t_1 < \cdots < t_m \leq 1$ of the unit interval, is called the total variation of the function $f$ on $[0,1]$. Note that the total variation of a monotone function is equal to $|f(1) - f(0)|$.

Thus, we get from (2.2.9) that

$$\sum_{k=0}^{2^j-1} |\theta_{jk}| \leq 2^{-1-j/2} \text{TV}(f).$$  \hspace{1cm} (2.2.10)

This inequality shows how the sum of absolute values of Haar coefficients at a resolution scale $j$ decreases as $j$ increases. Such behavior is typical for wavelet coefficients (see more in Section 2.5).

Absolutely similarly we establish that

$$\sum_{k=0}^{2^j-1} |\theta_{jk}|^2 \leq 2^{-2-j} (\text{QV}(f))^2,$$  \hspace{1cm} (2.2.11)

where

$$\text{QV}(f) := \lim_{m \to \infty} \sup \left( \sum_{k=1}^{m} |f(t_k) - f(t_{k-1})|^2 \right)^{1/2}$$  \hspace{1cm} (2.2.12)

is called the quadratic variation of $f$ on $[0,1]$. Here again the supremum is taken over all possible partitions $0 \leq t_0 < t_1 < \cdots < t_m \leq 1$ of the unit interval.

These are the fundamentals that we need to know about the series approximation. The topic of how the decay of Fourier coefficients depends on various properties of an underlying function and, conversely, what Fourier coefficients may tell us about an underlying function, is a well-developed branch of mathematics. We shall discuss more formally other interesting mathematical results and approaches in the following sections.
2.3 Special Topic: Geometry of Square Integrable Functions

It would be beneficial to know that square integrable functions, which are the primary target in nonparametric curve estimation, may be viewed like points or vectors in a finite-dimensional Euclidean space, only with their own notion of perpendicular coordinates, distance, angle, Pythagorean theorem, etc.

Denote by \( L_2 = L_2([0,1]) \) the space of all square integrable functions with domain \([0,1]\). In other words, \( L_2 \) is the set of all functions \( f \) such that \( \|f\| := (\int_0^1 |f(x)|^2 dx)^{1/2} < \infty \). Note that bounded functions belong to \( L_2 \) because if \(|f(x)| \leq c < \infty \), then \( \int_0^1 |f(x)|^2 dx \leq c^2 \int_0^1 dx = c^2 < \infty \).

Below, the geometry of \( L_2 \) is discussed via a sequence of steps that make the similarity between \( L_2 \) and \( k \)-dimensional Euclidean space \( \mathcal{E}_k \) of points \( \mathbf{v} = (v_1, \ldots, v_k) \) apparent. We shall also consider vectors in \( \mathcal{E}_k \) that are directed line segments like ones shown in Figure 2.9. In what follows we shall denote by \( \mathbf{v} \) the vector from the origin to the point \( \mathbf{v} \). Figure 2.9 reminds us the main rule of finding the difference (and respectively the sum) of two vectors.

- **\( L_2 \) is a linear space.** If \( \mathbf{v} \) and \( \mathbf{u} \) are two vectors in \( \mathcal{E}_k \), then \( a\mathbf{v} + b\mathbf{u} \in \mathcal{E}_k \) for any real numbers \( a \) and \( b \). A space with this property is called linear because any linear combination of its elements is again an element of this space. Let us verify that \( L_2 \) is linear. Using the Cauchy inequality

\[
2|abf(x)g(x)| \leq a^2 f^2(x) + b^2 g^2(x),
\]

which is a corollary of the elementary \(|af(x) - bg(x)|^2 \geq 0\), implies

\[
\|af + bg\|^2 = \int_0^1 (af(x) + bg(x))^2 dx \leq 2a^2\|f\|^2 + 2b^2\|g\|^2 . \tag{2.3.1}
\]

Thus, any linear combination of two square integrable functions is again a square integrable function. In short, \( af + bg \in L_2 \) whenever \( f, g \in L_2 \) and \( a \) and \( b \) are real numbers.

- **Distance between two square integrable functions.** If \( \mathbf{v} \) and \( \mathbf{u} \) are two points in \( \mathcal{E}_k \) then the Euclidean distance (length, norm) between them is \( \left[ \sum_{j=1}^k (v_j - u_j)^2 \right]^{1/2} \). Note that this definition is based on the Pythagorean theorem and the orthonormality of the basic vectors of the Cartesian coordinate system. In particular, the length (norm) of a vector \( \mathbf{v} \), which is the distance between the origin and the point \( \mathbf{v} \), is \( \left[ \sum_{j=1}^k v_j^2 \right]^{1/2} \). Also, the length of the difference \( \mathbf{v} - \mathbf{u} \) of two vectors corresponding to points \( \mathbf{v} \) and \( \mathbf{u} \) is \( \left[ \sum_{j=1}^k (v_j - u_j)^2 \right]^{1/2} \), which is exactly the distance between those points.

For functions we may define the distance between two square integrable functions \( f \) and \( g \) as \( ||f - g|| \). In particular, this definition implies that the norm of \( f \) is \( ||f|| \). Below, we shall see that this definition preserves all properties of classical Euclidean geometry.
• **The orthogonality (perpendicularity) of square integrable functions.** The crown jewel of Euclidean geometry is the Pythagorean theorem. Recall that this famous theorem is about a right triangle whose two sides are perpendicular, i.e., the angle between them is 90° (see Figure 2.9, where the angle $\gamma$ is 90°). Recall that the side opposed to the right angle is called the hypotenuse, and the other sides are called legs. The Pythagorean theorem states that the sum of the squares of the legs of a right triangle is equal to the square of the hypotenuse. Moreover, this is a property only of right triangles. In other words, to check that two sides are perpendicular it suffices to check that the sum of the squares of these sides is equal to the square of the other side.

Let us use this Pythagorean rule for introducing the notion of orthogonal (or one may say perpendicular) square integrable functions. Figure 2.9 illustrates the underlying idea. Let $f$ and $g$ be two square integrable functions, which may be thought as either points in $L^2$ or the corresponding vectors. As we have defined earlier, their lengths (norms) in $L^2$ are $\|f\|$ and $\|g\|$. The Pythagorean rule together with Figure 2.9 implies that if these two functions are orthogonal (perpendicular), then the equality $\|f\|^2 + \|g\|^2 = \|f - g\|^2$ must hold. Let us check when this happens:

$$\|f - g\|^2 = \int_0^1 (f(x) - g(x))^2 dx = \|f\|^2 + \|g\|^2 - 2 \int_0^1 f(x)g(x)dx. \quad (2.3.2)$$

Thus, we may say that two square integrable functions $f$ and $g$ are orthogonal (perpendicular) in $L^2$ if their inner product $\langle f, g \rangle := \int_0^1 f(x)g(x)dx$ is zero. Moreover, the angle $\gamma$ between two functions in $L^2$ may be defined via the relation

$$\cos(\gamma) := \langle f, g \rangle / [\|f\| \|g\|]. \quad (2.3.3)$$

The definition (2.3.3) fits the geometry of Euclidean space $\mathbb{E}_k$, where the inner product, also referred to as the dot product, is defined as $\langle \vec{v}, \vec{u} \rangle = \sum_{j=1}^k v_j u_j$. Let us check that the absolute value of the right side of (2.3.3)
is at most 1. This follows at once from the Cauchy–Schwarz inequality
\[ \langle f, g \rangle \leq \|f\| \|g\|, \tag{2.3.4} \]
where equality holds if and only if \( f = ag \) for some real number \( a \). Let us prove this assertion. First, note that if \( \|f\| \|g\| = 0 \), then the assertion clearly holds. Thus, consider the case \( \|f\| \|g\| > 0 \). As in (2.3.1), for \( t_1(x) := f(x)/\|f\| \) and \( t_2(x) := g(x)/\|g\| \) we may write
\[ 0 \leq \|t_1 - t_2\|^2 = \|t_1\|^2 + \|t_2\|^2 - 2\langle t_1, t_2 \rangle. \]
This together with \( \|t_1\| = \|t_2\| = 1 \) implies (2.3.4), with equality if and only if \( \|t_1 - t_2\| = 0 \), which is equivalent to \( f = ag \).

