21
Smoothing Using Orthogonal Functions

In this chapter we will study an approach to nonparametric curve estimation based on **orthogonal functions**. We begin with a brief introduction to the theory of orthogonal functions, then we turn to density estimation and regression.

21.1 Orthogonal Functions and L_2 Spaces

Let $v = (v_1, v_2, v_3)$ denote a three-dimensional vector, that is, a list of three real numbers. Let $\mathcal{V}$ denote the set of all such vectors. If a is a scalar (a number) and v is a vector, we define $av = (av_1, av_2, av_3)$. The sum of vectors v and w is defined by $v + w = (v_1 + w_1, v_2 + w_2, v_3 + w_3)$. The **inner product** between two vectors v and w is defined by $\langle v, w \rangle = \sum_{i=1}^{3} v_i w_i$. The **norm (or length)** of a vector v is defined by

$$||v|| = \sqrt{\langle v, v \rangle} = \sqrt{\sum_{i=1}^{3} v_i^2}. \tag{21.1}$$

Two vectors are **orthogonal (or perpendicular)** if $\langle v, w \rangle = 0$. A set of vectors are orthogonal if each pair in the set is orthogonal. A vector is **normal** if $||v|| = 1$.

L. Wasserman, *All of Statistics*

Let $\phi_1 = (1,0,0)$, $\phi_2 = (0,1,0)$, $\phi_3 = (0,0,1)$. These vectors are said to be an **orthonormal basis** for $\mathcal{V}$ since they have the following properties:
(i) they are orthogonal;
(ii) they are normal;
(iii) they form a basis for $\mathcal{V}$, which means that any $v \in \mathcal{V}$ can be written as a linear combination of ϕ_1, ϕ_2, ϕ_3:

$$v = \sum_{j=1}^{3} \beta_j \phi_j \quad \text{where} \quad \beta_j = \langle \phi_j, v \rangle. \tag{21.2}$$

For example, if $v = (12, 3, 4)$ then $v = 12\phi_1 + 3\phi_2 + 4\phi_3$. There are other orthonormal bases for $\mathcal{V}$, for example,

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right), \ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

You can check that these three vectors also form an orthonormal basis for $\mathcal{V}$. Again, if v is any vector then we can write

$$v = \sum_{j=1}^{3} \beta_j \psi_j \quad \text{where} \quad \beta_j = \langle \psi_j, v \rangle.$$

For example, if $v = (12, 3, 4)$ then

$$v = 10.97\psi_1 + 6.36\psi_2 + 2.86\psi_3.$$

Now we make the leap from vectors to functions. Basically, we just replace vectors with functions and sums with integrals. Let $L_2(a, b)$ denote all functions defined on the interval $[a, b]$ such that $\int_a^b f(x)^2 dx < \infty$:

$$L_2(a,b) = \left\{ f : [a,b] \to \mathbb{R}, \ \int_a^b f(x)^2 dx < \infty \right\}. \tag{21.3}$$

We sometimes write L_2 instead of $L_2(a, b)$. The inner product between two functions $f, g \in L_2$ is defined by $\int f(x)g(x)dx$. The norm of f is

$$||f|| = \sqrt{\int f(x)^2 dx.} \tag{21.4}$$

Two functions are orthogonal if $\int f(x)g(x)dx = 0$. A function is normal if $||f|| = 1$.

A sequence of functions $\phi_1, \phi_2, \phi_3, \ldots$ is **orthonormal** if $\int \phi_j^2(x)dx = 1$ for each j and $\int \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$. An orthonormal sequence is **complete** if the only function that is orthogonal to each ϕ_j is the zero function.

In this case, the functions $\phi_1, \phi_2, \phi_3, \ldots$ form in basis, meaning that if $f \in L_2$ then f can be written as[1]

$$f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x), \quad \text{where } \beta_j = \int_a^b f(x)\phi_j(x)dx. \tag{21.5}$$

A useful result is **Parseval's relation** which says that

$$||f||^2 \equiv \int f^2(x)\,dx = \sum_{j=1}^{\infty} \beta_j^2 \equiv ||\beta||^2 \tag{21.6}$$

where $\beta = (\beta_1, \beta_2, \ldots)$.

21.1 Example. An example of an orthonormal basis for $L_2(0,1)$ is the **cosine basis** defined as follows. Let $\phi_0(x) = 1$ and for $j \geq 1$ define

$$\phi_j(x) = \sqrt{2}\cos(j\pi x). \tag{21.7}$$

The first six functions are plotted in Figure 21.1. ■

21.2 Example. Let

$$f(x) = \sqrt{x(1-x)}\,\sin\left(\frac{2.1\pi}{(x+.05)}\right)$$

which is called the "doppler function." Figure 21.2 shows f (top left) and its approximation

$$f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$$

with J equal to 5 (top right), 20 (bottom left), and 200 (bottom right). As J increases we see that $f_J(x)$ gets closer to $f(x)$. The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically. ■

21.3 Example. The **Legendre polynomials** on $[-1,1]$ are defined by

$$P_j(x) = \frac{1}{2^j j!}\frac{d^j}{dx^j}(x^2-1)^j, \quad j = 0,1,2,\ldots \tag{21.8}$$

It can be shown that these functions are complete and orthogonal and that

$$\int_{-1}^{1} P_j^2(x)dx = \frac{2}{2j+1}. \tag{21.9}$$

[1] The equality in the displayed equation means that $\int (f(x) - f_n(x))^2 dx \to 0$ where $f_n(x) = \sum_{j=1}^{n} \beta_j \phi_j(x)$.

