Lecture 5. Basis Expansions and Regularization

Outline

- Background
- Piecewise-polynomial
- Splines
- Wavelet
- *Dictionary learning*

Background: Moving beyond Linear Model

- Linear regression, LDA, Logistic Regression and separating hyperplanes —— linear models
  - Why ? Simple? Taylor expansion? Non-Overfitting?
- Moving beyond linear model via transformation:\

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X),$$

$h_m : X \to R$

- $h_m(X)$: basis function.
- Beauty: Linear again!

nonlinear feature map

$$X \subseteq R^d \xrightarrow{\text{"}h\text{"}} R^M$$

Background: Examples

- $h_m(X) = \underline{X_m},\ m = 1, \ldots, p$
- $h_m(X) = \underline{X_j^2}$ or $h_m(X) = \underline{X_j X_k}$
- $h_m(X) = \underline{\log(X_j)},\ \underline{\sqrt{X_j}}, \ldots$
- $h_m(X) = \underline{I(L_m \leq X_k < U_m)},$

J^M

O(p^d) for a degree-d polynomial

分片常数

Harr

$\sin(mX)$.

Background: How many basis do we use?

$H : \mathbb{R}^p \rightarrow \mathbb{R}^{\overset{100}{M}}$

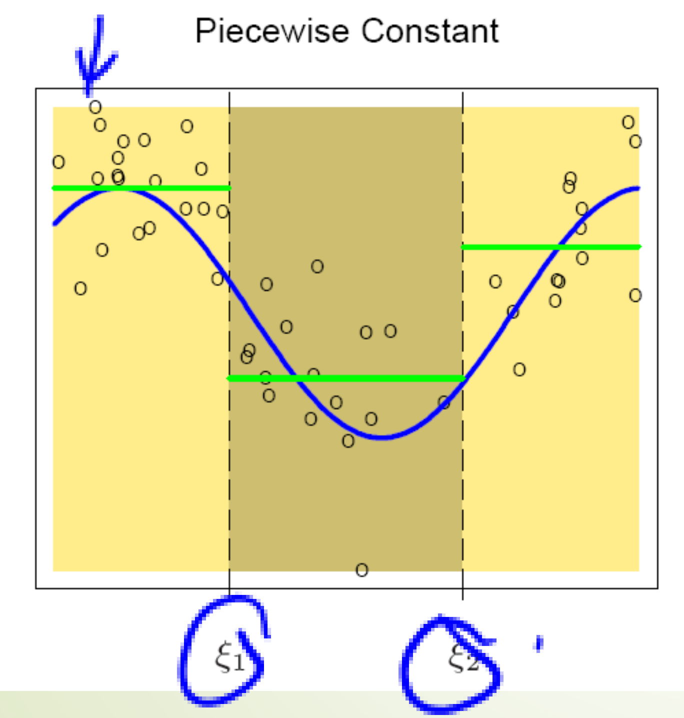
- Restriction methods

$$\begin{aligned} f(X) &= \sum_{j=1}^{p} f_j(X_j) \\ &= f_1(X_1) + f_2(X_2) + f_3(X_3) \\ &= \sum_{m=1}^{M_1} X_1^m + \sum_{m=1}^{M_2} X_2^m + \sum_{m=1}^{M_3} X_3^m \end{aligned}$$

- Selection methods: feature selection methods —— stagewise for example
- Regularization methods: ridge regression for example.

Piecewise Polynomials and “Splines”

- Piecewise Linear (constant)



- Supose that ξ_1 and ξ_2 are known
- $f(X) = \beta_1 h_1(X) + \beta_2 h_2(X) + \beta_3 h_3(X)$

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

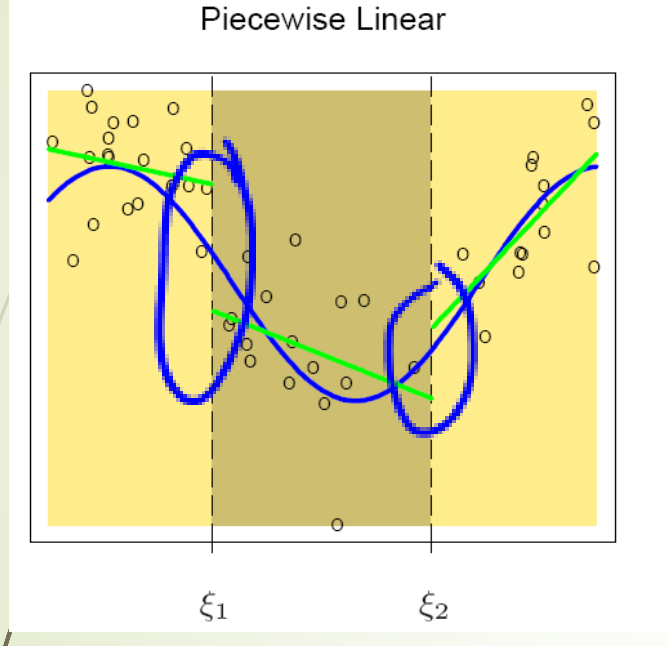
- Least square estimate:

$$\hat{\beta}_m = \bar{Y}_m$$

mean

- Degree of freedom: K+1

Piecewise Linear (Cont')



- Supose that ξ_1 and ξ_2 are known
- $f(X) = \sum_m \beta_m h_m(X)$, where

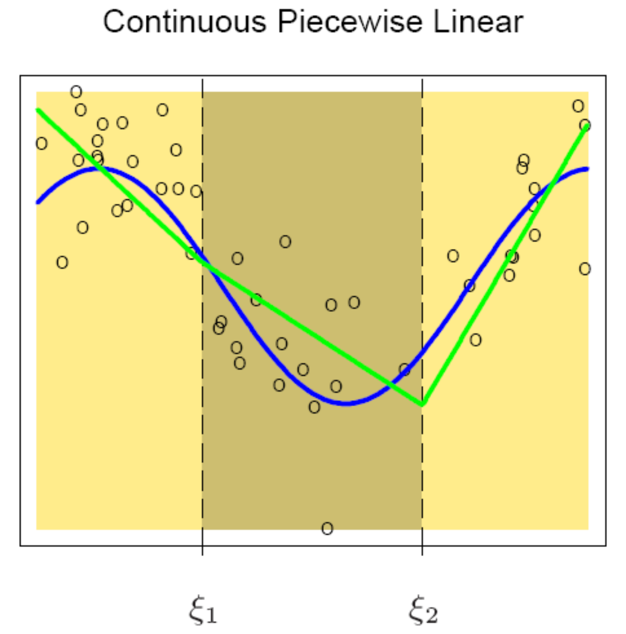
$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X).$$

$$h_{m+3} = h_m(X)X, \; m = 1, \ldots, 3.$$

- These parameters can be estimated via OLS
- Degree of freedom: 2(K+1)

