ESC FALL Week3

Spline

esl ch5

오현도

박상용

Contents

Part 1

Spline Introduciton

기저 함수 – Basis Function

비선형 데이터의 예측률을 높이는
비선형 모델을 만드려면 어떻게 해야 할까?

선형 모델 : 해석하기 쉬운 f(x)에 대한 일차 테일러 근사

기저 함수 : 설명변수를 함수 형태로 나타낸 것

기저 함수와 설명변수의 선형 결합 -> 모델 표현

기저 함수 – Basis Function

$$h_m(X) : R^p \rightarrow R, m = 1, ..., M$$
$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

기저함수 h에 상관없이,
이전과 같이 **선형**

-> 계수 추정하기 위해서
least squares, 선형 모델에서
사용하던 도구들 적용이 가능함

기저함수 $h_m(X)$ 예시

- $h_m(X) = X_m, m = 1, ..., p$

  : 선형 모델

- $h_m(X) = X_j^2$
- $h_m(X) = X_j X_k$

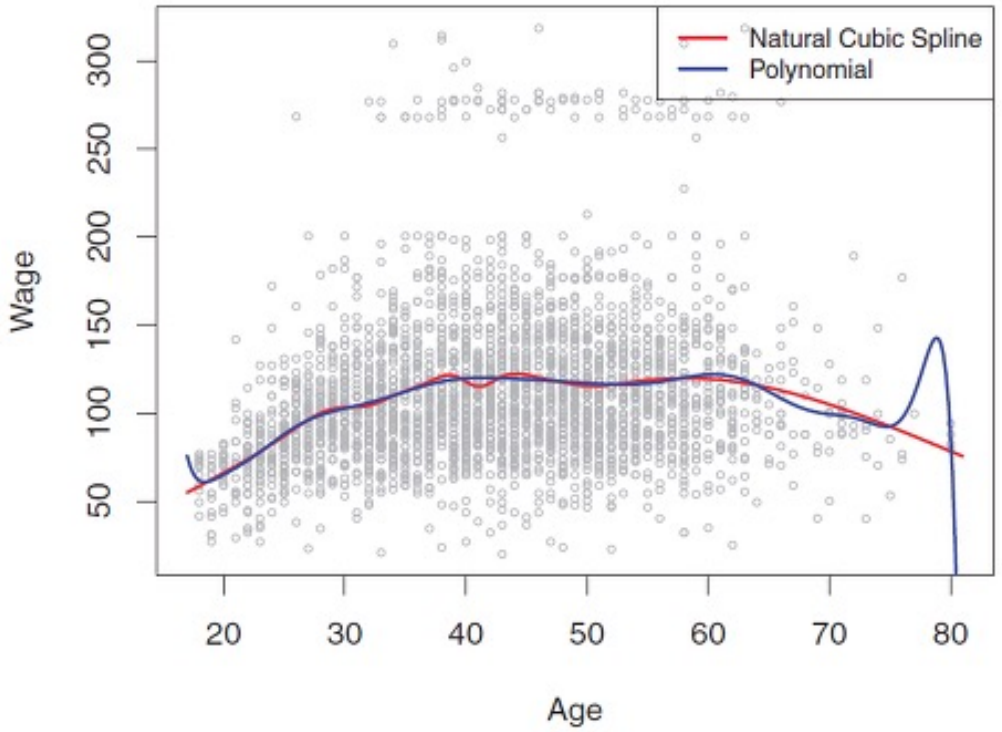
  : 고차 테일러 전개할 수 있도록 입력으로 다항식을 덧붙임

- $h_m(X) = log(X_j)$
- $h_m(X) = \sqrt{X_j}$

  : 비선형 변환

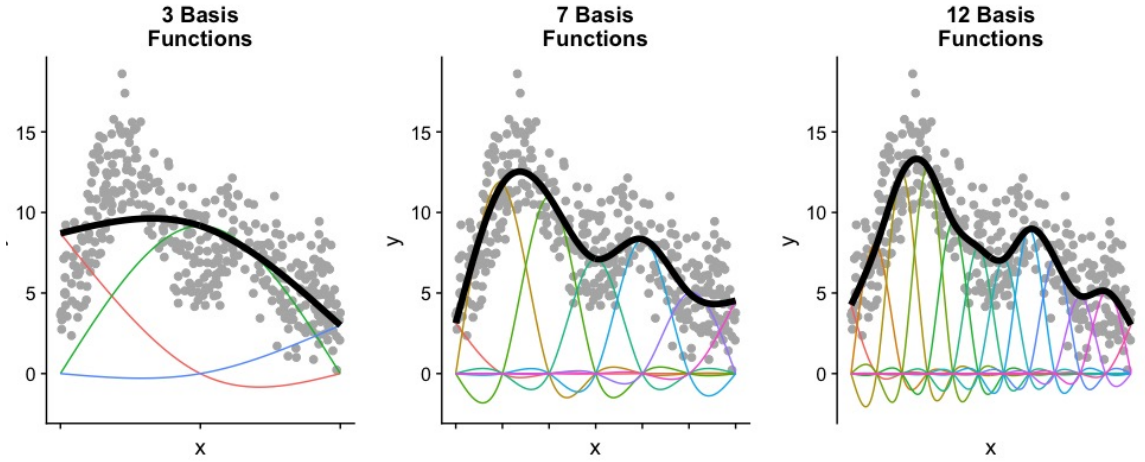
Polynomial Regression

비선형 데이터의 예측률을 높이는
비선형 모델을 만드려면 어떻게 해야 할까?



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$



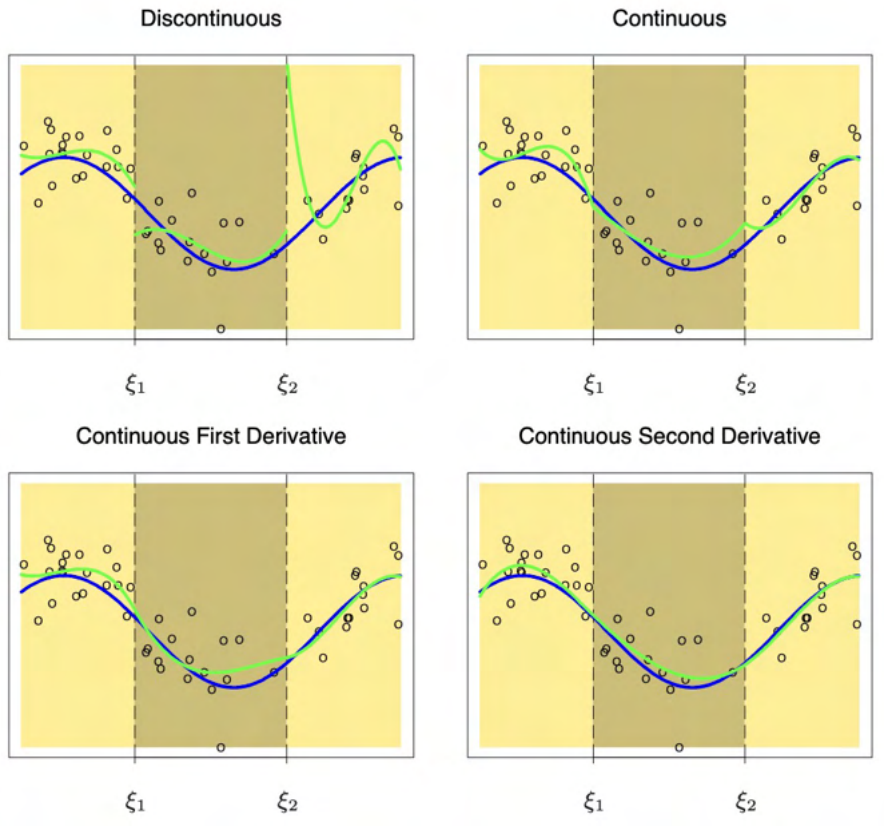
Spline

Knot : 매듭(점들)
Spline : 고정된 점들 사이를 잇는 방법
Piecewise polynomials : 고정된 점(knot) 사이를 부드럽게 연결

데이터의 문제에서는,
knot를 **어디에, 몇 개** 둬야 할까?

Degree-d Spline

각 knot에서 자기 자신과 도함수로부터 (d-1)계도함수까지
모두 연속인 piecewise degree-d polynomial



Cubic Spline

자기자신(3차), 도함수, 이계도함수까지
모두 연속인 piecewise degree-3 polynomial

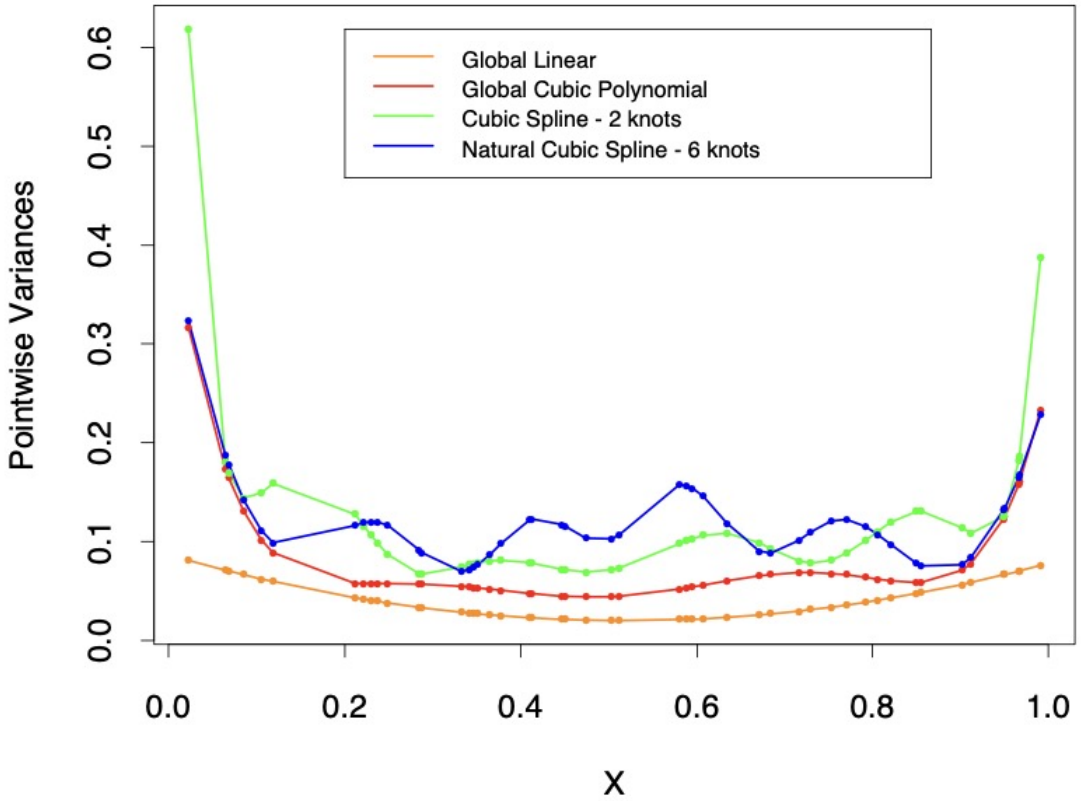
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

$$b_1(x_i) = x_i$$
$$b_2(x_i) = x_i^2$$
$$b_3(x_i) = x_i^3$$
$$b_{k+3}(x_i) = (x_i - \xi_k)_+^3 \ , \ k = 1, 2, \ldots , K$$

4+K 개의 predictors를 Least Squares로 적합
K : knot의 개수
4 : cubic function's unknowns
연속성 조건 -> knot 하나에 자유도 1 증가

