Basis Expansions and Regularization

Model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X) \ \ (X \text{ is a vector})$$

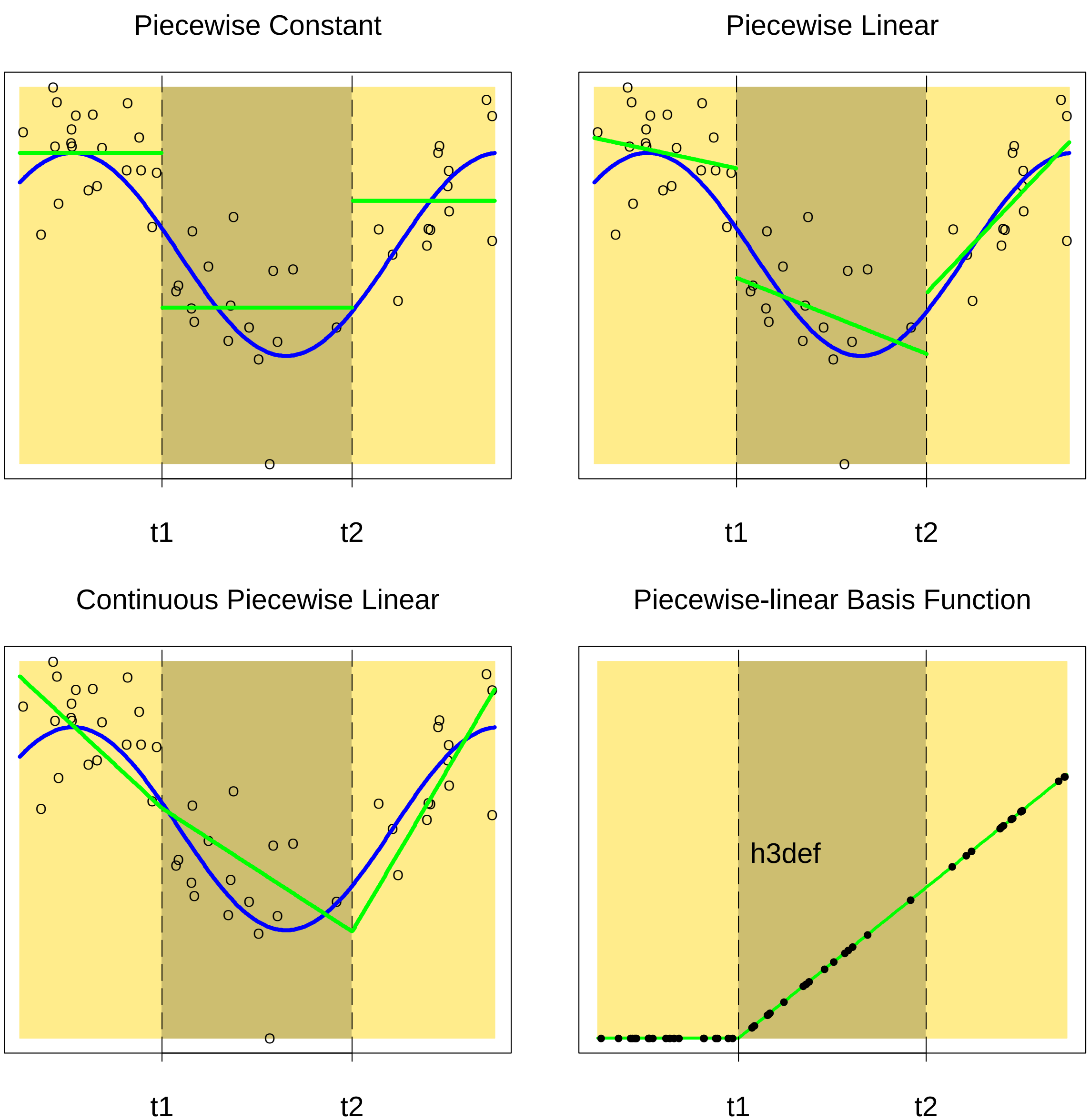
- $h_m(X) = X_j^2, X_j X_\ell, \ldots$
- $h_m(X) = ||X||, \log(X_j), \ldots.$
- $h_m(X) = I(L_m < X_k < U_m)$

