Basis Expansions and Regularization

For a vector X, we consider models of the form

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

Examples of h_m:

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

Fit by least squares or maximum-likelihood.

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} \beta_m h_m(x_i))^2$$

Often $h_m,\ m = 1, 2, \ldots$ is hierarchical, and m is tuning parameter.

Regularization

Sometimes we use a large expansion

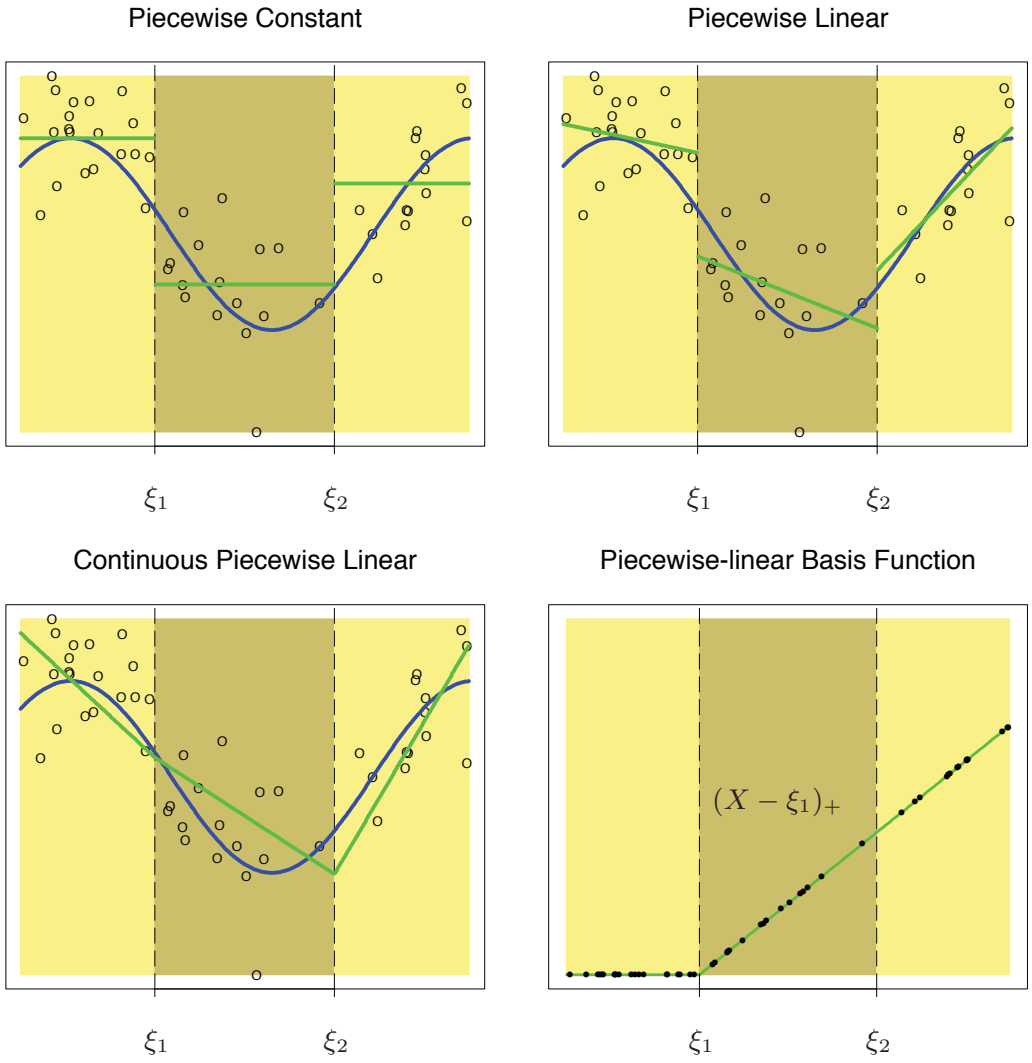
$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

and then control complexity by regularization:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda J(f)$$

$J(f)$ is a *roughness penalty* or other *regularizer*, and λ is tuning parameter.

For example, $J(f) = ||\beta||_2^2$, or $J(f) = \int [f''(t)]^2 dt$.



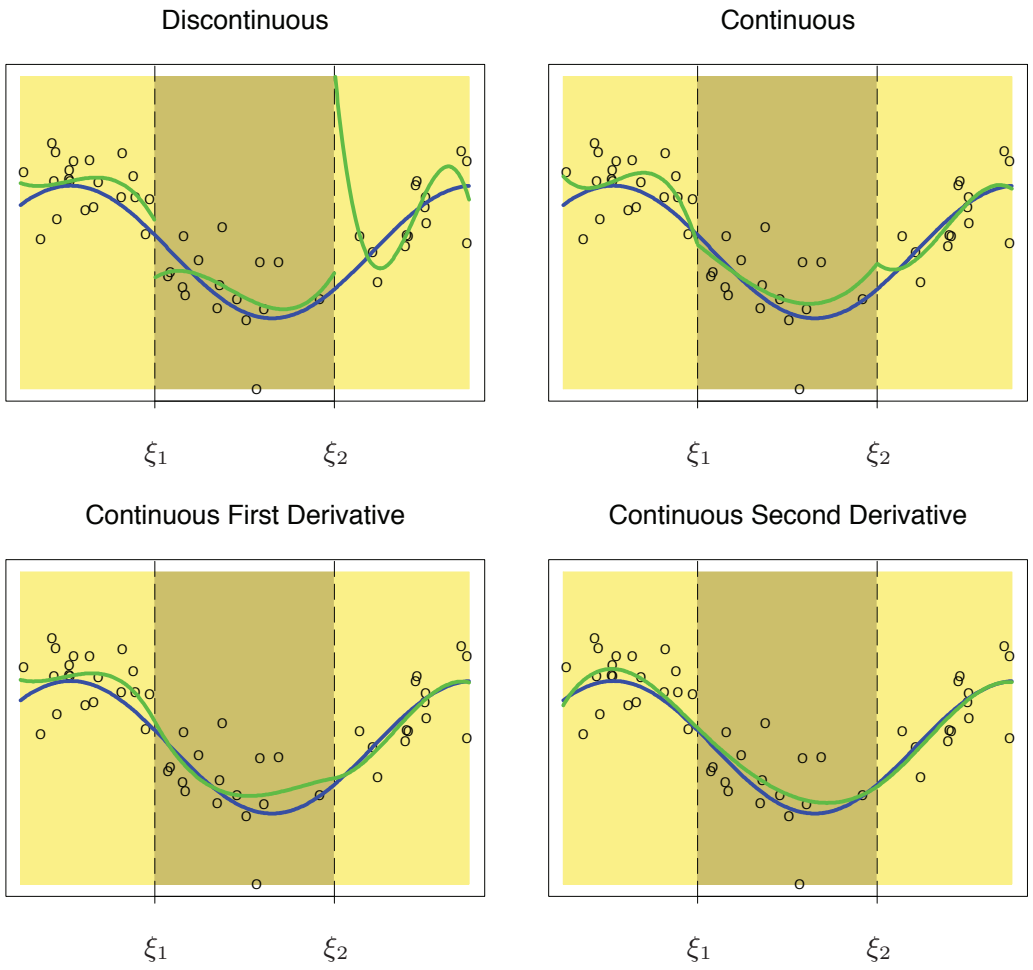
Piecewise Polynomials and Splines

Blue curve is truth.

Top row: piecewise constant and linear between the knots ξ_1 and ξ_2.

Bottom Row: Linear spline, and a piecewise linear basis function.