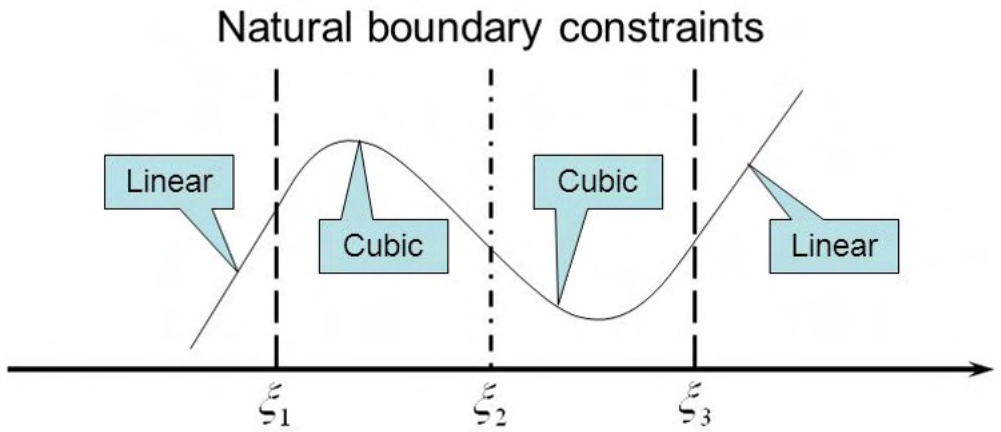
Natural Cubic Spline

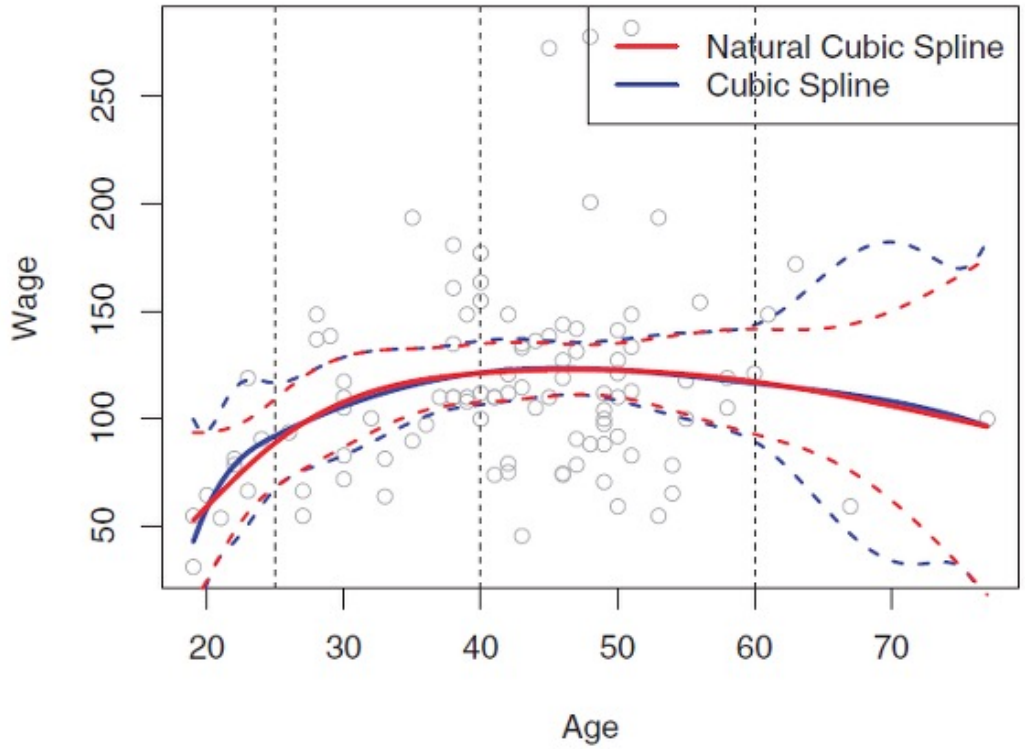
양 끝에서 분산이 튀는 문제는 어떻게 해결할까?



Natural Cubic Spline

Additional Boundary Constraints
: 가장 작은 knot보다 작거나, 가장 큰 knot보다 큰 부분에서는
함수가 직선이어야 한다는 조건 추가





Part 2

Smooth spline

Spline Method

Objective : find $\hat{f}$!

- Basis function -> linear vs non-linear
- Knot Selection -> Spline
- Bias – Variance trade-off

Smooth Spline

Smooth Spline

- **Avoids the knot selection problem**
  - → Completely using a maximal set of knots
- Penalty : smoothness of a function

Smooth Spline

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

Smooth Parameter

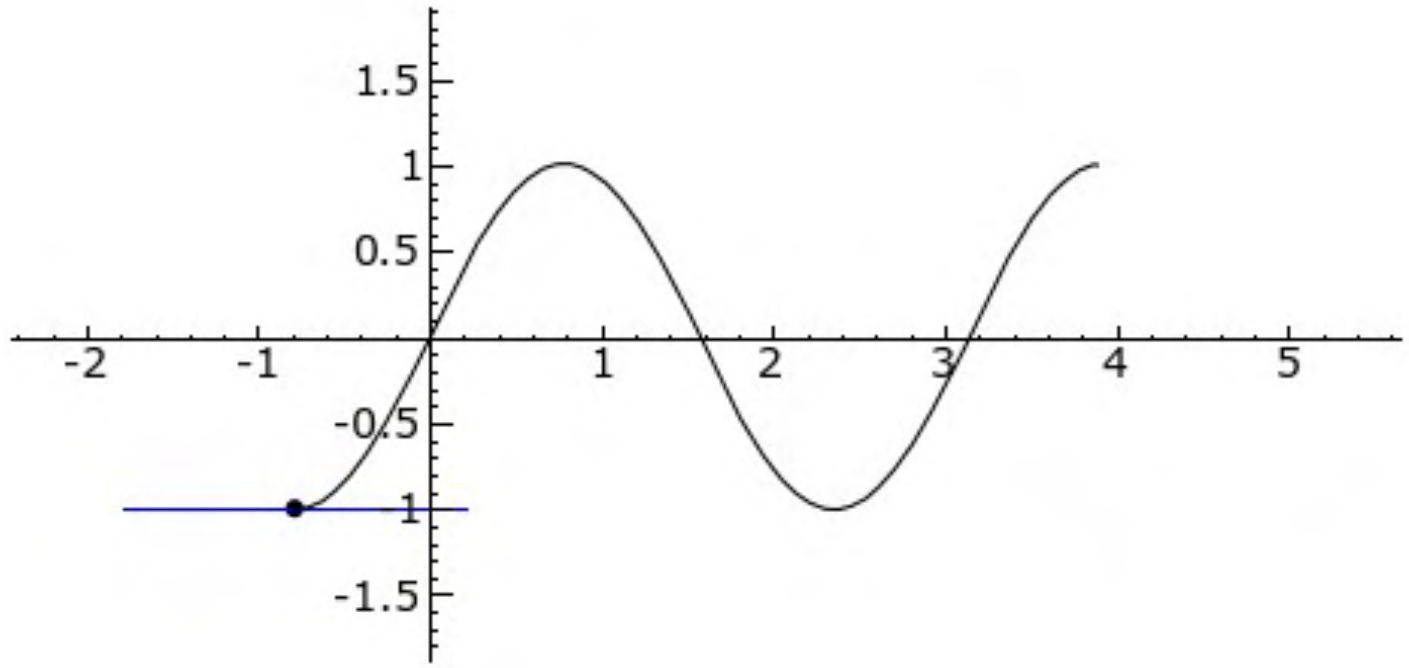
Measure of the smoothness

Second Derivative

이계도함수 f''

⇒ 함수의 오목성 측정!

⇒ 그래프의 곡률 &
볼록성 지표



$y = \sin(2x)$

Smoothing parameter

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

Smooth Parameter

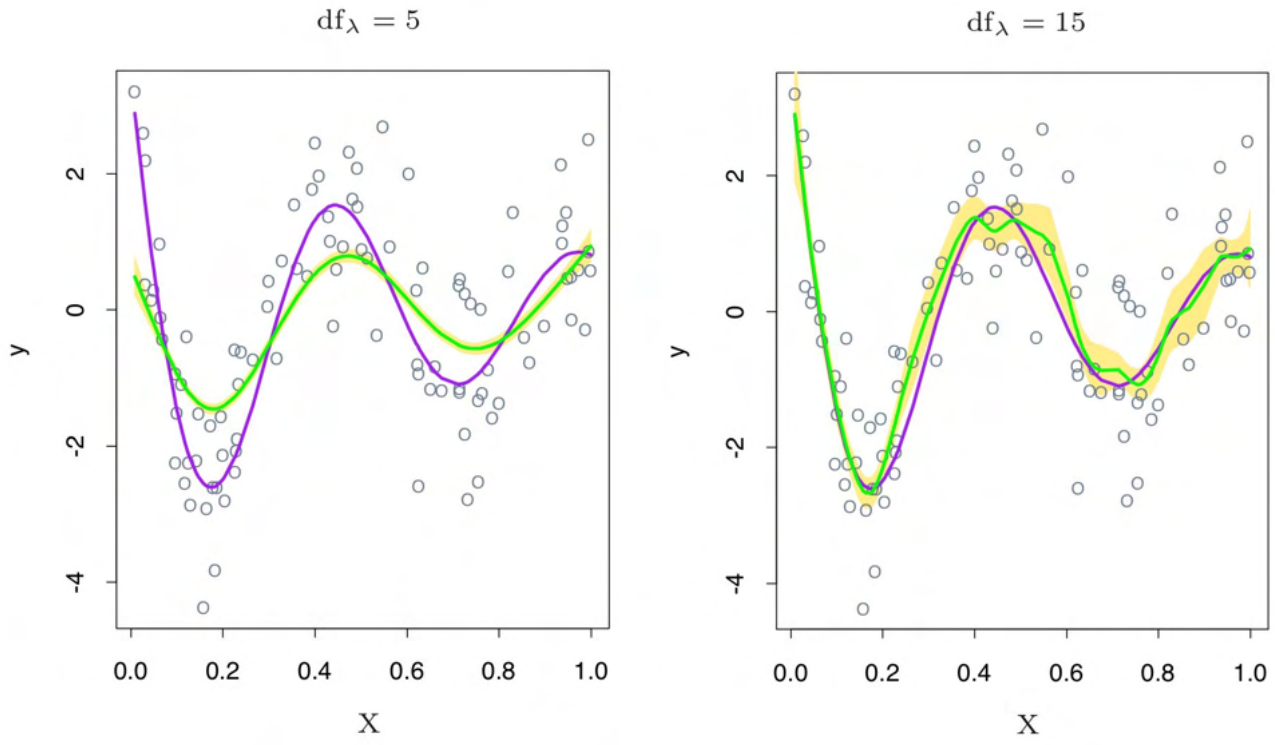
- λ -> 0 : no penalty : f = data interpolation
- λ -> ∞ : infinite penalty : f = simple least squares line fit

Effective degree of freedom

$$\hat{f} = S_\lambda y \quad : \quad df_\lambda = tr(S_y)$$

λ -> 0 : df -> N , S_λ -> I

λ -> ∞ : df -> 2, S_λ -> H



Smooth Spline solution

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{N}'\boldsymbol{N} + \boldsymbol{\lambda}\boldsymbol{\Omega}_{\boldsymbol{N}})^{-1}\boldsymbol{N}'\boldsymbol{y}$$

$$f(x) = \sum_{j}^{N} N_j(x)\theta_j$$

$$\{N\}_{ij} = N_j(x_i)$$

$$\{\Omega_N\}_{jk} = \int N_j''(t)N_k''(t)dt$$

Smooth Spline solution

$$f = N\theta \quad \Longrightarrow \quad RSS(\theta, \lambda) = (y - N\theta)'(y - N\theta) + \lambda\theta'\Omega_N\theta$$

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{N'N} + \boldsymbol{\lambda\Omega_N})^{-1}\boldsymbol{N'y}$$

$$\hat{f} = N\hat{\theta} = N(\boldsymbol{N'N} + \boldsymbol{\lambda\Omega_N})^{-1}\boldsymbol{N'y} := \boldsymbol{S_\lambda y}$$

Selection of degree of freedom

Integrated squared prediction error (EPE)

