Non-parametric regression

Sofia Olhede

EPFL

November 29, 2020

1. Non-parametric regression

Nonparametric relationships with x_i

Whatever happened to likelihood, though? Find $h \in C^2$ that minimises

$$\underbrace{\sum_{i=1}^{n}\{Y_i - h(x_i)\}^2}_{\text{Fit Penalty}} + \underbrace{\lambda \int_I \{h''(t)\}^2 dt}_{\text{Roughness Penalty}}$$

- This is a Gaussian likelihood with a roughness penalty
  - $\hookrightarrow$ If use only likelihood, any interpolating function is an MLE!
- λ to balance fidelity to the data and smoothness of the estimated h.

Remarkably, problem has unique explicit solution!

$\hookrightarrow$ Natural Cubic Spline with knots at $\{x_i\}_{i=1}^n$:

- piecewise polynomials of degree 3,
- with pieces defined at the knots,
- with two continuous derivatives at the knots,
- and linear outside the data boundary.

Nonparametric relationships with x_i

Can represent splines via natural spline basis functions B_j, as

$$s(x) = \sum_{j=1}^{n} \gamma_j B_j(x).$$

Defining matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ as

$$B_{ij} = B_j(x_i), \quad \Omega_{ij} = \int B_i''(x) B_j''(x)\, dx,$$

our penalised likelihood becomes

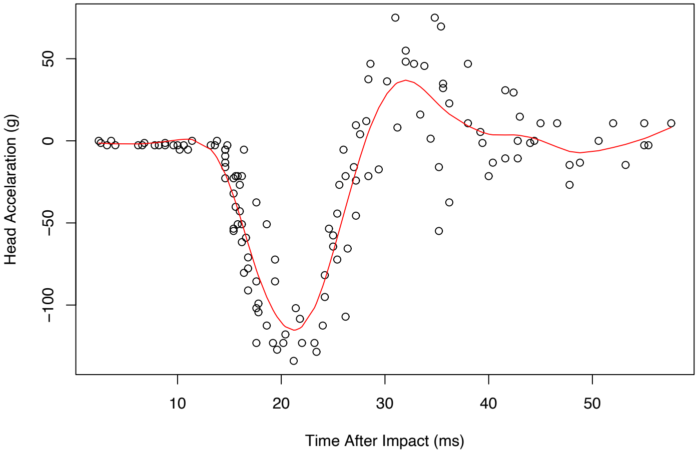
$$\text{min!} \left\{ (\boldsymbol{Y} - \boldsymbol{B}\gamma)^\top (\boldsymbol{Y} - \boldsymbol{B}\gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma \right\}.$$

Differentiating and equating with zero yields

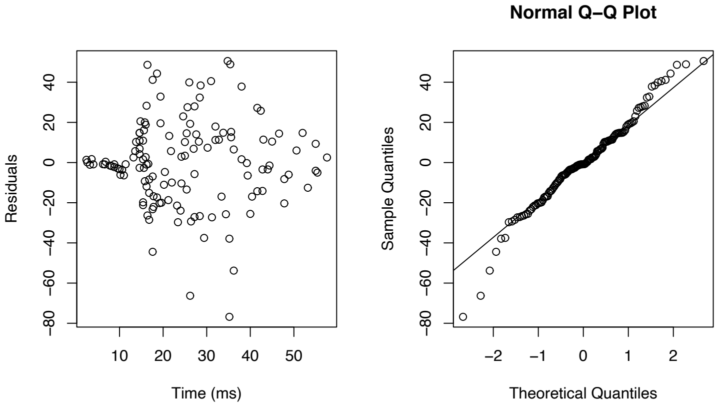
$$(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})\hat{\gamma} = \boldsymbol{B}^\top \boldsymbol{Y} \implies \hat{\gamma} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top \boldsymbol{Y}.$$

- The *smoothing matrix* is $\boldsymbol{S}_\lambda = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top$.
- The cubic spline fit is approximately a kernel smoother (keyword: equivalent kernel).

Nonparametric relationships with x_i



Nonparametric relationships with x_i



Nonparametric relationships with x_i

Equivalent degrees of freedom

hat matrix

- Least squares estimation: $\boldsymbol{Y} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we have $\hat{\boldsymbol{Y}} = \boldsymbol{H}\,\boldsymbol{Y}$, with $\mathrm{trace}(\boldsymbol{H}) = p$, in terms of the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

Linear smoothers: In the case that the smoothed values can be written as a linear transformation of the observed values, the smoothing operation is known as a linear smoother; the matrix representing the transformation is known as a smoother matrix or hat matrix.[citation needed]

The operation of applying such a matrix transformation is called convolution.

