Basis Expansion and Regularization

Dept. Computer Science & Engineering,
Shanghai Jiao Tong University

Outline

- Piece-wise Polynomials and Splines
- Smoothing Splines
- Automatic Selection of the Smoothing Parameters
- Nonparametric Logistic Regression
- Multidimensional Splines
- Regularization and Reproducing Kernel Hilbert Spaces
- Wavelet Smoothing

Piece-wise Polynomials and Splines

- Linear basis expansion $f(x) = \sum_{m=1}^{N} \beta_m h_m(x)$

- Some basis functions that are widely used

$$h_m(x) = x_m$$

$$h_m(x) = x^m$$

$$h_j(x) = \log(x_j);\ \sqrt{X_j}$$

$$h_m(x) = \sin(m\pi x); \cos(m\pi x)$$

Regularization

- Three approaches for controlling the complexity of the model.
  - Restriction

$$f(x)=\sum_{j=1}^{p} f_j(x_j)=\sum_{j=1}^{p}\sum_{k=1}^{M_j}\beta_{jk}h_{jk}(x_j)$$

  - Selection： The variable selection techniques

  - Regularization:

$$f(x)=\sum_{k=1}^{N}\beta_k h_k(x)$$

$$y=\sum_{k=1}^{m}\beta_k h_k(x)+\varepsilon$$

$$\min_\beta \sum_{i=1}^{M}\|\varepsilon(i)\|^2=\sum_{i=1}^{M}\left\|y_i-\sum_{k=1}^{m}\beta_k h_k(x_i)\right\|^2$$

$$\min_\beta \sum_{i=1}^{M}\left\|y_i-\sum_{k=1}^{m}\beta_k h_k(x_i)\right\|^2+\lambda J(\beta)$$

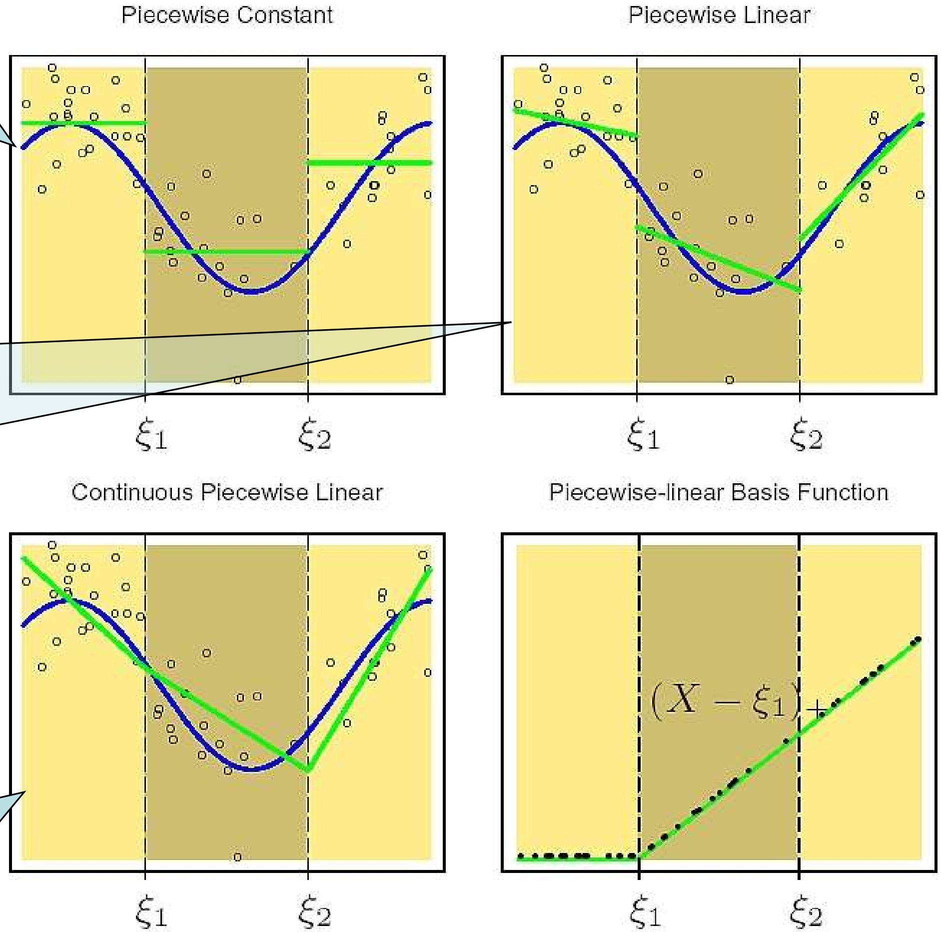
$$J(\beta)=\|\beta\|^2\ ?$$

Piecewise Polynomials and Splines

$$h_1(X) = I(X < \xi_1),$$
$$h_2(X) = I(\xi_1 \le X < \xi_2)$$
$$h_3(X) = I(\xi_2 \le X);$$

$$h_1(X) = I(X < \xi_1),$$
$$h_2(X) = I(\xi_1 \le X < \xi_2)$$
$$h_3(X) = I(\xi_2 \le X),$$
$$h_{m+3}(X) = h_m(X)X;$$

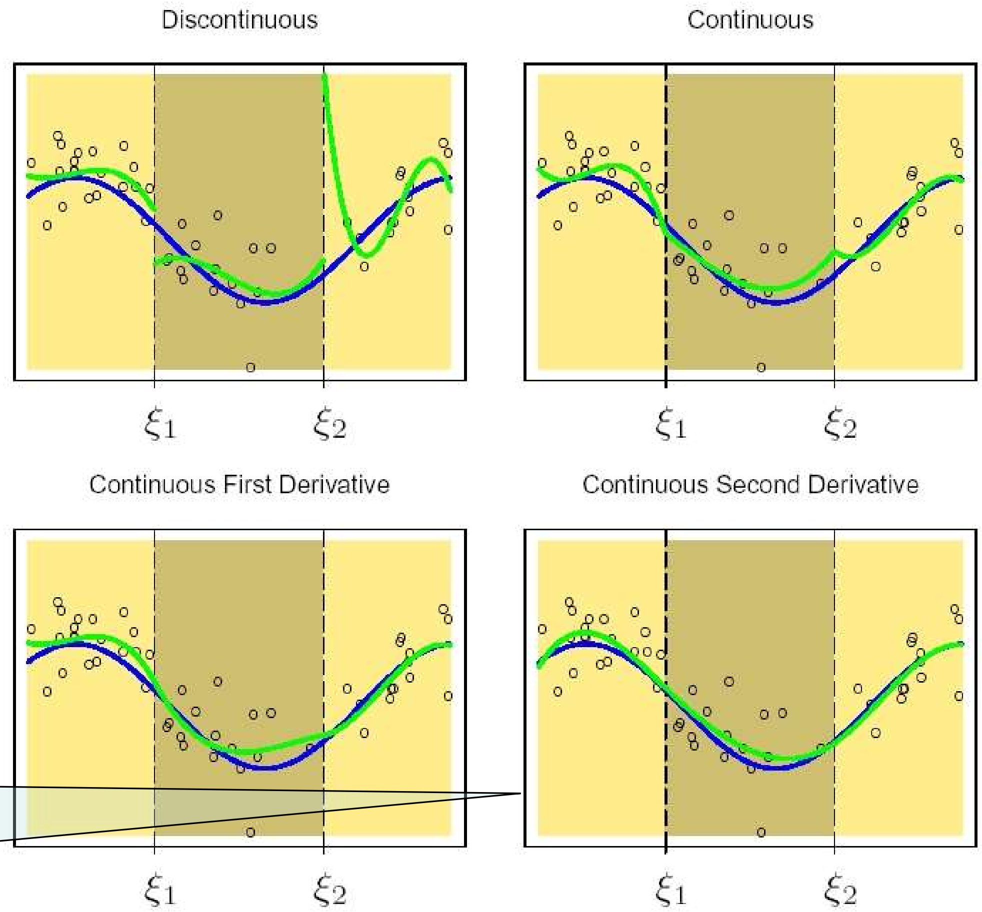
$$h_1(X) = 1, \quad h_2(X) = X,$$
$$h_3(X) = (X - \xi_1)_+,$$
$$h_4(X) = (X - \xi_2)_+$$



Piecewise Cubic Polynomials

- Increasing orders of continuity at the knots.
- A cubic spline with knots at ξ_1 and ξ_2:

$$1, X, X^2, X^3,$$
$$(X-\xi_1)_+^3, (X-\xi_2)_+^3;$$



Cubic spline truncated power basis

Piecewise Cubic Polynomials

- An order-M spline with knots , j=1,…,K is a piecewise-polynomial of order M, and has continuous derivatives up to order M-2.
- A cubic spline has $M=4$.
- Truncated power basis set:

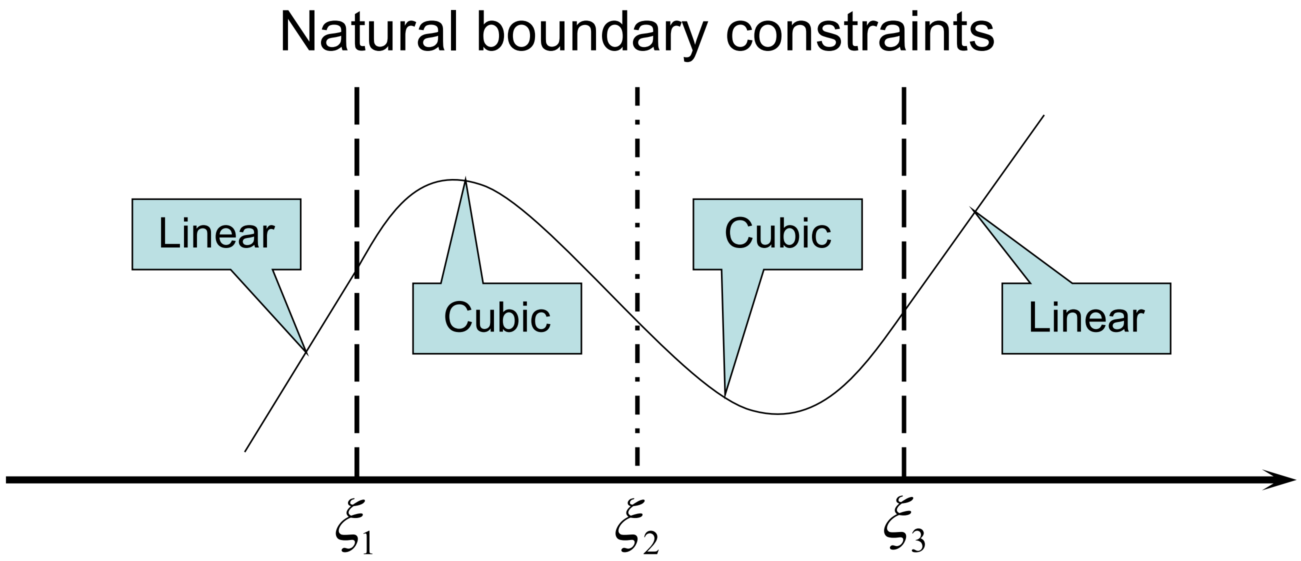
$$h_j(X) = X^{j-1}, \quad j = 1, \ldots, M$$

$$h_{M+l}(X) = (X - \xi_l)_+^{M-1}, \quad l = 1, \ldots, M$$

- Problem: are these basis functions good for generalization?
  - No
- Solution: to find piecewise polynomials with local supports.

Natural cubic spline（自然三次样条）

- Natural cubic spline adds additional constraints, namely that the function is linear beyond the boundary knots.