K=2

Piecewise Linear (Cont')



- Supose that ξ_1 and ξ_2 are known
- $f(X) = \sum_m \beta_m h_m(X)$, where

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

$$h_{m+3} = h_m(X)X, \; m = 1, \dots, 3.$$

- With two constraints:

$$f(\xi_1 -) = f(\xi_1 +) \; and \; f(\xi_2 -) = f(\xi_2 +)$$

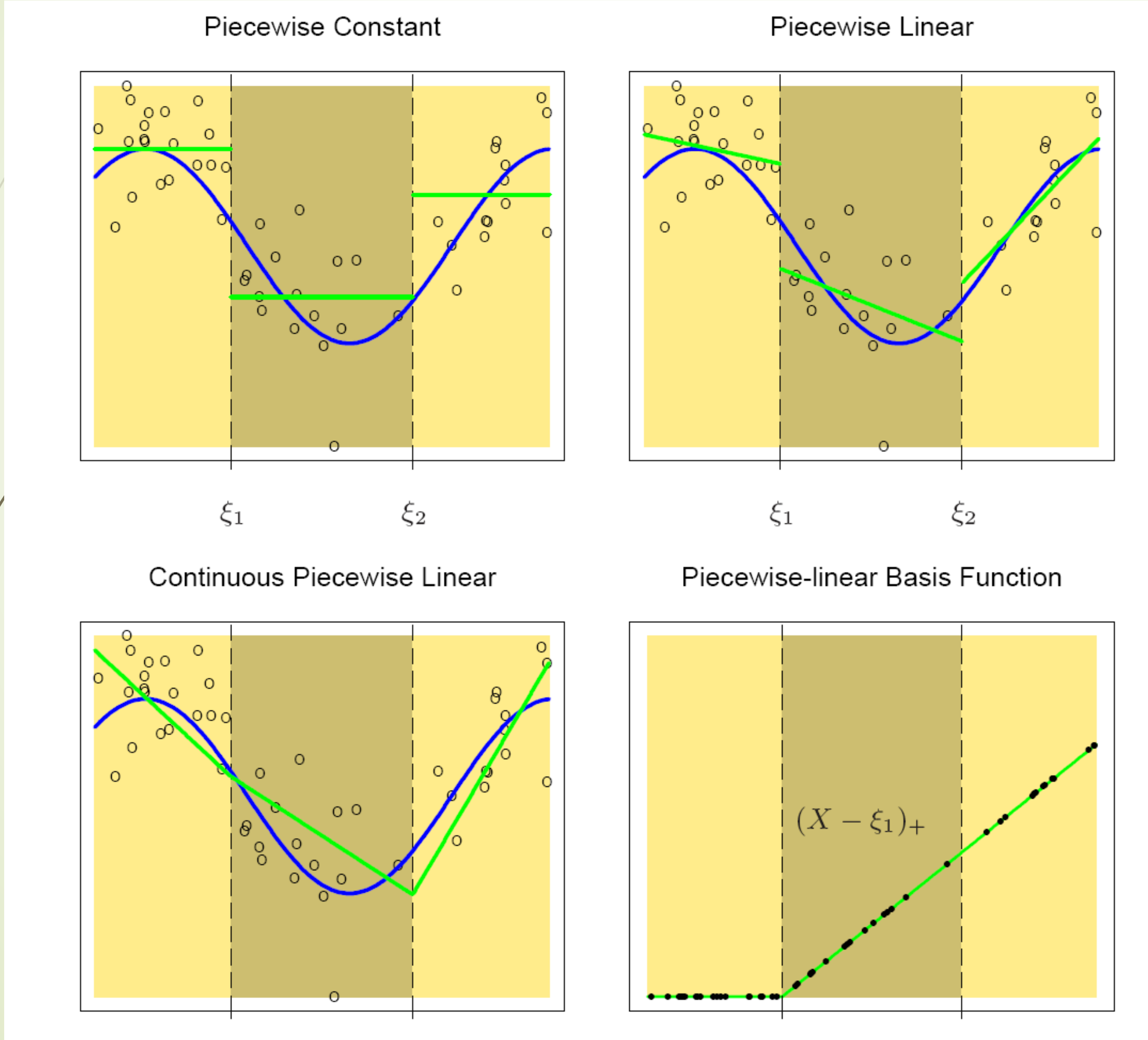
$$\beta_1 + \beta_4\xi_1 = \beta_2 + \beta_5\xi_1 \; and \; \beta_2 + \beta_5\xi_2 = \beta_3 + \beta_6 \, \xi_2$$

- These parameters can be estimated via OLS

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

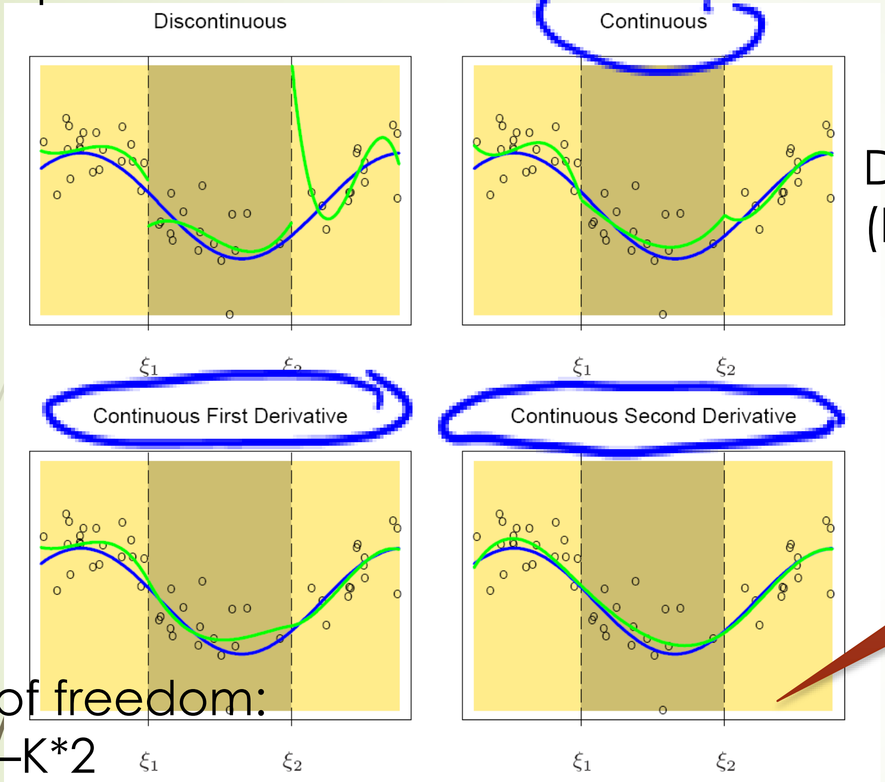
$2(K+1)-K$
$= K+1$

Piecewise Linear (Cont')



