Splines and additive models

MATH 6312
Department of mathematics, UTA

ESL 5.1-5.2.2, 5.3-5.6, 5.7, 9.1

Optional reading: ESL 5.8

Transforming data

- Let f(X) = E(Y|X).
  - It is very unlikely that the true function f(X) is linear in X.
- Transformations of the features X can help.
- Let $h_m(X)$: $R^p \rightarrow R$. Then we can use

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

basis function

Common basis functions

1. $h_m(X) = X_m$ (Usual linear regression).
2. $h_m(X) = X_j^2$ or $h_m(X) = X_j X_k$ (Taylor polynomials).
3. $h_m(X) = \log(X_j), \sqrt{X_j}$.
4. $h_m(X) = I(L_m \leq X_k < U_m)$ (Indicator functions in some intervals).

Curse of dimensionality

- These methods produce a dictionary D consisting of typically a very large number |D| of basis functions, far more than we can afford to fit to our data.

- Restriction methods – decide before-hand to limit the class of functions.

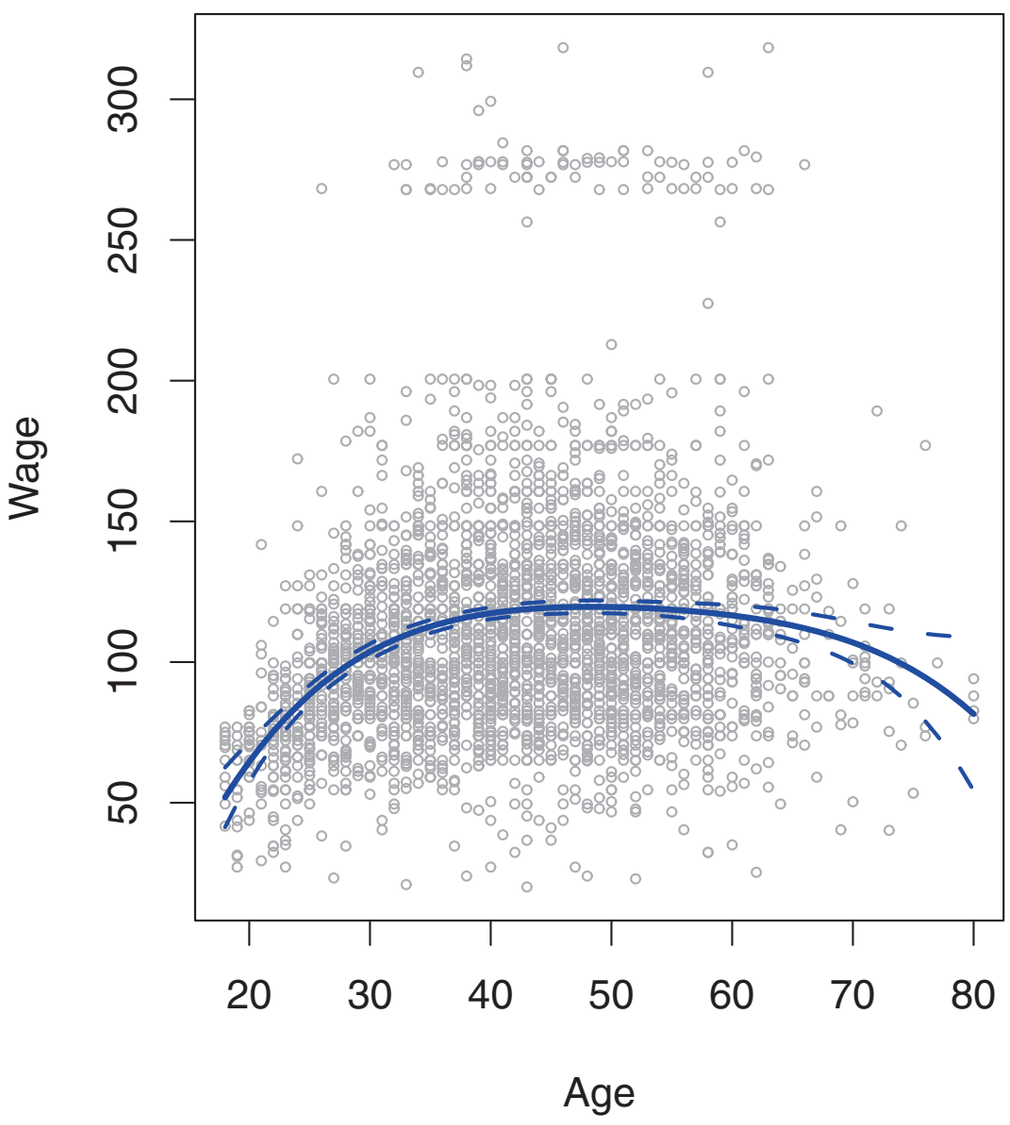
- Selection methods – adaptively scan the dictionary and include only those basis functions that contribute significantly to the fit of the model.

- Regularization methods – use the entire dictionary but restrict the coefficients

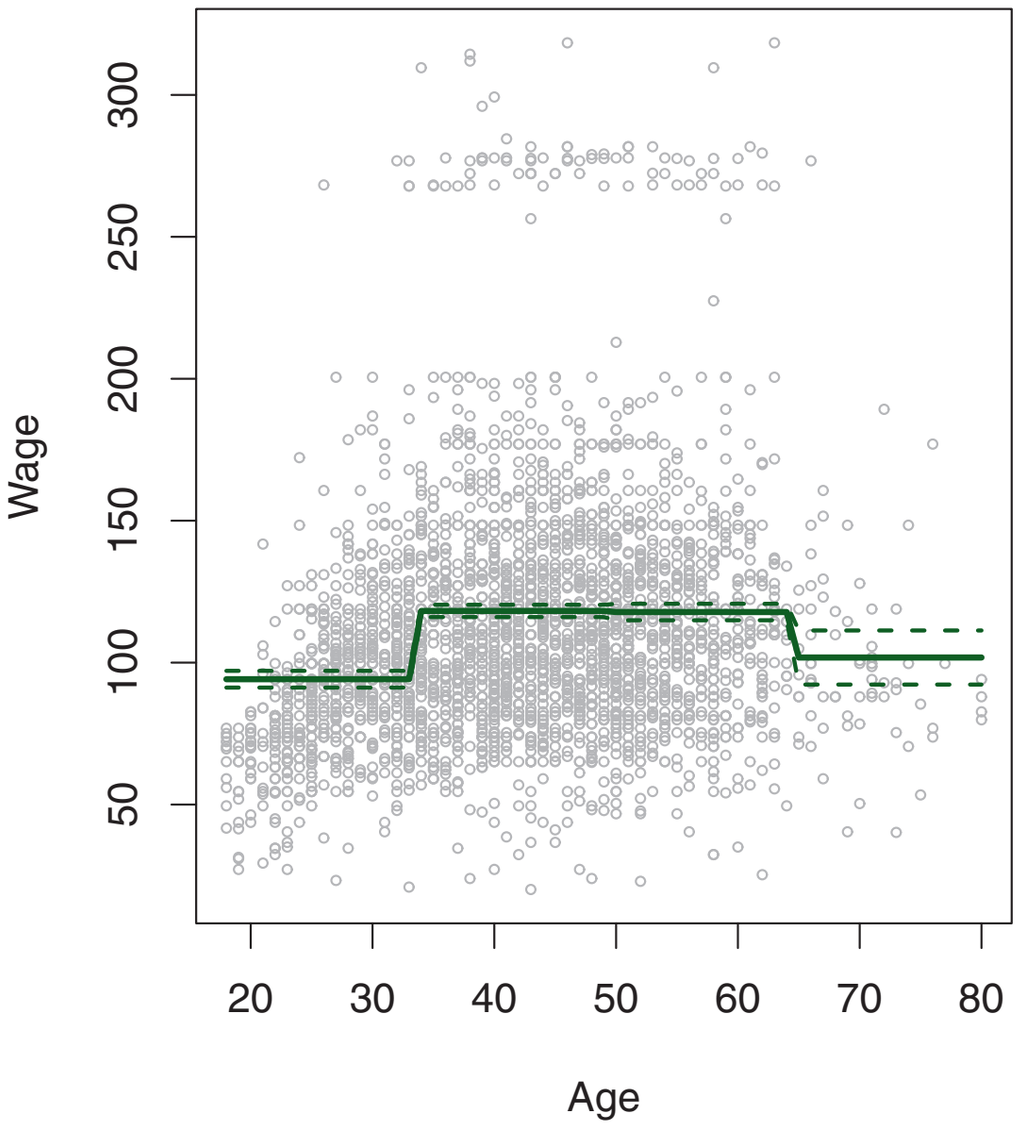
First, we assume X is one-dimensional. Later, we will discuss models for multi-dimensional X.