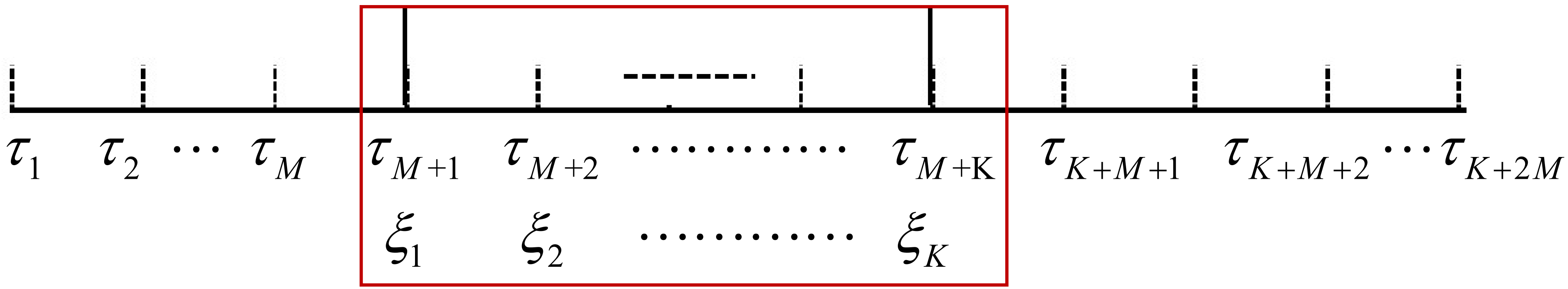
B-spline

- The augmented knot sequence τ:

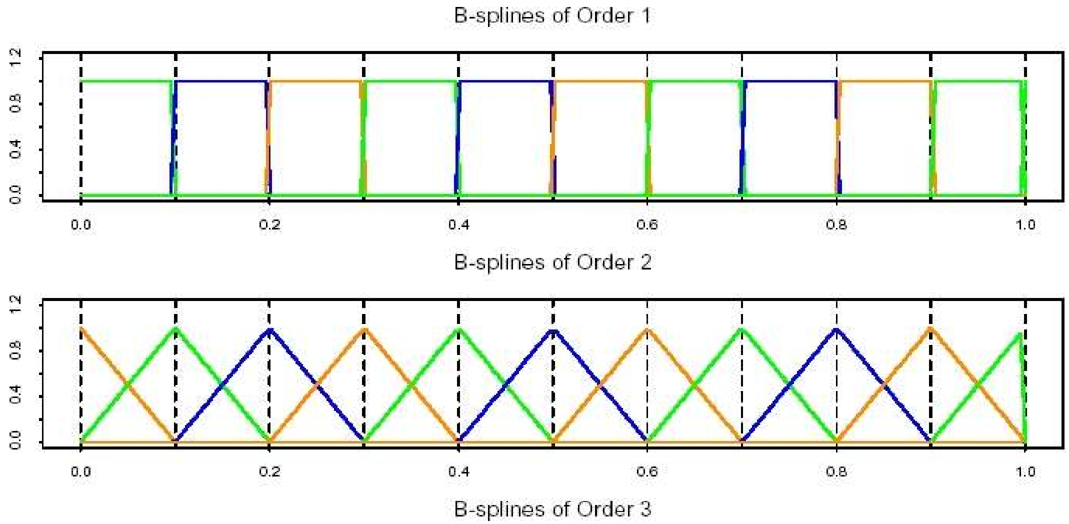
$$\tau_1 \le \tau_2 \le \cdots \le \tau_M \le \xi_0;$$

$$\tau_{j+M} = \xi_j, \quad j = 1, \cdots, K;$$

$$\xi_{K+1} \le \tau_{K+M+1} \le \tau_{K+M+2} \le \cdots \le \tau_{K+2M}$$



B-spline

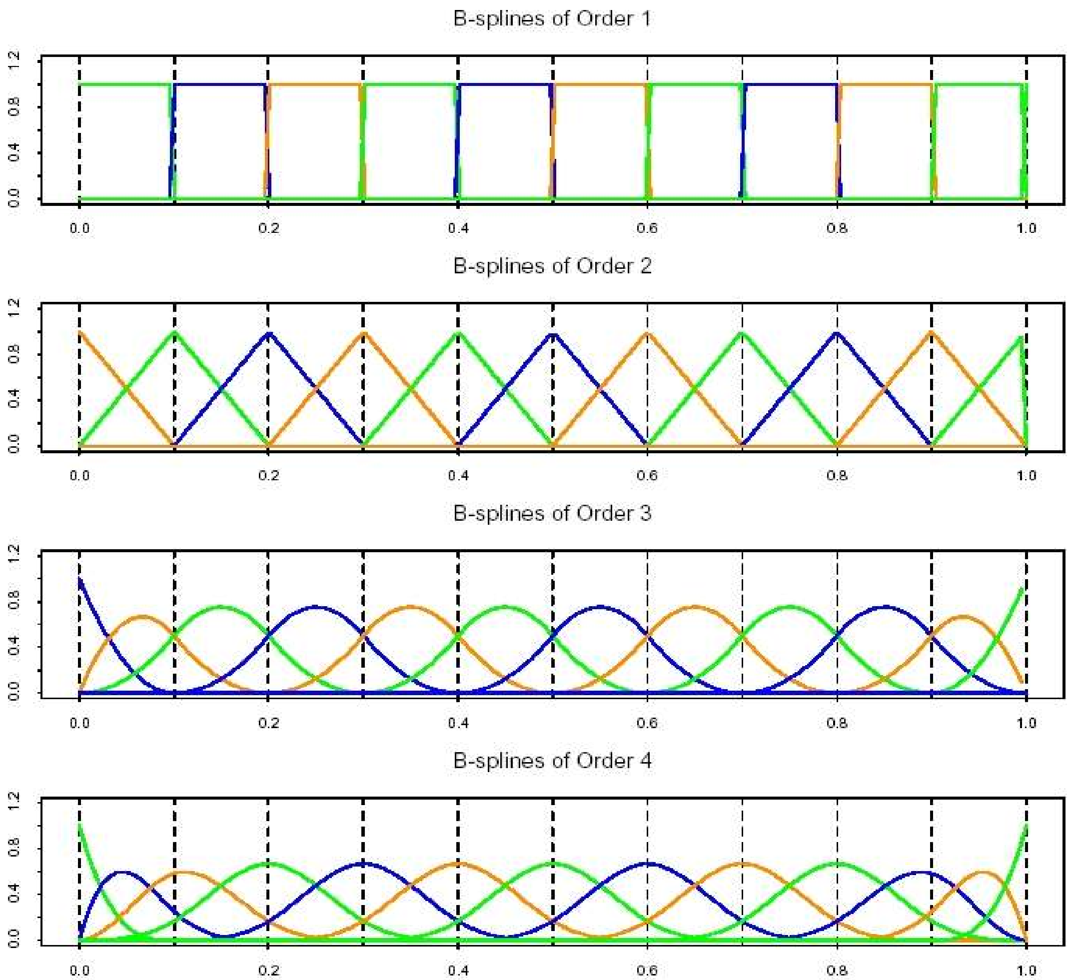


- $B_{i,m}(x)$, the i-th B-spline basis function of order m for the knot-sequence τ, $m \leq M$.

$$B_{i,1}(x) = \begin{cases} 1 & \tau_i \leq x < \tau_{i+1} \\ 0 & others \end{cases} \quad i = 1, \cdots, K + 2M - m$$

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i.m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x)$$

B-spline

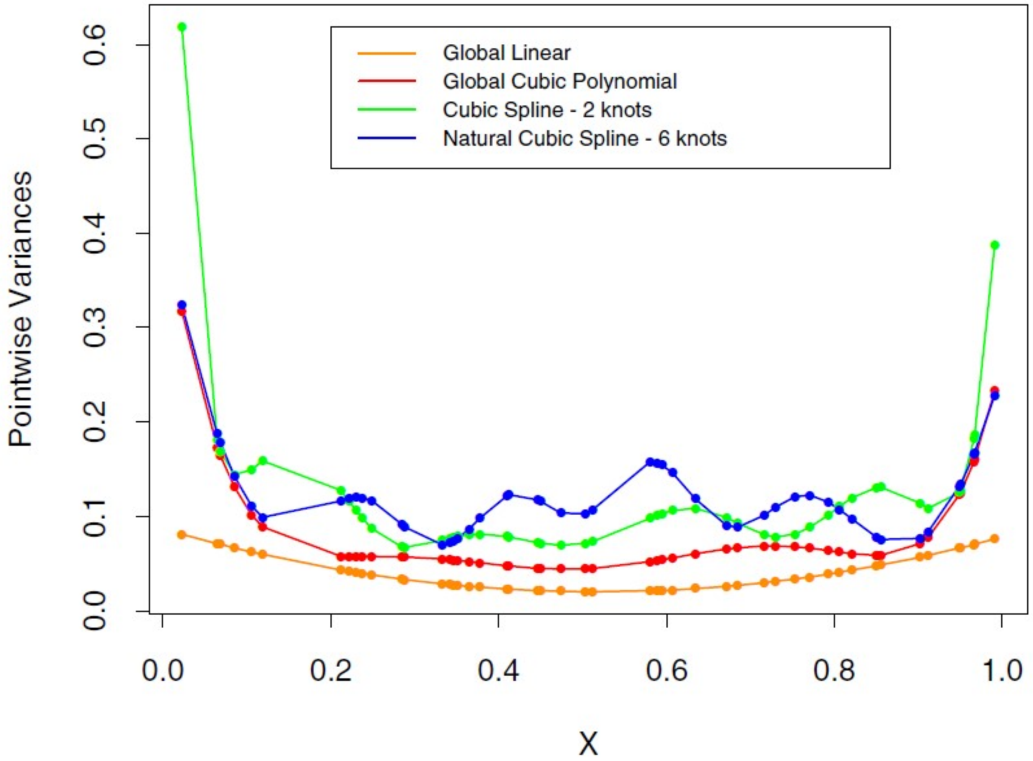


- The sequence of B-spline up to order 4 with ten knots evenly spaced from

  0 to 1

- The B-spline have local support; they are nonzero on an interval spanned by M+1 knots.

Boundary Effect in Variances



$$f(x) = \sum_{k=1}^{N} \beta_k h_k(x) = \beta^{\mathrm{T}} h(x)$$

$$\mathrm{var}(\hat{f}(x)) = h(x)^T (H^T H)^{-1} h(x) \sigma^2$$

Smoothing Splines（平滑样条）

- Base on the spline basis method: $f(x)=\sum_{k=1}^{N}\beta_k h_k(x)$
- So $y=\sum_{k=1}^{m}\beta_k h_k(x)+\varepsilon$, ε is the noise.
- Minimize the penalized residual sum of squares

$$RSS(f,\lambda)=\sum_{i=1}^{N}\{y_i-f(x_i)\}^2+\lambda\int\{f''(t)\}^2dt$$

λ is a fixed smoothing parameter

$\lambda=0$: f can be any function that interpolates the data

$\lambda=\infty$: the simple least squares line fit

Smoothing Splines

- The solution is a natural spline: $$f(x)=\sum_{j=1}^{N} N_j(x)\theta_j$$
- Then the criterion reduces to:

$$RSS(\theta,\lambda)=\left(y-\mathbf{N}\theta\right)^T\left(y-\mathbf{N}\theta\right)+\lambda\theta^T\Omega_N\theta$$