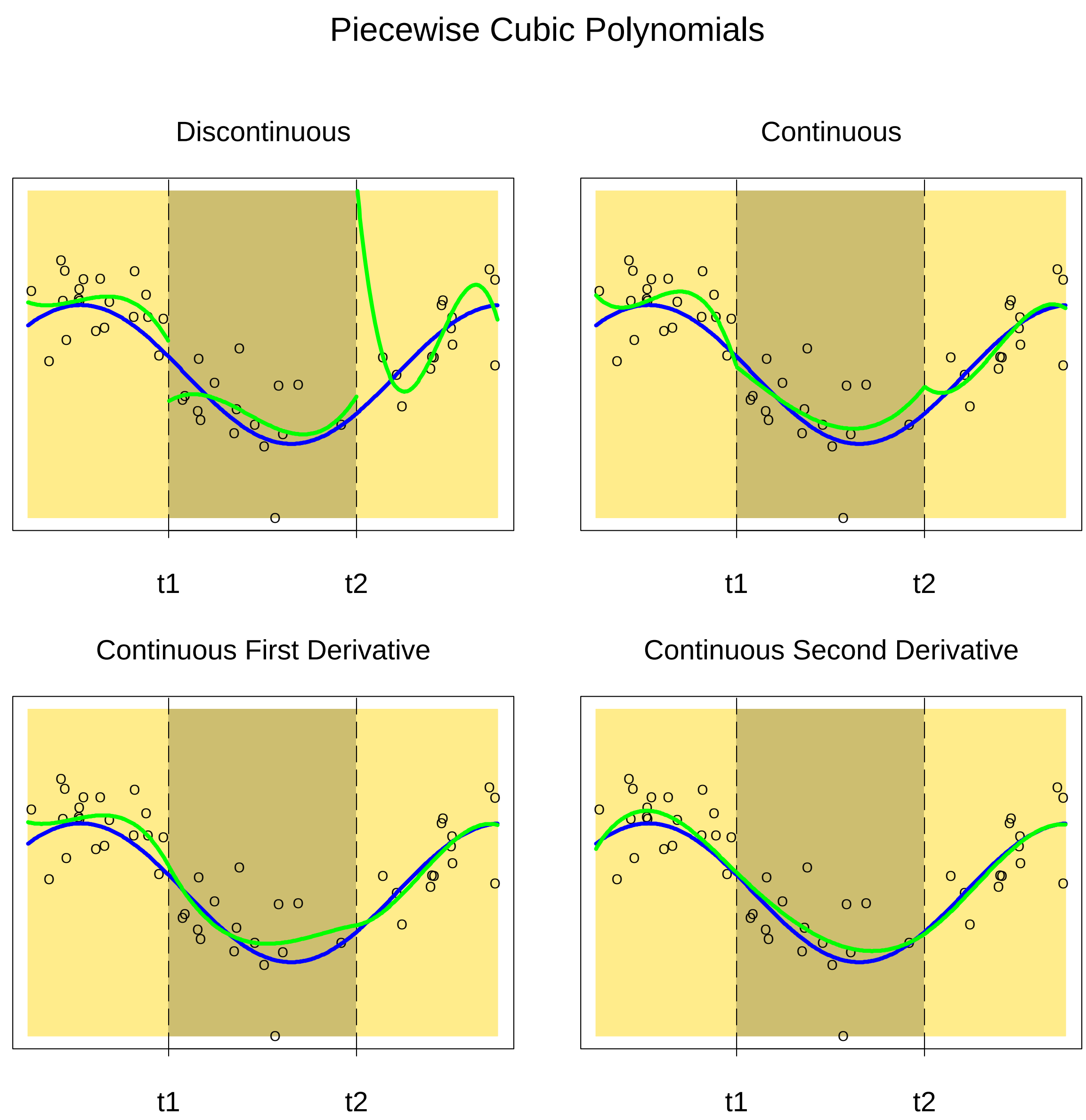
Regularization

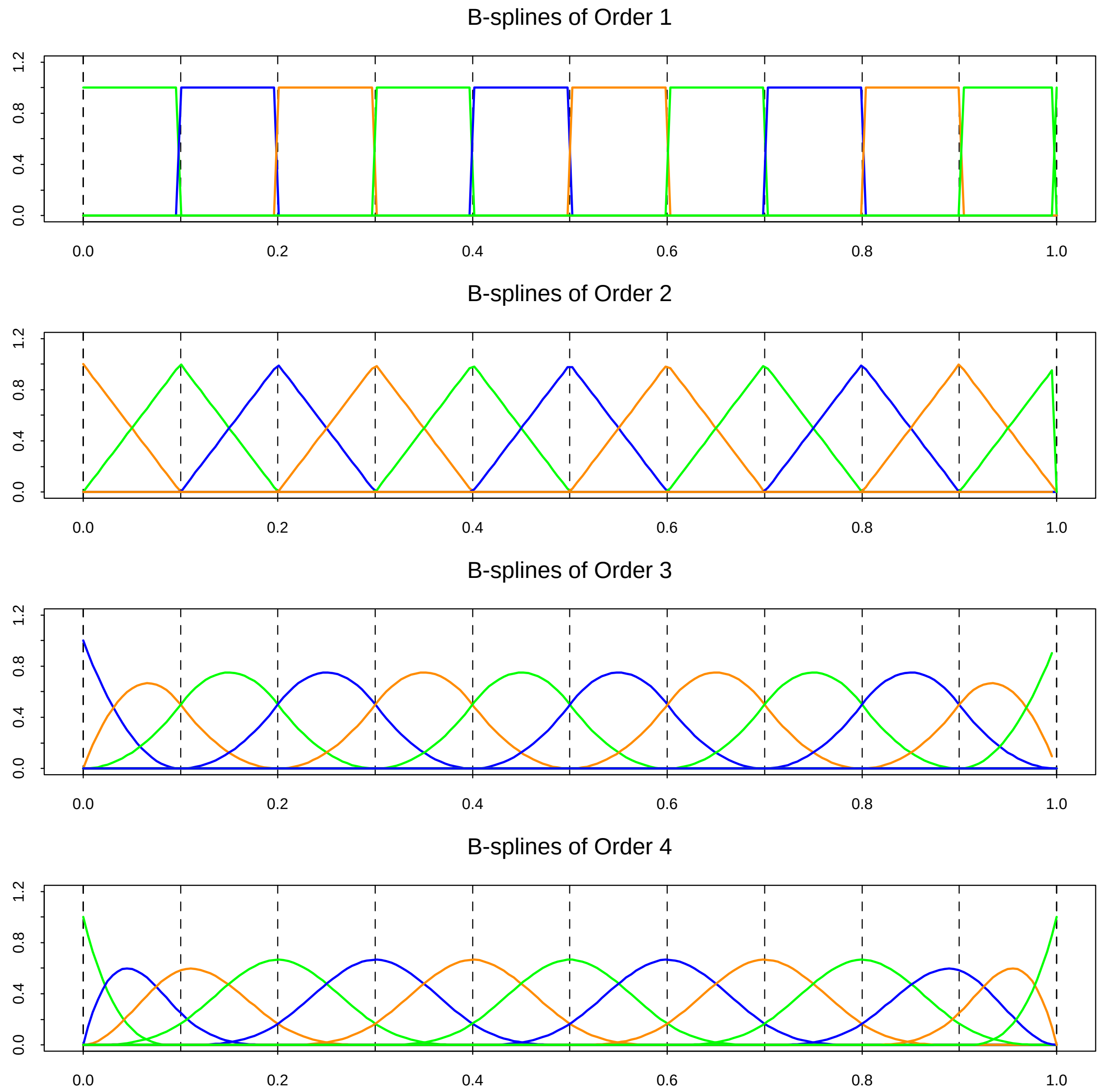
$$\min \sum_{i=1}^{n} (y_i - \sum_{m=1}^{M} \beta_m h_m(x_i))^2 + \lambda J(f)$$

where $f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$.

$J(f) = ||\beta||_2^2$ for example.


Piecewise Constant
Piecewise Linear
t1
t2
Continuous Piecewise Linear
Piecewise-linear Basis Function
h3def


Piecewise Cubic Polynomials
Discontinuous
Continuous
Continuous First Derivative
Continuous Second Derivative
t1
t2

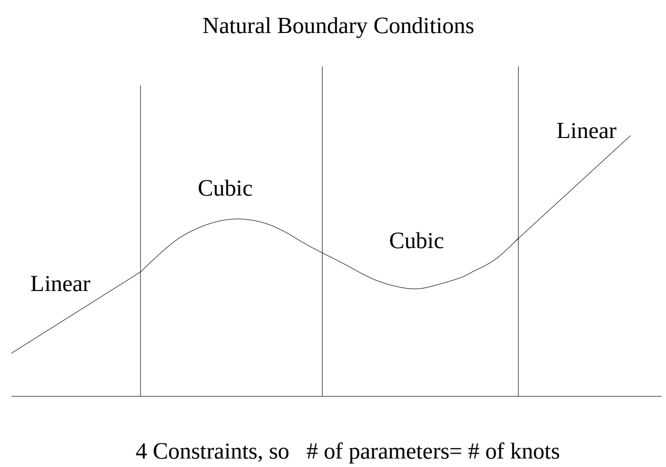

B-splines of Order 1
B-splines of Order 2
B-splines of Order 3
B-splines of Order 4
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0

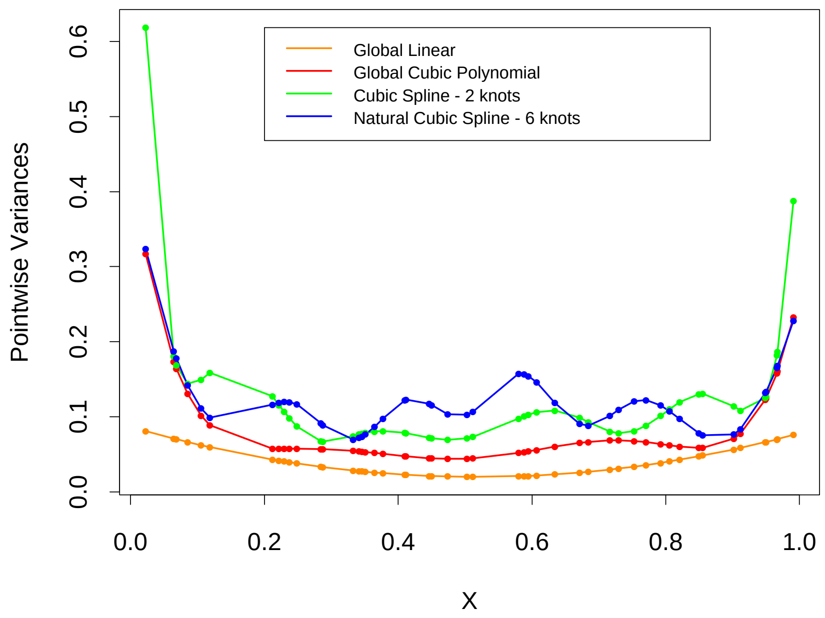
Cubic splines and natural cubic splines

In R,

```
bs(x, degree=3, knots=c(,2.,4.,6))
```

Should return an $N \times 7$ matrix (actually $N \times 6$ since intercept$= F$ is default).





