Eco 523 - Class 4

Nonparametric Econometrics

Carolina Caetano*

Contents

1 Wavelets

Wavelets are basis systems that approximate simultaneously the position of the function variation and the frequency of the variation itself. They are automatically adapted for approximating desired levels of frequency in a sequential manner. This means that you can control exactly the level of resolution that you wish your approximation to have. Series estimation with wavelets is as simple as any other method. In practice, though, much fewer elements are necessary to achieve excellent approximation results. Computationally, there are optimized algorithms that make estimation with wavelets even faster than other orthogonal series. Among all the series approaches, none is as efficient in catching local behavior as wavelets. It is in fact excellent for fitting kinks and discontinuities.

A wavelet is a wave oscillation that starts at zero, increases, then decreases and goes back to zero. In order to explain how these bases work, it is a good idea to begin with an example. The Haar wavelet basis is the oldest, and probably best known of all. It is not the best wavelet, because it is not differentiable, but if the function is known to be discontinuous it is an excellent choice. The Haar basis was proposed in 1909, and at the

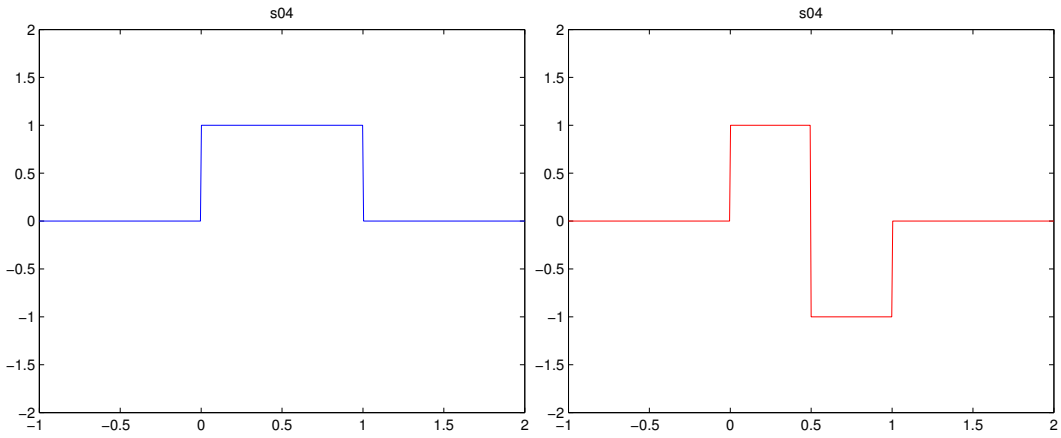
*Special thanks to Andrew Baynes for help with plots and editing.

time wavelets were not known. Only later researchers realized that these were in fact the most basic wavelets of all.

Consider the following functions:

$$\varphi(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\psi(x) = \begin{cases} 1 & \text{if } 0 < x < 1/2 \\ \text{-1} & \text{if } 1/2 \leqslant x < 1 \\ 0 & \text{otherwise} \end{cases}$$



The function φ is known as the **father wavelet**, and the function ψ is known as the **mother wavelet**. The father wavelet is not a real wavelet, it gives the general level of the function. It is the mother wavelet that gives the variation of the function around the level. Observe that the mother and father wavelets are orthogonal.

