Basis Expansions and Regularization

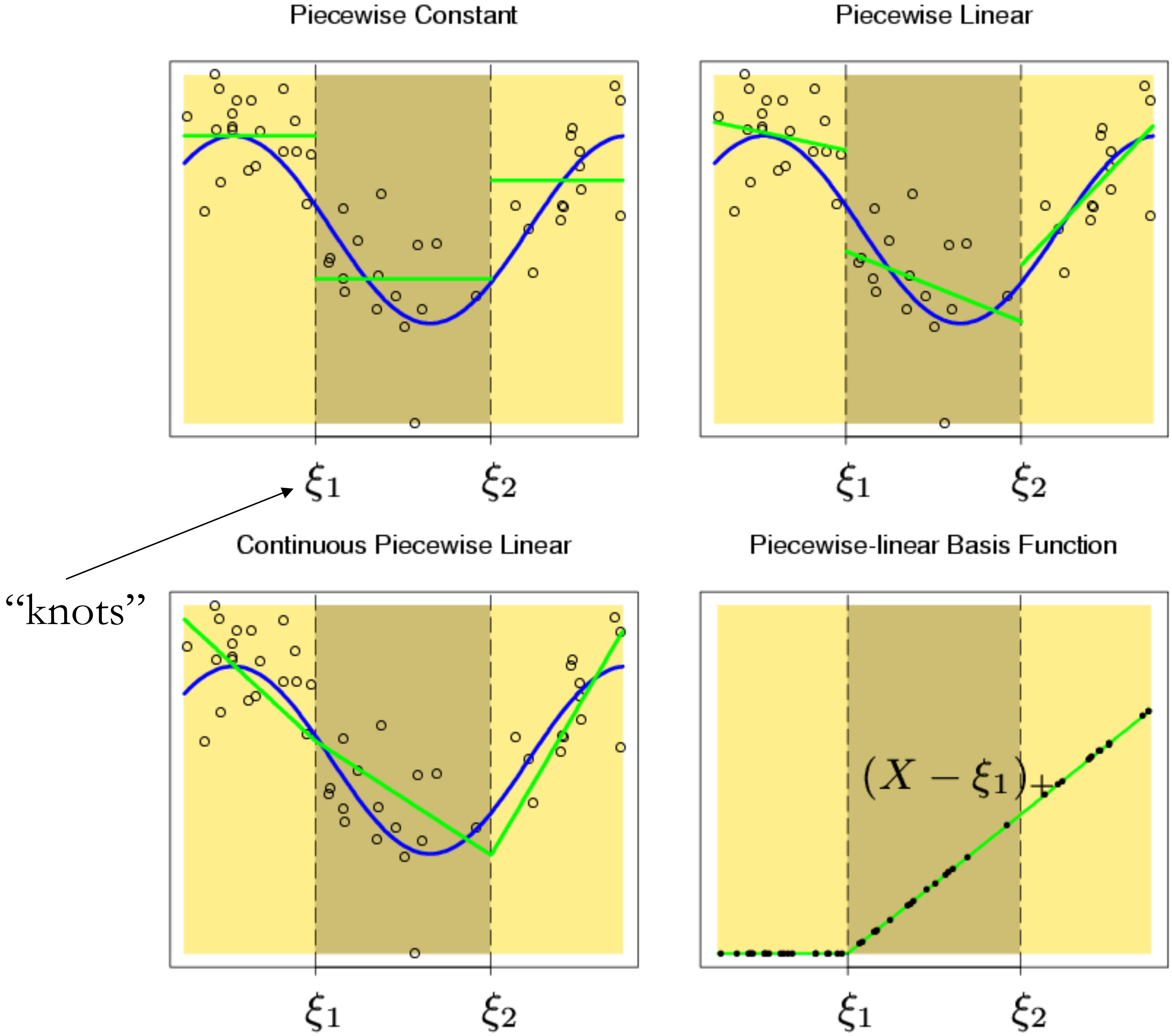
Based on Chapter 5 of
Hastie, Tibshirani and Friedman
(Prepared by J. Cabrera and David Madigan)

Basis Expansions for Linear Models

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

Here the h_m's might be:

- $h_m(X)=X_m$, $m=1,\ldots,p$ recovers the original model
- $h_m(X)=X_j^2$ or $h_m(X)= X_jX_k$
- $h_m(X)=\mathrm{I}(L_m \leq X_k \leq U_m)$,
- $h_m(X)$ = *Kernel function*



Piecewise Constant

Piecewise Linear

Continuous Piecewise Linear

Piecewise-linear Basis Function

"knots"