Pointwise variance curves for four different models, with X consisting of 50 points drawn at random from $U[0,1]$, and an assumed error model with constant variance. The linear and cubic polynomial fits have two and four degrees of freedom, respectively, while the cubic spline and natural cubic spline each have six degrees of freedom. The cubic spline has two knots at 0.33 and 0.66, while the natural spline has boundary knots at 0.1 and 0.9, and four interior knots uniformly spaced between them.

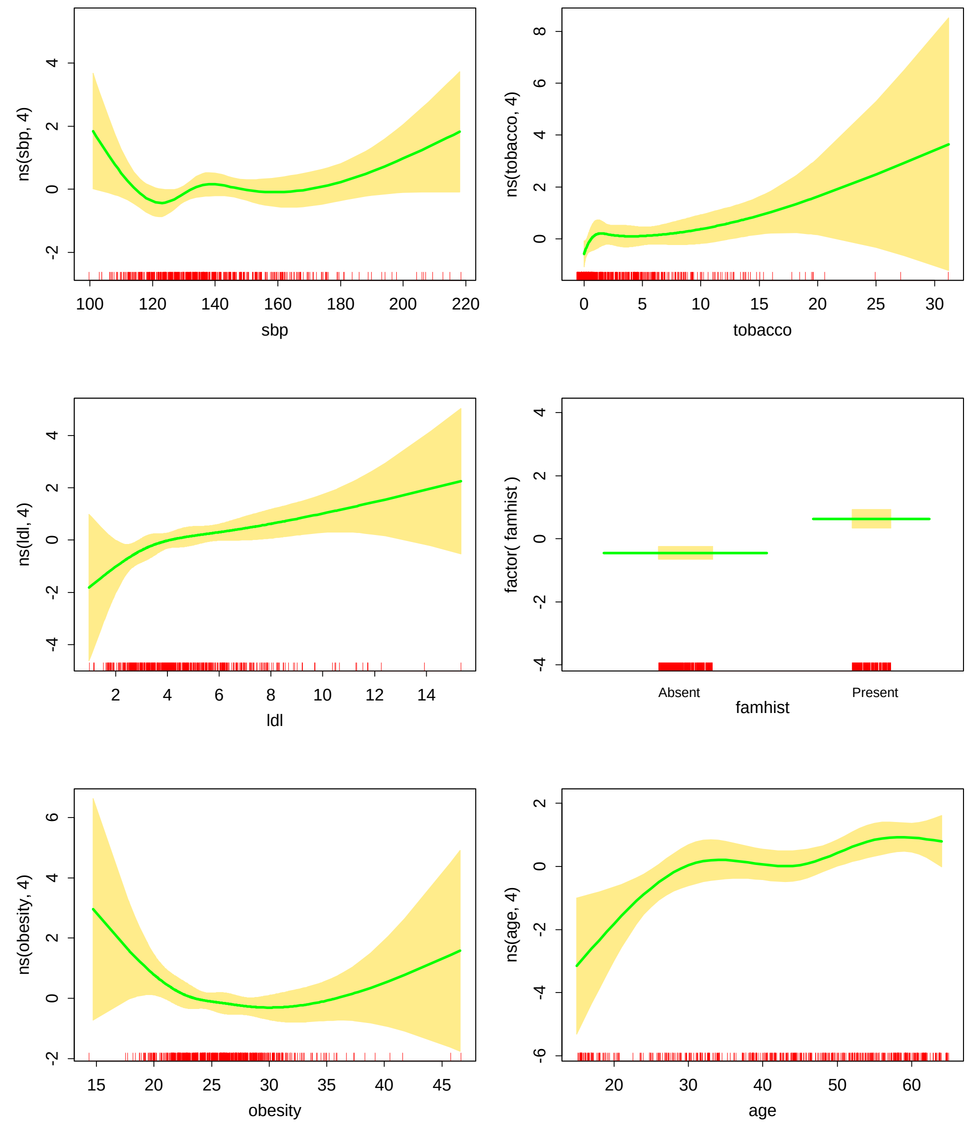
South African Heart Disease data

$$\begin{aligned}\text{logit}[Pr(chd|x)] &= \theta_0 + h_1(x_1)^T + h_2(x_2)^T\theta_2 + \cdots h_p(x_p)^T\theta_p \\ &= h(x)^T\theta\end{aligned}$$

- $h_j(x_j) = \mathrm{ns(x_j, df = 4)}$
- Basis matrix H, $n \times (1 + \sum_{j=1}^{p} \mathrm{df}_j)$
- $\hat{\theta}$ obtained from binomial maximum likelihood (logistic regression)
- $\widehat{\mathrm{Cov}} = (H^T W H)^{-1} = \hat{\Sigma}$, $W = \mathrm{diag[\hat{p}_i(1 - \hat{p}_i])}$.
- $\hat{f}_j(x_j) = h_j(x_j)^T\hat{\theta}_j$, $\widehat{\mathrm{var}} = h_j(x_j)^T\hat{\Sigma}_{jj}h_j(x_j)$.

Table 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
`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
`age`	4	481.86	517.86	23.768	0.000