Finally, to finish our “triangles” business, recall that Euclidean geometry tells us that a side of a triangle is not longer than the sum of the other two sides. Such a property is called the triangle inequality. This property also holds in \( L_2 \). Indeed, (2.3.2) together with the Cauchy–Schwarz inequality implies
\[ \|f \pm g\| \leq \|f\| + \|g\|. \tag{2.3.5} \]

- **Coordinate system in** \( L_2 \). What makes a Euclidean space so transparent and intuitively clear? Why is the procedure of depicting a point in this space so simple? The answer is obvious: the familiar Cartesian (rectangular) coordinates make this space so convenient. Thus, let us briefly recall this coordinate system and then try to introduce its analogue for \( L_2 \).

Cartesian coordinates in \( \mathcal{E}_k \) are defined by \( k \) perpendicular basic unit vectors \( \{\vec{b}_1, \ldots, \vec{b}_k\} \). By definition, the basic vectors proceed from the origin to the points whose Cartesian coordinates are \( \{(1,0,\ldots,0),\ldots,(0,\ldots,0,1)\} \). Then a vector \( \vec{v} \) may be written as \( \vec{v} = \sum_{j=1}^k v_j \vec{b}_j \), and it is easy to check that \( v_j = \langle \vec{v}, \vec{b}_j \rangle \).

Thus, “translation” of the Cartesian system of coordinates into the “language” of the space of square integrable functions is straightforward. Let \( \{\varphi_1, \varphi_2, \ldots\} \) be a system of square integrable functions that are pairwise orthogonal and have unit norms, that is, \( \langle \varphi_j, \varphi_l \rangle = 0 \) if \( j \neq l \) and \( \|\varphi_j\| = 1 \). This system is called orthonormal. Also, let us assume that this system spans \( L_2 \), that is, for any \( f \in L_2 \) and \( \varepsilon > 0 \) there exist a positive integer \( n \) and numbers \( c_1, \ldots, c_n \) such that \( \|f - \sum_{j=1}^n c_j \varphi_j\| \leq \varepsilon \). If such a system of functions exists, then it is called an orthonormal basis, or simply basis.

Then the elements of a basis may be declared as the basic unit vectors in \( L_2 \). Indeed, they are orthonormal, and it is possible to show that as is the case for a finite-dimensional Euclidean space,
\[ \left\| f - \sum_{j=1}^n c_j \varphi_j \right\| = \min_{\{c_j\}} \left\| f - \sum_{j=1}^n c_j \varphi_j \right\|, \quad \text{where } \theta_j = \langle f, \varphi_j \rangle. \tag{2.3.6} \]

In other words, the best representation of a function by a linear combination of \( n \) basic unit vectors is one in which the coefficients are Fourier coeffi-
cients. Thus Fourier coefficients play the role of coordinates of a function in the space $L_2$ where the coordinate system is created by the orthonormal elements of the basis.

Let us prove (2.3.6). Write

$$
\left\| f - \sum_{j=1}^{n} c_j \varphi_j \right\|^2 = \left\| f - \sum_{j=1}^{n} \theta_j \varphi_j + \sum_{j=1}^{n} (\theta_j - c_j) \varphi_j \right\|^2
$$

$$
= \left\| f - \sum_{j=1}^{n} \theta_j \varphi_j \right\|^2 + \left\| \sum_{j=1}^{n} (\theta_j - c_j) \varphi_j \right\|^2 + 2 \left\langle f - \sum_{j=1}^{n} \theta_j \varphi_j, \sum_{j=1}^{n} (\theta_j - c_j) \varphi_j \right\rangle.
$$

The orthonormality of the elements $\{\varphi_j\}$ together with the definition of the Fourier coefficients $\theta_j$ implies that the inner product term is zero. This yields (2.3.6).

- **Gram–Schmidt orthonormalization.** How can one construct a basis in $L_2$? To answer this question, let us assume that a countable system $\{\psi_1, \psi_2, \ldots\}$ of square integrable functions spans $L_2$ (in other words, this system is dense in $L_2$). Note that we can always discard an element $\psi_n$ of this system if it is a linear combination of the previous elements $\psi_1, \ldots, \psi_{n-1}$, that is, if $\psi_n(x) = \sum_{l=1}^{n-1} c_l \psi_l(x)$. Thus, let us assume that this system contains only linearly independent elements.

Then a basis may be constructed using the **Gram–Schmidt orthonormalization procedure**. The first element $\varphi_1$ is defined by

$$
\varphi_1(x) := \psi_1(x)/\|\psi_1\|.
$$

Then all the following elements are defined by the recursion

$$
\varphi_j(x) := \frac{\psi_j(x) - \sum_{l=1}^{j-1} (\psi_j, \varphi_l) \varphi_l(x)}{\|\psi_j(x) - \sum_{l=1}^{j-1} (\psi_j, \varphi_l) \varphi_l(x)\|}.
$$

- **The projection theorem.** The notion of a projection of a point onto a set of points is well known for Euclidean spaces. The projection of a point onto a set is defined as the point of this set that is nearest to the point. If there is more than one such point, then all these points are called projections. For instance, in a plane, the projection of a point onto a straight line is the foot of the perpendicular from the point to the line. In this case the (orthogonal) projection is unique. In three-dimensional space, the projection of a point onto a plane is also unique: This is the foot of the perpendicular from the point to the plane. Of course, there are plenty of examples where a projection is not unique. For instance, a projection of the center of a circle onto the circle is not unique, because all points of the circle are equidistant from its center. Note that the difference between the line–plane case and the circle case is that a line and a plane are linear...
subspaces of two-dimensional Euclidean space, while the circle is not. Also, the projection may not exist. For instance, consider points on the real line and let us try to project the point 2 onto the interval \([0, 1)\). There is no nearest point, because the point 1, which could be a natural projection, does not belong to the interval. On the other hand, the projection onto \([0, 1]\) is well-defined, and it is the point 1. Note that the first interval is half open, while the second is closed, and this is what makes the difference.

Keeping these examples in mind, we would like to formulate the result about a unique projection in \(L^2\) of a function \(f\) onto a linear subspace. Let us say that a linear subspace \(L\) of \(L^2\) is a closed subspace if \(L\) contains all of its limits points, that is, if \(f_n \in L\) and \(\|g - f_n\| \to 0\) then \(g \in L\). The following result states that the projection of a function \(f\) onto a closed linear subspace is always unique, and moreover, the geometry of this projection is absolutely similar to the examples for Euclidean spaces.

**The projection theorem.** Let \(L\) be a closed linear subspace of \(L^2\). Then for each \(f \in L^2\) there exists a unique element \(f^* \in L\) (the projection of \(f\) onto \(L\)) such that

\[
\|f - f^*\| = \inf_{g \in L} \|f - g\|. \tag{2.3.9}
\]

Moreover, \(f^*\) is the projection iff the difference \(f - f^*\) is orthogonal to all elements of \(L\). (Thus, the projection \(f^*\) is unique, and it may be called the orthogonal projection of \(f\) onto \(L\).)

The proof of this theorem may be found, for instance, in the textbook by Debnath and Mikusinski (1990).

- **Hilbert space.** The previous step finished our discussion of the geometry of \(L^2\). On the other hand, we are so close to understanding the notion of a Hilbert space that it is irresistible to discuss it here because both Euclidean spaces and \(L^2\) are particular examples of a Hilbert space.

Let \(H\) be a linear space with an inner product. Denote by \(x, y,\) and \(z\) any 3 elements of \(H\) and by \(a\) any real number. An inner product \(\langle x, y \rangle\) should satisfy the following properties: \(\langle x, y \rangle = \langle y, x \rangle\); \(\langle x + y, z \rangle = \langle x, y \rangle + \langle y, z \rangle\); \(\langle ax, y \rangle = a \langle y, x \rangle\); \(\langle x, x \rangle \geq 0\), with equality iff \(x = 0\). The distance in \(H\) between two elements is declared to be \(\|x - y\| := \langle x - y, x - y \rangle^{1/2}\).

Then the space \(H\) is called a Hilbert space if for any sequence of elements \(x_n \in H\) the fact \(\|x_n - x_m\| \to 0\) as \(n, m \to \infty\) implies that \(x_n\) converges to some \(x \in H\) (in other words, any Cauchy sequence converges to an element of the Hilbert space, and this property is called completeness).

One more definition is due: A Hilbert space is called separable if there exists a countable system of elements that approximates any other element from the space.

It is possible to show that \(L^2\), with the inner product defined via the Lebesgue integral, is a separable Hilbert space. A proof of this result may be found in the textbook by Debnath and Mikusinski (1990), and a sketch of a proof will be given in the next section. We do not discuss here the
Lebesgue integral but note that for all practically interesting functions discussed in the book it is equal to the Riemann integral.

