Basis Expansions and Splines[1]

Germán G. Creamer

June 16, 2022

[1] Sources: Introduction to Statistical Learning with R, and Thomas Lonon

©2021 The Trustees of the Stevens Institute of Technology

Linear Basis Expansion: basis function approach

Denote by $h_m(X) : \mathbb{R}^p \mapsto \mathbb{R}$, the m^{th} transformation of X.

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

1. $h_m(X) = X_m, m = 1, \ldots, p$ recovers the original linear model
2. $h_m(X) = X_j^2$ or $h_m(X) = X_j X_k$ allows us to augment with polynomial terms of higher order
3. $h_m(X) = \log(X_j), \sqrt{X_j}, \ldots$ permits other nonlinear transformations
4. $h_m(X) = \mathbb{I}_{\{L_m \leq X_k \leq U_m\}}$ indicators of region X_k which allows models with piecewise contributions

[1] $h_m(X)$ is a basis function.

©2021 The Trustees of the Stevens Institute of Technology

Methods

- Restriction Methods: decide before-hand to limit the class of functions. For example if we assume our model has the form

$$f(X) = \sum_{j=1}^{p} f_j(X_j) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm} h_{jm}(X_j)$$

  the size of the model is limited by the basis functions M_j
- Selection Methods: Adaptively scan the dictionary and include only those basis functions h_m that contribute significantly to the fit of the model
- Regularization Methods: Use the entire dictionary but restrict the coefficients

©2021 The Trustees of the Stevens Institute of Technology

Step Functions

A simple basis for representing a function would be using indicator functions:

$$h_1(X) = \mathbb{I}_{\{X<\xi_1\}}$$
$$h_2(X) = \mathbb{I}_{\{\xi_1 \leq X<\xi_2\}}$$
$$h_3(X) = \mathbb{I}_{\{X \geq \xi_2\}}$$

The least squares estimate of the model

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

amount to $\hat{\beta}_m = \bar{Y}_m$ the mean of Y in the mth region. I: indicator function. Returns 1 if true. Generates dummy variables.

©2021 The Trustees of the Stevens Institute of Technology

To improve upon this model we could allow for more versatility by including the basis functions:

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

If we wanted this piecewise function to be continuous at the knots, we can include constraints such as

$$f(\xi_1^-) = f(\xi_1^+)$$

which implies

$$\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$$

ξ_i is the knot: points where there is a major change in the dataset.

©2021 The Trustees of the Stevens Institute of Technology

We can also incorporate the constraints into the basis, giving us at this point:

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

$t+$ denotes the positive part.

$(X - \xi_i)_+$: $(X - \xi_i)$, if $X > \xi_i$; 0 otherwise.