Piecewise Cubic

Degree of freedom:
(K+1) *4



Degree of freedom:
(K+1) *4 -K

Degree of freedom:
(K+1) *4 –K*2

1990, Wahba
Spline models
for obs.
data

Cubic Spline

Exercise!

$$h_1(X) = 1, \quad h_3(X) = X^2, \quad h_5(X) = (X - \xi_1)^3_+,$$
$$h_2(X) = X, \quad h_4(X) = X^3, \quad h_6(X) = (X - \xi_2)^3_+.$$

Degree of freedom: ✓
(K+1) *4 – K*3

Piecewise Polynomial

- K knots, order M spline:

$$\begin{aligned} h_j(X) &= X^{j-1},\ j = 1, \ldots, M, \\ h_{M+\ell}(X) &= (X - \xi_\ell)_+^{M-1},\ \ell = 1, \ldots, K. \end{aligned}$$

- It is claimed that cubic splines are the lowest order splines for which the knot discontinuity is not visible to the human eye!
- Widely used: piecewise constant, piecewise linear and cubic spline
- Basis functions are not unique! B-spline basis is more efficient
- DF: M+K

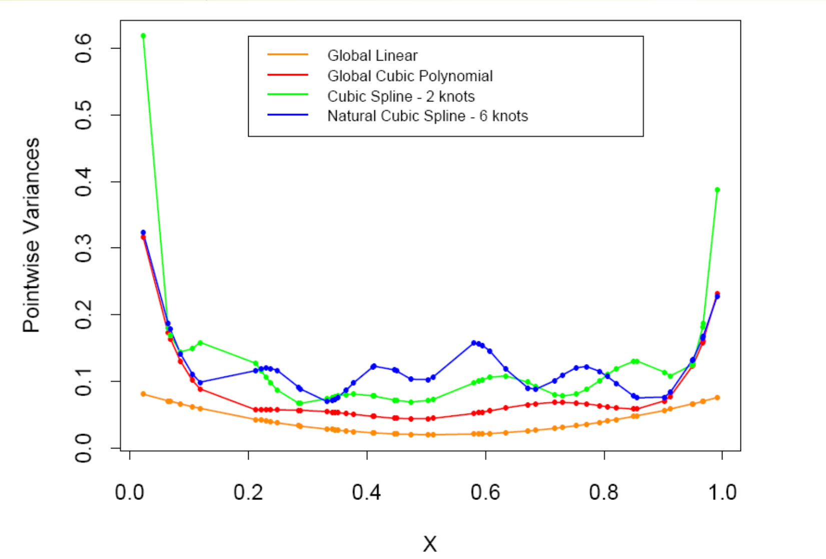
Piecewise Polynomial (Cont')

- These fixed-knot splines are also known as regression splines.
- Regression splines are determined by

the order of spline, the number of knots and their placement

- R: bs(x,df=7) generates a basis matrix of cubic-spline functions
- M = 4, K = df – M + 1 =7-3 = 4 knots
- By default, the four knots are (20th,40th ,60th and 80th) percentiles of x
- bs(x,degree = 1, knots= c(0.2,0.4,0.6)) generates an $N \times 4$ matrix

Natural Cubic Splines



Pointwise variance curves

$$X \sim U[0,1]$$
$$Y = X + N(0,1)$$

n= 50

Cubic spline: two knots at 0.33 and 0.66

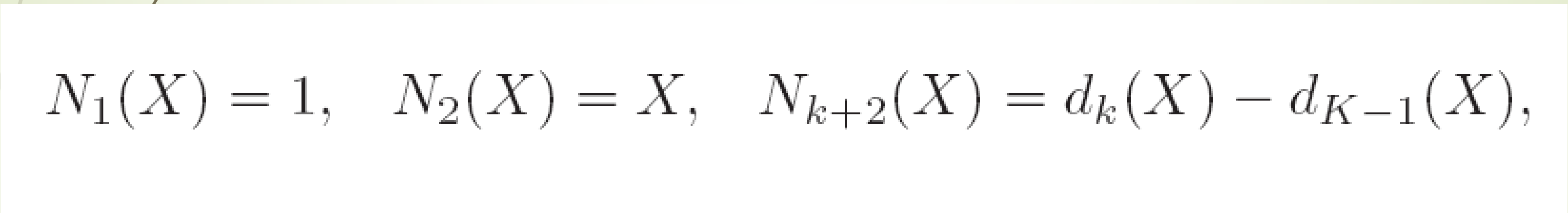
Natural spline: two boundary kn
0.1 and 0.9, four interior knots
uniformly spaced between the

$$Y = \sum h_m(X)\,\beta_m + \epsilon = H\beta + \epsilon$$
$$\hat{\beta} = (H^T H)^{-1} H^T Y$$
$$var(\hat{\beta}) = \underline{(H^T H)^{-1}\sigma^2}$$
$$\underline{var(H\hat{\beta}) = H(H^T H)^{-1} H^T \sigma^2}$$

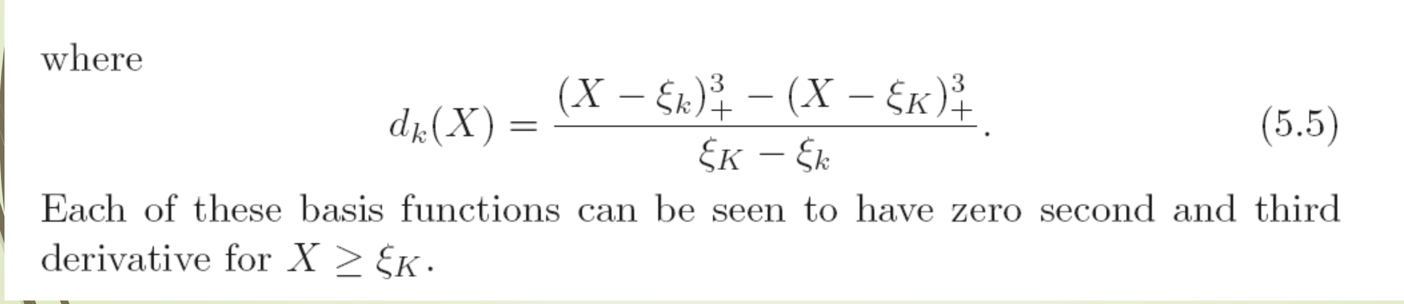
Natural Cubic Splines

- Two more constraints: linear beyond the boundary knots: frees 4 parameters
- K knots, K basis:

  K + 4 -4



$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$



where

$$d_k(X) = \frac{(X - \xi_k)^3_+ - (X - \xi_K)^3_+}{\xi_K - \xi_k}. \qquad (5.5)$$

Each of these basis functions can be seen to have zero second and third derivative for $X \geq \xi_K$.

