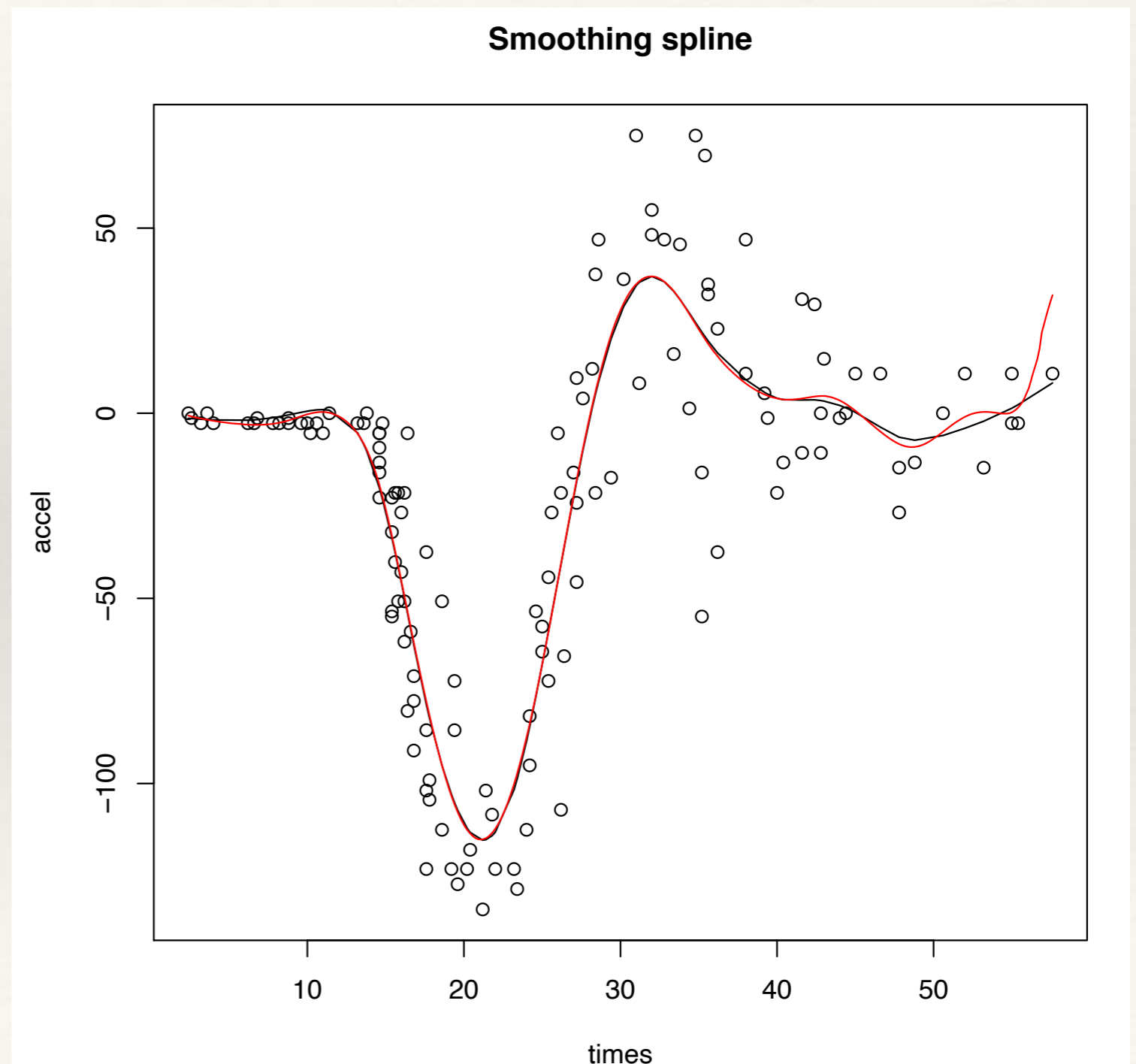


Making sense of data: introduction to statistics for gravitational wave astronomy

Lecture 11: Nonparametric regression

AEI IMPRS Lecture Course

Jonathan Gair jgair@aei.mpg.de



Nonparametric regression

- ❖ The idea of nonparametric regression is to infer the mean value of an observable, Y , as a function of some dependent variable, X , given pairs of observations (x_i, y_i) for $i=1, \dots, n$.

- ❖ In parametric regression we assume a form for the mean that depends on a (small) finite number of parameters and analysis is based on inference of those parameters.

- ❖ In nonparametric regression we instead aim to constrain a function $f(x)$ such that

$$f(x) = \mathbb{E}(Y_i | X_i = x)$$

- ❖ We typically assume data of the form

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

- ❖ with $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ and the support of the function assumed to be $[0,1]$.