©2021 The Trustees of the Stevens Institute of Technology

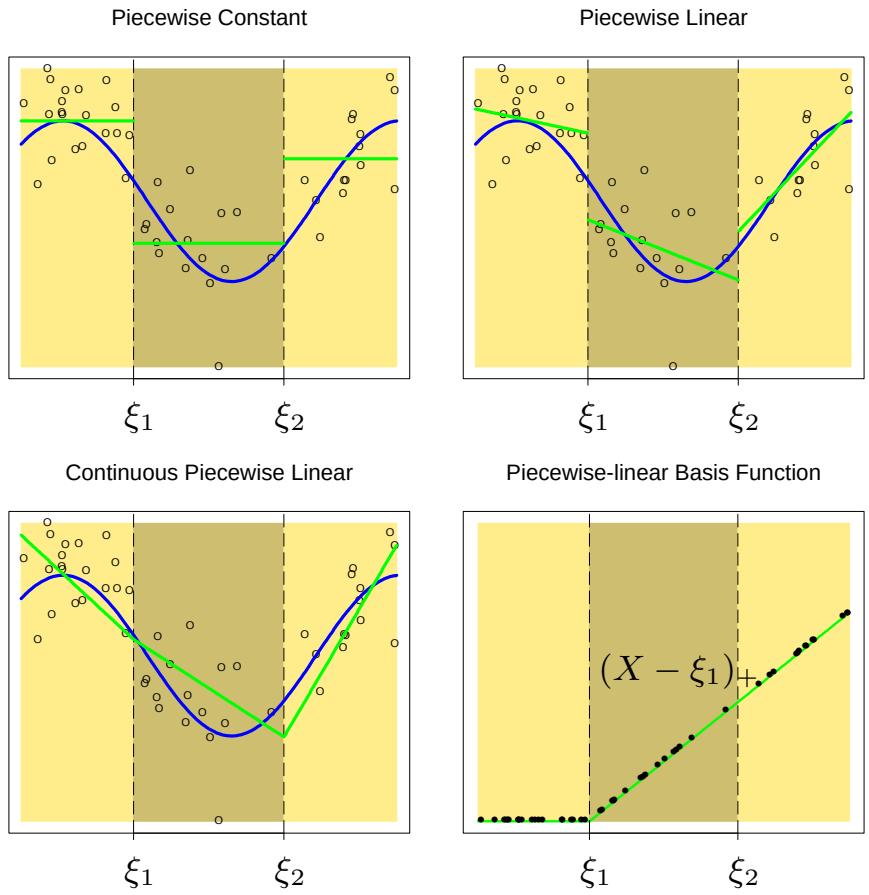


FIGURE 5.1. *The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two* knots *ξ_1 and ξ_2. The blue curve represents the true function, from which the data were generated with Gaussian noise. The remaining two panels show piecewise linear functions fit to the same data—the top right unrestricted, and the lower left restricted to be continuous at the knots. The lower right panel shows a piecewise-*

To smooth out these splines, we can include higher order terms. To improve the fit at the knots we also look at higher order terms there. This leads us to the basis for a **cubic spline** with two knots ξ_1 and ξ_2 as:

$$
\begin{aligned}
h_1(X) &= 1 \\
h_2(X) &= X \\
h_3(X) &= X^2 \\
h_4(X) &= X^3 \\
h_5(X) &= (X - \xi_1)_+^3 \\
h_6(X) &= (X - \xi_2)_+^3
\end{aligned}
$$

h_5 and h_6 are examples of a *truncated power basis*.
$(X - \xi_i)_+^3 : (X - \xi_i)^3$, if $X > \xi_i$; 0 otherwise.
Cubic spline: continuous 1st and 2nd derivatives at knots ξ_i.