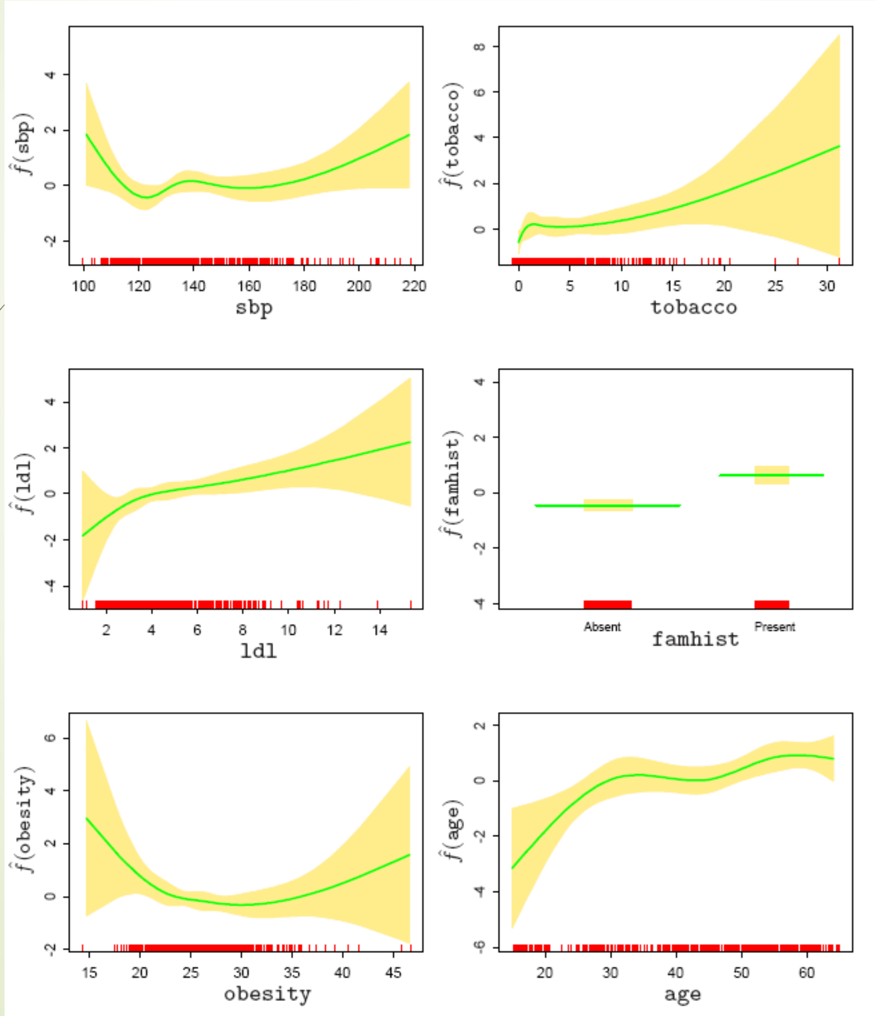
Example: South African Heart Disease

$$\text{logit}[\Pr(\text{chd}|X)] = \theta_0 + h_1(\underset{\uparrow}{X_1})^T\theta_1 + h_2(\underset{\uparrow}{X_2})^T\theta_2 + \cdots + h_p(\underset{\uparrow}{X_p})^T\theta_p,$$

- Four natural spline bases for each term are used
- 5 ? knots (3 chosen at random as interior knots, 2 boundary knots at the extremes) [?—exclude the constant term for each h_j]
- Binary variable is kept as itself

generalised additive model.

Example: South African Heart Disease (pointwise variance)



Example: South African Heart Disease -- Backward selection

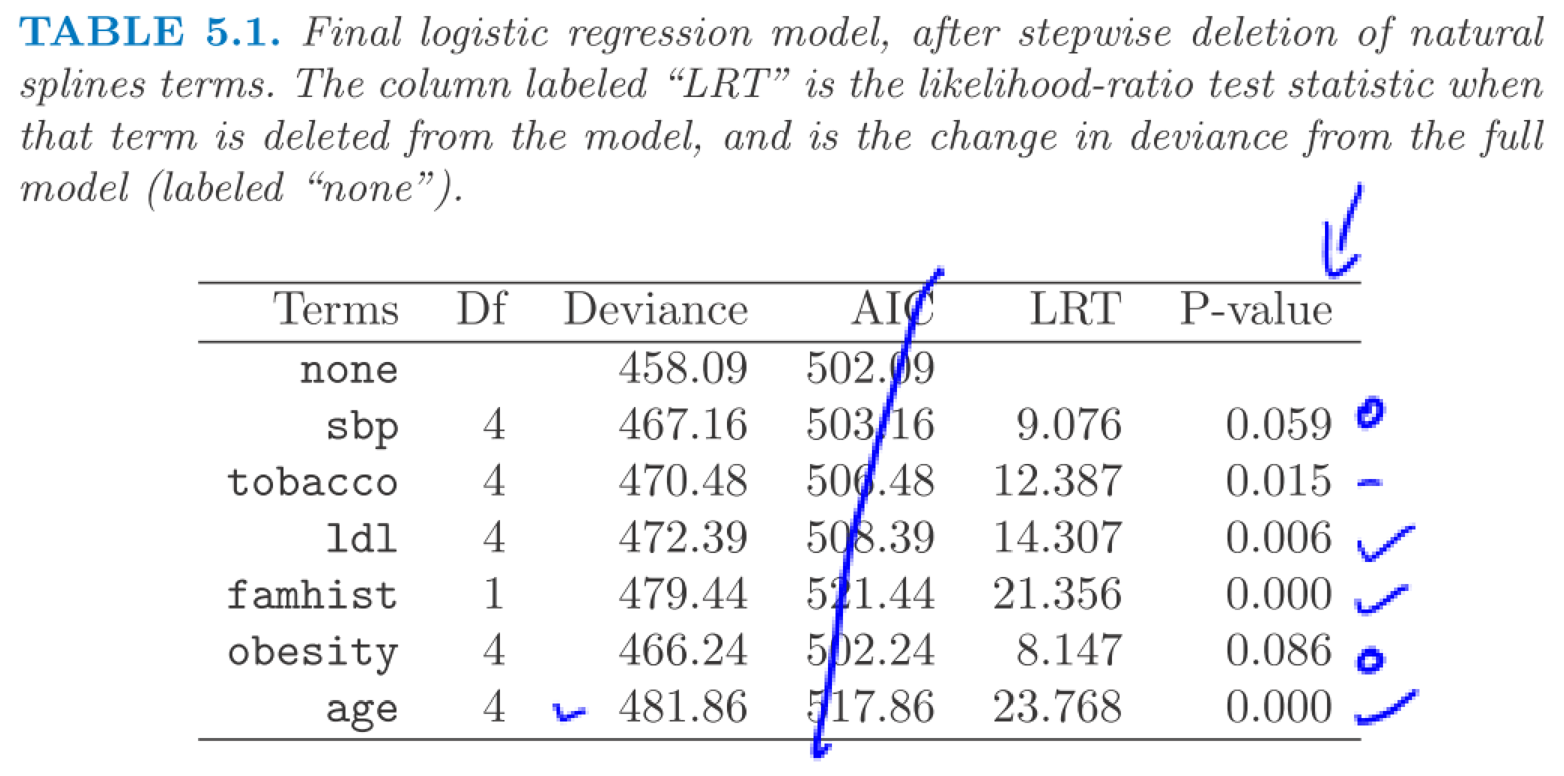


TABLE 5.1. *Final logistic regression model, after stepwise deletion of natural splines terms. The column labeled "LRT" is the likelihood-ratio test statistic when that term is deleted from the model, and is the change in deviance from the full model (labeled "none").*