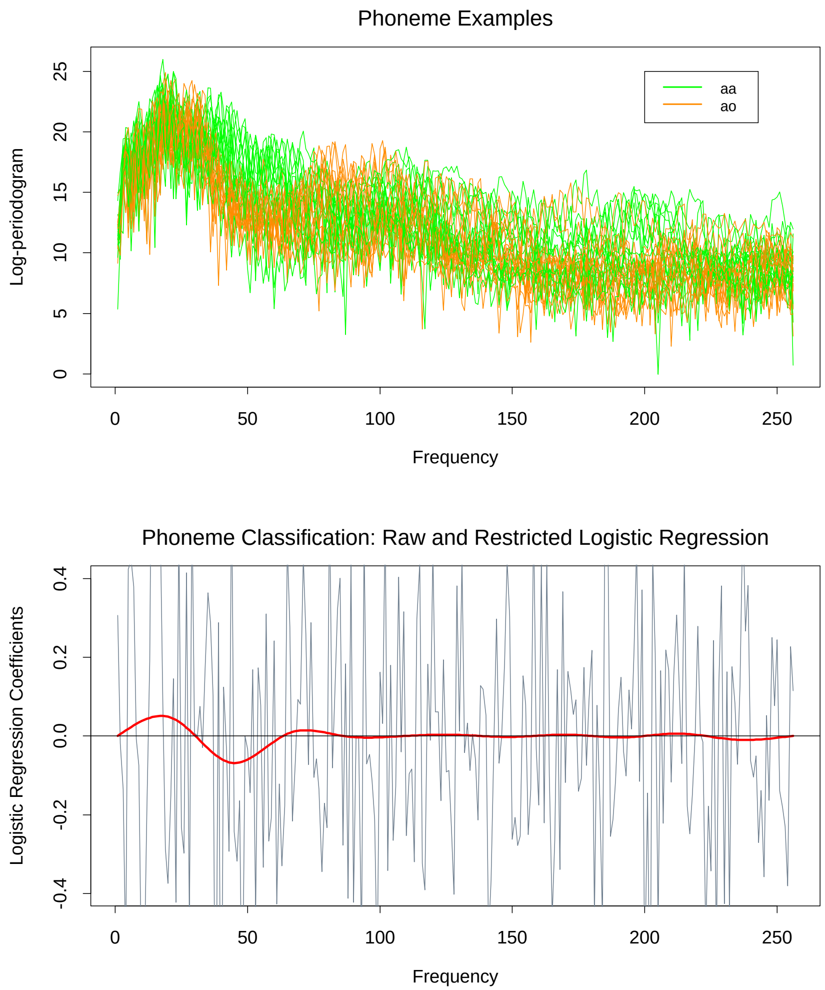

ns(sbp, 4)
sbp
100
120
140
160
180
200
220
ns(tobacco, 4)
tobacco
0
5
10
15
20
25
30
ns(ldl, 4)
ldl
2
4
6
8
10
12
14
factor(famhist)
famhist
Absent
Present
ns(obesity, 4)
obesity
15
20
25
30
35
40
45
ns(age, 4)
age
20
30
40
50
60

Example: Phoneme recognition

- $X(f)$ observed over grid of frequencies, $x_j = X(f_j)$.
- Two classes "aa" (695) and "ao" (1022)
- $\log \frac{Pr(aa|x)}{Pr(ao|x)} = \int X(f)\beta(f)df \approx \sum_{j=1}^{256} x(f_j)\beta(f_j) = \sum_{j=1}^{256} x_j\beta_j$
- $\beta(f) = \sum_{m=1}^{M} h_m(f)\theta_m$

CV Misclassification rates:

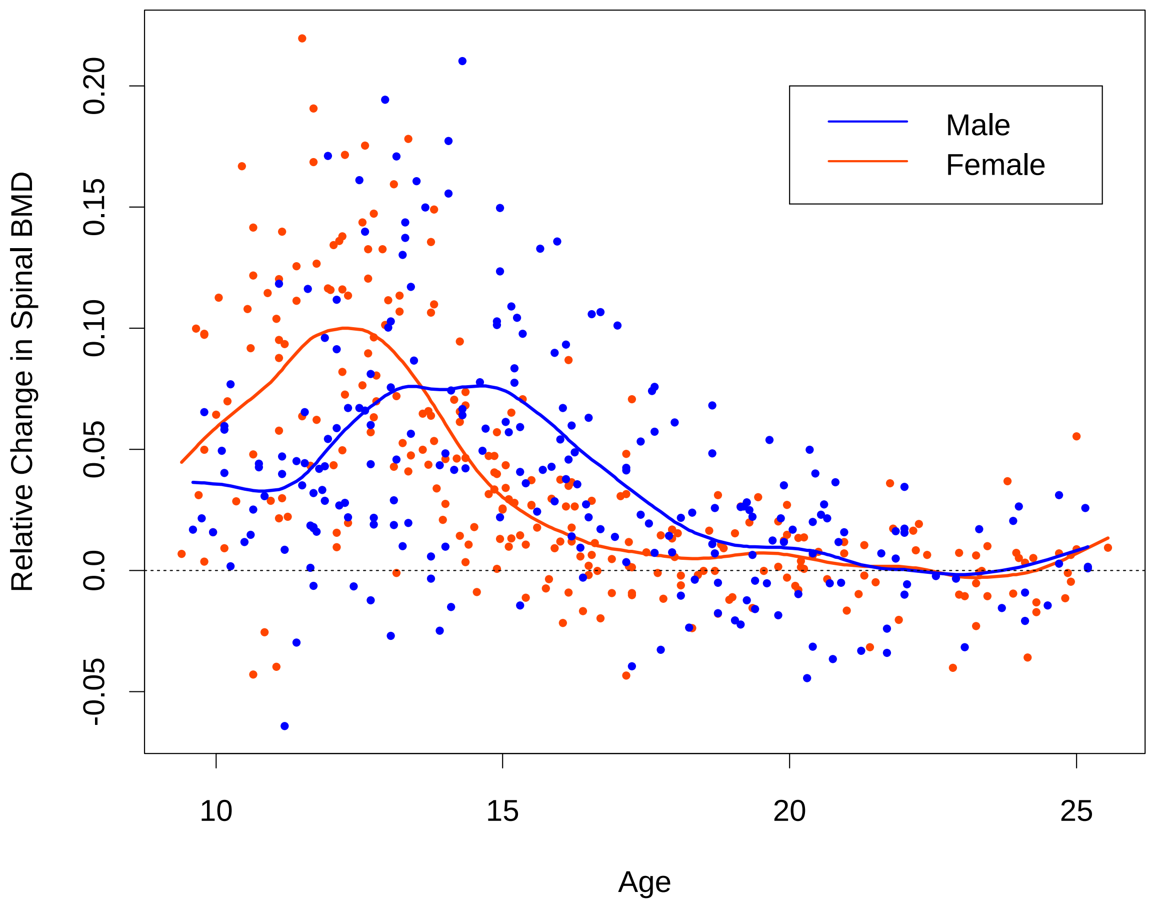
	Raw	Regularized
Train	0.080	0.185
Test	0.255	0.158


Phoneme Examples
aa
ao
Log-periodogram
Frequency
Phoneme Classification: Raw and Restricted Logistic Regression
Logistic Regression Coefficients
Frequency

Smoothing splines

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

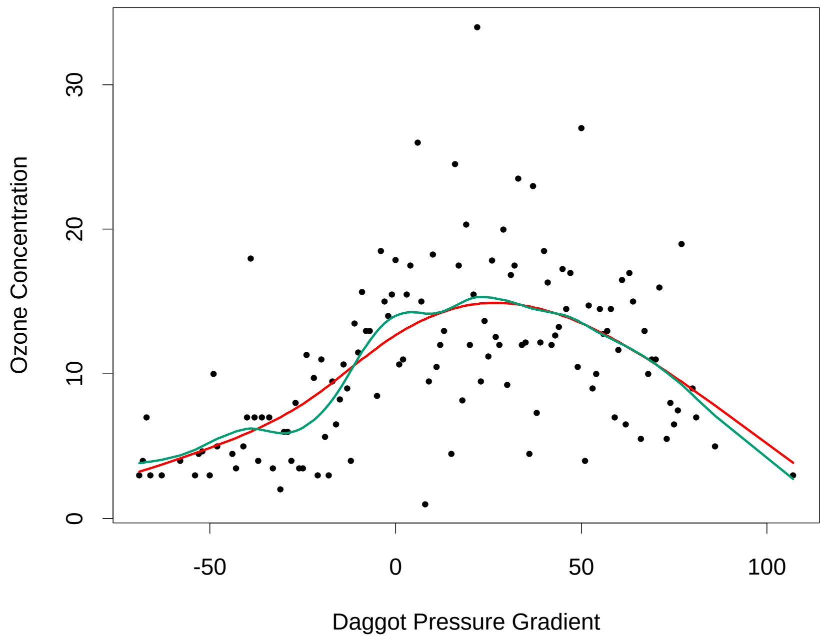
- when $\lambda = 0$ solution interpolates data
- when $\lambda = \infty$ solution is linear least solution line
- in general, $\hat{f}(x) = \sum j = 1^n N_j(x)\theta_j$. This is a natural cubic spline with knots at each of the unique x_i values
- $\mathrm{RSS}(f, \lambda) = (Y - N\theta)^T(Y - N\theta) + \lambda\theta^T\Omega\theta$
- $\{N\}_{ij} = N_j(x_i), \Omega_{ij} = \int N_i^{''}(t)N_j^{''}(t)dt$
- $\hat{\theta} = (N^T N + \lambda\Omega)^{-1}N^T y$

