Chapter 3

Splines

3.1 Introduction

Consider again the regression model

$$y_i = m(x_i) + e_i \ , \qquad i = 1, 2, ..., n,$$

where the e_i are assumed to be independently and identically distributed with $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$. As before we wish to estimate the function $m(x)$, the conditional expectation of the response variable, y, when the covariate takes on the value x. For convenience we will assume that the observations have been reordered so that $x_1 \leq x_2 \leq \cdots \leq x_n$. In the previous chapter we have considered two general approaches to the problem of estimating m, namely parametric modelling and kernel regression. The latter, being local in nature, is more flexible than parametric-based methods. In this chapter we consider another flexible approach to estimate m, based on splines.

The term spline comes from the ship building domain. It is the name used for a flexible strip of wood that is used to draw smooth curves through a set of points on a section of the ship. In that context the curves pass through all the points and are referred to as "interpolating splines". In our context we will not require the curves to pass through all the points. Instead we wish to use splines to provide a smooth curve that describes the general shape of the relationship between x and y. In the following, we will consider regression splines and smoothing splines.

3.2 Regression Splines

The idea here is to partition the range of the covariate into intervals and to fit a polynomial to the data in each interval. In what follows we restrict our attention to the case of **cubic functions**, which turn out to have desirable properties.

One begins by defining a convenient interval $[a, b]$ that contains all the x–values and then defines K values, called **knots**, $a < \xi_1 < \xi_2 < ... < \xi_K < b$ which specify the boundaries of the partition of $[a, b]$. For convenience we also define $\xi_0 = a$ and $\xi_{K+1} = b$, which are called the exterior knots. We can now fit a polynomial to the observations in each of the $K + 1$ intervals $(\xi_k, \xi_{k+1}]$, $k = 0, 1, .., K$. A cubic polynomial is determined by 4 parameters, and so the estimator, $\hat{m}$, depends on $4(K + 1)$ parameters.

Panel 1 at the top left in Figure 3.1 illustrates the case with $K = 2$ knots. In its unconstrained form this model has $3 \times 4 = 12$ parameters.