Mid-Atlantic Wage Data

- Wage and other data for a group of 3000 male workers in the Mid-Atlantic region. A data frame with 3000 observations on the following 11 variables.
- year: Year that wage information was recorded
- age: Age of worker
- maritl: A factor with levels 1. Never Married 2. Married 3. Widowed 4. Divorced and 5. Separated indicating marital status
- race: A factor with levels 1. White 2. Black 3. Asian and 4. Other indicating race
- education: A factor with levels 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree indicating education level
- region: Region of the country (mid-atlantic only)
- jobclass: A factor with levels 1. Industrial and 2. Information indicating type of job
- health: A factor with levels 1. <=Good and 2. >=Very Good indicating health level of worker
- health_ins: A factor with levels 1. Yes and 2. No indicating whether worker has health insurance
- logwage: Log of workers wage
- wage: Workers raw wage



Degree 4 polynomial



Indicator functions

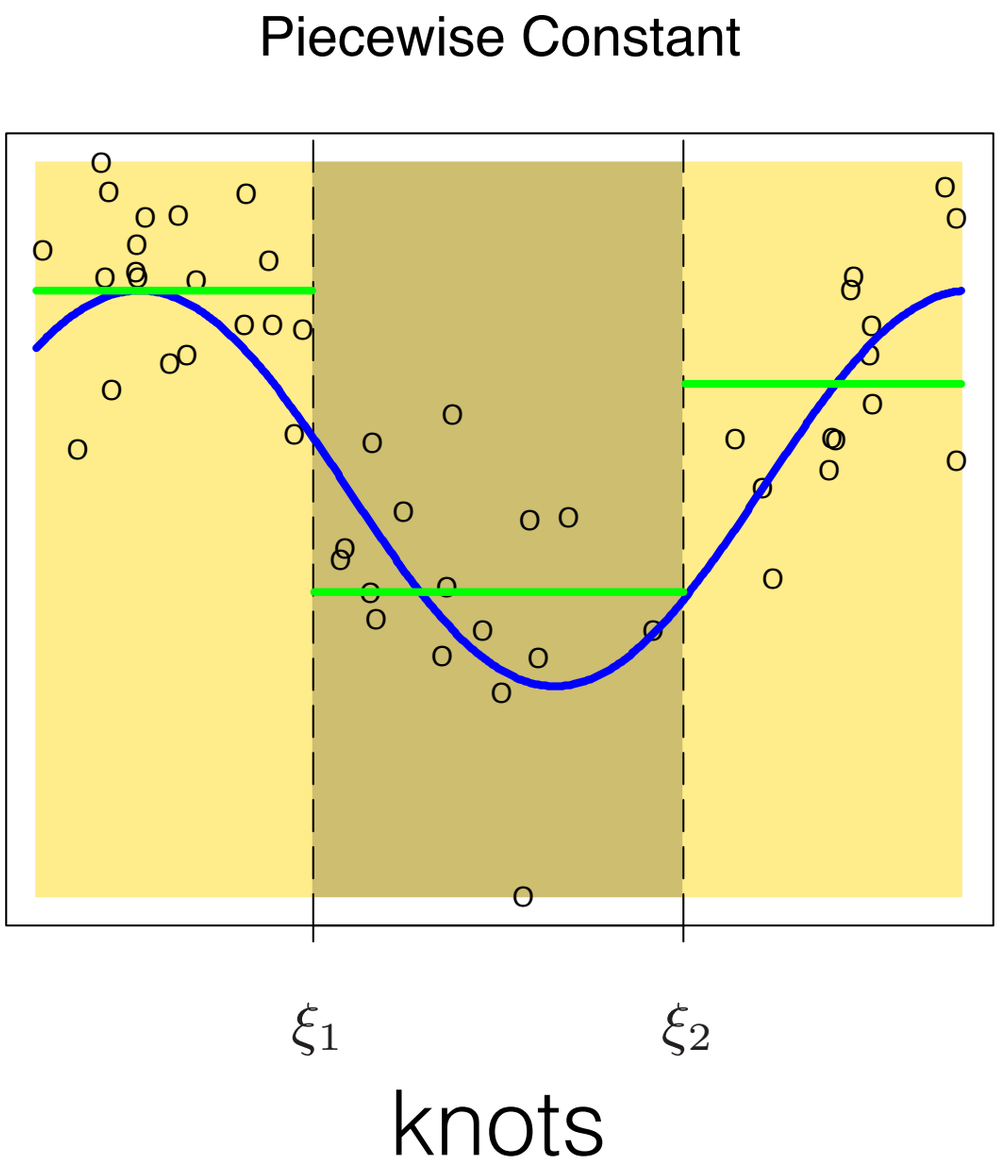
How about combining these basis functions?

- Basis

$$h_1(X) = I(X < \varsigma_1)$$
$$h_2(X) = I(\varsigma_1 < X < \varsigma_2)$$
$$h_3(X) = I(\varsigma_2 \leq X)$$

$$f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$$

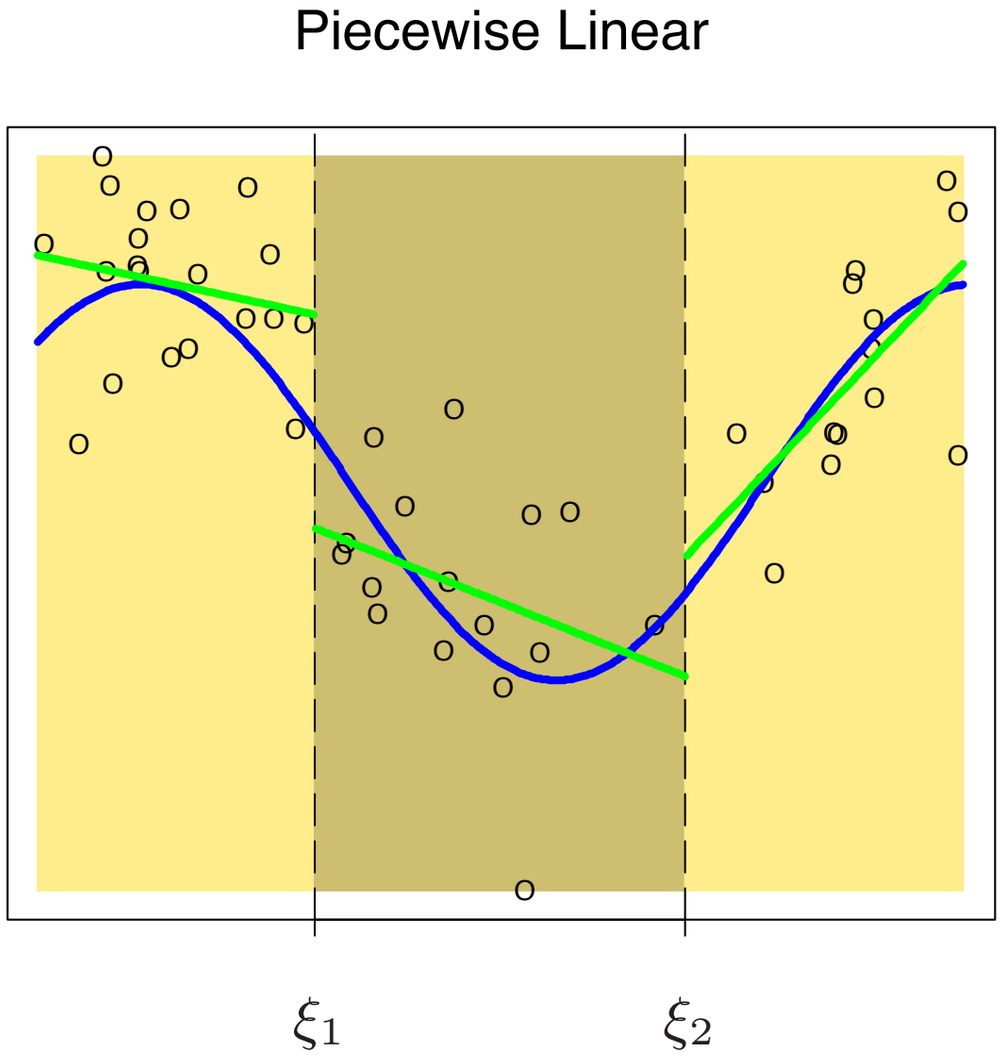
- Degree of freedom



- Basis

$$h_1(X) = I(X < \varsigma_1)$$
$$h_2(X) = I(\varsigma_1 < X < \varsigma_2)$$
$$h_3(X) = I(\varsigma_2 \le X)$$
$$h_{m+3}(X) = h_m(X)X \text{ for } m = 1,2,3.$$

