Chapter 11
Spline and Kernel Methods

11.1 Introduction

Spline models are based on *piecewise* polynomial fitting, while kernel regression models are based on *local* polynomial fitting. These two approaches to modeling are extremely popular, and so we dedicate a whole chapter to their description.

The layout of this chapter is as follows. In Sect. 11.2, a variety of approaches to spline modeling are described, while Sect. 11.3 discusses kernel-based methods. For inference, an estimate of the error variance is required; this topic is discussed in Sect. 11.4. In this chapter we concentrate on a single x variable only. However, we do consider general responses and, in particular, the class of generalized linear models. Approaches for these types of data are described in Sect. 11.5. Concluding comments appear in Sect. 11.6. There is an extensive literature on spline and kernel modeling; Sect. 11.7 gives references to key contributions and book-length treatments.

11.2 Spline Methods

11.2.1 Piecewise Polynomials and Splines

For *continuous responses*, splines are simply linear models, with an enhanced basis set that provides flexibility.[1] Let $h_j(x) : \mathbb{R} \to \mathbb{R}$ denote the jth function of x, for $j = 1, \ldots, J$. A generic linear model consists of the *linear basis expansion* in x:

$$f(x) = \sum_{j=1}^{J} \beta_j h_j(x).$$

[1] Appendix C gives a brief review of bases.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_11,

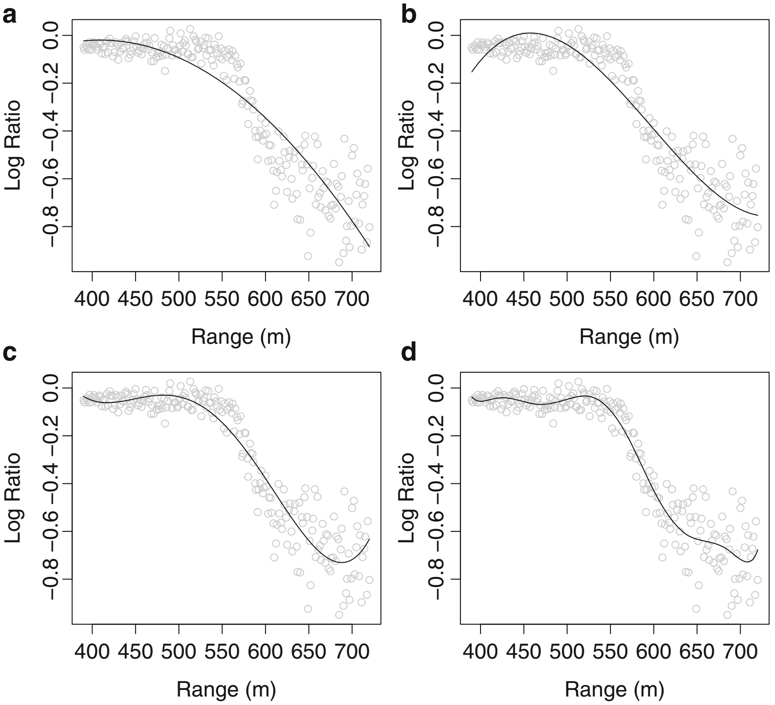

Fig. 11.1 Polynomial fits to the LIDAR data: (**a**) quadratic, (**b**) cubic, (**c**) quartic, and (**d**) degree-8 polynomial

An obvious choice of basis is a polynomial of degree $J - 1$, but the global behavior of such a choice can be poor in the sense that the polynomial will not provide a good fit over the complete range of x. However, *local* behavior can be well represented by relatively low-order polynomials.

Example: Light Detection and Ranging

Figure 11.1 shows degree 2, 3, 4, and 8 polynomial fits to the LIDAR data. The quadratic and cubic models fit very badly, while the quartic model produces a poor fit for ranges of 500–560 m. The degree-8 polynomial fit is also not completely satisfactory with wiggles at the extremes of the range variable due to the global nature of the fitting.

To motivate spline models, we fit piecewise-constant, linear, quadratic, and cubic models using least squares, with three pieces in each case. The fits are displayed in Fig. 11.2. We focus on the piecewise linear model, as shown in Fig. 11.2(b). By forcing the curve to be continuous but only allowing linear segments, we see that