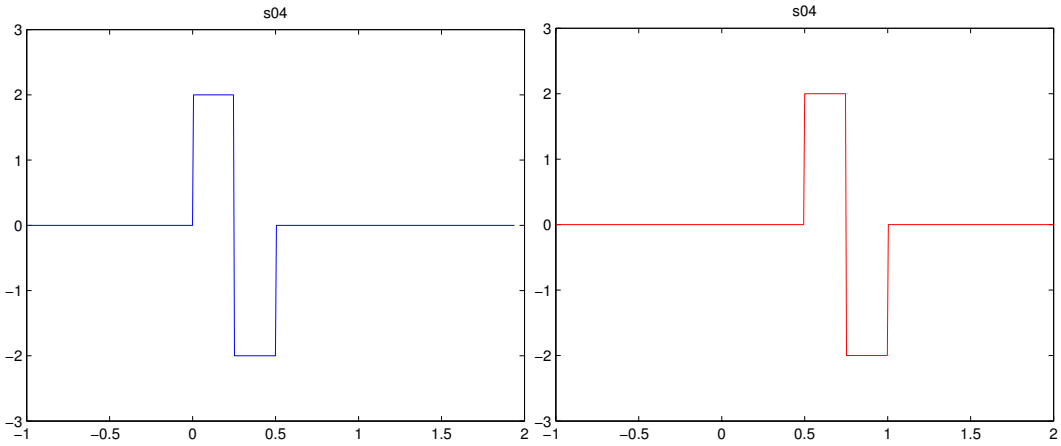
The wavelet basis is built by dividing the support of the mother wavelet by 2, while keeping the wavelet norm equal to one, and then doubling the number of wavelets in the basis. Suppose that the support is $[0, 1]$. Then the first resolution level of the basis is simply composed of the mother and father wavelets. We will name them:

$$\varphi_0(x) = \varphi(x), \quad \text{and} \quad \psi_{0,0}(x) = \psi(x).$$

The second resolution level (finer) is composed by adding the following wavelets:

$$\psi_{1,0}(x) = \begin{cases} \sqrt{2} & \text{if } 0 < x < 1/4 \\ -\sqrt{2} & \text{if } 1/4 \leqslant x < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_{1,1}(x) = \begin{cases} \sqrt{2} & \text{if } 1/2 \leqslant x < 3/4 \\ -\sqrt{2} & \text{if } 3/4 \leqslant x < 1 \\ 0 & \text{otherwise} \end{cases}$$



Observe that all these wavelets are orthogonal to each other and to the previous resolution wavelets. If you will use the next level of resolution in the basis, you must include all the elements of the basis of this resolution level. Observe also that all the new resolution wavelets still have norm one.

Each new step to increase the resolution consists of the inclusion of the new wavelets. For example, resolution level j means the inclusion of 2^j new ψ, with the values:

$$\psi_{j,k}(x) = 2^{j/2}\psi\left(2^j x - k\right), \quad k = 0, \ldots, 2^j - 1$$

If we did this expansion *ad infinitum*, the resulting basis spans $L^2[0,1]$. Hence, a function $g \in L^2[0,1]$ can be expressed as

$$g(x) = c_0\,\varphi_0(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^{j-1}} d_{j,k}\,\psi_{j,k}(x).$$

In order to span the entire L^2, we can do the same that we we did for $[0,1]$ for the rest of $\mathbb{R}$. For this, we just need to use the locators for the rest of the real line. For example,

let $\varphi_k(x) = \varphi(x-k)$, then the expansion for $g \in L^2$ is

$$g(x) = \sum_{k=-\infty}^{\infty} c_k \, \varphi_k(x) + \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} d_{j,k} \, \psi_{j,k}(x).$$

Though formally the expansion result is valid for L^2, wavelets are able to approximate other functions equally well. For instance, polynomials are not in L^2, and yet Haar wavelets can approximate them.

The same result that holds for the Haar basis holds for other functions as well. The wavelet arises from the definition of the mother wavelet, which should satisfy:

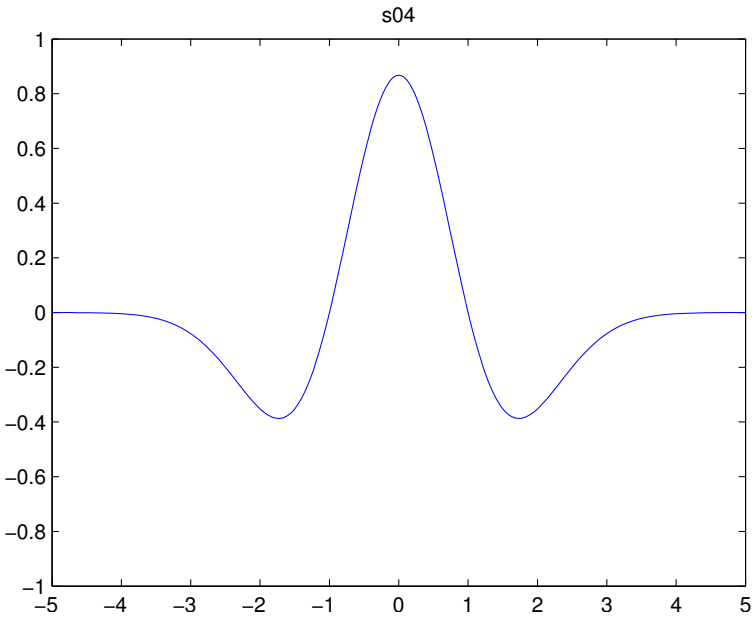
$$(i) \int |\psi(x)| dx < \infty$$
$$(ii) \int \psi(x)^2 dx = 1$$
$$(iii) \int x^m \psi(x) = 0$$

While the father wavelet is defined as

$$\varphi(x) = |\psi(x)|.$$

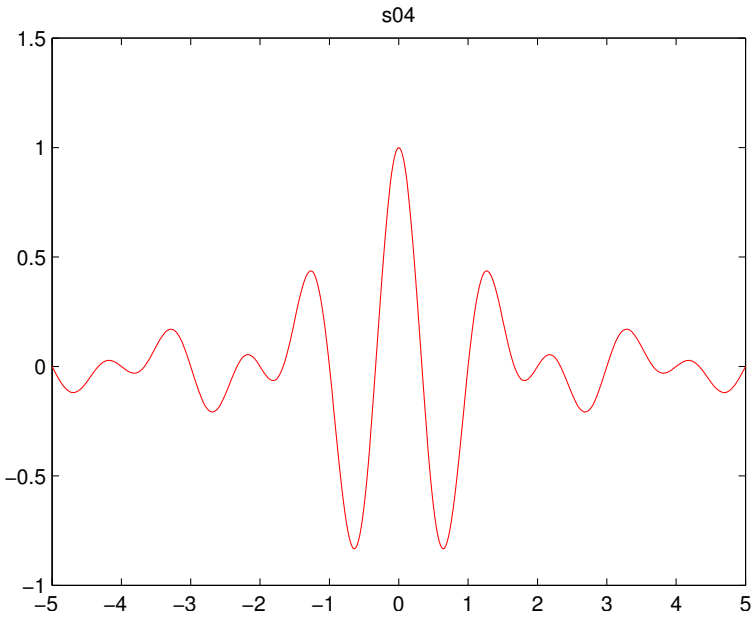
There are many famous wavelet functions. Some of the better known are the Mexican hat wavelet:

$$\psi(x) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}} \left(1 - \frac{x^2}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}},$$



the Shannon wavelet:

$$\psi(x) = \left(\frac{\pi x}{2}\right)^{-1} \cdot \sin\left(\frac{\pi x}{2}\right) \cdot \cos\left(\frac{3\pi x}{2}\right),$$



among others. Perhaps one of the most successful is the Daubechies wavelet, which has a fractal structure (when you zoom in the micro-behavior of this wavelet, it is an exact

reproduction of its macro-behavior, a property called self-similarity).