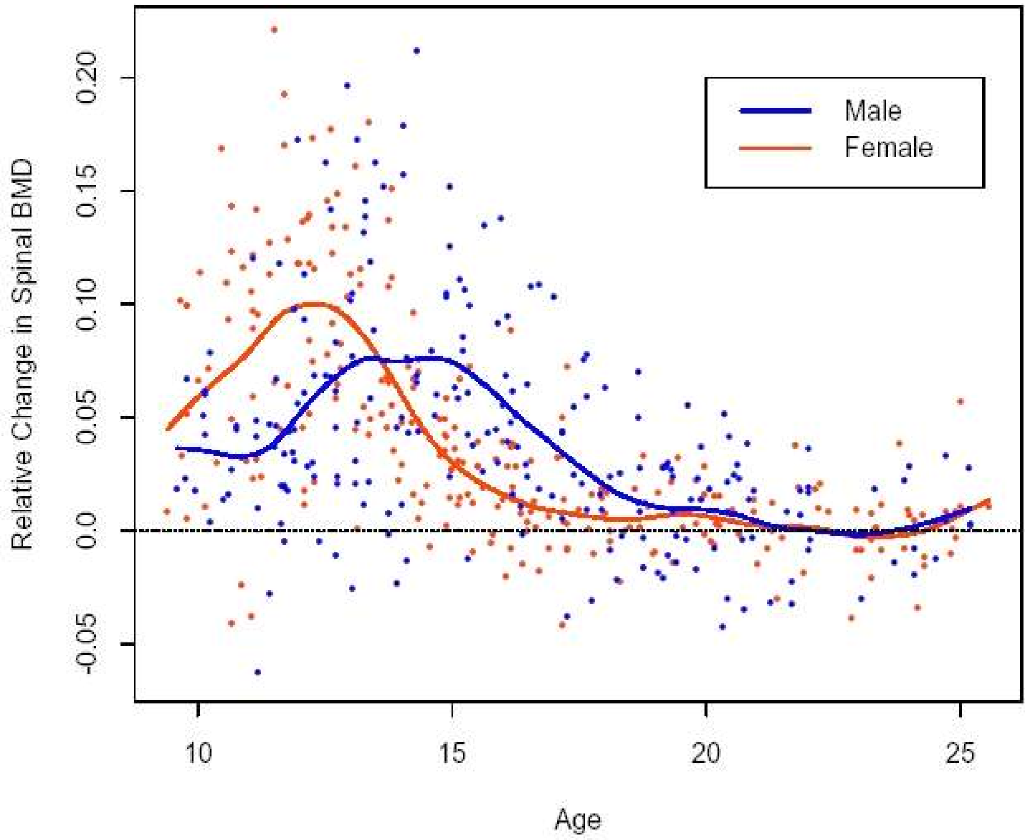
  – where $$\mathbf{N}=\{N_j(x_i)\}_{N\times N};\quad \Omega_{Nij}=\int N_i''(t)N_j''(t)dt$$

- So the solution: $$\hat{\theta}=(\mathbf{N}^T\mathbf{N}+\lambda\Omega_N)^{-1}\mathbf{N}^Ty$$

- The fitted smoothing spline:

$$\hat{f}(x)=\sum_{j=1}^{N} N_j(x)\hat{\theta}_j$$

Smoothing Splines



脊骨BMD—骨质密度

- The relative change in bone mineral density measured at the spline in adolescents
- Separate smoothing splines fit the males and females,
- 12 degrees of freedom

$$\lambda \approx 0.00022$$

Smoothing Matrix（平滑矩阵）

- $\hat{f}$ the N-vector of fitted values $\quad \hat{f} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T y = S_\lambda y$

- The finite linear operator S_λ — the smoother matrix

- Compare with the linear operator in the LS-fitting:

  M cubic-spline basis functions, knot sequence ξ

$$\hat{f} = B_\xi (B_\xi{}^T B_\xi)^{-1} B_\xi{}^T y = H_\xi y, \quad B_\xi \ is\ N \times M\ matrix$$

- Similarities and differences:
  - Both are symmetric, positive semidefinite matrices
  - $H_\xi H_\xi = H_\xi$ idempotent（幂等的）; $S_\lambda S_\lambda \le S_\lambda$ shrinking
  - rank: $r(S_\lambda) = N, r(H_\xi) = M$

Smoothing Matrix

- **Effective degrees of freedom** of a smoothing spline

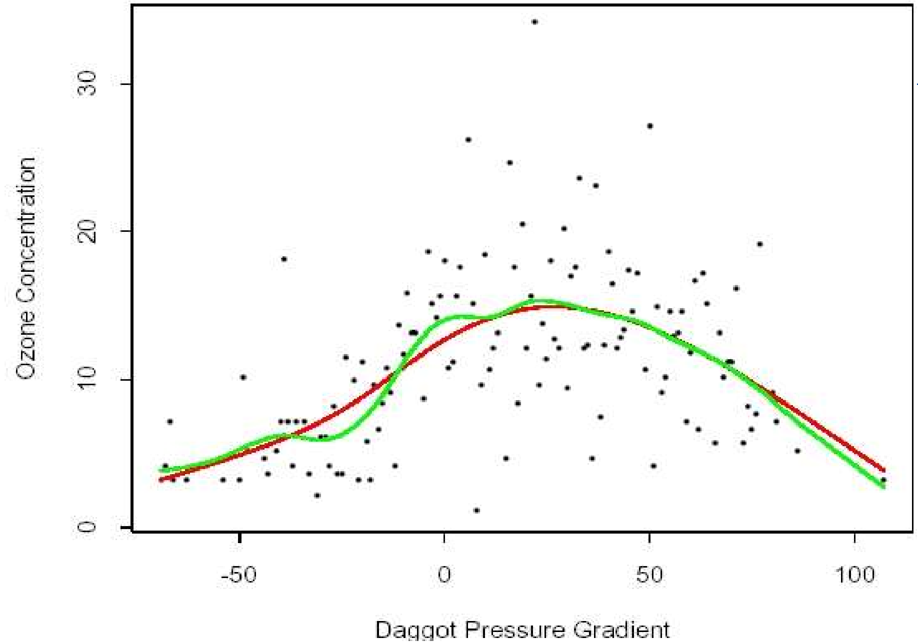
$$df_\lambda = trace(S_\lambda)$$

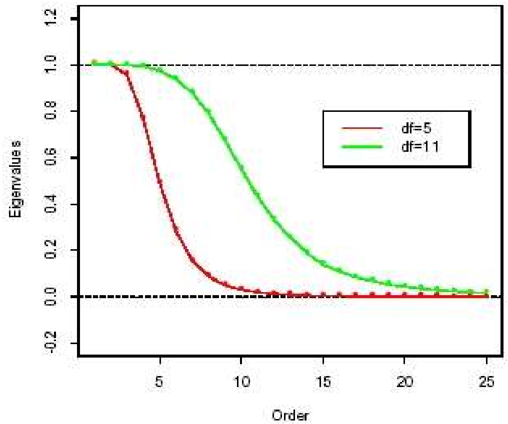
- S_λ in the ***Reinsch*** form: $S_\lambda = (I + \lambda K)^{-1}$
- Since $\hat{f} = S_\lambda y$, solution: $\min_f \|y - f\|^2 + \lambda f^T K f$
- S_λ is symmetric and has a real eigen-decomposition

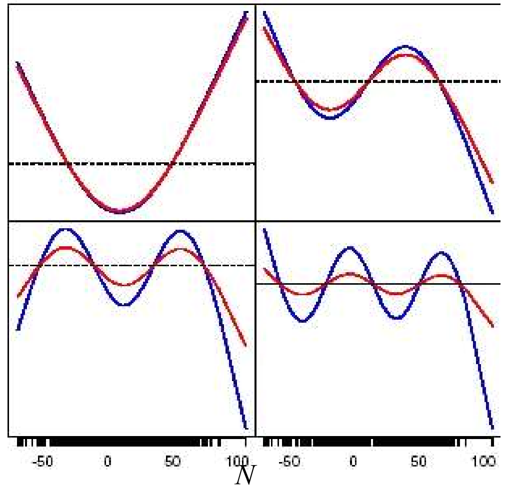
$$S_\lambda = \sum_{k=1}^{N} \rho_k(\lambda) u_k u_k^T, \qquad \rho_k(\lambda) = \frac{1}{1 + \lambda d_k}$$

  – d_k is the corresponding eigenvalue of K

- Smoothing spline fit of ozone(臭氧) concentration versus Daggot pressure gradient.
- Smoothing parameter df=5 and df=10.
- The 3rd to 6th eigenvectors of the spline smoothing matrices



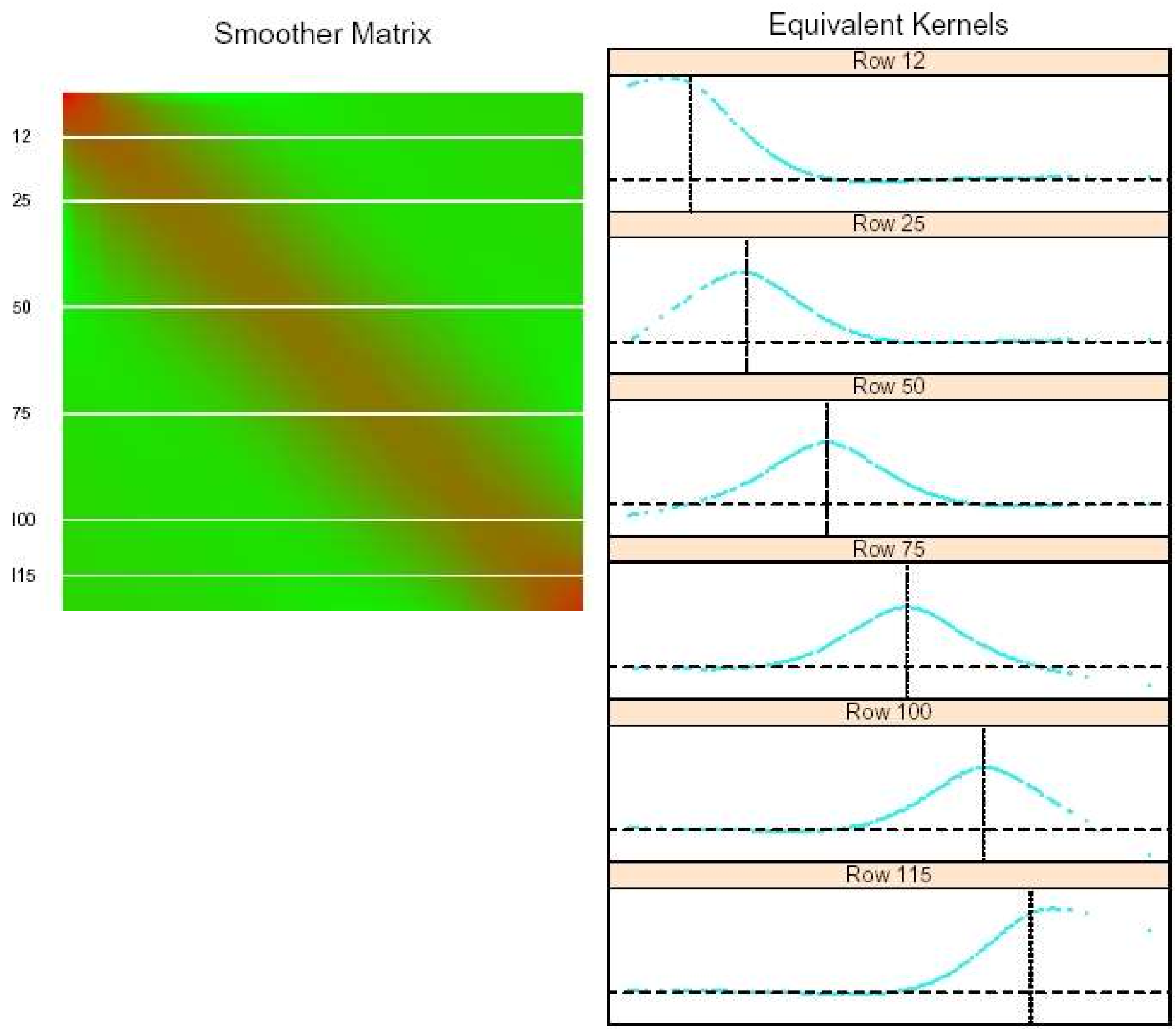




$$\rho_k(\lambda) = \frac{1}{1 + \lambda d_k}$$

$$S_\lambda = \sum_{k=1}^{N} \rho_k(\lambda) u_k u_k^T$$

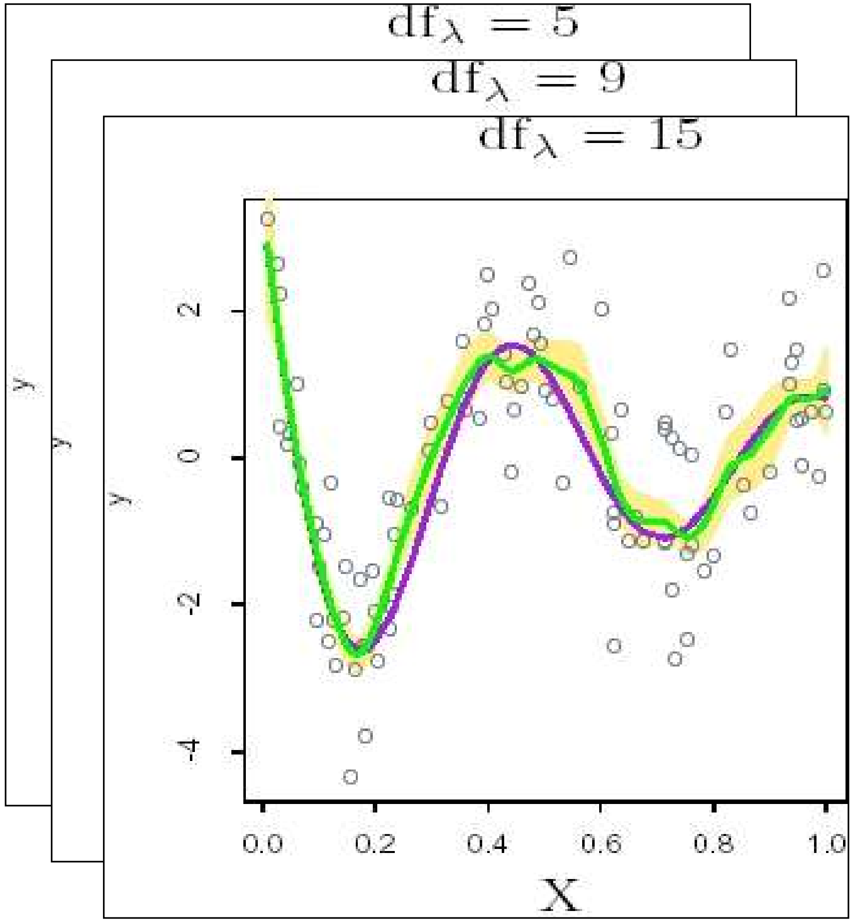
- The smoother matrix for a smoothing spline is nearly banded, indicating an equivalent kernel with local support.