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

$$Bias(\hat{f}) = f - E(\hat{f}) = f - S_\lambda f$$

$$Cov(\hat{f}) = S_\lambda Cov(y) S_\lambda' = S_\lambda S_\lambda'$$

$$, when\ Y = f(X) + \epsilon,$$

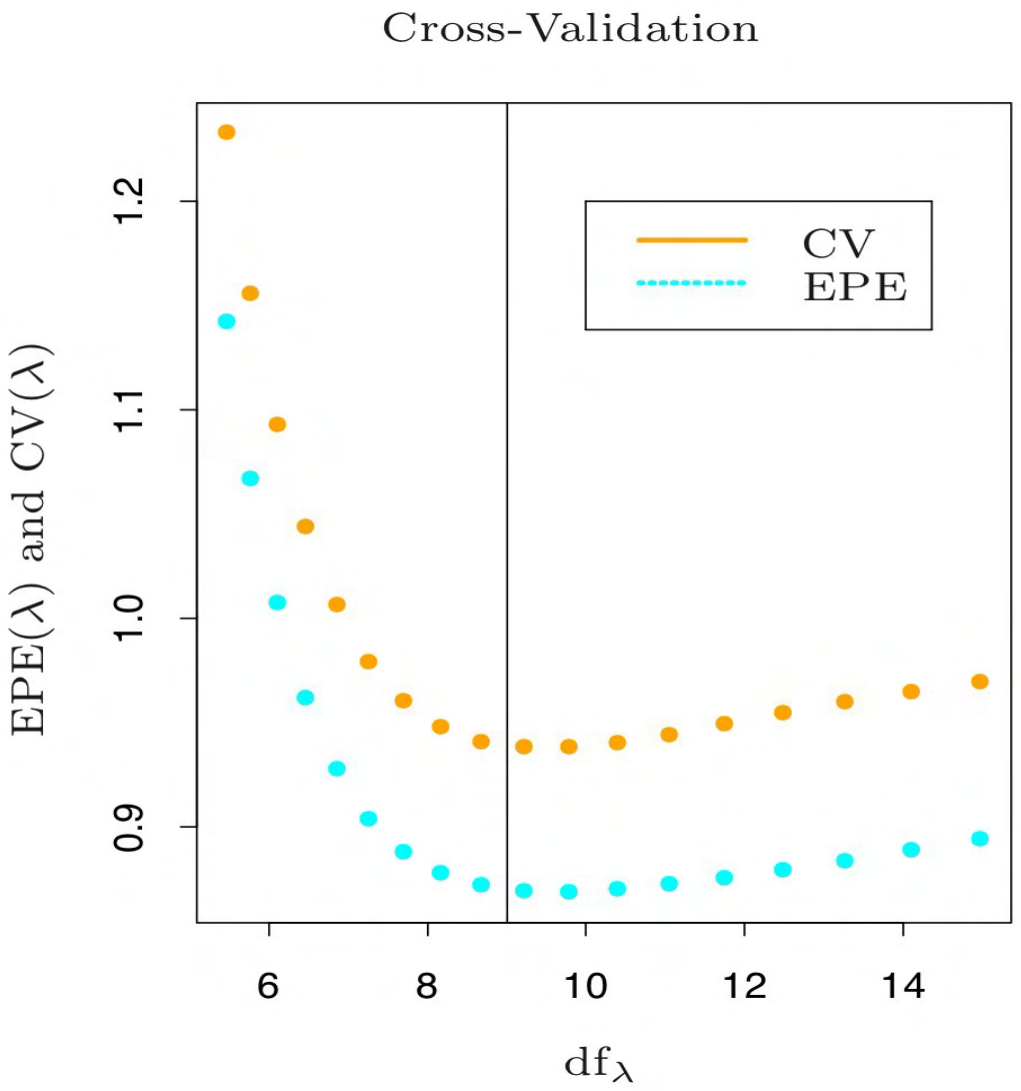
$$f(X) = \frac{\sin(12(X + 0.2))}{x + 0.2}$$

Selection of degree of freedom

Cross Validation

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

Selection of degree of freedom



Part 3

Reproducing Kernel Hilbert Space(RKHS)

Motivation

- Smooth Spline 문제의 solution
- SVM Kernel trick

Background. Hilbert Space

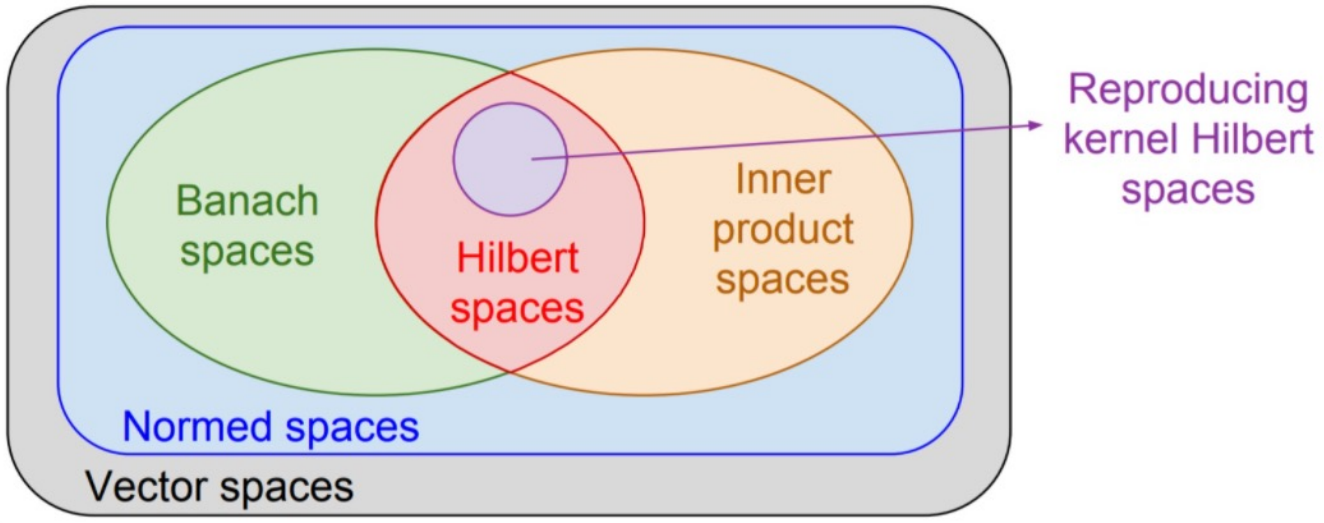
Hilbert Space = "Complete" "Inner Product" "linear"

Cauchy Sequence

$$For\ any\ \epsilon > 0, \exists\ n \in Z^{+}\ s.t.\ d(x_k, x_m) < \epsilon\ whenever\ n \geq k, m$$

Background. Hilbert Space

Hilbert Space = “Complete” “Inner Product” “linear”



$$\int_{\chi} f(x)d(x)dx = < f, g >_{\mathrm{H}}$$

$$\sqrt{< f, f >_{H}} = \left|\left| f \right|\right|_{H}$$

Background. Hilbert Space

Hilbert Space = “Complete” “Inner Product” “linear”

$$For\ all\ f, g \in H, f + g \in H$$

$$For\ all\ f, g \in H, fg \in H$$

General Optimization Problem

$$\boldsymbol{min_{f \in H}}\left[\sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{L}\left(\boldsymbol{y_i}, \boldsymbol{f}(\boldsymbol{x_i})\right) + \boldsymbol{\lambda J}(\boldsymbol{f})\right]$$

$$\boldsymbol{J}(\boldsymbol{f}) = \int \frac{\left|\widetilde{\boldsymbol{f}(\boldsymbol{s})}\right|^2}{\widetilde{\boldsymbol{G}(\boldsymbol{s})}} \boldsymbol{ds}$$

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

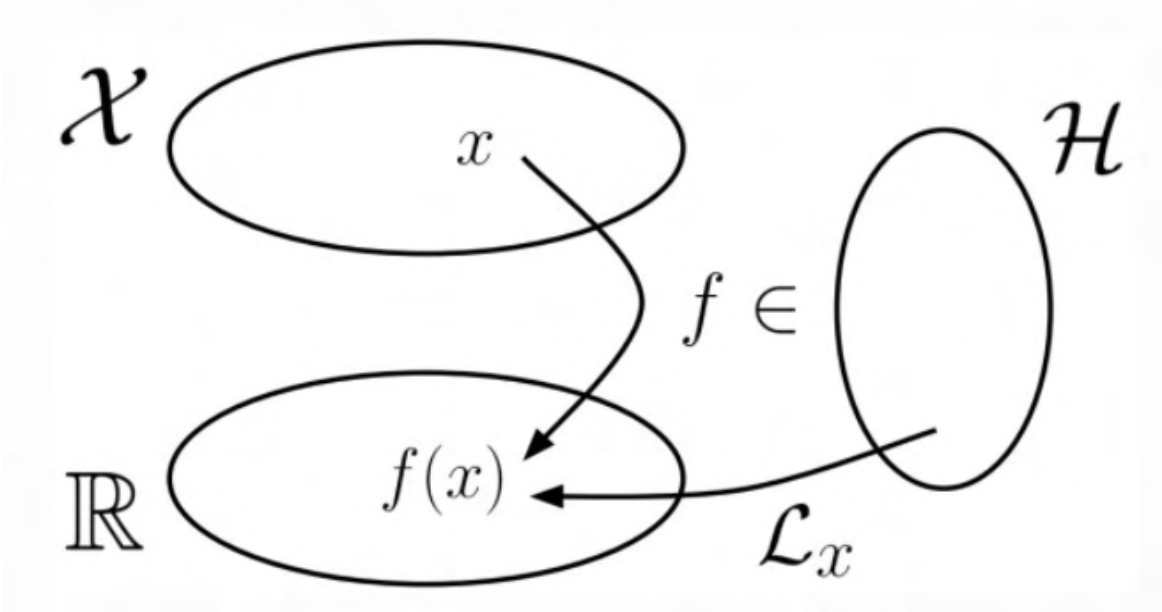
Reproducing Kernel Hilbert Space (RKHS)

Produce RKHS

- Feature map
- Integral as a linear operator

Reproducing Kernel Hilbert Space (RKHS) – Method 1

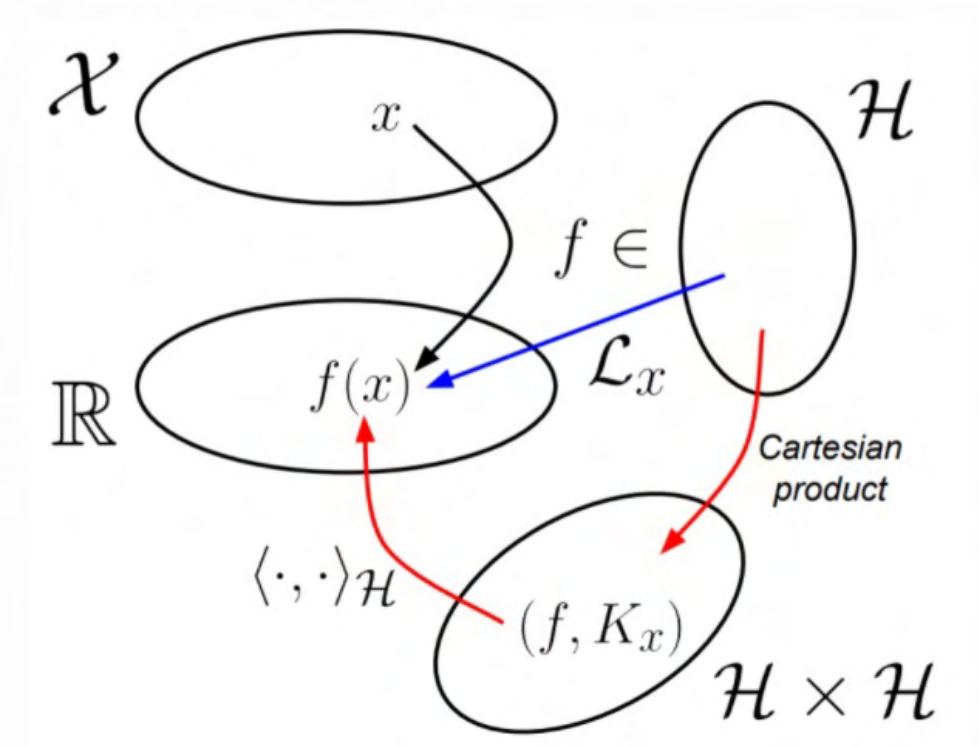
$$\textbf{Let } \chi \textbf{ be an arbitrary set, and } H \textbf{ a Hilbert space of all functions } f : \chi \to R.$$
$$\textbf{For each element } x \in \chi, \textbf{the evaluation map functional is a linear functional}$$
$$\textbf{that evaluates each } f \in H \textbf{ at the point } x, \textbf{written}$$
$$L_x : H \to R, \textbf{where } L_x(f) = f(x) \textbf{for all } f \in H$$
$$L_x \textbf{ is continuous at every } f \in H$$



Q. evaluation map?

Reproducing Kernel Hilbert Space (RKHS) – Method 1

$$\boldsymbol{If\ L_x\ is\ continuous\ at\ every\ f \in H, then, for\ each\ L_x,}$$
$$\textbf{there is a unique function } \mathbf{K(.,x)} \in \boldsymbol{H\ s.t.}\ \textbf{for every } \mathbf{f} \in \boldsymbol{H},$$
$$\boldsymbol{L_x(f) = f(x) = < f, K(.,x) >_H}$$
This equation is known as the reproducing property.