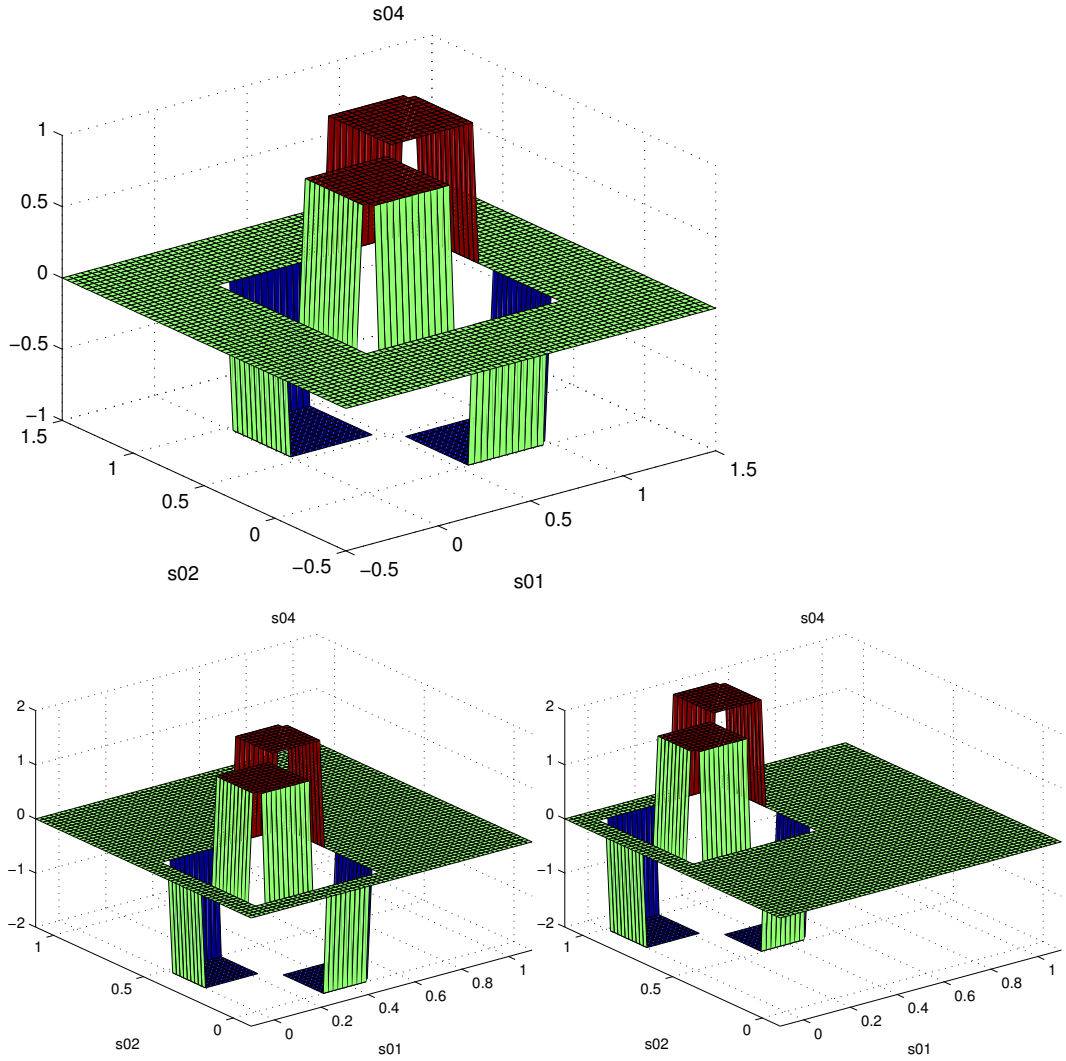
The wavelet structure of starting from zero and then returning to it is suited for estimation of singularities, such as kinks and discontinuities. As a general rule, smoother wavelets are preferable to discontinuous ones, but if we know that the function is discontinuous it is a good idea to fit it with a discontinuous wavelet, particularly if we plan to use a small amount of elements. Though fitting discontinuous functions with smooth wavelets is slightly less desirable, the boundary bias does disappear at an impressive rate. Usually a resolution of 5 or 6 is enough to kill all the naked-eye visible bias in plots printed in letter-sized paper sheets. There is no real Gibb's effect because the oscillation of the wavelet is always equal to the oscillation of the mother wavelet, never increasing.