$$f(X) = \sum_{m=1}^{6} \beta_m h_m(X)$$

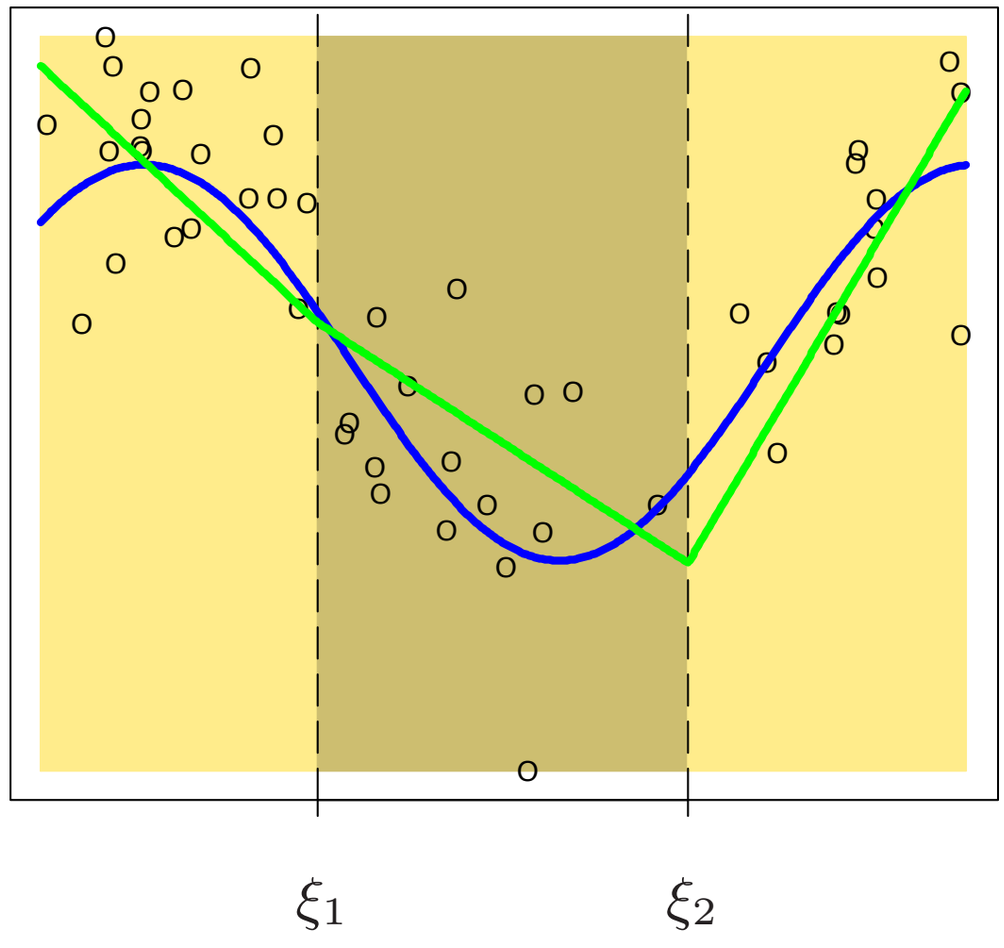


- Degree of freedom

- Basis

$$h_1(X) = I(X < \varsigma_1)$$
$$h_2(X) = I(\varsigma_1 < X < \varsigma_2)$$
$$h_3(X) = I(\varsigma_2 \le X)$$
$$h_{m+3}(X) = h_m(X)X \text{ for } m = 1,2,3.$$

$$f(X) = \sum_{m=1}^{6} \beta_m h_m(X)$$



- Constraints

$$f(\varsigma_1^-) = f(\varsigma_1^+), f(\varsigma_2^-) = f(\varsigma_2^+)$$

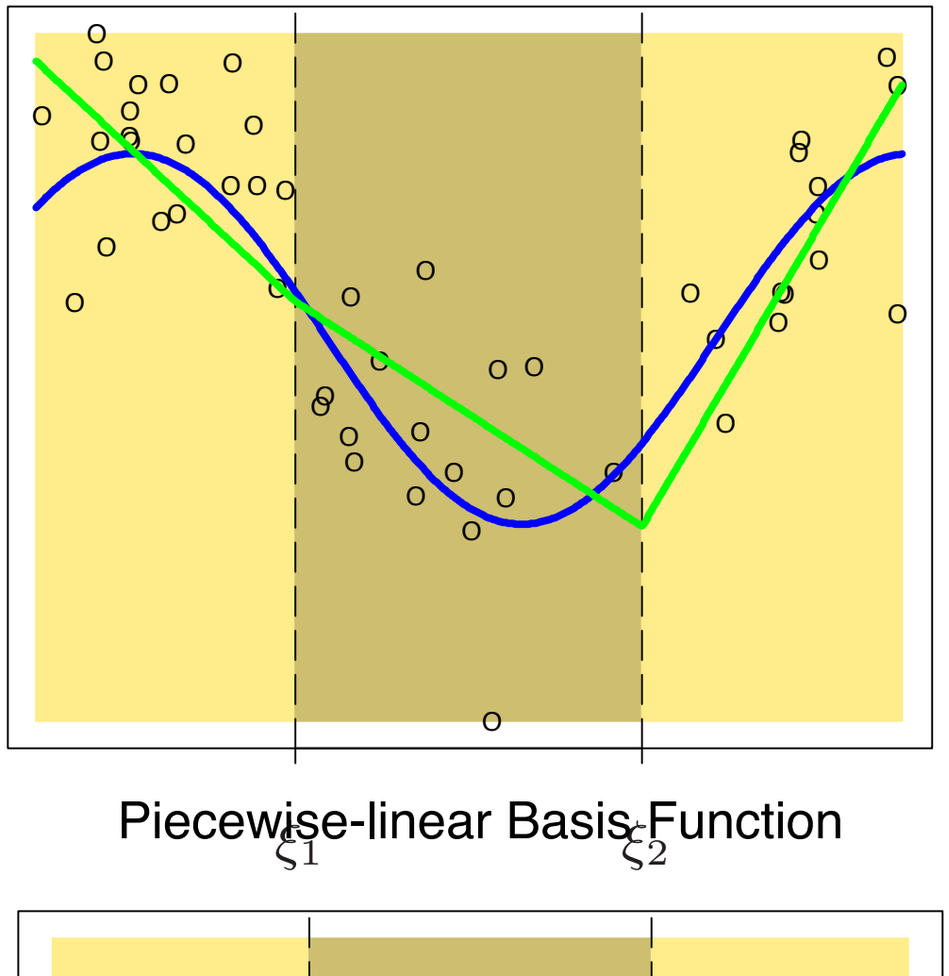
- Degree of freedom

- Alternatively,

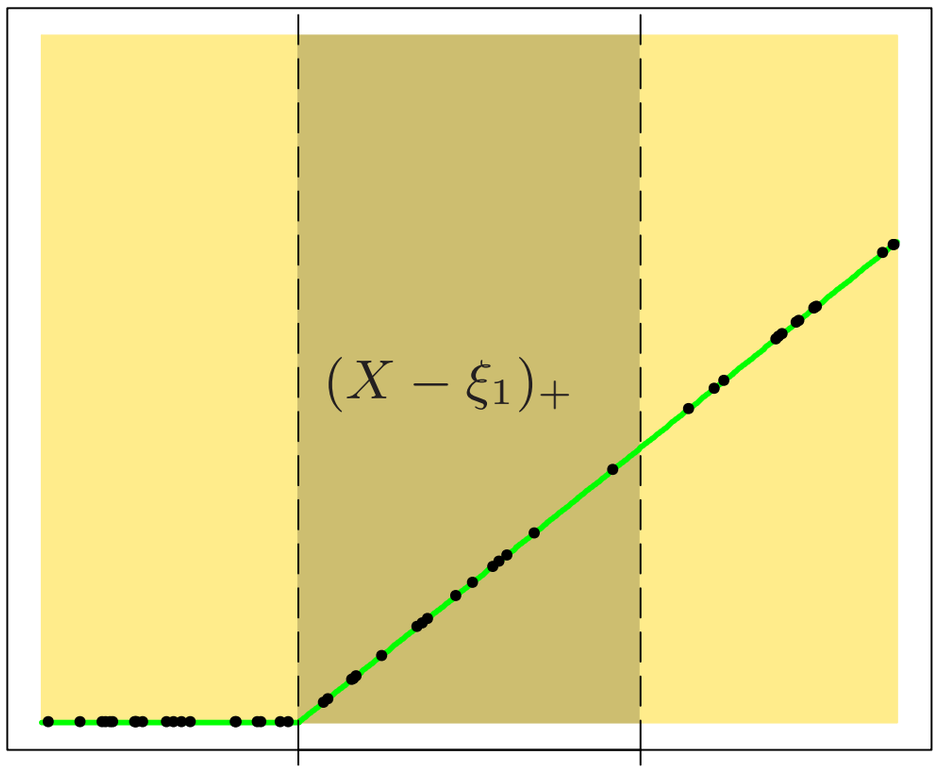
$$h_1(X) = 1$$
$$h_2(X) = X$$
$$h_3(X) = (X - \varsigma_1)_+$$
$$h_4(X) = (X - \varsigma_2)_+$$
$$f(X) = \sum_{m=1}^{4} \beta_m h_m(X)$$

- Find f(X) in each segment

Continuous Piecewise Linear



Piecewise-linear Basis Function



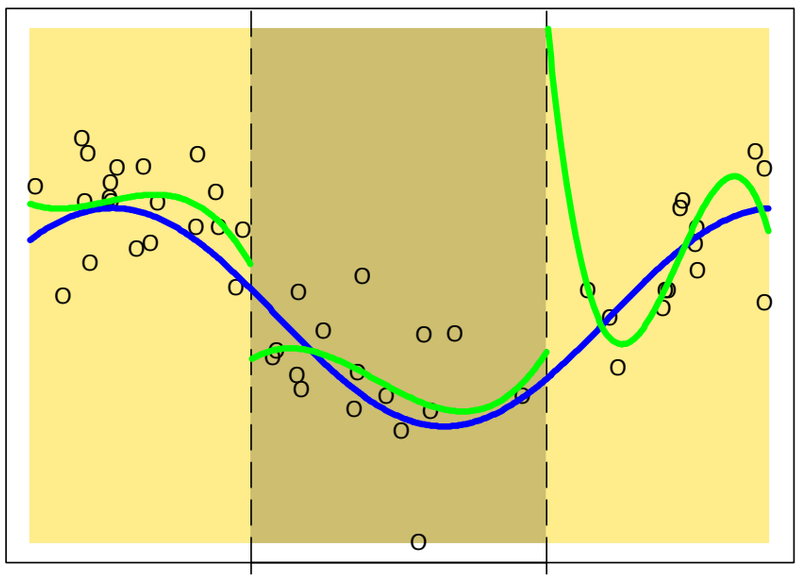
Piecewise polynomial

- Now, imagine we use piecewise polynomial (degree 3) instead of piecewise linear.