©2021 The Trustees of the Stevens Institute of Technology

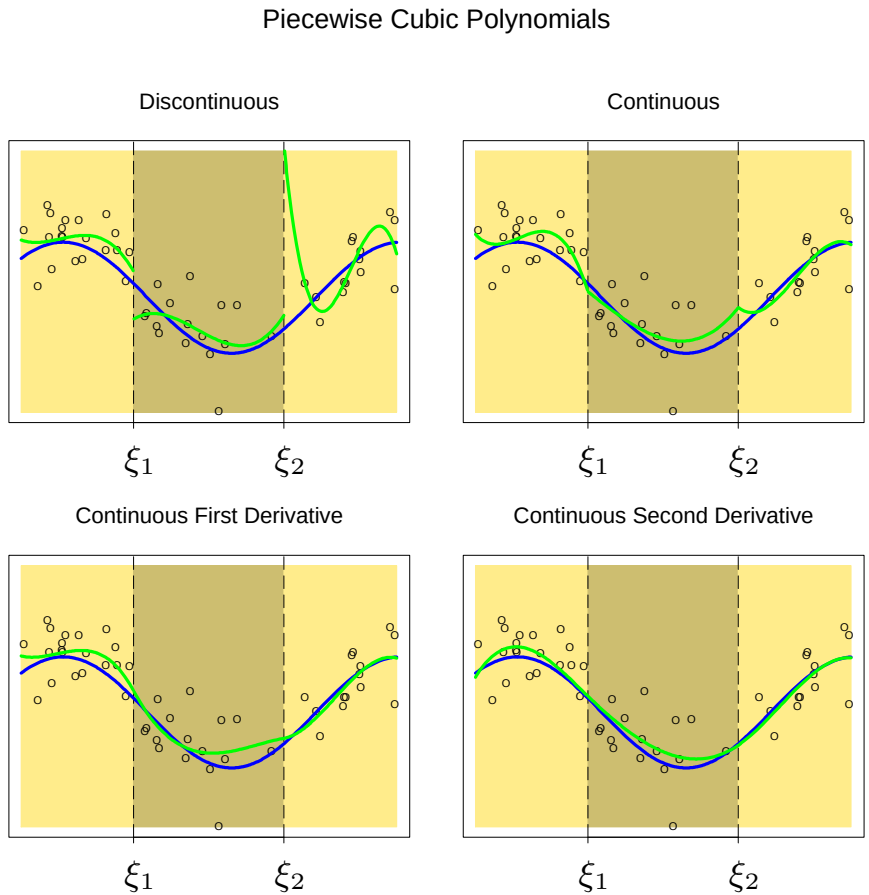


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

We can express this generally for an M-order spline with knots $\xi_j, j = 1, \ldots, K$ (which will have continuous derivatives up to order $M - 2$) as:

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

a cubic spline has $M = 4$ and has continuous derivatives up to order $M - 2$. Most common splines' orders are M = 1, 2 and 4. These fixed knot splines are also known as **regression splines**[1]
Suggestion: parameterize a family of splines by the number of basis functions or degrees of freedom, and have the observations x_i determine the positions of the knots.

©2021 The Trustees of the Stevens Institute of Technology

Splines cause the fitting of the polynomials in the area near the boundary to be very erratic.

A **natural cubic spline** adds the additional constraints that the function is linear or extrapolates linearly beyond the boundary knots: second derivative of spline polynomial are set equal to zero at end points of interval of interpolation.

This adds 4=2*2 extra constraints, and allows to put more internal knots for the same degrees of freedom as a regular cubic spline.

The tradeoff of this assumption is an increase in the bias near the boundaries. However, assuming that the function is linear near the boundaries is reasonable.

©2021 The Trustees of the Stevens Institute of Technology

A natural cubic spline with K knots is represented by K basis functions. Starting with the basis for cubic splines and reducing the basis by imposing the boundary constraints gives us:

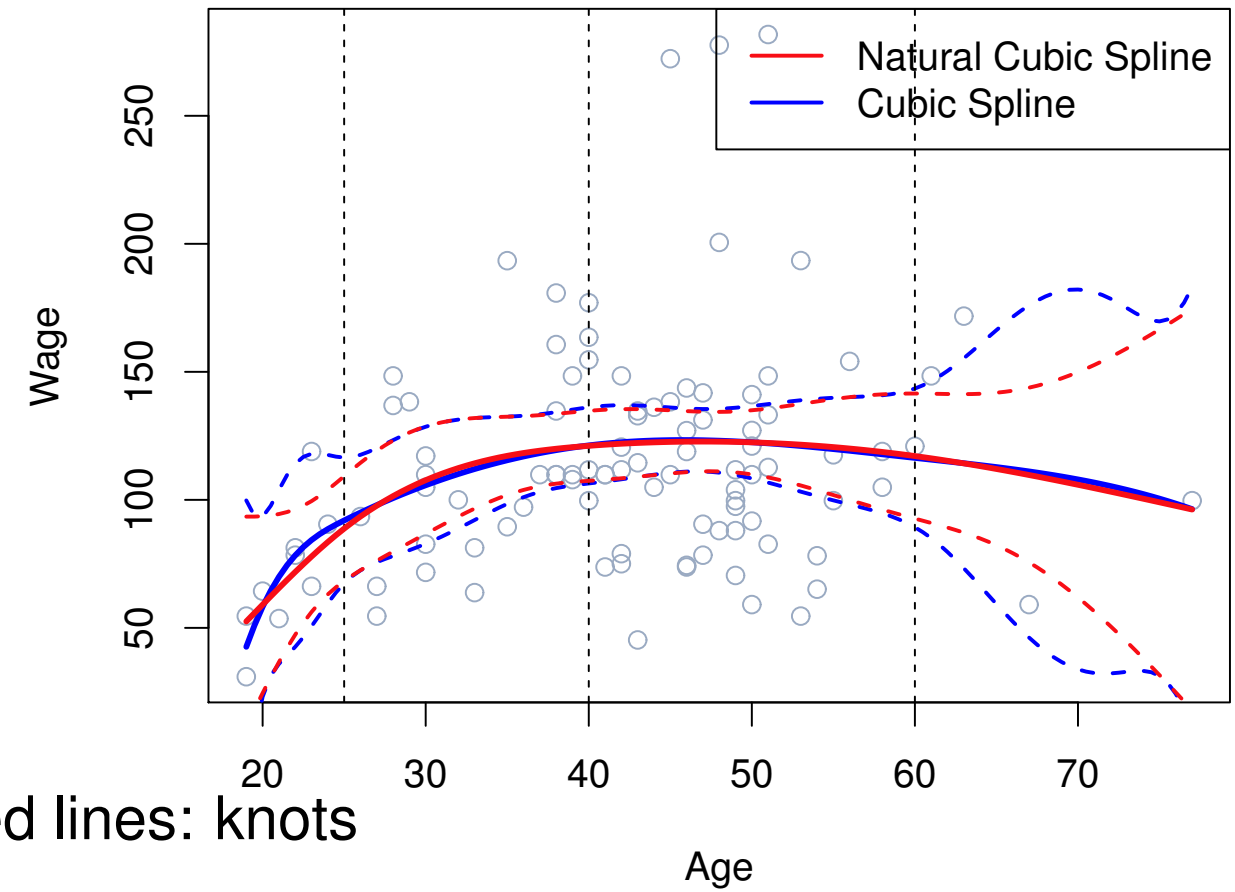
$$\begin{aligned} N_1(X) &= 1 \\ N_2(X) &= X \\ N_{k+2}(X) &= d_k(X) - d_{K-1}(X) \end{aligned}$$