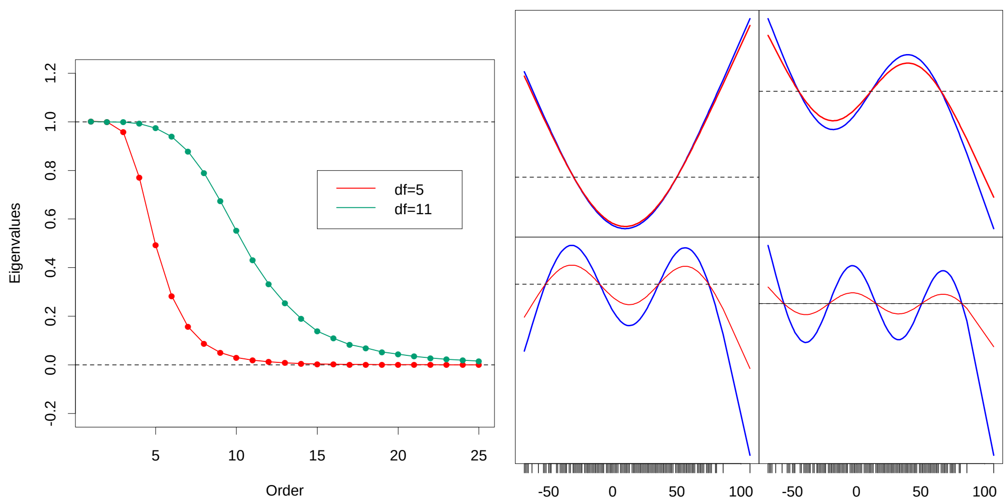

Male
Female
Relative Change in Spinal BMD
Age
-0.05
0.0
0.05
0.10
0.15
0.20
10
15
20
25

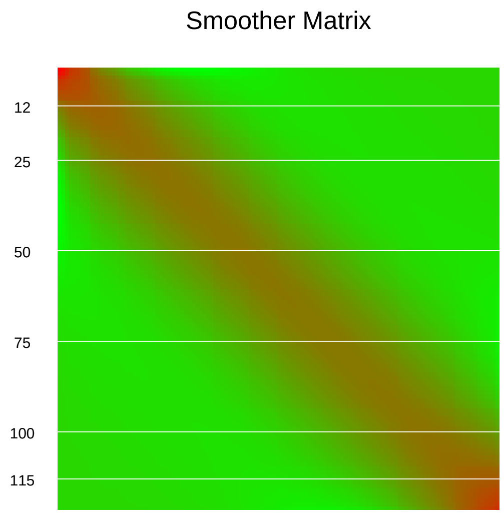
Smoothing matrices

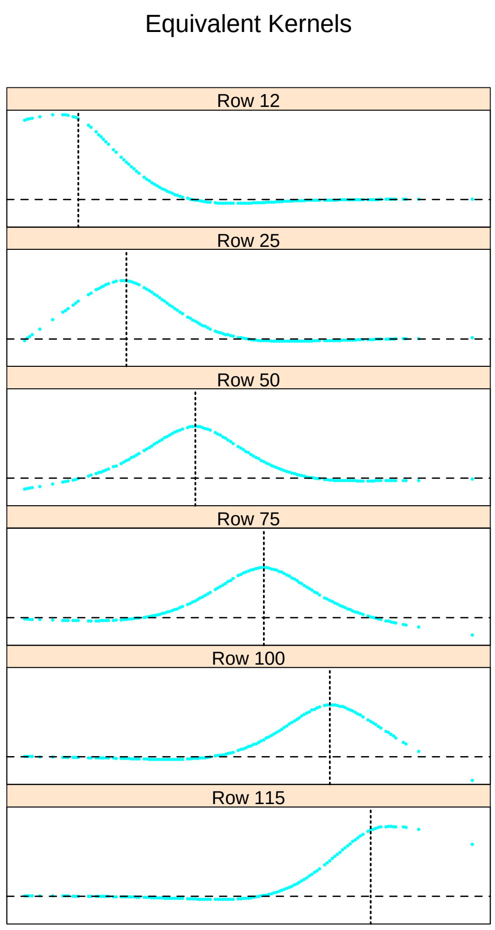
$$\hat{f} = N(N^TN + \lambda\Omega)^{-1}N^Ty \equiv S_\lambda y$$

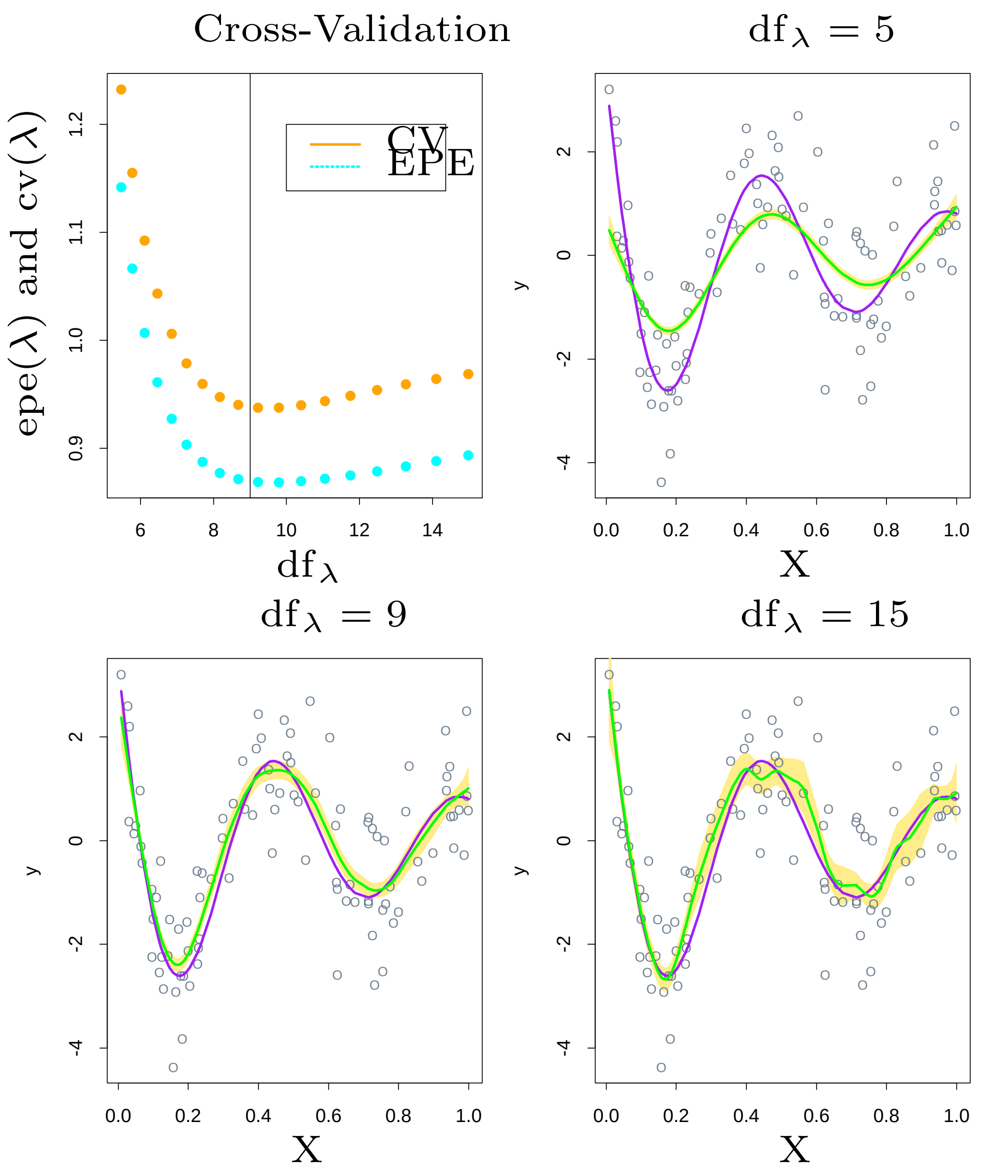
- symmetric
- positive definite
- eigenvalues in $(0, 1]$, rank n.
- $\mathrm{df}(\lambda)$ is defined to be $\mathrm{trace}(S_\lambda)$
- Reinsch form $S_\lambda = (I + \lambda K)^{-1}$; minimizer of $||y - f||^2 + \lambda f^T K f$.
- $S_\lambda = \sum_{k=1}^n p_k(\lambda) u_k u_k^T; p_k(\lambda) = 1/(1 + \lambda d_k)$; $\mathrm{df}(\lambda) = \sum p_k(\lambda)$, d_k is kth eigenvalue of K.