Piecewise Constant

Linear Spline

Quadratic Spline

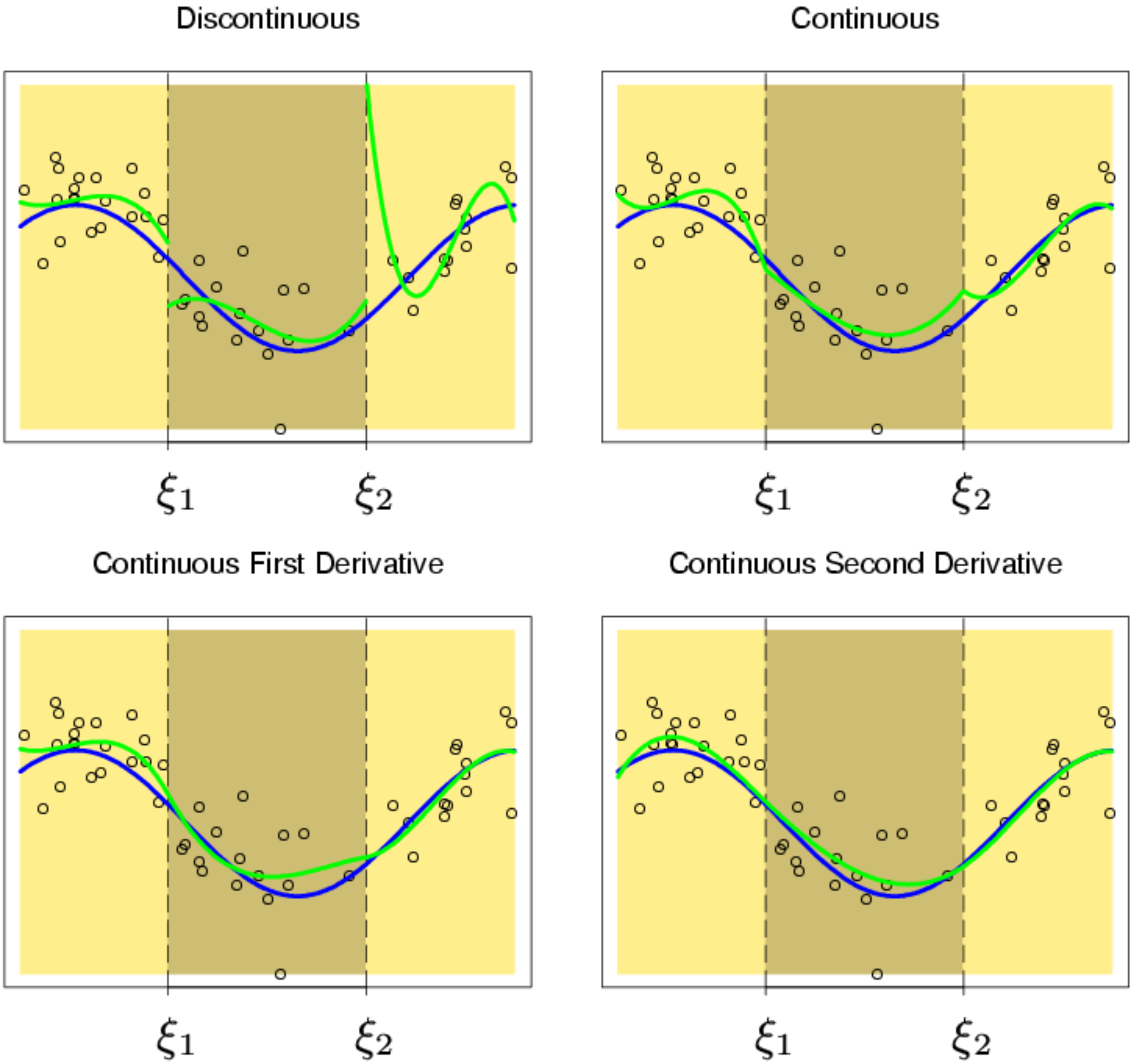
Cubic Spline

Regression Splines

Bottom left panel uses:

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

Number of parameters = (3 regions) X (2 params per region)
- (2 knots X 1 constraint per knot)
= 4



cubic spline

Cubic Spline

continuous first and second derivatives

$$h_1(X) = 1$$

$$h_2(X) = X$$

$$h_3(X) = X^2$$

$$h_4(X) = X^3$$

$$h_5(X) = (X - \xi_1)_+^3$$

$$h_6(X) = (X - \xi_2)_+^3$$

Number of parameters = (3 regions) X (4 params per region)
- (2 knots X 3 constraints per knot)
= 6

Knot discontinuity essentially invisible to the human eye

Natural Cubic Spline

Adds a further constraint that the fitted function is linear beyond the boundary knots