- ❖ The set of points $\{x_i\}$ is called a **design** and may be **random** or **fixed**.

Nonparametric regression

- ❖ There are three main approaches to nonparametric regression
 - **kernel estimators**
 - **smoothing splines**
 - **wavelet estimators.**
- ❖ We will give an overview of all three approaches. Further details, and proofs of some of the results that will be quoted, may be found in the lecture notes on the course webpage.

Kernel Estimators

- ❖ **Definition:** a kernel is a function $K(x)$ satisfying

$$\int_{-\infty}^{\infty} K(x) d(x) = 1$$

- ❖ **Definition:** a symmetric kernel is one for which $K(x)=K(-x)$.
- ❖ **Definition:** the order of a kernel is m if $\int_{-\infty}^{\infty} x^l K(x) dx = 0$ for all $l = 1, \dots, m-1$ and $\int_{-\infty}^{\infty} x^m K(x) dx \neq 0$.
- ❖ If $K(x)$ is a kernel, then so is $K_h(x) = K(x/h)/h$. h is called the **bandwidth**.
- ❖ **Examples**
 - ❖ **Uniform (box, rectangular) kernel** $K(x) = I(|x| < 1)/2$.
 - ❖ **Triangular kernel** $K(x) = (1 - |x|) I(|x| < 1)$.
 - ❖ **Gaussian kernel** $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Nadaraya-Watson Estimator

- ❖ Given a kernel $K(x)$ and bandwidth h , the **Nadaraya-Watson estimator** is

$$\hat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{j=1}^n K_h(X_j - x)}, \text{ when } \sum_{i=1}^n K_h(X_i - x) \neq 0,$$

- ❖ and the estimator is zero otherwise.
- ❖ This estimator can be tuned by choosing the kernel and bandwidth to give the smallest **asymptotic mean integrated squared error (MISE)**

$$\text{MISE}(\hat{f}_n) = \mathbb{E}[||\hat{f}_n - f||^2] = \mathbb{E} \left[\int_0^1 |\hat{f}_n(x) - f(x)|^2 dx \right] = \int_0^1 v(x) dx + \int_0^1 [b(x)]^2 dx$$

- ❖ Asymptotically, the variance and bias of the Nadaraya-Watson estimator can be approximated by

$$v(x) \approx \frac{\sigma^2}{nh} ||K||_2^2 \qquad b(x) \approx \frac{\mu_2(K)h^2}{2} f''(x)$$

Nadaraya-Watson Estimator

- ❖ Giving these asymptotic results, we can choose the bandwidth, for a given kernel, that minimises the ASIME. The optimal choice of bandwidth is

$$h_{\text{opt}} = \left(\frac{C_2}{4nC_1} \right)^{\frac{1}{5}} = \left(\frac{\sigma^2 \|K\|_2^2}{n \|f''(x)\|_2^2 \mu_2(K)^2} \right)^{\frac{1}{5}}$$

- ❖ We can now minimise the resulting AMISE over the choice of the kernel. The optimal kernel is the **Epanechnikov kernel**

$$K^{\text{opt}}(x) = \frac{3}{4} \frac{1}{\sqrt{5}} \left(1 - \frac{x^2}{5} \right) \mathbf{1}(|x| \leq \sqrt{5})$$

Nadaraya-Watson Estimator

- ❖ The previous results were valid asymptotically. It is also possible to obtain non-asymptotic results by making some assumption about the smoothness of the function being constrained. One common assumption is that it belongs to the **Hölder class**

Definition 10.11. *The Hölder Class $\mathbf{H}^\beta(M)$ of functions on $[0, 1]$ with $\beta > 0$, $M > 0$ is defined as the set of functions f that satisfy the following conditions with $k = \lfloor \beta \rfloor$:*

1. $|f^{(k)}(x)| \leq M$ for all $x \in [0, 1]$,
2. $|f^{(k)}(x) - f^{(k)}(y)| \leq M|x - y|^{\beta - k}$, $\forall x, y \in [0, 1]$,
where $f^{(k)}$ is the k th derivative of f .

If $\beta \in (0, 1)$, $k = 0$ and $f^{(0)}(x) = f(x)$.

Nadaraya-Watson Estimator

- ❖ Under certain assumptions and defining

$$0 \leq K(u) \leq K_{\max}$$

$$\frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \geq \lambda_0$$

- ❖ The bias and variance of the NW estimator can be bounded by

$$|b(x)| \leq Mh^\beta$$

$$v(x) \leq \frac{\sigma^2 K_{\max}}{nh\lambda_0}$$

- ❖ These can be used to bound the MISE, providing a bound on the **convergence rate** of the NW estimator

$$\mathbb{E} \left[\left(\hat{f}_n^{NW}(x) - f(x) \right)^2 \right] \leq Cn^{\frac{-2\beta}{2\beta+1}}$$

- ❖ In fact, it can be shown that this is the best possible convergence rate that any estimator can achieve for the Hölder class.