- **Two useful relations.** Let \( \{ \varphi_j \} \) be an orthonormal basis in \( L_2 \) and let \( \theta_j = \langle f, \varphi_j \rangle = \int_0^1 f(x) \varphi_j(x) \, dx \) be the \( j \)th Fourier coefficient of \( f \in L_2 \). Then the following relations hold: The *Bessel inequality*
  \[
  \sum_{j=1}^n \theta_j^2 \leq \| f \|^2, \quad n = 1, 2, \ldots ,
  \]
  and the *Parseval identity*
  \[
  \| f \|^2 = \sum_{j=1}^\infty \theta_j^2.
  \]
  The Bessel inequality is implied by the line
  \[
  0 \leq \left\| f - \sum_{j=1}^n \theta_j \varphi_j \right\|^2 = \| f \|^2 + \left\| \sum_{j=1}^n \theta_j \varphi_j \right\|^2 - 2 \langle f, \sum_{j=1}^n \theta_j \varphi_j \rangle = \| f \|^2 - \sum_{j=1}^n \theta_j^2.
  \]
  The fact that \( \{ \varphi_j \} \) is a basis in \( L_2 \) means that \( \| f - \sum_{j=1}^n \theta_j \varphi_j \| \to 0 \) as \( n \to \infty \). This together with the last line yields the Parseval identity.

### 2.4 Special Topic: Classical Trigonometric Series

The classical orthonormal trigonometric Fourier system is defined by

\[
\begin{align*}
\varphi_0(x) & := 1, \quad \varphi_{2j-1}(x) := \sqrt{2} \sin(2\pi j x), \\
\varphi_{2j}(x) & := \sqrt{2} \cos(2\pi j x), \quad j = 1, 2, \ldots .
\end{align*}
\]  

(2.4.1)

Our first object is to discuss how the *partial trigonometric (Fourier)* sums

\[
S_J(x) := \sum_{j=0}^{2J} \theta_j \varphi_j(x)
\]

approximate an underlying integrable function \( f \) and why this system is a basis in \( L_2 \). In this section \( \theta_j := \int_0^1 f(x) \varphi_j(x) \, dx \) denote the Fourier coefficients, which are well-defined for integrable \( f \).

Fourier sums for the corner functions are shown in Figure 2.10. Be aware that here the Fourier sum \( S_J \) is based on \( 1 + 2J \) Fourier coefficients. Because the trigonometric elements (2.4.1) are 1-periodic, the Fourier sums are 1-periodic as well. This definitely shows up in the approximation of functions like the Monotone and the Steps (Section 2.6 explains how to improve approximations of aperiodic functions). Also, the approximations of the Steps again exhibit the Gibbs phenomenon of overshooting. Interestingly, as \( J \to \infty \), the overshoot approaches approximately 9% of a jump. (Historical
FIGURE 2.10. Approximation of the corner functions (solid lines) by Fourier sums: Dotted, short-dashed, and long-dashed lines correspond to \( J = 2 \), \( J = 3 \), and \( J = 5 \), respectively. \( \{ \text{set.} J = 2, 3, 5 \} \)

notes and discussion of this interesting phenomenon may be found in Dym and McKean 1972, Section 1.6.) In general, the Gibbs phenomenon occurs in the vicinity of any jump of a piecewise smooth function, and at this point the Fourier sums converge to the average value of the function. This is clearly exhibited in Figure 2.10.7. (It is worthwhile to know that even some wavelet expansions, discussed in the following section, suffer from overshooting. Thus, in one way or another nonsmooth functions present a challenge for any orthonormal system.)

Let us now focus on the theory of convergence of Fourier sums \( S_J(x) \) to \( f(x) \) at a given point \( x \in [0, 1] \) as \( J \to \infty \). Such convergence is called pointwise. Substituting the expressions for Fourier coefficients \( \theta_j \) into the right-hand side of (2.4.2) yields

\[
S_J(x) = 2 \int_0^1 f(t) \left[ \frac{1}{2} + \sum_{k=1}^J (\cos(2\pi k x) \cos(2\pi k t) + \sin(2\pi k x) \sin(2\pi k t)) \right] dt,
\]

and because \( \cos(\alpha - \beta) = \cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta) \), we get

\[
S_J(x) = 2 \int_0^1 f(t) \left[ \frac{1}{2} + \sum_{k=1}^J \cos(2\pi k(t - x)) \right] dt. \tag{2.4.3}
\]
The expression inside the square brackets may be simplified using the following trigonometric formula and notation,

\[
\frac{1}{2} + \sum_{k=1}^{J} \cos(2\pi ku) = \frac{1}{2} \frac{\sin(\pi(2J+1)u)}{\sin(\pi u)} = \frac{1}{2} D_J(u),
\]

with the understanding that \( D_J(0) := 2J + 1 \). The function \( D_J(u) \) is called the Dirichlet kernel, and it plays a central role in the study of pointwise convergence.

Note that graphs of \( D_J(u - 0.5) \) resemble the approximations in Figure 2.10.5. Namely, as \( J \to \infty \), the peak tends to infinity and the symmetric oscillations to either side of the peak become increasingly rapid, and while they do not die away, on the average they cancel each other. In short, as \( J \to 0 \), the Dirichlet kernels approximate the theoretical delta function.

From now on let us assume that \( f(x) \) is 1-periodic, that is, \( f(x+1) = f(x) \) for all \( x \) (in this case the unit interval may be thought as a unit circular circumference with identified endpoints 0 and 1). Then the function \( f(t)D_J(t-x) \) is also 1-periodic in \( t \). The substitution \( z = t-x \) gives

\[
S_J(x) = \int_0^1 f(x+z)D_J(z)dz.
\]

Recall that the theoretical delta function is integrated to 1, and from (2.4.5) it is easy to see by choosing \( f(x) \equiv 1 \) that the Dirichlet kernel has the same property,

\[
\int_0^1 D_J(z)dz = 1.
\]

Thus we may write

\[
S_J(x) - f(x) = \int_0^1 [f(x+z) - f(x)]D_J(z)dz.
\]

An important conclusion from (2.4.7) is that a pointwise approximation should crucially depend on the local smoothness of an approximated function \( f \) in the vicinity of \( x \).

The expression (2.4.7) is the key to all the main properties of the Fourier sum. For instance, assume that \( f \) is a twice differentiable function, and set \( g(x,z) := (f(x-z) - f(x))/\sin(\pi z) \). Under the assumption, the partial derivative \( \partial g(x,z)/\partial z \) exists, and let us additionally assume that this derivative is bounded. Then integration by parts implies

\[
S_J(x) - f(x) = \int_0^1 g(x,z)\sin(\pi(2J+1)z)dz
\]

\[
= (\pi(2J+1))^{-1}\left[ g(x,1) + g(x,0) + \int_0^1 \frac{\partial g(x,z)}{\partial z} \cos(\pi(2J+1)z)dz \right].
\]
Thus, under our assumption (recall that \( C \) is a generic positive constant)
\[
\max_x |S_J(x) - f(x)| < CJ^{-1}. \tag{2.4.8}
\]

This result allows us to make the following two conclusions. First, if an approximated function is sufficiently smooth, then it may be uniformly approximated by Fourier sums. Second, because twice differentiable functions may approximate any square integrable function in the \( L_2 \)-norm, this together with (2.4.8) implies that the trigonometric system is a basis in \( L_2 \).

More properties of Fourier sums may be found in Exercises 2.4.2–4.

For pointwise convergence it might be a good idea to smooth (shrink) the Fourier coefficients, that is, to multiply them by some real numbers between 0 and 1. Smoothing (shrinkage) is also a key idea of adaptive nonparametric series estimation (as well as many other statistical approaches). Let us consider two famous smoothing procedures.