where

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

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

©2021 The Trustees of the Stevens Institute of Technology



Dashed lines: knots

Smoothing Splines

An approach that avoids the knot selection problem (by using the maximal set). Among all functions f that have two continuous derivatives, find the one that minimizes

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

where λ is a fixed **smoothing parameter**: larger λ leads to a smoother f, and controls bias-variance trade-off.
Similar to regularization (Ridge or Lasso): first term: accuracy; second term: penalizes curvature or instability in the 2nd. derivative or changes of the slope (1st derivative). This penalty is on the spline coefficients: shrunk some toward the linear fit.
$\lambda = 0$: f can be any function that interpolates the data
$\lambda = \infty$: simple least squares fit (no second order derivative tolerated). Perfectly smooth.
f: function f that minimizes RSS: smoothing spline.

©2021 The Trustees of the Stevens Institute of Technology

It can be shown that this representation of $RSS(f, \lambda)$ has an explicit, finite-dimensional unique minimizer which is a natural cubic spline with knots at $x_i, i = 1, \ldots, N$

Since the solution is a natural spline, we can express it as

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

where the $N_j(x)$ are an N-dimensional set of basis functions. θ_j is the parameter j.

©2021 The Trustees of the Stevens Institute of Technology

The criterion reduces to:

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

where $\{\mathbf{N}\}_{ij} = N_j(x_i)$ and $\{\Omega_N\}_{jk} = \int N_j''(t)N_k''(t)dt$. The solution is:

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

a generalized Ridge regression.

The fitted smoothing spline is:

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

©2021 The Trustees of the Stevens Institute of Technology

Smoother Matrices

A smoothing spline with a prechosen λ is a **linear smoother**: parameters are a liner combination of y_i. Denote $\hat{\mathbf{f}}$ the N-vector of fitted values $\hat{f}(x_i)$ at the predictors x_i:x

$$\hat{\mathbf{f}} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T\mathbf{y} = \mathbf{S}_\lambda\mathbf{y}$$

Fit is linear in $\mathbf{y}$. The finite linear operator $\mathbf{S}_\lambda$ is known as the **smoother matrix** (and only depends on λ and x_i)

©2021 The Trustees of the Stevens Institute of Technology

Fixing the Degrees of Freedom

The **effective degrees of freedom** is given by:

$$\begin{aligned} df_\lambda &= \text{trace}(\mathbf{S}_\lambda) \\ &= \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii} \end{aligned}$$