Reproducing Kernel Hilbert Space (RKHS) – Method 1

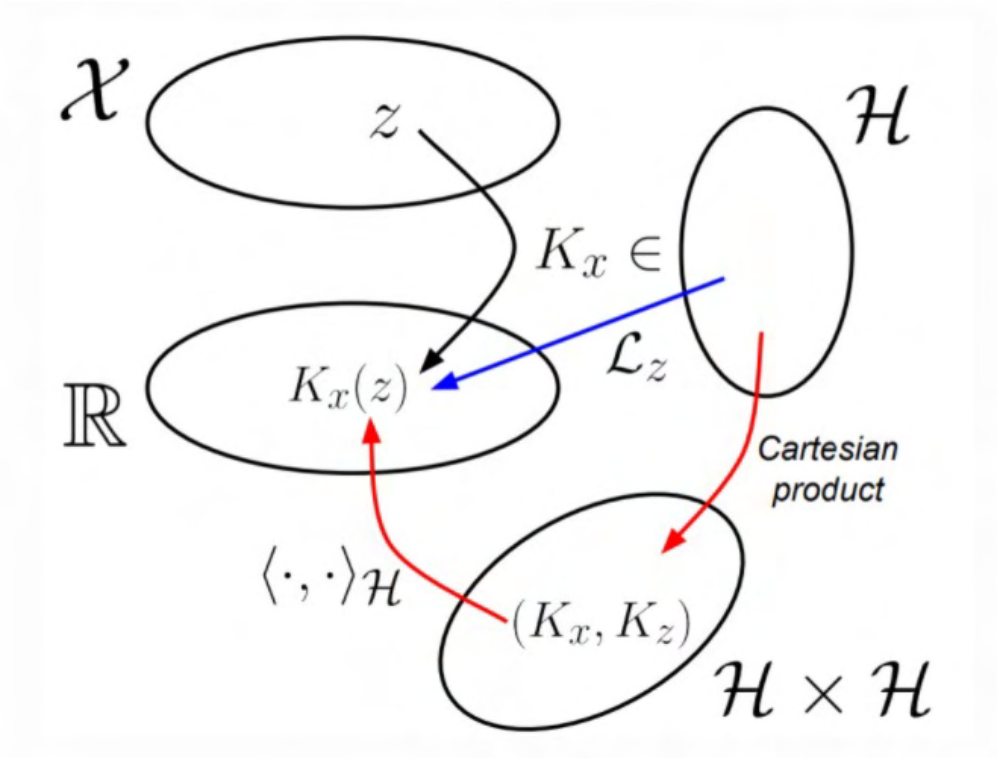
If L_x is continuous at every $f \in H$, then, for each L_x,
there is a unique function $K(.,x) \in H$ s.t. for every $f \in H$,
$L_x(f) = f(x) = < f, K(.,x) >_H$
This equation is known as the reproducing property.

$$L_x(f) = f(x) = < f, K(.,x) >_H$$

$$L_x(K(.,y)) = K(x,y) = < K(.,x), K(.,y) >_H$$

Reproducing Kernel Hilbert Space (RKHS) – Method 1

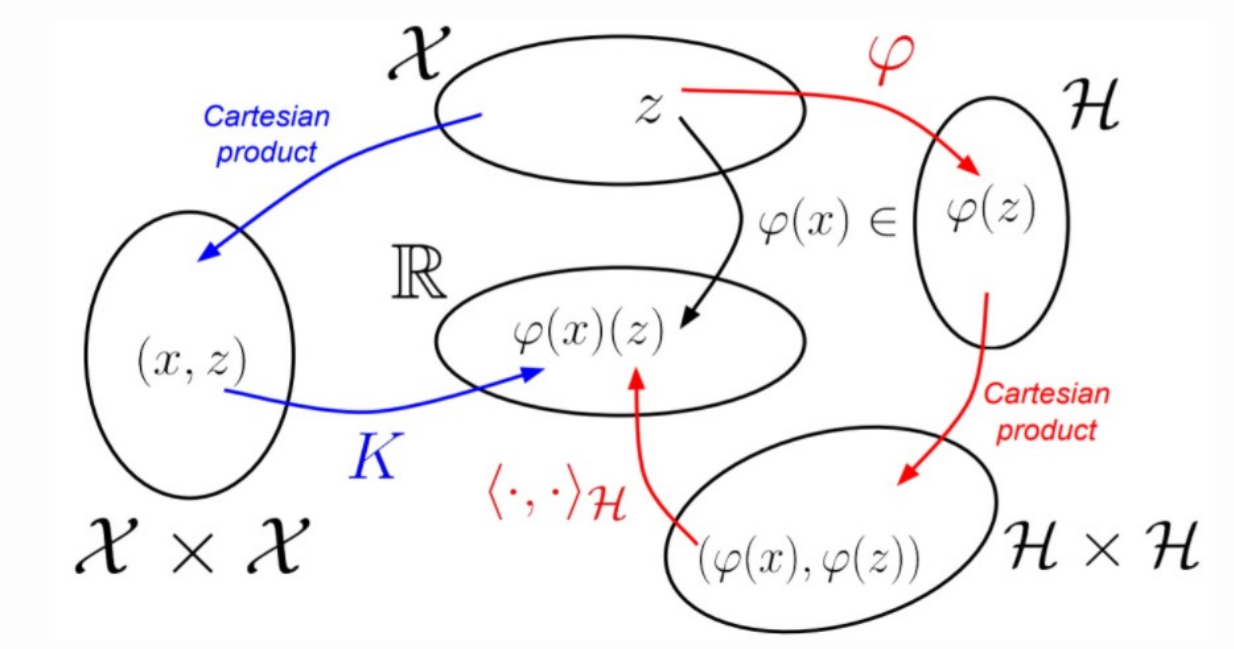
Let χ be an arbitrary set, and H a Hilbert space of all functions $f : \chi \to R$.
If, for all x ∈ χ, *the linear evaluation functional* L_x: H → R is continuous at every f ∈ *H*,
we can construct the reproducing kernel*, which is a bivariate function K: $\chi \times \chi \to R$ *defined by
$$K(x, y) = < K(., x), K(., y) >_H,$$
and the Hilbert space H is called a reproducing kernel Hilbert space (RKHS).



$$K(x, y) = < K(., x), K(., y) >_H$$

Reproducing Kernel Hilbert Space (RKHS) – Method 1

Given the RKHS H, we can find a function $\phi: \chi \to H$ that links up χ and H, and defined as $\phi(x) = K(.,x)$ for all $x \in \chi$.
In machine learning, ϕ is frequentily refered to as the 'feature map',
and H as the 'feature space'.



$$K(x,y) = <\phi(x), \phi(y)>_H$$

$$H := \overline{span}\{\phi(x) : x \in \chi\}$$
$$= \overline{span}\,\{K(x,.) : x \in \chi\}$$

Properties of Kernel

Symmetry (assumption)

$$K(x, y) = K(y, x) \;\; for\ all\ x, y \in \chi$$

Positive Semi-definite (from the reproducing property)

$$Gram\ matrix\ \{K\}_{ij}\ is\ positive\ semidefinite,$$
$$i.e.\ a'Ka \geq 0$$

$$\int \int K(x, y) dx dy \;\geq 0$$

Reproducing Kernel Hilbert Space (RKHS) – Method 2

$$(\boldsymbol{Mercer's\ theoerm})$$

$$\boldsymbol{Fix\ a\ symmetric\ kernel\ function\ K\ on\ a\ compact\ set\ \chi}$$

$$\boldsymbol{K : \chi^2 \to R}$$

$$\boldsymbol{T_k : L_2(\chi) \to L_2(\chi)}$$

$$\boldsymbol{T_k f(x) = \int K(x,y) f(y) dy}$$

$$\boldsymbol{The\ integral\ operator\ T\ is\ characterized\ by\ K.}$$

Reproducing Kernel Hilbert Space (RKHS) – Method 2

$(\boldsymbol{Mercer's\ Theoerm})$

$\boldsymbol{If\ K\ is\ continuous\ and\ T_k\ is\ positive\ semidefinite,}$
$\boldsymbol{then\ T\ has\ eigenfunction\ \psi_i, with\ eigenvalue\ \lambda_i\ s.t.}$

$$\boldsymbol{K(u,v) = \sum_{j}^{\infty} \lambda_j \psi_j(u)\psi_j(v)}$$

$\boldsymbol{That\ series\ converges\ uniformly.}$

$$\psi_i \in L_2(\chi), \qquad \lambda \geq 0, \qquad \left|\left|\psi_i\right|\right|_{L^2} = 1$$

$$(T_K \boldsymbol{\psi_a})(\boldsymbol{x_i}) = \boldsymbol{\lambda_a \psi_a(x_i)}$$

Reproducing Kernel Hilbert Space (RKHS) – Method 2

$$(\boldsymbol{Mercer's\ Theoerm})$$

$$\boldsymbol{For\ \chi \in R^d\ compact\ and\ K: \chi^2 \to R\ continuous\ and\ symmetric,\ the\ followings\ are\ equivalent:}$$

$$\boldsymbol{1.\ Every\ Gram\ matrix\ is\ positive\ semidefinite.}$$

$$\boldsymbol{2.\ The\ integral\ operator\ T_k\ is\ positive\ semidefinite.}$$

$$\boldsymbol{3.\ We\ can\ express\ K\ as}$$
$$\boldsymbol{K(u,v) := \Sigma \lambda_i \psi_i(u)\psi_i(v) for\ fixed\ \lambda_i \geq 0\ and\ orthonormal\ \psi_i: \chi \to R}$$

$$\boldsymbol{4.\ K\ is\ the\ reproducing\ kernel\ of\ an\ RKHS\ on\ \chi}$$

Reproducing Kernel Hilbert Space (RKHS) – Method 2

$$By\ the\ reproducing\ property, < f, K(.,x) >_H = f(x)$$

$$Thus, < \psi_j, \psi_j' >_H = \begin{cases} \frac{1}{\lambda_j} & if\ j = j' \\ 0 & otherwise \end{cases}$$

$$\theta_j = < f, \psi_j >, thus\ f = \sum_j^{\infty} \theta_j \psi_j \quad \Rightarrow \quad ||f||_H^2 = < f, f >_H = \sum_j^{\infty} \frac{\theta_j^2}{\lambda_j}$$

$$(\boldsymbol{Representer\ Theoerm})$$

$$\min Q_\lambda(f) = \sum_i^n \left(y_i - f(x_i)\right)^2 + \lambda ||f||_H^2 \quad \Rightarrow \quad solution\ \hat{f}(x) = \sum_i^n \alpha_i K(x, x_i)$$

Non-uniqueness of basis function

$$\mathbf{K}(\mathbf{x},\mathbf{y}) = <\mathbf{K}(.,\mathbf{x}),\mathbf{K}(.,\mathbf{y})>_{\mathbf{H}}$$

$$\boldsymbol{\Phi}_{\mathbf{1}}(\mathbf{x}) = \mathbf{K}(.,\mathbf{x})$$

$$\boldsymbol{\Phi}_{\mathbf{2}}(\mathbf{x}) : \mathbf{x} \rightarrow (\sqrt{\boldsymbol{\lambda}_{\mathbf{1}}}\psi_{\mathbf{1}}(\mathbf{x}), \sqrt{\boldsymbol{\lambda}_{\mathbf{2}}}\psi_{\mathbf{2}}(\mathbf{x}))$$

$$K(x,y) = <\Phi_1(x),\Phi_1(y)>_H = <\Phi_2(x),\Phi_2(y)>_H$$

But… Kernel ⇔ RKHS : one-to-one by the theorem not in this ppt

Meaning of RKHS

- 무한 차원의 문제 => 유한 차원의 solution!