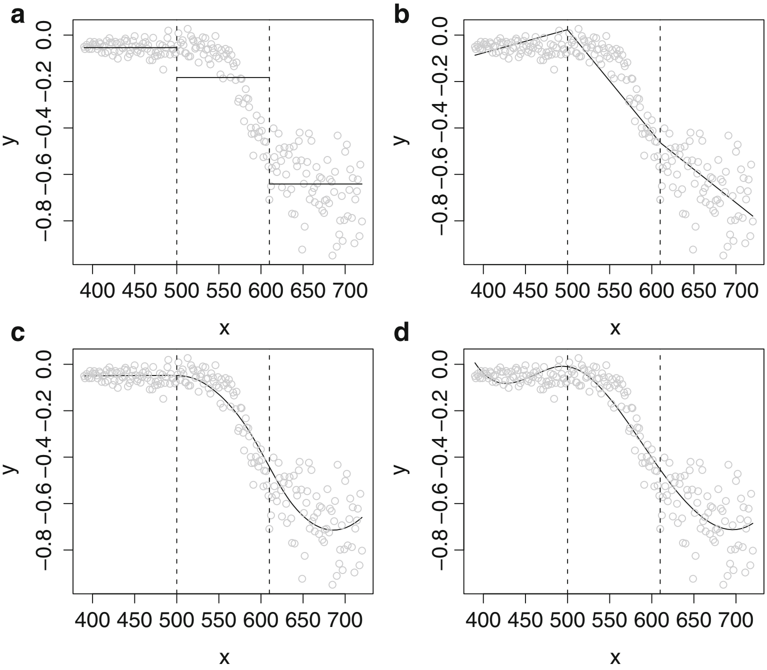

Fig. 11.2 Piecewise polynomials for the LIDAR data: (**a**) constant, (**b**) linear, (**c**) quadratic, and (**d**) cubic

the fit is not good (particularly in the first segment). The lack of smoothness is also undesirable. The quadratic and cubic fits in panels (c) and (d) are far more appealing visually, though neither provide satisfactory fits because we have only allowed three piecewise polynomials. In particular, in panel (d), the cubic fit is still poor at the left endpoint. □

We now start the description of spline models by introducing some notation. Let $\xi_1 < \xi_2 < \ldots < \xi_L$ be a set of ordered points, called *knots*, contained in some interval $[a, b]$. An *M-th order spline* is a piecewise $M - 1$ degree polynomial with $M - 2$ continuous derivatives at the knots.[2] Splines are very popular in nonparametric modeling though, as we shall see, care is required in choosing the degree of smoothing. The latter depends on a variety of factors including the order of the spline and the number and position of the knots.

We begin with a discussion on the order of the spline. The most basic piecewise polynomial is a piecewise-constant function, which is a first-order spline. With two knots, ξ_1 and ξ_2, one possible set of three basis functions is

[2] From the Oxford dictionary, a *spline* is a "flexible wood or rubber strip, for example, used in drawing large curves especially in railway work."

$$h_1(x) = I(x < \xi_1), \quad h_2(x) = I(\xi_1 \le x < \xi_2), \quad h_3(x) = I(\xi_2 \le x)$$

where $I(\cdot)$ is the indicator function. Note that there are no continuous derivatives at the knots; Fig. 11.2(a) clearly shows the undesirability of this aspect.

To obtain linear models in each of the intervals, we may introduce three additional bases

$$h_{3+j} = h_j(x)x, \quad j = 1, 2, 3,$$

to give the model

$$f(x) = I(x < \xi_1)(\beta_1 + \beta_4 x) + I(\xi_1 \le x < \xi_2)(\beta_2 + \beta_5 x) + I(\xi_2 \le x)(\beta_3 + \beta_6 x),$$

which contains six parameters. Lack of continuity is a problem with this model, but we can impose two constraints to enforce $f(\xi_1^-) = f(\xi_1^+)$ and $f(\xi_2^-) = f(\xi_2^+)$, which imply the two conditions

$$\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$$
$$\beta_2 + \xi_2\beta_5 = \beta_3 + \xi_2\beta_6,$$

to give four parameters in total. A neater way of incorporating these constraints is with the basis set:

$$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = (x - \xi_1)_+, \quad h_4(x) = (x - \xi_2)_+ \tag{11.1}$$

where t_+ denotes the positive part. We refer to the generic basis $(x - \xi)_+$ as a *truncated line*.[3] The resultant function

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)_+ + \beta_3(x - \xi_2)_+$$

is continuous at the knots since all prior basis functions are contributing to the fit up to any single x value. The model defined by the basis (11.1) is an order-2 spline, and the first derivative is discontinuous. Figure 11.3 shows the basis functions for this representation and Fig. 11.2(b) the fit of this model to the LIDAR data.

We now consider how the piecewise linear model may be extended. Naively, we might assume the quadratic form:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(x - \xi_1)_+ + \beta_4(x - \xi_1)_+^2 + \beta_5(x - \xi_2)_+ + \beta_6(x - \xi_2)_+^2, \tag{11.2}$$

which is continuous but has first derivative

$$f'(x) = \beta_1 + 2\beta_2 x + \beta_3 I(x > \xi_1) + 2\beta_4(x - \xi_1)_+ + \beta_5 I(x > \xi_2) + 2\beta_6(x - \xi_2)_+,$$

[3] It is conventional to define the truncated lines with respect to bases that take the positive part, but we could have defined the same model with respect to bases taking the negative part.