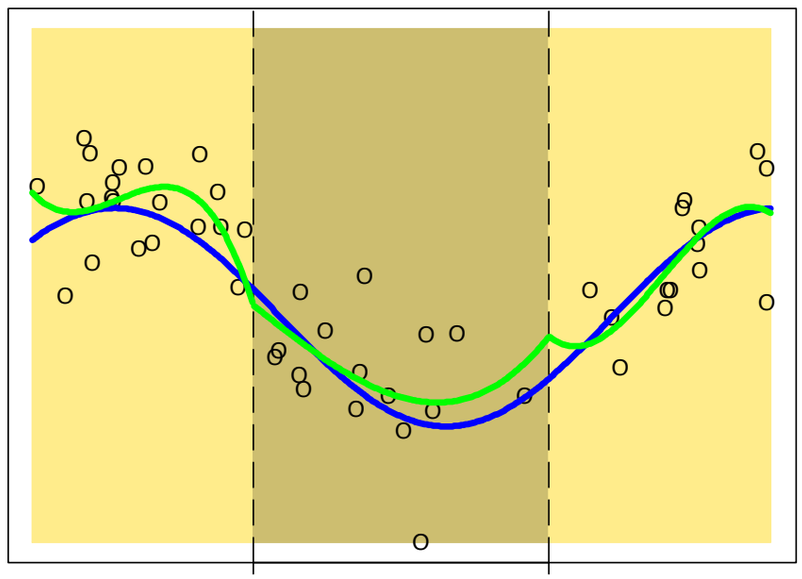
- Continuity constraints can be imposed as well.

- Stronger constraints such as continuous derivatives at knots make f(X) smoother.

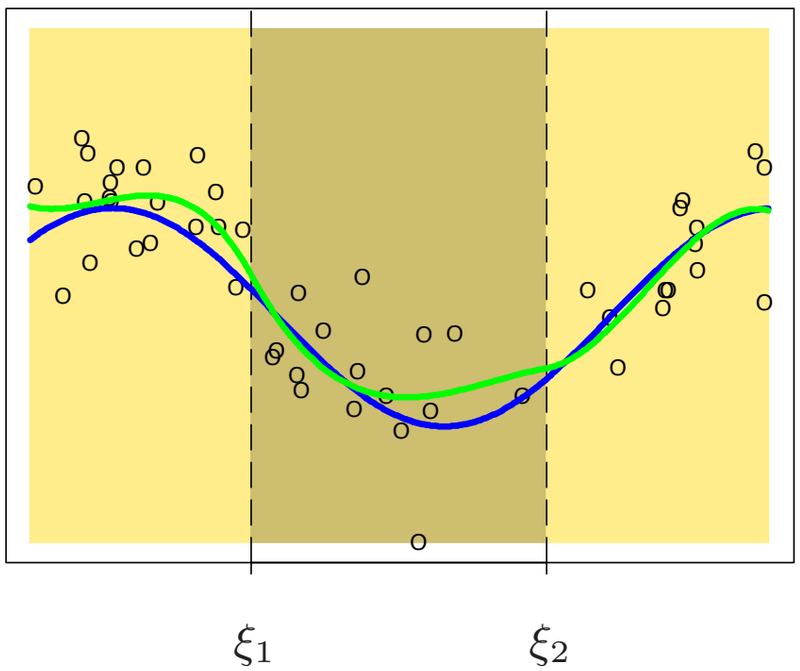
Discontinuous



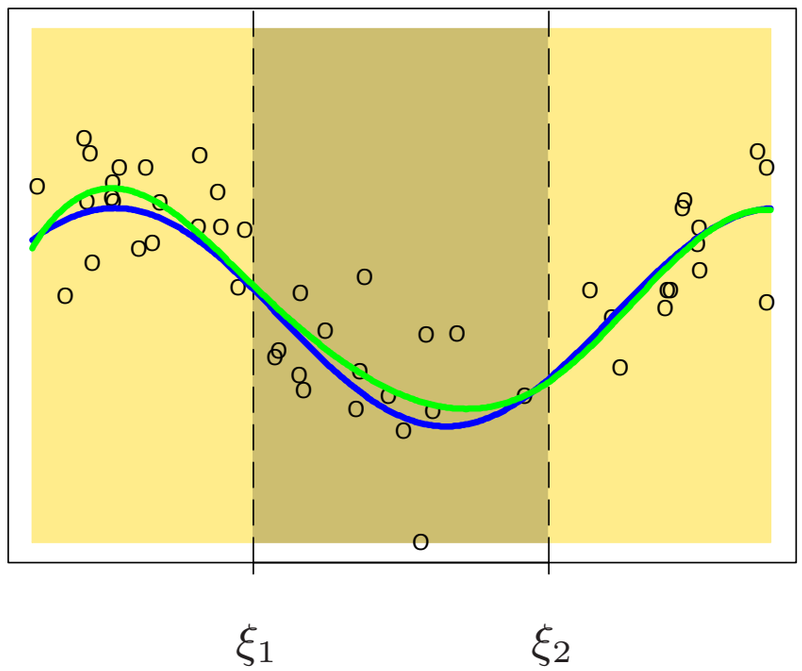
Continuous



Continuous First Derivative



Continuous Second Derivative



Can you find degree of freedom?

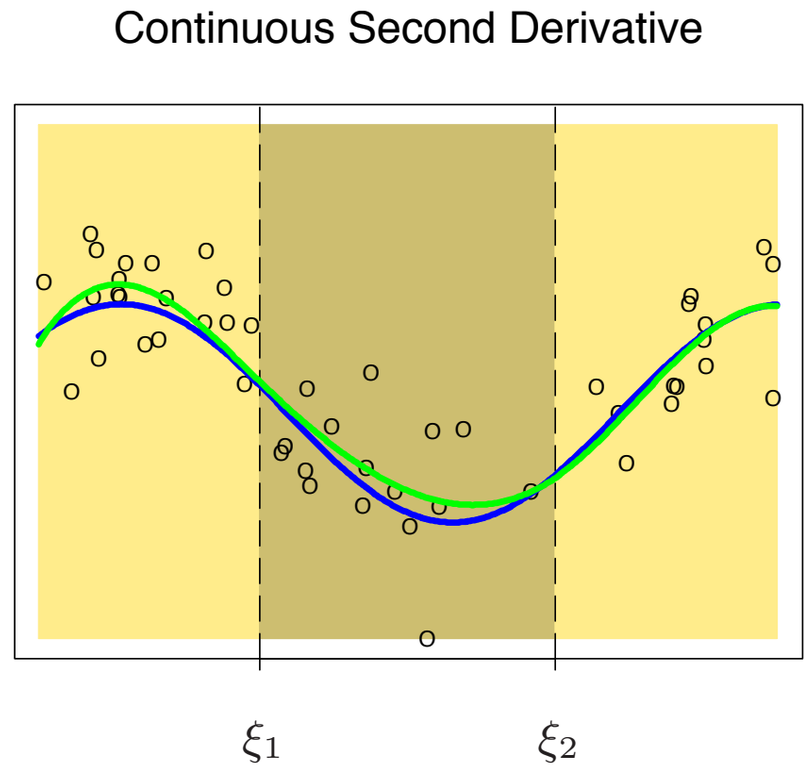
Cubic Spline

- Piecewise cubic (n=3) polynomial f with continuous 1st and 2nd derivatives of f

- The smallest n for which it is impossible to detect the location of the knots by eye.

- Basis (with 2 knots):

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



Basis of cubic splines

- Basis (with 2 knots):

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

- More generally, we M knots, add

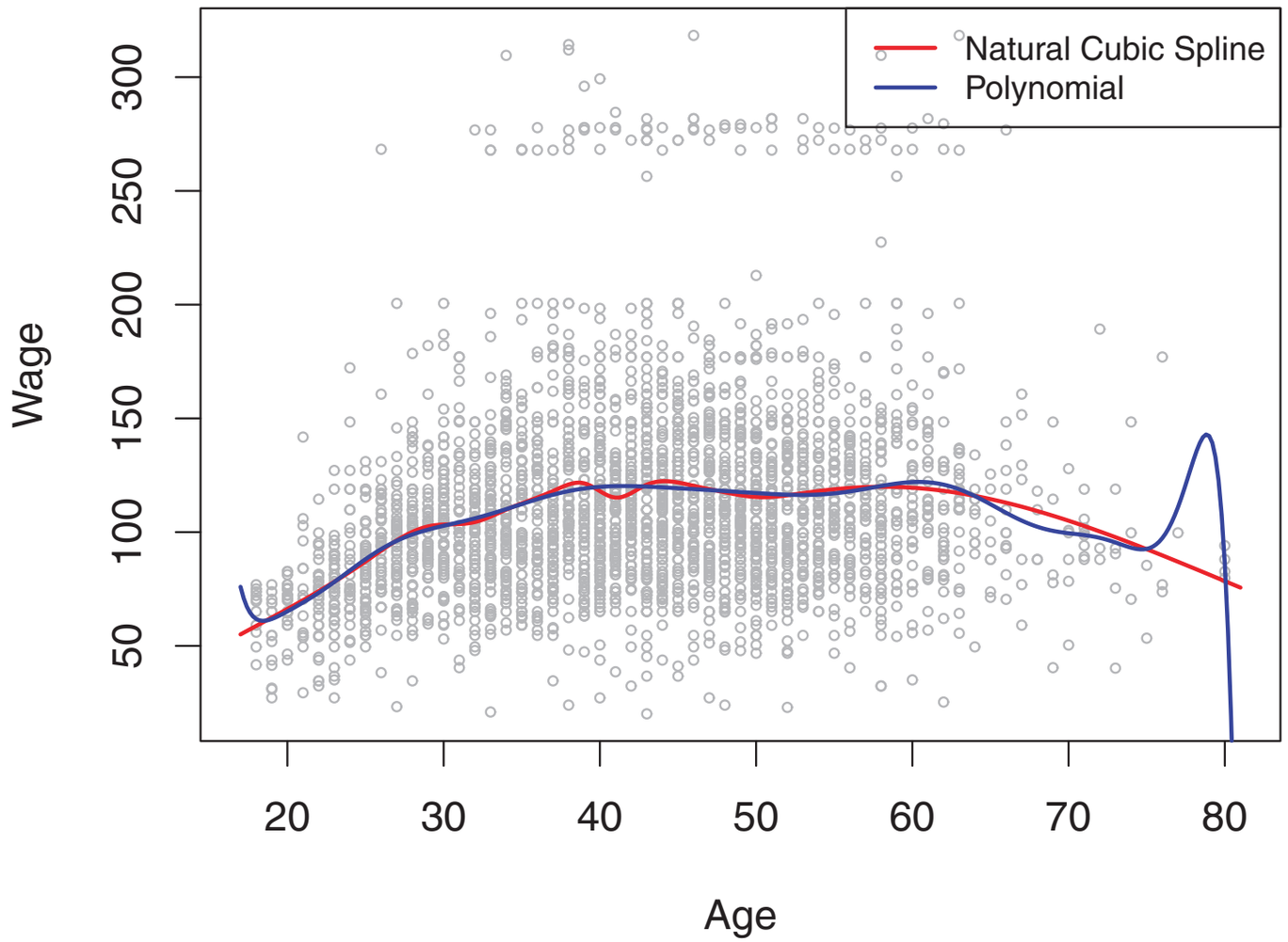
$$(X - \xi_3)_+^3, \ldots, (X - \xi_M)_+^3.$$