- X : input space -> 모집단

  $\{x_1, x_2, \ldots, x_N\}$: sample, or training data

  모집단 전체를 고려하는 최적화 문제
  => sample data만을 가지고 solution 도출 가능

Bayesian interpretation : RKHS

A function f = a realization of a zero-mean stationary Gaussian process
With prior covariance function K

$$H = H_0 \oplus H_1$$

The direct sum of null space & orthogonal space

Back to Smooth Spline

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

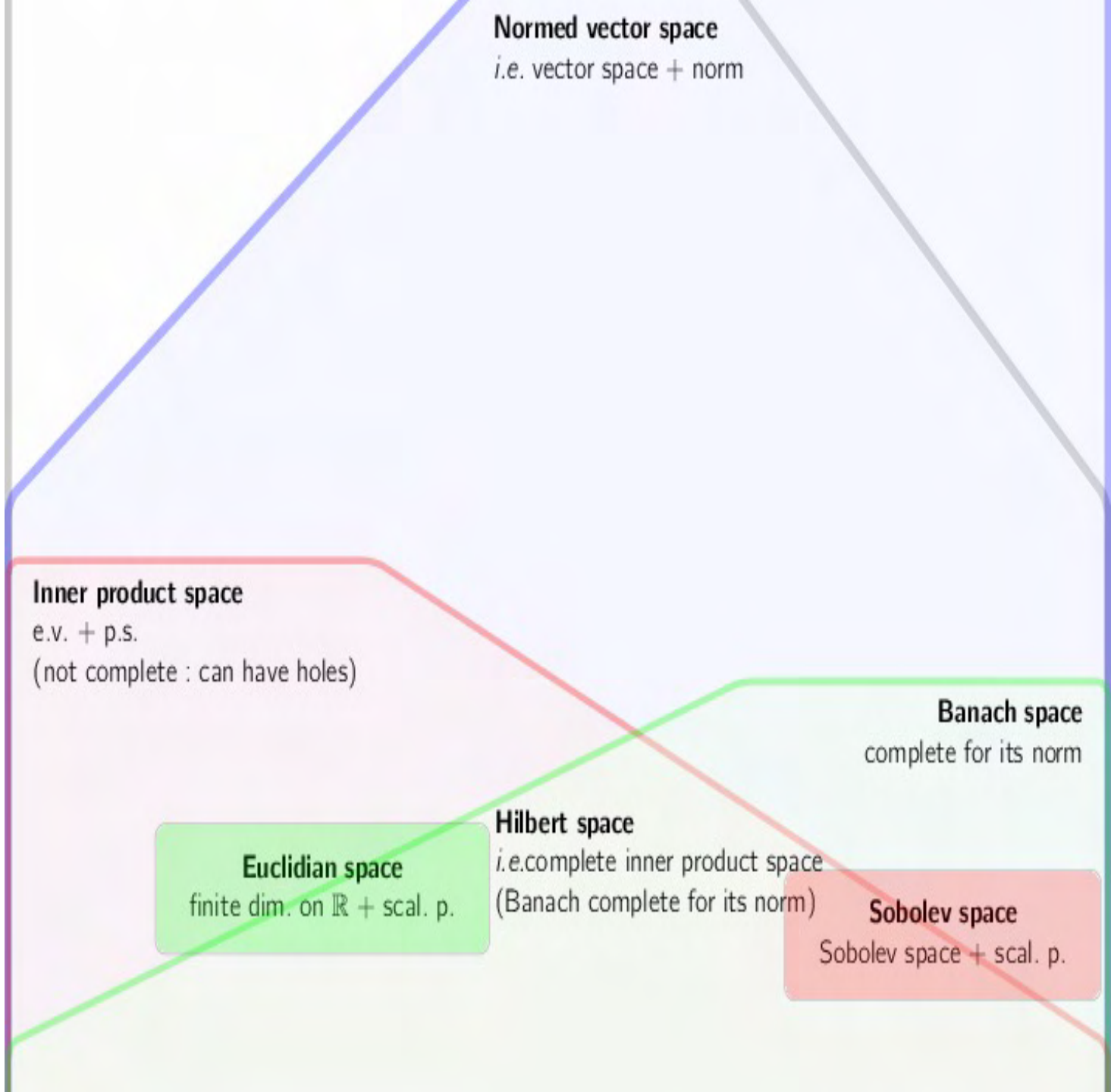
$$f(x) = \sum_{j}^{N} N_j(x)\theta_j$$

In “Sobolev Space”

Back to Smooth Spline

$$Banach\ Space \supset Sobolev\ Space$$

Sobolev Space
vs Hilbert Space



Sobolev Space

$$W^{m,p}(\Omega) := \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega) \; for \; 0 \leq |\alpha| \leq m\}$$

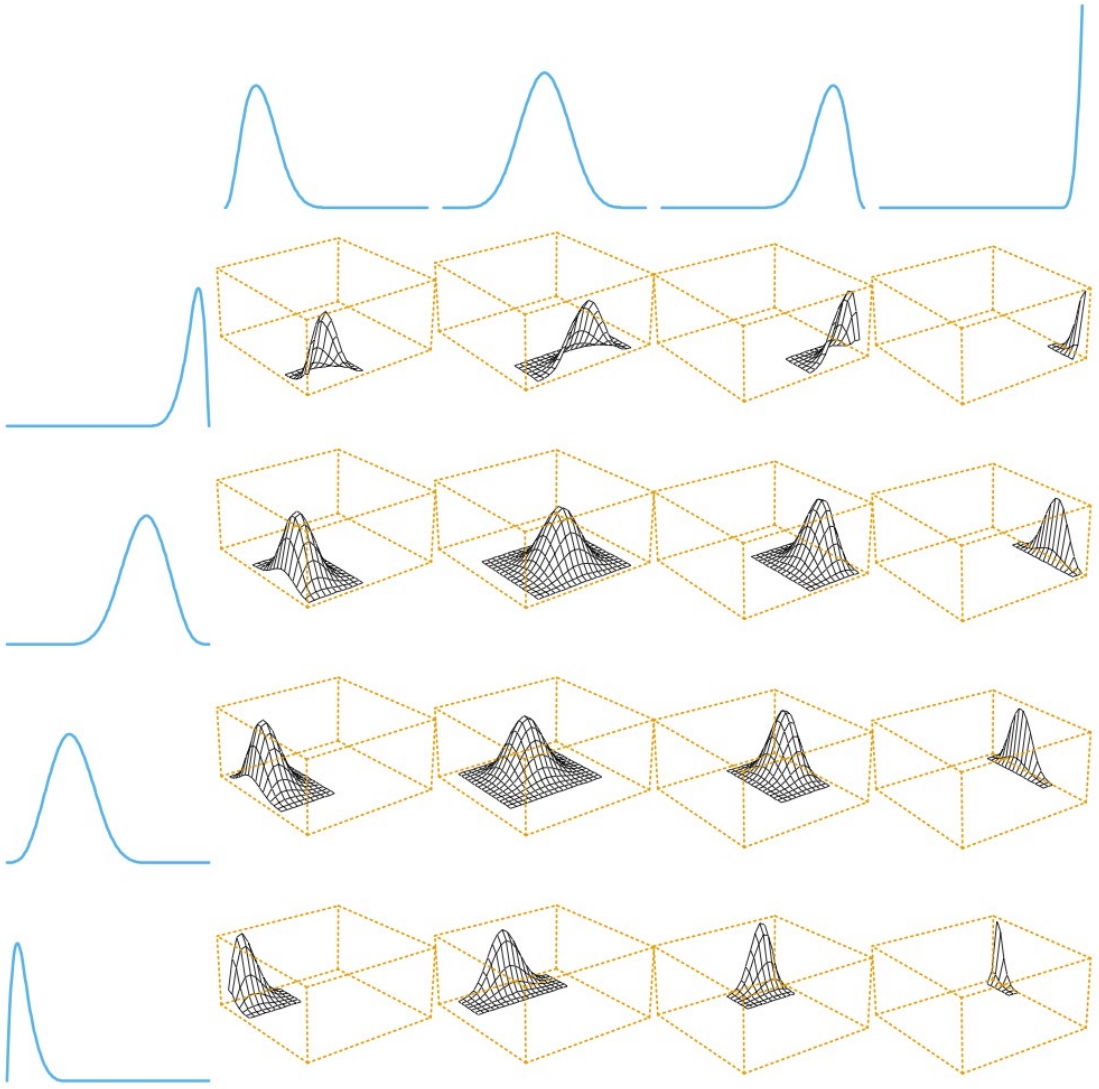
함수의 n차 미분
smoothness 가정!

When p = 2, W = H (Hilbert Space)

Multivariate smooth spline – Tensor product basis

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

$$,where\ h_{1k}(X_1) is\ a\ basis\ function\ of\ X_1,$$
$$h_{2k}(X_2) is\ a\ basis\ function\ of\ X_2$$



Multivariate smooth spline

$$\boldsymbol{min_f Q_\lambda(f)} = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{\left(y_i - f(x_i)\right)^2} + \boldsymbol{\lambda J(f)}$$

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

$$f(x) = \beta_0 + \beta' x + \sum_j^N \alpha_j h_j(x)$$

$$h_j(x) = \left|\left|x - x_j\right|\right|^2 \log||x - x_j|| \quad : \quad \text{Radial Basis}$$

Part 4

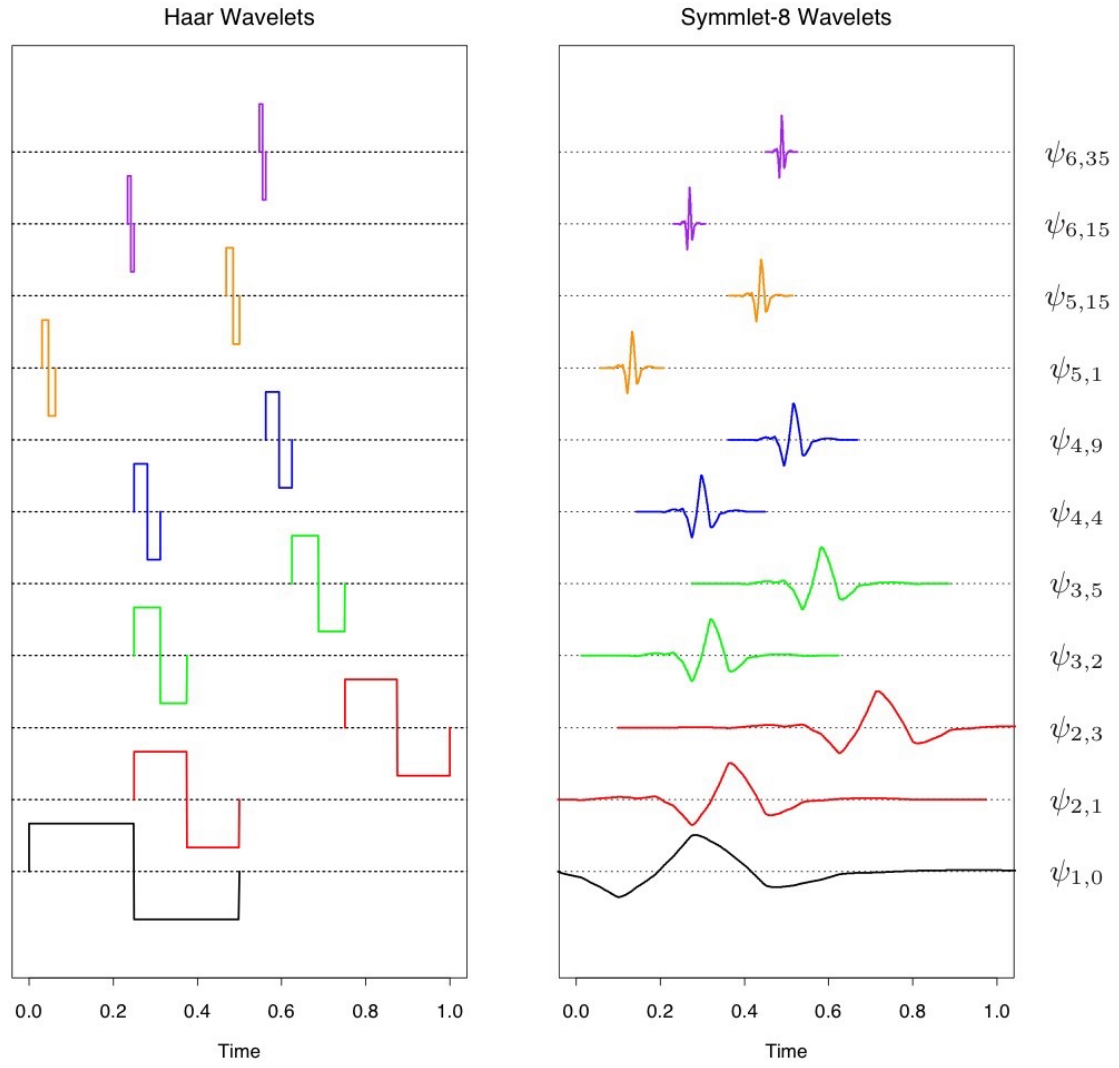
Wavelet

Spline

Wavelet Smoothing

"Signal Processing"