Local polynomial estimators

- ❖ The Nadaraya-Watson estimator can be thought of as a locally constant estimator. This can be generalised to the notion of a **local polynomial estimator**.

Definition 10.15. A local polynomial estimator of $f(x)$ of order k , denoted $LP(k)$ estimator, is defined by

$$\hat{f}_n^{LP}(x) = \hat{\theta}_0(x)$$

where for each x $\hat{\theta}(x) = \left(\hat{\theta}_0(x), \hat{\theta}_1(x), \dots, \hat{\theta}_k(x)\right)^T$ is the solution of

$$\hat{\theta}(x) = \arg \min_{\theta_x \in \mathbb{R}^{k+1}} \left\{ \sum_{i=1}^n (Y_i - U_{x,i}^T \theta_x)^2 K \left(\frac{X_i - x}{h} \right) \right\}.$$

For each $m = 1, \dots, k$, $\hat{\theta}_m(x)/h^m$ is an estimator of $f^{(m)}(x)$.

- ❖ where

$$U_{x,i} = \left(1, \frac{x_i - x}{h}, \frac{1}{2!} \left(\frac{x_i - x}{h} \right)^2, \dots, \frac{1}{k!} \left(\frac{x_i - x}{h} \right)^k \right)^T$$

Local polynomial estimators

- ❖ The local polynomial estimator can be evaluated explicitly by noting

$$\hat{\theta}_x = \arg \min_{\theta_x} \{ \theta_x^T \cdot B(x) \cdot \theta_x - 2\theta_x^T \cdot a(x) \}$$

- ❖ where

$$\begin{aligned} B(x) &= \frac{1}{nh} \sum_{i=1}^n U_{x,i} U_{x,i}^T K \left(\frac{X_i - x}{h} \right) \\ &= \frac{1}{n} \sum_{i=1}^n U_{x,i} U_{x,i}^T K_h(X_i - x) \end{aligned}$$

$$\begin{aligned} a(x) &= \frac{1}{nh} \sum_{i=1}^n Y_i U_{x,i} K \left(\frac{X_i - x}{h} \right) \\ &= \frac{1}{n} \sum_{i=1}^n Y_i U_{x,i} K_h(X_i - x) \end{aligned}$$

- ❖ The solution is

$$\hat{\theta}_x = B^{-1}(x) a(x)$$

- ❖ which makes it obvious that the local polynomial estimator is also linear.

Smoothing splines

- ❖ A second method of nonparametric curve fitting is to use **smoothing splines**. These are defined as penalised least squares estimators

$$\hat{f}_n^{\text{pen}}(x) = \arg \min_{f \in \mathcal{C}^2} \left[\sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \text{pen}(f) \right]$$

- ❖ Typically the penalty function is quadratic

$$\text{pen}(f) = \int [f''(x)]^2 dx = \|f''\|_2^2.$$

- ❖ The parameter $\lambda > 0$ is called the **regularisation parameter**. Large values of lambda give more weight to smoothness, while small values give more weight to the observed data.
- ❖ **Theorem 10.3:** the solution to this minimisation problem with a quadratic penalty function is a **natural cubic spline** with knots at the data points.

Cubic splines

Definition 10.17. Let $a \leq t_1 < \dots < t_N \leq b$ be a set of ordered points - called knots. A cubic spline is a continuous function g such that

- $g(x)$ is cubic on $[t_j, t_{j+1}]$, for each $j = 1, \dots, N - 1$:

$$g(x) = b_{j0} + b_{j1}x + b_{j2}x^2 + b_{j3}x^3, \quad x \in [t_j, t_{j+1}],$$

- both g' and g'' are continuous at t_i , $i = 1, \dots, N$.

A spline that is linear beyond the boundary knots is called a natural spline.

- $g(x)$ is linear on $[a, t_1]$ and $[t_N, b]$

$$g(x) = b_{00} + b_{01}x, \quad x \in [a, t_1]$$

$$g(x) = b_{N0} + b_{N1}x, \quad x \in [t_N, b]$$

Fitting cubic splines

Theorem 10.4. *Let knots $a \leq t_1 < \dots < t_N \leq b$. For $j = 3, \dots, N$, define*

$$h_1(x) = 1, h_2(x) = x,$$

$$h_j(x) = (x - t_{j-2})_+^3 - \frac{(t_N - t_{j-2})}{(t_N - t_{N-1})} (x - t_{N-1})_+^3 + \frac{(t_{N-1} - t_{j-2})}{(t_N - t_{N-1})} (x - t_N)_+^3, \quad \forall 3 \leq j \leq N,$$

where $(x - y)_+^3 = \max \{ (x - y)^3, 0 \}$

The set of functions $(h_j)_{j=1}^N$ forms a basis for the set of natural cubic splines at these knots.