Sum of the diagonal elements of $\mathbf{S}_\lambda$.
If we fix the degrees of freedom, we can specify a λ or λ controls df_λ

In R, we can specify the amount of smoothing through the degrees of freedom such as in commands like:
smooth.spline(x,y,df=6)

©2021 The Trustees of the Stevens Institute of Technology

Bias-Variance Tradeoff

Since $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$,

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

The diagonal contains the pointwise variances at the training x_i. The bias is given by:

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

where $\mathbf{f}$ is the vector of evaluations of the true f at the training X's

©2021 The Trustees of the Stevens Institute of Technology

The integrated squared prediction error (EPE) combines both bias and variance in a single summary:

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

©2021 The Trustees of the Stevens Institute of Technology

In reality we don't know the real function and so can't use the EPE. Instead we could look at some of our alternate methods, such as the LOOCV (N-fold):

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

©2021 The Trustees of the Stevens Institute of Technology

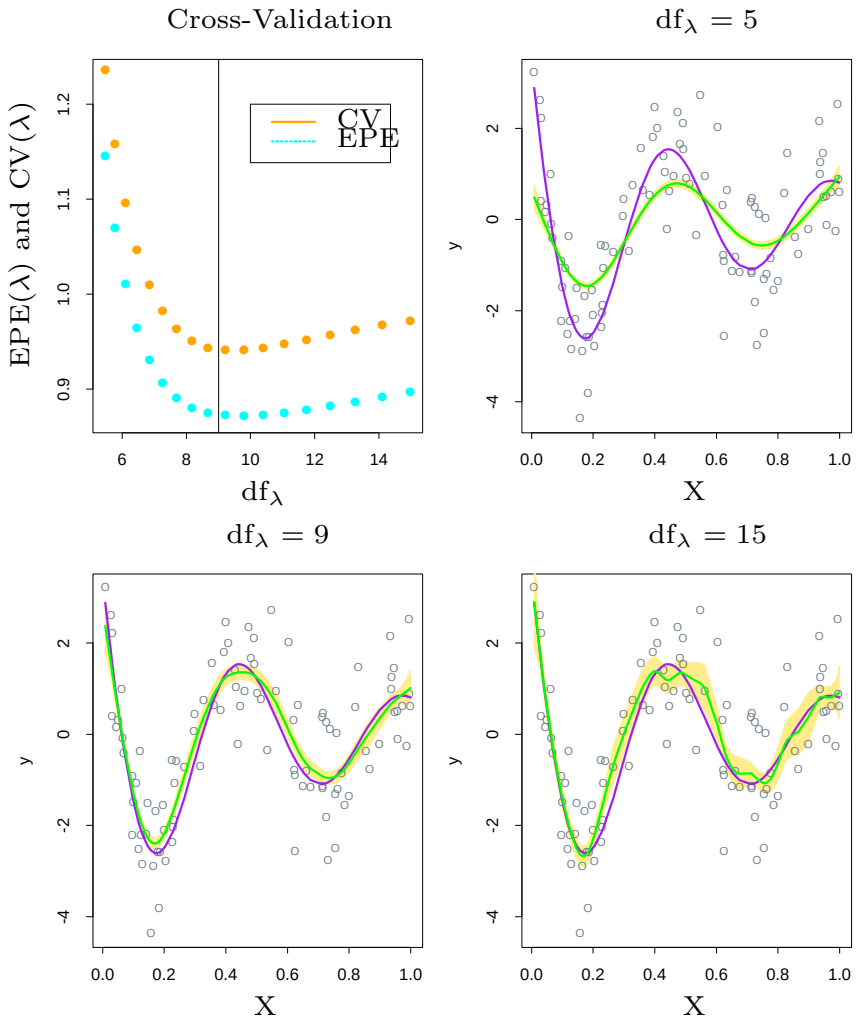


FIGURE 5.9. *The top left panel shows the* EPE(λ) *and* CV(λ) *curves for a realization from a nonlinear additive error model (5.22). The remaining panels show the data, the true functions (in purple), and the fitted curves (in green) with yellow shaded* $\pm 2\times$ *standard error bands, for three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

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer-Verlag, New York, 2 edition, 2008.

©2021 The Trustees of the Stevens Institute of Technology