- In spline smoothing

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B}^\top}_{S_\lambda}\ \boldsymbol{Y}.$$

— from before (smoothing matrix)

suggesting definition of equivalent degrees of freedom of smoother as

$$\mathrm{edf} = \mathrm{trace}(\boldsymbol{S}_\lambda)$$

- $\mathrm{trace}(\boldsymbol{S}_\lambda)$ is monotone decreasing in λ, with $\mathrm{trace}(\boldsymbol{S}_\lambda) \to 2$ as $\lambda \to \infty$ (will always have two nonzero eigenvalues) and $\mathrm{trace}(\boldsymbol{S}_\lambda) \to n$ as $\lambda \to 0$.
- Note 1–1 map $\lambda \leftrightarrow \mathrm{trace}(\boldsymbol{S}_\lambda) =$ df, so usually determine roughness using edf (interpretation easier).
- Each eigenvalue of $\boldsymbol{S}_\lambda$ lies in $(0, 1)$, so this is a smoothing matrix, not a projection matrix.

Nonparametric relationships with x_i

Bias/Variance Tradeoff and Cross Validation

Focus on the fit for the given grid $x_1, \dots, x_n$:

$$\hat{\mathrm{g}} = (\hat{g}(x_1), \dots, \hat{g}(x_n)), \quad \mathrm{g} = (g(x_1), \dots, g(x_n))$$

Consider the mean squared error:

$$\mathbb{E}(||\mathrm{g} - \hat{\mathrm{g}}||^2) = \underbrace{\mathbb{E}\{||\mathbb{E}(\hat{\mathrm{g}}) - \hat{\mathrm{g}}||^2\}}_{\text{variance}} + \underbrace{||\mathrm{g} - \mathbb{E}(\hat{\mathrm{g}})||^2}_{\text{bias}^2}.$$

In the case of a linear smoother, for which $\hat{\mathrm{g}} = \boldsymbol{S}_\lambda \boldsymbol{Y}$, we easily calculate

$$\mathbb{E}(||\mathrm{g} - \hat{\mathrm{g}}||^2) = \frac{\mathrm{trace}(\boldsymbol{S}_\lambda \boldsymbol{S}_\lambda^\top)}{n}\sigma^2 + \frac{(\mathrm{g} - \boldsymbol{S}_\lambda \mathrm{g})^\top (\mathrm{g} - \boldsymbol{S}_\lambda \mathrm{g})}{n},$$

so

- $\lambda \uparrow \implies$ variance $\downarrow$ but bias $\uparrow$,
- $\lambda \downarrow \implies$ bias $\downarrow$ but variance $\uparrow$.
- Would like to choose λ to find optimal bias-variance tradeoff:
  - $\hookrightarrow$ Unfortunately, optimal λ will depend on unknown g!

Nonparametric relationships with x_i

- Fitted values are $\hat{\boldsymbol{Y}} = \boldsymbol{S}_\lambda \boldsymbol{Y}$.
- Fitted value $\hat{Y}_j^-$ obtained when (Y_j, x_j) is dropped from fit is

$$S_{jj}(\lambda)(Y_j - \hat{Y}_j^-) = \hat{Y}_j - \hat{Y}_j^-.$$

- Cross-validation sum of squares is

$$\mathrm{CV}(\lambda) = \sum_{j=1}^n (Y_j - \hat{Y}_j^-)^2 = \sum_{j=1}^n \left\{ \frac{Y_j - \hat{Y}_j}{1 - S_{jj}(\lambda)} \right\}^2,$$

and generalised cross-validation sum of squares is

$$\mathrm{GCV}(\lambda) = \sum_{j=1}^n \left\{ \frac{Y_j - \hat{Y}_j}{1 - \mathrm{trace}(S_\lambda)/n} \right\}^2,$$

where $S_{jj}(\lambda)$ is (j, j) element of $\boldsymbol{S}_\lambda$.

Nonparametric relationships with x_i

If $\mathcal{F} \ni g(\cdot)$ is a separable Hilbert space,we can write:

$$g(x) = \sum_{k \in \mathbb{Z}} \beta_k \psi_k(x) \quad \text{(in an appropriate sense)},$$

?

with $\{\psi\}_{k=1}^{\infty}$ known (orthogonal) basis functions for $\mathcal{F}$, e.g.,

- $\mathcal{F} = L^2(-\pi, \pi)$,
- $\{\psi_k\} = \{e^{-ikx}\}_{k \in \mathbb{Z}}$, $\psi_i \perp \psi_j$, $i \neq j$.
- Gives Fourier series expansion, $\beta_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) e^{-ikx} dx$.