A natural cubic spline model with *K* knots is represented by *K* basis functions:

$$N_1(X) = 1$$

$$N_2(X) = X$$

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

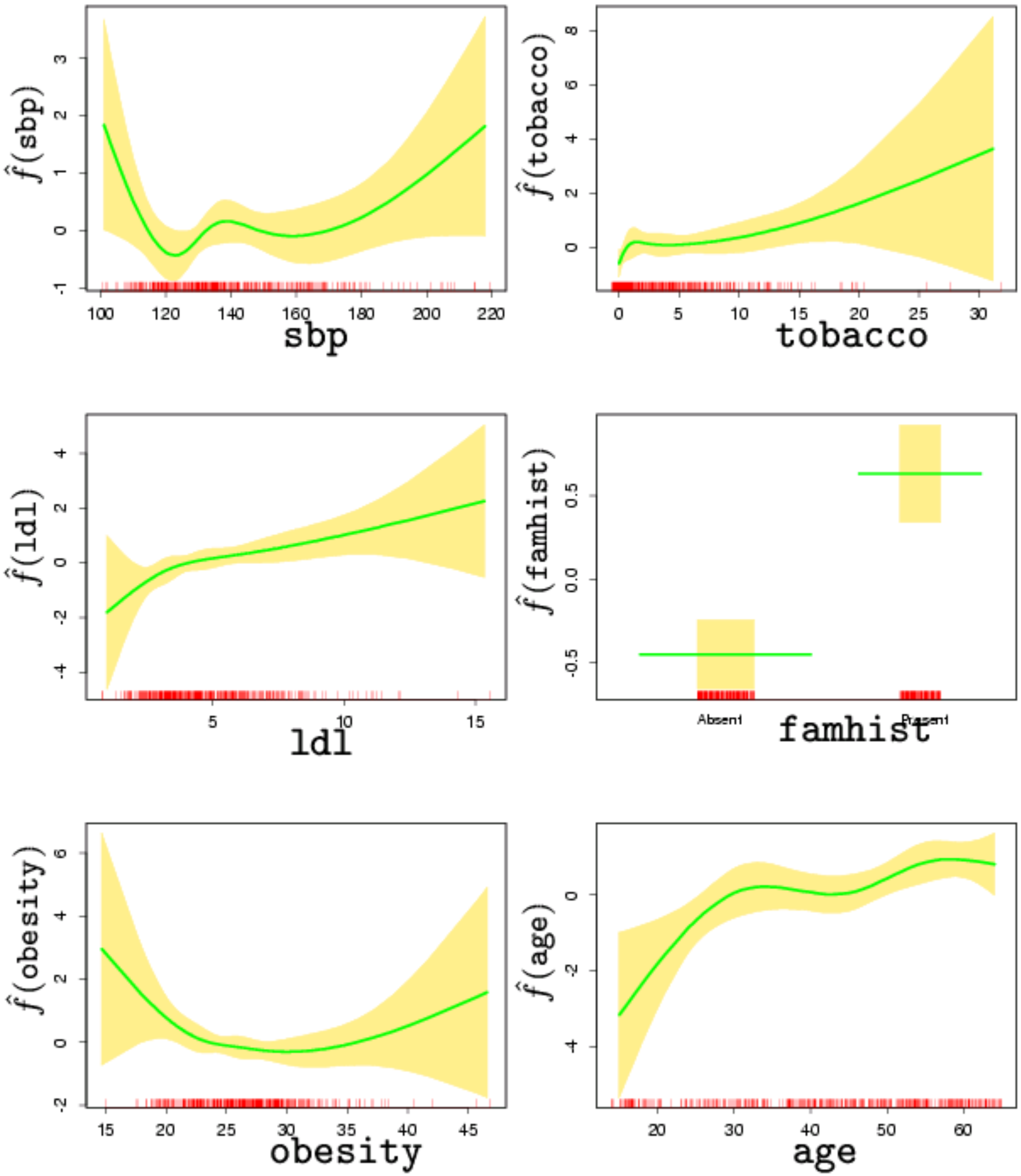
$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}$$

Each of these basis functions has zero 2nd and 3rd derivative outside the boundary knots

Natural Cubic Spline Models

Can use these ideas in, for example, regression models.