Bias-Variance Tradeoff

- Example: $Y = f(X) + \varepsilon, \qquad \operatorname{cov}(\varepsilon) = \sigma_\varepsilon^2$
  $$f(X) = \frac{\sin(12(X+0.2))}{X+0.2}$$
- For $\hat{f} = S_\lambda y, \quad then \ \operatorname{cov}(\hat{f}) = S_\lambda \operatorname{cov}(y) S_\lambda^T = \operatorname{cov}(\varepsilon) S_\lambda S_\lambda^T = \boxed{\sigma_\varepsilon^2 S_\lambda S_\lambda^T}$
- The diagonal contains the pointwise variances at the training x_i
- Bias is given by $Bias(\hat{f}) = f - E(\hat{f}) = f - S_\lambda f$
- f is the (unknown) vector of evaluations of the true $\boldsymbol{f}$

Bias-Variance Tradeoff



- df=5, bias high, standard error band narrow
- df=9, bias slight, variance not increased appreciably
- df=15, over learning, standard error widen

*

Bias-Variance Tradeoff

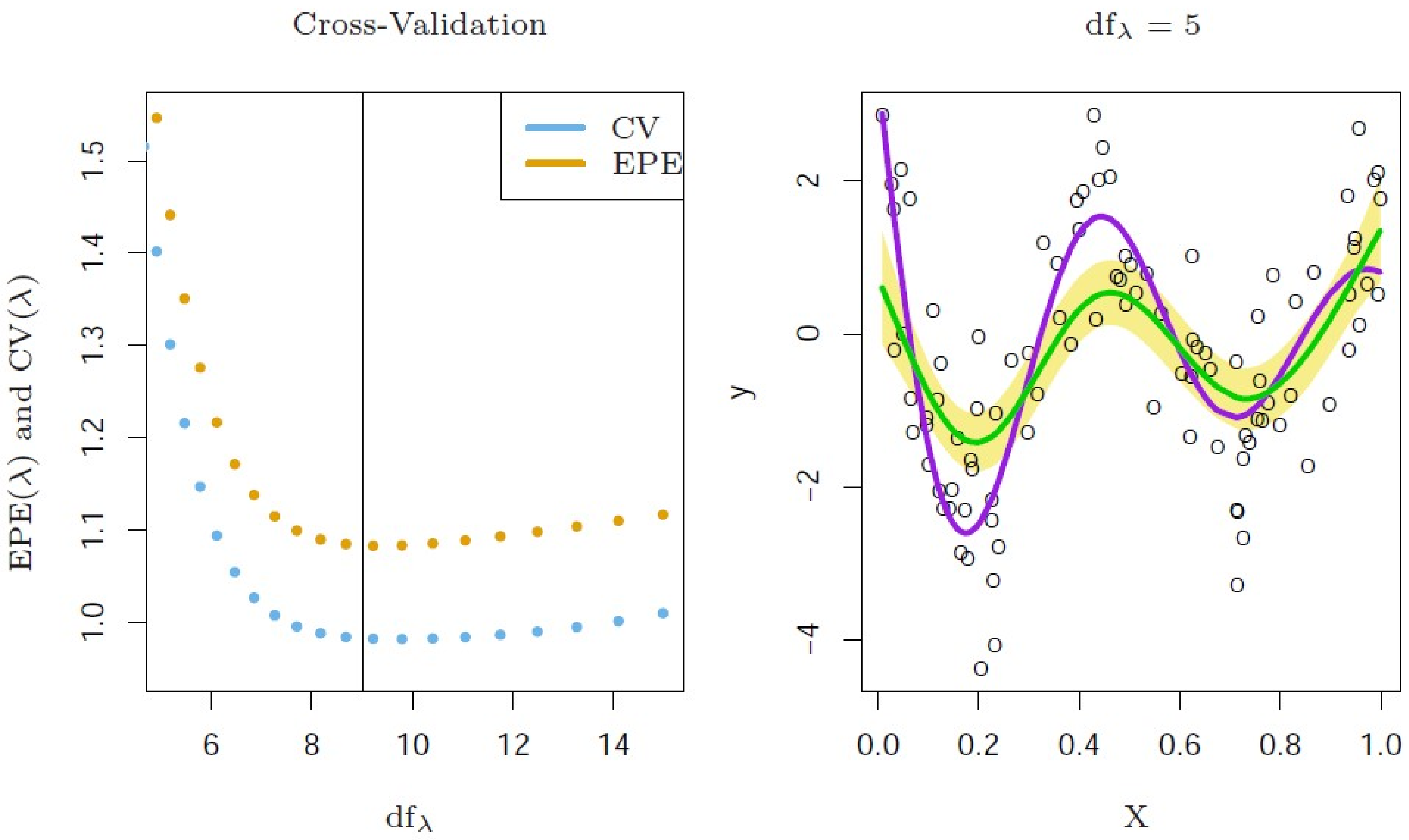
- The integrated squared prediction error (EPE) combines both bias and variance in a single summary:

$$
\begin{aligned}
EPE(\hat{f}_\lambda) &= E(Y - \hat{f}_\lambda(X))^2 \\
&= Var(Y) + E\left[Bias^2(\hat{f}_\lambda(X)) + Var(\hat{f}_\lambda(X)) \right] \\
&= \sigma^2 + MSE(\hat{f}_\lambda)
\end{aligned}
$$

- **N fold (leave one)** cross-validation:

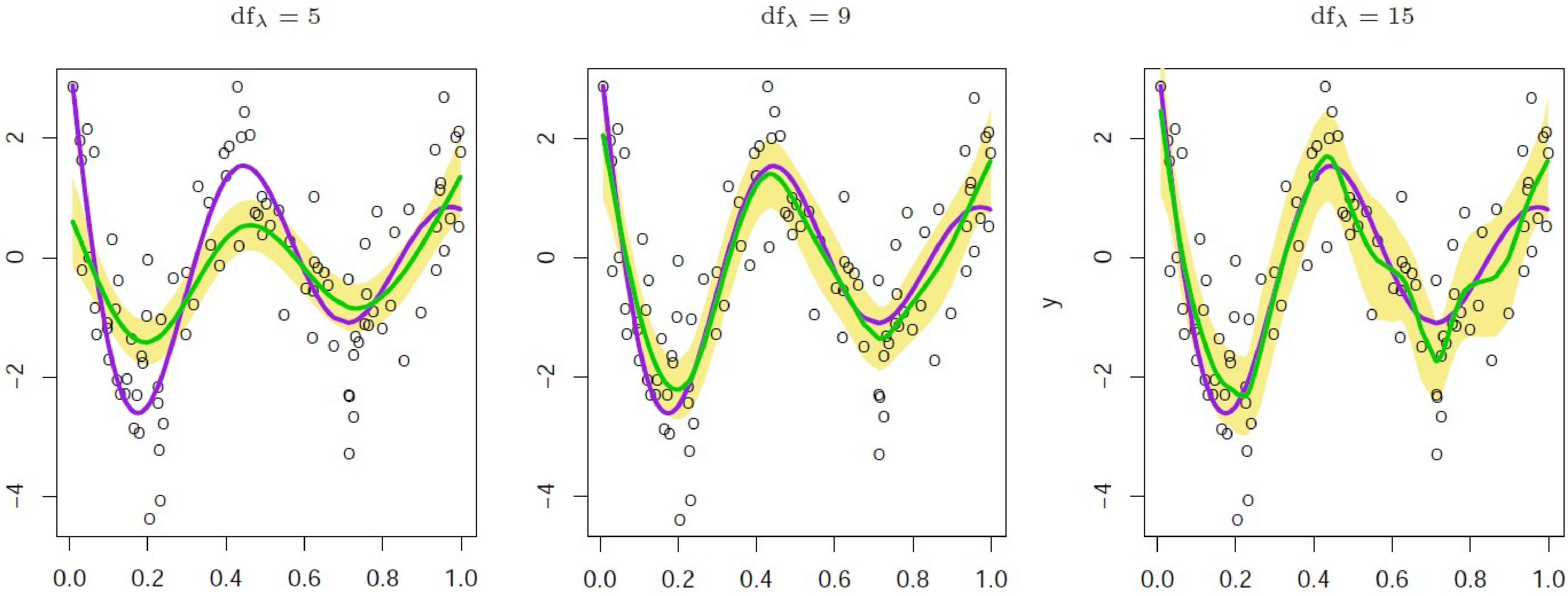
$$
CV(\hat{f}_\lambda) = \sum_{i=1}^{N} (y - \hat{f}_\lambda^{-i}(x_i))^2 = \sum_{i=1}^{N} \left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_\lambda(i,i)} \right)^2
$$

Bias-Variance Tradeoff



The EPE and CV curves have the a similar shape.
And, overall the CV curve is approximately unbiased as an estimate of the EPE curve

Bias-Variance Tradeoff



Outline

- Piece-wise Polynomials and Splines
- Smoothing Splines
- Automatic Selection of the Smoothing Parameters
- Nonparametric Logistic Regression
- Multidimensional Splines
- Regularization and Reproducing Kernel Hilbert Spaces
- Wavelet Smoothing

App: Logistic Regression

- Logistic regression with a single quantitative input X

$$\log \frac{\Pr(Y=1 \mid X=x)}{\Pr(Y=0 \mid X=x)} = f(x)$$

$$\Pr(Y=1 \mid X=x) = \frac{e^{f(x)}}{1+e^{f(x)}}$$

- The penalized log-likelihood criterion

$$l(f;\lambda) = \boxed{\sum_{i=1}^{N} \left[y_i \log p(x_i) + (1-y_i)\log(1-p(x_i)) \right]} \boxed{-\frac{1}{2}\lambda \int \{f''(t)\}^2 dt}$$

$$= \sum_{i=1}^{N} \left[y_i f(x_i) + \log(1+e^{f(x_i)}) \right] - \frac{1}{2}\lambda \int \{f''(t)\}^2 dt$$