Terms	Df	Deviance	~~AIC~~	LRT	P-value ✓
`none`		458.09	~~502.09~~		
`sbp`	4	467.16	~~503.16~~	9.076	0.059 o
`tobacco`	4	470.48	~~506.48~~	12.387	0.015 -
`ldl`	4	472.39	~~508.39~~	14.307	0.006 ✓
`famhist`	1	479.44	~~521.44~~	21.356	0.000 ✓
`obesity`	4	466.24	~~502.24~~	8.147	0.086 o
`age`	4	✓ 481.86	~~517.86~~	23.768	0.000 ✓

Smoothing Splines

Wahba

- To avoid Knot selection
- Regularization

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

- λ is called smooth parameter, because

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

$\lambda=\infty$: the simple least squares line fit, since no second derivative can be tolerated.

- The solution of $\min_{f} RSS(f,\lambda)$is a natural cubic spline with knots at x_i. ——Exercise!

"RKHS"

Smoothing Splines

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

$$f(x)=\sum_{j=1}^{N}N_j(x)\theta_j,$$

Representer Thm

$N_j(x) = [K(x, x_j)]$

$$\mathrm{RSS}(\theta,\lambda)=(\mathbf{y}-\mathbf{N}\theta)^T(\mathbf{y}-\mathbf{N}\theta)+\lambda\theta^T\mathbf{\Omega}_N\theta,$$

$$\{\mathbf{\Omega}_N\}_{jk}=\int N_j''(t)N_k''(t)dt.$$

≥ 0

RKHS

"Kernel"

$$\hat{\theta}=(\mathbf{N}^T\mathbf{N}+\lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y},$$

→ $\mathrm{Var}(\hat{\theta})$

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

Smoothing Parameter Selection

- Df: degree of freedom.

$$\begin{aligned}\boxed{\hat{\mathbf{f}}} &= \underline{\mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T}\mathbf{y}\\ &= \mathbf{S}_\lambda\mathbf{y}.\end{aligned}$$

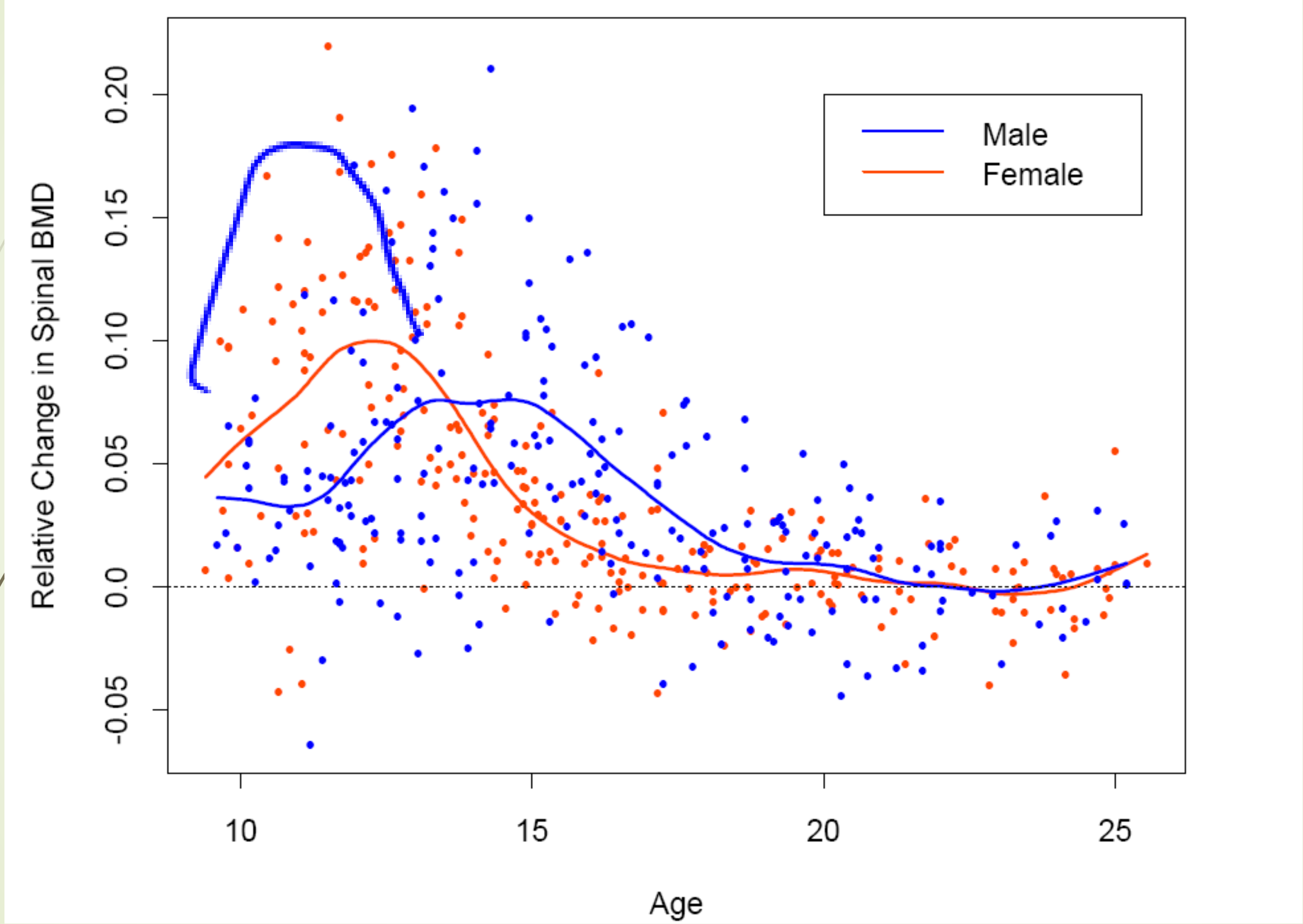
linear map.