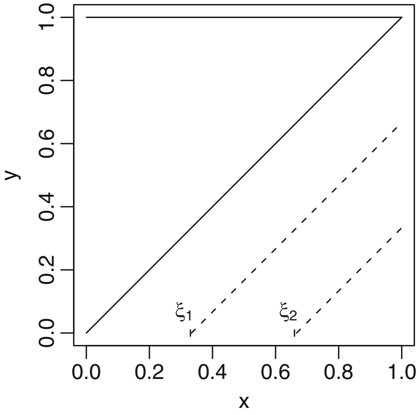

Fig. 11.3 Basis functions for a piecewise linear model with two knots at ξ_1 and ξ_2. The *solid lines* are the bases 1 and x, and the *dashed lines* are the bases $(x - \xi_1)_+$ and $(x - \xi_2)_+$

which is discontinuous at the knot points ξ_1 and ξ_2 due to the linear truncated bases associated with β_3 and β_5 in (11.2). This lack of smoothness at the knots is undesirable. Hence, we drop the truncated linear bases to give the regression model

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \xi_1)_+^2 + \beta_4 (x - \xi_2)_+^2$$

which has continuous first derivative:

$$f'(x) = \beta_1 + 2\beta_2 x + 2\beta_3 (x - \xi_1)_+ + 2\beta_4 (x - \xi_2)_+.$$

The second derivative is discontinuous, however, which may also be undesirable. Consequently, a popular form (which we justify more rigorously shortly) is a cubic spline. We will concentrate on cubic splines in some detail, and so we introduce a slight change of notation with respect to the truncated cubic parameters. With two knots the function and first three derivatives are

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + b_1 (x - \xi_1)_+^3 + b_2 (x - \xi_2)_+^3 \\ f'(x) &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + 3b_1 (x - \xi_1)_+^2 + 3b_2 (x - \xi_2)_+^2 \\ f''(x) &= 2\beta_2 + 6\beta_3 x + 6b_1 (x - \xi_1)_+ + 6b_2 (x - \xi_2)_+ \\ f'''(x) &= 6\beta_3 + 6b_1 I(x > \xi_1) + 6b_2 I(x > \xi_2). \end{aligned}$$

The latter is discontinuous, with a jump at the knots. Figure 11.4 shows the basis functions for the cubic spline, with two knots, and Fig. 11.2(d) shows the fit to the LIDAR data.

For L knots, we write the cubic spline function as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{l=1}^{L} b_l (x - \xi_l)_+^3, \tag{11.3}$$

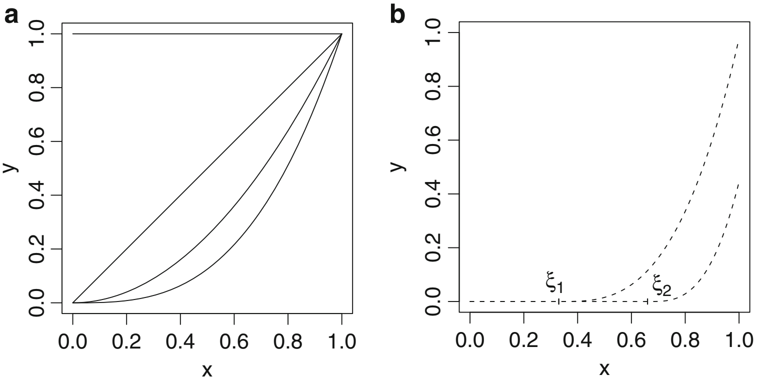

Fig. 11.4 Basis functions for a piecewise cubic spline model with two knots at ξ_1 and ξ_2. Panel (**a**) shows the bases 1, x, x^2, and x^3 and panel (**b**) the bases $(x-\xi_1)_+^3$ and $(x-\xi_2)_+^3$. Note that in (**b**) the bases have been scaled in the vertical direction for clarity

so that we have $L+4$ coefficients. The key to implementation is to recognize that we simply have a linear model, $f(x) = \text{E}[\boldsymbol{Y} \mid \boldsymbol{z}] = \boldsymbol{z}\boldsymbol{\gamma}$, where $\boldsymbol{z} = \boldsymbol{z}(x)$ and

$$\boldsymbol{z} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1-\xi_1)_+^3 & \cdots & (x_1-\xi_L)_+^3 \\ 1 & x_2 & x_2^2 & x_2^3 & (x_2-\xi_1)_+^3 & \cdots & (x_2-\xi_L)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n-\xi_1)_+^3 & \cdots & (x_n-\xi_L)_+^3 \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ b_1 \\ \vdots \\ b_L \end{bmatrix}.$$

The obvious estimator is therefore $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{z}^{\text{T}}\boldsymbol{z})^{-1}\boldsymbol{z}^{\text{T}}\boldsymbol{Y}$, which gives the linear smoother $\widehat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$, where $\boldsymbol{S} = \boldsymbol{z}(\boldsymbol{z}^{\text{T}}\boldsymbol{z})^{-1}\boldsymbol{z}^{\text{T}}$.