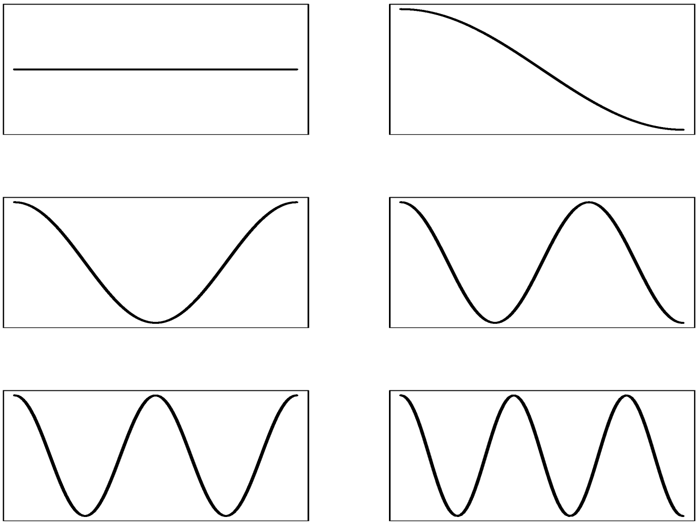

FIGURE 21.1. The first six functions in the cosine basis.

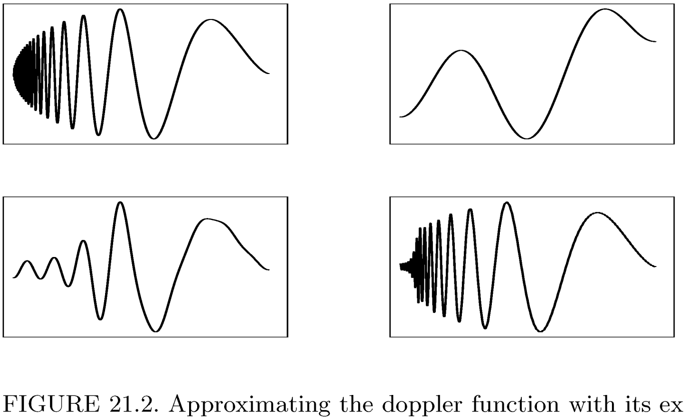

FIGURE 21.2. Approximating the doppler function with its expansion in the cosine basis. The function f (top left) and its approximation $f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ with J equal to 5 (top right), 20 (bottom left), and 200 (bottom right). The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically.

It follows that the functions $\phi_j(x) = \sqrt{(2j+1)/2}P_j(x)$, $j = 0, 1, \ldots$ form an orthonormal basis for $L_2(-1,1)$. The first few Legendre polynomials are:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{1}{2}\Big(3x^2 - 1\Big), \text{ and} \\ P_3(x) &= \frac{1}{2}\Big(5x^3 - 3x\Big). \end{aligned}$$

These polynomials may be constructed explicitly using the following recursive relation:

$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1}. \quad \blacksquare \tag{21.10}$$

The coefficients $\beta_1, \beta_2, \ldots$ are related to the smoothness of the function f. To see why, note that if f is smooth, then its derivatives will be finite. Thus we expect that, for some k, $\int_0^1 (f^{(k)}(x))^2 dx < \infty$ where $f^{(k)}$ is the k^{th} derivative of f. Now consider the cosine basis (21.7) and let $f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x)$. Then,

$$\int_0^1 (f^{(k)}(x))^2 dx = 2\sum_{j=1}^{\infty} \beta_j^2 (\pi j)^{2k}.$$

The only way that $\sum_{j=1}^{\infty} \beta_j^2 (\pi j)^{2k}$ can be finite is if the β_j's get small when j gets large. To summarize:

> If the function f is smooth, then the coefficients β_j will be small when j is large.

For the rest of this chapter, assume we are using the cosine basis unless otherwise specified.

21.2 Density Estimation

Let $X_1, \ldots, X_n$ be IID observations from a distribution on $[0,1]$ with density f. Assuming $f \in L_2$ we can write

$$f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x)$$

where $\phi_1, \phi_2, \ldots$ is an orthonormal basis. Define

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i). \tag{21.11}$$

21.4 Theorem. *The mean and variance of $\widehat{\beta}_j$ are*

$$\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j, \quad \mathbb{V}\left(\widehat{\beta}_j\right) = \frac{\sigma_j^2}{n} \tag{21.12}$$

where

$$\sigma_j^2 = \mathbb{V}(\phi_j(X_i)) = \int (\phi_j(x) - \beta_j)^2 f(x)dx. \tag{21.13}$$

PROOF. The mean is

$$\begin{aligned} \mathbb{E}\left(\widehat{\beta}_j\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\phi_j(X_i)\right) \\ &= \mathbb{E}\left(\phi_j(X_1)\right) \\ &= \int \phi_j(x) f(x)dx = \beta_j. \end{aligned}$$

The calculation for the variance is similar. ■

Hence, $\widehat{\beta}_j$ is an unbiased estimate of β_j. It is tempting to estimate f by $\sum_{j=1}^{\infty} \widehat{\beta}_j \phi_j(x)$ but this turns out to have a very high variance. Instead, consider the estimator

$$\widehat{f}(x) = \sum_{j=1}^{J} \widehat{\beta}_j \phi_j(x). \tag{21.14}$$

The number of terms J is a smoothing parameter. Increasing J will decrease bias while increasing variance. For technical reasons, we restrict J to lie in the range

$$1 \leq J \leq p$$

where $p = p(n) = \sqrt{n}$. To emphasize the dependence of the risk function on J, we write the risk function as $R(J)$.