- Wavelet Bases
- Wavelet smoothing

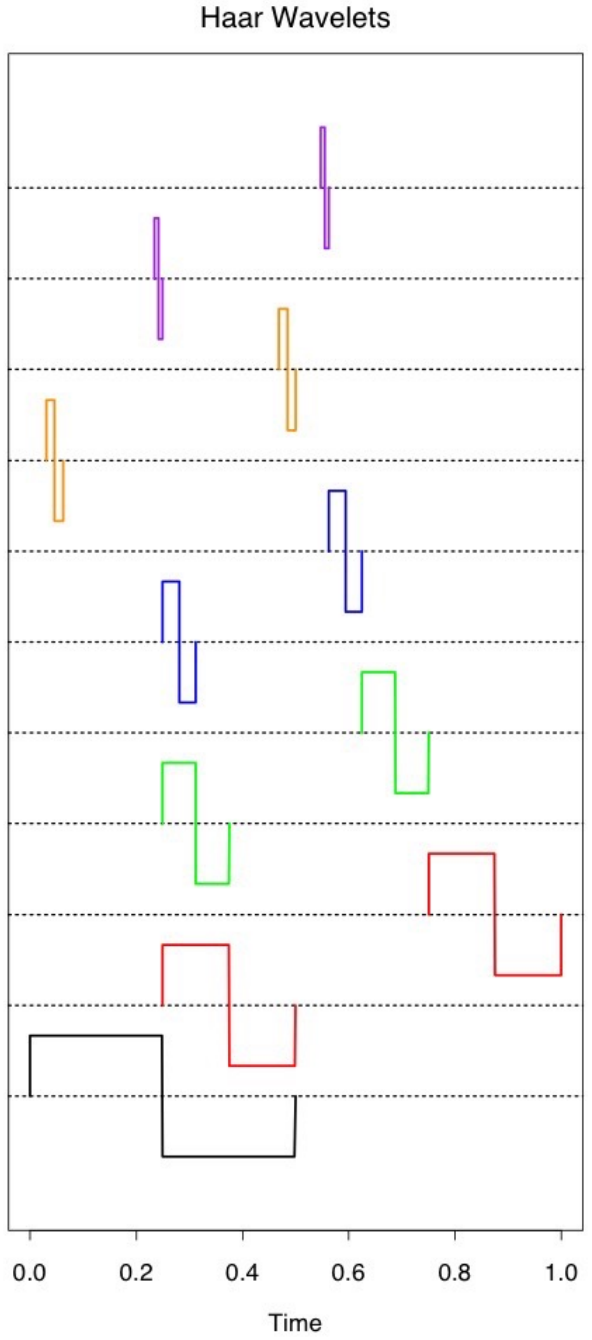


Wavelet Bases

$$Let\ \phi(x) = I(x \in [0,1]).$$
$$\phi_{0,k} = \phi(x-k)$$
$$: orthonormal\ basis\ of\ reference\ space\ V_0$$

$$The\ dilations\ \phi_{1,k}(x)$$
$$= \sqrt{2}\,\phi(2x-k): orthonormal\ basis\ of\ V_1\ \supset V_0$$

$$\ldots \supset V_1 \supset V_0 \supset V_{-1} \supset \ \ldots$$
$$each\ V_i = span\{\phi_{i,k} 2^{\frac{i}{2}} \phi(2^i x - k)\}$$



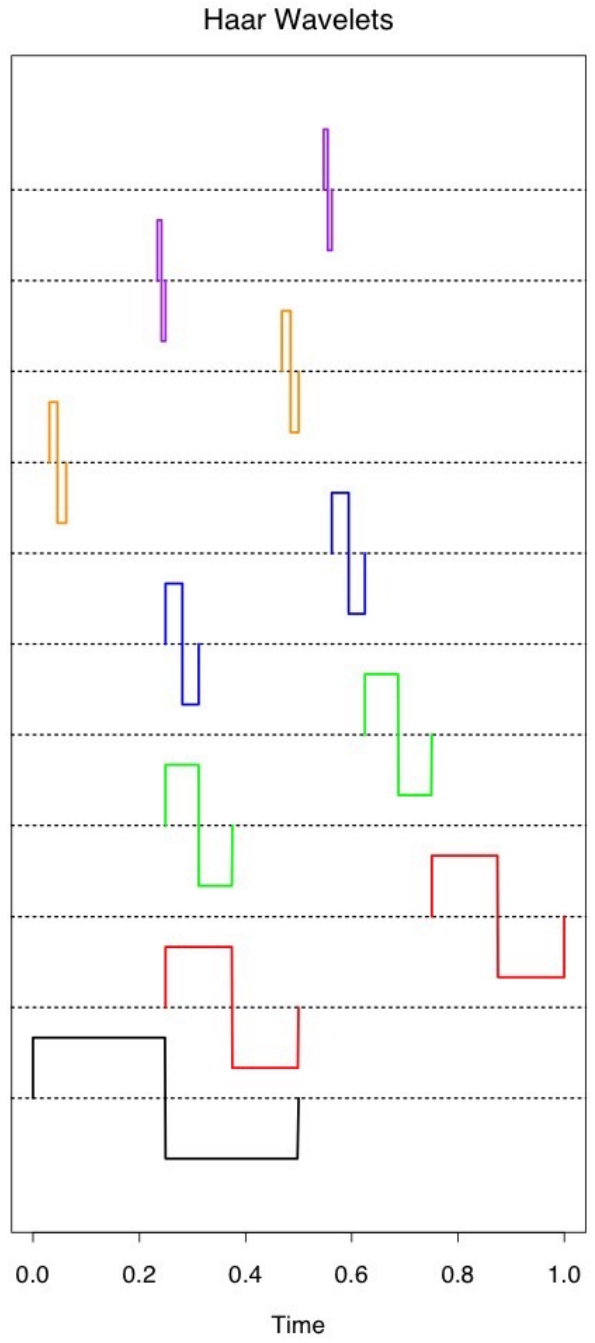
Wavelet Bases

$$\psi(x) = \phi(2x) - \phi(2x-1)$$
$$: an\ orthonormal\ basis\ of\ W_0$$

$$inductively : \psi_{j,k}$$
$$= 2^{\frac{j}{2}}\psi\left(2^j x - k\right) : a\ basis\ for\ W_j$$

W : “Detail”

$$V_J = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{J-1}$$



Adaptive Wavelet Filtering

Wavelet Transform

$$y^* = W'y \ , where\ W\ is\ the\ N * N\ orthonormal\ basis$$

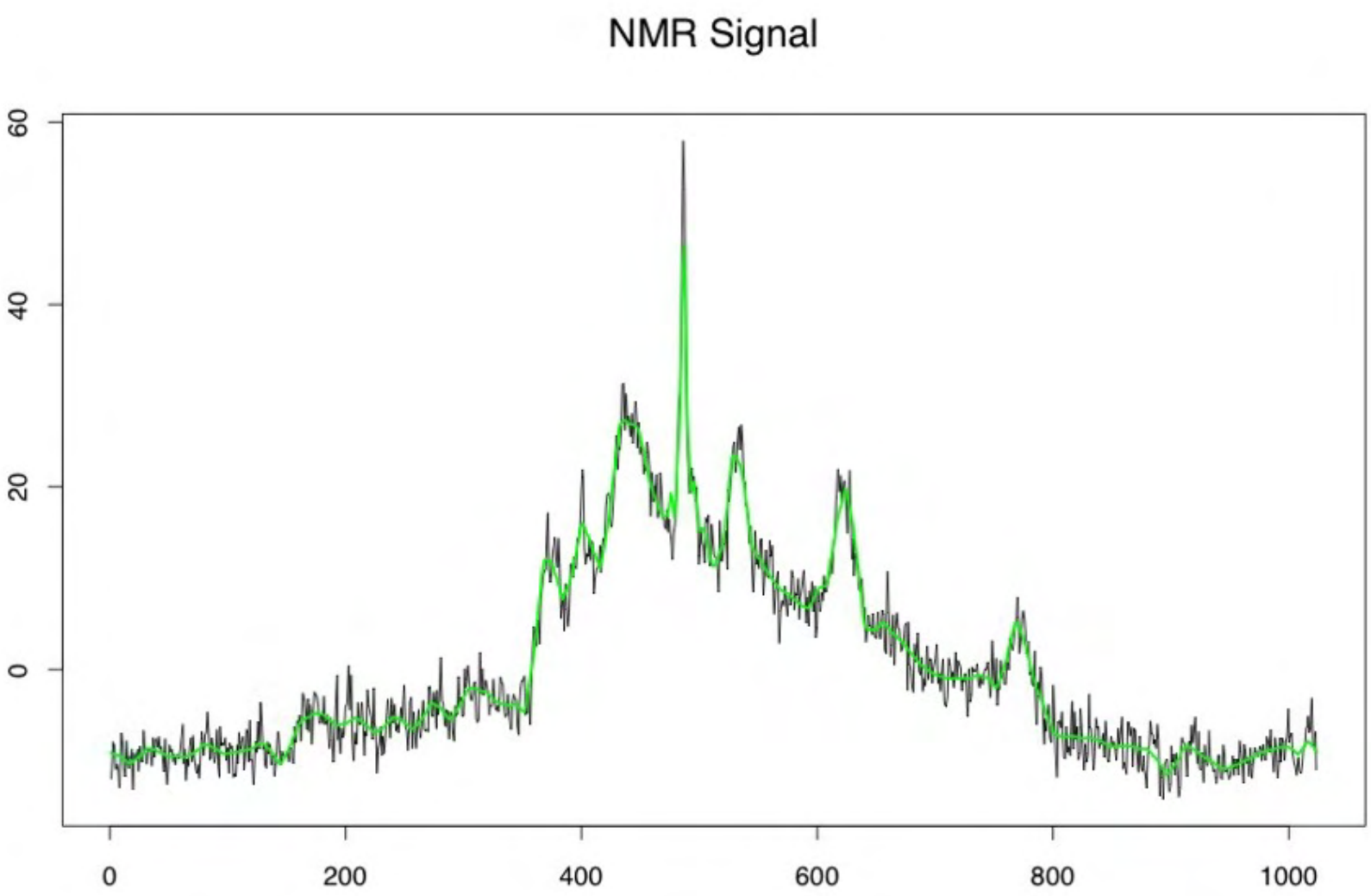
Inverse Wavelet Transform $\hat{f} = W\hat{\theta}$

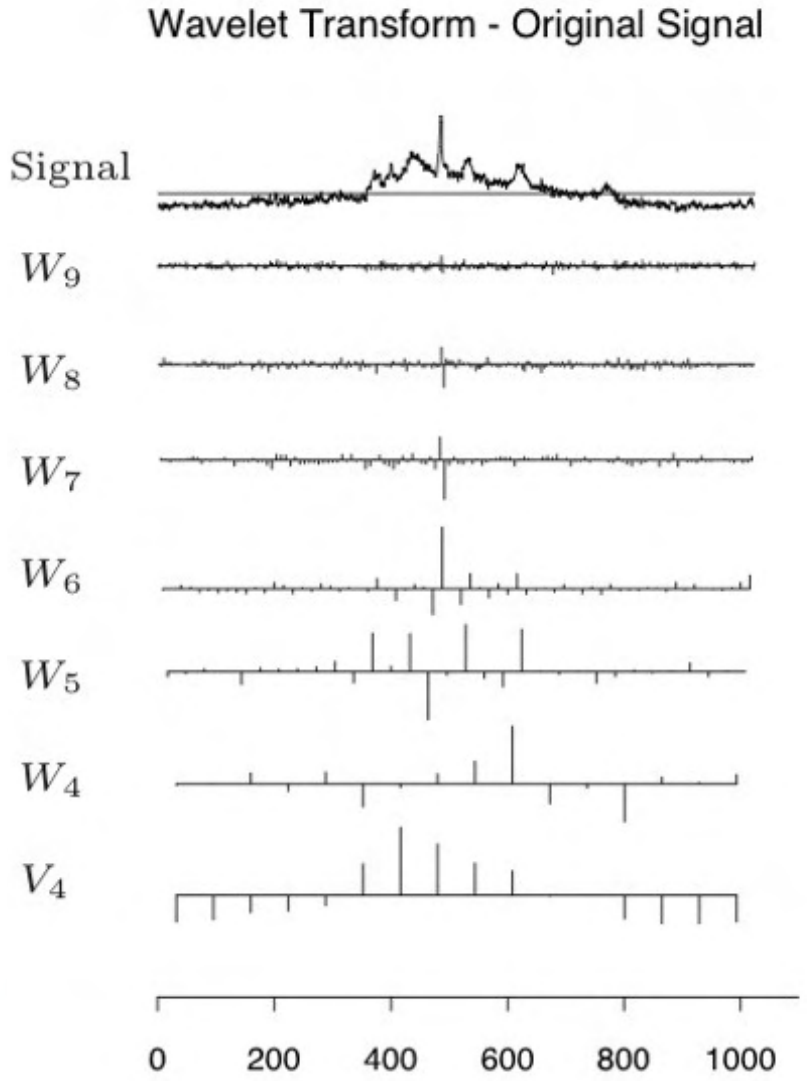
$\lambda = \sigma\sqrt{2logN}$: $Simple$