The Fejér (Cesáro) sum is the average of Fourier sums,
\[
\sigma_J(x) := [S_0(x) + S_1(x) + \cdots + S_{J-1}(x)]/J. \tag{2.4.9}
\]

It is easy to see that \( \sigma_J \) is a smoothed partial sum \( S_{J-1} \). Indeed,

\[
\sigma_J(x) = \theta_0 \varphi_0(x) + \sum_{j=1}^{J-1} (1 - j/J) [\theta_{2j-1} \varphi_{2j-1}(x) + \theta_{2j} \varphi_{2j}(x)].
\]

A remarkable property of the Fejér sum is that if \( f \) is nonnegative, then \( \sigma_J \) is also nonnegative (thus the Fejér sum is a bona fide approximation for probability densities). To check this we use (2.4.3)–(2.4.4) and write
\[
\sigma_J(x) = J^{-1} \int_0^1 \left[ \sum_{k=0}^{J-1} \sin(\pi(2k + 1)z)/\sin(\pi z) \right] f(x + z)dz.
\]

This equality together with \( \sum_{k=0}^{J-1} \sin(\pi(2k + 1)z) = \sin^2(\pi Jz)/\sin(\pi z) \) implies
\[
\sigma_J(x) = J^{-1} \int_0^1 [\sin(\pi Jz)/\sin(\pi z)]^2 f(x + z)dz. \tag{2.4.10}
\]

This expression yields the nonnegativity of the Fejér sum whenever \( f \) is nonnegative. The function \( \Phi_J(z) := J^{-1}[\sin(\pi Jz)/\sin(\pi z)]^2 \) is called the Fejér kernel. Thus \( \sigma_J(x) = \int_0^1 \Phi_J(z) f(x + z)dz \).

Another useful property of the Fejér sum is that if \( f \) is continuous and 1-periodic, then \( \sigma_n(x) \to f(x) \) uniformly over all \( x \in [0,1] \). Also, the Fejér sum does not “overshoot” (does not suffer from the Gibbs phenomenon). The proof may be found in Dym and McKean (1972, Theorem 1.4.3).

The performance of Fejér sums is shown in Figure 2.11 (note that the number of elements of the trigonometric basis used is the same as in Figure 2.10). Due to the smoothing, the Fejér approximations are worse for visualizing modes than the similar trigonometric approximations (just compare...
the approximations for the Bimodal, the Strata, and the Angle). On the other hand, the approximations of the Delta are nonnegative, the first step in the Steps is shown much better, and there are no overshoots. In short, we see exactly what has been predicted by the theory.

One more interesting property of Fejér sums is that $|\sigma_J(x)| \leq \max_x |f(x)|$; see Exercise 2.4.7 and Figure 2.11. Fourier sums have no such property. On the other hand, if $f$ is a trigonometric polynomial of degree $n$, then Fourier sums $S_J$ are equal to $f$ for $J \geq n$, while Fejér sums have no such property. The next smoothing procedure has both these properties.

The de la Vallée Poussin sum, which is a trigonometric polynomial of degree $2J - 1$, is given by

$$V_J(x) := (S_J + S_{J+1} + \ldots + S_{2J-1})/J \quad \text{(2.4.11)}$$

$$= \theta_0 \varphi_0(x) + \sum_{j=1}^{J-1} [\theta_{2j-1} \varphi_{2j-1}(x) + \theta_{2j} \varphi_{2j}(x)]$$

$$+ \sum_{j=J}^{2J-1} (2 - j/J)[\theta_{2j-1} \varphi_{2j-1}(x) + \theta_{2j} \varphi_{2j}(x)].$$

It is clear that $V_J(x) = f(x)$ if $f$ is a trigonometric polynomial of degree $n \leq J$. Also, direct calculations show that $V_J(x) = 2\sigma_J(x) - \sigma_J(x)$, which implies $|V_J(x)| \leq 3 \max_x |f(x)|$. 
Another remarkable property of this sum is that under very mild conditions it converges uniformly to \( f \), and it is within a factor 4 of the best sup-norm approximation by trigonometric polynomials of degree \( J \). Also, for the case of smooth functions there is a simple formula for the pointwise approximation error.

To describe these properties mathematically, define the (inhomogeneous) Lipschitz function space (class) \( \text{Lip}_{0, \alpha, L}, 0 < \alpha \leq 1, 0 < L < \infty \), of 1-periodic functions:

\[
\text{Lip}_{0, \alpha, L} := \{ f : \sup_x |f(x)| < \infty, \sup_{x, h} |f(x + h) - f(x)| h^{-\alpha} \leq L < \infty \}.
\]

(2.4.12)

Here \( \alpha \) is the order and \( L \) is the constant of the Lipschitz space.

Also, we define a Lipschitz space \( \text{Lip}_{r, \alpha, L} \) of \( r \)-fold differentiable and 1-periodic (including the derivatives) functions:

\[
\text{Lip}_{r, \alpha, L} := \{ f : \sup_x |f(x)| < \infty, f^{(r)} \in \text{Lip}_{0, \alpha, L} \}.
\]

(2.4.13)

Here \( f^{(r)} \) denotes the \( r \)th derivative of \( f \). (A Lipschitz space of order \( \alpha < 1 \) is often referred to as a Hölder space. We shall use this notion in the next section which is devoted to wavelets, because wavelet coefficients characterize Hölder functions but not Lipschitz functions of order \( \alpha = 1 \).)

**Proposition 2.4.1.** Let us restrict our attention to 1-periodic functions \( f \). Then for any trigonometric polynomial \( T_J \) of degree \( J \), i.e., \( T_J(x) = \sum_{j=0}^{2^J} c_j \varphi_j(x) \), the de la Vallée Poussin inequality holds:

\[
\sup_{f \in L_p} (\|V_J - f\|_p/\|T_J - f\|_p) \leq 4, \quad 1 \leq p \leq \infty.
\]

(2.4.14)

Here \( L_p \)-norms are defined as \( \|g\|_p := (\int_0^1 |g(x)|^p dx)^{1/p}, 1 \leq p < \infty \), \( \|g\|_\infty := \sup_{x \in [0, 1]} |f(x)| \) is the sup-norm of \( g \), and \( L_p := \{ g : \|g\|_p < \infty \} \) is the \( L_p \) space of functions with finite \( L_p \)-norm.

**Proposition 2.4.2.** For any integer \( r \geq 0 \), \( 0 < \alpha \leq 1 \), and a finite \( L \) there exists a constant \( c \) such that

\[
\sup_{f \in \text{Lip}_{r, \alpha, L}} \sup_{x \in [0, 1]} |V_J(x) - f(x)| \leq c J^{-\beta}, \quad \beta := r + \alpha.
\]

(2.4.15)

The proofs of these propositions may be found in Temlyakov (1993, p. 68) and DeVore and Lorentz (1993, p. 205), respectively.

Due to these properties, de la Vallée Poussin sums are the primary tool in pointwise estimation of functions. Figure 2.12 exhibits these sums.

Only for the case \( p = 2 \) do Fourier sums “enjoy” the nice properties formulated in these propositions. This is not a big surprise, because Fourier coefficients and Fourier sums are specifically designed to perform well for square-integrable functions. On the other hand, it is amazing that a simple smoothing (shrinkage) of Fourier coefficients allows one to attain the best
2.4 Special Topic: Classical Trigonometric Series

FIGURE 2.12. Approximation of the corner functions (solid lines) by de la Vallée Poussin sums: Dotted, short-dashed, and long-dashed lines correspond to $J = 2$, $J = 3$, and $J = 4$, respectively. [set $J = c(2, 3, 4)$]

possible convergence within a reasonable factor in any $L_p$-norm. This fact tells us that it is worthwhile to use Fourier coefficients as building blocks in approximation of functions.

Let us return to Fourier sums and approximations in the $L_2$-norm. By the Parseval identity (2.3.11),

$$\|f - S_J\|^2 = \sum_{j > 2J} \theta_j^2. \tag{2.4.16}$$

Recall that the approximation theory refers to the left-hand side of (2.4.16) as the integrated squared error, but we shall use here the statistical notion of the integrated squared bias (ISB),

$$\text{ISB}_J(f) := \|f - S_J\|^2 = \sum_{j > 2J} \theta_j^2. \tag{2.4.17}$$

According to (2.3.5), $S_J$ is the optimal trigonometric polynomial of degree $J$ for approximation of a square integrable function $f$ under the $L_2$-norm. Thus, all known results about optimal approximation of functions from specific function classes may be applied to ISB. The next proposition states that $\text{ISB}_J$ converges similarly to (2.4.15). It is discussed in DeVore and Lorentz (1993, p. 205), and its direct proof may be found in Bary (1964, Section 2.3).
Proposition 2.4.3. For any integer \( r \geq 0 \), real \( \alpha \in (0, 1] \) and finite \( L \) there exists a finite constant \( c \) such that
\[
\sup_{f \in \text{Lip}_{r, \alpha, L}} \text{ISB}_J(f) \leq cJ^{-2\beta}, \quad \beta := r + \alpha.
\] (2.4.18)