- ❖ Smoothing splines fits to data can be found by writing the target function as a linear combination of these basis functions.

$$g(x) = \sum_{j=1}^N \beta_j h_j(x)$$

Fitting cubic splines

- ❖ Substituting this expansion into the penalised least squares expression we find the solution

$$\hat{\beta} = \left[(H^T H + \lambda \Omega)^{-1} H^T Y \right]$$

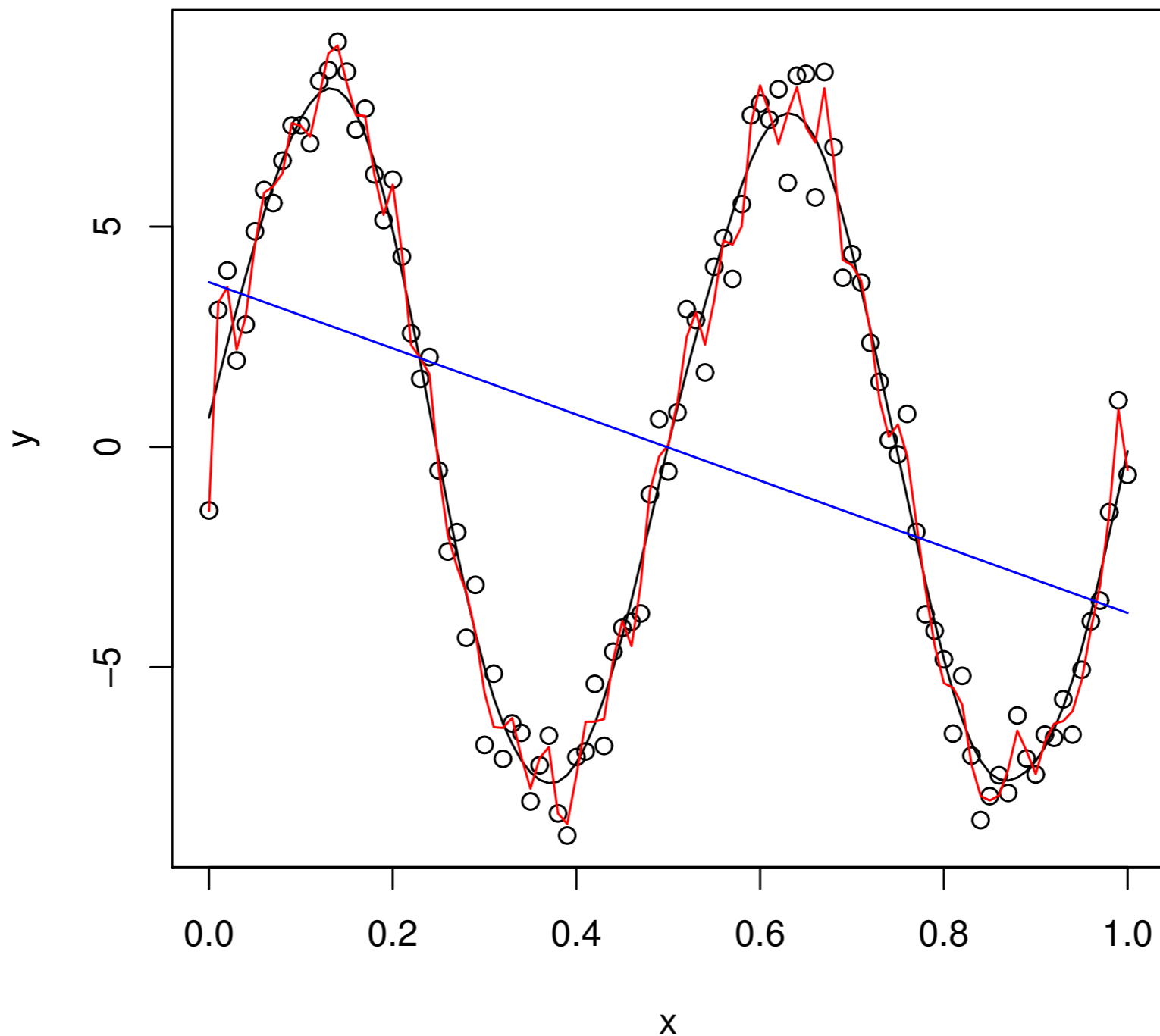
- ❖ where

$$H_{ij} = h_j(x_i), \quad \Omega_{jl} = \int_a^b h_j''(x) h_l''(x) dx, \quad i \in 1, \dots, n, \quad j, l \in 1, \dots, N$$