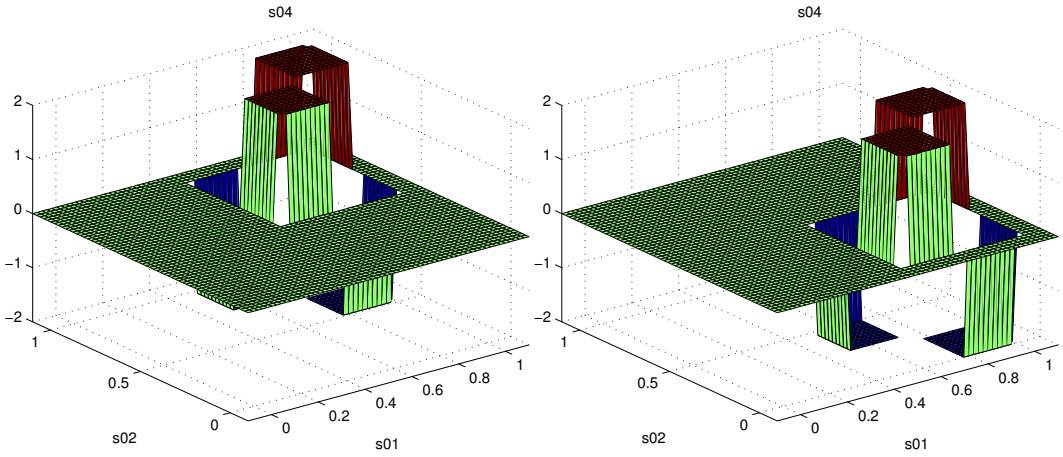
Wavelet bases are also desirable in that they are automatically linked to how good is the intended approximation. By definition each new level increases the resolution by twice the definition of the last step. Hence, if the scale is measured in inches, a 5 level approximation shows almost no naked-eye difference between the real signal and the approximation. Because of their relation to resolution, coefficients of wavelet approximations decrease at an incredibly large speed. Usually we start seeing zero coefficients around 6 or 7 degrees of resolution for discontinuous functions, and a lot earlier for smooth functions. No other method can claim this level of speed in approximation. This characteristic is robust for very wide classes of functions. Because of this, wavelet bases are qualified as **unconditional bases**, which means that they have certain optimality characteristics in very wide classes of functions.

From the computational point of view, it is usually not a good idea to program the wavelet bases on your own. The reason is that wavelet have excellent mathematical properties that allow computation to be optimized in many ways. For example, it is possible to calculate the coefficients of a wavelet expansion from the coefficients of the higher resolution terms through "filter banks." Because of this, computer burden can be optimized to $O(n)$ levels. Hence, it is a good idea to use an algorithm that is already optimized for these calculations. I am not aware of ready-made Stata codes for wavelets, but there is a wavelet toolbox for MatLab. In the problem set you will be asked to program a Haar basis yourself, so you can have a closer understanding of the workings of wavelets, but if you would like to use wavelets in actual applied work, it is a good idea to familiarize yourself with a ready package.

Multivariate wavelets can be built using tensor products (as seen with splines), or by using a multivariate wavelet. Many multivariate versions of the common wavelets have been developed for two-dimensional spaces, mainly because wavelets are widely used in digital image transmission for HDTV. The transformation to multidimensional spaces is straight-forward. It simply requires partitioning the space in equal cells and defining the

mother wavelet in these cells. Then the increasing resolutions are achieved by dividing each cell in two and scaling the mother wavelet so its support is the new cell, and so on. This can be very easily done by hand for multiple dimensions with a Haar basis.





2 Choice of K

For this section we will denote the number of elements from the basis used in the regression by K. The real choice depends on the kind of basis used. For the Fourier basis, one increases 2 basis elements at a time: $\sin(kx)$ and $\cos(kx)$. For Power series, include the next degree of the polynomial, which will be a function of x^k. For spline basis, the basis size is directly related to the partition structure. In the case of wavelets, each new resolution level increases the basis by two times the number of elements added on the previous resolution level. The point is that though the theory is made based on the number of elements in the basis K, each type of basis must increase in a manner that is coherent with the structure of the basis, which is not always one element at a time.

There are three methods commonly used. All are based on the cross-validation technique. The first two methods shown below combine a minimization of the squared errors with a penalty for lack of smoothness (determined by K). The third is the classic cross-validation.