There are two important function spaces defined via Fourier coefficients. The **Sobolev function space** (ellipsoid) \( W^{\beta, Q} \), \( 0 \leq \beta, Q < \infty \), is
\[
W^{\beta, Q} := \left\{ f : \theta_0^2 + \sum_{j=1}^{\infty} (1 + (2\pi j)^{2\beta})[\theta_{2j-1}^2 + \theta_{2j}^2] \leq Q \right\}.
\] (2.4.19)

Clearly, if \( \beta = 0 \), then according to the Parseval identity, \( W^{\beta, Q} \) is the space of functions whose \( L_2 \)-norm is at most \( Q^{1/2} \). If \( f \) is \( r \)-fold differentiable and 1-periodic (including the derivatives), then the inequality \( \|f + f^{(r)}\|^2 \leq Q \) together with the Parseval identity implies \( f \in W^{r,Q} \). Recall that \( f^{(r)} \) denotes the \( r \)th derivative. Exercise 2.4.6 shows that a Sobolev space is larger than just a set of functions whose \( r \)th derivatives are square integrable; on the other hand, this set of functions is the main reason why Sobolev functions are considered in statistical applications. (Rules of integration and differentiation of Fourier sums may be found in Bary 1964, Sections 1.23.8–9, 1.24.)

Another important example of a function space that may be defined via Fourier coefficients is the space of **analytic** functions,
\[
A^{\gamma, Q} := \left\{ f : \theta_0 \leq Q, \ |\theta_{2j-l}| \leq Qe^{-\gamma j}, \ l = 0, 1, \ j = 1, 2, \ldots \right\}.
\] (2.4.20)

Analytic functions are 1-periodic and infinitely differentiable (i.e., they are extremely smooth), and the parameters \((\gamma, Q)\) define a region in the \( xy \)-plane where a complex-valued function \( f(x + iy) \) may be expanded into a convergent power series.

Note that a function belongs to one of the above-defined spaces if and only if the absolute values of the Fourier coefficients satisfy some restrictions. In other words, the signs of Fourier coefficients play no role in the characterization of these function spaces. In this case the basis used is called **unconditional**.

Now let us consider two different bases closely related to the classical trigonometric one.

**Half-range trigonometric systems on** \([0, 1]\). A shortcoming of the classical trigonometric basis is that any partial sum is periodic. The following half-range trigonometric (cosine) basis is popular among statisticians because it allows one to approximate aperiodic functions very nicely. This is also the reason why we introduced it in Section 2.1.

The underlying idea is that a function \( f(x) \), \( 0 \leq x \leq 1 \), is considered as an even 2-periodic function on the interval \([-1, 1]\); that is, \( f(-x) = f(x) \). Then the classical trigonometric basis is used, and because \( \int_{-1}^{1} f(x) \sin(\pi j x) \, dx = 0 \) for any integrable even function, the only nonzero Fourier coefficients correspond to cosine functions.
Thus, we get the half-range trigonometric (cosine) basis on $[0,1]$ defined by $\{1, \sqrt{2}\cos(\pi x), \sqrt{2}\cos(2\pi x), \ldots \}$. To see that the elements are orthonormal it suffices to recall that $\cos(j\pi x)$ is an even function.

For the case of functions vanishing at the boundary points, that is, when $f(0) = f(1) = 0$, the half-range sine basis is defined as $\{\sqrt{2}\sin(\pi x), \sqrt{2}\sin(2\pi x), \ldots \}$. To see that the elements are orthonormal, recall that $\sin(j\pi x)$ is an odd function.

**Complex trigonometric basis on $[0,2\pi]$.** For the case of the interval of support $[0,2\pi]$, the classical trigonometric orthonormal system in $L_2([0,2\pi])$ and its Fourier coefficients are defined similarly to (2.4.1)–(2.4.2), only here $\varphi_0(x) = (2\pi)^{-1/2}$, $\varphi_{2j-1}(x) = \pi^{-1/2}\sin(jx)$, $\varphi_{2j}(x) = \pi^{-1/2}\cos(jx)$, and $\theta_j = \int_0^{2\pi} f(x)\varphi_j(x)dx$.

Then, the famous Euler’s formulae

\[
\cos(jx) = \frac{e^{ijx} + e^{-ijx}}{2}, \quad \sin(jx) = \frac{e^{ijx} - e^{-ijx}}{2i},
\]

where $i^2 := -1$ is the complex unit, imply the expansion

\[
f(x) = \sum_{k=-\infty}^{\infty} c_k (2\pi)^{-1/2} e^{-ikx}
\]

of a function $f(x)$ supported on $[0,2\pi]$. Here

\[
c_0 = \theta_0, \quad c_k = [\theta_{2k} + i\theta_{2k-1}]/\sqrt{2}, \quad c_{-k} = [\theta_{2k} - i\theta_{2k-1}]/\sqrt{2}, \quad k > 0.
\]

This gives us the **complex trigonometric** system $\{e^{isx}, s = 0, \pm1, \pm2, \ldots\}$. For complex functions the inner product is defined by $\langle f, g \rangle := \int_0^{2\pi} f(x)\bar{g}(x)dx$, where $\bar{g}$ is the complex conjugate of $g$ (i.e., $a + ib = a - ib$). For example, $c_k = \langle f, e^{-ikx} \rangle = \int_0^{2\pi} f(x)e^{ikx}dx$ because $e^{ikx} = e^{-ikx}$, and $\langle f, g \rangle = \langle g, f \rangle$. While similar to the sine–cosine basis, the complex basis is more convenient in some statistical applications like regression with measurement errors in predictors or density estimation with indirectly observed data.

### 2.5 Special Topic: Wavelets

In Section 2.1 the Haar basis was introduced, which is the simplest example of a wavelet basis. We have seen that the Haar basis has an excellent localization property. On the other hand, because its elements are not smooth, stepwise Haar approximations of smooth functions may be confusing. Thus, if a smooth approximation is desired, then smooth father and mother wavelets should be used.

Smooth wavelets are relative newcomers to the orthogonal approximation scene. Their name itself was coined in the mid 1980s, and in the 1990s interest in them among the statistical community has grown at an explosive rate.
There is both bad and good news about smooth wavelets. The bad news is that there are no simple mathematical formulae to describe them. Figure 2.13 depicts four smooth mother functions. The functions have a continuous, wiggly, localized appearance that motivates the label \textit{wavelets}. Just by looking at these graphs it is clear why there are no nice formulae for these wiggly functions.

The good news is that there are software packages that allow one to employ wavelets and calculate wavelet coefficients very rapidly and accurately (the last is a very delicate mathematical problem by itself, but fortunately, nice solutions have been found). Here we use the module S+WAVELETS mentioned in Section 2.1 Also, we shall see that these wiggly functions are good building blocks for approximation of a wide variety of functions.

Four types of smooth wavelets are supported by S+WAVELETS. The first is the familiar Haar. The second is the \textit{Daublets}. These wavelets are continuous, have bounded support, and they are identified by “d$^j$” where $j$ is an even integer between 4 and 20. The mother wavelets “d4” and “d12” are shown in the first two plots in Figure 2.13. The number $j$ of a wavelet indicates its width and smoothness. Wavelets with larger indices are typically wider and smoother.

The third type of supported wavelets is \textit{Symmlets}, which are also continuous, have bounded support, and are more symmetric than the Daublets. \textit{Symmlet 8} (“s8”), which is one of the most popular among statisticians, is shown in the third diagram in Figure 2.13. Here again, the larger the index of the Symmlet, the wider and smoother the mother function.

The last type is \textit{Coiflets}. These wavelets have an additional property of vanishing moments. Coiflet “c12” is shown in Figure 2.13.