11.2.2 Natural Cubic Splines

Spline models such as (11.3) can produce erratic behavior beyond the extreme knots. A *natural* spline enforces linearity beyond the boundary knots, that is,

$$f(x) = a_1 + a_2 x \quad \text{for} \quad x \le \xi_1$$

$$f(x) = a_3 + a_4 x \quad \text{for} \quad x \ge \xi_L.$$

The first condition only considers values of x before the knots, and therefore, the b_l parameters in (11.3) are irrelevant. Consequently, it is straightforward to see that we require

$$\beta_2 = \beta_3 = 0. \tag{11.4}$$

For $x \geq \xi_L$,

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{l=1}^{L} b_l (x - \xi_l)^3 \\ &= \beta_0 + \beta_1 x + \sum_{l=1}^{L} b_l (x^3 - 3x^2 \xi_l + 3x\xi_l^2 - \xi_l^3), \end{aligned}$$

and so, for linearity,

$$\sum_{l=1}^{L} b_l = \sum_{l=1}^{L} b_l \xi_l = 0. \tag{11.5}$$

Hence, we have four additional constraints in total, so that the basis for a natural cubic spline has L elements. Exercise 11.3 describes an alternative basis.

11.2.3 Cubic Smoothing Splines

So far we have examined splines in a heuristic way, as flexible functions with certain desirable properties in terms of the continuity of the function and the first and second derivatives at the knots. We now present a formal justification for the natural cubic spline.

Result. Consider the penalized least squares criterion

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int f''(x)^2 dx, \tag{11.6}$$

where the second term penalizes the *roughness* of the curve and λ controls the degree of this roughness. It is clear that without the penalization term, we could choose an infinite number of curves that interpolate the data (in the case of unique x values, at least), with arbitrary behavior in between. Quite remarkably, the $f(\cdot)$ that minimizes (11.6) is the *natural cubic spline* with knots at the unique data points; we call this function $g(x)$.

Proof. The proof has two parts and is based on Green and Silverman (1994, Chap. 2). We begin by showing that a natural cubic spline minimizes (11.6) amongst all interpolating functions and then extend to non-interpolating functions. Assume that $x_1 < \ldots < x_n$. We consider all functions that are continuous in $[x_1, x_n]$ with continuous first and second derivatives and which interpolate $[x_i, y_i]$, $i = 1, \ldots, n$. Since the first term of (11.6) is zero, we need to show that the natural cubic spline, $g(x)$, minimizes

$$\int_{x_1}^{x_n} f''(x)^2 dx.$$

Let $\widetilde{g}(x)$ be another interpolant of (x_i, y_i), and define $h(x) = \widetilde{g}(x) - g(x)$. Then,

$$\begin{aligned}\int_{x_1}^{x_n} \widetilde{g}\,''(x)^2 \, dx &= \int_{x_1}^{x_n} \left[g''(x) + h''(x)\right]^2 \, dx \\ &= \int_{x_1}^{x_n} \left[g''(x)^2 + 2g''(x)h''(x) + h''(x)^2\right] \, dx.\end{aligned}$$

Applying integration by parts to the cross term,

$$\begin{aligned}\int_{x_1}^{x_n} g''(x)h''(x)dx &= \left[g''(x)h'(x)\right]_{x_1}^{x_n} - \int_{x_1}^{x_n} g'''(x)h'(x) \, dx \\ &= -\int_{x_1}^{x_n} g'''(x)h'(x) \, dx \;\; \text{since} \;\; g''(x_1) = g''(x_n) = 0 \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x) \, dx \\ &\quad \text{since } g'''(x) \text{ is constant in, and } x_i^+ \text{ is a point in, } [x_i, x_{i+1}] \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+) \left[h(x_{i+1}) - h(x_i)\right] \\ &= 0\end{aligned}$$

since $h(x_{i+1}) = \widetilde{g}(x_{i+1}) - g(x_{i+1})$ and both are interpolants (and similarly for $h(x_i)$). We have shown that

$$\begin{aligned}\int_{x_1}^{x_n} \widetilde{g}''(x)^2 dx &= \int_{x_1}^{x_n} g''(x)^2 dx + \int_{x_1}^{x_n} h''(x)^2 \, dx \\ &\geq \int_{x_1}^{x_n} g''(x)^2 \, dx\end{aligned}$$

with equality if and only if $h''(x) = 0$ for $x_1 < x < x_n$. The latter implies $h(x) = a + bx$, but $h(x_1) = h(x_n) = 0$, and so $a = b = 0$. Consequently, any interpolant that is not identical to $g(x)$ will have a higher integrated squared second derivative. Therefore, the natural cubic spline with knots at the unique x values is the smoothest interpolant in the sense of minimizing $\int f''(x)^2 \, dx$. This is of use in, for example, numerical analysis, where interpolation of $[x_i, y_i]$ is of interest. But, in statistical applications, the data are measured with error, and we typically do not wish to restrict attention to interpolating functions.[4]

[4] There are some analogies here with bias, variance, and mean squared error. The penalized sum of squares (11.6) is analogous to the mean squared error, and interpolating functions are "unbiased"