For example, if you use 4 knots and hence 4 basis functions per predictor variable, then simply fit logistic regression model with four times the number of predictor variables…


$\hat{f}(\texttt{sbp})$
-1
0
1
2
3
100
120
140
160
180
200
220
sbp
$\hat{f}(\texttt{tobacco})$
0
2
4
6
8
0
5
10
15
20
25
30
tobacco
$\hat{f}(\texttt{ldl})$
-4
-2
0
2
4
5
10
15
ldl
$\hat{f}(\texttt{famhist})$
-0.5
0.0
0.5
Absent
Present
famhist
$\hat{f}(\texttt{obesity})$
-2
0
2
4
6
15
20
25
30
35
40
45
obesity
$\hat{f}(\texttt{age})$
-4
-2
0
20
30
40
50
60
age

Smoothing Splines

Consider this problem: among all functions *f*(*x*) with two continuous derivatives, find the one that minimizes the penalized residual sum of squares:

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

smoothing
parameter

λ=0 : *f* can be any function that interpolates the data
λ=infinity : least squares line

Smoothing Splines

Theorem: The unique minimizer of this penalized RSS is a natural cubic spline with knots at the unique values of x_i, i=1,…,N

Seems like there will be N features and presumably overfitting of the data. But,… the smoothing term shrinks the model towards the linear fit

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

$$RSS(\theta,\lambda) = (y - H\theta)^T (y - H\theta) + \lambda\theta^T \Omega_H \theta \text{ where}$$

$$\{\mathrm{H}\}_{ij} = H_j(x_i) \quad \text{and} \quad \{\Omega_H\}_{jk} = \int H_j''(t) H_k''(t) dt$$

$$\hat{\theta} = (H^T H + \lambda\Omega_H)^{-1} H^T y = S_\lambda y$$

This is a generalized ridge regression

Can show that $S_\lambda = (I + \lambda K)^{-1}$ where K does not depend on λ

We can also estimate $\mathrm{df}_\lambda = \mathrm{trace}(S_\lambda)$.

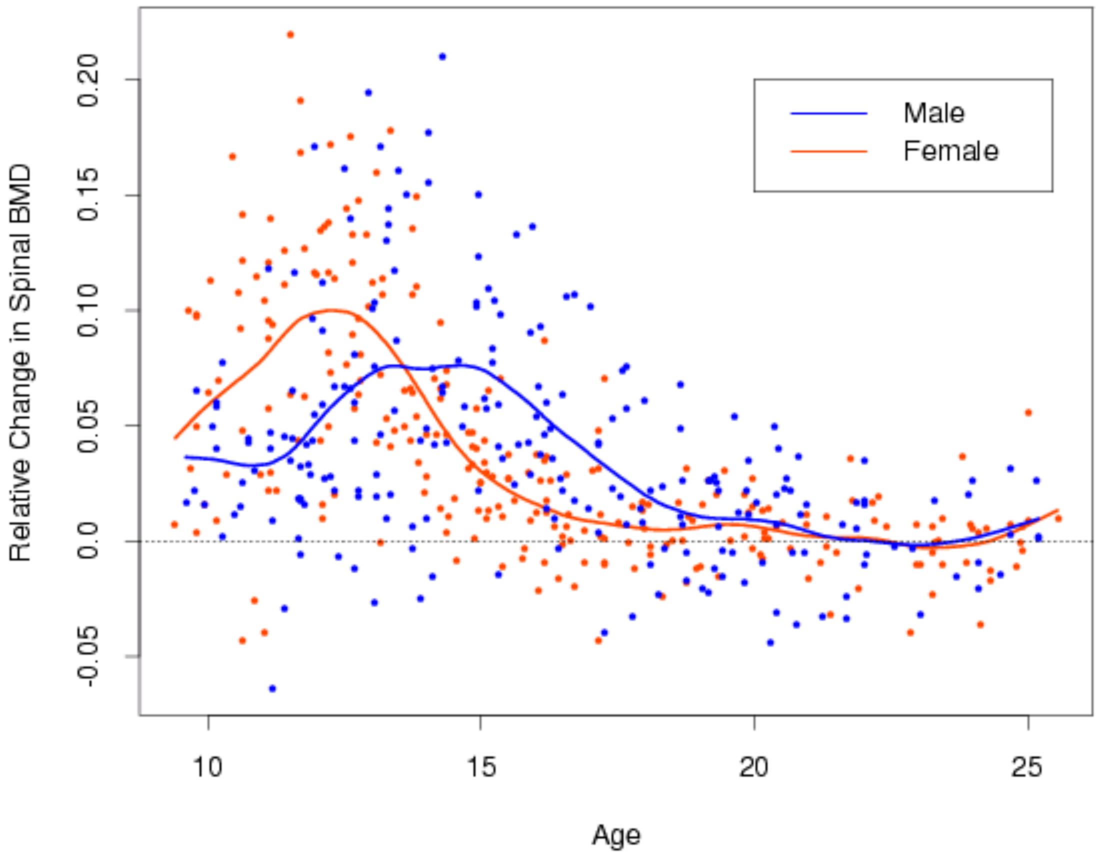
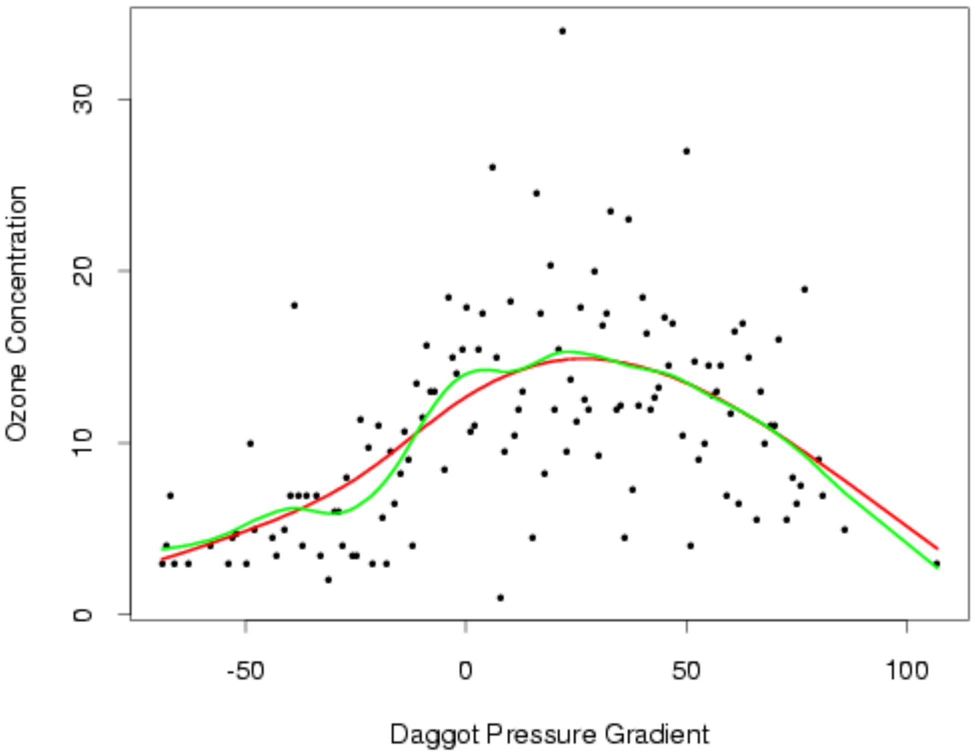