If we truncate series, then we reduce to linear regression:

$$Y_i = \sum_{|k| < \tau} \beta_k \psi_k(x_i) + \varepsilon_i, \quad \tau < \infty$$

Notice: truncation has implications, e.g., in Fourier case:

- Truncating implies assume $g \in \text{span}\{\psi_{-\tau}, ..., \psi_\tau\} \subset L^2$.
- Interpret this as a smoothness assumption on g.
- How to choose τ optimally?

Nonparametric relationships with x_i

Classical exercise in Fourier analysis shows that

$$\sum_{k=-\tau}^{\tau} \beta_k e^{-ikx} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(y) D_\tau(x-y)\, dy$$

?

with the Dirichlet kernel of order τ, $D_\tau(u) = \sin\{(\tau + 1/2)\, u\}/\sin(u/2)$.

Recall kernel smoother:

$$\hat{g}(x_0) = \sum_{i=1}^{n} \frac{Y_i K_\lambda(x_i - x_0)}{\sum_{i=1}^{n} K_\lambda(x_i - x_0)} = \frac{1}{c} \int_I y(x) K_\lambda(x - x_0)\, dx,$$

with

$$y(x) = \sum_{i=1}^{n} Y_i \delta(x - x_i).$$

- So if K is the Dirichlet kernel, we can do series approximation via kernel smoothing.
- Works for other series expansions with other kernels (e.g., Fourier with convergence factors)

Orthogonal functions

- Suppose again that we observe

$$Y_i = h(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Here $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid.
- Initially we assume $x_i = i/n$ namely a regular design for $i = 1, \ldots, n$.
- Let $\phi_1(x)$, $\phi_2(x), \ldots$ be an orthogonal basis for the interval $[0, 1]$. Often the cosine basis is used

$$\underbrace{\phi_1}_{\cos(0)=1}(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(\{j-1\}\pi x), \; j\cancel{x} = 2,\ 3 \ldots.$$

- Here we expand $h(x)$ as

$$h(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x),$$

where $\theta_j = \int_0^1 h(x)\phi_j(x)\, dx$.

Orthogonal functions II

like regular
bias functions in ML

- We approximate

$$h_n(x) = \sum_{j=1}^{n} \theta_j \phi_j(x),$$

which is a projection of $h(x)$ into the span of $\{\phi_1(x), \phi_2(x), \ldots, \phi_n\}$.

- This introduces an integrated squared bias of

$$B_n(\theta) = \int_0^1 \{r(x) - r_n(x)\}^2 \, dx = \sum_{j=n+1}^{\infty} \theta_j^2.$$

- We can understand this further.

Orthogonal functions III

- This can be quantified.
  Lemma: Let $\Theta(m, c)$ be a Sobolev ellipsoid. Then

?

θ

$$\sup_{\theta \in \Theta(m,c)} B_n(\theta) = O\left(\frac{1}{n^{2m}}\right).$$

- A Sobolev ellipsoid is a set of functions for which $\theta_j^2 \sim (\pi j)^{2m}$; an ellipsoid is defined by

$$\Theta = \left\{ \theta : \sum_j a_j^2 \theta_j^2 \leq c^2 \right\}.$$

- Therefore if $m > 1/2$ we find $B_n = o(1/n)$.
- The bias is negligible and we shall ignore it for the rest of the chapter. We will therefore focus on estimating $h_n(x)$ rather than $h(x)$.

Orthogonal functions IV

- We define

$$Z_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(x_i), \quad j = 1, 2, 3, \ldots .$$

?

- We can then ask what is the distribution of Z_j?
- We note that

$$\begin{aligned} Z_j &= \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \{h(x_i) + \varepsilon_i\}\phi_j(x_i) \\ &= \theta_j + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \phi_j(x_i) = \theta_j + \nu_j. \end{aligned} \tag{1}$$

Using earlier results we can deduce that $\nu_j \sim N(0, \frac{\sigma^2}{n})$.