Ozone Concentration
0
10
20
30
-50
0
50
100
Daggot Pressure Gradient


Eigenvalues
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5
10
15
20
25
Order
df=5
df=11
-50
0
50
100
-50
0
50
100


Smoother Matrix
12
25
50
75
100
115


Equivalent Kernels
Row 12
Row 25
Row 50
Row 75
Row 100
Row 115


Cross-Validation
$\mathrm{df}_\lambda = 5$
epe(λ) and cv(λ)
CV
EPE
df_λ
y
X
$\mathrm{df}_\lambda = 9$
$\mathrm{df}_\lambda = 15$

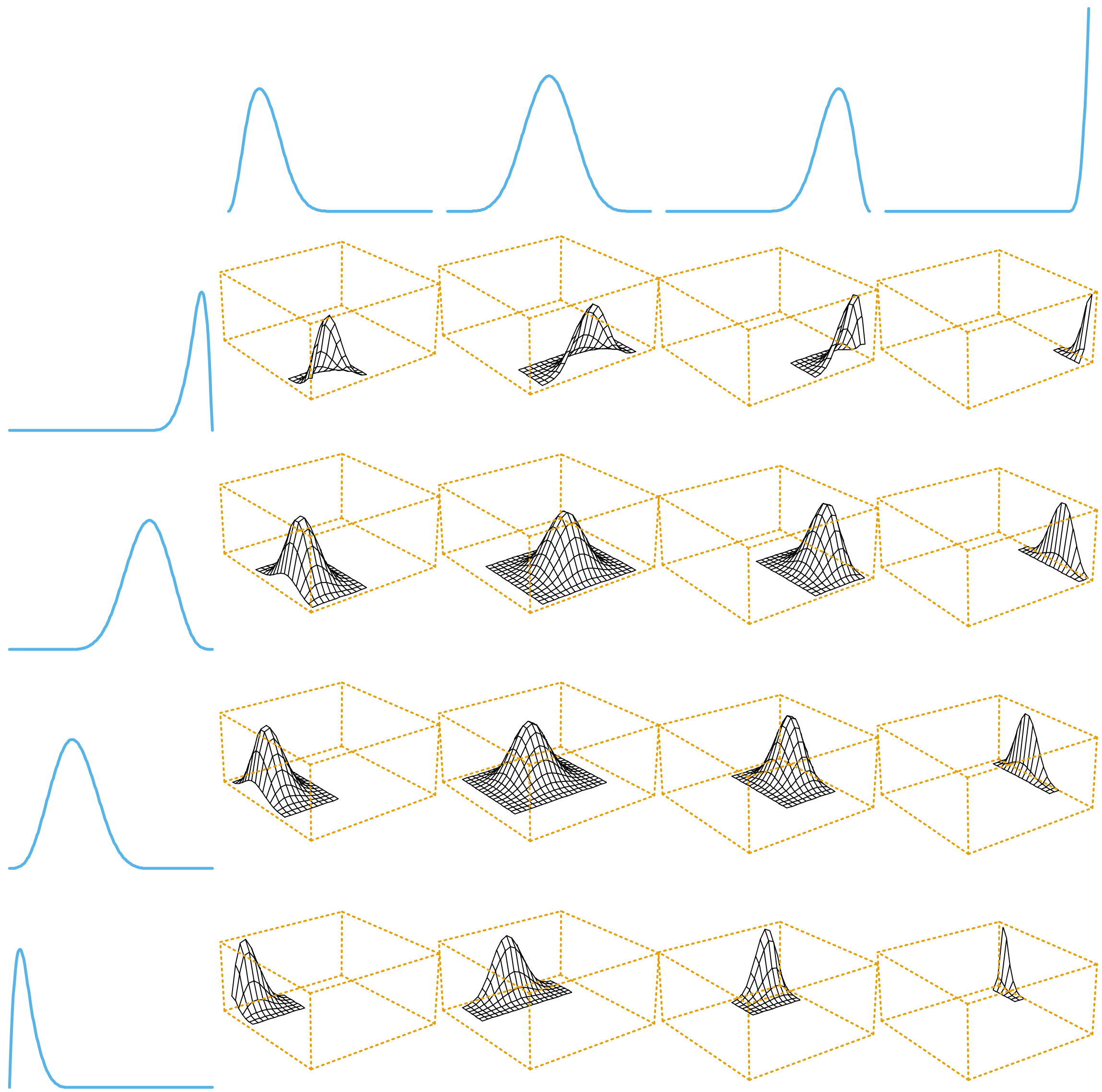
Cross-validation for Smoothing splines

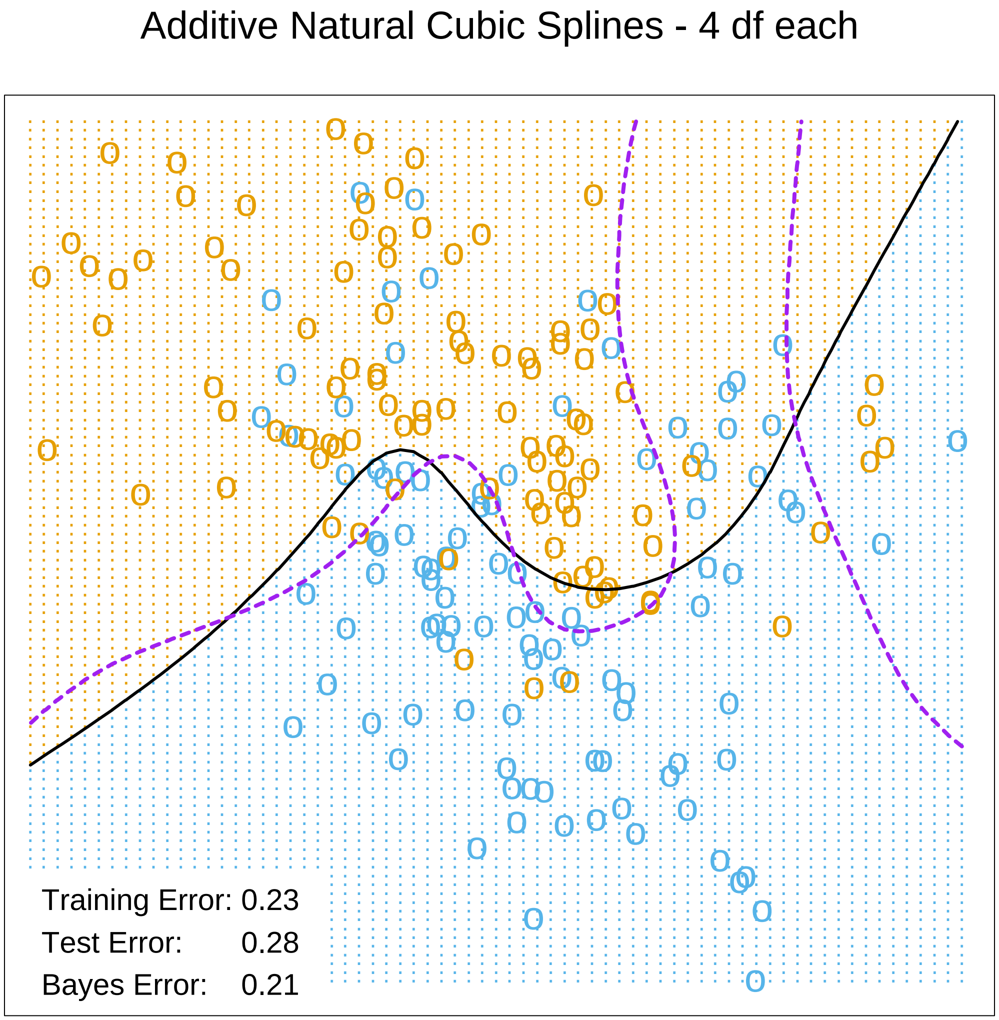