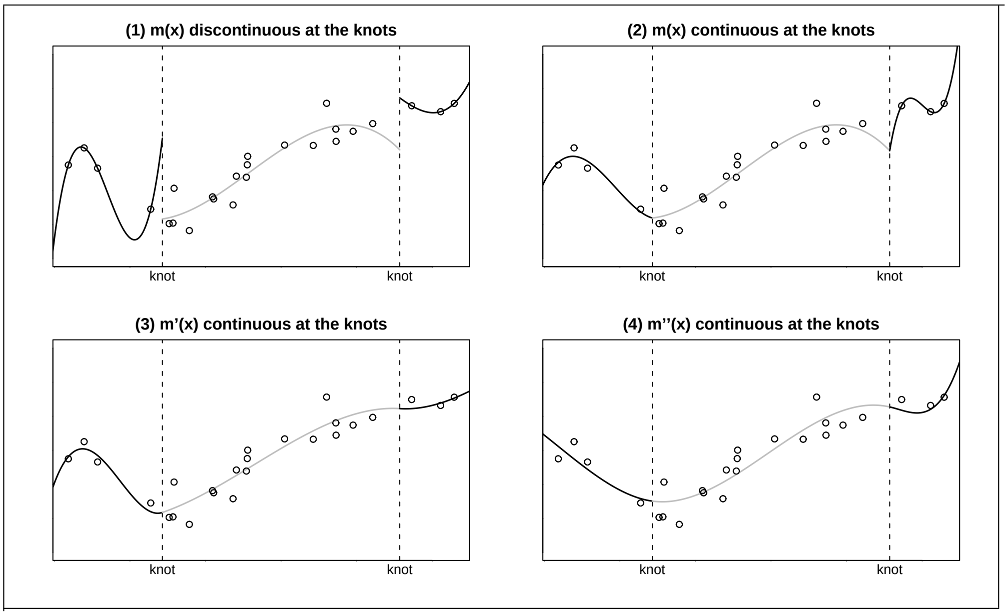

Figure 3.1: *Regression splines with increasing degrees of continuity.*

The resulting curve is not continuous. In most applications we wish to assume that m is continuous. We can ensure that $\hat{m}$ is continuous by imposing K restrictions on the polynomials, namely that they meet at the interior knots, $\xi_1, \xi_2, \ldots, \xi_K$. These restrictions reduce the number of free parameters by K. Panel 2 of the figure shows the resulting smooth, which illustrates that continuity alone does not guarantee that the resulting curve, $\hat{m}$, looks smooth. To make it smoother we can impose K further restrictions on the polynomials, namely that the derivatives at the interior knots should be equal. This improves the appearance of the estimate, but it still doesn't look entirely smooth (see Panel 3). This can be improved by requiring that the second derivatives at the interior knots should also be equal. This results in K further restrictions and so the number of free parameters has been reduced to $4(K + 1) - 3K = K + 4$. The resulting curve is called a cubic spline and has the following properties:

- it is a cubic polynomial on each interval $(\xi_k, \xi_{k+1}]$, $k = 0, 1, .., K$,

- it is smooth in the sense that $m(x)$, $m'(x)$ and $m''(x)$ are continuous on $(a, b]$,
- $m'''(x)$ is piecewise constant.

One can further reduce the number of parameters to $K + 2$ by imposing the additional restrictions that $\hat{m}$ is a straight line for $x \leq \xi_0$ and for $x > \xi_{K+1}$, i.e. $m''(\xi_0) = 0$ and $m''(\xi_{K+1}) = 0$. The result is known as a **natural cubic spline**.

3.2.1 Details for the case $K = 1$

We illustrate the above ideas for the case in which there is a single knot ξ. For convenience we assume that $\xi = 0$.

In its unconstrained form the model has $4(K + 1) = 8$ parameters and the conditional expectation is of the form:

$$m(x) = \left\{ \begin{array}{lcll} A(x) & = & \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 & \text{for} \quad x \leq 0, \\ B(x) & = & \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 & \text{for} \quad x > 0. \end{array} \right.$$

The continuity conditions impose restrictions on the 8 parameters:

$$\begin{array}{lll} \text{continuity of } m(x) \text{ at } \xi = 0 & \Longrightarrow A(0) = B(0) & \Longrightarrow \alpha_0 = \beta_0, \\ \text{continuity of } m'(x) \text{ at } \xi = 0 & \Longrightarrow A'(0) = B'(0) & \Longrightarrow \alpha_1 = \beta_1, \\ \text{continuity of } m''(x) \text{ at } \xi = 0 & \Longrightarrow A''(0) = B''(0) & \Longrightarrow \alpha_2 = \beta_2. \end{array}$$

Thus there are only 5 free parameters, namely $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_3$:

$$m(x) = \left\{ \begin{array}{lcll} A(x) & = & \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 & \text{for} \quad x \leq 0, \\ B(x) & = & \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \beta_3 x^3 & \text{for} \quad x > 0. \end{array} \right.$$

Writing $\beta_3 = \alpha_3 + \theta$ one has that

$$m(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \theta (x)_+^3, \quad \text{where } (x)_+ = \left\{ \begin{array}{ll} 0 & \text{for } x \leq 0, \\ x & \text{for } x > 0. \end{array} \right.$$

Suppose that the first m x-values are smaller than or equal to zero and the remaining $n - m$ are greater than zero. Thus the model can be represented as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & 0 \\ 1 & x_2 & x_2^2 & x_2^3 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & x_m^3 & 0 \\ 1 & x_{m+1} & x_{m+1}^2 & x_{m+1}^3 & x_{m+1}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_{m+1}^3 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \theta \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \\ e_{m+1} \\ \vdots \\ e_n \end{pmatrix}$$

This is a linear model and has the form

$$y = X\eta + e \ , \tag{3.1}$$

where η represents the vector of parameters. The least squares estimator of η, and the fitted values, are given by

$$\begin{aligned} \hat{\eta} &= (X'X)^{-1}X'y & (3.2)\\ \hat{y} &= X(X'X)^{-1}X'y & (3.3)\\ &= Sy \end{aligned}$$

The above expression for $\hat{y}$ shows that the fitted values are linear functions of the observations, i.e. $\hat{y}$ is a linear smooth of the y–values.