21.5 Theorem. *The risk of $\widehat{f}$ is*

$$R(J) = \sum_{j=1}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2. \tag{21.15}$$

An estimate of the risk is

$$\widehat{R}(J) = \sum_{j=1}^{J} \frac{\widehat{\sigma}_j^2}{n} + \sum_{j=J+1}^{p} \left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}_j^2}{n}\right)_+ \tag{21.16}$$

where $a_+ = \max\{a, 0\}$ and

$$\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\phi_j(X_i) - \widehat{\beta}_j\right)^2. \tag{21.17}$$

To motivate this estimator, note that $\widehat{\sigma}_j^2$ is an unbiased estimate of σ_j^2 and $\widehat{\beta}_j^2 - \widehat{\sigma}_j^2$ is an unbiased estimator of β_j^2. We take the positive part of the latter term since we know that β_j^2 cannot be negative. We now choose $1 \leq \widehat{J} \leq p$ to minimize $\widehat{R}(\widehat{f}, f)$. Here is a summary:

Summary of Orthogonal Function Density Estimation

1. Let
$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i).$$
2. Choose $\widehat{J}$ to minimize $\widehat{R}(J)$ over $1 \leq J \leq p = \sqrt{n}$ where $\widehat{R}$ is given in equation (21.16).
3. Let
$$\widehat{f}(x) = \sum_{j=1}^{\widehat{J}} \widehat{\beta}_j \phi_j(x).$$

The estimator $\widehat{f}_n$ can be negative. If we are interested in exploring the shape of f, this is not a problem. However, if we need our estimate to be a probability density function, we can truncate the estimate and then normalize it. That is, we take $\widehat{f}^* = \max\{\widehat{f}_n(x), 0\} / \int_0^1 \max\{\widehat{f}_n(u), 0\} du$.

Now let us construct a confidence band for f. Suppose we estimate f using J orthogonal functions. We are essentially estimating $f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ not the true density $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$. Thus, the confidence band should be regarded as a band for $f_J(x)$.

21.6 Theorem. *An approximate* $1 - \alpha$ *confidence band for* f_J *is* $(\ell(x), u(x))$ *where*
$$\ell(x) = \widehat{f}_n(x) - c, \quad u(x) = \widehat{f}_n(x) + c \tag{21.18}$$
where
$$c = K^2 \sqrt{\frac{J \chi^2_{J,\alpha}}{n}} \tag{21.19}$$
and
$$K = \max_{1 \leq j \leq J} \max_x |\phi_j(x)|.$$
For the cosine basis, $K = \sqrt{2}$.

Proof. Here is an outline of the proof. Let $L = \sum_{j=1}^{J} (\widehat{\beta}_j - \beta_j)^2$. By the central limit theorem, $\widehat{\beta}_j \approx N(\beta_j, \sigma_j^2/n)$. Hence, $\widehat{\beta}_j \approx \beta_j + \sigma_j \epsilon_j / \sqrt{n}$ where

$\epsilon_j \sim N(0,1)$, and therefore

$$L \approx \frac{1}{n}\sum_{j=1}^{J}\sigma_j^2\epsilon_j^2 \leq \frac{K^2}{n}\sum_{j=1}^{J}\epsilon_j^2 \overset{d}{=} \frac{K^2}{n}\chi_J^2. \tag{21.20}$$

Thus we have, approximately, that

$$\mathbb{P}\left(L > \frac{K^2}{n}\chi_{J,\alpha}^2\right) \leq \mathbb{P}\left(\frac{K^2}{n}\chi_J^2 > \frac{K^2}{n}\chi_{J,\alpha}^2\right) = \alpha.$$

Also,

$$\begin{aligned}
\max_x |\widehat{f}_J(x) - f_J(x)| &\leq \max_x \sum_{j=1}^{J} |\phi_j(x)|\, |\widehat{\beta}_j - \beta_j| \\
&\leq K\sum_{j=1}^{J} |\widehat{\beta}_j - \beta_j| \\
&\leq \sqrt{J}\,K\sqrt{\sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j)^2} \\
&= \sqrt{J}\,K\sqrt{L}
\end{aligned}$$

where the third inequality is from the Cauchy-Schwartz inequality (Theorem 4.8). So,

$$\begin{aligned}
\mathbb{P}\left(\max_x |\widehat{f}_J(x) - f_J(x)| > K^2\sqrt{\frac{J\chi_{J,\alpha}^2}{n}}\right) &\leq \mathbb{P}\left(\sqrt{J}\,K\sqrt{L} > K^2\sqrt{\frac{J\chi_{J,\alpha}^2}{n}}\right) \\
&= \mathbb{P}\left(\sqrt{L} > K\sqrt{\frac{\chi_{J,\alpha}^2}{n}}\right) \\
&= \mathbb{P}\left(L > \frac{K^2\chi_{J,\alpha}^2}{n}\right) \\
&\leq \alpha. \quad \blacksquare
\end{aligned}$$

21.7 Example. Let

$$f(x) = \frac{5}{6}\phi(x;0,1) + \frac{1}{6}\sum_{j=1}^{5}\phi(x;\mu_j,.1)$$

where $\phi(x;\mu,\sigma)$ denotes a Normal density with mean μ and standard deviation σ, and $(\mu_1,\ldots,\mu_5) = (-1,-1/2,0,1/2,1)$. Marron and Wand (1992) call this