It is necessary to know that the toolkit S+WAVELETS was created for the analysis of time series (it is assumed that observations are $f(1), f(2), \ldots, f(n)$), so the following multiresolution expansion is very special. Under the assumption that $n$ is divisible by $2^j_0$, the wavelet partial

\[ \text{FIGURE 2.13. Four different mother wavelets “d4,” “d12,” “s8,” and “c12.”} \]

{[Recall that before any figure with wavelets is used, the S+WAVELETS module should be loaded with the command > \texttt{module(wavelets).} [set.wav=c("d4", "d12", "s8", "c12")]}
FIGURE 2.14. Wavelet coefficients and default multiresolution approximations of the Normal and the Steps corner functions. Functions are given at \( n = 128 \) equidistant points, and the wavelet used is Symmlet 8. \{The choice of a wavelet and two corner functions is controlled by the arguments wav and set.cf, respectively.\} \([n=128, \text{set.cf} = c(2,8), \text{wav} = "s8"]\)

sum for time series is defined by

\[
    f_{j_0}(x) := \sum_{k=1}^{n/2^{j_0}} s_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=1}^{j_0} \sum_{k=1}^{n/2^j} d_{j,k} \psi_{j,k}(x). \quad (2.5.1)
\]

Here \( j_0 \) is the number of multiresolution components (or scales), \( \psi_{j,k}(x) := 2^{-j/2} \phi(2^{-j} x - k), \phi_{j,k}(x) := 2^{-j/2} \phi(2^{-j} x - k), \psi(x) \) is the wavelet function (mother wavelet), and \( \phi(x) \) is the scaling function (father wavelet); \( s_{j_0,k} \) and \( d_{j,k} \) are wavelet coefficients. The notation is the same as in the toolkit.

As we explained in Section 2.1, S+WAVELETS allows us to visualize wavelet coefficients and multiresolution approximations. As in Figure 2.8, let us consider approximation of the Normal and the Steps by the Symmlet 8 ("s8") shown in Figure 2.14. Here the particular case \( n = 128 \) and \( j_0 = 5 \) is exhibited.

We see that approximation of the Normal is significantly improved in comparison to the Haar approximation shown in Figure 2.8. Here even approximation by the four father functions (see the second diagram in the bottom row, S5, which is called the low-frequency part) gives us a fair visualization. However, only the partial sum S3 gives us an approximation that resembles the underlying function. Recall that in Figure 2.3 a good
approximation had been obtained by using only 5 Fourier coefficients; here we need at least 16 wavelet coefficients.

The outcome changes for the Steps function. Here the approximations are better than by trigonometric bases but much worse than by the Haar basis. Also, the Gibbs phenomenon (the overshoot of jumps), familiar from trigonometric approximations, is pronouncedly represented.

Now let us consider a wavelet expansion of a function $f(x), -\infty < x < \infty$. Let $\phi'$ be a father wavelet and let $\psi'$ be a mother wavelet. Denote by $\theta_{j,k} := \int_{-\infty}^{\infty} f(x) \psi_{j,k}'(x) dx$ the wavelet coefficient that corresponds to $\psi_{j,k}' := 2^{j/2} \psi(2^j x - k)$ and by $\kappa_{j,k} := \int_{-\infty}^{\infty} f(x) \phi_{j,k}'(x) dx$ the coefficient that corresponds to $\phi_{j,k}'(x) := 2^{j/2} \phi(2^j x - k)$. Then, for any integer $j_1$, the wavelet multiresolution expansion of a square integrable function $f$ is

$$f(x) = \sum_{j=-\infty}^{\infty} \kappa_{j_1,j,k} \phi_{j_1,j,k}'(x) + \sum_{j=1}^{\infty} \sum_{k=-\infty}^{\infty} \theta_{j,k} \psi_{j,k}'. \quad (2.5.2)$$

Note that if a function $f$ vanishes beyond a bounded interval and the wavelets also vanish beyond a bounded interval, then the number of nonzero wavelet coefficients at the $j$th resolution level is at most $C 2^j$.

Let us consider two function spaces that can be characterized by absolute values of wavelet coefficients (when a wavelet basis is an unconditional basis). The first one is the Hölder space $H_{r,\alpha}$ with $0 < \alpha < 1$, which is defined as the space (2.4.13), only here the assumption about 1-periodicity is dropped. Then the following characterization result holds (the proof may be found in Meyer 1992, Section 6.4): There exist wavelets such that

$$f \in H_{r,\alpha} \iff |\theta_{j,k}| < c_1 2^{-j(r+\alpha+1/2)}, \quad |\kappa_{j_0,k}| < c_2,$$  \quad (2.5.3)

where $c_1$ and $c_2$ are some constants. No characterization of Lipschitz spaces with $\alpha = 1$ exists. In this case a larger Zygmund function space may be considered; see Meyer (1992, Section 6.4).

The second function space is a Besov space $B_{pqQ}^\sigma$, $1 \leq p, q \leq \infty$, $0 < \sigma, Q < \infty$, which includes both smooth and discontinuous functions like Hölder functions and functions of bounded total variation. The definition of this space may be found in Meyer (1992, Section 2.9), and it is skipped here. Instead, its characterization via wavelet coefficients is presented (the mathematics of this characterization and assumptions may be found in Meyer 1992, Section 6.10),

$$B_{pqQ}^\sigma := \left\{ f : \left[ \sum_{k=-\infty}^{\infty} |\kappa_{j_1,k}|^p \right]^{1/p} + \left( \sum_{j=j_1}^{\infty} \left[ 2^{j(\sigma+1/2-1/p)} \left[ \sum_{k=-\infty}^{\infty} |\theta_{j,k}|^p \right]^{1/p} \right]^q \right]^{1/q} < Q \right\}. \quad (2.5.4)$$
For instance, a Hölder space $H_{r,\alpha}$ corresponds to $B^{\beta}_{\infty\infty}$, $\beta := r + \alpha$, and the space of functions of bounded total variation is a superset of $B^{1}_{11}$ and a subset of $B^{1}_{1\infty}$.$^\prime$. These two examples shed light on the meaning of the parameters $\sigma$, $p$, $q$, and $Q$.

2.6 Special Topic: More Orthonormal Systems

This section reviews several orthonormal systems that may be useful for approximation of functions from particular spaces.

- **Polynomials on a bounded interval.** For classical polynomials the customarily studied bounded interval is $[-1,1]$. An orthonormal basis for the space $L_{2}([-1,1])$, with the inner product $\langle f,g \rangle := \int_{-1}^{1} f(x)g(x)dx$, is generated by applying the Gram–Schmidt orthonormalization procedure (2.3.7)–(2.3.8) to the powers $\{1, x, x^{2}, \ldots\}$. Also, the $j$th element $G_{j}(x)$ of this basis may be calculated via the formula

$$G_{j}(x) = \frac{1}{j!2} \sqrt{\frac{(2j+1)}{2}} \frac{d^{j}}{dx^{j}} (x^{2} - 1)^{j}. \quad (2.6.1)$$

It is worthwhile to note that the well-known Legendre polynomials $P_{j}(x) = \sqrt{2/(2j+1)} G_{j}(x)$, which are built-in functions in many software packages, are orthogonal but not orthonormal. To compute Legendre polynomials, the recurrence formula

$$P_{n}(x) = n^{-1}[(2n-1)xP_{n-1}(x) - (n-1)P_{n-2}(x)], \quad (2.6.2)$$

together with the facts that $P_{0}(x) = 1$ and $P_{1}(x) = x$, is especially useful.

The following assertion, whose proof may be found in DeVore and Lorentz (1993, Section 7.6), shows how the integrated squared bias of the polynomial approximation decreases.

**Proposition 2.6.1.** Let $f$ be $r$-fold differentiable and

$$|f^{(r)}(t) - f^{(r)}(s)| \leq Q|t - s|^{\alpha}, \quad \text{where} \quad t, s \in [-1,1], \quad 0 < \alpha \leq 1. \quad (2.6.3)$$

Then there exists a constant $c$ such that the polynomial partial sums $S_{J}^{*}(x) := \sum_{j=0}^{J} \langle f, Q_{j} \rangle Q_{j}(x)$ satisfy the relation

$$\int_{-1}^{1} (f(x) - S_{J}^{*}(x))^{2}dx \leq cJ^{-2(r+\alpha)}. \quad (2.6.4)$$

- **Polynomials on $[0,\infty)$.** Sometimes it is convenient to approximate an underlying function on a half-line $[0,\infty)$. In this case the idea is to modify the inner product in such a way that the integration over the half-line is well-defined. The customary approach is to consider the inner product $\langle f,g \rangle := \int_{0}^{\infty} f(x)g(x)e^{-x^{2}}dx$ and then apply the Gram–Schmidt orthonormalization procedure to the polynomials $\{1, x, x^{2}, \ldots\}$. This defines the Laguerre basis.
• Polynomials on \((-\infty, \infty)\). The approach is the same and the inner product is defined by \[ \langle f, g \rangle := \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2} \, dx. \] Then the Gram-Schmidt procedure is applied to the polynomials \(\{1, x, x^2, \ldots\}\), and the system obtained is called the **Hermite** basis.