$$\underline{\mathrm{df}_\lambda = \mathrm{trace}(\mathbf{S}_\lambda)},$$

$S_\lambda = \underline{I_d}$

N

Example



with $\lambda \approx 0.00022.$

df_λ

Smoothing Parameter Selection

df =

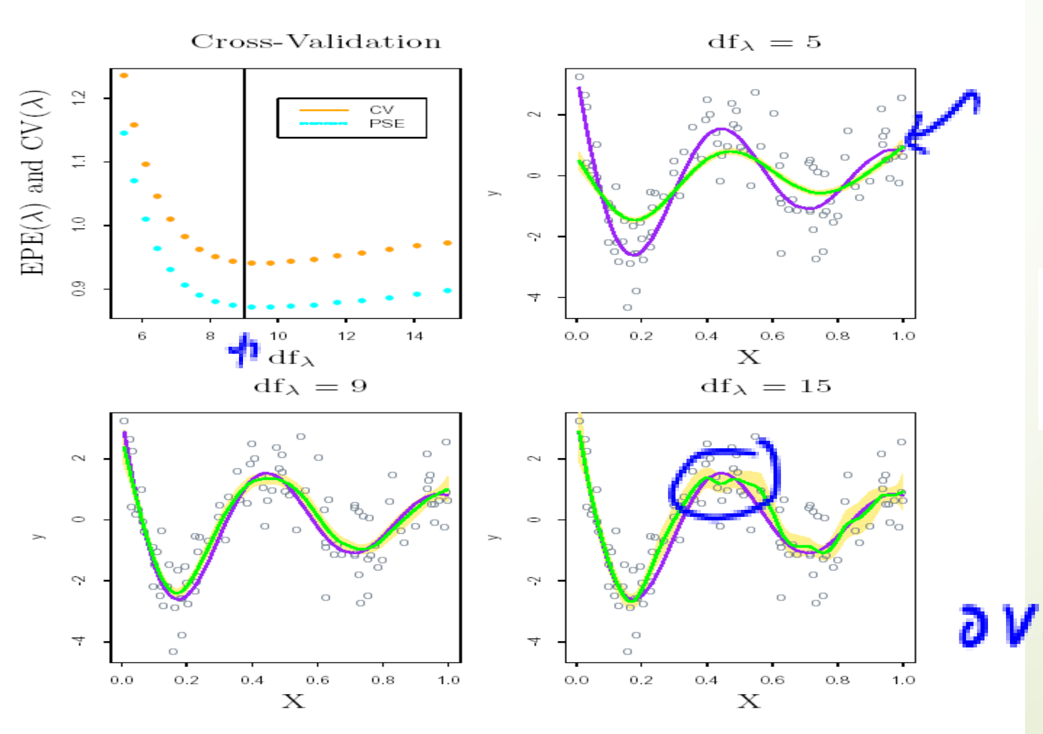
- Specify fix degree of freedom $Tr(S_\lambda)$

  2

  - R> smooth.spline(x,y,df=??)
- Try a couple of values of df. and choose one based on a model selection criteria
  - Integrated EPE
  - K-fold CV to choose the value of λ

Smoothing Parameter Selection(Cont')

$$
\begin{aligned}
Y &= f(X) + \varepsilon,\\
f(X) &= \frac{\sin(12(X+0.2))}{X+0.2},
\end{aligned}
$$



overfit

True Function

Fitted Function

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

Nonparametric Logistic Regression

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

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

$$\begin{aligned} \ell(f;\lambda) &= \sum_{i=1}^{N} [y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))] - \frac{1}{2}\lambda \int \underline{\{f''(t)\}^2 dt} \\ &= \underline{\sum_{i=1}^{N} \left[y_i f(x_i) - \log(1+e^{f(x_i)})\right]} \underline{- \frac{1}{2}\lambda \int \{f''(t)\}^2 dt}, \qquad (5.30) \end{aligned}$$

Multidimensional Splines

$$h_{1k}(X_1),\ \ k = 1, \ldots, M_1$$

$$h_{2k}(X_2)$$

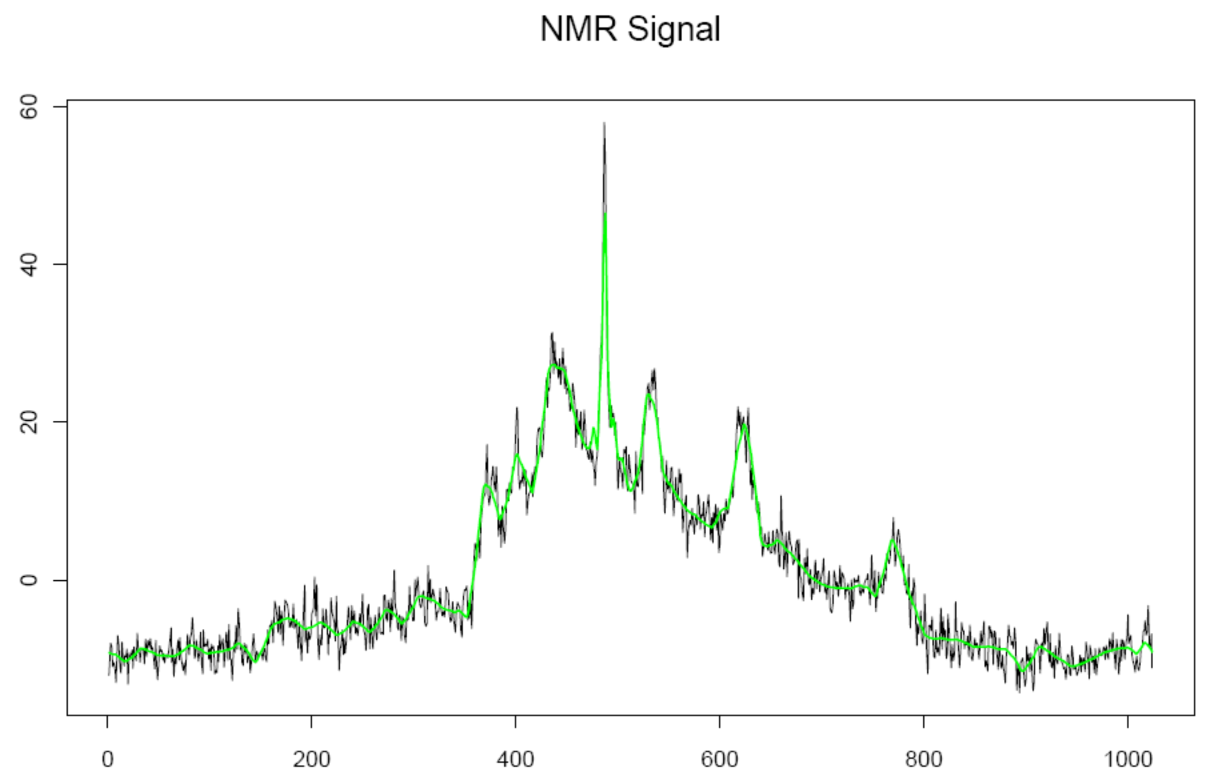
$$g_{jk}(X) = h_{1j}(X_1)h_{2k}(X_2),\ j = 1, \ldots, M_1,\ \ k = 1, \ldots, M_2$$