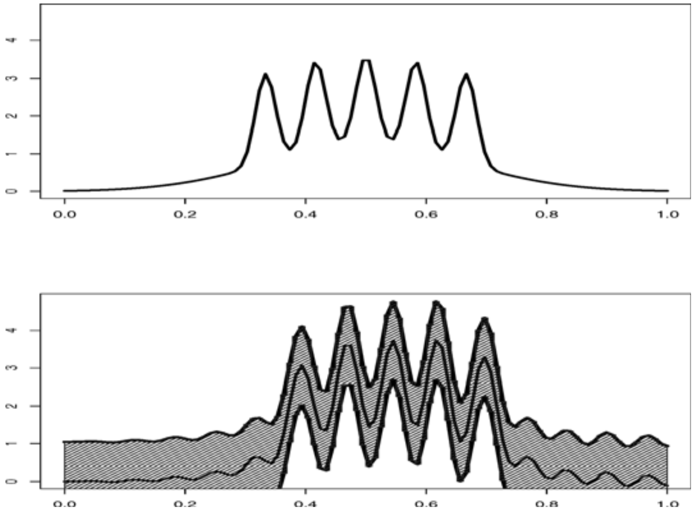

FIGURE 21.3. The top plot is the true density for the Bart Simpson distribution (rescaled to have most of its mass between 0 and 1). The bottom plot is the orthogonal function density estimate and 95 percent confidence band.

"the claw" although the "Bart Simpson" might be more appropriate. Figure 21.3 shows the true density as well as the estimated density based on $n = 5,000$ observations and a 95 percent confidence band. The density has been rescaled to have most of its mass between 0 and 1 using the transformation $y = (x + 3)/6$. ■

21.3 Regression

Consider the regression model

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{21.21}$$

where the ϵ_i are independent with mean 0 and variance σ^2. We will initially focus on the special case where $x_i = i/n$. We assume that $r \in L_2(0, 1)$ and hence we can write

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \quad \text{where} \quad \beta_j = \int_0^1 r(x)\phi_j(x)dx \tag{21.22}$$

where $\phi_1, \phi_2, \ldots$ where is an orthonormal basis for $[0, 1]$.

Define

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\,\phi_j(x_i), \quad j = 1, 2, \ldots \tag{21.23}$$

Since $\widehat{\beta}_j$ is an average, the central limit theorem tells us that $\widehat{\beta}_j$ will be approximately Normally distributed.

21.8 Theorem.

$$\widehat{\beta}_j \approx N\left(\beta_j, \frac{\sigma^2}{n}\right). \tag{21.24}$$

Proof. The mean of $\widehat{\beta}_j$ is

$$\begin{aligned}\mathbb{E}(\widehat{\beta}_j) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i)\phi_j(x_i) = \frac{1}{n}\sum_{i=1}^{n} r(x_i)\phi_j(x_i) \\ &\approx \int r(x)\phi_j(x)dx = \beta_j\end{aligned}$$

where the approximate equality follows from the definition of a Riemann integral: $\sum_i \Delta_n h(x_i) \to \int_0^1 h(x)dx$ where $\Delta_n = 1/n$. The variance is

$$\begin{aligned}\mathbb{V}(\widehat{\beta}_j) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(Y_i)\phi_j^2(x_i) \\ &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\phi_j^2(x_i) = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(x_i) \\ &\approx \frac{\sigma^2}{n}\int \phi_j^2(x)dx = \frac{\sigma^2}{n}\end{aligned}$$

since $\int \phi_j^2(x)dx = 1$. ■

Let

$$\widehat{r}(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x),$$

and let

$$R(J) = \mathbb{E}\int (r(x) - \widehat{r}(x))^2\,dx$$

be the risk of the estimator.

21.9 Theorem. *The risk $R(J)$ of the estimator $\widehat{r}_n(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x)$ is*

$$R(J) = \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty}\beta_j^2. \tag{21.25}$$

To estimate for $\sigma^2 = \mathbb{V}(\epsilon_i)$ we use

$$\widehat{\sigma}^2 = \frac{n}{k}\sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \tag{21.26}$$

where $k = n/4$. To motivate this estimator, recall that if f is smooth, then $\beta_j \approx 0$ for large j. So, for $j \geq k$, $\widehat{\beta}_j \approx N(0, \sigma^2/n)$ and thus, $\widehat{\beta}_j \approx \sigma Z_j/\sqrt{n}$ for for $j \geq k$, where $Z_j \sim N(0,1)$. Therefore,

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{n}{k}\sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \approx \frac{n}{k}\sum_{i=n-k+1}^{n} \left(\frac{\sigma}{\sqrt{n}}\widehat{\beta}_j\right)^2 \\ &= \frac{\sigma^2}{k}\sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 = \frac{\sigma^2}{k}\chi_k^2\end{aligned}$$

since a sum of k Normals has a χ_k^2 distribution. Now $\mathbb{E}(\chi_k^2) = k$ and hence $\mathbb{E}(\widehat{\sigma}^2) \approx \sigma^2$. Also, $\mathbb{V}(\chi_k^2) = 2k$ and hence $\mathbb{V}(\widehat{\sigma}^2) \approx (\sigma^4/k^2)(2k) = (2\sigma^4/k) \to 0$ as $n \to \infty$. Thus we expect $\widehat{\sigma}^2$ to be a consistent estimator of σ^2. There is nothing special about the choice $k = n/4$. Any k that increases with n at an appropriate rate will suffice.

We estimate the risk with

$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n}\left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n}\right)_+ . \tag{21.27}$$

21.10 Example. Figure 21.4 shows the doppler function f and $n = 2,048$ observations generated from the model

$$Y_i = r(x_i) + \epsilon_i$$

where $x_i = i/n$, $\epsilon_i \sim N(0, (.1)^2)$. The figure shows the data and the estimated function. The estimate was based on $\widehat{J} = 234$ terms. ■

We are now ready to give a complete description of the method.