We have shown that a natural cubic spline minimizes (11.6) amongst all interpolating functions but the minimizing function need not necessarily be an interpolant since an interpolating function may have a large associated penalty contribution. The second part of the proof considers functions that do not necessarily interpolate the data but have n free parameters $g(x_i)$ with the aim being minimization of (11.6). The resulting $g(x)$ is known as a *smoothing spline*. Suppose some function $f^\star(x)$, other than the cubic smoothing spline, minimizes (11.6). Let $g(x)$ be the natural cubic spline that interpolates $[\,x_i, f^\star(x_i)\,]$, $i = 1, \ldots, n$. Obviously, $f^\star$ and g produce the same residual sum of squares in (11.6) since $f^\star(x_i) = g(x_i)$. But, by the first part of the proof,

$$\int f^{\star''}(x)^2 dx > \int g''(x)^2 dx.$$

Hence, the natural cubic spline is the function that minimizes (11.6); this spline is known as a *cubic smoothing spline*.

The above result has shown us that if we wish to minimize (11.6), we should take as model class the cubic smoothing splines. The coefficient estimates of the fit will depend on the value chosen for λ. We stress that the fitted natural cubic smoothing spline will not typically interpolate the data, and the level of smoothness will be determined by the value of λ chosen. Small values of λ, which correspond to a large effective degrees of freedom (Sect. 10.5.1), impose little smoothness and bring the fit closer to interpolation, while large values will result in the fit being close to linear in x (in the limit, a zero second derivative is required).

In terms of interpretation, if a thin piece of flexible wood (a mechanical spline) is placed over the points $[\,x_i, y_i\,]$, $i = 1, \ldots, n$, then the position taken up by the piece of wood will be of minimum energy and will describe a curve that approximately minimizes $\int f''^2$ over curves that interpolate the data.

Example: Light Detection and Ranging

We fit a natural cubic spline to the LIDAR data. Figure 11.5 shows the ordinary and generalized cross-validation scores (as described in Sects. 10.6.2 and 10.6.3, respectively) versus the effective degrees of freedom. The curves are very similar with well-defined minima since these data are abundant and the noise level is relatively low. The OCV and GCV scores are minimized at 9.3 and 9.4 effective degrees of freedom, respectively. Figure 11.6 shows the fit (using the GCV minimum corresponding to $\widehat{\lambda} = 959$), which appears good. In particular, we note that the boundary behavior is reasonable.

but may have large variability. However, we can obtain a better estimator if we are prepared to examine "biased" (i.e., non-interpolating) functions.

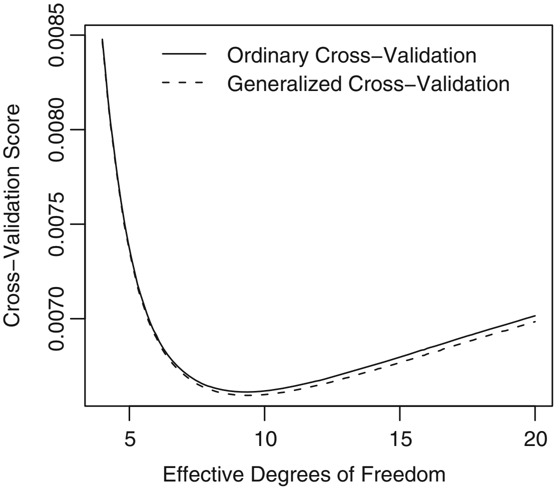

Fig. 11.5 Ordinary and generalized cross-validation scores versus effective degrees of freedom for the LIDAR data and a natural cubic spline model

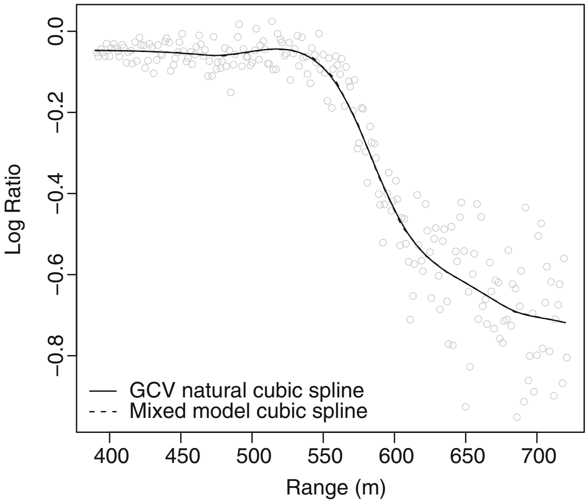

Fig. 11.6 Cubic spline fits to the LIDAR data. The natural cubic spline fit has smoothing parameter chosen by generalized cross-validation. The mixed model cubic spline has smoothing parameter chosen by REML

11.2.4 B-Splines

There are many ways of choosing a basis to represent a cubic spline; the so-called B-spline basis functions are popular, a primary reason being that they are nonzero over a limited range which aids in computation. B-splines also form the building blocks for other spline models as we describe in Sect. 11.2.5. The classic text on B-splines is de Boor (1978).