Multidimensional Splines

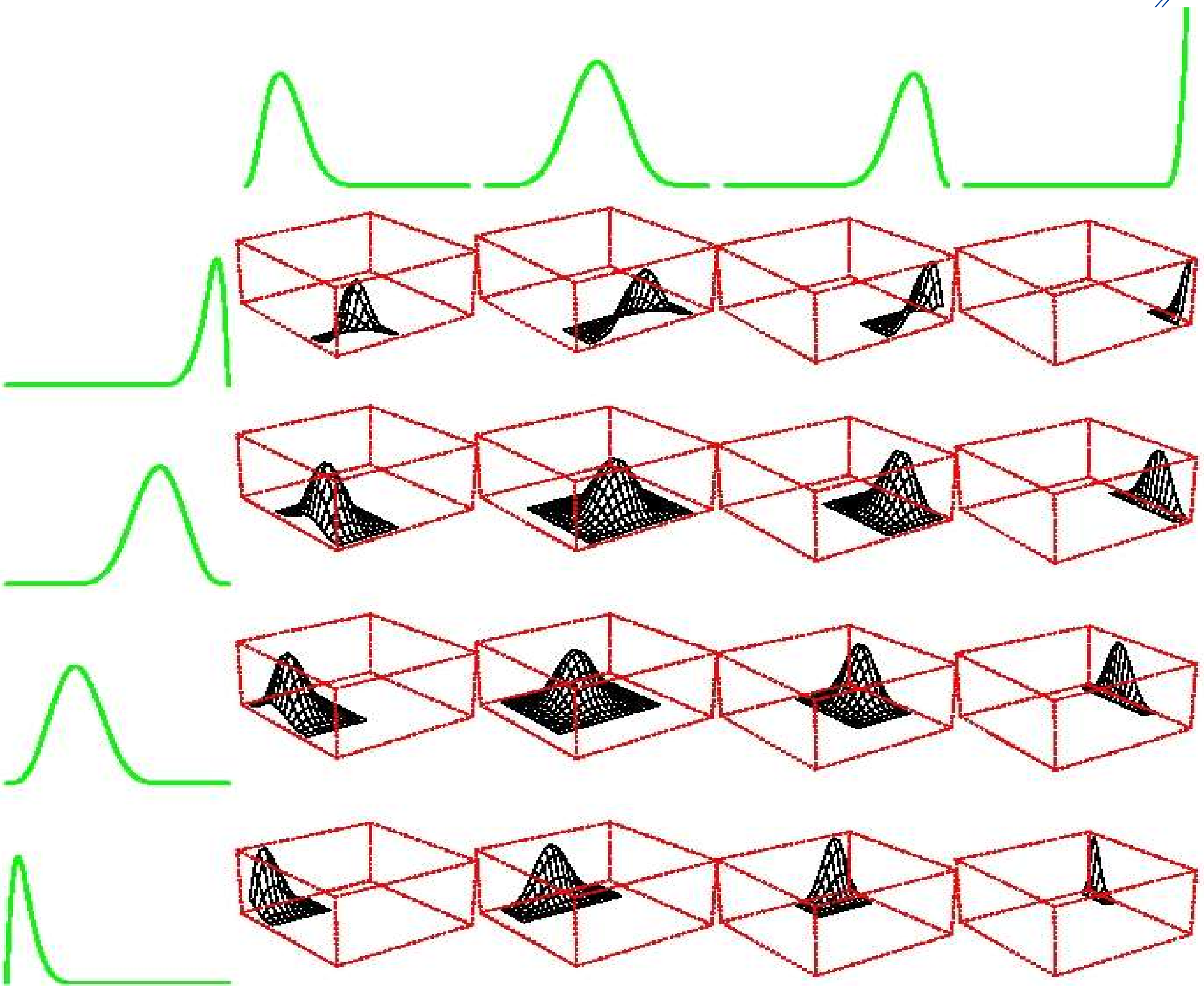
- Tensor product basis
  - The M1×M2 dimensional tensor product basis

$$g_{jk}(X) = h_{1j}(X_1)h_{2k}(X_2), \quad j = 1,\ldots,M_1, \quad k = 1,\ldots,M_2$$

  - $h_{1j}(X_1)$, basis function for coordinate X1
  - $h_{2k}(X_2)$, basis function for coordinate X2

$$g(X) = \sum_{j=1}^{M_1}\sum_{k=1}^{M_2}\theta_{jk}g_{jk}(X)$$

Tenor product basis of B-splines



Multidimensional Splines

- High dimension smoothing Splines

$$\min_f \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda J[f], \quad x_i \in IR^d$$

  - J is an appropriate penalty function

$$J[f] = \int\int_{IR^2} \left[\left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2\left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^2 \right] dx_1 dx_2$$

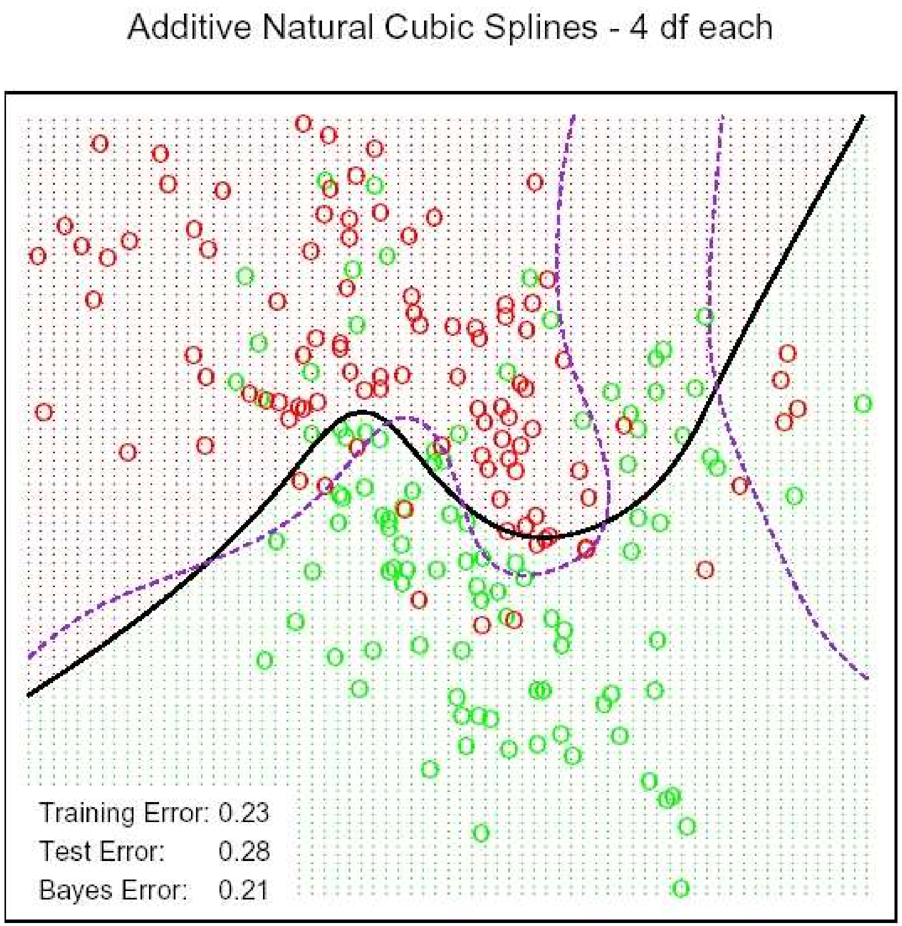
a smooth two-dimensional surface, a thin-plate spline.

- The solution has the form

$$f(x) = \beta_0 + \beta^T x + \sum_{j=1}^{N} \alpha_j h_j(x)$$

$$h_j(x) = \|x - x_j\|^2 \log\|x - x_j\| \quad \text{---- radial basis functions}$$

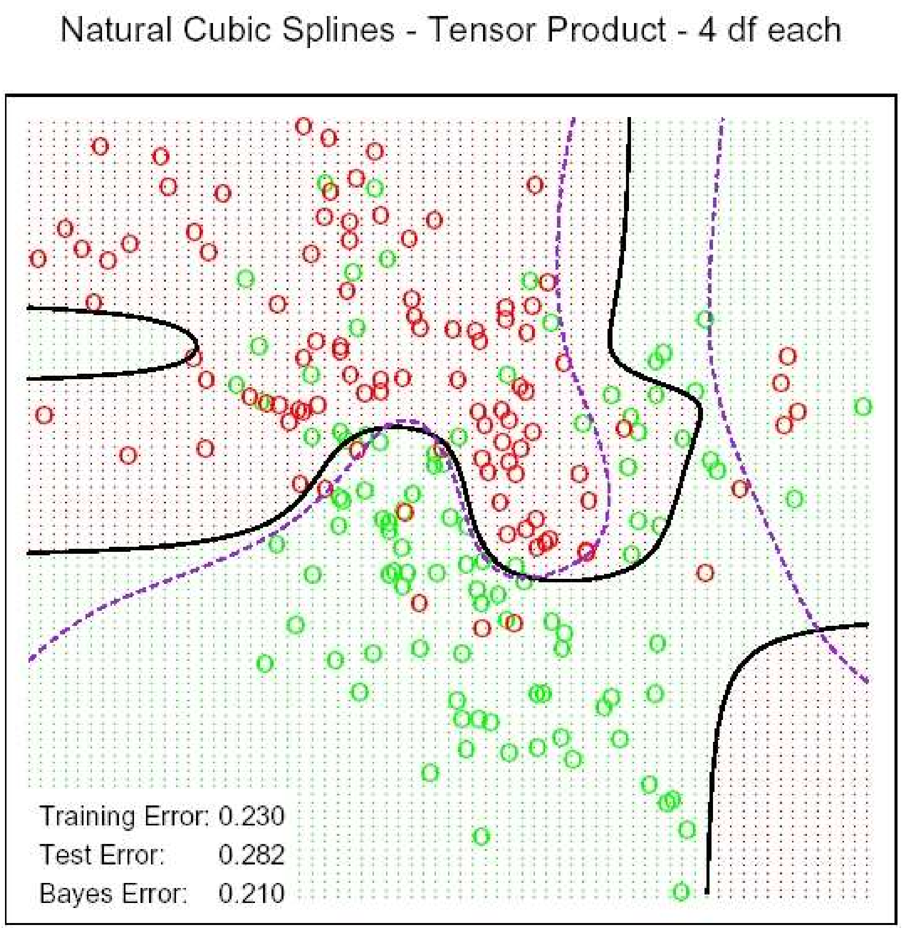
Multidimensional Splines



- The decision boundary of an additive logistic regression model. Using natural splines in each of two coordinates.
- df = 1 +(4-1) + (4-1) = 7

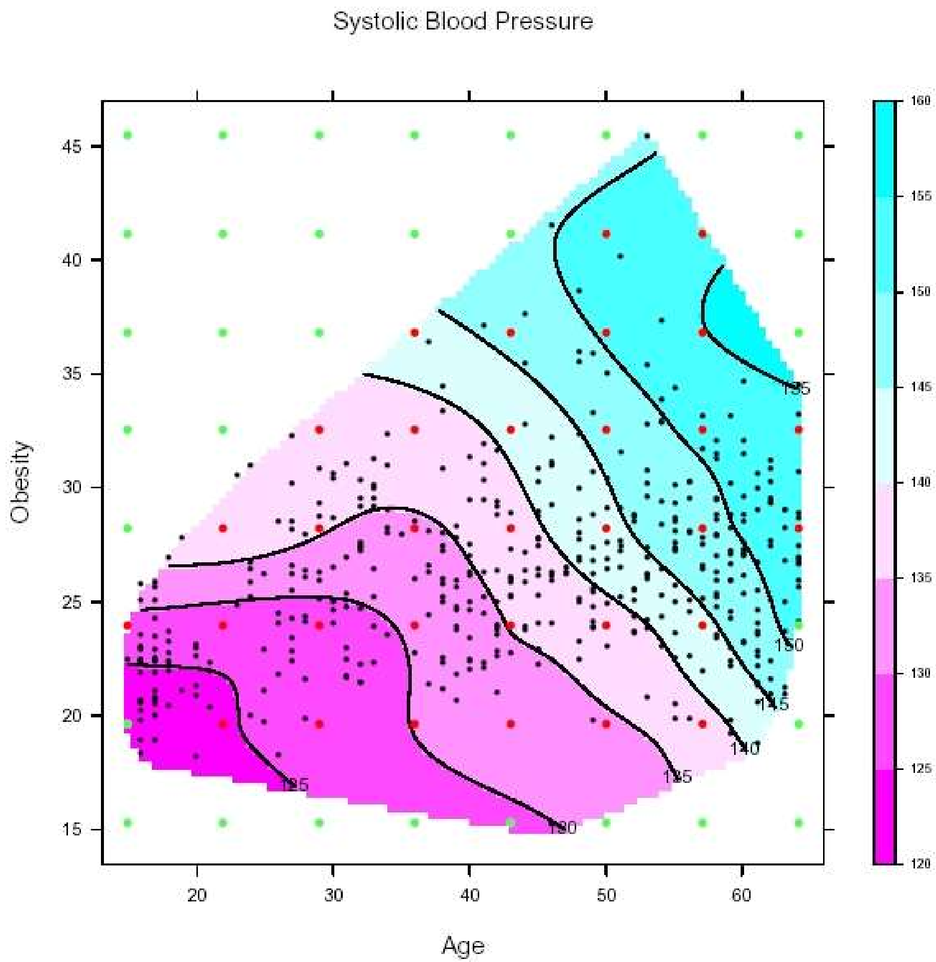
Multidimensional Splines

- The results of using a tensor product of natural spline basis in each coordinate.
- df = 4 x 4 = 16



Multidimensional Splines

- A thin-plate spline fit to the heart disease data.
- The data points are indicated, as well as the lattice of points used as knots.