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X \\ &\quad + \beta_2 (X-\xi_1)_+ \\ &\quad + \beta_3 (X-\xi_2)_+ \end{aligned}$$

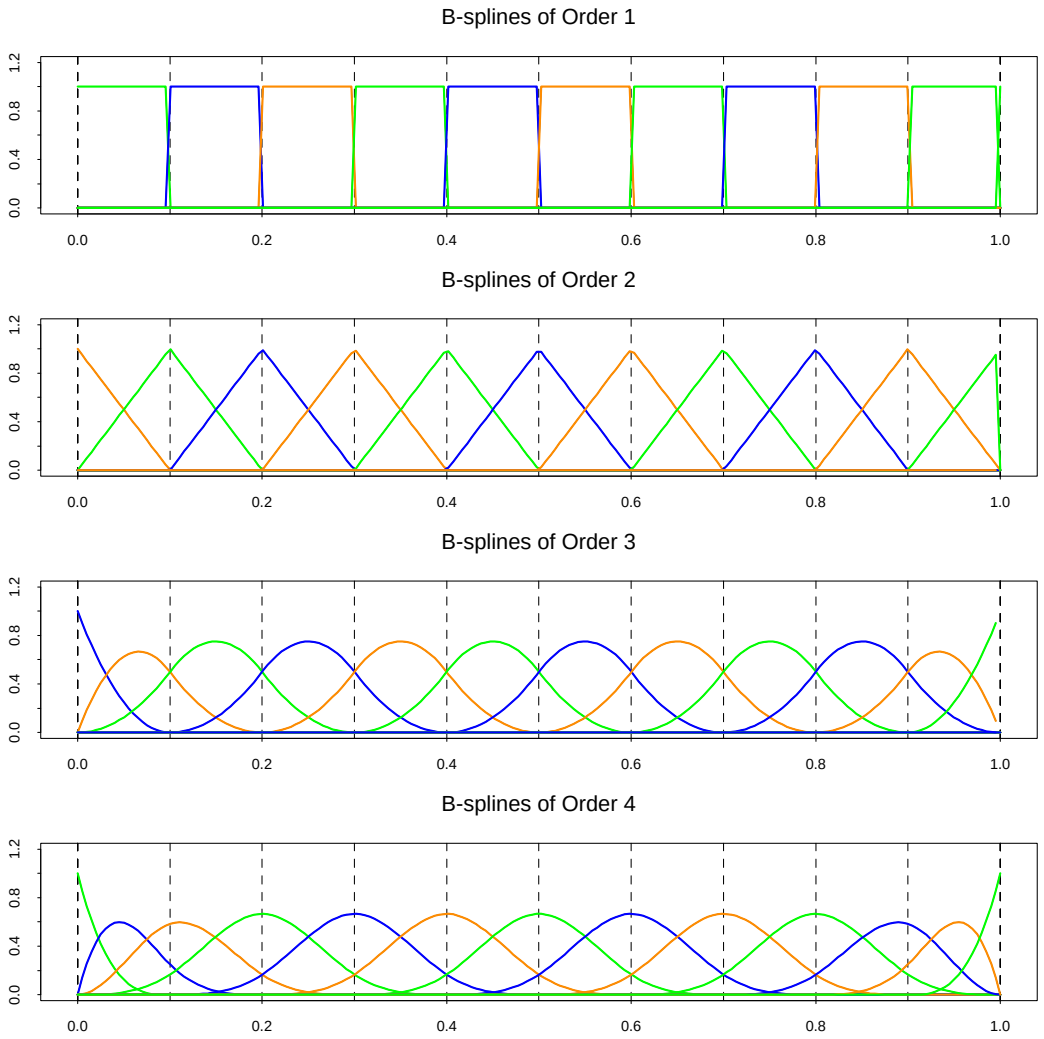


Piecewise Cubics and Splines

Panel shows different orders of continuity.
Bottom right is cubic spline.

$$f(X) = \beta_0 + \sum_{j=1}^{3} \beta_j X^j + \beta_4 (X - \xi_1)_+^3 + \beta_5 (X - \xi_2)_+^3$$

Other bases exist for cubic splines.



B-splines up to order $M = 4$ with evenly spaced knots. B-splines have *local support*, and are nonzero on interval spanned by $M + 1$ knots.

Cubic splines in R

A cubic spline with M knots has $M + 4$ basis functions.
$(M + 1) \times 4$ parameters $-M \times 3$ constraints.

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

Should return a $N \times 7$ matrix.
Instead returns a $N \times 6$ matrix, since argument `intercept=FALSE` is default for use in modelling software.

```
bs(x, df=4)
```

Defaults to cubic, and places knots at uniform quantiles to achieve 4 columns (with intercept removed).

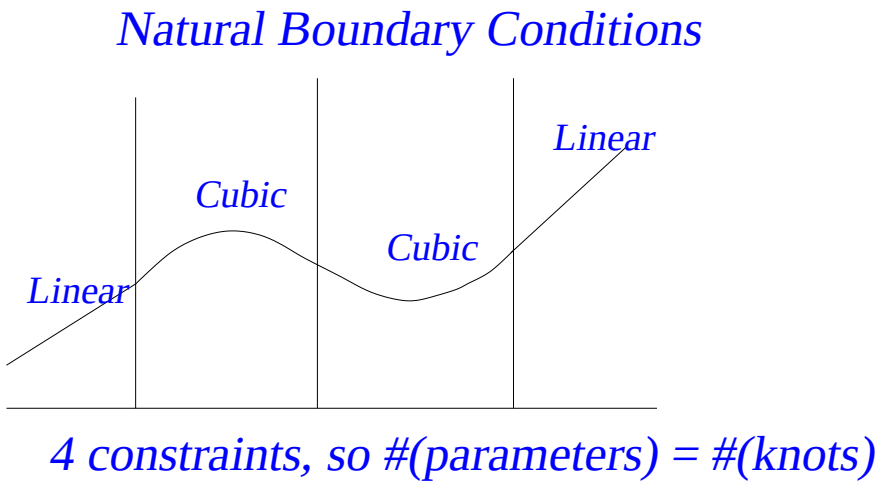
Natural cubic splines in R

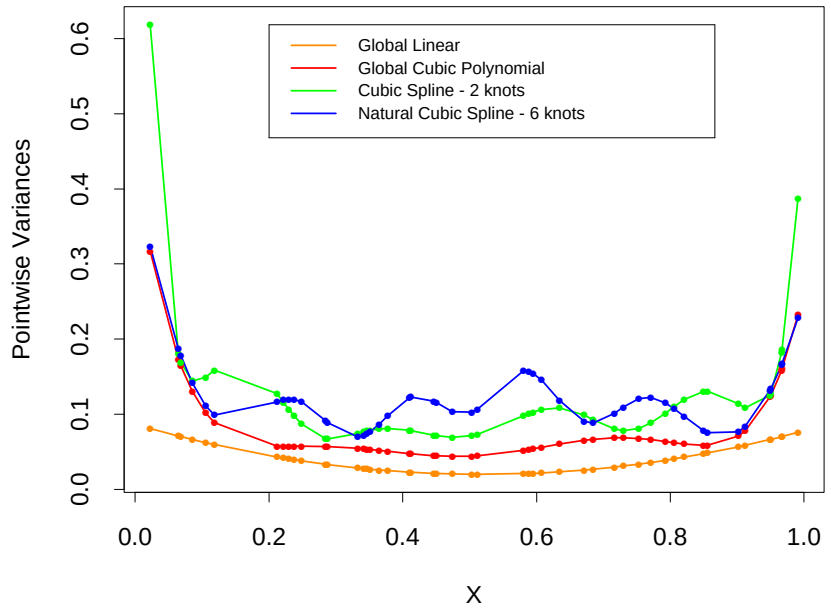
`ns(x, knots=c(.2,.4,.6))`

$N \times 4$ matrix

`intercept=FALSE` is default, and two boundary knots at the extremes of the data.

`ns(x, df=4)`





$$y_i = f(x_i) + \epsilon_i$$
$$\epsilon_i \sim \text{iid}(0, \sigma^2)$$
$$\text{Var}\hat{f}(x) = h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)\sigma^2$$

This is conditional variance, holding the training x_i fixed.

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.