• A set of discrete points. So far, we have discussed the case of functions from \(L_2\) that are defined on intervals. For many practical problems a function is defined only at a set of discrete points, and there is no interest in values of this function at other points. Let there be \(m\) such points \(\{x_1, x_2, \ldots, x_m\}\). Then the inner product may be defined as

\[
\langle f, g \rangle := \sum_{k=1}^{m} p_k f(x_k)g(x_k),
\]

(2.6.5)

where \(p_k\) are some positive “weights.” If these weights are summable to 1, then the inner product is just \(E\{f(X)g(X)\}\), where \(X\) is a discrete random variable with probability mass function \(p_k\), and \(E\{\cdot\}\) denotes the expectation.

Thus, for any system \(\{\psi_1(x), \ldots, \psi_J(x)\}\), a problem of best approximation, which is the analogue of the \(L_2\)-approach, becomes the familiar problem of finding the coefficients \(\{c_j\}\) minimizing \(\sum_{k=1}^{J} p_k [f(x_k) - \sum_{j=1}^{m} c_j \psi_j(x_k)]^2\).

It was shown by Chebyshev (the famous probabilistic inequality (A.26) also bears his name) that orthonormalization of the polynomials \(\{1, x, x^2, \ldots\}\) gives a complete orthonormal system (basis) in this setting. More details may be found in Kolmogorov and Fomin (1957, Section 7.3.8). Similarly, orthonormalization of trigonometric functions also leads to a basis at discrete points. Note that for the case of identical weights and equidistant points on \([0, 1]\) the trigonometric system is the orthonormal basis.

• Enriched bases. So far, we have discussed only classical bases. In some cases it is worthwhile to enrich a classical basis by elements like linear, quadratic, or step functions, which allow one to approximate a set of targeted functions.

As an example, let us begin with the case of the trigonometric sine–cosine system. We have seen in Section 2.4 that its approximations of aperiodic functions were terrible.

The issue is that the elements of the basis are periodic, and the derivatives of the elements are also periodic. Thus, to fit aperiodic functions, this basis should be enriched by aperiodic elements, for instance, by the linear function \(x\) and the quadratic function \(x^2\).

Since both the linear and quadratic functions are not orthonormal to the elements of the trigonometric system, the Gram–Schmidt procedure should be used.

Approximations of the corner functions by the trigonometric system enriched by the linear function are shown in Figure 2.15. The partial sums for the Monotone and the Steps look much better. The only remaining pattern
FIGURE 2.15. Approximation of the corner functions (solid lines) by Fourier sums enriched by the linear function: dotted, short-dashed, and long-dashed lines correspond to the cutoffs $J = 2$, $J = 3$, and $J = 5$, respectively. [set.$J = c(2,3,5)$]

FIGURE 2.16. Approximation of the corner functions (solid lines) by Fourier sums enriched by the linear and quadratic polynomial functions: Dotted, short-dashed, and long-dashed lines correspond to cutoffs $J = 2$, $J = 3$, and $J = 5$, respectively. [set.$J = c(2,3,5)$]
that catches eye is that near the edges, the Monotone (and the Angle) are not represented very well. This is because the derivative of a partial sum is still periodic.

Thus, let us additionally enrich this new basis by the quadratic function. The corresponding approximations are shown in Figure 2.16. As we see, now the Angle and the Monotone are represented near the boundaries much better. On the other hand, some other corner functions are fitted worse (especially for the smallest cutoff) near the edges. So, for the case of small cutoffs there are pros and cons in this enrichment.

Now let us consider the more challenging problem of enriching the cosine basis by a function that mimics a step function at a point \( a \). The aim is to get a perfect fit for the first jump in the Steps.

Set \( \phi(x,a) = 1 \) if \( 0 \leq x \leq a \) and \( \phi(x,a) = 0 \) if \( a < x \leq 1 \); that is, \( \phi(x,a) \) is a step function with unit jump at the point \( a \). We add the step function to the set of the first \( 1 + J \) cosine functions \( \{ \varphi_0 = 1, \varphi_j = \sqrt{2} \cos(\pi j x), j = 1, \ldots, J \} \) and then use the Gram–Schmidt orthogonalization procedure to get the \( (2 + J) \)th element of the enriched system,

\[
\phi(x,a,J) := \frac{\phi(x,a) - a - \sum_{j=1}^{J} 2^{1/2}(\pi j)^{-1} \sin(\pi j a) \varphi_j(x)}{\left[ \int_0^1 (\phi(u,x_0) - x_0 - \sum_{j=1}^{J} 2^{1/2}(\pi j)^{-1} \sin(\pi j a) \varphi_j(u))^2 du \right]^{1/2}}.
\]

![FIGURE 2.17. Approximation of the corner functions (solid lines) by the cosine basis enriched with the step function \( \phi(x,a) \) with the default value \( a = \frac{1}{3} \). Dotted, short-dashed, and long-dashed lines correspond to \( J = 3, J = 5, \) and \( J = 10 \), respectively. [\( a=1/3, \ set.J=c(3,5,10) \)]](image)
Denote $\theta_j = \int_0^1 f(u) \varphi_j(u) du$ and $\kappa(a, J) = \int_0^1 f(u) \phi(u, a, J) du$. Then, a partial sum

$$S_J(x, a) := \sum_{j=0}^J \theta_j \varphi_j(x) + \kappa(a, J) \phi(x, a, J)$$

is used to approximate $f$. Partial sums are shown in Figure 2.17.

Let us begin the discussion of these approximations with the Uniform. The numerical errors are apparently large because the trapezoid rule is used for nonsmooth functions. Also note how “aggressively” the step function tries to find its place for the smallest cutoffs. The same pattern is clearly seen in the Bimodal and the Strata diagrams. This is because there is plenty of room for the step function when only several cosine functions are used to approximate a spatially inhomogeneous function. On the other hand, the enriched basis does a superb job in visualizing the first jump in the Steps.

### 2.7 Exercises

**2.1.1** Suggest corner functions with 3 and 4 modes using mixtures of normal densities.

**2.1.2** Repeat Figure 2.3 with different cutoffs. Answer the following questions: (a) What are the minimal cutoffs (if any) for each corner function that imply a reasonable fit? (b) How does the cosine system approximate a constant part of a function? (c) Indicate graphs that exhibit the Gibbs phenomenon.

**2.1.3** Verify that for the polynomial system, $\varphi_1(x) = \sqrt{3}(2x - 1)$. Also, calculate $\varphi_2(x)$.

**2.1.4** Find an antiderivative for: (a) $3x - 5x^2$; (b) $5 \cos(2x) - 3 \sin(5x)$.

**2.1.5** Verify (2.1.5).

**2.1.6** Repeat Figure 2.5 with different cutoffs. What are the minimal cutoffs for each corner function (if any) that give a reasonable fit? Compare these cutoffs with those obtained for the cosine system in Exercise 2.1.2.

**2.1.7** Repeat Figure 2.8 for different corner functions, and discuss the outcomes.

**2.1.8** Explain the multiresolution approximation of the Delta function shown in Figure 2.8.

**2.2.1** Let $f_J(x) = \sum_{j=0}^J \theta_j \varphi_j(x)$. Find $\int_0^1 (f_J + L(x) - f_J(x))^2 dx$.

**2.2.2** For $f_J$ from the previous exercise, check that $\int_0^1 f_J^2(x) dx = \sum_{j=0}^J \theta_j^2$.

**2.2.3** Verify (2.2.3).

**2.2.4** Assume that the boundary condition $f^{(1)}(0) = f^{(1)}(1)$ holds, and $f$ has either 3 or 4 derivatives. How fast do the Fourier coefficients (for the cosine basis) decrease? Hint: Continue (2.2.7) using integration by parts.

**2.2.5** Find how fast the Fourier coefficients (for the cosine system) of the functions $x, x^2, x^3, x^4$ decrease.
2.2.6 Verify (2.2.9).
2.2.7 Verify (2.2.11).
2.2.8 Find the total and quadratic variations of $\cos(j\pi x)$ on $[0,1]$.

2.3.1 Prove that if $f_1$ and $f_2$ are square integrable, then $f_1f_2 \in L_1$.
2.3.2 Show that both the inner product $\langle f, g \rangle$ in $L_2$ space and the dot product $\langle \vec{v}, \vec{u} \rangle$ in $\mathcal{E}_k$ satisfy the properties of an inner product formulated in the subsection Hilbert space.