B-splines are available for splines of general order, which we again denote by M (so that for a cubic spline, $M = 4$). The number of basis functions is $L + M$ since we have an $M-1$ degree polynomial (giving M bases) and one basis for each knot. The original set of knots are denoted ξ_l, $l = 1, \ldots, L$, and we let $\xi_0 < \xi_1$ and $\xi_L < \xi_{L+1}$ represent two boundary knots. We define an augmented set of knots, τ_j, $j = 1, \ldots, L + 2M$, with

$$\begin{aligned} &\tau_1 \leq \tau_2 \leq \ldots \leq \tau_M \leq \xi_0 \\ &\qquad\qquad\tau_{j+M} = \xi_j,\ j = 1, \ldots, L \\ &\qquad\qquad\xi_{L+1} \leq \tau_{L+M+1} \leq \tau_{L+M+2} \leq \cdots \leq \tau_{L+2M} \end{aligned}$$

where the choice of the additional knots is arbitrary and so we may, for example, set $\tau_1 = \ldots = \tau_M = \xi_0$ and $\xi_{L+1} = \tau_{L+M+1} = \ldots = \tau_{L+2M}$. These additional knots ensure the basis functions detailed below are defined close to the boundaries. To construct the bases, first define

$$B_j^1(x) = \begin{cases} 1 \text{ if } \tau_j \leq x < \tau_{j+1} \\ 0 \text{ otherwise} \end{cases} \tag{11.7}$$

for $j = 2, \ldots, L + 2M - 1$. For $1 < m \leq M$, define

$$B_j^m(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_j^{m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1}^{m-1} \tag{11.8}$$

for $j = 1, \ldots, L + 2M - m$. If we divide by zero, then we define the relevant basis element to be zero. The B-spline bases are nonzero over a domain spanned by at most $M + 1$ knots. For example, the support of cubic B-splines ($M = 4$) is at most five knots. At any x, M of the B-splines are nonzero.

The cubic B-spline model is

$$f(x) = \sum_{j=1}^{L+4} B_j^4(x)\beta_j. \tag{11.9}$$

For further details on computation, see Hastie et al. (2009, p. 186). Figure 11.7 shows the cubic B-spline basis (including the intercept) for $L = 9$ knots.

11.2.5 Penalized Regression Splines

Although the result of Sect. 11.2.3 is of theoretical interest, in general, we would like to have a functional form that has less parameters than data points. ***Regresssion splines*** are defined with respect to a reduced set of $L < n$ knots. Automatically deciding on the number and location of knots is difficult. For example, starting with n knots and then selecting via stepwise methods (Sect. 4.8.1) is fraught with difficulties since there are 2^n models to choose from (assuming the intercept

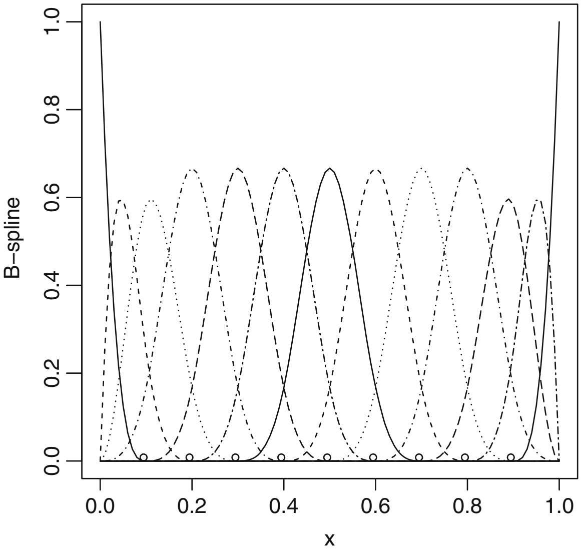

Fig. 11.7 B-spline basis functions corresponding to a cubic spline ($M = 4$) with $L = 9$ equally spaced knots (whose positions are shown as *open circles* on the x-axis). There are $L + M = 13$ bases in total. Note that six distinct line types are used so that, for example, there are three splines represented by *solid curves*: the leftmost, the central, and the rightmost

and linear terms are always present). An alternative *penalized regression spline* approach, with $L < n$ knots, is to choose sufficient knots for flexibility and then to penalize the parameters associated with the knot bases. If this approach is followed, the number and selection of knots is far less important than the choice of smoothing parameter. An obvious choice is to place an L_2 penalty on the coefficients, that is, to include the term $\lambda \sum_{l=1}^{L} b_l^2$ in a penalized least squares form. So-called *low-rank* smoothers use considerably fewer than n basis functions.