$$\begin{aligned} \mathrm{CV}(\hat{f}_\lambda) &= \sum (y_i - \hat{f}_\lambda^{-i}(x_i))^2 \\ &= \frac{\sum (y_i - \hat{f}_\lambda(x_i))^2}{1 - S_\lambda(i,i))^2} \end{aligned}$$

Smoothing spline logistic regression

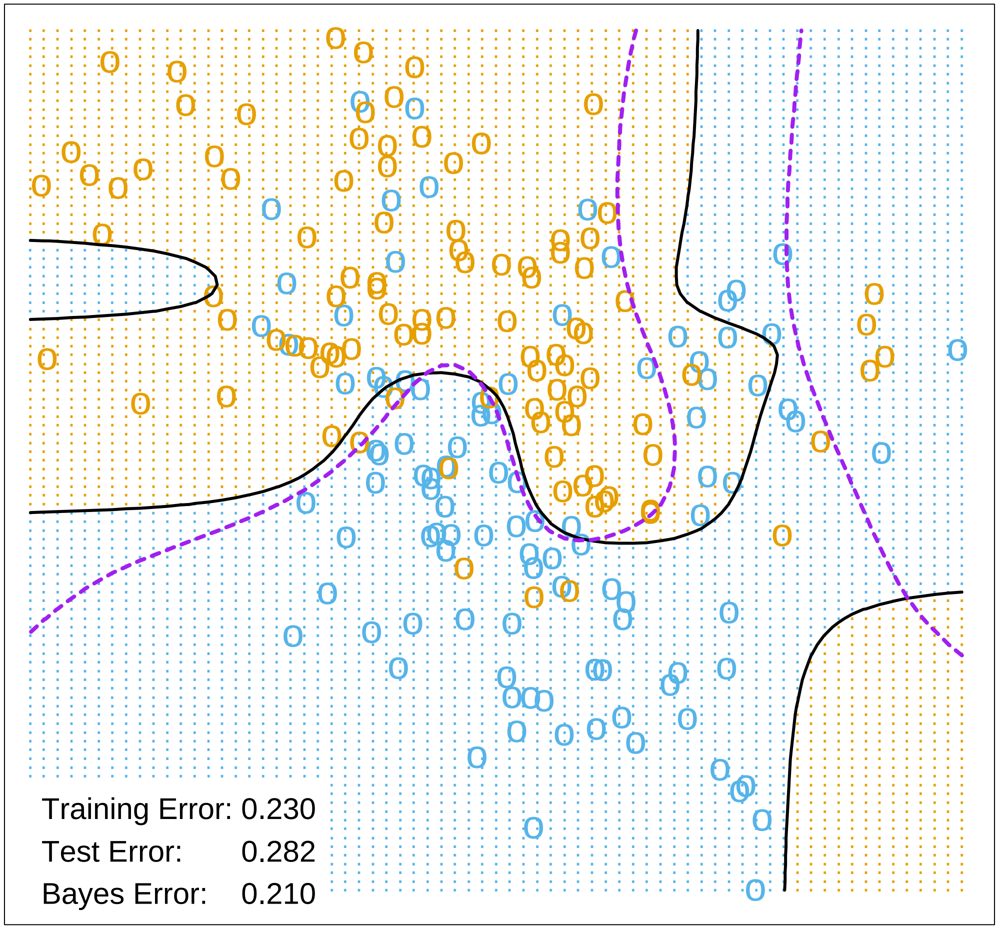
$$\begin{aligned} \Pr(y=1|x) &= \frac{exp(f(x))}{1+exp(f(x))} \\ \ell(f;\lambda) &= \sum_{i=1}^{n} [y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))] - \frac{1}{2}\lambda \int f^{''2}(t)^2 dt \end{aligned}$$

Algorithm $f^{new} \leftarrow S_{\lambda,w}(f^{old} + W^{-1}(y-p))$ where $S_{\lambda,w} = N(N^T W N + \lambda\Omega)^{-1} N^T W$ fits a weighted cubic smoothing spline.