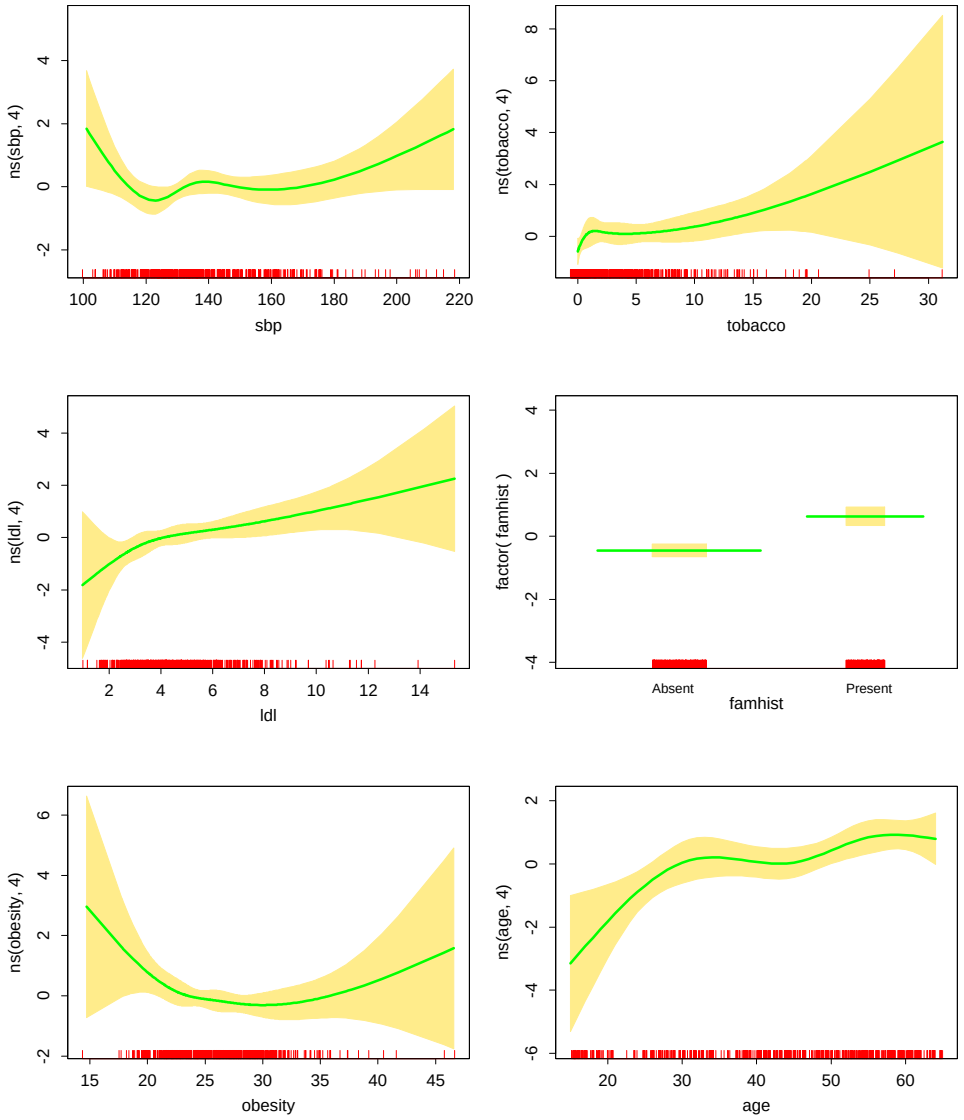
South African Heart Disease data

$$\begin{aligned}\text{logit}\left[\Pr(\text{CHD}|X=x)\right] &= \theta_0 + h_1(x_1)^T\theta_1 + h_2(x_2)^T\theta_2 + \cdots h_p(x_p)^T\theta_p \\ &= h(x)^T\theta\end{aligned}$$

- $h_j(x_{ij}) \leftarrow$ `ns (x[,j], df=4)`
- Basis matrix $\mathbf{H} = \{h_\ell(x_i)\}$, $N \times (1 + \sum_{j=1}^{p} \text{df}_j)$
- $\hat{\theta}$ obtained from binomial maximum likelihood (logistic regression)
- $\hat{\Sigma} = \widehat{\text{Cov}}(\hat{\theta}) = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}$ where
  - $\mathbf{W} = \text{diag}\{\hat{p}_i(1-\hat{p}_i)\}$, $N \times N$ diagonal weight matrix.
- $\hat{f}_j(x_j) = h_j(x_j)^T\hat{\theta}_j$
- $\widehat{\text{Var}}\hat{f}_j(x_j) = h_j(x_j)^T\hat{\Sigma}_{jj}h_j(x_j)$.

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



Fitted Natural Spline Model

Each term has 4 df (except binary `famhist`)

Shown in yellow are pointwise standard-error curves ($\pm 2\times$).

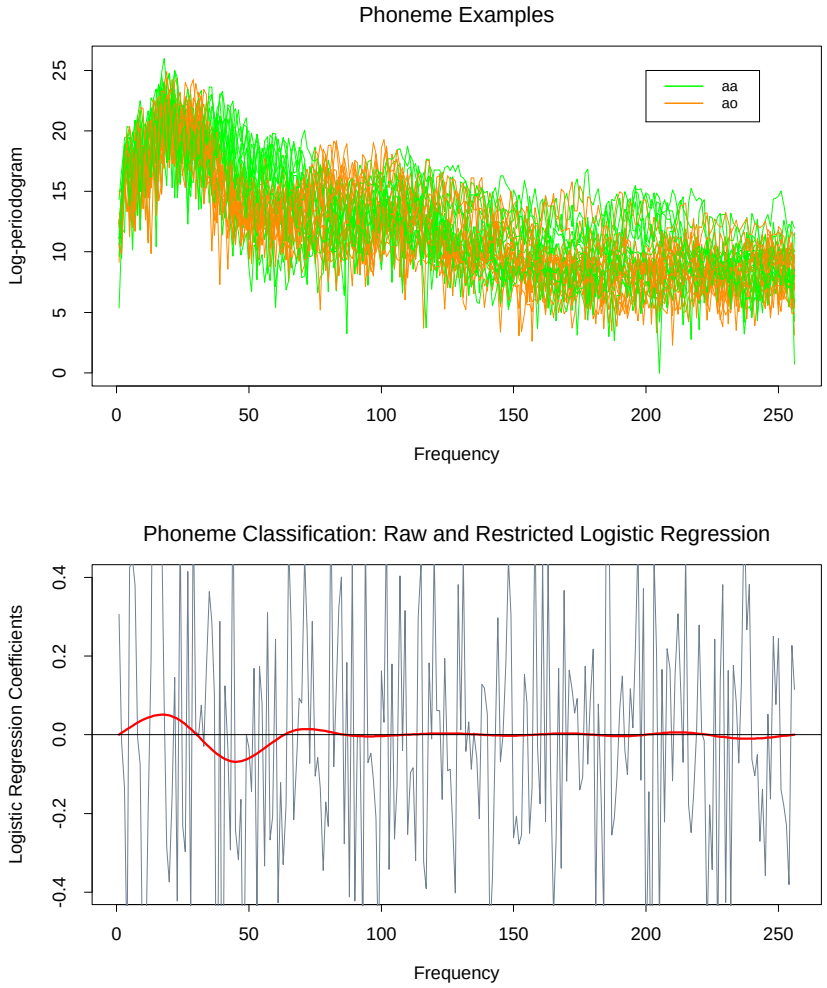
Rugplot in red shows location of x values in sample.

Example: Phoneme recognition

- $X(f)$ observed over grid of frequencies, $x_j = X(f_j)$.
- Two classes aa ($N_1 = 695$) and ao ($N_2 = 1022$)
- $\log \frac{\Pr(aa|X)}{\Pr(ao|X)} = \int X(f)\beta(f)df \approx \sum_{j=1}^{256} X(f_j)\beta(f_j) = \sum_{j=1}^{256} x_j\beta_j$
- $\beta(f) = \sum_{m=1}^{M} h_m(f)\theta_m$

CV Misclassification rates:

	Raw	Regularized
Train	0.080	0.185
Test	0.255	0.158



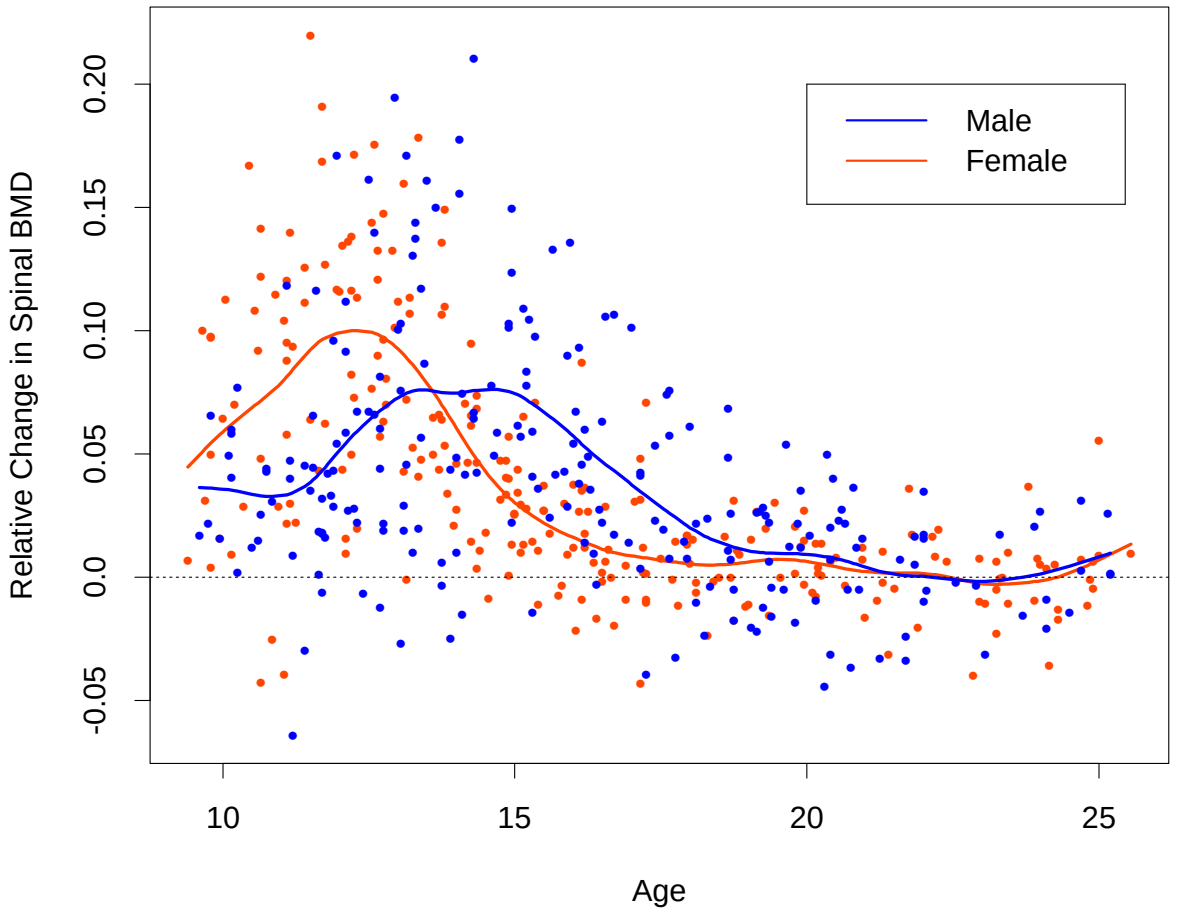
Top: Log-periodogram as a function of frequency for 15 examples each of the phonemes aa and ao sampled from a total of 695 aas and 1022 aos. Each log-periodogram is measured at 256 uniformly spaced frequencies.

Bottom: Coefficients (as a function of frequency) of a logistic regression fit to the data by maximum likelihood, using the 256 log-periodogram values as inputs. The coefficients are restricted to be smooth in the red curve, and are unrestricted in the jagged gray curve.

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-squares solution line.
- in general, $\hat{f}(x) = \sum_{j=1}^{N} h_j(x)\theta_j$. This is a natural cubic spline with knots at each of the unique x_i values (Ex. 5.7 ESL).
- $\mathrm{RSS}(f,\lambda) = (\mathbf{y} - \mathbf{H}\theta)^T(\mathbf{y} - \mathbf{H}\theta) + \lambda\theta^T\mathbf{\Omega}\theta$
- $\{\mathbf{H}\}_{ij} = h_j(x_i),\ \{\mathbf{\Omega}\}_{ij} = \int h_i''(t)h_j''(t)dt$
- $\hat{\theta} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{\Omega})^{-1}\mathbf{H}^T\mathbf{y}$

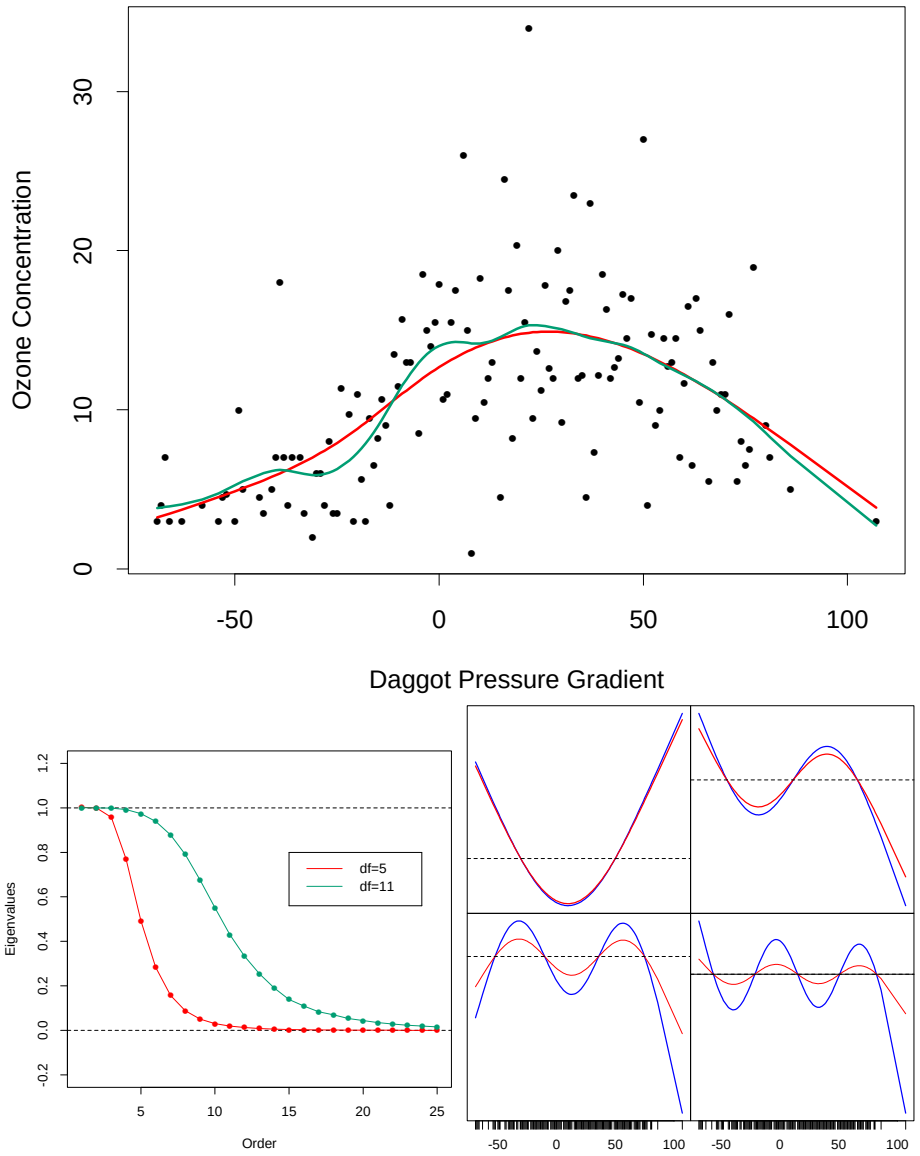


The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with $\lambda \approx 0.00022$. This choice corresponds to about 12 degrees of freedom.

Smoothing matrices

$$\hat{\mathbf{f}} = \mathbf{H}(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{\Omega})^{-1}\mathbf{H}^T\mathbf{y} \equiv \mathbf{S}_\lambda\mathbf{y}$$

- symmetric
- positive definite
- eigenvalues in $(0, 1]$, rank N (or # unique values of x_i).
- $\mathrm{df}(\lambda)$ is defined to be $\mathrm{trace}(\mathbf{S}_\lambda)$.
- Reinsch form $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$; minimizer of $||\mathbf{y} - \mathbf{f}||^2 + \lambda\mathbf{f}^T\mathbf{K}\mathbf{f}$.
- $\mathbf{S}_\lambda = \sum_{k=1}^N p_k(\lambda)\mathbf{u}_k\mathbf{u}_k^T, \quad p_k(\lambda) = 1/(1 + \lambda d_k)$,
  $\mathrm{df}(\lambda) = \sum_{k=1}^N p_k(\lambda), \quad d_k$ is kth eigenvalue of $\mathbf{K}$.