We now consider linear smoothers of the form:

$$f(x) = \sum_{j=1}^{J} h_j(x)\beta_j = \boldsymbol{h}(x)\boldsymbol{\beta},$$

where $\boldsymbol{h}(x)$ is a $1 \times J$ vector. A general *penalized* regression spline is $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{h}(x)$, where $\widehat{\boldsymbol{\beta}}$ is the minimizer of

$$\sum_{i=1}^{n} (y_i - \boldsymbol{h}_i\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\beta}, \tag{11.10}$$

with $\boldsymbol{h}_i = \boldsymbol{h}(x_i)$, $\boldsymbol{D}$ is a symmetric-positive semi-definite matrix, and $\lambda > 0$ is a scalar. If we let $\boldsymbol{h} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n]^{\mathrm{T}}$ represent the $n \times J$ design matrix, then

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} + \lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{Y}. \tag{11.11}$$

The penalty

$$\lambda \int f''(x)^2 dx \tag{11.12}$$

is of the form (11.10) since, for a linear smoother $f(x)$,

$$\begin{aligned}\int f''(x)^2 dx &= \boldsymbol{\beta}^{\mathrm{T}} \left[\int \boldsymbol{h}''(x) \boldsymbol{h}''(x)^{\mathrm{T}} dx \right] \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\beta}\end{aligned}$$

with $\boldsymbol{D}$ a matrix of known coefficients. The penalty is measuring complexity: For $\lambda = 0$, there is no cost to fitting a very complex function, while $\lambda = \infty$ gives the simple linear least squares line.

O'Sullivan splines (O'Sullivan 1986) use the cubic B-spline basis representation (11.9), combined with the penalty (11.12), which takes the form:

$$\lambda \int \left(\sum_{j=1}^{L+4} B_j^4(x)'' \beta_j \right)^2 dx.$$

Hence, the penalty matrix $\boldsymbol{D}$ has (j,k)-th element $\int B_j^4(x)'' B_k^4(x)''\, dx$. O'Sullivan splines correspond to cubic smoothing splines for $L = n$ and distinct x_i (Green and Silverman 1994, Sect. 3.6).

The construction of P-splines is based on a different penalty in which a set of B-spline basis functions are used with a collection of equally spaced knots (Eilers and Marx 1996). The form of the penalty is

$$\lambda \sum_{j=k+1}^{J} (\Delta^k \beta_j)^2 \tag{11.13}$$

with $\Delta\beta_j = \beta_j - \beta_{j-1}$, the difference operator, and where k is a positive integer. For $k = 2$, the penalty is

$$\lambda \sum_{j=1}^{J-1} (\beta_{j+1} - \beta_j)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 + \ldots + 2\beta_{J-1}^2 - 2\beta_{J-1}\beta_J + \beta_J^2,$$

which corresponds to the general penalty $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\beta}$ with

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

This form penalizes large changes in adjacent coefficients, providing an alternative representation of smoothing. The P-spline approach was heavily influenced by the derivation of O'Sullivan splines (O'Sullivan 1986), and the P-spline penalty is an approximation to the integrated squared derivative penalty. See Eilers and Marx (1996) for a careful discussion of the two approaches. Wand and Ormerod (2008) also contrast O'Sullivan splines (which they refer to as *O-splines*) with P-splines and argue that O-splines are an attractive option for nonparametric regression.

With respect to penalized regression splines, a number of suggestions exist for the number and location of the knots. For example, Ruppert et al. (2003) take as default choice:

$$L = \min\left(\frac{1}{4} \times \text{ number of unique } x_i, 35\right),$$

with knots ξ_l taken at the $(l+1)/(L+2)$th points of the unique x_i. These authors say that these choices "work well in most of the examples we come across" but urge against the unquestioning use of these rules.

11.2.6 A Brief Spline Summary

The terminology associated with splines can be confusing, so we provide a brief summary. For simplicity, we assume that the covariate x is univariate and that $x_1, \ldots, x_n$ are unique. A *smoothing spline* contains n knots, and a *cubic smoothing spline* is piecewise cubic. A *natural spline* is linear beyond the boundary knots. If there are $L < n$ knots, we have a *regression spline*. A *penalized regression spline* imposes a penalty on the coefficients associated with the piecewise polynomial. The penalty terms may take a variety of forms.

The number of basis functions that define the spline depends on the number of knots and the degree of the polynomial; natural splines have a reduced number of bases. Spline models may be parameterized in many different ways.

11.2.7 Inference for Linear Smoothers

Nonparametric regression may be used for a variety of purposes. The simplest use is as a scatterplot smoother for pure exploration. In such a context, a plot of $\widehat{f}(x)$ versus x is perhaps all that is required. In other instances, we may wish to produce interval estimates, either pointwise or simultaneous, in order to examine the uncertainty as a function of x.