Orthogonal Series Regression Estimator

1. Let

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_j(x_i), \quad j = 1, \ldots, n.$$

2. Let

$$\widehat{\sigma}^2 = \frac{n}{k}\sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \tag{21.28}$$

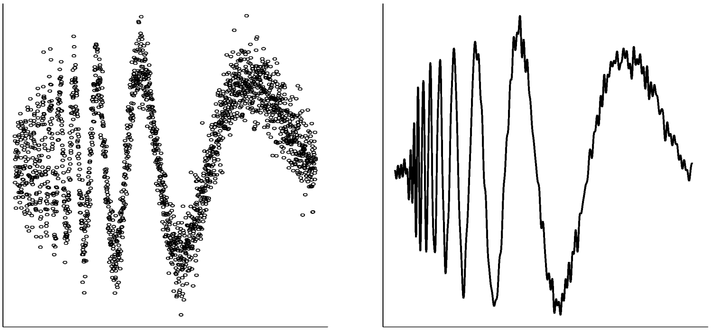

FIGURE 21.4. Data from the doppler test function and the estimated function. See Example 21.10.

where $k \approx n/4$.

3. For $1 \leq J \leq n$, compute the risk estimate

$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n} \left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n}\right)_+ .$$

4. Choose $\widehat{J} \in \{1, \ldots n\}$ to minimize $\widehat{R}(J)$.

5. Let

$$\widehat{r}(x) = \sum_{j=1}^{\widehat{J}} \widehat{\beta}_j \phi_j(x).$$

Finally, we turn to confidence bands. As before, these bands are not really for the true function $r(x)$ but rather for the smoothed version of the function $r_J(x) = \sum_{j=1}^{\widehat{J}} \beta_j \phi_j(x)$.

21.11 Theorem. *Suppose the estimate $\widehat{r}$ is based on J terms and $\widehat{\sigma}$ is defined as in equation (21.28). Assume that $J < n - k + 1$. An approximate $1 - \alpha$ confidence band for r_J is (ℓ, u) where*

$$\ell(x) = \widehat{r}_n(x) - c, \quad u(x) = \widehat{r}_n(x) + c, \tag{21.29}$$

where

$$c = \frac{a(x)\ \widehat{\sigma}\ \chi_{J,\alpha}}{\sqrt{n}}, \quad a(x) = \sqrt{\sum_{j=1}^{J} \phi_j^2(x)},$$

and $\widehat{\sigma}$ is given in equation (21.28).

Proof. Let $L = \sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j)^2$. By the central limit theorem, $\widehat{\beta}_j \approx N(\beta_j, \sigma^2/n)$. Hence, $\widehat{\beta}_j \approx \beta_j + \sigma\epsilon_j/\sqrt{n}$ where $\epsilon_j \sim N(0,1)$ and therefore

$$L \approx \frac{\sigma^2}{n}\sum_{j=1}^{J}\epsilon_j^2 \overset{d}{=} \frac{\sigma^2}{n}\chi_J^2.$$

Thus,

$$\mathbb{P}\left(L > \frac{\sigma^2}{n}\chi_{J,\alpha}^2\right) = \mathbb{P}\left(\frac{\sigma^2}{n}\chi_J^2 > \frac{\sigma^2}{n}\chi_{J,\alpha}^2\right) = \alpha.$$

Also,

$$\begin{aligned}
|\widehat{r}(x) - r_J(x)| &\leq \sum_{j=1}^{J}|\phi_j(x)|\,|\widehat{\beta}_j - \beta_j| \\
&\leq \sqrt{\sum_{j=1}^{J}\phi_j^2(x)}\sqrt{\sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j^2)} \\
&\leq a(x)\,\sqrt{L}
\end{aligned}$$

by the Cauchy-Schwartz inequality (Theorem 4.8). So,

$$\begin{aligned}
\mathbb{P}\left(\max_x \frac{|\widehat{f}_J(x) - \overline{f}(x)|}{a(x)} > \frac{\widehat{\sigma}\chi_{J,\alpha}}{\sqrt{n}}\right) &\leq \mathbb{P}\left(\sqrt{L} > \frac{\widehat{\sigma}\chi_{J,\alpha}}{\sqrt{n}}\right) \\
&= \alpha
\end{aligned}$$

and the result follows. ■

21.12 Example. Figure 21.5 shows the confidence envelope for the doppler signal. The first plot is based on $J = 234$ (the value of J that minimizes the estimated risk). The second is based on $J = 45 \approx \sqrt{n}$. Larger J yields a higher resolution estimator at the cost of large confidence bands. Smaller J yields a lower resolution estimator but has tighter confidence bands. ■

So far, we have assumed that the x_i's are of the form $\{1/n, 2/n, \ldots, 1\}$. If the x_i's are on interval $[a, b]$, then we can rescale them so that are in the interval $[0, 1]$. If the x_i's are not equally spaced, the methods we have discussed still apply so long as the x_i's "fill out" the interval [0,1] in such a way so as to not be too clumped together. If we want to treat the x_i's as random instead of fixed, then the method needs significant modifications which we shall not deal with here.

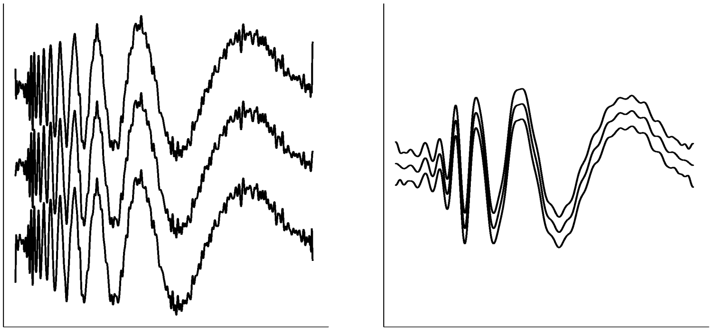

FIGURE 21.5. Estimates and confidence bands for the doppler test function using $n = 2,048$ observations. First plot: $J = 234$ terms. Second plot: $J = 45$ terms.