Orthogonal functions V

- We know from a previous section (Lecture 7) that shrinkage estimators can reduce the mean square error.
- We shall discuss James-Stein estimators a bit further.
- A modulator is a vector $b = \begin{pmatrix} b_1 & \dots & b_n \end{pmatrix}$ such that $0 \leq b_j \leq 1$ for $j = 1, \dots, n$.
- A modulation estimator takes the form

$$\begin{aligned} \widehat{\theta} &= b \odot Z \\ &= \begin{pmatrix} b_1 Z_1 \\ \dots \\ b_n Z_n \end{pmatrix}. \end{aligned} \tag{2}$$

- A constant modulator is a modulation of the form $\begin{pmatrix} b & \dots & b \end{pmatrix}$.
- A nested subset selection modulator is a modulator of the form $\begin{pmatrix} b & \dots & b & 0 & \dots & 0 \end{pmatrix}$.

Orthogonal functions VI

- A monotone modulator is of the form

$$1 \geq b_1 \geq b_2 \geq \cdots \geq b_n \geq 0.$$

- The function estimator provided by a modulator is

i.e. we're reweighting our model

because it removes the higher frequency behavior

$$\hat{h}_n(x) = \sum_{j=1}^{n} \underbrace{\widehat{\theta}_j \phi_j(x)}_{\text{like in CNNs: } \sum_k \sum_l w_{kl} \phi(x_{ij})} = \sum_{j=1}^{n} b_j Z_j \phi_j(x).$$

This is a linear smoother.

- Modulators shrink Z_j towards 0. This smoothes the function estimates.
- We define the risk as

$$R(b) = \mathbb{E}_\theta\{\sum_{j=1}^{n}(b_j Z_j - \theta_j)^2\}$$

Orthogonal functions VII

- To estimate b we need to estimate σ. There are reasons why we would take

$$\widehat{\sigma}^2 = \frac{1}{n - J_n} \sum_{i=n-J_n+1}^{n} Z_i^2.$$

variability in terms of wiggles vs variability in terms of true function

- Often we take $J_n = n/4$.
- Theorem: The risk of a modulator b is

$$R(b) = \sum_{j=1}^{n} \theta_j^2 (1 - b_j)^2 + \frac{\sigma^2}{n} \sum_{j=1}^{n} b_j^2.$$

- The SURE estimator of $R(b)$ are

$$\widehat{R}(b) = \sum_{j=1}^{n} \left(Z_j^2 - \frac{\widehat{\sigma}^2}{n} \right)_+ (1 - b_j)^2 + \frac{\widehat{\sigma}^2}{n} \sum_{j=1}^{n} b_j^2.$$

Orthogonal functions VIII

as before we used $\theta = bZ$ when re-writting $h(x)$ (slide 17)

- The modulation estimator of θ is

$$\theta = \left(\widehat{b}_1 Z_1,\ \widehat{b}_2 Z_2,\ \ldots\right).$$

where b minimises $\widehat{R}(b)$. This yields

$$\hat{h}_n(x) = \sum_{j=1}^{n} \widehat{\theta}_j \phi_j(x) = \sum_{j=1}^{n} b_j Z_j \phi_j(x).$$

For a fixed b we expect that $\widehat{R}(b)$ approximates $R(b)$. We need more, as $\hat{b}$ will depends on the same data as $\hat{R}(b)$. We therefore need $\widehat{R}(b)$ to approximate $R(b)$ uniformly.

- We shall assume that the modulator takes the form

$$\left(1 \quad \ldots 1 \quad 0 \quad \ldots \quad 0\right).$$

Orthogonal functions IX

- This corresponds to picking J to minimize

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

- We note that $\widehat{R}(b)$ is

$$\widehat{R}(b) = \sum_{i=1}^{n} \{b_i - g_i\}^2 Z_i^2 + \frac{\widehat{\sigma}^2}{n} \sum_{i=1}^{n} g_i .$$

- Here

$$g_i = \{Z_i^2 - \frac{\widehat{\sigma}^2}{n}\}/Z_i^2 .$$

We therefore minimize $\sum_{i=1}^{n} \{b_i - g_i\}^2 Z_i^2$.