2.3.3 Check the orthogonality (in $L_2$) between the following two functions:
(a) $1$ and $\sin(ax)$; (b) $\sin(ax)$ and $\cos(bx)$; (c) $\cos(ax)$ and $\cos(bx)$; (d) $1$ and $x+a$; (e) $x$ and $ax+bx^2$.

2.3.4 We say that a sequence of functions $f_n$ converges to $f$ in the $L_2$-norm if and only if $\|f_n - f\| \to 0$ as $n \to \infty$. Let sequences $f_n$ and $g_n$ converge in the $L_2$-norm to $f$ and $g$, respectively. Prove that (a) the sum of two sequences $f_n + g_n$ converges to the sum of their limits $f + g$; (b) if $a_n$ is a sequence of real numbers converging to $a$, then the sequence of functions $a_nf$ converges to $af$; (c) the following convergence holds,

$$\langle f_n, g_n \rangle \to \langle f, g \rangle. \quad (2.7.1)$$

2.3.5 Let $n$ functions $\{f_j, j = 1, 2, \ldots, n\}$ be orthogonal in $L_2$. Prove that these functions are also linearly independent, that is, the identity $\sum_{j=1}^{n} a_j f_j(x) \equiv 0, x \in [0,1]$ implies, $a_1 = a_2 = \cdots = a_n = 0$.
2.3.6 Establish that if a series $\sum_{j=0}^{\infty} \theta_j \varphi_j(x)$ converges to a function $f$ in $L_2$ as $n \to \infty$ and $\{\varphi_j\}$ is an orthonormal system in $L_2$, then

$$\theta_j = \langle f, \varphi_j \rangle = \int_0^1 f(x) \varphi_j(x) dx. \quad (2.7.2)$$

2.3.7 Let $f_1, f_2, \ldots, f_k$ be pairwise orthogonal, i.e., $\langle f_i, f_j \rangle = 0$ whenever $l \neq j$. Verify that $\|\sum_{l=1}^{k} f_l\|^2 = \sum_{l=1}^{k} \|f_l\|^2$.
2.3.8 Check that for any orthonormal system $\{f_1, f_2, \ldots\}$ the equality $\|f_l - f_j\| = \sqrt{2}$ holds for $l \neq j$.

2.3.9 Using the Gram–Schmidt procedure, orthogonalize the set of functions $\{1, x, x^2\}$. As a result, the first three elements of the Legendre polynomial basis on $[0,1]$ are obtained.

2.3.10 Using the Gram–Schmidt procedure, orthogonalize the following set of trigonometric functions enriched by the power functions $\{1, \sin(2\pi x), \ldots, \sin(2\pi Nx), \cos(2\pi x), \ldots, \cos(2\pi Nx), x, x^2\}$. As a result, you get a so-called trigonometric-polynomial system.

2.3.11 Show that the element (2.3.8) is orthogonal to the elements $\varphi_s$, $s = 1, \ldots, j - 1$.

2.3.12 Verify (2.3.6) using the projection theorem.

2.3.13 Find the orthogonal projection (in $L_2$) of a function $f \in L_2$ onto a subspace of all linear combinations of the functions $\{1, \cos(\pi x), \cos(2\pi x)\}$.

2.4.1 Repeat Figure 2.10 with different cutoffs, and then for every corner function find a minimal cutoff that gives a fair fit.
2.4.2 Let \( f \in L_1 \) (i.e., \( \int_0^1 |f(x)|dx < \infty \)) and let for some \( \delta > 0 \) the Dini condition \( \int_{-\delta}^\delta |(f(x + t) - f(x))/t|dt < \infty \) hold. Then \( S_J(x) \rightarrow f(x) \) as \( J \rightarrow \infty \). Hint: See Theorem 8.1.1 in Kolmogorov and Fomin (1957).

2.4.3 Suppose that \( f \) is bounded, has only simple discontinuities, and at every point has left and right derivatives. Then \( S_J(x) \) converges to \( \lim_{\delta \rightarrow 0}[f(x + \delta) + f(x - \delta)]/2 \). Hint: See Remark 8.1.1 in Kolmogorov and Fomin (1957).

2.4.4 Let \( f \) be bounded and differentiable and let its derivative be square integrable. Then \( S_J(x) \) converges to \( f(x) \) uniformly over all \( x \in [0, 1] \). Hint: See Theorem 8.1.2 in Kolmogorov and Fomin (1957).

2.4.5 Check (2.4.10).

2.4.6 The function \( f(x) = |x - 0.5| \) is not differentiable. Nevertheless, show that it belongs to a Sobolev class \( W_{1,Q} \) with some \( Q < \infty \).

2.4.7 Show that Fejér sums satisfy \( |\sigma_J(x)| \leq \max_x |f(x)| \). Hint: Begin with considering \( f(x) = 1 \) and then using (2.4.10) show that the Fejér kernel is integrated to 1.

2.4.8 Use Figure 2.11 to find cutoffs for Fejér sums that give the same visualization of modes as the Fourier sums in Figure 2.10.

2.4.9 Use Figure 2.12 to find cutoffs for de la Vallée Poussin sums that give the same visualization of modes as the Fourier sums in Figure 2.10.

2.5.1 Repeat Figure 2.14 with two other wavelets having the parameter \( j \) different from 8. Discuss how this parameter affects the data compression property of wavelet approximations.

2.5.2 Repeat Figure 2.14 for two other corner functions. Discuss how smoothness of an underlying function affects the data compression.

2.5.3 Repeat Figure 2.14 with different \( n \), different corner functions, and different wavelets. Find the best wavelets for the corner functions.

2.6.1 Explain how to calculate the polynomial basis for \( L_2([0, 1]) \).

2.6.2 Find \( G_1 \), \( G_2 \), and \( G_3 \) using (2.6.1).

2.6.3 For the subsection “Polynomials on \((-\infty, \infty)\)” find the first four elements of the Hermite basis. Hint: Recall that \( \pi^{-1/2}e^{-x^2} \) is the normal \( N(0, .5) \) density.

2.6.4 Prove that a trigonometric basis is orthonormal on a set of equidistant points.

2.6.5 Use Figure 2.17 to analyze how the cutoff \( J \) affects the visualization of pseudo-jumps in the smooth corner functions.

2.6.6 Try different parameters \( a \) in Figure 2.17. Explain the results.

2.8 Notes

The basic idea of Fourier series is that “any” periodic function may be expressed as a sum of sines and cosines. This idea was known to the Babylonians, who used it for the prediction of celestial events. The history of the
subject in more recent times begins with d’Alembert, who in the eighteenth century studied the vibrations of a violin string. Fourier’s contributions began in 1807 with his studies of the problem of heat flow. He made a serious attempt to prove that any function may be expanded into a trigonometric sum. A satisfactory proof was found later by Dirichlet. These and other historical remarks may be found in Dym and McKean (1972). Also, Section 1.1 of that book gives an excellent explanation of the Lebesgue integral, which should be used by readers with advanced mathematical background. The textbook by Debnath and Mikusinski (1990) is devoted to Hilbert spaces. The books by Bary (1964) and Kolmogorov and Fomin (1957) give a relatively simple discussion (with rigorous proofs) of Fourier series.

The simplest functions of the variable $x$ are the algebraic (ordinary) polynomials $P_n = c_0 + c_1 x + \cdots + c_n x^n$ of degree $n$. Thus, there is no surprise that they became the first and powerful tool for approximation of other functions. Moreover, a theorem about this approximation, discovered in the nineteenth century by Weierstrass, became the cornerstone of modern approximation theory. For a continuous function $f(x), x \in [0, 1]$, it asserts that there exists a sequence of ordinary polynomials $P_n(x)$ that converge uniformly to $f(x)$ on $[0, 1]$. There are many good books on approximation theory (but do not expect them to be simple). Butzer and Nessel (1971) is the classical reference. Among recent ones, DeVore and Lorentz (1993), Temlyakov (1993), and Lorentz, Golitschek and Makovoz (1996) may be recommended as solid mathematical references.

The mathematical theory of wavelets was developed in the 1980s and it progressively appeared to be useful in approximation of spatially inhomogeneous functions. There are several relatively simply written books by statisticians for statisticians about wavelets, namely, the books by Ogden (1997) and Vidacovic (1999), as well as the more mathematically involved book by Härdle et al. (1998). The book by Mallat (1998) discusses the application to signal processing.

The book by Walter (1994) gives a rather balanced approach to the discussion of all orthogonal systems, and it is written on a level accessible to graduate students with good mathematical background.