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

These are naturally extended to ANOVA spline decompositions,

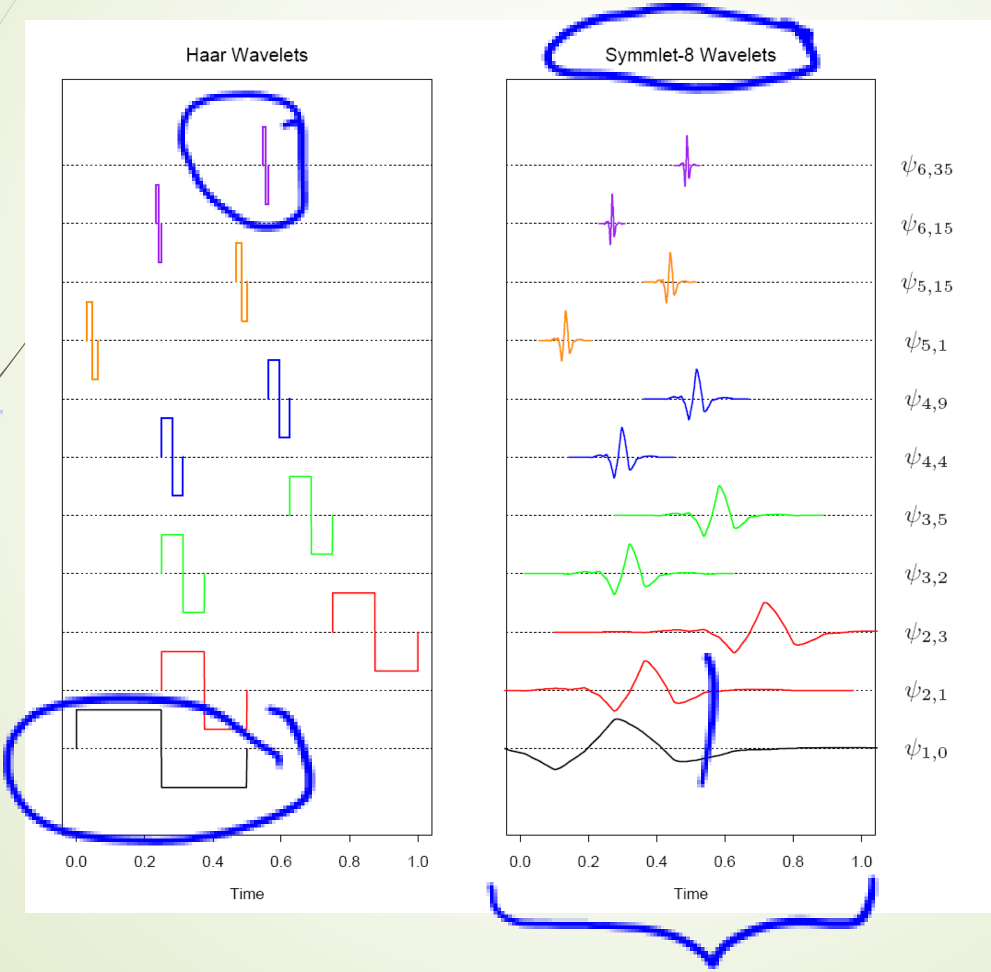
$$f(X) = \alpha + \underline{\sum_j \overset{\downarrow}{f_j}(X_j)} + \sum_{j<k} f_{jk}(\underline{X_j, X_k}) + \cdots ,$$

Wavelet Smoothing



Wavelet Smoothing

Scale j

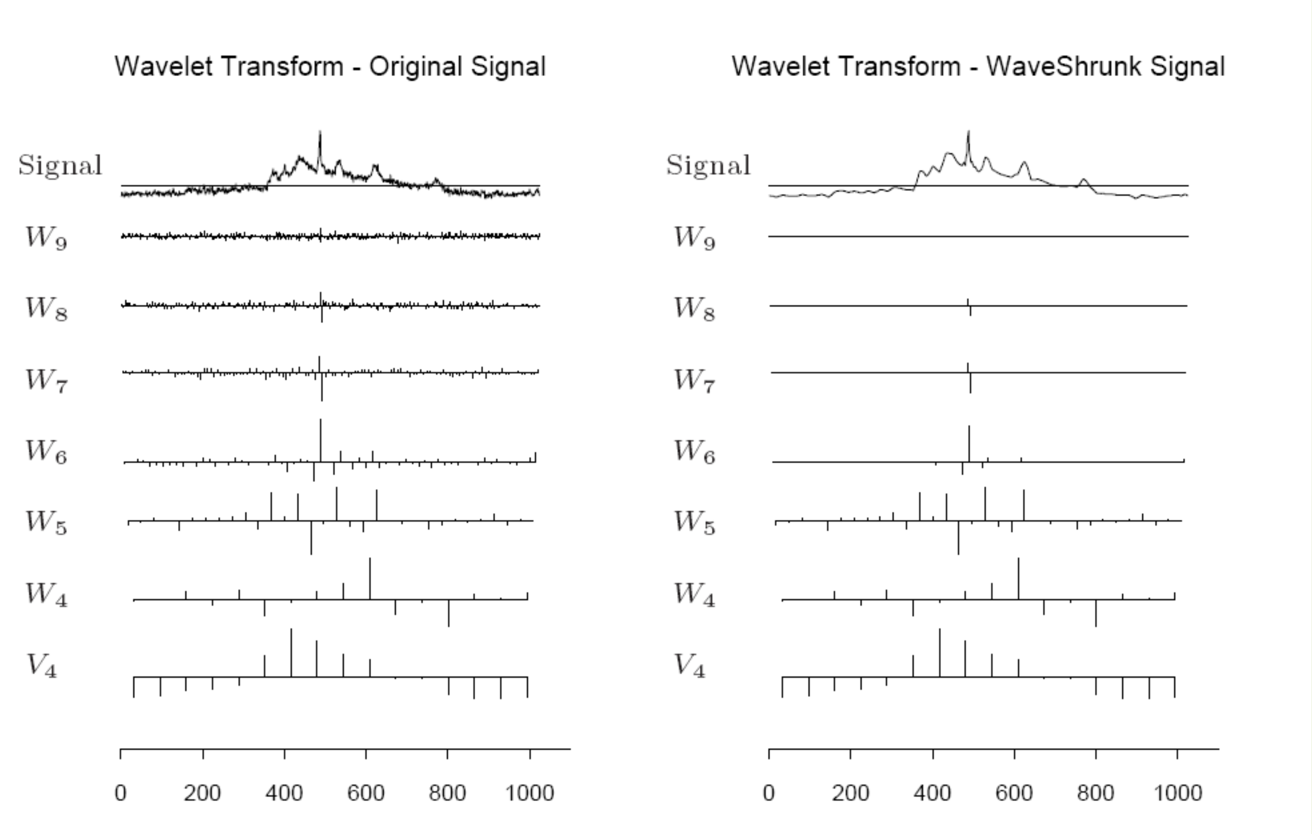


'Daubechies'
1990s

'Compact support wavelet'

Smoothness

Wavelet Smoothing



Dictionary Learning

Consider a signal X. In many cases X can be represented by some "atoms" (keywords, topics).