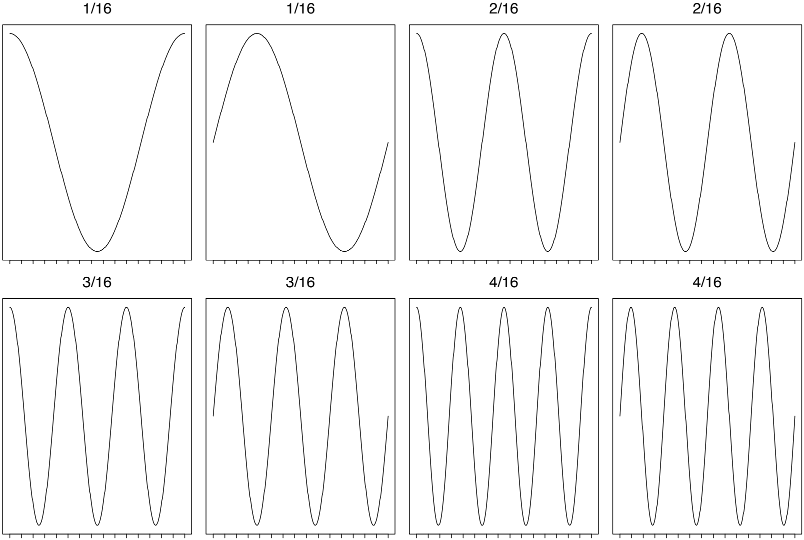
Orthogonal functions X

- This then produces an estimator.
- The first generalization of this problem uses a basis that is orthogonal with respect to the design points $x_1, \ldots, x_n$.
- We define

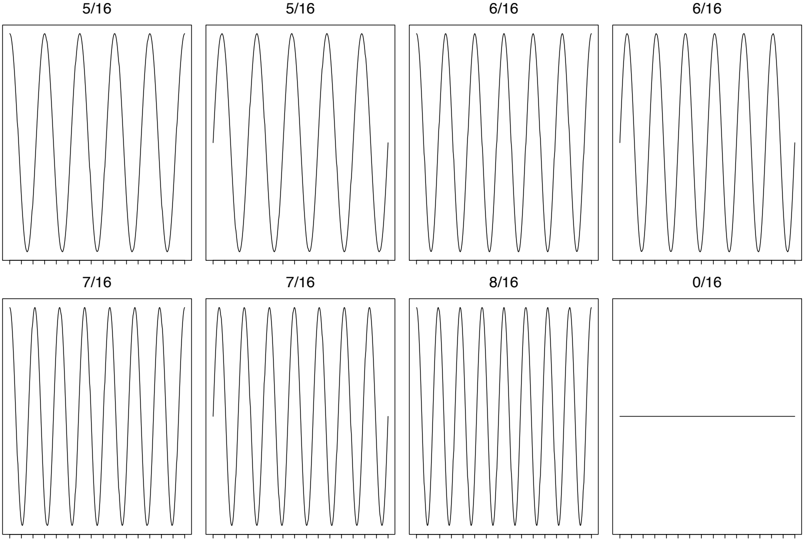
$$Z_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi(x_i).$$

- We can still use the developed methodology.