- Degree of freedom: 4 + M

Natural cubic splines

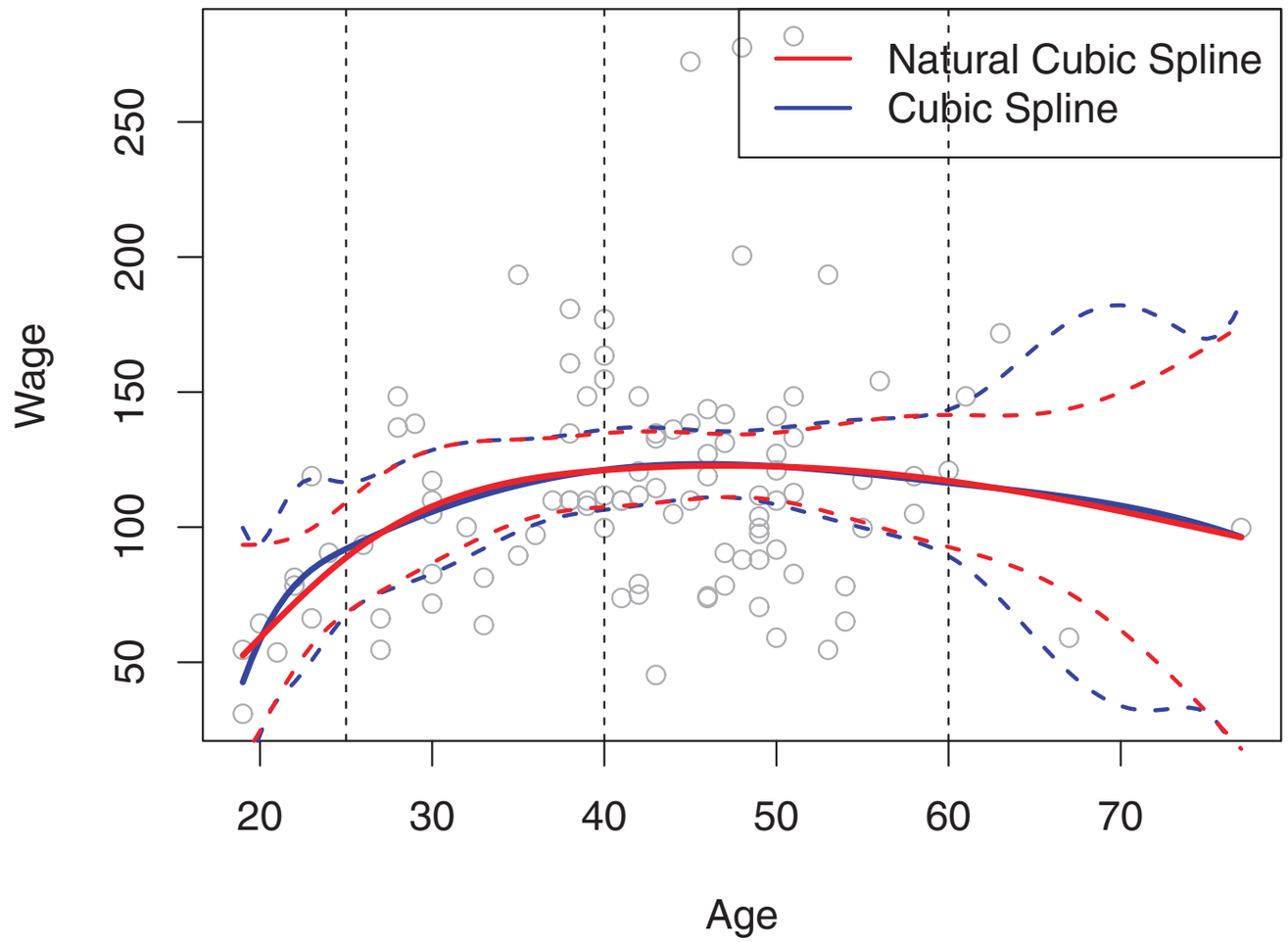
- Polynomials fit to data tends to be erratic near the boundaries
- Natural splines can be used to avoid the erratic behavior of polynomials beyond the knots.
- A natural cubic spline imposes the supplementary conditions that the spline is linear beyond the boundary knots.
- Degree of freedom

Polynomial vs Natural cubic spline



Natural cubic spline vs cubic spline

What is degree of freeom?



- Cubic spline and natural cubic spline (three knots)
- subset of the Wage data

Natural cubic splines

- Basis (with K knots), k=1,2,…,K-2.

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

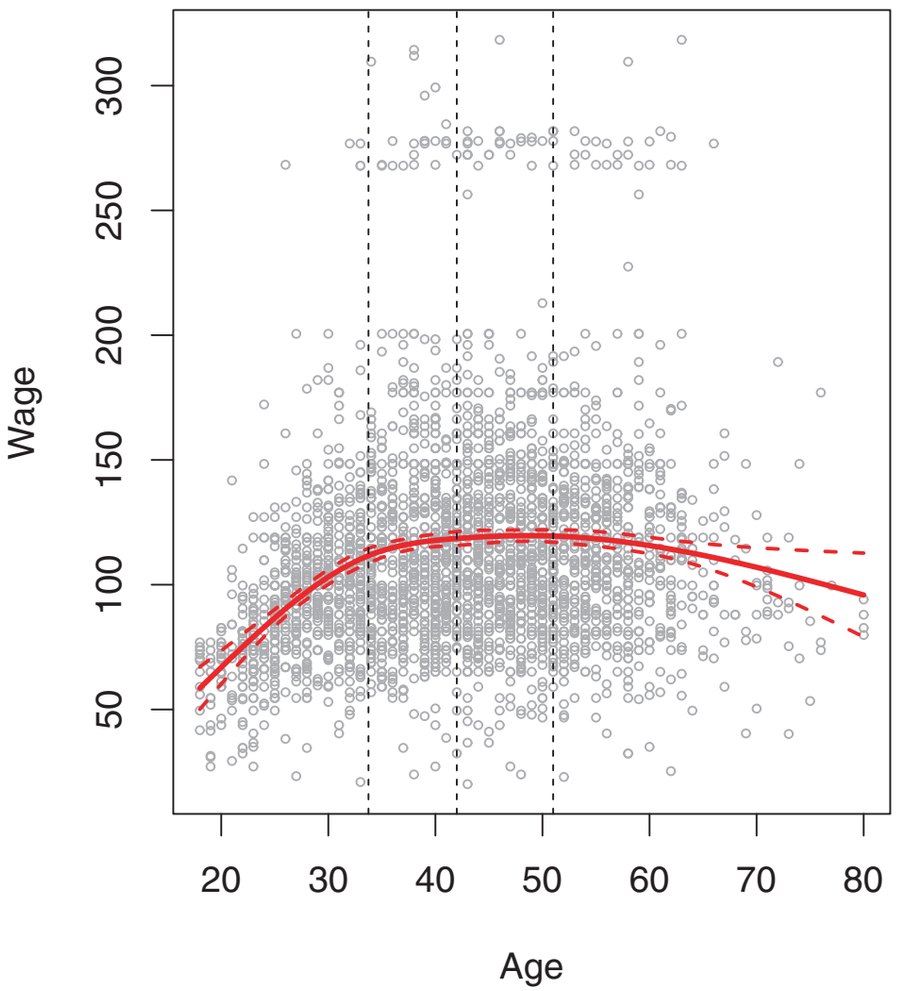
$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}$$