Figure 5.6: *The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with $\lambda \approx 0.00022$. This choice corresponds to about 12 degrees of freedom.*



Smoothing spline fit of ozone concentration versus Daggot pressure gradient. The two fits correspond to different values of the smoothing parameter, chosen to achieve five and eleven effective degrees of freedom, defined by $df_\lambda = \text{trace}(S_\lambda)$.

Bias/Variance Tradeoff

$\boldsymbol{y = f(x) + \varepsilon}$ $\qquad \hat{\theta} = \hat{f} = (H^T H + \lambda \Omega_H)^{-1} H^T y = S_\lambda y$

Variance:

$$\begin{aligned} \mathrm{Cov}(\hat{\mathbf{f}}) &= \mathbf{S}_\lambda \mathrm{Cov}(\mathbf{y}) \mathbf{S}_\lambda^T \\ &= \mathbf{S}_\lambda \mathbf{S}_\lambda^T. \end{aligned}$$

Bias:

$$\begin{aligned} \mathrm{Bias}(\hat{\mathbf{f}}) &= \mathbf{f} - \mathrm{E}(\hat{\mathbf{f}}) \\ &= \mathbf{f} - \mathbf{S}_\lambda \mathbf{f}, \end{aligned}$$

Example:

$$Y = f(X) + \varepsilon,$$

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

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

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

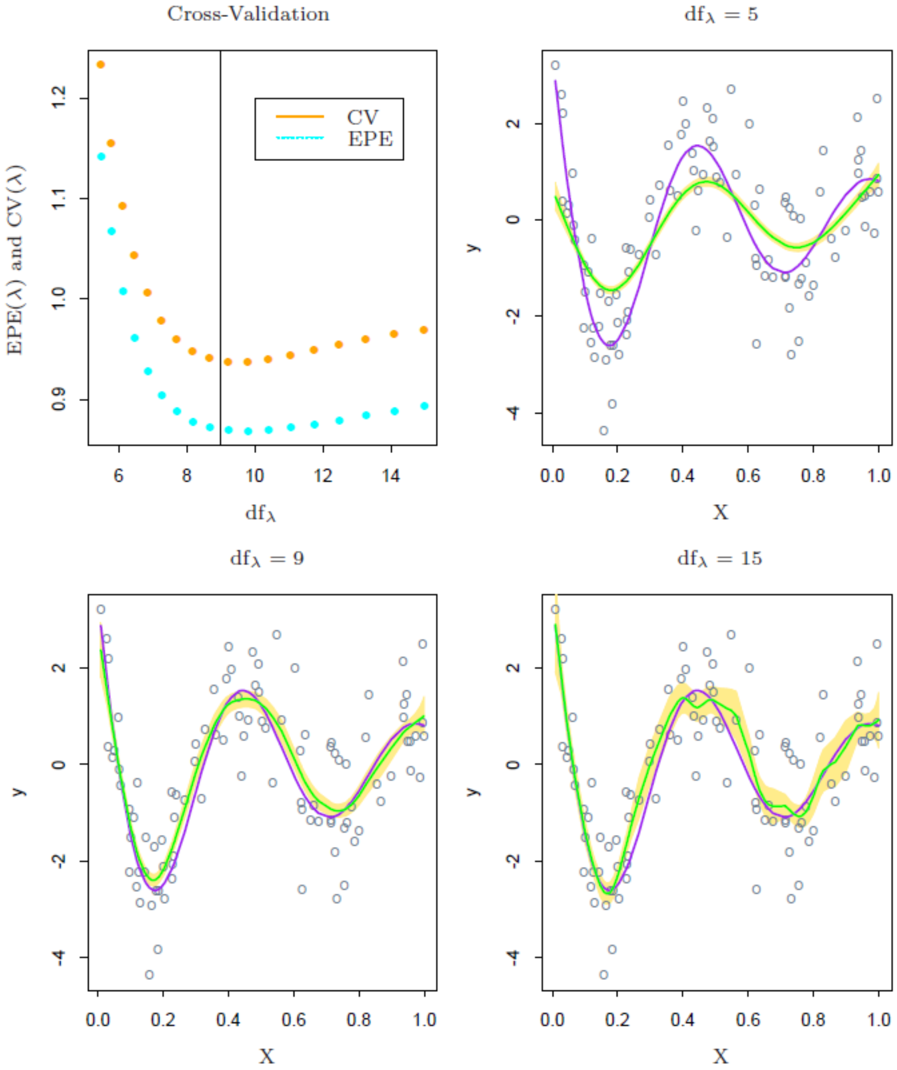


FIGURE 5.9. *The top left panel shows the* EPE(λ) *and* CV(λ) *curves for a realization from a nonlinear additive error model (5.22). The remaining panels show the data, the true functions (in purple), and the fitted curves (in green) with yellow shaded* ±2× *standard error bands, for three different values of* df_λ.

Homework 3

Build your own function to fit a cubic spline that uses 1-20 knots and estimates the optimal number of knots by leave-one-out cross-validation

1. Pick number of knots 1...20
2. Estimate RSS for each number of knots by CV
3. Find the optimal # of knots
4. Produce an object with residuals, df's, and fitted values and a vector with CV'd RSS
5. generate 2 graphs one with the fit and points the other with the RSS as a function of df's

Please use only lsfit or lm to calculate the splines

Nonparametric Logistic Regression

Consider logistic regression with a single *x*:

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

and a penalized log-likelihood criterion:

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

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

Again can show that the optimal *f* is a natural spline with knots at the datapoint

Can use Newton-Raphson to do the fitting.