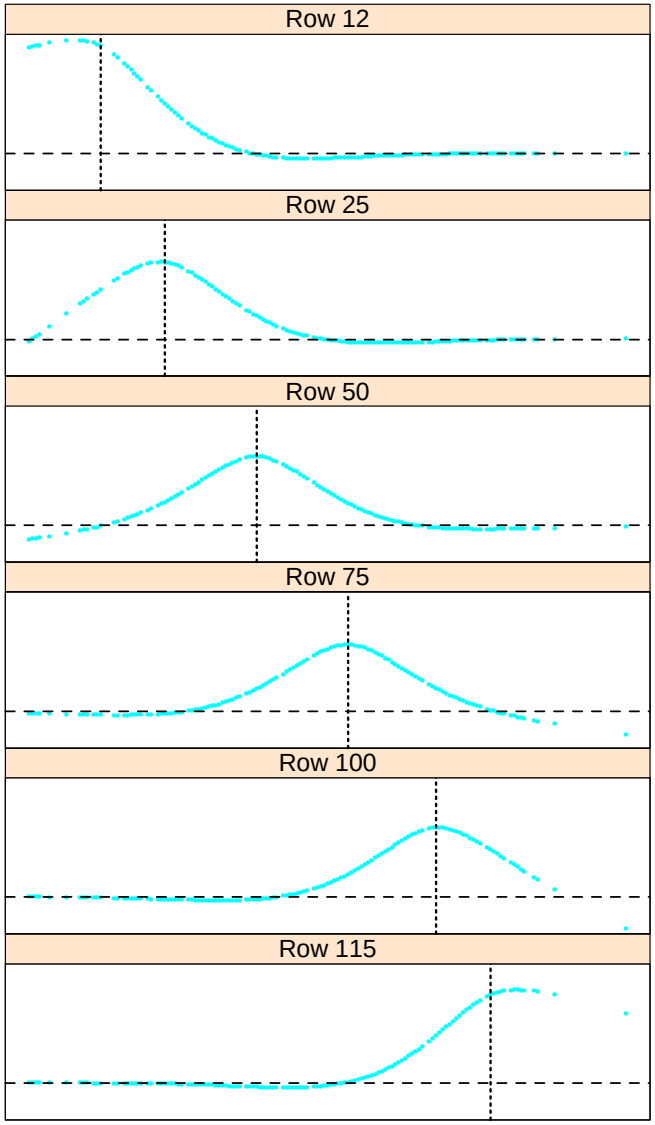


Top: Smoothing spline fit of ozone concentration versus Daggot pressure gradient. $\mathrm{df}_\lambda = 5$ (red) and 11 (green)

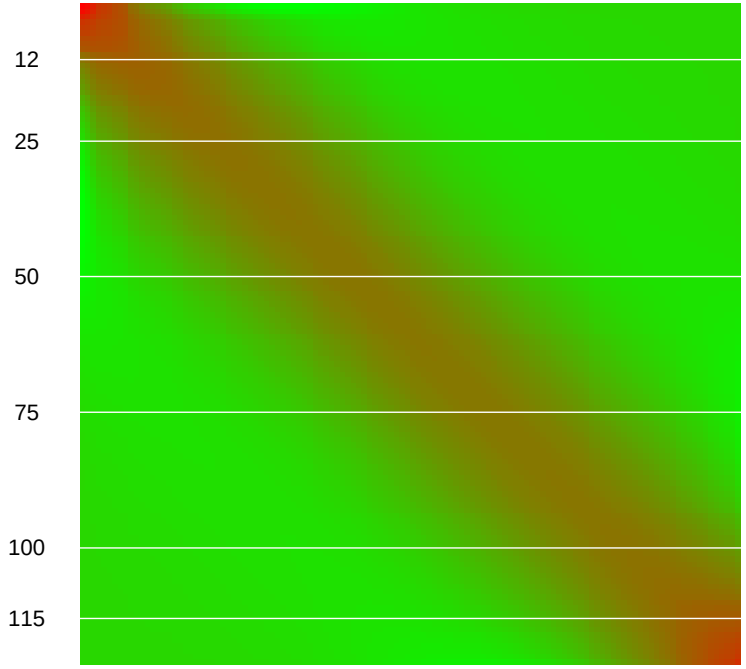
Lower left: First 25 eigenvalues for the two smoothing-spline matrices. The first two are exactly 1, and all are ≥ 0.

Lower right: 3rd to 6th eigenvectors of the spline smoother matrices. $\mathbf{u}_k$ is plotted against $\mathbf{x}$ (viewed as a function of x). The damped functions represent the smoothed versions of these functions (using the 5 df smoother).

Equivalent Kernels

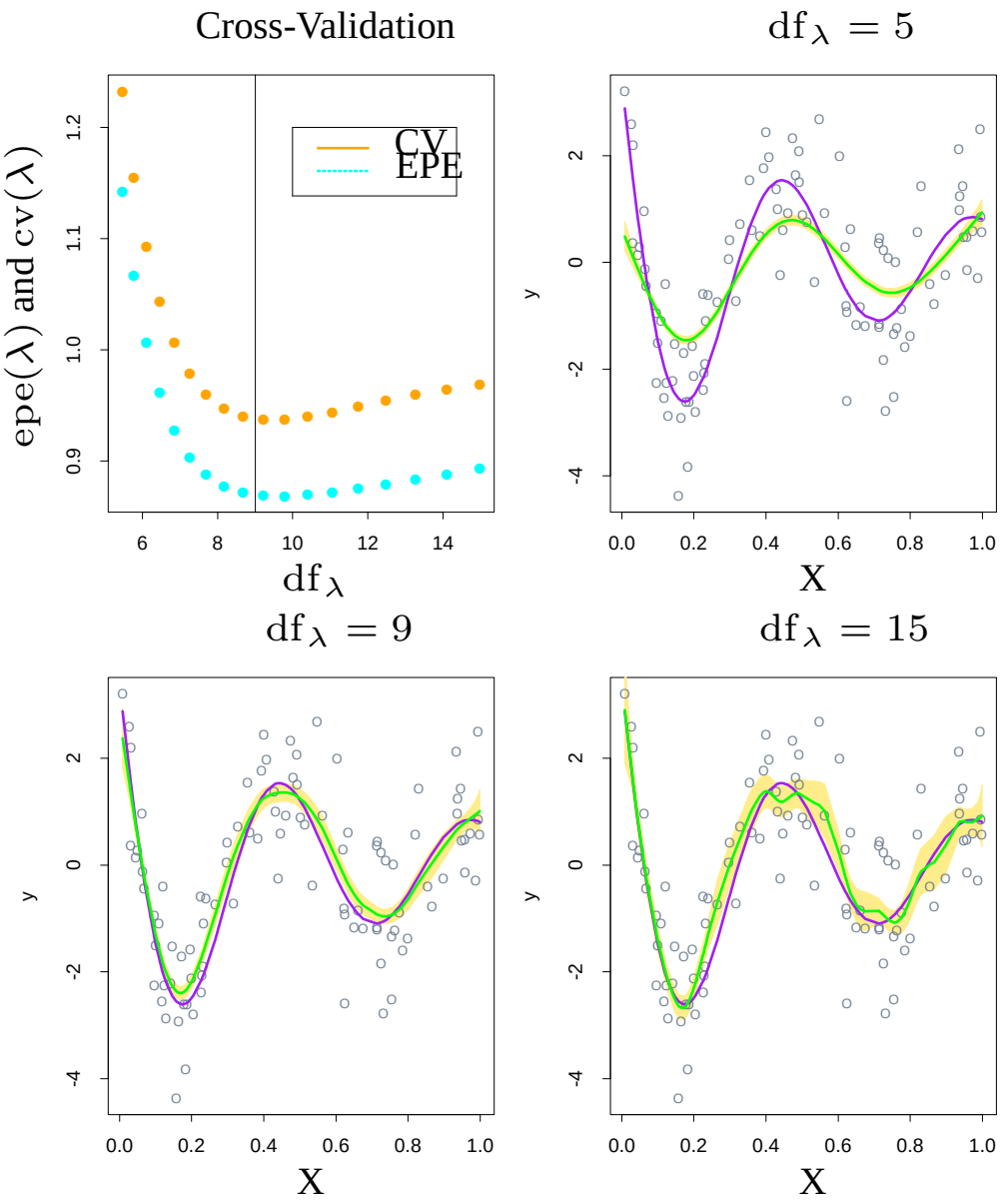


Smoother Matrix



Left: Particular rows of a smoothing spline matrix $\mathbf{S}$, plotted as a function of x. This is the *equivalent kernel* view of the smoothing spline.

Above: Smoother matrix represented as an image.