- ❖ and $Y^T = (Y_1, Y_2, \dots, Y_n)$ is the observed data.
- ❖ This expression makes it clear that the smoothing spline is also a linear estimator.
- ❖ In the limit $\lambda \rightarrow 0$, the smoothing spline becomes a natural cubic spline that passes through all the data points.
- ❖ In the limit $\lambda \rightarrow \infty$, the smoothing spline is a straight line, which is the best fit (in a least squares sense) straight line through the observed data.

Fitting cubic splines

Smoothing splines



Fitting cubic splines

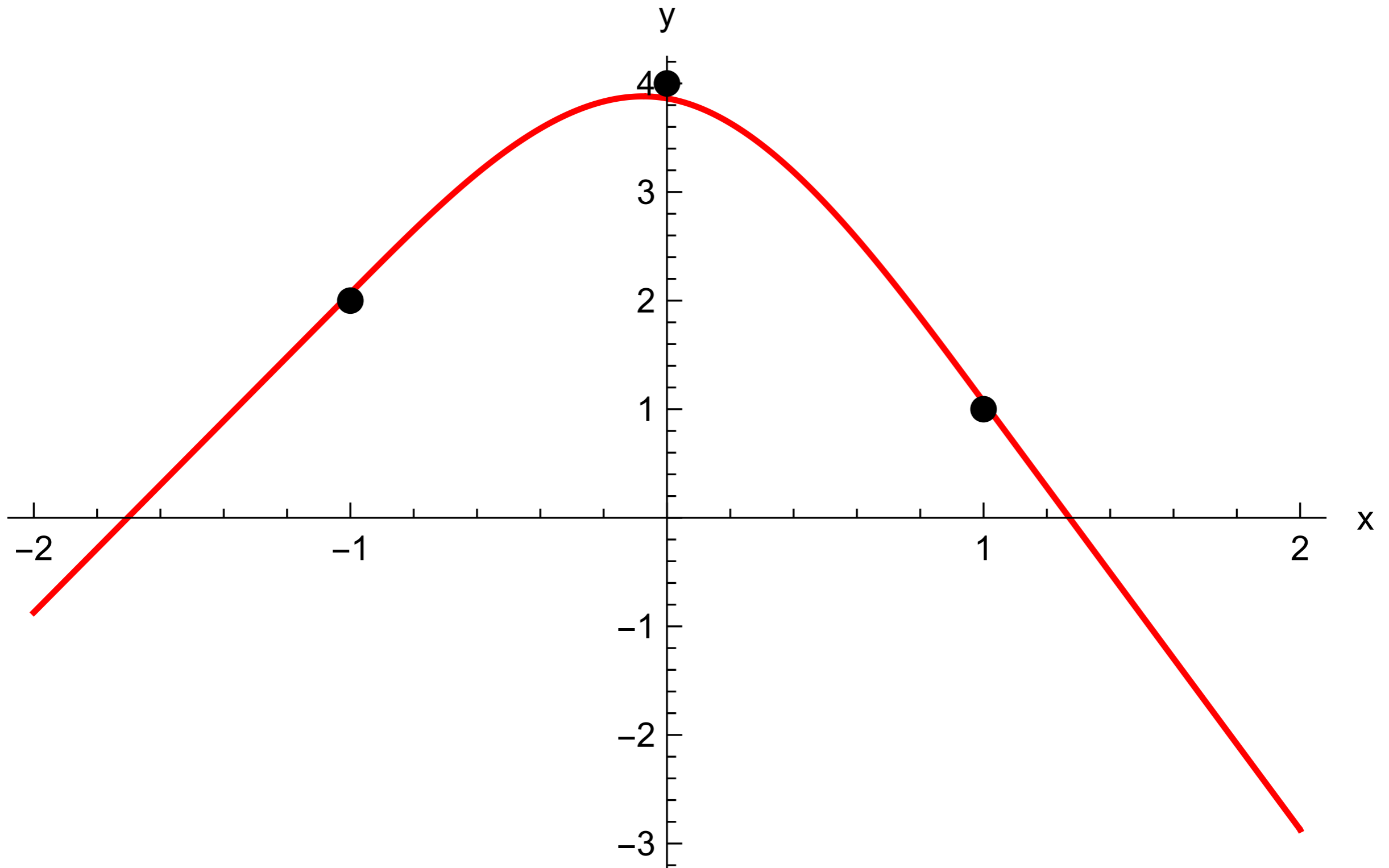
- ❖ The “best” choice of the regularisation parameter can be guided by the observed data using the process of **leave-one-out cross-validation**, i.e., using fits of the data to all but one point to estimate the MISE of the fit.