Outline

- Piece-wise Polynomials and Splines
- Smoothing Splines
- Automatic Selection of the Smoothing Parameters
- Nonparametric Logistic Regression
- Multidimensional Splines
- Regularization and Reproducing Kernel Hilbert Spaces
- Wavelet Smoothing

Reproducing Kernel Hilbert space

- A regularization problems has the form:

$$\min_{f \in H} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \qquad J(f) = \int \frac{\left|\tilde{f}(s)\right|^2}{\tilde{G}(s)} ds$$

  - $L(y_i, f(x_i))$ is a loss-function.
  - $J(f)$ is a penalty functional, and H is a space of functions on which $J(f)$ is defined.

- The solution

$$f(x) = \sum_{k=1}^{K} \alpha_k \phi_k(X) + \sum_{i=1}^{N} \theta_i G(X - x_i)$$

  - ϕ_k span the null space of the penalty functional J

Spaces of Functions Generated by Kernel

- Important subclass are generated by the positive kernel $K(x,y)$.
- The corresponding space of functions H_k is called reproducing kernel Hilbert space.
- Suppose that K has an eigen-expansion

$$K(x,y)=\sum_{i=1}^{\infty}\gamma_i\varphi_i(x)\varphi_i(y),\quad \gamma_i>0,\ \sum_{i=1}^{\infty}\gamma_i^2<\infty$$

- Elements of H have an expansion

$$f(x)=\sum_{i=1}^{\infty}c_i\phi_i(x),\quad \|f\|_{H_k}^2\triangleq\sum_{i=1}^{\infty}c_i^2/\gamma_i<\infty$$

*

Spaces of Functions Generated by Kernel

- The regularization problem become

$$\min_{f \in H_k} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|_{H_k}^2 \right]$$

$$\min_{\{C_j\}_1^\infty} \left[\sum_{i=1}^{N} L(y_i, \sum_{j=1}^{\infty} c_j \phi_j(x_i)) + \lambda \sum_{j=1}^{\infty} c_j^2 / \gamma_j \right]$$

- The finite-dimension solution(Wahba,1990)

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$$

- Reproducing properties of kernel function

$$< K(\bullet, x_i), f >_{H_K} = f(x_i), \quad < K(\bullet, x_i), K(\bullet, x_j) >= K(x_i, x_j)$$

Spaces of Functions Generated by Kernel

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$$

- The penalty functional

$$J(f) = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \alpha_i \alpha_j$$

- The regularization function reduces to a finite-dimensional criterion

$$\min_{\alpha} L(y, K\alpha) + \lambda \alpha^T K \alpha, \quad K = \left[K(x_i, x_j)\right]$$

  – K is NxN matrix

RKHS

- Penalized least squares

$$\min_{\alpha}(y-K\alpha)^T(y-K\alpha)+\lambda\alpha^T K\alpha$$

- The solution: $\hat{\alpha}=(K+\lambda I)^T y$
-
- The fitted values: $\hat{f}(x)=\sum_{k=1}^{N}\hat{\alpha}_k K(x,x_k)$

- The vector of N fitted value is given by

$$\hat{f}=K\hat{\alpha}=K(K+\lambda I)^{-1}y$$
$$=(I+\lambda K^{-1})^{-1}y$$

Example of RKHS

- Polynomial regression
  - Suppose $h(x): IR^p \to IR^M$, M huge
  - Given $x_1, x_2, \ldots, x_N, \quad with\ M >> N, \quad H = \{h_j(x_i)\}$
  - Loss function: $R(\beta) = (\mathbf{y} - H\beta)^T(\mathbf{y} - H\beta) + \lambda\beta^T\beta$
  - **The penalty polynomial regression:**

$$\min_{\{\beta_m\}_1^M} \sum_{i=1}^{N}\left(y_i - \sum_{m=1}^{M}\beta_m h_m(x_i)\right)^2 + \lambda\sum_{m=1}^{M}\beta_m^2$$

–The solution:

$$\hat{f}(x) = h(x)^T\beta = \sum_{i=1}^{N}\hat{\alpha}_i K(x, x_i), \quad \hat{\alpha} = (K + \lambda I)^{-1}\mathbf{y}$$

$$\frac{\partial L(\beta)}{\partial \beta} = 0 \Rightarrow -H^T(y - H\hat{\beta}) + \lambda\hat{\beta} = 0$$

$$-HH^T(y - H\hat{\beta}) + \lambda H\hat{\beta} = 0$$

$$H\beta = (HH^T + \lambda I)^{-1}HH^T y$$

$$\{HH^T\}: < h(x_i), h(x_j) >= K(x_i, x_j)$$

Penalized Polynomial Regression

- Kernel: $K(x, x') = (1 + \langle x, x' \rangle)^d$ has $M = \binom{p+d}{d}$ eigen-functions
- E.g. d=2, p=2: M=6

$$K(x, x') = (1 + x_1 x_1' + x_2 x_2')^2 = h(x)^T h(x')$$

$$= 1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2'$$

$$h(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)$$

- The penalty polynomial regression:

$$\min_{\{\beta_m\}_1^M} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} \beta_m h_m(x_i) \right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

Penalized Polynomial Regression

Relations between polynomial basis and eigen-functions

$$\boxed{h(x) = VD_{\gamma}^{1/2}\phi(x)}$$

By definition

$$K(x, y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x)\phi_i(y) = \sum_{m=1}^{M} h_m(x)h_m(y) \qquad (1)$$

$$\gamma_k \phi_k(y) = \sum_{m=1}^{M} g_{km} h_m(y) \qquad (2)$$

where $g_{km} = \int h_m(x)\phi_k(x)dx$

Penalized Polynomial Regression

Taking inner product with eigen-functions $\phi_l(y)$ on both sides

$$\sum_{m=1}^{M} g_{km} g_{lm} = \gamma_k \delta_{kl} = \sqrt{\gamma_k}\sqrt{\gamma_l}\,\delta_{kl}$$

$$\left(\gamma_k^{-1/2} g_k^T\right)\left(\gamma_l^{-1/2} g_l\right) = \delta_{kl};\ G = \begin{pmatrix} g_1^T \\ \vdots \\ g_1^T \end{pmatrix}; \qquad G D_\gamma^{-1} G^T = I$$

i.e. $G = (g_{km})_{\infty * M}$ is orthogonal with rows.

Denote $D_\gamma = diag(\gamma_1, \gamma_2, \ldots, \gamma_M)$, $\quad V^T = D_\gamma^{-1/2} G_{M\times M}; \qquad VV^T = I_{M\times M}$

$$h(x) = [h_1(x), h_2(x), \ldots, h_M(x)]^T$$

$$\phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_M(x)]^T,$$

Penalized Polynomial Regression

Rewrite the following equation into matrix form

$$\gamma_k \phi_k(y) = \sum_{m=1}^{M} g_{km} h_m(y) \qquad (2)$$

$$D_\gamma \phi(y) = G h(y)$$

Using the relation $V^T = D_\gamma^{-1/2} G_{M \times M}$, we have

$$D_\gamma^{-1/2} D_\gamma \phi(x) = D_\gamma^{-1/2} G h(x) = V^T \phi(x)$$

$$h(x) = V V^T h(x) = V D_\gamma^{1/2} \phi(x)$$

*

RBF kernel & SVM kernel

- Gaussian Radial Basis Functions

$$K(x,y) = e^{-\|x-y\|^2/2\sigma^2};$$

$$h_j(x) = K(x,x_j); \quad j = 1,\dots,M$$

- Support Vector Machines

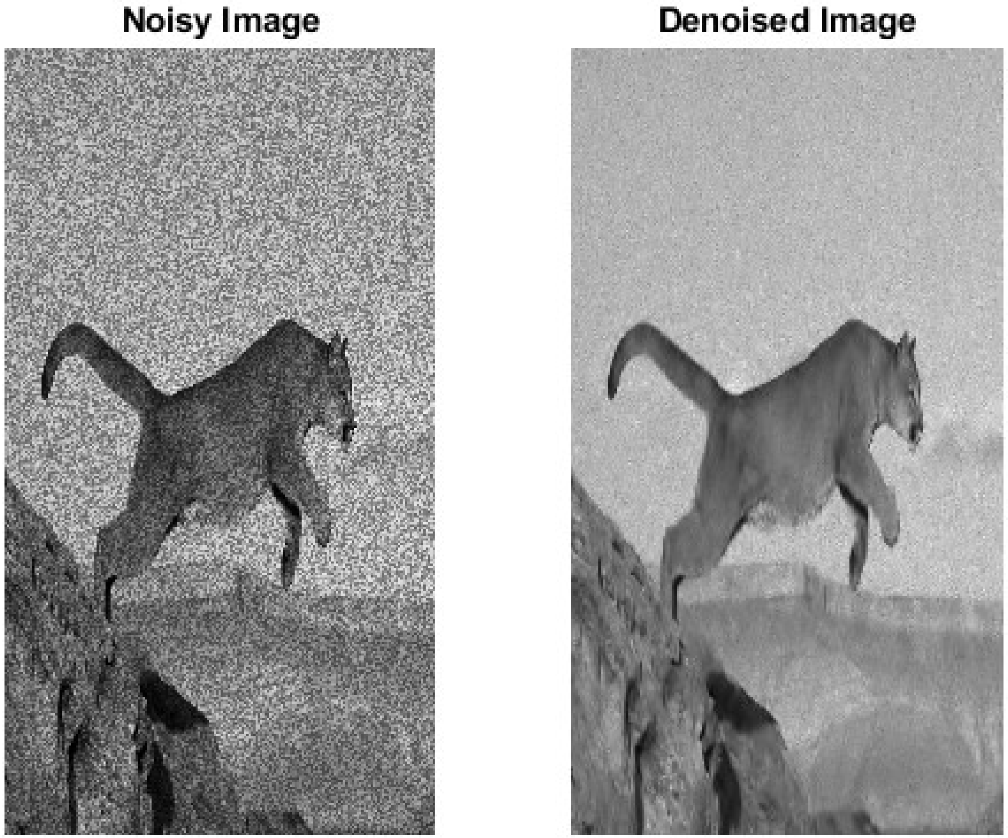
$$f(x) = \alpha_0 + \sum_{j=1}^{N} \alpha_j K(x,x_j)$$

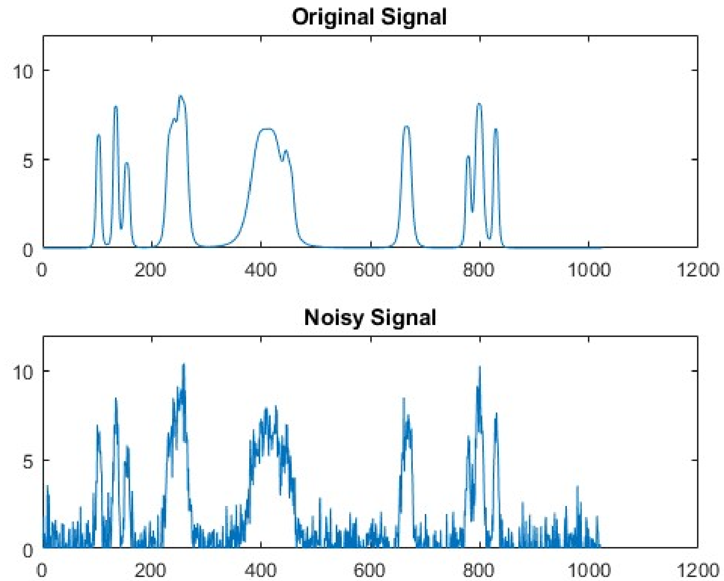
$$\min_{\alpha_0,\alpha} \left[\sum_{i=1}^{N} \left(1 - y_i f(x_i)\right)_+ + \lambda \alpha^T K \alpha \right]$$