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

LOO Cross-validation for Smoothing splines

LOO=”leave one out”

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

“Proof”

Consider $\mathcal{T}^{-i}$, and augment it with the point $(x_i, \hat{f}_\lambda^{-i}(x_i))$. The smoothing spline solution on this size-N dataset is $\hat{f}_\lambda^{-i}$. Hence

$$\hat{f}_\lambda^{-i}(x_i) = \sum_{\ell \neq i} S_\lambda(i,\ell) y_\ell + S_\lambda(i,i)\hat{f}_\lambda^{-i}(x_i)$$

Smoothing spline logistic regression

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

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

IRPLS Algorithm

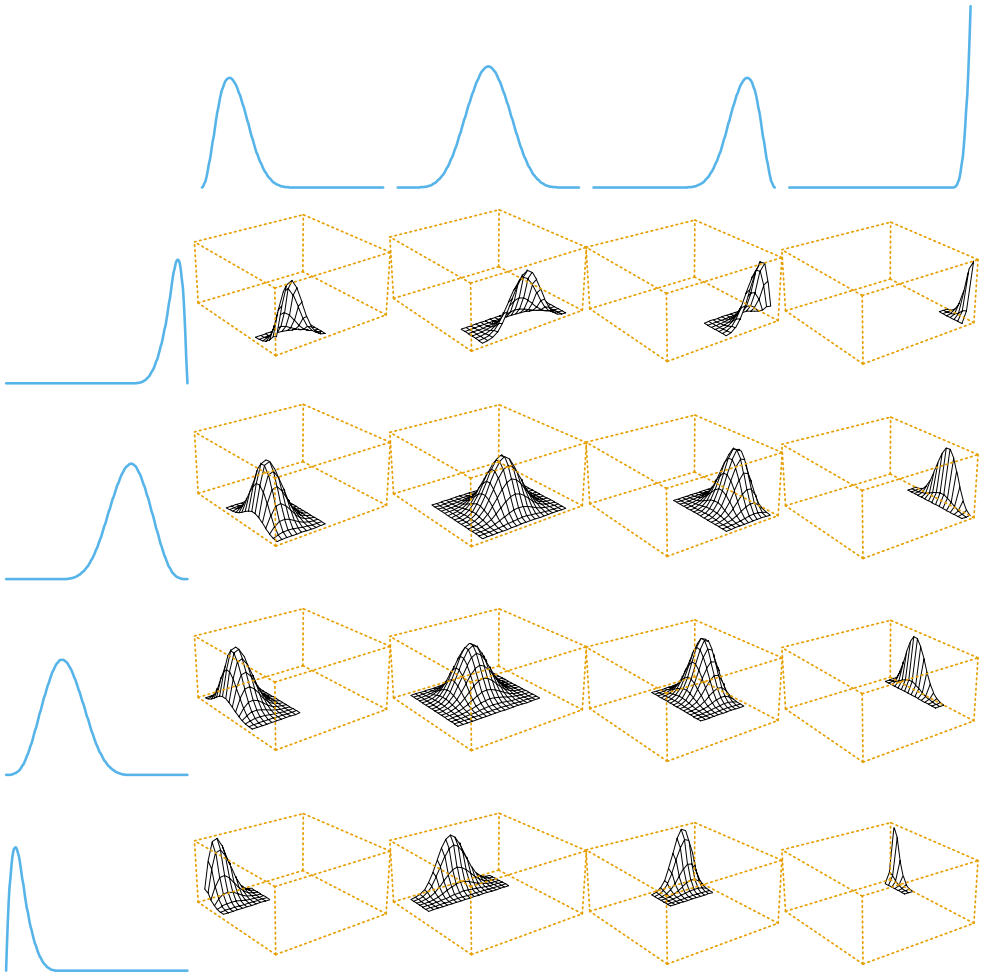
$$\mathbf{f}^{new} \leftarrow \mathbf{S}_{\lambda,w}(\mathbf{f}^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}))$$

where

$$\mathbf{S}_{\lambda,w} = \mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H} + \lambda\mathbf{\Omega})^{-1}\mathbf{H}^T\mathbf{W}$$

fits a weighted cubic smoothing spline.

Tensor-Product Basis



Let $h_j^1(X_1),\ j \in \mathcal{J}$ be a basis for $X_1 \in \mathbb{R}$ of size $|\mathcal{J}|$.

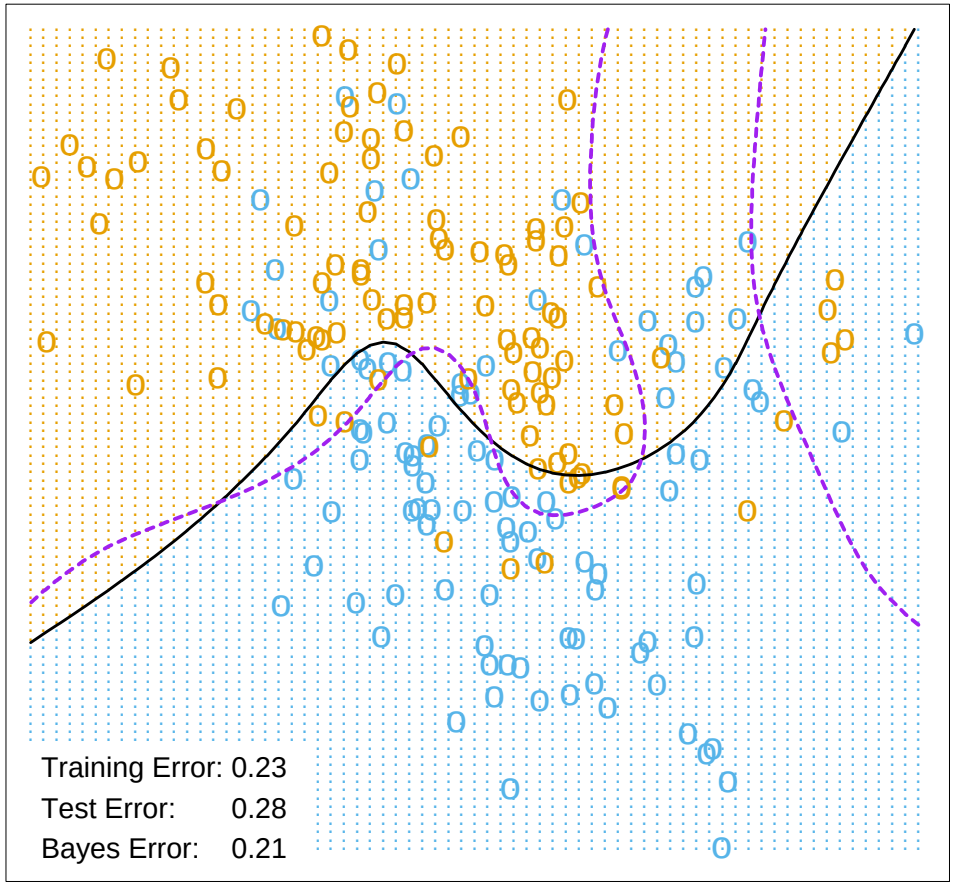
Let $h_k^2(X_2),\ k \in \mathcal{K}$ be a basis for $X_2 \in \mathbb{R}$ of size $|\mathcal{K}|$.

Then

$$h_{jk}^{12}(X_1, X_2) = h_j^1(X_1) h_k^2(X_2)$$

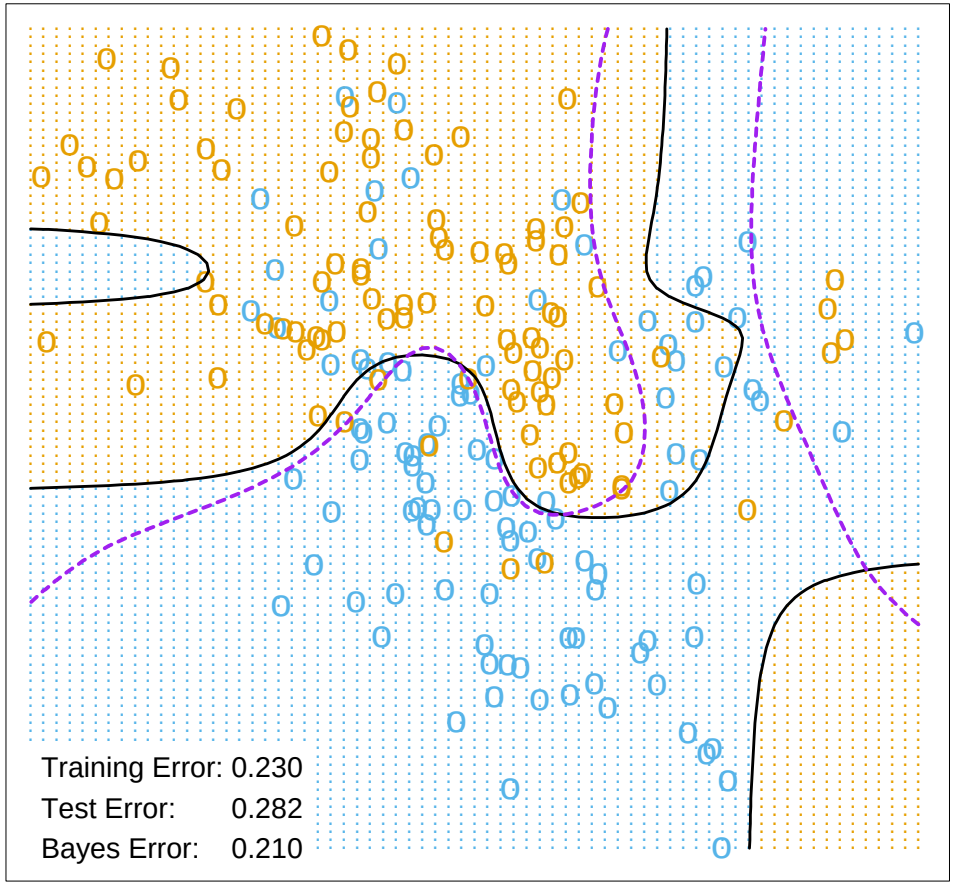
for $i, j \in \mathcal{J} \times \mathcal{K}$ is a *tensor-product* basis over the joint space $(X_1, X_2) \in \mathbb{R}^2$, of size $|\mathcal{J}| \cdot |\mathcal{K}|$.