2.1 Mallows' C_L (Mallows (1973))

$$\hat{K}_C = arg\min_K \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2 + 2\sigma^2\frac{K}{n},$$

where the undersmoothing penalty σ^2 can be estimated by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2$.

Hence,

$$\hat{K}_C = arg\min_K \left(1 + \frac{2K}{n}\right) \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}(X_i))^2.$$

2.2 Generalized cross-validation (Craven and Wahba (1979))

$$\hat{K}_{GCV} = arg\min_K \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2}{\left(1 - \frac{K}{n}\right)^2}.$$

2.3 Leave-one-out cross-validation (Stone (1974))

Select $\hat{K}_{CV}$ to minimize

$$CV_K = \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2,$$

where $\hat{g}_{-i}(X_i)$ is obtained by doing exactly the same procedure as for $\hat{g}(X_i)$, except we remove the i-th observation. This method is more intuitive, but is as hard to perform here as it was for the kernel cases, because CV_K requires n regressions per K. However, the previous two methods require only one regression at all. They are in fact so easy that it is worth to check them both and see what they yield, and only perform the cross-validation if there is a large discrepancy. The three methods are asymptotically equivalent.

3 Asymptotic behavior of series estimators

Theorem: Suppose that

1. X has compact support $\mathcal{S}$.
2. $E(|Y - g(X)|^4|X)$ is bounded, and $\sigma^2(x)$ is bounded away from zero. Define $\Sigma = E(\boldsymbol{u}\boldsymbol{u}^T|X_1, \ldots, X_n)$, where $\boldsymbol{u} = (u_1, \ldots, u_n)^T$.
3. For every K, P^K must have full rank. Moreover, the smallest eigenvalue of $E[P^K(X_i)P^K(X_i)^T]$ is bounded away from zero uniformly in K. Finally, there exists a sequence of constants $\zeta_0(K)$ such that $\sup_{x\in\mathcal{S}} ||P^K(x)|| \leqslant \zeta_0(K)$, and $\zeta_0(K)^2 K/n \to 0$ as $n \to \infty$.
4. There exists $\alpha > 0$ such that $\sup_{x\in\mathcal{S}} |g(x) - p^K(x)^T\beta| < CK^{-\alpha}$.
5. The sample (Y_i, X_i), $i = 1, \ldots, n$ is i.i.d.
6. As $n \to \infty$, $K \to \infty$, $K/n \to 0$, and $n/K^{1+2\alpha} \to 0$.

Then,

$$\sqrt{n}\hat{V}_K(x)^{-1/2}(\hat{g}(x) - g(x)) \xrightarrow{d} N(0,1)$$

where

$$\hat{V}_K(x) = \frac{1}{n} P^K(x)^T (P^T P)^{-1} (P^T \hat{\Sigma} P)(P^T P)^{-1} P^K(x).$$

and $\hat{\Sigma} = \begin{pmatrix} \hat{u}_1^2 & & & \\ & \hat{u}_2^2 & & 0 \\ 0 & & \ddots & \\ & & & \hat{u}_n^2 \end{pmatrix}$.

***

We begin to analyze this theorem by the assumptions, as usual.

1. Assumptions 1 can be relaxed, but relaxing it is immaterial. From a practical point of view, it is important to know the support of X if the basis requires this knowledge, such as with splines and wavelets.

2. Assumption 2 is a regularity condition, and we don't need to be concerned with it. Observe that the results allow for deviations from homoskedasticity.

3. Assumption 3 is entirely determined by the basis. The condition that P has full rank is there to guarantee that $(P^T P)$ be invertible. The constants $\zeta(K)$ are different depending on the basis chosen. More on that later.

4. Assumption 4 depends on the space to which g belongs and on the basis itself. It is a direct expression of the ability of the basis to approximate g at a polynomial rate. More on that later.

5. Assumption 5 is the usual. There are results in series regression theory for dependent data, but we are not going to cover them.