SURE shrinkage : Stein Unbiased Risk Estimation

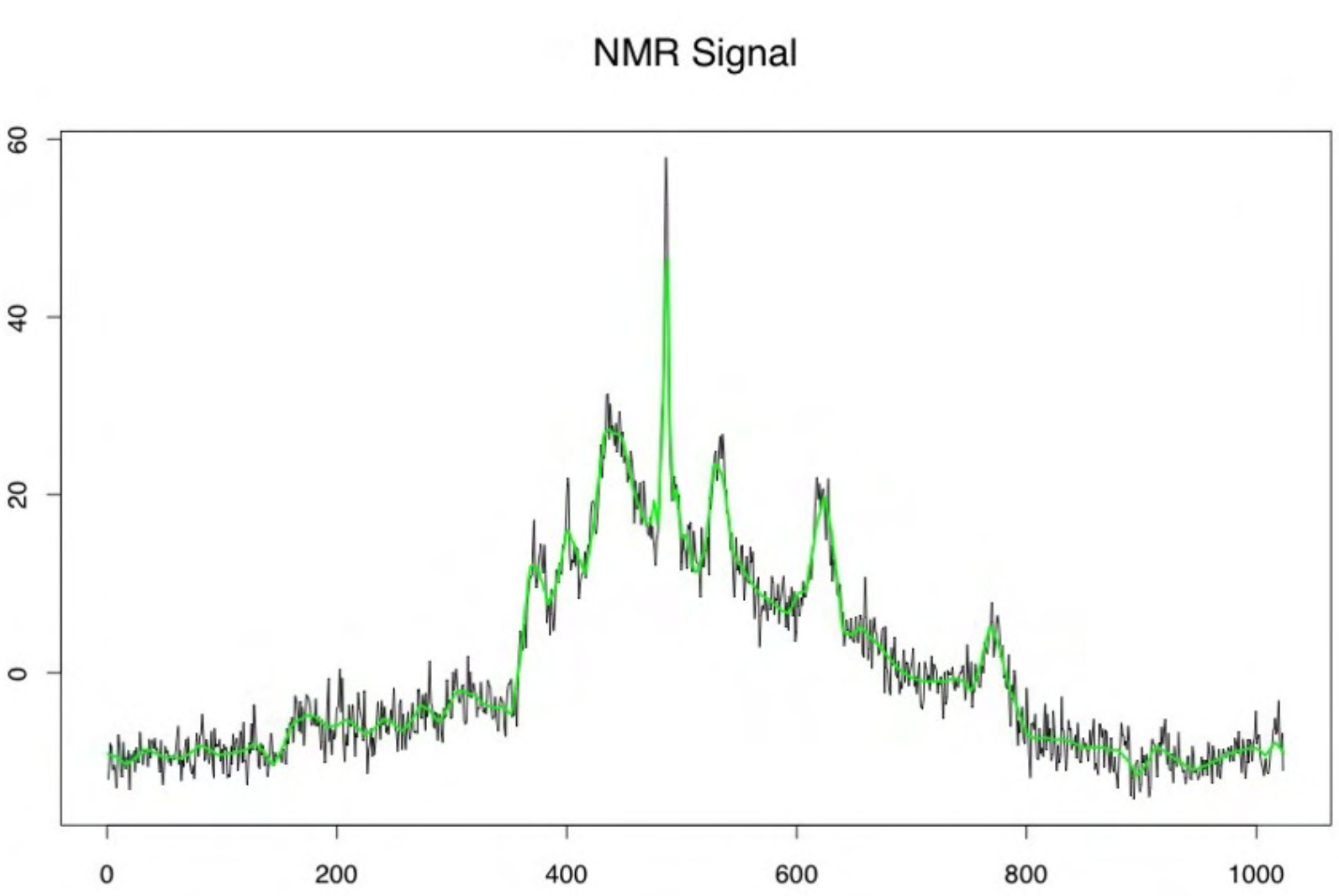
$$\min_{\theta} ||y - W\theta||_2^2 + 2\lambda||\theta||_1 \longrightarrow \widehat{\theta_j} = sign(y_j^*)(|y_j^*| - \lambda)_+$$