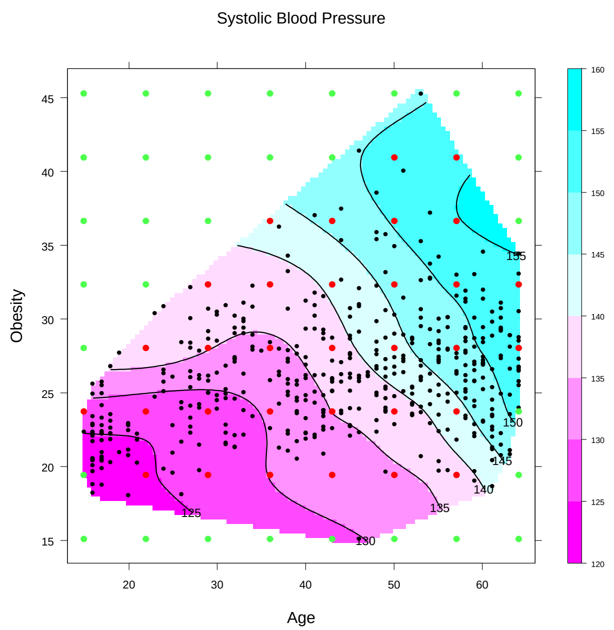

Additive Natural Cubic Splines - 4 df each
Training Error: 0.23
Test Error: 0.28
Bayes Error: 0.21

Natural Cubic Splines - Tensor Product - 4 df each



Thin plate splines

$$J[f] = \int\int_{R^2} \Big[\left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^2 \Big] dx_1 dx_2. \tag{1}$$



$df = 15$; red points are knots

The Kernel property and reproducing kernel Hilbert spaces

- Example: polynomial regression
- Suppose $h(x) : R^p \leftarrow R^M$, M huge
- Given $x_1, x_2, \ldots x_n$ with $M \gg n$, let $H = \{h_j(x_i)\}_{M\times n}$.
- $R(\beta) = (y - H\beta)^T(y - H\beta) + \lambda\beta^T\beta$
- Then

$$
\begin{aligned}
\hat{y} &= H\hat{\beta} \\
-H^T(y - H\hat{\beta}) + \lambda\hat{\beta} &= 0 \\
-HH^T(y - H\hat{\beta}) + \lambda H\hat{\beta} &= 0 \\
H\hat{\beta} &= (HH^T + \lambda I)^{-1}y
\end{aligned}
$$

where HH^T is $n \times n$. $\{HH^T\}_{i,i'} = \langle h(x_i), h(x_{i'})\rangle = K(x_i, x_{i'})$.

$$\hat{f}(x) = h(x)^T\hat{\beta} = \sum \hat{\alpha}_i K(x, x_i)$$

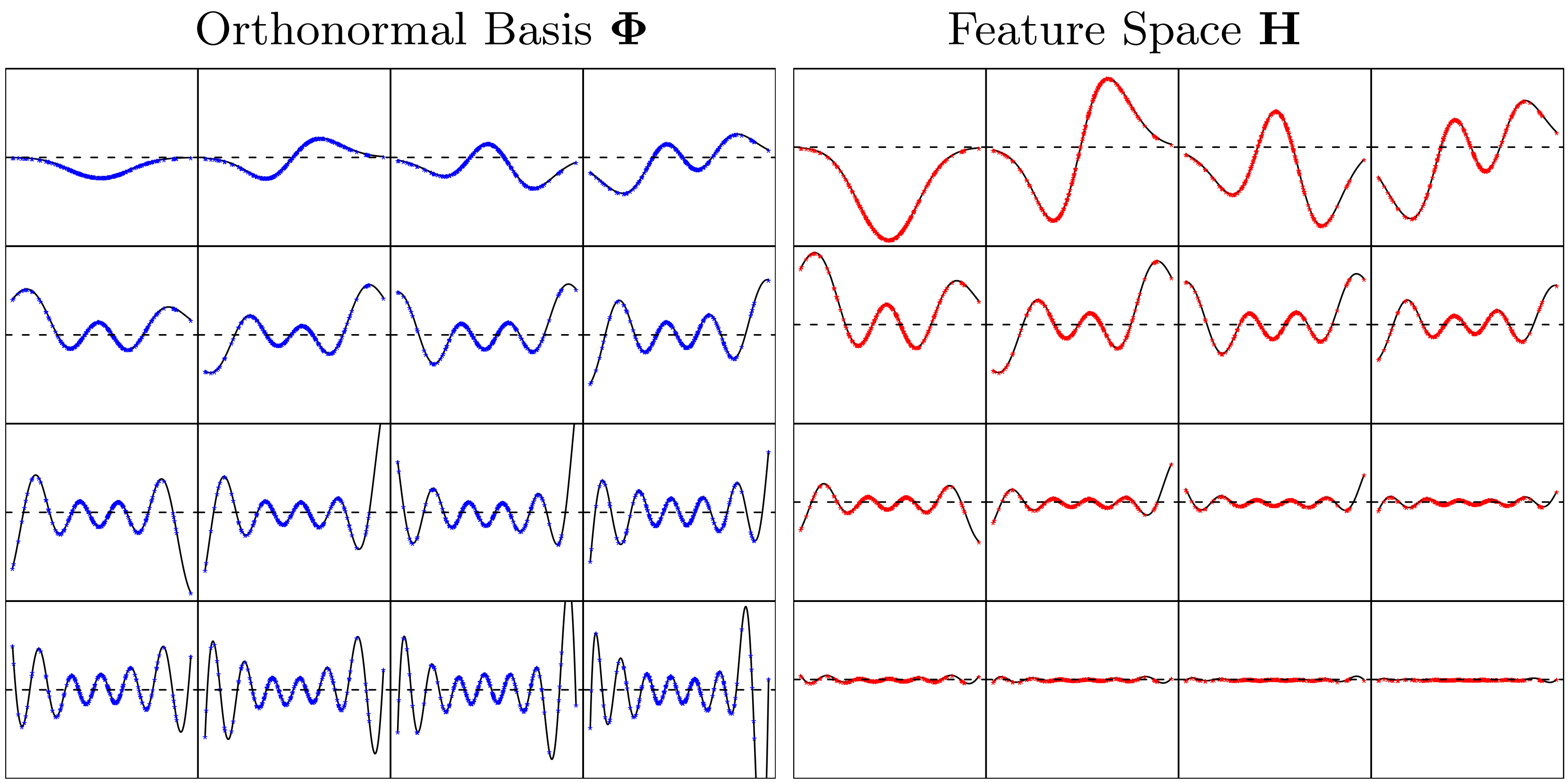
and $\hat{\alpha}_i = (K + \lambda I)^{-1}y$.

Polynomial kernels

- $K(x, x') = (1 + \langle x, x' \rangle)^d$
- e.g. if $x \in R^2$, $d = 2$,

$$
\begin{aligned}
K(x, x') &= (1 + x_1 x_1' + x_2 x_2')^2 \\
&= 1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2'
\end{aligned}
$$

- then $M = 6$ and $h_1(x) = 1, h_2(x) = \sqrt{2}x_1, h_3(x) = \sqrt{2}x_2, h_4(x) = x_1^2, h_5(x) = x_2^2, h_6(x) = \sqrt{2}x_1 x_2$


Orthonormal Basis $\mathbf{\Phi}$
Feature Space $\mathbf{H}$

Reproducing kernel Hilbert spaces

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

$\gamma_i \geq 0$, $\sum \gamma^2 < \infty$.

Definition; $f \in H_K$ if $f(x) = \sum i = 1^{\infty} c_i \phi_i(x)$ with

$$||f||^2_{\mathcal{H}_K} \equiv \sum_{i=1}^{\infty} c_i^2 / \gamma_i < \infty,$$

where $||f||_{\mathcal{H}_K}$ is the norm induced by K.

Rewriting, we have

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

or equivalently

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

It can be shown that

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

which is finite dimensional!

Properties

- $K_i(x) = K(x, x_i)$ "Representer of evaluation at x_i".
- $\langle K(x, x_i), f \rangle_{H_K} = f(x_i)$
- $\langle K(x, x_i), K(x, x_j) \rangle_{H_K} = K(x_i, x_j)$ "Reproducing property"
- $J(\hat{f}) = \sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i x_j)\hat{\alpha}_i, \hat{\alpha}_j$