Outline

- Piece-wise Polynomials and Splines
- Smoothing Splines
- Automatic Selection of the Smoothing Parameters
- Nonparametric Logistic Regression
- Multidimensional Splines
- Regularization and Reproducing Kernel Hilbert Spaces
- Wavelet Smoothing

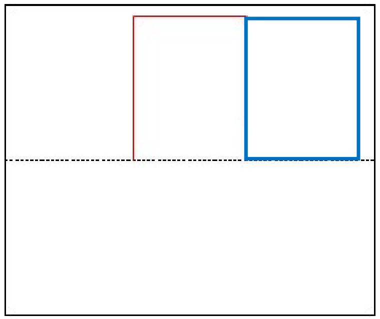
Smoothing for Denoising





Wavelet smoothing

- Another type of bases——Wavelet bases
- Wavelet bases are generated by translations and dilations of a single scaling function $\phi(x)$.
- If $\phi(x) = I\ (x \in [0\ 1])$,then $\boxed{\phi_{0,k}(x) = \phi(x-k)}$ generates an orthonormal basis for functions with jumps at the integers.
- $\phi_{0,k}(x)$ form a space called reference space V_0



$\phi(x)$

Wavelet smoothing

- The dilations $\boxed{\phi_{1,k}(x) = \sqrt{2}\phi(2x-k)}$ form an orthonormal basis for a space $V_1 \supset V_0$

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$$

$$V_0 \subset V_1 \subset V_2 \subset \cdots \qquad V_{j+1} = V_j \oplus W_j$$

$$W_j \text{ --Signal details, orthognal to } V_j.$$

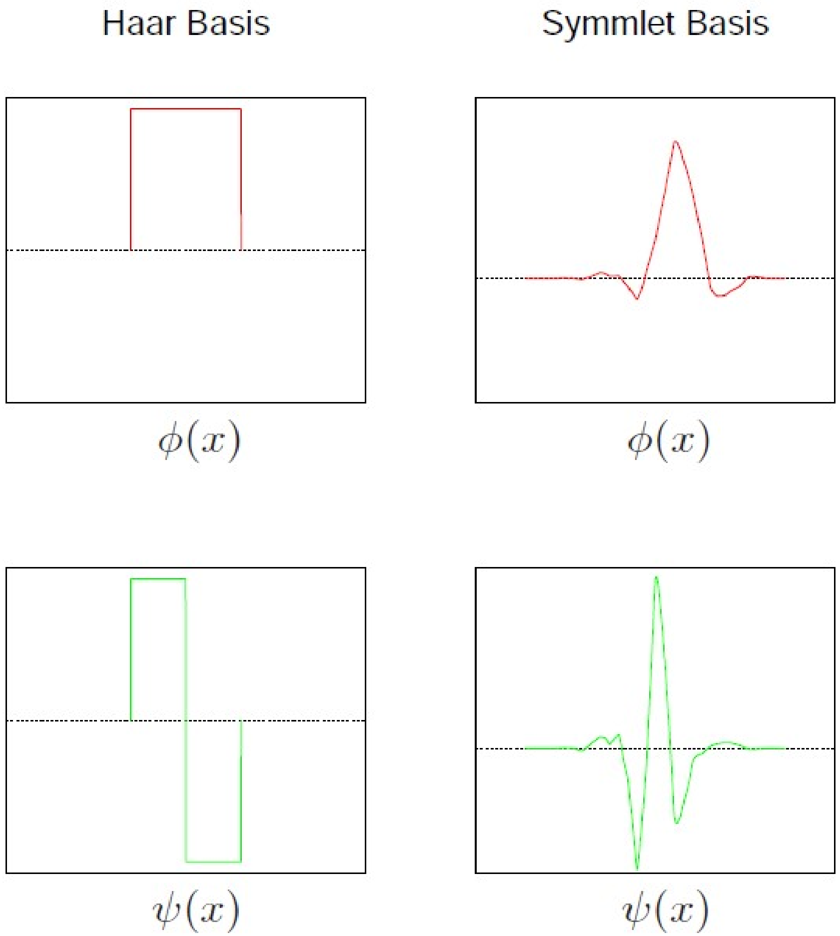
$$\psi(x) = \varphi(2x) - \varphi(2x-1) \text{ --} \mathit{Wavelet} \text{ mother bases on } W_0$$

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \text{ --} \mathit{Wavelet} \text{ bases on } W_j$$

Wavelet smoothing

- The symmlet-*p* wavelet:
  - A support of 2*p*-1 consecutive intervals.
  - *p* vanishing moments:

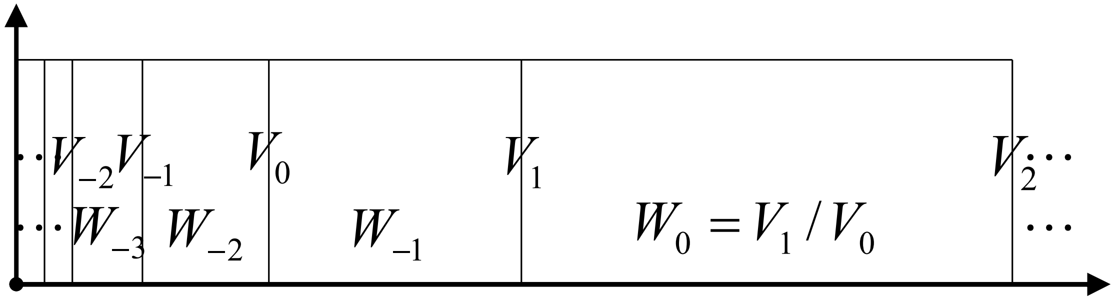
$$\int \phi(x) x^j dx = 0, \quad j = 1, \ldots, p$$



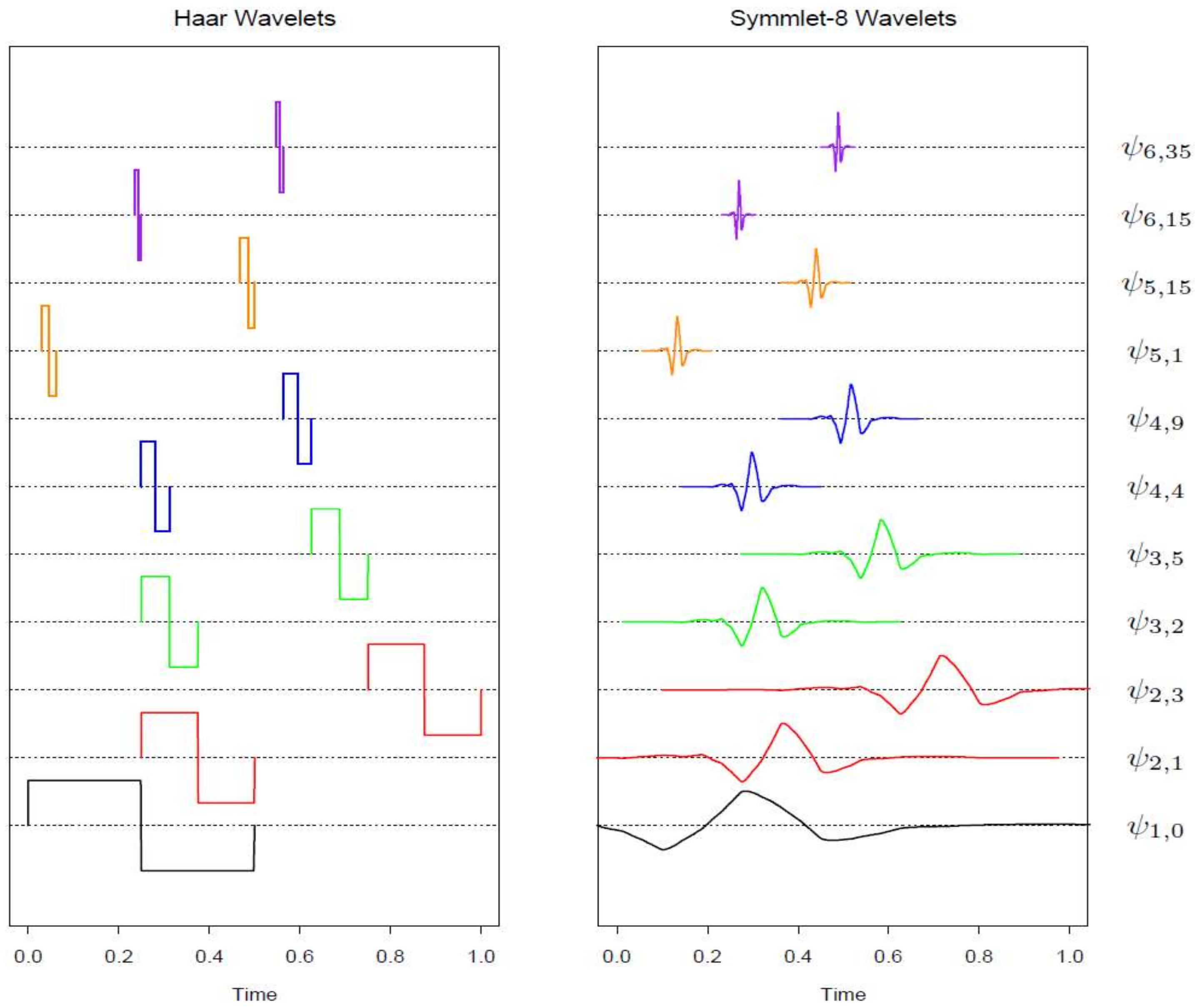
Wavelet smoothing

- The L_2 space dividing

$$V_{j+1} = V_j \oplus W_j = V_{j-1} \oplus W_{j-1} \oplus W_j$$
$$= V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_j$$



- Mother wavelet $\boxed{\psi(x) = \varphi(2x) - \varphi(2x-1)}$ generate function $\psi_{0,k} = \psi(x-k)$ form an orthonormal basis for W_0. Likewise $\boxed{\psi_{j,k} = 2^{j/2}\psi(2^j x - k)}$ form a basis for W_j.