$$\hat{\lambda} = \arg \min_{\lambda > 0} \left\{ \sum_{i=1}^n \left(Y_i - \hat{f}_{\lambda, -i}(x_i) \right)^2 \right\}$$

- ❖ Smoothing splines are related to kernel estimators. In the limit of large N , the smoothing spline estimator coincides with the Nadaraya-Watson estimator with bandwidth $h = \lambda^{\frac{1}{4}}$ and using the **Silverman kernel**.

$$K(z) = \frac{1}{2} e^{-|z|/\sqrt{2}} \sin(|z|/\sqrt{2} + \pi/4)$$

Smoothing spline: example



Additive models

- ❖ In the preceding discussion we have focussed on fitting models to univariate data, but more commonly the observed data will depend on multiple covariates. There are extensions of kernel estimators and smoothing splines to higher dimensions, but they do not scale well. An alternative is to use an **additive model** of the form

$$Y_i = \alpha + \sum_{j=1}^m f_j(x_j) + \epsilon_i, \quad i = 1, \dots, n$$

- ❖ where, to make the model **identifiable**, we impose the constraints

$$\hat{\alpha} = \bar{Y} = \sum_{i=1}^n Y_i / n \quad \sum_{i=1}^n \hat{f}_j(x_{ji}) = 0$$

- ❖ A **generalised additive model** takes a similar form but now

$$\eta(\mathbf{x}) = g(\mathbb{E}(Y)) = \alpha + \sum_{j=1}^m f_j(x_j)$$

Fitting additive models

- ❖ Additive models can be fitted using the **backfitting algorithm**:

Definition 10.20. *The backfitting algorithm obtains estimates of $\hat{f}_j(x_j)$ in the additive model as follows. Fix the estimator $\hat{\alpha} = \bar{Y}$ and choose initial guesses for $\hat{f}_1, \dots, \hat{f}_m$. Then*

1. For $j = 1, \dots, m$:

(a) Compute $\tilde{Y}_i = Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ki}), i = 1, \dots, n$.

(b) Apply a one-dimensional nonparametric fitting procedure (smoother) to \tilde{Y}_i as a function of x_j . Set \hat{f}_j equal to the output of this procedure.

(c) Renormalise by setting $\hat{f}_j(x)$ equal to $\hat{f}_j(x) - \sum_{i=1}^n \hat{f}_j(x_{ji})/n$.

2. Repeat step 1 until the estimators converge.

Wavelet estimators

❖ Kernel estimators and smoothing splines are nonparametric techniques that rely on smoothing. An alternative approach is to use **orthogonal projection estimators**. The idea is to represent an arbitrary curve as a linear combination of basis functions.

❖ A **wavelet basis** is defined by two functions

• $\phi(x)$ the **father wavelet** or **scaling function** satisfying $\int_0^1 \phi(x) dx = 1$

• $\psi(x)$ the **mother wavelet** or **wavelet function** satisfying $\int_0^1 \psi(x) dx = 0$