21.4 Wavelets

Suppose there is a sharp jump in a regression function f at some point x but that f is otherwise very smooth. Such a function f is said to be **spatially inhomogeneous**. The doppler function is an example of a spatially inhomogeneous function; it is smooth for large x and unsmooth for small x.

It is hard to estimate f using the methods we have discussed so far. If we use a cosine basis and only keep low order terms, we will miss the peak; if we allow higher order terms we will find the peak but we will make the rest of the curve very wiggly. Similar comments apply to kernel regression. If we use a large bandwidth, then we will smooth out the peak; if we use a small bandwidth, then we will find the peak but we will make the rest of the curve very wiggly.

One way to estimate inhomogeneous functions is to use a more carefully chosen basis that allows us to place a "blip" in some small region without adding wiggles elsewhere. In this section, we describe a special class of bases called **wavelets**, that are aimed at fixing this problem. Statistical inference using wavelets is a large and active area. We will just discuss a few of the main ideas to get a flavor of this approach.

We start with a particular wavelet called the **Haar wavelet.** The **Haar father wavelet** or **Haar scaling function** is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{21.30}$$

The **mother Haar wavelet** is defined by

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \le x \le \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < x \le 1. \end{cases} \tag{21.31}$$

For any integers j and k define

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{21.32}$$

The function $\psi_{j,k}$ has the same shape as ψ but it has been rescaled by a factor of $2^{j/2}$ and shifted by a factor of k.

See Figure 21.6 for some examples of Haar wavelets. Notice that for large j, $\psi_{j,k}$ is a very localized function. This makes it possible to add a blip to a function in one place without adding wiggles elsewhere. Increasing j is like looking in a microscope at increasing degrees of resolution. In technical terms, we say that wavelets provide a **multiresolution analysis** of $L_2(0,1)$.

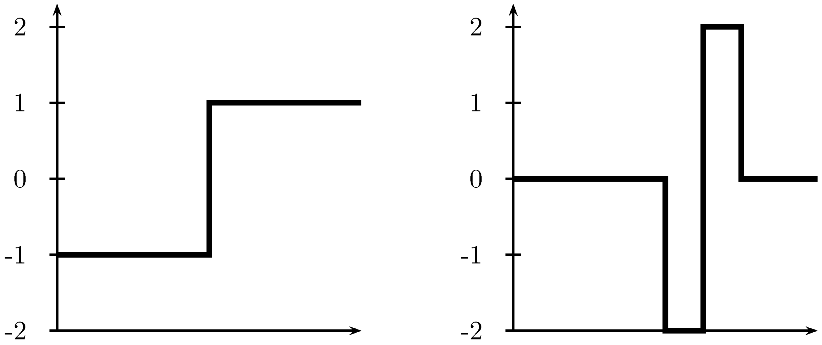

FIGURE 21.6. Some Haar wavelets. Left: the mother wavelet $\psi(x)$; Right: $\psi_{2,2}(x)$.

Let

$$W_j = \{\psi_{jk},\ k = 0, 1, \ldots, 2^j - 1\}$$

be the set of rescaled and shifted mother wavelets at resolution j.

21.13 Theorem. *The set of functions*

$$\left\{\phi, W_0, W_1, W_2, \ldots, \right\}$$

is an orthonormal basis for $L_2(0,1)$.

It follows from this theorem that we can expand any function $f \in L_2(0,1)$ in this basis. Because each W_j is itself a set of functions, we write the expansion as a double sum:

$$f(x) = \alpha\,\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x) \tag{21.33}$$

where

$$\alpha = \int_0^1 f(x)\phi(x)\,dx, \quad \beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)\,dx.$$

We call α the **scaling coefficient** and the $\beta_{j,k}$'s are called the **detail coefficients**. We call the finite sum

$$f_J(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x) \tag{21.34}$$

the **resolution** J approximation to f. The total number of terms in this sum is

$$1 + \sum_{j=0}^{J-1} 2^j = 1 + 2^J - 1 = 2^J.$$

21.14 Example. Figure 21.7 shows the doppler signal, and its reconstruction using $J = 3, 5$ and $J = 8$. ■

Haar wavelets are localized, meaning that they are zero outside an interval. But they are not smooth. This raises the question of whether there exist smooth, localized wavelets that from an orthonormal basis. In 1988, Ingrid Daubechie showed that such wavelets do exist. These smooth wavelets are difficult to describe. They can be constructed numerically but there is no closed form formula for the smoother wavelets. To keep things simple, we will continue to use Haar wavelets.

Consider the regression model $Y_i = r(x_i) + \sigma\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $x_i = i/n$. To simplify the discussion we assume that $n = 2^J$ for some J.