Haar Wavelets
Symmlet-8 Wavelets
$\psi_{6,35}$
$\psi_{6,15}$
$\psi_{5,15}$
$\psi_{5,1}$
$\psi_{4,9}$
$\psi_{4,4}$
$\psi_{3,5}$
$\psi_{3,2}$
$\psi_{2,3}$
$\psi_{2,1}$
$\psi_{1,0}$
0.0
0.2
0.4
0.6
0.8
1.0
Time

Adaptive Wavelet Filtering

- Wavelet transform: $\boxed{y^* = W^T y}$
  - y: response vector, W: *N*x*N* orthonormal wavelet basis matrix

- Stein Unbiased Risk Estimation (SURE)

$$\min_\theta \|y - W\theta\|_2^2 + 2\lambda\|\theta\|_1$$

  - The solution: $\hat{\theta}_j = sign(y_j^*)(|y_j^*| - \lambda)_+$

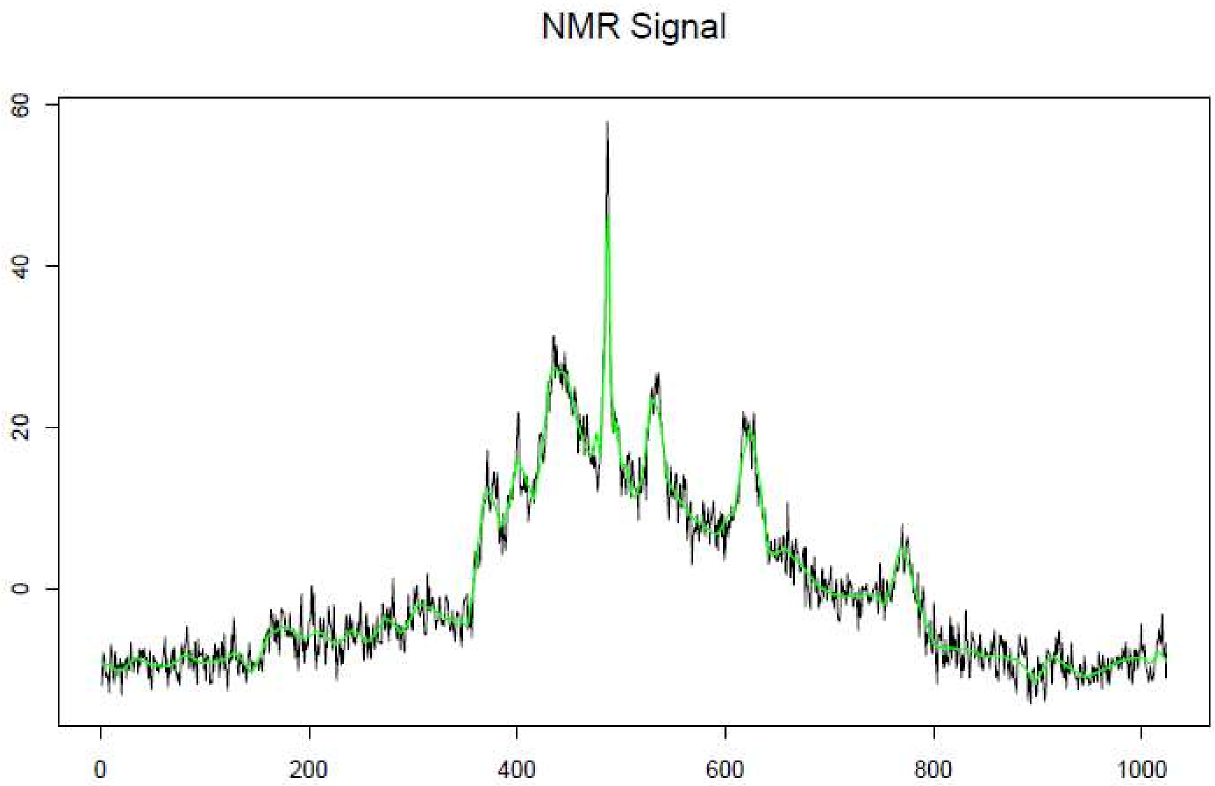
  - Fitted function is given by inverse wavelet transform: $\boxed{\hat{f} = W\hat{\theta}}$, LS coefficients truncated to 0

$$\lambda: \quad \lambda = \sigma\sqrt{2\log N}$$

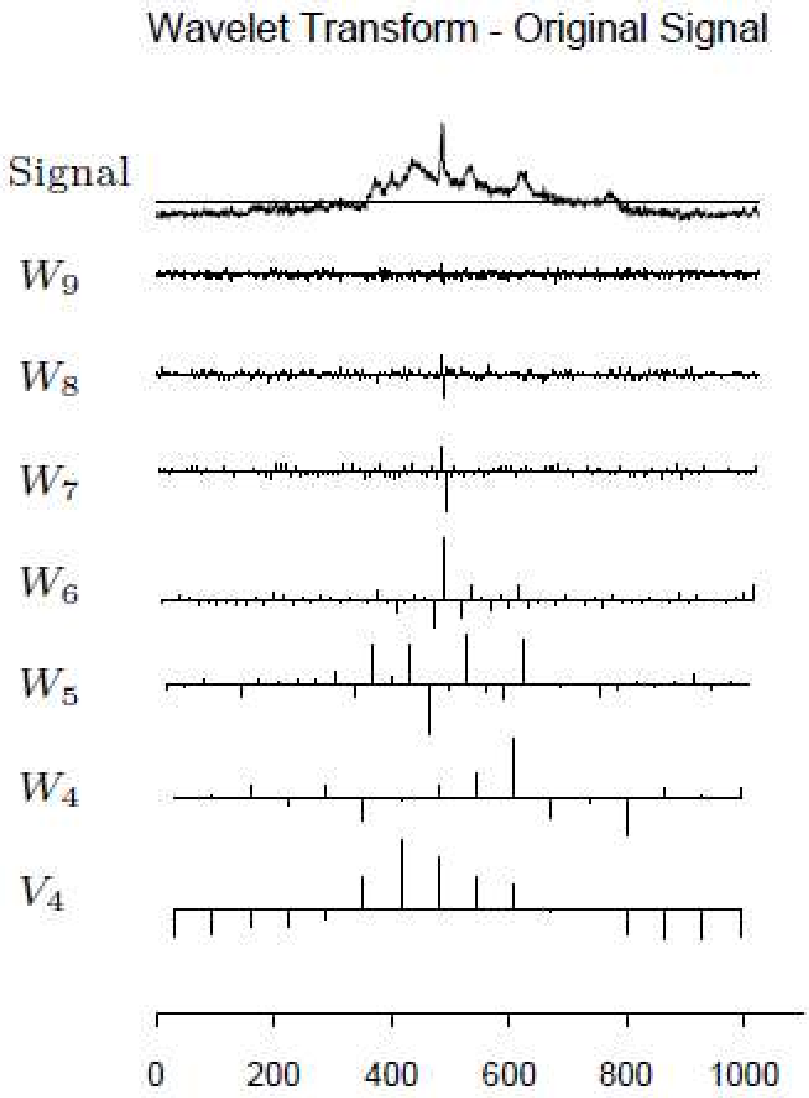
Parameter Selection

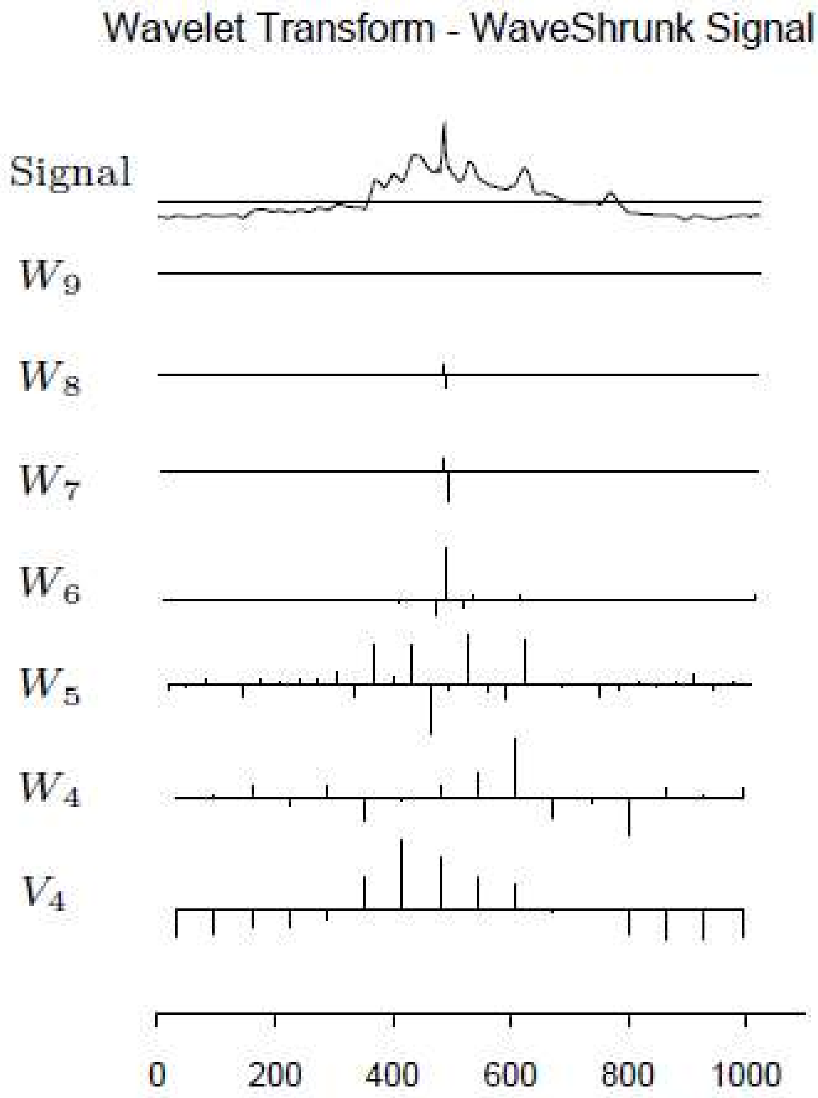
- Simple choice for $\lambda: \quad \lambda = \sigma\sqrt{2\log N}$
- If random variables $Z_1,\ Z_2,\cdots\cdots,\ Z_N$ are white noise with variance σ, the expected maximum of $|Z_i|$ is approximately $\sigma\sqrt{2\log N}$.
- Hence all coefficients below $\sigma\sqrt{2\log N}$ are likely to be noise and are set to zero.

Noise Reduction



*


Wavelet Transform - Original Signal
Signal
W_9
W_8
W_7
W_6
W_5
W_4
V_4
0
200
400
600
800
1000


Wavelet Transform - WaveShrunk Signal
Signal
W_9
W_8
W_7
W_6
W_5
W_4
V_4
0
200
400
600
800
1000

The End

Key Points in the Talk

- Good representation of function spaces
  - Easy to implement (efficient in space and time)
  - Good for generalization
  - Easy to select good models
- Good parameter for model selection
  - Effective degrees of freedom
  - CV for Model selection
- Reproducing Kernel Hilbert Space
  - Polynomial Kernel
- Spline & Wavelet

Problem

- Let $\hat{f} = Sy$ be a linear smoothing of y.

1. If S_{ii} is the i-th diagonal element of S

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}$$

Proof.

$$Smoothing\ matrx$$

$$S_\lambda = X(X^T X + \lambda\Omega)^{-1} X^T$$

$$S_{ii} = x_i^T (X^T X + \lambda\Omega)^{-1} x_i$$

$$The\ linear\ fitting\ function$$

$$\hat{f}(x_i) = x_i^T (X^T X + \lambda\Omega)^{-1} X^T y$$

Problem

Denote $X_{(-i)}$ *and* $y_{(-i)}$ be the matrix and the corresponding vectors after the removal of x_i and y_i

The linear fitting function

$$\hat{f}^{(-i)}(x_i) = x_i^T (X_{(-i)}^T X_{(-i)} + \lambda\Omega)^{-1} X_{(-i)}^T y_{(-i)}$$

Note that

$$X_{(-i)}^T X_{(-i)} = X^T X - x_i x_i^T$$

Lemma :

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + u^T A^{-1} v}$$

$$Denote \qquad A = X^T X + \lambda\Omega$$

$$
\begin{aligned}
x_i^T (X_{(-i)}^T X_{(-i)} + \lambda\Omega)^{-1} &= x_i^T (A - x_i x_i^T)^{-1} \\
&= x_i^T A^{-1} + \frac{x_i^T A^{-1} x_i \, x_i^T A^{-1}}{1 - x_i^T A^{-1} x_i} \\
&= x_i^T A^{-1} + \frac{S_{ii} x_i^T A^{-1}}{1 - S_{ii}}
\end{aligned}
$$

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + u^T A^{-1} v}$$

$$f^{(-i)}(x_i) = \left(x_i^T A^{-1} + \frac{S_{ii} x_i^T A^{-1}}{1 - S_{ii}} \right) (X^T y - x_i y_i)$$

$$= x_i^T A^{-1} X^T y + \frac{S_{ii} x_i^T A^{-1}}{1 - S_{ii}} X^T y - \frac{S_{ii} x_i^T A^{-1}}{1 - S_{ii}} x_i y_i - x_i^T A^{-1} x_i y_i$$

$$= \hat{f}(x_i) + \frac{S_{ii} \hat{f}(x_i)}{1 - S_{ii}} - \frac{S_{ii} S_{ii}}{1 - S_{ii}} y_i - S_{ii} y_i$$

$$= \frac{\hat{f}(x_i)}{1 - S_{ii}} - \frac{y_i S_{ii}}{1 - S_{ii}}$$

Then we have

$$y_i - \hat{f}^{(-i)}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}$$