For example: (1) PCA.

$$\underline{X = U DV^T.}$$

$$\min_{A_r} \|X - A_r\|_2^2$$

U can be treat as the "keywords" and DV^T is the loadings of the "keywords".

(2) Wavelet or DCT.

$$X = \underline{\overset{\swarrow}{D}}\alpha$$

$$X = \underset{\overset{\uparrow}{\text{orthognal}}}{U} \cdot *$$

(3) Documents.

$$X = D\alpha$$

*

Dictionary Learning

Y = AB

Y^2

- Cane we learn a good dictionary?

$$\arg\min_{D,\alpha} \frac{1}{N} \sum_i \| \underset{=}{y_i} - \overset{\downarrow}{\underset{?}{D}} \underset{?}{\alpha_i} \|_2^2 + \lambda \sum_i \| \underset{\uparrow}{\alpha_i} \|_1 .$$

"LASSO" ?

nonconvex

(1) Fix $\underset{=}{D}$ and update α_i. → LASSO

(2) Fix $\alpha_i, i = 1, \ldots, N$ and update D. PCA

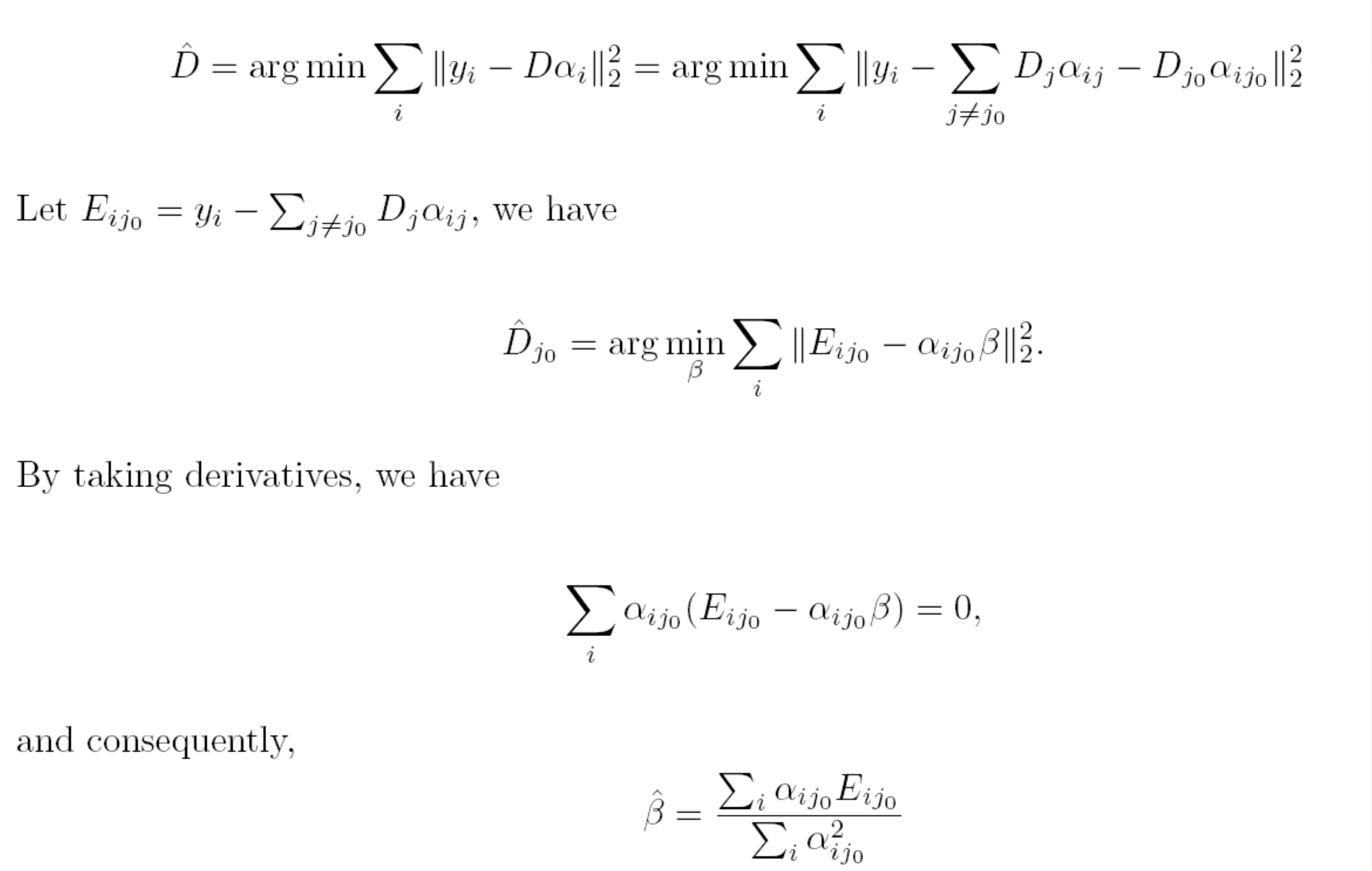
"LSQ"

① Convergence !

< Arora · "anchor"

Tensor

Dictionary Learning



$$\hat{D} = \arg\min \sum_i \|y_i - D\alpha_i\|_2^2 = \arg\min \sum_i \|y_i - \sum_{j \neq j_0} D_j \alpha_{ij} - D_{j_0}\alpha_{ij_0}\|_2^2$$

Let $E_{ij_0} = y_i - \sum_{j \neq j_0} D_j \alpha_{ij}$, we have

$$\hat{D}_{j_0} = \arg\min_{\beta} \sum_i \|E_{ij_0} - \alpha_{ij_0}\beta\|_2^2.$$

By taking derivatives, we have

$$\sum_i \alpha_{ij_0}(E_{ij_0} - \alpha_{ij_0}\beta) = 0,$$

and consequently,

$$\hat{\beta} = \frac{\sum_i \alpha_{ij_0} E_{ij_0}}{\sum_i \alpha_{ij_0}^2}$$

Homework

- Due Oct 26
- ESLII, Chapter 5, Exercise: 5.1, 5.3, 5.4, 5.5, 5.7