There is one major difference between estimation using wavelets instead of a cosine (or polynomial) basis. With the cosine basis, we used all the terms $1 \leq j \leq J$ for some J. The number of terms J acted as a smoothing parameter. With wavelets, we control smoothing using a method called **thresholding** where we keep a term in the function approximation if its coefficient is large,

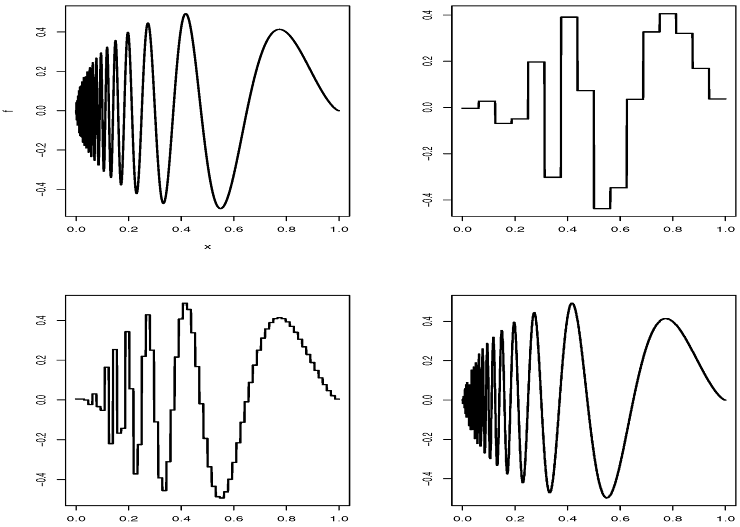

FIGURE 21.7. The doppler signal and its reconstruction $f_J(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_k \beta_{j,k}\psi_{j,k}(x)$ based on $J = 3$, $J = 5$, and $J = 8$.

otherwise, we throw out that term. There are many versions of thresholding. The simplest is called hard, universal thresholding. Let $J = \log_2(n)$ and define

$$\widehat{\alpha} = \frac{1}{n}\sum_i \phi_k(x_i)Y_i \quad \text{and} \quad D_{j,k} = \frac{1}{n}\sum_i \psi_{j,k}(x_i)Y_i \tag{21.35}$$

for $0 \leq j \leq J - 1$.

Haar Wavelet Regression

1. Compute $\widehat{\alpha}$ and $D_{j,k}$ as in (21.35), for $0 \leq j \leq J - 1$.
2. Estimate σ; see (21.37).
3. Apply universal thresholding:

$$\widehat{\beta}_{j,k} = \left\{ \begin{array}{ll} D_{j,k} & \text{if } |D_{j,k}| > \widehat{\sigma}\sqrt{\frac{2\log n}{n}} \\ 0 & \text{otherwise.} \end{array} \right\} \tag{21.36}$$

4. Set $\widehat{f}(x) = \widehat{\alpha}\phi(x) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\widehat{\beta}_{j,k}\psi_{j,k}(x)$.

In practice, we do not compute S_k and $D_{j,k}$ using (21.35). Instead, we use the **discrete wavelet transform (DWT)** which is very fast. The DWT for Haar wavelets is described in the appendix. The estimate of σ is

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\left(|D_{J-1,k}| : \ k = 0, \ldots, 2^{J-1} - 1\right)}{0.6745}. \tag{21.37}$$

The estimate for σ may look strange. It is similar to the estimate we used for the cosine basis but it is designed to be insensitive to sharp peaks in the function.

To understand the intuition behind universal thresholding, consider what happens when there is no signal, that is, when $\beta_{j,k} = 0$ for all j and k.

21.15 Theorem. *Suppose that $\beta_{j,k} = 0$ for all j and k and let $\widehat{\beta}_{j,k}$ be the universal threshold estimator. Then*

$$\mathbb{P}(\widehat{\beta}_{j,k} = 0 \text{ for all } j, k) \to 1$$

as $n \to \infty$.

Proof. To simplify the proof, assume that σ is known. Now $D_{j,k} \approx N(0, \sigma^2/n)$. We will need Mill's inequality (Theorem 4.7): if $Z \sim N(0,1)$ then $\mathbb{P}(|Z| > t) \le (c/t)e^{-t^2/2}$ where $c = \sqrt{2/\pi}$ is a constant. Thus,

$$\begin{aligned}
\mathbb{P}(\max |D_{j,k}| > \lambda) &\le \sum_{j,k} \mathbb{P}(|D_{j,k}| > \lambda) = \sum_{j,k} \mathbb{P}\left(\frac{\sqrt{n}|D_{j,k}|}{\sigma} > \frac{\sqrt{n}\lambda}{\sigma}\right) \\
&\le \sum_{j,k} \frac{c\sigma}{\lambda\sqrt{n}} \exp\left\{-\frac{1}{2}\frac{n\lambda^2}{\sigma^2}\right\} \\
&= \frac{c}{\sqrt{2\log n}} \to 0. \quad \blacksquare
\end{aligned}$$

21.16 Example. Consider $Y_i = r(x_i) + \sigma\epsilon_i$ where f is the doppler signal, $\sigma = .1$ and $n = 2,048$. Figure 21.8 shows the data and the estimated function using universal thresholding. Of course, the estimate is not smooth since Haar wavelets are not smooth. Nonetheless, the estimate is quite accurate. ■

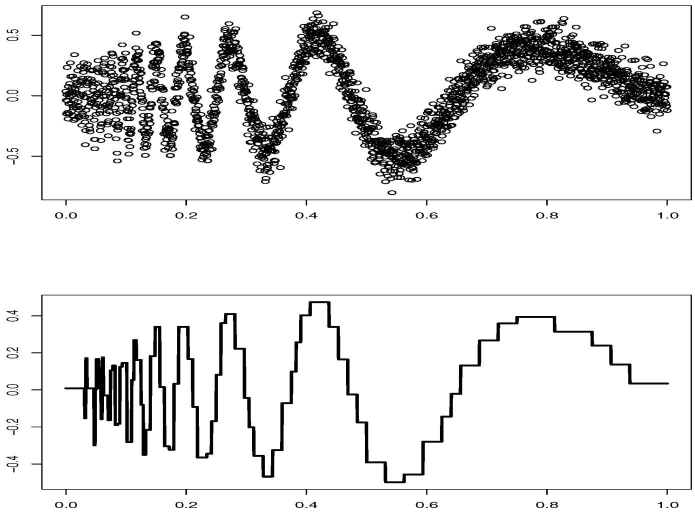

FIGURE 21.8. Estimate of the Doppler function using Haar wavelets and universal thresholding.