3.2.2 The general case with K knots

It can be shown that a cubic spline with K knots at $\xi_1, \xi_2, ..., \xi_K$ can be represented by the formula

$$m(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \sum_{K=1}^{K} \theta_k (x - \xi_k)_+^3 \ , \qquad \text{where} \quad z_+ = \begin{cases} 0 & \text{if} \quad z \le 0 \\ z & \text{if} \quad z > 0 \end{cases}$$

Thus the model can be represented as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1-\xi_1)_+^3 & \dots & (x_1-\xi_K)_+^3 \\ 1 & x_2 & x_2^2 & x_2^3 & (x_2-\xi_1)_+^3 & \dots & (x_2-\xi_K)_+^3 \\ 1 & x_3 & x_3^2 & x_3^3 & (x_3-\xi_1)_+^3 & \dots & (x_3-\xi_K)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n-\xi_1)_+^3 & \dots & (x_n-\xi_K)_+^3 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \theta_1 \\ \vdots \\ \theta_K \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

This is also a linear model of the form (3.1), namely a multivariate regression with the basis $P_1(x) = 1$, $P_2(x) = x$, $P_3(x) = x^2$, $P_4(x) = x^3$, $P_5(x) = (x - \xi_1)_+^3$, $\dots$, $P_{K+4}(x) = (x - \xi_K)_+^3$. The least squares estimator of η is given by (3.2), and the fitted values are given by (3.3). However, this estimator is numerically unstable because, in many applications, the matrix $(X'X)$ is nearly singular. In practice one uses so-called B–bases instead, which lead to the same fit in theory, but which are more stable numerically. Details of B-bases and their computations are given, for example, in Chapter 2 of Hastie and Tibshirani (1990). The `R`–function `bs` in the software library `splines` can be used to fit regression splines using B–bases.

3.2.3 Example

Figure 3.2 gives scatterplots of the annual salary of the chief executive officers for 59 small highly ranked firms plotted against the covariate age (source: Forbes, November 8, 1993, "America's Best Small Companies"). The panels in the figure display four regression splines, each based on two knots, which have been placed at different points.