Additive Natural Cubic Splines - 4 df each

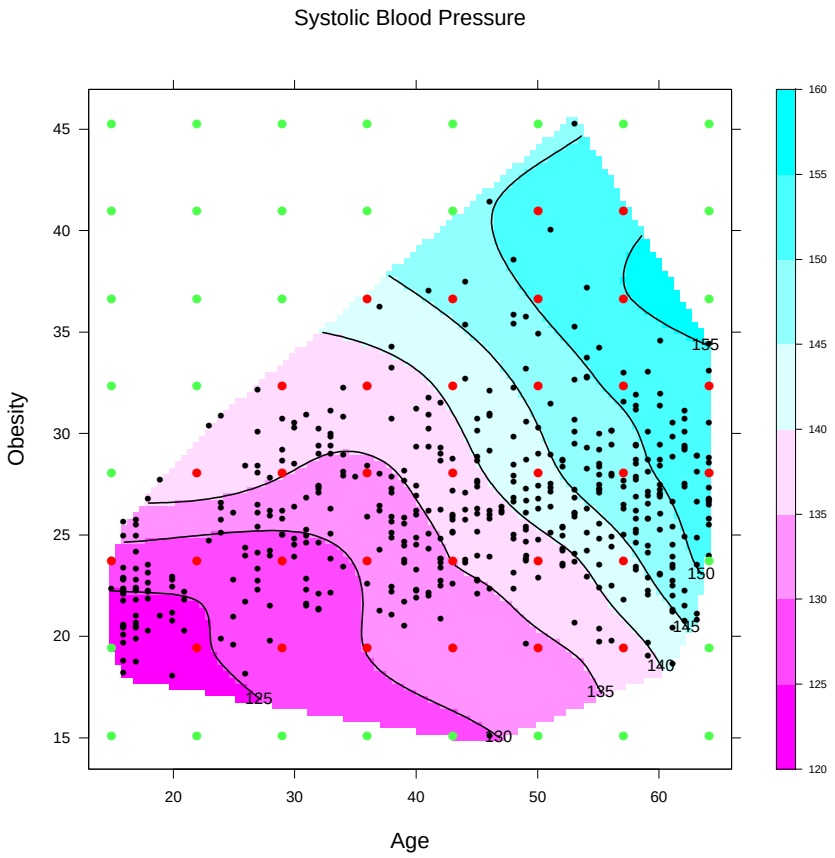


$$df = 1 + (4 - 1) + (4 - 1) = 7$$

Natural Cubic Splines - Tensor Product - 4 df each



$$df = 4 \times 4 = 16$$



Thin plate splines

$$df = 15$$

Red points are knots

$$\max_f \text{loglik}(\mathbf{y}, f) - \frac{1}{2} J(f)$$

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

The "Kernel Property" and Reproducing-Kernel Hilbert Spaces (RKHS)

Example: polynomial ridge regression. Suppose $h(x) : \mathbb{R}^p \mapsto \mathbb{R}^M$ with M huge, and our model is $f(x) = h(x)^T\beta$. Given $x_1, x_2, \ldots x_N$ with $M \gg N$, construct the *wide* basis matrix $\mathbf{H} = \{h_j(x_i)\}_{N\times M}$.

- Objective: $R(\beta) = (\mathbf{y} - \mathbf{H}\beta)^T(\mathbf{y} - \mathbf{H}\beta) + \lambda\beta^T\beta$
- Solution is given by

$$\hat{\mathbf{f}} = \mathbf{H}(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I}_M)^{-1}\mathbf{H}^T\mathbf{y}$$

  The matrix in the "middle" is $M \times M$!

  Can write solution as

$$\hat{\mathbf{f}} = \mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T + \lambda\mathbf{I}_N)^{-1}\mathbf{y}$$

  $N \times N$ system — the kernel trick!

Proof

$$\frac{\partial R}{\partial \beta} = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\beta) + 2\lambda\beta = 0$$

This implies that $\beta = \mathbf{H}^T\alpha$ for some α. Plugging in, and premultiplying by $\mathbf{H}$, we get

$$\mathbf{H}\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{H}^T\alpha) = \lambda\mathbf{H}\mathbf{H}^T\alpha.$$

This gives

$$\hat{\alpha} = (\mathbf{H}\mathbf{H}^T + \lambda\mathbf{I}_N)^{-1}\mathbf{y}$$

Hence

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

In previous example $\mathbf{H}\mathbf{H}^T$ is $N \times N$, and is the matrix of all inner-products $\langle h(x_i), h(x_{i'})\rangle$. Suppose we have a bivariate function $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}_+$ such that $K(x, x') = \langle h(x), h(x')\rangle$. Let $\mathbf{K} = \mathbf{H}\mathbf{H}^T$.

Then

$$\hat{f}(x) = h(x)^T\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i)$$

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

Note: $\hat{\beta}^T\hat{\beta} = \hat{\alpha}^T\mathbf{K}\hat{\alpha}$.

Example: Polynomial kernel

- $K(x, x') = (1 + \langle x, x' \rangle)^d$
- 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 $h_1(x) = 1,\ h_2(x) = \sqrt{2}x_1,\ h_3(x) = \sqrt{2}x_2,$ $h_4(x) = x_1^2,\ h_5(x) = x_2^2,\ h_6(x) = \sqrt{2}x_1 x_2$
- More generally, a basis for degree-d polynomial regression in $\mathbb{R}^p$, with dimension $M = \binom{p+d}{d}$.

Radial Kernels

- The *radial kernel* is defined

$$K(x, x_i) = e^{-\gamma ||x - x_i||^2}.$$

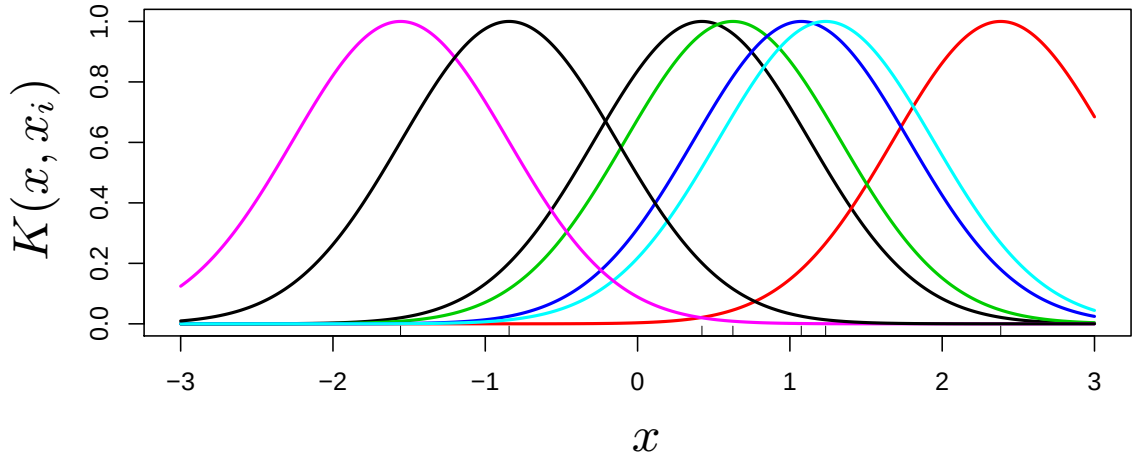
  Also known as a radial basis function (RBF).