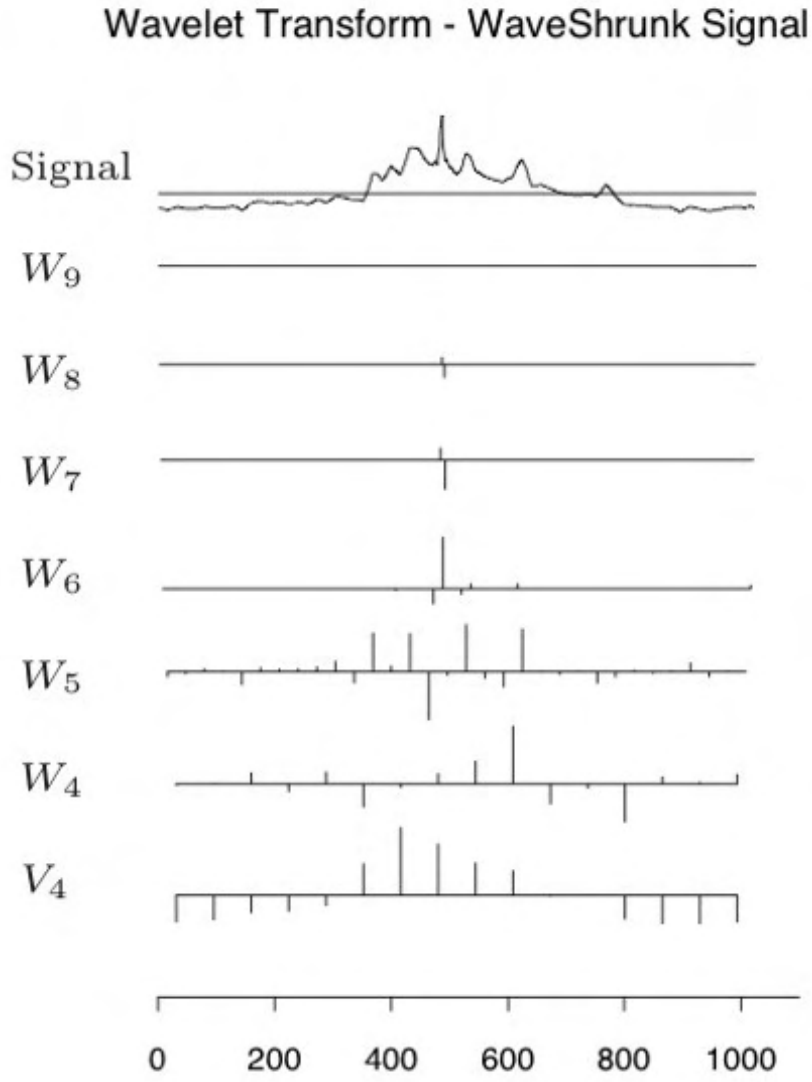
Wavelet Smoothing



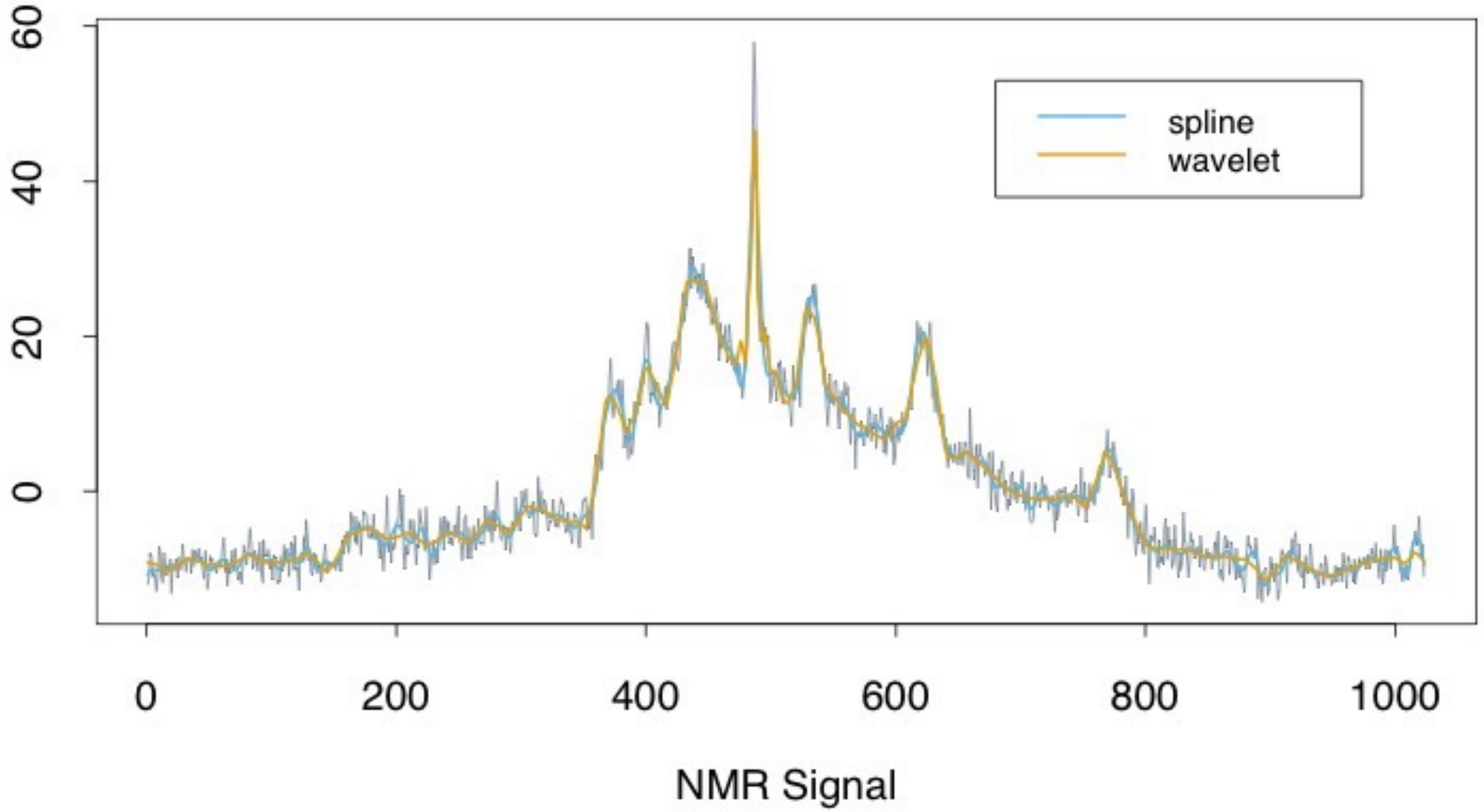


Wavelet Smoothing

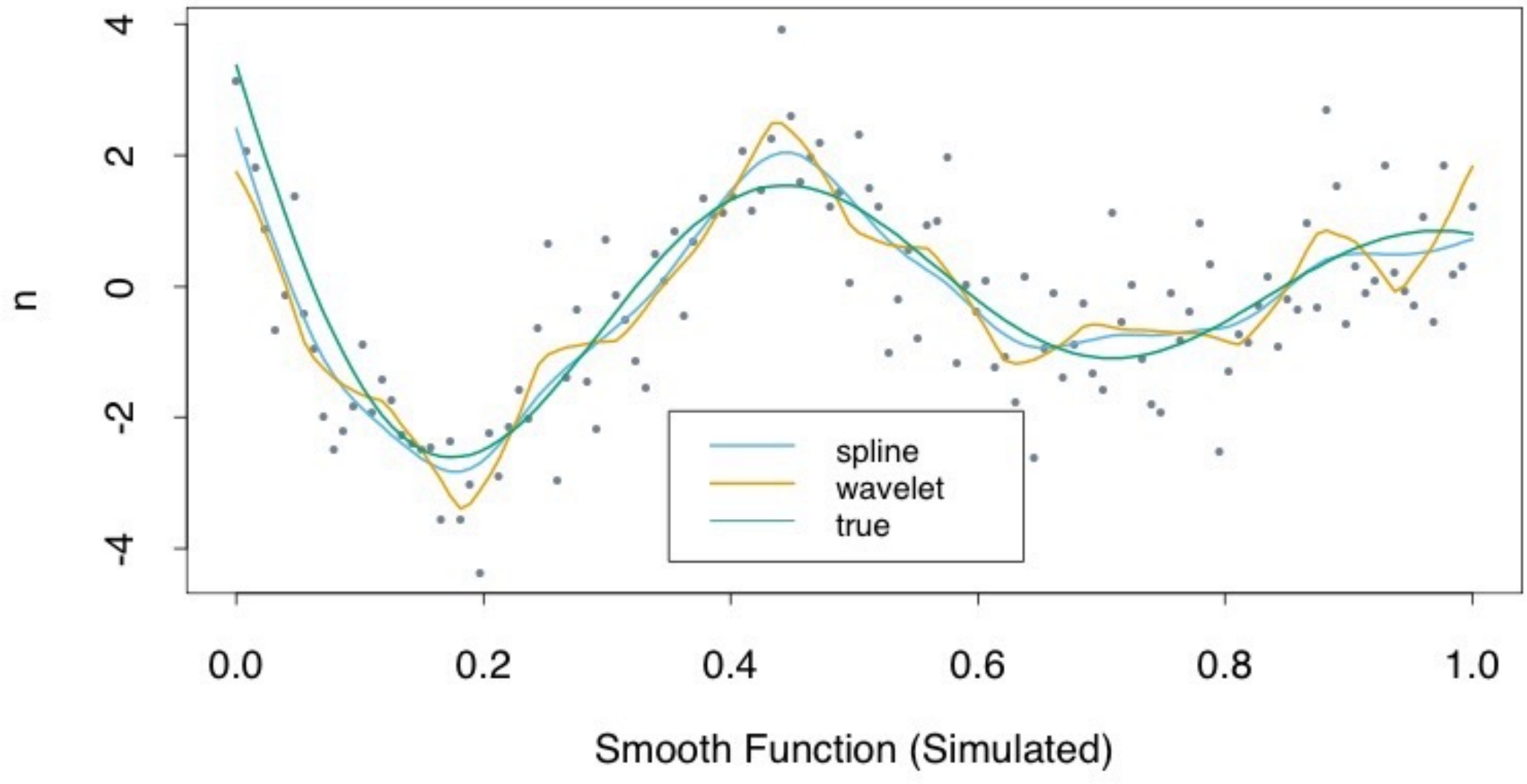




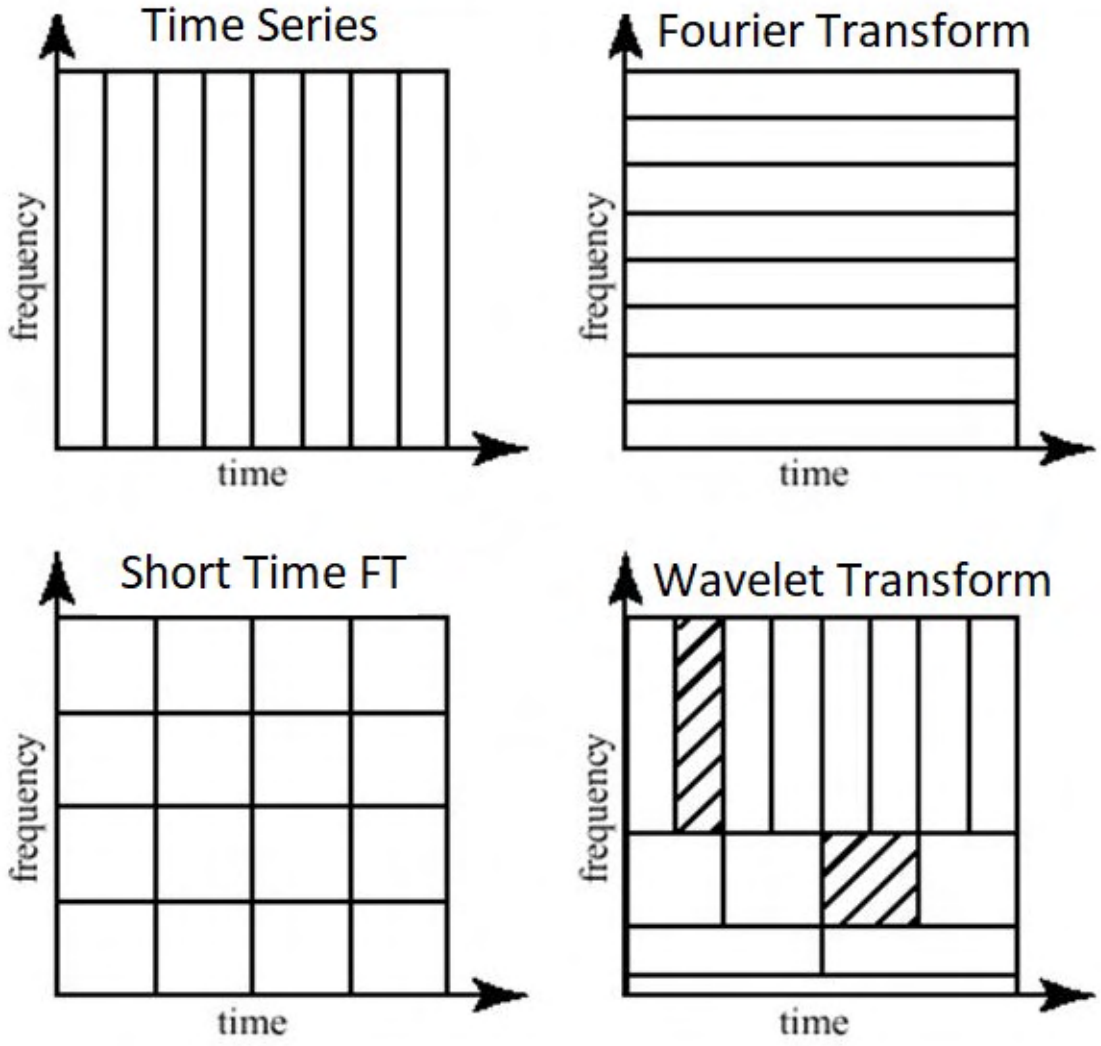
Comparison : **Smooth Spline** vs **Wavelet Spline**



Comparison : **Smooth Spline** vs **Wavelet Spline**



Comparison : **FFT** vs **Wavelet Spline**



Time Complexity Difference

FFT : O(NlogN)

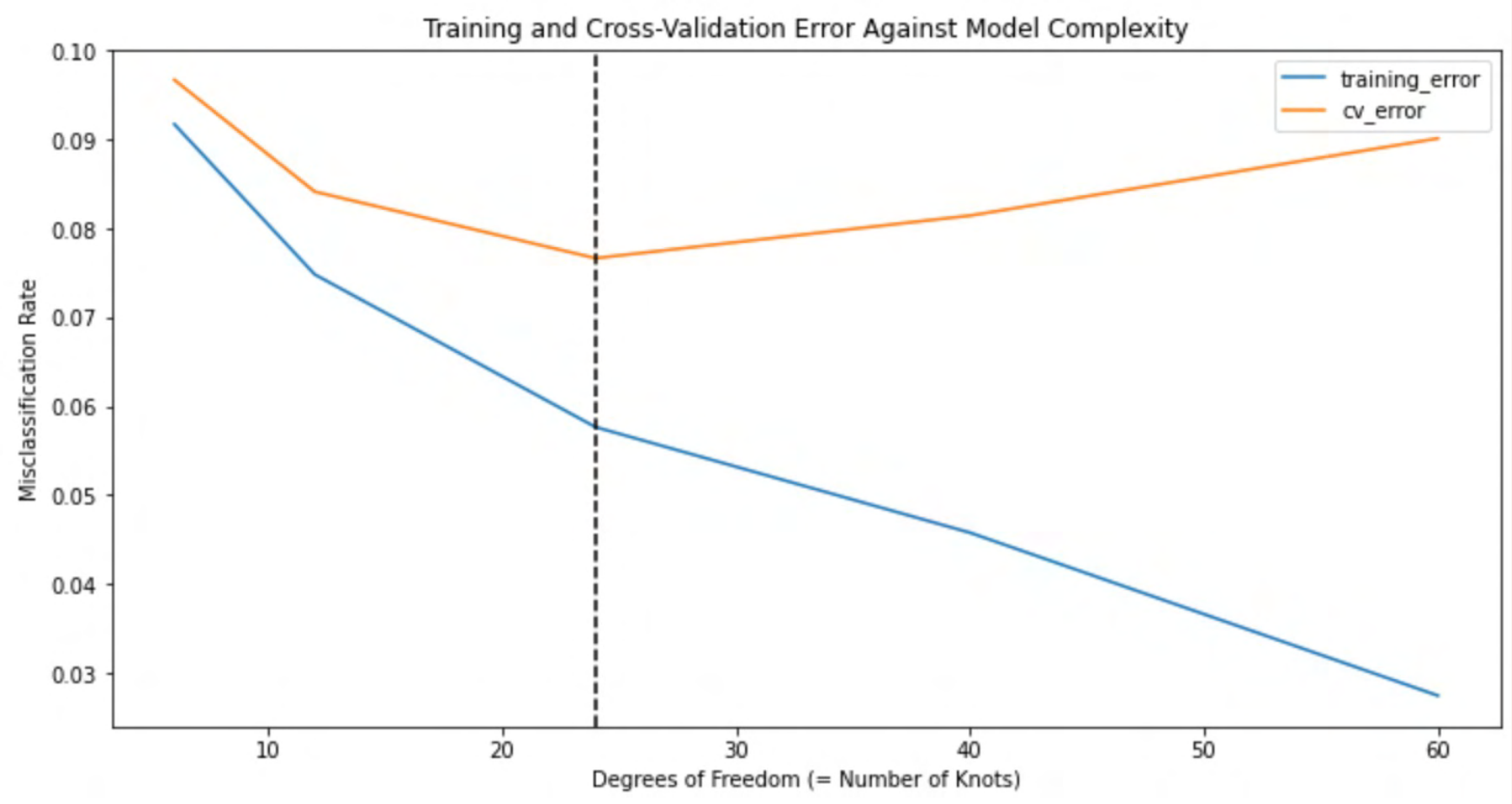
VS

Wavelet Transform : O(N)
Inverse Wavelet Transform : O(N)

Homework - ESL

Phoneme data

자세한 문제 설명은 깃허브에!



Homework - ESL

Ex. 5.12 Characterize the solution to the following problem,

$$\min_f \mathrm{RSS}(f,\lambda) = \sum_{i=1}^{N} w_i\{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt, \qquad (5.73)$$

where the $w_i \geq 0$ are observation weights.

Characterize the solution to the smoothing spline problem (5.9) when the training data have ties in X.

References

- https://www.youtube.com/watch?v=Es0aXmCshm4&index=13&list=PLUXWLrdMaU9vK0mvcRRbg83RqwFvTjHJq&t=0s
- https://jiminsun.github.io/2018-05-27/RKHS/
- https://ngilshie.github.io/jekyll/update/2018/02/01/RKHS.html
- https://ngilshie.github.io/jekyll/update/2018/02/01/RKHS.html
- https://www.youtube.com/watch?v=lgvcyCPxNDQ
- https://www.youtube.com/watch?v=0aZHkrF55cc


Q&A