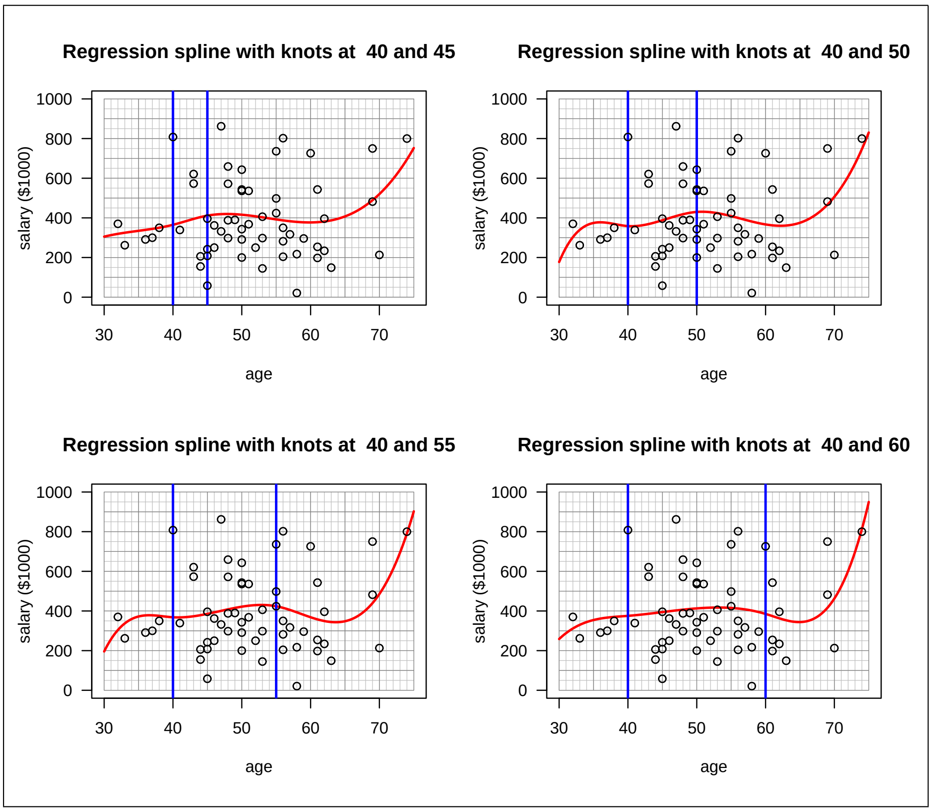

Figure 3.2: *Regression splines with two knots, placed at different points.*

The `R`–library `splines` was used to create the displays in Figure 3.2.

Note that the fitted regression spline depends on the number of knots and where the knots are placed. I.e. one has to decide how many knots to use, and where to place them. In practice the knots are usually placed at appropriate quantiles of the covariate. For example for $K = 3$ the knots are placed at the three quartiles of the x–values. Regarding the number of knots to use one has to balance the bias against the variance of the estimator. The number of knots, K, determines the number of parameters and thus plays a similar role to the degree of the polynomial in polynomial regression, or to the bandwidth in kernel regression. In general, increasing K leads to a reduction of the bias. That's because the increased number of parameters allows for more flexibility of the estimator. However, increasing K also increases the variance.

3.3 Smoothing Splines

Our objective is to estimate m by means of a function that (a) fits the data well, and (b) is as smooth as possible. A measure of smoothness of m is the integral of the square of its second derivative. The following criterion takes both (a) and (b) into account:

$$\sum_{i=1}^{n}(y_i - m(x_i))^2 + \lambda \int_a^b (m''(x))^2 \, dx \tag{3.4}$$

where $\lambda > 0$ is a fixed constant and $x_i \in [a, b]$, $i = 1, 2, \ldots, n$.

The first term is the sum of squares of the residuals; it provides a measure of how well the function m fits the data. The integral is a measure for the roughness/smoothness of the function m. Functions which are highly curved will result in a large value of the integral; straight lines result in the integral being zero. The roughness penalty, λ, controls how much emphasis one wishes to place on smoothness. By increasing λ one places more emphasis on smoothness; as λ becomes large the function approaches a straight line. On the other hand a small value of λ emphasises the fit of m to the data points; as λ approaches zero m approaches a function that interpolates the data points. A remarkable result is that the criterion (3.4) can be minimized analytically:

> **Result**: Among all functions with two continuous derivatives, there is a unique function that minimizes criterion (3.4); it is a natural cubic spline with knots at the unique values of x.

If all the values of the covariate are different, then the natural cubic spline has as many knots as there are observations, i.e. $K = n$. That means that the function that minimizes criterion (3.4) would seem to be overparameterized. In fact this is not the case because the parameters are highly dependent, thereby rendering the "effective number of parameters" much smaller than n. This is a consequence of the smoothing requirement which is imposed by using $\lambda > 0$.

3.3.1 Degrees of freedom

The smoothing parameter, λ, determines the degree of smoothing. It plays a similar role as did the bandwidth in the case of kernel regression. In general, selecting a small value of λ leads to a small bias but to a large variance. On the other hand, increasing λ increases the bias but reduces the variance. The value of λ that is estimated to minimize the mean squared error can be estimated using the method of cross-validation (we will return to this in the next sub-section). Although λ is the controlling parameter, its value is not easy to interpret. An easier (equivalent) quantity to interpret is the number of "degrees of freedom".