- The *implicit basis* $h(x)$ for a radial kernel is in theory infinite dimensional, and in practice very high dimensional.

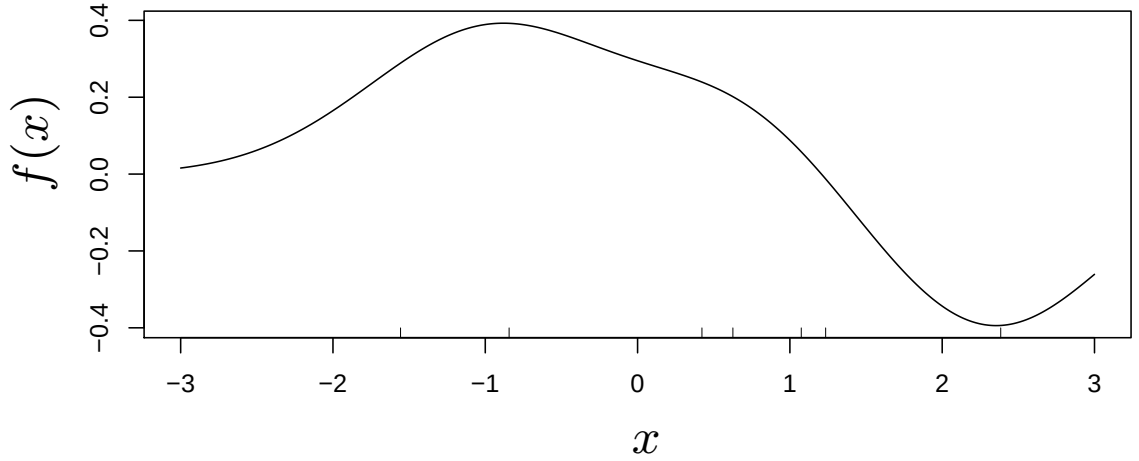
- The kernel parameter γ controls the width of the kernel.

More generally consider space of functions generated by $\{K(\cdot, z),\ z \in \mathbb{R}^p)\}$.

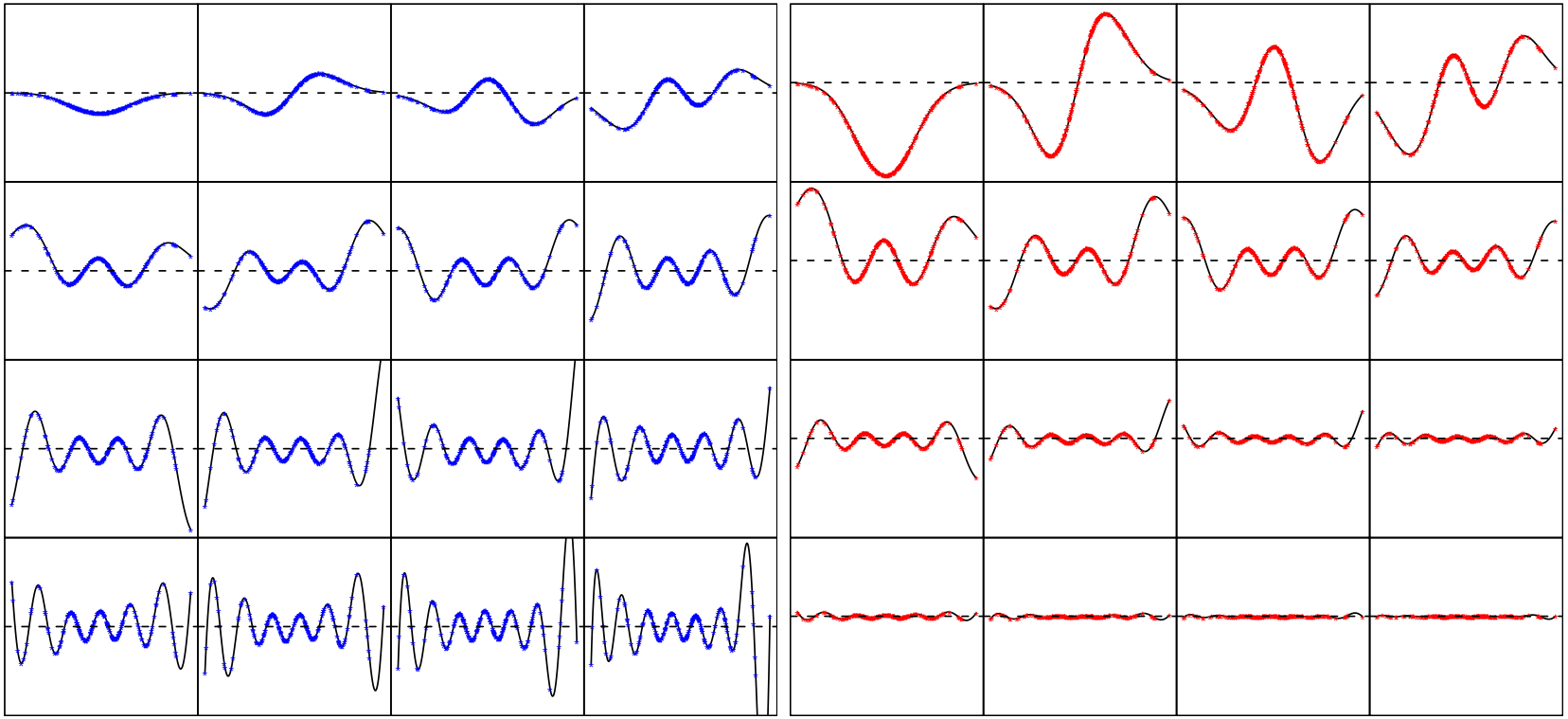
Radial Basis Functions $K(x, x') = e^{-\gamma||x-x'||^2}$



$f(x) = \alpha_0 + \sum_i \alpha_i K(x, x_i)$



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



Reproducing Kernel Hilbert Spaces

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

$\gamma_i \geq 0, \sum_{i=1}^{\infty} \gamma_i^2 < \infty.$

Definition:

$f \in \mathcal{H}_K$ if $f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$

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

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

We now solve

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

$\mathbf{K} = \mathbf{H}\mathbf{H}^T = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$

$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ (with $N > p$)

$\mathbf{f} = \mathbf{H}\beta = \mathbf{U}\mathbf{D}\mathbf{V}^T\beta = \mathbf{U}c$

$f(x) = h(x)^T\beta = u(x)^T c,$

hence $u(x) = \mathbf{D}^{-1}\mathbf{V}^T h(x)$

$$||\mathbf{f}||^2_K = \sum_{i=1}^{N} \frac{c_i^2}{d_i^2} = ||\mathbf{V}^T\beta||^2 = ||\beta||^2$$

or $||\mathbf{f}||^2_K = \mathbf{f}^T\mathbf{K}^-\mathbf{f}$

Here $\mathbf{K}^- = \mathbf{U}\mathbf{D}^{-2}\mathbf{U}^T$

$$\min_{\beta} \left[\sum_{i=1}^{N} L(y_i, h(x_i)^T\beta) + \lambda ||\beta||^2 \right]$$

or equivalently

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

$$\min_{\{c_j\}_1^N} \left[\sum_{i=1}^N L(y_i, \sum_{j=1}^N c_j u_j(x_i)) + \lambda \sum_{j=1}^N \frac{c_j^2}{d_j^2} \right]$$

It can be shown that

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

which is finite dimensional!

Properties

- $k_i(x) = K(x, x_i)$ is called the *Representer of evaluation at x_i:* for $f \in \mathcal{H}_K$ we have $\langle k_i, f\rangle_{\mathcal{H}_K} = f(x_i)$
- $\langle k_i, k_j\rangle_{\mathcal{H}_K} = \langle K(x, x_i), K(x, x_j)\rangle_{\mathcal{H}_K} = K(x_i, x_j)$ —*Reproducing property*
  $J(\hat{f}) = \sum_{i=1}^N \sum_{j=1}^N K(x_i, x_j)\hat{\alpha}_i, \hat{\alpha}_j$

The optimization problem is now

$$\min_{\alpha} L(\mathbf{y}, \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha},$$

where $\mathbf{K} = \{K(x_i, x_j)\}$

Penalized regressions in infinite-dimensional spaces.