21.5 Appendix

The DWT for Haar Wavelets. Let y be the vector of Y_i's (length n) and let $J = \log_2(n)$. Create a list D with elements

$$D[[0]],\ \ldots,\ D[[J-1]].$$

Set:

$$temp \leftarrow y/\sqrt{n}.$$

Then do:

```
$for(j\ \ in\ \ (J-1):0)\{$
    $m \leftarrow 2^j$
    $I \leftarrow (1:m)$
    $D[[j]] \leftarrow \Big(temp[2*I] - temp[(2*I)-1]\Big)/\sqrt{2}$
    $temp \leftarrow \Big(temp[2*I] + temp[(2*I)-1]\Big)/\sqrt{2}$
$\}$
```

21.6 Bibliographic Remarks

Efromovich (1999) is a reference for orthogonal function methods. See also Beran (2000) and Beran and Dümbgen (1998). An introduction to wavelets is given in Ogden (1997). A more advanced treatment can be found in Härdle et al. (1998). The theory of statistical estimation using wavelets has been developed by many authors, especially David Donoho and Ian Johnstone. See Donoho and Johnstone (1994), Donoho and Johnstone (1995), Donoho et al. (1995), and Donoho and Johnstone (1998).

21.7 Exercises

1. Prove Theorem 21.5.

2. Prove Theorem 21.9.

3. Let

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right),\ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right),\ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

   Show that these vectors have norm 1 and are orthogonal.

4. Prove Parseval's relation equation (21.6).

5. Plot the first five Legendre polynomials. Verify, numerically, that they are orthonormal.

6. Expand the following functions in the cosine basis on $[0,1]$. For (a) and (b), find the coefficients β_j analytically. For (c) and (d), find the coefficients β_j numerically, i.e.

$$\beta_j = \int_0^1 f(x)\phi_j(x) \approx \frac{1}{N}\sum_{r=1}^{N} f\left(\frac{r}{N}\right)\phi_j\left(\frac{r}{N}\right)$$

   for some large integer N. Then plot the partial sum $\sum_{j=1}^{n} \beta_j \phi_j(x)$ for increasing values of n.

   (a) $f(x) = \sqrt{2}\cos(3\pi x)$.

   (b) $f(x) = \sin(\pi x)$.

   (c) $f(x) = \sum_{j=1}^{11} h_j K(x - t_j)$ where $K(t) = (1 + \text{sign}(t))/2$, $\text{sign}(x) = -1$ if $x < 0$, $\text{sign}(x) = 0$ if $x = 0$, $\text{sign}(x) = 1$ if $x > 0$,

$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81),$

$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2).$

(d) $f = \sqrt{x(1-x)} \sin\left(\frac{2.1\pi}{(x+.05)}\right).$

7. Consider the glass fragments data from the book's website. Let Y be refractive index and let X be aluminum content (the fourth variable).

   (a) Do a nonparametric regression to fit the model $Y = f(x) + \epsilon$ using the cosine basis method. The data are not on a regular grid. Ignore this when estimating the function. (But do sort the data first according to x.) Provide a function estimate, an estimate of the risk, and a confidence band.

   (b) Use the wavelet method to estimate f.

8. Show that the Haar wavelets are orthonormal.

9. Consider again the doppler signal:

$$f(x) = \sqrt{x(1-x)} \sin\left(\frac{2.1\pi}{x + 0.05}\right).$$

   Let $n = 1{,}024$, $\sigma = 0.1$, and let $(x_1, \ldots, x_n) = (1/n, \ldots, 1)$. Generate data

$$Y_i = f(x_i) + \sigma\epsilon_i$$

   where $\epsilon_i \sim N(0, 1)$.

   (a) Fit the curve using the cosine basis method. Plot the function estimate and confidence band for $J = 10, 20, \ldots, 100$.

   (b) Use Haar wavelets to fit the curve.

10. (Haar density Estimation.) Let $X_1, \ldots, X_n \sim f$ for some density f on $[0, 1]$. Let's consider constructing a wavelet histogram. Let ϕ and ψ be the Haar father and mother wavelet. Write

$$f(x) \approx \phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x)$$

   where $J \approx \log_2(n)$. Let

$$\widehat{\beta}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \psi_{j,k}(X_i).$$

(a) Show that $\widehat{\beta}_{j,k}$ is an unbiased estimate of $\beta_{j,k}$.

(b) Define the Haar histogram

$$\widehat{f}(x) = \phi(x) + \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \widehat{\beta}_{j,k} \psi_{j,k}(x)$$

for $0 \leq B \leq J - 1$.

(c) Find an approximate expression for the MSE as a function of B.

(d) Generate $n = 1,000$ observations from a Beta (15,4) density. Estimate the density using the Haar histogram. Use leave-one-out cross validation to choose B.

11. In this question, we will explore the motivation for equation (21.37). Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. Let

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\,(|X_1|, \ldots, |X_n|)}{0.6745}.$$

(a) Show that $\mathbb{E}(\widehat{\sigma}) = \sigma$.

(b) Simulate $n = 100$ observations from a N(0,1) distribution. Compute $\widehat{\sigma}$ as well as the usual estimate of σ. Repeat 1,000 times and compare the MSE.

(c) Repeat (b) but add some outliers to the data. To do this, simulate each observation from a N(0,1) with probability .95 and simulate each observation from a N(0,10) with probability .95.

12. Repeat question 6 using the Haar basis.