The concept of "degrees of freedom" is straight–forward in the context of parametric linear models. Thus, for example, if we fit a simple linear regression model

$$y_i = \alpha_0 + \alpha_1 x_i + e_i \ , \quad i = 1, 2, ..., n \ ,$$

using the method of least squares, then the estimated residuals

$$\hat{e}_i = y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i) \ , \quad i = 1, 2, ..., n \ ,$$

can be described as points in a $(n-2)$ dimensional space, they have $(n-2)$ degrees of freedom. In general if we fit a parametric model with p free parameters then the estimated residuals have $n-p$ degrees of freedom; the model "uses up" p degrees of freedom.

In the context of non–parametric models the notion of degrees of freedom has to be generalized because one does not have a well–defined fixed number of parameters as one does in the parametric case. A number of such generalizations have been proposed. We will restrict our attention to the following that is applicable to linear smoothers, i.e. smoothers for which the fitted values are of the form $\hat{y} = Sy$, where S is a given $n \times n$ matrix.

> The degrees of freedom of a linear smoother is the trace of S,$tr(S)$.

This definition is consistent with the familiar definition for degrees of freedom for a parametric model, $y = X\beta + e$. To see this first note that the matrix S of a paramteric model is given by $X(X'X)^{-1}X'$, and it holds that $tr(S) = rank(X)$. We assume that there are more observations than there are parameters (i.e. that $p < n$), and that the covariates are linearly independent (i.e. that it is not possible to compute all the values of a given covariate as a linear combination of the others). Then rank$(X) = p$, and so the definiton of the degrees of freedom associated with the model is given by $df = tr(S) = rank(X) = p$, i.e. the number of parameters in the model; the estimated residuals have $n-p$ degrees of freedom.

In the case of spline smoothing it is more convenient to specify the degrees of freedom than it is to specify λ. That's because λ depends on the units that are used to quantify the data. Secondly, by specifying the degrees of freedom we can compare the resulting smooth with parametric regressions having the same degrees of freedom. The degrees of freedom determine λ, and vice versa.

3.3.2 Example

We illustrate the use of smoothing splines using the data set `cars` that is provided in **R** (for details key in `help(cars)`). Given are the speed of cars and the distances taken to stop. Figures 3.3 and 3.4 contrast parametric regression and smoothing splines for the car data with comparable degrees of freedom.