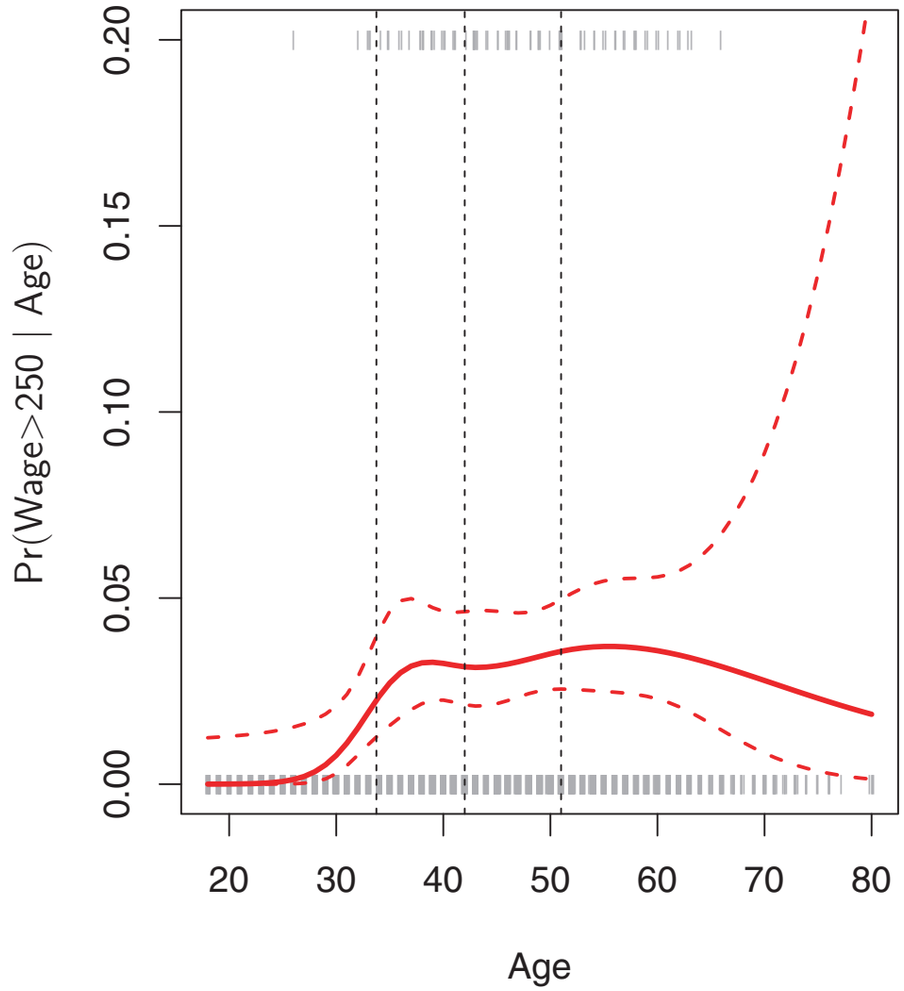
- With K knots, how many basis (=degree of freedom) are in the natural cubic spline?

- How to fit? We can include spline basis in linear regression, or more generally in GLM.

- How to choose the knots? Not always obvious…

Wage data


Wage
50
100
150
200
250
300
20
30
40
50
60
70
80
Age


Pr(Wage>250 | Age)
0.00
0.05
0.10
0.15
0.20
20
30
40
50
60
70
80
Age

South African heart disease

- Model:

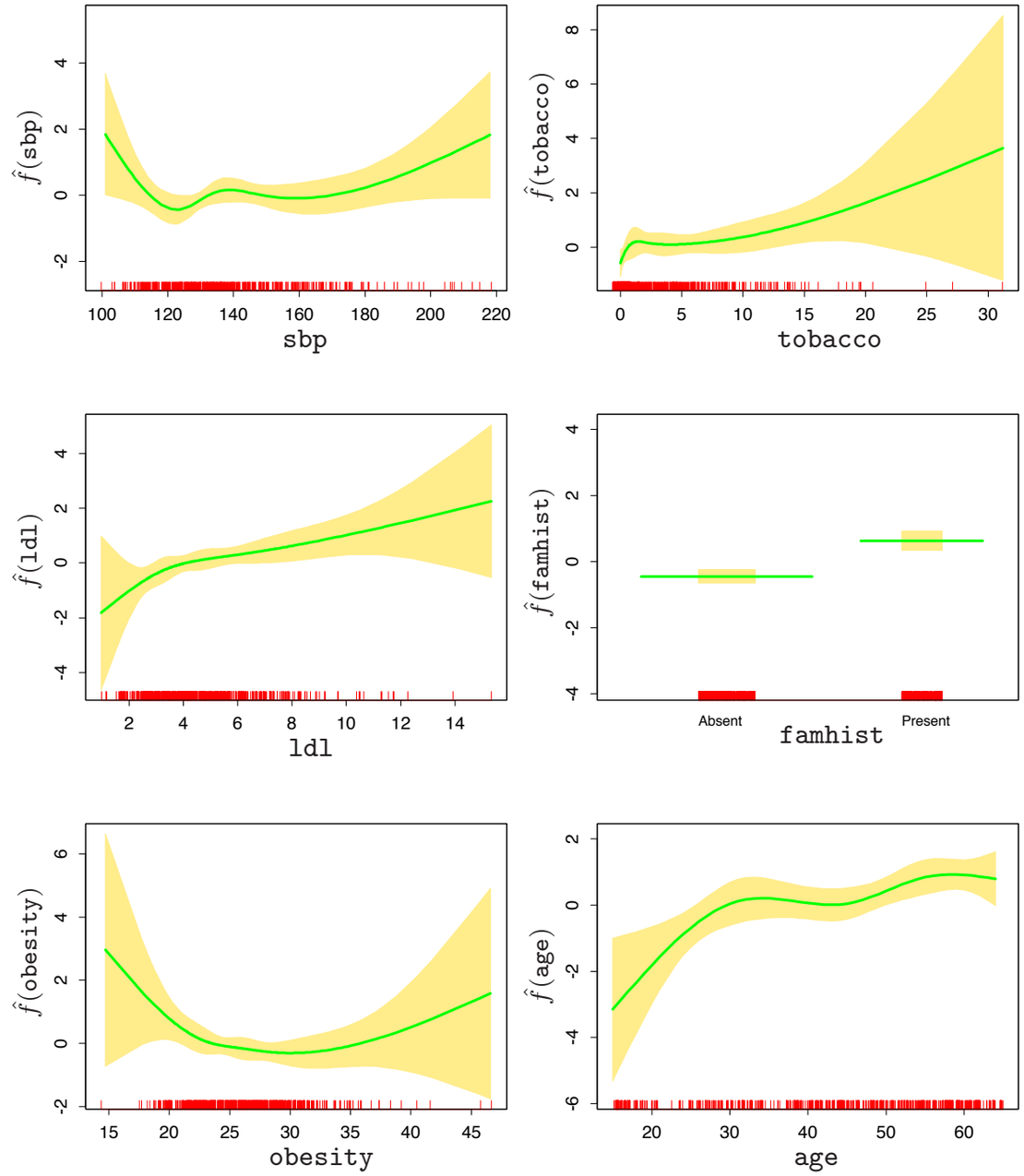
$$\text{logit}[\Pr(\texttt{chd}|X)] = \theta_0 + h_1(X_1)^T\theta_1 + h_2(X_2)^T\theta_2 + \cdots + h_p(X_p)^T\theta_p$$

- Each θ_j is a vector of coefficients

- $h_j(X_j)$ is a vector of basis functions (natural cubic splines with 4 knots).

- What is the dimension of θ_j (i=1,2,…p)?

Smooth functions



Selected model

- Backward stepwise selection (according to AIC)
- Final selected model:

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

Summary

- In the previous example, we fitted a logistic regression to transformed inputs.
- Non-linear transformations are very useful for preprocessing data.
- Powerful method for improving the performance of a learning algorithm
- How to choose the knots in an optimal way?

Smoothing splines

- Find a function $f \in C^2$ such that minimizes

wiggliness

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

smoothing parameter

- First term controls closeness to data.
- Second term controls curvature of the function.
  - $\lambda = 0$: any function that interpolates the data.
  - $\lambda = \infty$: the simple least squares fit

Smoothing splines

- To compute a smoothing spline, we need to optimize on an infinite dimensional space of functions.

- Remarkably, it can be shown that the problem has an explicit, finite-dimensional, unique minimizer which is a natural cubic spline with knots at the unique x_i, i = 1,...,N.

- The penalty term translates to a penalty on the spline coefficients, which are shrunk some of the way toward the linear fit.

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

- Since the solution is a natural cubic spline, it can be written as

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

- We want to find θ such that minimizes

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

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

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

- This is a quadratic function of θ, and the minimizer is (recall the ridge estimate)

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

Effective degrees of freedom

- The prediction is linear in y, and the finite linear operator S_λ is the smoother matrix.