❖ Defining translations and dilations of the wavelets through

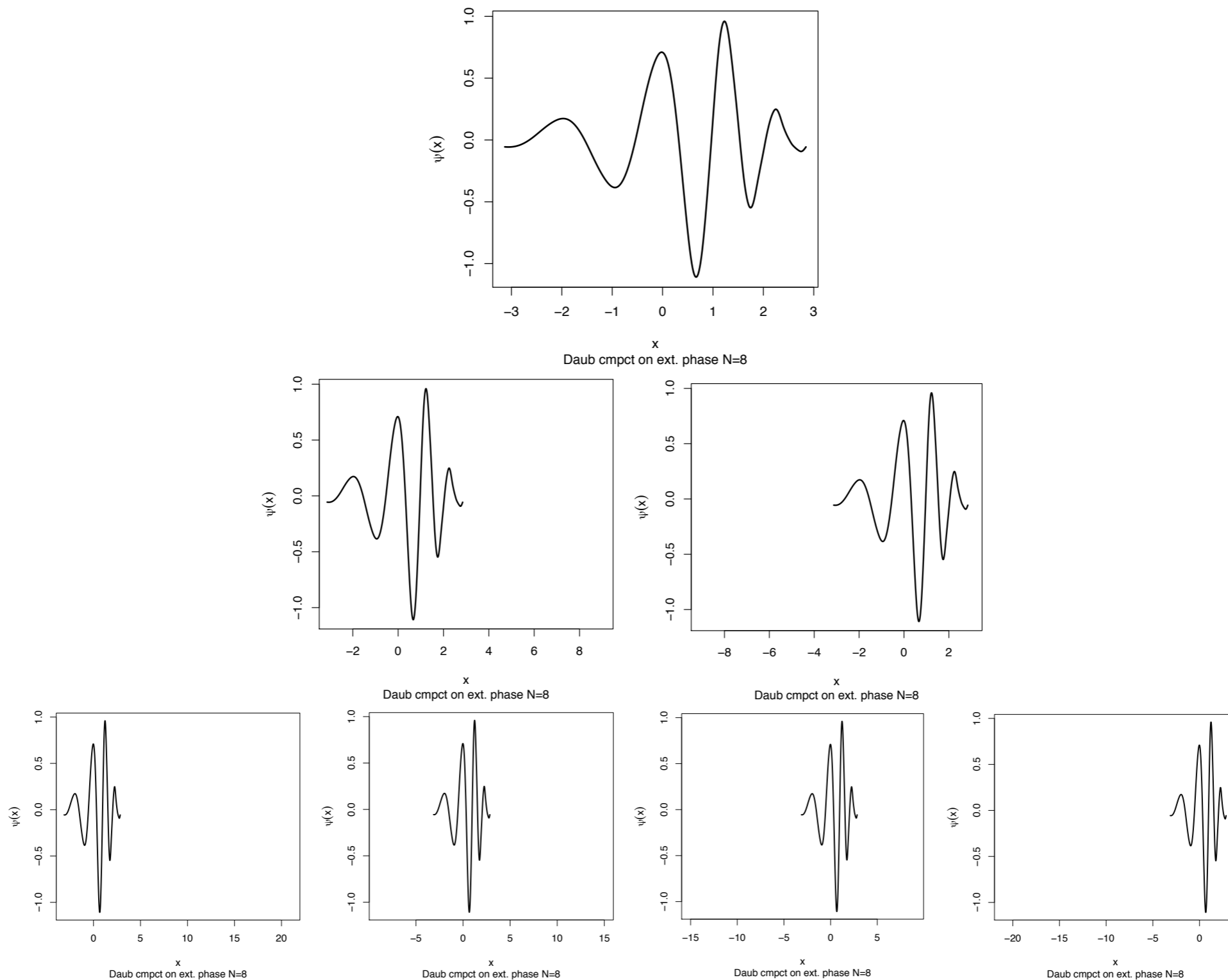
$$\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$$

❖ If the father and mother wavelet are defined appropriately then the set

$$\{\phi, \psi_{jk}, j = 0, 1, \dots, k = 0, \dots, 2^j - 1\}$$

❖ is an orthonormal basis for the space of square integrable functions.

Wavelet basis example: Daubechies, $s = 8$



Wavelet estimators

- ❖ To define a valid wavelet basis the scaling function must obey the **scaling equation**

$$\phi(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} h_k \phi(2x - k)$$

- ❖ for some coefficients $\{h_k\}$ and the wavelet function must obey the **wavelet equation**

$$\psi(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} g_k \phi(2x - k)$$

- ❖ for some coefficients $\{g_k\}$. The coefficients must obey the constraints

$$\sum_k h_k = \sqrt{2}, \quad \sum_k h_k h_{k-2l} = \delta_{0l}$$

$$\sum_k g_k h_{k+2m} = 0 \quad \forall m \in \mathbb{Z}, \quad \sum_k g_k g_{k-2l} = \delta_{0l}$$

- ❖ The latter two equations are automatically satisfied by the choice $g_k = (-1)^k h_{1-k}$

Wavelet estimators

- ❖ We can use the wavelet basis to write any function as an expansion

$$f(x) = \theta_0 \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x)$$

- ❖ To estimate the coefficients in this expansion from observed data we can define a **wavelet projection estimator** given $n = 2^{J_0}$ observations by computing

$$\hat{\theta}_0 = \frac{1}{n} \sum_{i=1}^n Y_i \phi(x_i), \quad \hat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^n Y_i \psi_{jk}(x_i), \quad j < J_0$$

- ❖ and then constructing

$$\hat{f}_{J_0}(x) = \hat{\theta}_0 \phi(x) + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk} \psi_{jk}(x)$$

Thresholding estimators

- ❖ In general, we do not need all the possible coefficients to represent the function, and so we use **thresholding estimators** by dropping coefficients that are close to zero.