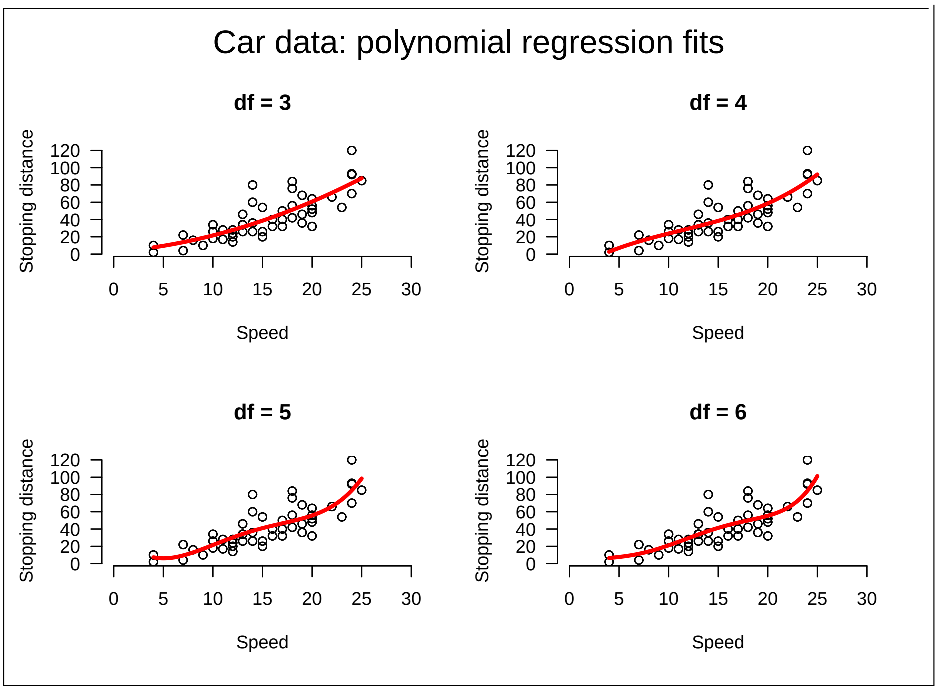

Figure 3.3: *Polynomial regression of different orders.*

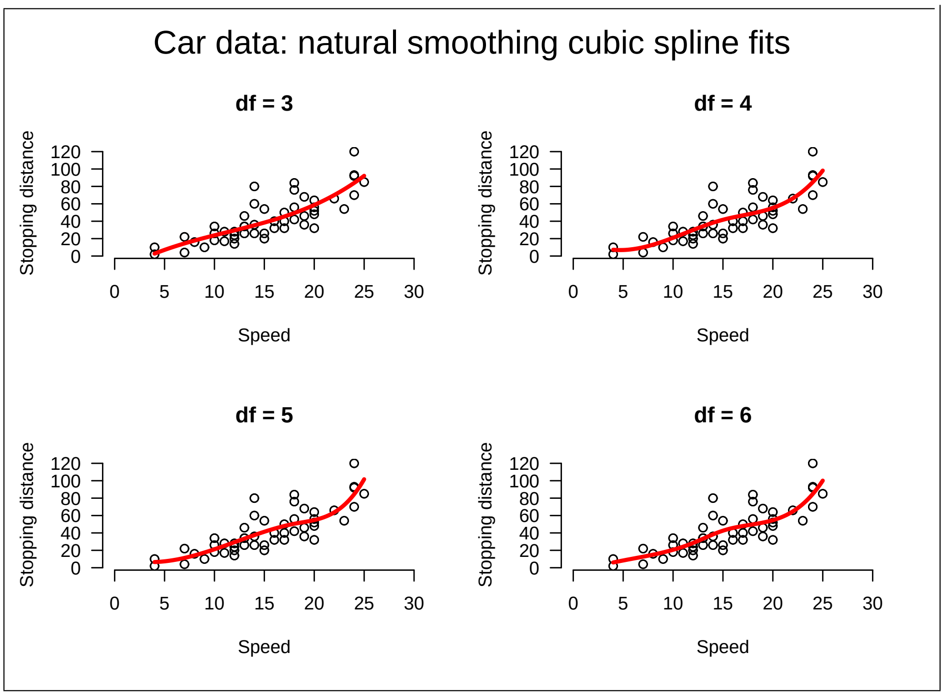

Figure 3.4: *Smoothing splines with different degrees of freedom.*

Figure 3.3 was created using the standard commands for parametric regression provided by `R`, and 3.4 was obtained applying the `bs`–command contained in the `splines`–library.

3.4 Cross–validation for linear smoothers

The method of cross-validation, discussed in Chapter 2, is applicable for determining the value of λ (or equivalently the degrees of freedom) that is estimated to minimize the integrated prediction squared error. One minimizes the one-item-out cross-validation criterion with respect to λ:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{m}_\lambda^{(-i)}(x_i) \right)^2$$

where $\hat{m}_\lambda^{(-i)}$ represents the estimator of m using the original sample but with the i-th observation left out.

Thus to compute $CV(\lambda)$ it would seem that one needs to fit the model n times for each value of λ. In fact it is not necessary to do so for *linear smoothers*, it is only necessary to fit the model once for each value of λ. Recall that for linear smoothers, which include cubic smoothing splines, the fitted values are of the form $\hat{y} = Sy$, where S is an $n \times n$ matrix of weights whose rows sum to one. It can be shown (see, e.g., Hastie and Tibshirani (1990), Section 3.4.3) that

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{m}_\lambda(x_i)}{1 - S_{ii}} \right)^2$$

where S_{ii} is the i-th diagonal element of the matix S. Thus once S has been computed for a given λ it is easy to compute $CV(\lambda)$. Of course this presupposes that an efficient algorithm exists to compute S. Such an algorithm does indeed exist for smoothing splines.