Multidimensional Splines

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

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

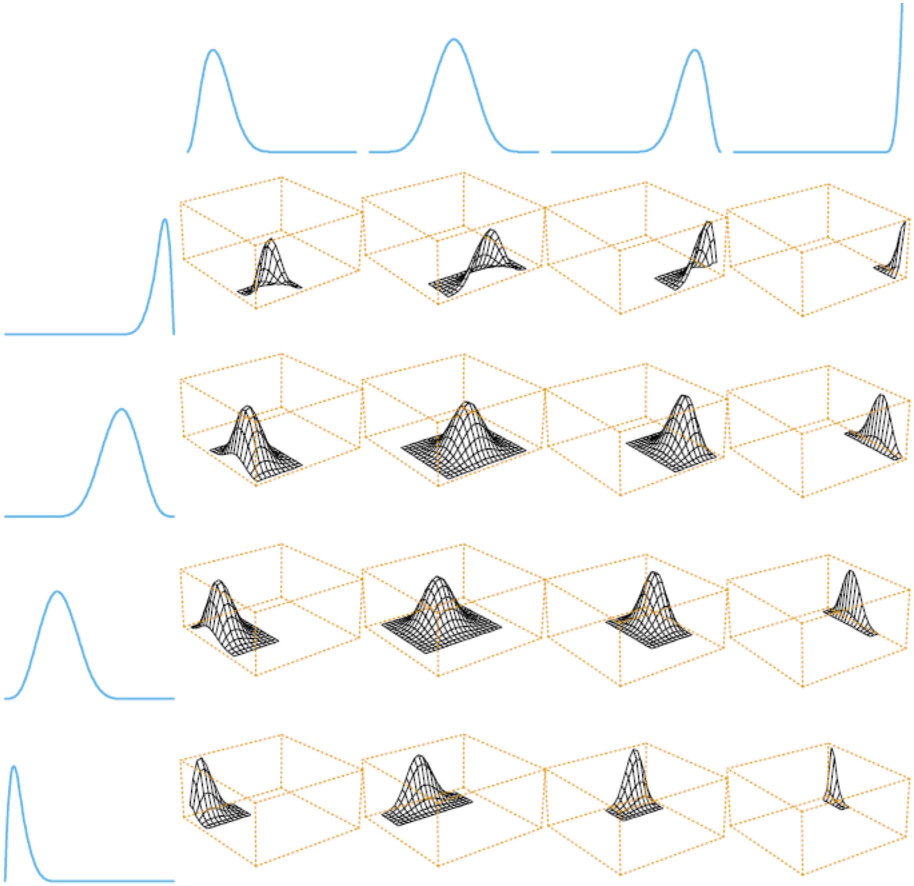
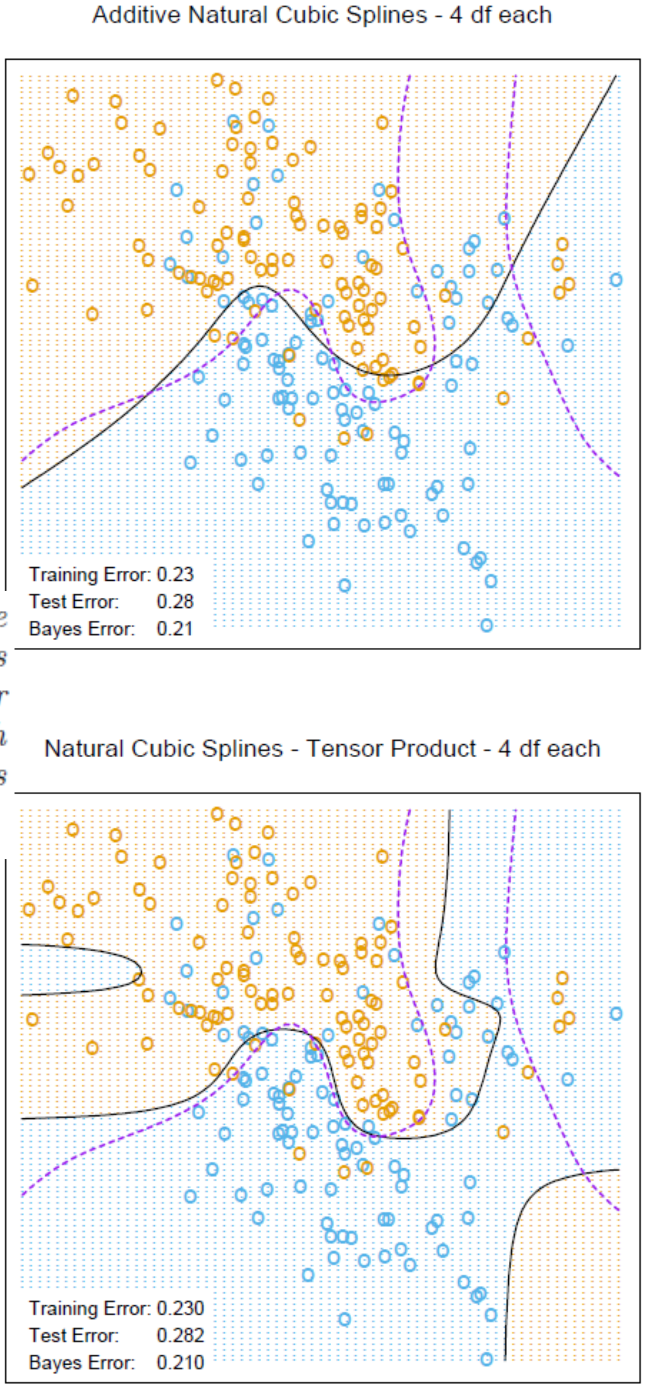


FIGURE 5.10. *A tensor product basis of B-splines, showing some selected pairs. Each two-dimensional function is the tensor product of the corresponding one dimensional marginals.*