- ❖ A **soft thresholding estimator** uses

$$\hat{\theta}_{jk} = \text{sign}(y_{jk})(|y_{jk}| - \lambda)_+$$

- ❖ A **hard thresholding estimator** uses

$$\hat{\theta}_{jk} = y_{jk} \mathbf{I}(|y_{jk}| > \lambda)$$

- ❖ A common choice for the threshold is the **universal threshold**

$$\lambda = \sigma \sqrt{2 \log n}$$

- ❖ This relies on having an estimate of the variance in the noise, which can be found from the **median absolute deviation**.

$$\hat{\sigma} = 1.4826 \text{ MAD}(d_{J-1,0}, \dots, d_{J-1,2^J-1}) \quad \text{MAD}(x_1, \dots, x_n) = \text{median}(|x_i - \text{median}(x_i)|)$$

Cascade algorithm

- ❖ Wavelet coefficients can be estimated quickly using the Cascade algorithm

Cascade algorithm

1. Set $c_{Jk} = Y_{k+1}$ for $k = 0, 1, \dots, 2^J - 1$, set $j = J - 1$;
2. Set

$$c_{jk} = \sum_{m \in \mathbb{Z}} h_m c_{j+1, 2k+m}, \quad d_{jk} = \sum_{m \in \mathbb{Z}} g_m c_{j+1, 2k+m};$$

3. if $j = 0$ stop; else set $j := j - 1$ and repeat step 2.

- ❖ It is particularly efficient when using **Haar wavelets**

$$\phi(x) = \mathbf{1}_{(0,1]}(x) \quad \psi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x)$$

- ❖ for which the only non-zero $\{h_k, g_k\}$ are $h_0 = h_1 = 1/\sqrt{2}$ and $g_0 = 1/\sqrt{2}$, $g_1 = -1/\sqrt{2}$

Inference: point estimates

- ❖ The nonparametric estimators can be used to construct confidence intervals for the unknown function at a specific point. These can be **asymptotic** or **conservative**.
- ❖ If we know $|b(x)| \leq b_0(x)$ & $v(x) \leq v_0(x)$ then a $(1 - \alpha)100\%$ **conservative confidence interval** based on a linear estimator takes the form

$$f(x) \in \hat{f}(x) \pm \left(b_0(x) + z_{\frac{\alpha}{2}} \sqrt{v_0(x)} \right)$$

- ❖ If the asymptotic bias is small relative to the variance $b^2(x) \ll v(x)$ then a $(1 - \alpha)100\%$ **asymptotic confidence interval** takes the form

$$f(x) \in \hat{f}(x) \pm z_{\frac{\alpha}{2}} \sqrt{v(x)}$$

Inference: confidence bands

- ❖ A **confidence band** is a statement about the global properties of a function rather than its value at a specified point. Assuming that the bias is much smaller than the standard deviation we have the following confidence band for linear estimators

$$\left\{ f : |f(x) - \hat{f}(x)| \leq c_\alpha \sqrt{v(x)}, \forall x \in [a, b] \right\}$$

- ❖ where

$$c_\alpha \approx \sqrt{2 \log \left(\frac{a_0}{\alpha h} \right)}, \text{ where } a_0 = \frac{|b - a|}{\pi} \frac{\|K'\|_2}{\|K\|_2}$$

- ❖ This is a special case of a more general result for linear estimators $\hat{g}(x) = \sum l_i(x) Y_i$

$$|\hat{g}(x) - g(x)| \leq c_\alpha \sqrt{\text{var}(\hat{g})} \quad \forall x \in [a, b]$$

$$\text{where } c_\alpha = \sqrt{2 \ln \left(\frac{\kappa_0}{\alpha \pi} \right)}, \quad \kappa_0 = \int_a^b \|\mathbf{T}'(x)\| dx$$

- ❖ where $T_i(x) = l_i(x)/\|l(x)\|$. The more general result can be used to obtain confidence bands on derivatives of a function when using local polynomial estimators.

Inference: hypothesis testing

- ❖ The above results apply for all linear estimators, and hence all three types of estimator that we discussed in this lecture (except the hard thresholding wavelet estimator which is not linear).
- ❖ A wavelet estimator based on Haar wavelets can also be used to test if a function is **constant on sub-intervals**. Specifically, the hypothesis

$$H_0 : f(x) = \text{constant on } (a, b)$$

- ❖ is equivalent to

$$H_0 : \theta_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k + 1/2}{2^j} < b$$

- ❖ and can be tested using

$$T = \sigma^{-2} \sum_{j,k: a < \frac{k+1/2}{2^j} < b, j < J} d_{jk}^2$$

- ❖ where d_{jk} are the estimators of the parameters obtained from, for example, the cascade algorithm. T follows a chi-squared distribution under H_0 .