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

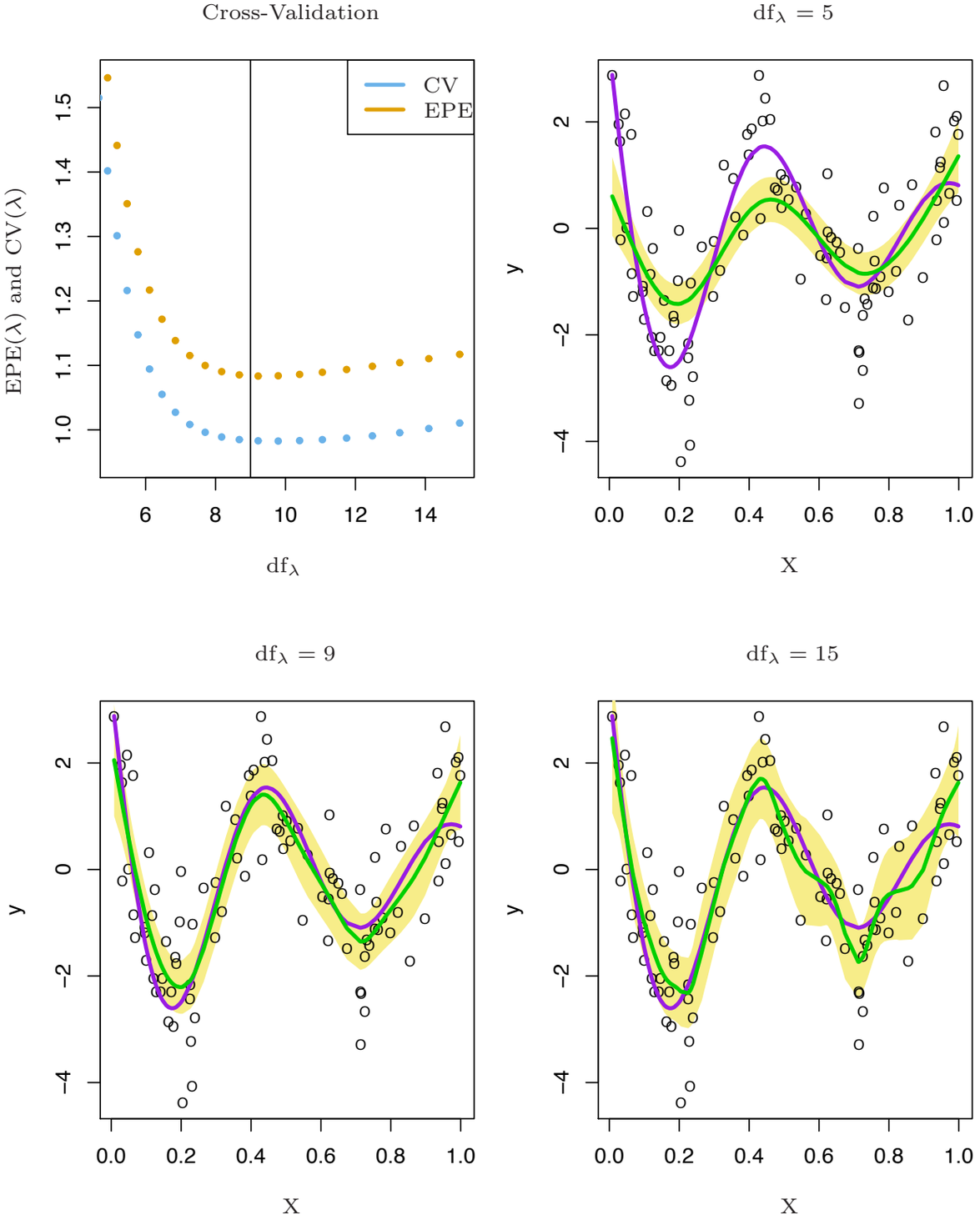
- One can show that the smoother matrix is a symmetric, positive semi-definite matrix.

- Recall that the trace of the hat matrix is used in the ridge regression to define the effective degree of freedom.

- By analogy we define the effective degrees of freedom of a smoothing spline to be

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

Bias–Variance Tradeoff



Artificial data

True functions (in purple), and fitted curves (in green) with yellow shaded ±2×standard error band

Bias–Variance Tradeoff

- $df_\lambda = 5$: The spline under fits.The standard error band is very narrow, so we estimate a badly biased version of the true function with great reliability!

- $df_\lambda = 9$: The spline is close to the true function, and the standard error band is narrow.

- $df_\lambda = 15$: The spline is wiggly, but close to the true function. The wiggliness also accounts for the increased width of the standard error bands — the curve is starting to follow some individual points too closely.

How to choose λ?

- We need to estimate EPE (expected prediction/ generalized error).

- K-fold cross-validation and C_p are in common use.

- Luckily, the leave-one-out CV error has a closed form solution, so it is almost free!

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

Nonparametric logistic regression

- Consider the logistic regression:

$$\log \frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = f(x) \qquad \Pr(Y=1|X=x) = \frac{e^{f(x)}}{1+e^{f(x)}}$$

- Before, we assumed a linear model for f(x), and coefficients were obtained using maximum likelihood.

- Consider the penalized log-likelihood criterion

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

- The optimal f(x) is a natural spline with knots at the unique x_i's

Generalized additive model (GAM)

- $X_1, X_2, \ldots, X_p$ represent predictors and Y is the outcome. f_j's are unspecified smooth functions.

- GAM in the regression setting:

$$\mathrm{E}(Y|X_1, X_2, \ldots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

- GAM in the logistic regression setting:

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(X_1) + \cdots + f_p(X_p) \qquad \mu(X) = \Pr(Y = 1|X)$$

- GAM in general

- $$g[\mu(X)] = \alpha + f_1(X_1) + \cdots + f_p(X_p)$$

Link function

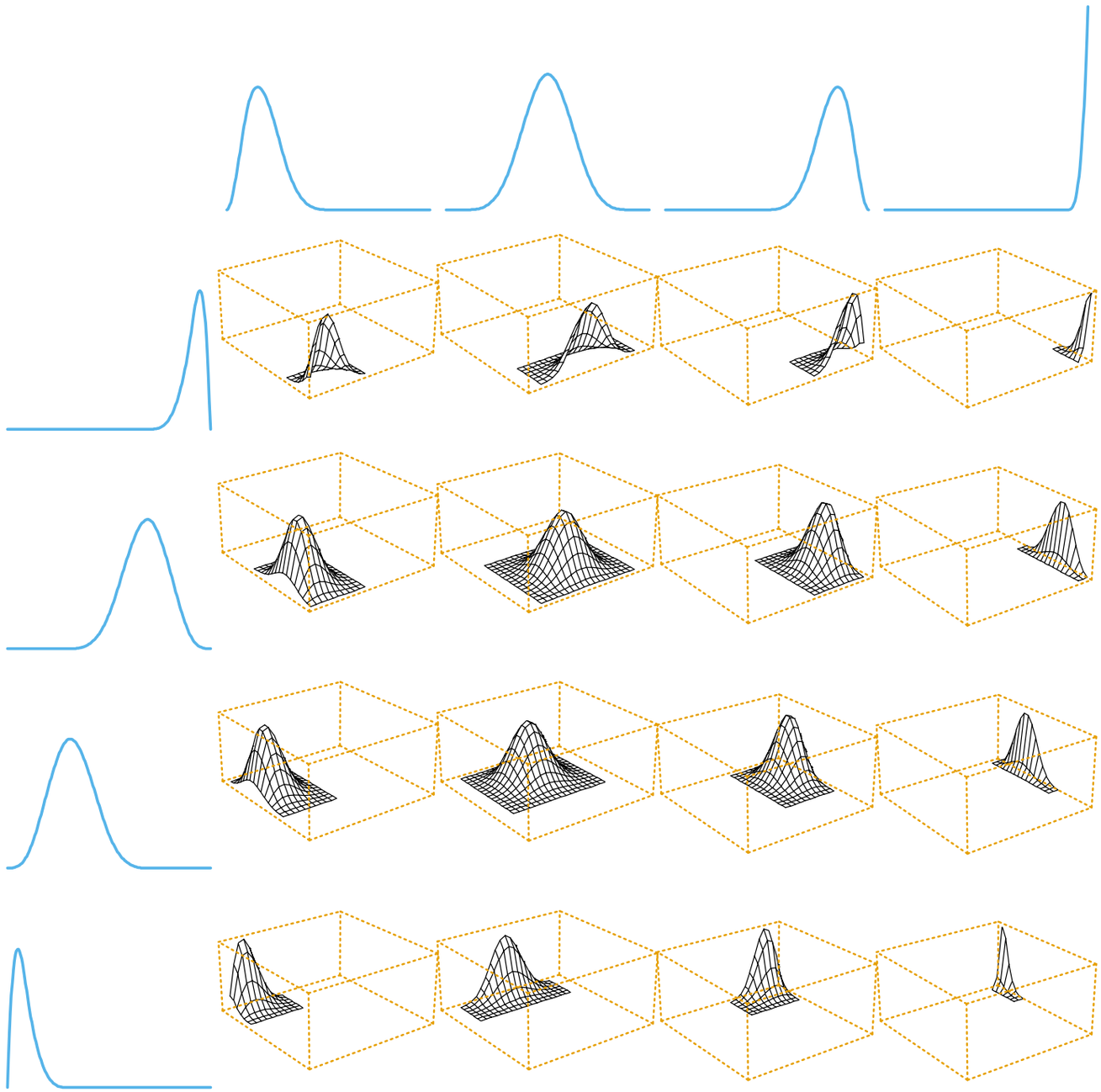
- Examples of g(), link functions
- g(μ) = μ is the identity link, used for linear and additive models for Gaussian response data.
- g(μ) = logit(μ) or g(μ) = probit(μ), the probit link function, for modeling binomial probabilities. The probit function is the inverse Gaussian cumulative distribution function: probit(μ) = $\Phi^{-1}(\mu)$.
- g(μ) = log(μ) for log-linear or log-additive models for Poisson count data.