Multidimensional Splines



Additive Natural Cubic Splines - 4 df each

Natural Cubic Splines - Tensor Product - 4 df each

FIGURE 5.11. *The simulation example of Figure 2.1. The upper panel shows the decision boundary of an additive logistic regression model, using natural splines in each of the two coordinates (total df* $= 1 + (4-1) + (4-1) = 7$*). The lower panel shows the results of using a tensor product of natural spline bases in each coordinate (total df* $= 4 \times 4 = 16$*). The broken purple boundary is the Bayes decision boundary for this problem.*

Thin-Plate Splines

The discussion up to this point has been one-dimensional. The higher-dimensional analogue of smoothing splines are "thin-plate splines." In 2-D, instead of minimizing:

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

minimize:

$$RSS(f,\lambda) = \sum_{i=1}^{N} \left\{y_i - f(x_i)\right\}^2 + \lambda J(f) \quad \text{where}$$

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

Thin-Plate Splines

The solution has the form:

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

$$h_j(x) = \eta(\|x - x_j\|) \quad \text{and} \quad \eta(z) = z^2 \log z^2$$

a type of “radial basis function”

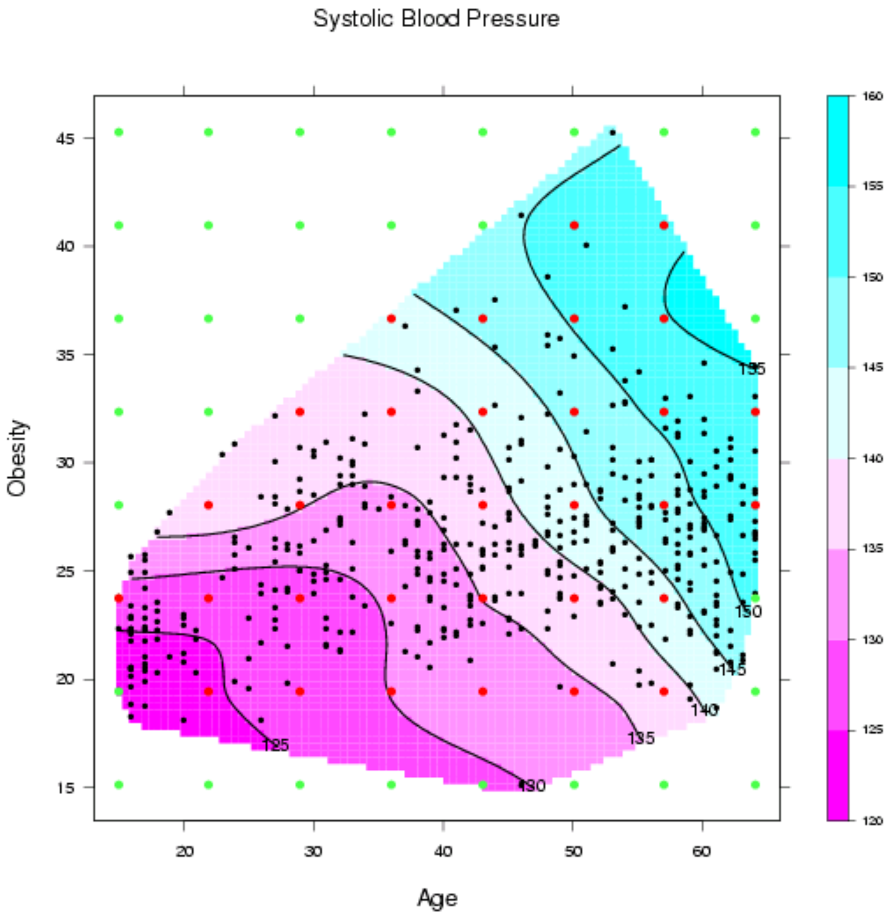


Figure 5.12: *A thin-plate spline fit to the heart disease data, displayed as a contour plot. The response is* **systolic blood pressure**, *modeled as a function of* **age** *and* **obesity**. *The data points are indicated, as well as the lattice of points used as knots. Care should be taken to use knots from the lattice inside the convex hull of the data (red), and ignore those outside (green).*

General Regularization problem

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

Kernel methods

$$f(x) = \textstyle\sum_m \alpha_m K(x, y_m)$$

Where some example of kernel is

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

Kernel methods

Gaussian Radial basis function

Using the Gaussian Kernel : $K(x,y) = e^{-\nu||x-y||^2}$

Radial Basis: $k_m(x) = e^{-\nu||x-x_m||^2},\ m = 1,\ldots,N,$

Another Kernel: $K(x,y) = ||x-y||^2 \log(||x-y||).$

Support Vector Classifiers

In general: $f(x) = \alpha_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i)$

The parameters are chosen to minimiuzeMinimize:

$$\min_{\alpha_0, \boldsymbol{\alpha}} \left\{ \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right\}$$