6. Assumption 6 is the equivalent for series of the assumptions over h in kernel regression. It states that the number of elements in the basis must increase, but not as fast as n. The condition $n/K^{1+2\alpha} \to 0$ is an undersmoothing condition. In reality, it is possible to get asymptotic normality when $K \to \infty$ a lot slower than required by this condition. However, when K increases faster it is possible to kill the bias asymptotically, and this is why we can write the theorem with no bias term as we did. In fact, uniform consistency holds, and the uniform speed of convergence of the entire function g to a gaussian process is $\zeta_0(K)\sqrt{K}/(\sqrt{n} + K^{-\alpha})$.

The reason why we avoid presenting the bias term here is that I expect you to use series regressions for the estimation of plugins inside of some form of average, where the increase in variance due to undersmoothing is not important. At the same time, in such cases it is always necessary to guarantee the consistency of $g(x)$, since the bias cannot be averaged out.

Now to the understanding of the theorem. The first thing to notice is that the entire result does not differ in practice from an OLS regression. That implies that from a computational point of view, the calculation of the standard errors can be made by the following procedure:

1. Regress Y onto P, using a "robust" estimator for the standard errors. Make sure to calculate the entire covariance matrix of $\hat{\beta}$. This will yield both $\hat{\beta}$ and the term
$$\hat{V}_{\hat{\beta}} = \frac{1}{n}(P^T P)^{-1}(P^T \hat{\Sigma} P)(P^T P)^{-1}.$$
2. The t-statistic is hence
$$\frac{P(x)^T \hat{\beta} - g(x)}{\sqrt{P(x)^T \hat{V}_{\hat{\beta}} P(x)}} \sim t_{n-1}$$

Observe that the variance term is in fact a summation of K terms, and $\hat{V}_{\hat{\beta}}$ is of the order of n^{-1}, as is usually the case with the OLS variance matrix. Hence, though the theorem looks like a simple linear combination of the coefficients of an OLS regression, with convergence speed $\sqrt{n}$, the fact that $K \to \infty$ makes the real convergence speed be $\sqrt{n/K}$, because the terms in the variance increase at the rate K when n increases.

If the basis used is orthonormal, then it is possible to substitute $\hat{V}_{\hat{\beta}}$ by

$$\hat{V}_{\hat{\beta}} = \frac{1}{n}(P^T \hat{\Sigma} P).$$

If the errors are homoskedastic, then $\Sigma = \sigma^2 I_n$, and therefore we can substitute $\hat{V}_{\hat{\beta}}$ by

$$\hat{V}_{\hat{\beta}} = \frac{\hat{\sigma}^2}{n}(P^T P)^{-1},$$

and if the basis is orthonormal,

$$\hat{V}_{\hat{\beta}} = \frac{\hat{\sigma}^2}{n} I_K,$$

where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2$.

As stated above, assumptions 3 and 4 depend on the basis chosen and on the space to which g belongs. Here are some thoughts for the bases that we studied:

1. For Power series, if X has a density which is always positive in $\mathcal{S}$, g is continuously differentiable of order m in $\mathcal{S}$, and $K^3/n \to 0$ as $n \to \infty$, then Assumptions 3 and 4 hold with $\zeta_0(K) = K$ and $\alpha = m/q$.[1] The conditions and results are exactly the same for Fourier series.

2. For B-splines, if X has a density which is always positive in $\mathcal{S}$, g is continuously differentiable of order m in $\mathcal{S}$, and $K^2/n \to 0$ as $n \to \infty$, then Assumptions 3 and 4 hold with $\zeta_0(K) = \sqrt{K}$ and $\alpha = m/q$.[2]

3. For Wavelets, if X has a density which is always positive in $\mathcal{S}$, g is continuously differentiable of order m in $x \in \mathcal{S}$, and $K^2/n \to 0$ as $n \to \infty$, then Assumptions 3 and 4 hold with $\zeta_0(K) = \sqrt{K}$ and $\alpha = m/q$.[3] Observe that the result here is pointwise for differentiable points.

[1] The variable q here is the number of covariates.

[2] The variable q here is the number of covariates, and it assumes that we used tensor products for multivariate splines.

[3] The variable q here is the number of covariates, and it assumes that we used tensor products for multivariate wavelets.