Cosines & Sines



Cosines & Sines II

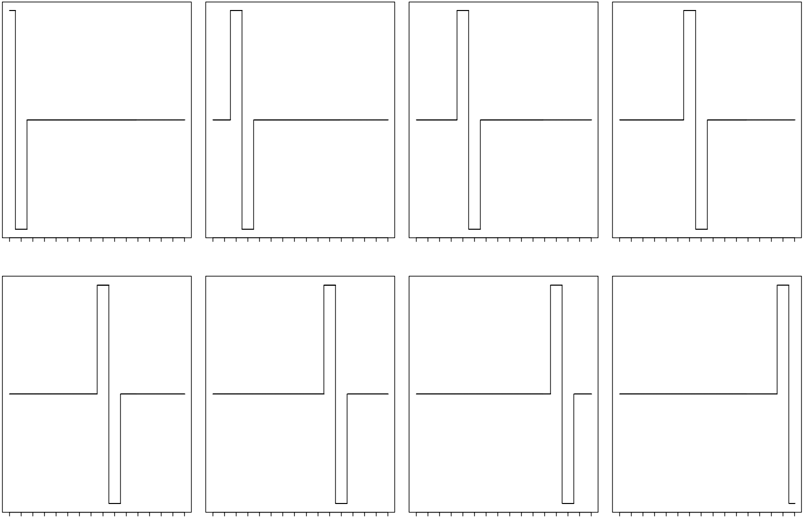


Orthogonal functions X

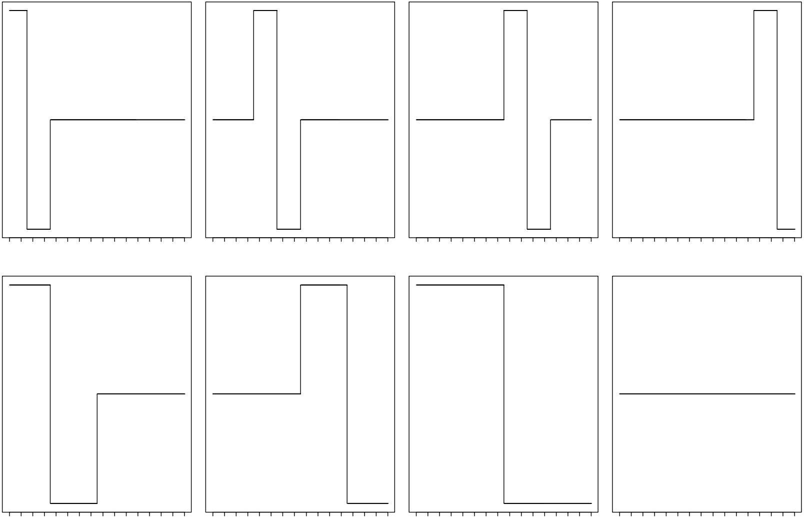
- We could use other functions than those based on trigonometric functions.
- We could start from set $\{\psi_{j,k}\}$ both associated with locality and scale.
- Until the 1980's the only well known orthogonal decompositions available were the Fourier bases, and orthogonal polynomials , which cannot make this time distinction.
- In the 1980's Ingrid Daubechies developed new projections which make this possible. These projections, or filters, are called *wavelets*, and form a substantial part of modern signal analysis.

wavelets are better than sin /cos functions
as they introduce locality.

Haar wavelets



Haar wavelets II



Orthogonal functions X

General set-up to decompose into wavelets or general forms.....

- We model

$$\begin{aligned}\mathcal{W}\underline{Y} &= \mathcal{W}\underline{\mu} + \mathcal{W}\eta \\ \underline{W} &= \mathcal{W}\underline{\mu} + \underline{\epsilon}\end{aligned}$$

?

σ

where

$$\mathbb{V}\mathrm{ar}\{\underline{\epsilon}\} = \mathcal{W}\mathbb{V}\mathrm{ar}\{\eta\}\mathcal{W}^T = \sigma^2\mathcal{W}\mathcal{W}^T = \sigma^2 \mathsf{I}_n.$$

Use our knowledge of $\underline{W}$ to find a good estimate of $\underline{\mu}$ via $\mathcal{W}$.

$$\tilde{\sigma}_{\mathrm{mad}} = \frac{\mathrm{median}\{|W_1|, \ldots, |W_{n/2}|\}}{0.6745}.$$

We shall threshold all but the final 2^j entries by

$$W_j^{(ht)} = \begin{cases} 0 & \mathrm{if} |W_j| \leq \lambda \\ W_j & \mathrm{if} |W_j| > \lambda \end{cases}$$

The only problem remains is how to choose λ well; look at order statistics of Gaussians;

Orthogonal functions X

?

- We would wish as $n \to \infty$

$$P(\max\{|W_i|\} > \lambda) \to 0$$

- So as we collect more observations we can guarantee that there is no noise left. We thus choose

$$\lambda = \sigma\sqrt{2\ln(n)}$$

Nonparametric relationships with x_i

So far: how to estimate $g : \mathbb{R} \to \mathbb{R}$ (assumed smooth) in

$$Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \text{ given data } \quad \{(Y_i, x_i)\}_{i=1}^n.$$

What happens if noise is not Gaussian?

Can extend to GLM setting as:

$$Y_i | x_i \overset{indep}{\sim} \exp\{g(x_i)y - \gamma(g(x_i)) + S(y)\}$$

- Parametrise candidate g via spline

$$s(x) = \sum_{j=1}^n \gamma_j B_j(x).$$

- Define matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ as before,

$$B_{ij} = B_j(x_i), \quad \Omega_{ij} = \int B_i''(x) B_j''(x) \, dx$$

- And consider penalised likelihood, similarly as with penalised GLM

$$\ell_n(\gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma = \gamma^\top \boldsymbol{B}^\top \boldsymbol{Y} - \sum_{i=1}^n \gamma(\boldsymbol{b}_i^\top \gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma.$$

Nonparametric relationships with x_i

How can we generalise to multivariate covariates?

- "Immediate" Generalisation: $g : \mathbb{R}^p \to \mathbb{R}$ (smooth)

$$Y_j = g(x_{j1}, \ldots, x_{jp}) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Estimation by (e.g.) multivariate kernel method.

- Two basic drawbacks of this approach ...

$\hookrightarrow$ Shape of kernel? (definition of *local*)

$\hookrightarrow$ *Curse of dimensionality*