Multidimensional splines

- Now we assume X is multi-dimensional!
- Suppose $X \in R^2$, and we have a basis of functions $h_{1k}(X_1)$, $k = 1,...,M_1$ for representing functions of coordinate X_1, and likewise a set of M_2 functions $h_{2k}(X_2)$, $k = 1,\ldots,M_2$.
- $M_1 x M_2$ dimensional tensor product basis defined by

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

Tensor product basis



Curse of dimensionality

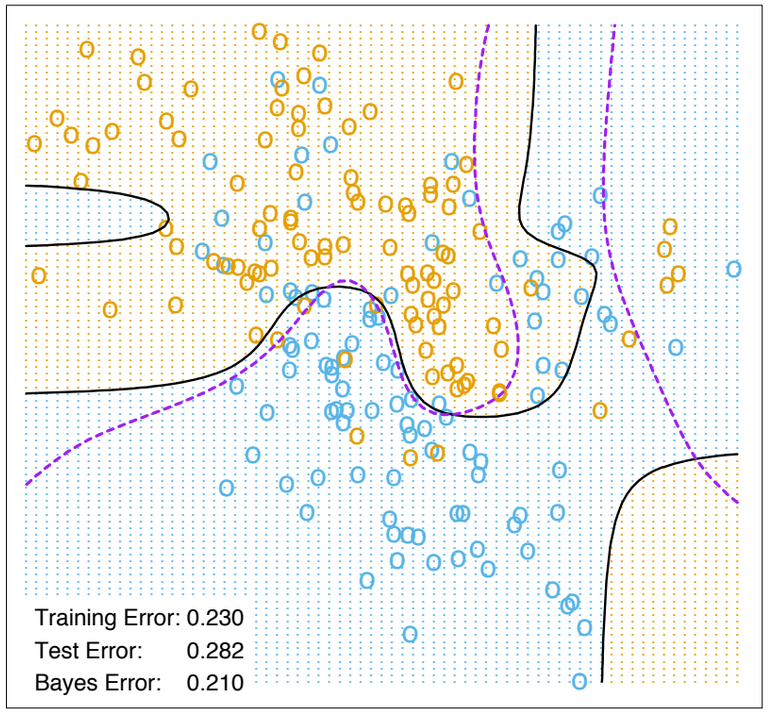
- The regression function now becomes

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

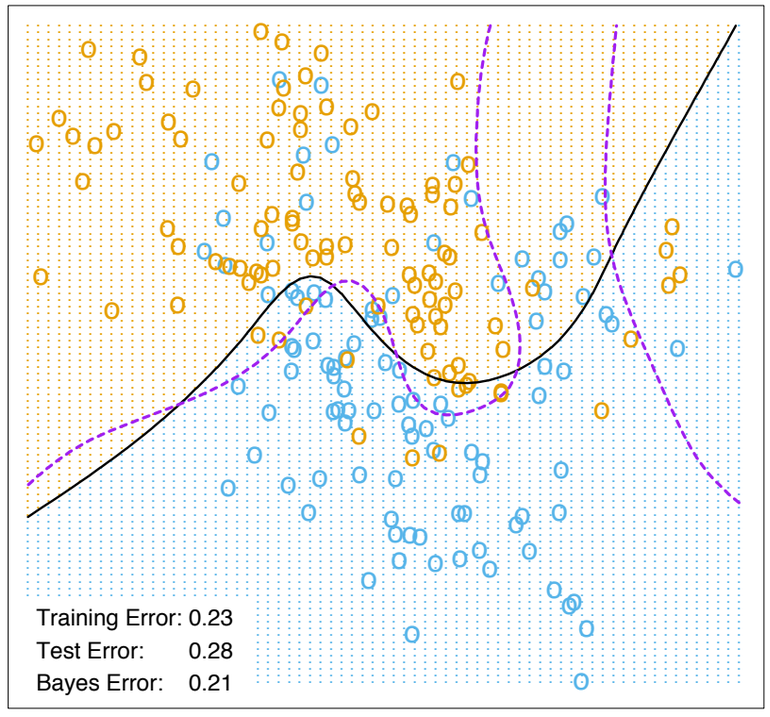
- The coefficients can be fit by least squares, as before.

- This can be generalized to d dimensions, but note that the dimension of the basis grows exponentially fast.

Natural Cubic Splines - Tensor Product - 4 df each



Additive Natural Cubic Splines - 4 df each



- Additive (df = 1 + (4-1) + (4-1)) and tensor product (natural) splines (df = 4 x 4) on a classification example

- The tensor product basis can achieve more flexibility at the decision boundary, but introduces some false structure along the way.

Multi-dimensional smoothing splines

- Suppose $x_i \in R^d$

- We seek a d-dimensional regression function f(x) such that

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

J is an appropriate penalty functional for stabilizing a function f in R^d

- $\lambda \to 0$, the solution approaches an interpolating function

- $\lambda \to \infty$, the solution approaches the least squares plane

Regularization 

- General class of regularization problems

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

$\mathcal{H}$ is a space of functions on which J(f) is defined

$$J(f) = \int_{\mathbb{R}^d} \frac{|\tilde{f}(s)|^2}{\tilde{G}(s)} ds \qquad \tilde{G}(s) > 0, \forall s. \quad \lim_{\|s\| \to 0} \tilde{G}(s) = \infty \qquad \tilde{f} \text{ is the Fourier transform of } f$$

$1/\tilde{G}$ increases the penalty for high-frequency components of f

- Remarkably, the solution is finite-dimensional.

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

- ϕ_k span the null space of the penalty functional J, and G is the inverse Fourier transform of G-tilde.

Reproducing Kernel Hilbert Spaces (RKHS) 

- Suppose K(x,y) is a positive definite kernel, and the corresponding space of functions $\mathcal{H}_K$ is called a reproducing kernel Hilbert space (RKHS)

- Any function in $\mathcal{H}_K$ has the following form

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \qquad (1.1)$$

holds for any $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}, c_1, \ldots, c_n \in \mathbb{R}$.

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

- Eigen-expansion of the kernel

$$K(x,y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x) \phi_i(y) \qquad \gamma_i \geq 0, \sum_{i=1}^{\infty} \gamma_i^2 < \infty$$

- f(x) can be expressed by these eigenfunctions with constraints

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x) \qquad ||f||^2_{\mathcal{H}_K} \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} c_i^2/\gamma_i < \infty$$

-

Example of $\mathcal{H}_K$

- Suppose x,y ∈ R^2

- Define $K(x,y)=\langle (x_1,x_2)',(y_1,y_2)'\rangle^2 = \langle (x_1^2,x_2^2,\sqrt{2}x_1x_2)',(y_1^2,y_2^2,\sqrt{2}y_1y_2)'\rangle$

  $\phi(x)$ $\phi(y)$

  - We can show this is a positive definite kernel.

  - $$K(x,y)=\langle \phi(x),\phi(y)\rangle = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 = \sum_{i=1}^{3}\phi_i(x)\phi_i(y)$$

$$f(x)=\sum_m \alpha_m K(x,y_m)=\sum_m \alpha_m \langle \phi(x),\phi(y_m)\rangle = \left\langle \phi(x),\sum_m \alpha_m \phi(y_m)\right\rangle = c_1x_1^2 + c_2x_2^2 + c_3\sqrt{2}x_1x_2$$

$$f(\cdot)=(c_1,c_2,c_3)' \Rightarrow \| f \|_{\mathcal{H}_k}^2 = \sum_{i=1}^{3} c_i^2$$

reproducing property

$$\langle f,K(\cdot,x_j)\rangle_{\mathcal{H}_k} = \langle f,\phi(x_j)\rangle_{\mathcal{H}_k} = \langle f,\phi(x_j)\rangle = \sum_{i=1}^{3} c_i^2\phi(x_j) = f(x_j)$$

- Problem:

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

$$J(f) = ||f||^2_{\mathcal{H}_K}$$

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda ||f||^2_{\mathcal{H}_K} \right] \quad \Longleftrightarrow \quad \min_{\{c_j\}_1^\infty} \left[\sum_{i=1}^{N} L(y_i, \sum_{j=1}^{\infty} c_j \phi_j(x_i)) + \lambda \sum_{j=1}^{\infty} c_j^2 / \gamma_j \right]$$

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$$

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

$$f(x) = \sum_{j=1}^{N} \alpha_j \left\langle \phi(x), \phi(x_j) \right\rangle = \left\langle \phi(x), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right\rangle$$

$$f = (c_1, c_2, \ldots)', \text{ where } c_l = \sum_{j=1}^{N} \alpha_j \phi_l(x_j)$$

$$\left\langle K(\cdot, x_i), f \right\rangle_{\mathcal{H}_k} = \left\langle \phi(x_i), f \right\rangle_{\mathcal{H}_k} = \left\langle \phi(x_i), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right\rangle = \sum_{j=1}^{N} \alpha_j K(x_i, x_j) = f(x_i)$$

$$\| f \|_{\mathcal{H}_k}^2 = \left\langle f, f \right\rangle_{\mathcal{H}_k} = \left\langle \sum_{i=1}^{N} \alpha_i \phi(x_i), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right\rangle = \sum_{i=1}^{N} \left\langle \phi(x_i), \phi(x_j) \right\rangle \alpha_i \alpha_j = \sum_{i=1}^{N} K(x_i, x_j) \alpha_i \alpha_j$$

reproducing property

$$J(f) = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \alpha_i \alpha_j$$

- The f(x) can be obtained by solving

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

$$K \text{ is } N \times N \text{ matrix where } [K]_{ij} = K(x_i, x_j)$$

Kernel trick

- The infinite-dimensional criterion reduces to a finite-dimensional criterion

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

- This phenomenon has been called the kernel property in support-vector machines (Ch 12).