We consider linear smoothers with J basis functions and write $f(x) = \boldsymbol{h}(x)\boldsymbol{\beta}$ for a prediction at x with $\boldsymbol{\beta}$ a $J \times 1$ vector and $\boldsymbol{h}(x)$ the $J \times 1$ design matrix associated with x. Further, assume $Y(x) = f(x) + \epsilon(x)$, with the error terms $\epsilon(x)$ uncorrelated and with constant variance σ^2. We emphasize that J is not equal to the effective degrees of freedom, which is given by $p^{(\lambda)} = \text{tr}\left[\boldsymbol{h}(\boldsymbol{h}^{\text{T}}\boldsymbol{h} + \lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\text{T}}\right]$ where $\boldsymbol{h} = [\,\boldsymbol{h}(x_1), \ldots, \boldsymbol{h}(x_n)\,]^{\text{T}}$. Differentiation of (11.10) with respect to $\boldsymbol{\beta}$ and setting equal to zero gives

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{h}^{\text{T}}\boldsymbol{h} + \lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\text{T}}\boldsymbol{Y}.$$

Assuming a fixed λ, asymptotic inference for $\boldsymbol{\beta}$ is straightforward since

$$\left[(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}+\lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}+\lambda\boldsymbol{D})^{-1}\right]^{-1/2}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\to \mathrm{N}_J(\mathbf{0},\sigma^2\mathbf{I}).$$

In a nonparametric regression context, interest often focuses on inference for the underlying function; we first consider inference at a single point x, $f(x)$.

Since the estimator is linear in the data,

$$\widehat{f}(x)=\boldsymbol{h}(x)\widehat{\boldsymbol{\beta}}=\boldsymbol{S}(x)\boldsymbol{Y}=\sum_{i=1}^{n}S_i(x)Y_i \tag{11.14}$$

where $\boldsymbol{S}(x)=\boldsymbol{h}(x)(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}+\lambda\boldsymbol{D})^{-1}\boldsymbol{h}^{\mathrm{T}}$ is the $1\times n$ vector with elements $S_i(x)$, $i=1,\ldots,n$. This estimator has mean

$$\mathrm{E}\left[\widehat{f}(x)\right]=\sum_{i=1}^{n}S_i(x)f(x_i)$$

and variance

$$\mathrm{var}\left(\widehat{f}(x)\right)=\sigma^2\sum_{i=1}^{n}S_i(x)^2=\sigma^2||\boldsymbol{S}(x)||^2. \tag{11.15}$$

A major difficulty with (11.14) is that there will be bias $b(x)$ present in the estimator. If this bias were known, then

$$\frac{\widehat{f}(x)-f(x)-b(x)}{\sigma||S(x)||}\to_d \mathrm{N}(0,1), \tag{11.16}$$

via a central limit theorem. Note that it is "local" sample size that is relevant here, with a precise definition depending on the smoothing technique used (which defines $S(x)$. Estimation of the bias is difficult since it involves estimation of $f''(x)$ (for a derivation in the context of density estimation, see Sect. 11.3.4).

Often the bias is just ignored. The interpretation of the resultant confidence intervals is that they are confidence intervals for $\overline{f}(x)=\mathrm{E}\left[\widehat{f}(x)\right]$, which may be thought of as a smoothed version of $f(x)$. We have

$$\begin{aligned}\frac{\widehat{f}(x)-f(x)}{\sigma||\boldsymbol{S}(x)||}&=\frac{\widehat{f}(x)-\overline{f}(x)}{\sigma||\boldsymbol{S}(x)||}+\frac{\overline{f}(x)-f(x)}{\sigma||\boldsymbol{S}(x)||}\\&=Z_n(x)+\frac{b(x)}{\sigma||\boldsymbol{S}(x)||},\end{aligned} \tag{11.17}$$

which is a restatement of (11.16) and where $Z_n(x)$ converges to a standard normal. Hence, a $100(1-\alpha)\%$ asymptotic confidence interval for $\overline{f}(x)$ is $\widehat{f}(x)\pm c_\alpha\sigma||\boldsymbol{S}(x)||$, where c_α is the appropriate cutoff point of a standard normal distribution. In parametric inference, the bias is usually much smaller than the standard deviation of the estimator, so the bias term goes to zero as the sample size increases.[5] In a smoothing context, we have repeatedly seen that optimal smoothing corresponds to balancing bias and variance, and the second term does not disappear from (11.17), even for large sample sizes (recall that $S(x)$ will depend on λ, whose choice will depend on sample size).

We now turn to simultaneous confidence bands of the function $f(x)$ over an interval $x \in [a, b]$ with $a = \min(x_i)$ and $b = \max(x_i)$, $i = 1, \ldots, n$. In the following, we will assume that the confidence bands are for the smoothed function $\overline{f}(x) = \mathrm{E}\left[\widehat{f}(x)\right]$, thus sidestepping the bias issue. We again assume linear smoothers so that (11.14) holds.

One way to think about a simultaneous confidence band is to begin with a finite grid of x values: $x_j = a + j(b - a)/m$, $j = 1, \ldots, m$. Now suppose we wish to obtain a simultaneous confidence band for $\overline{f}(x_j)$, $j = 1, \ldots, m$. One way of approaching this problem is to consider the probability that each of the m estimated functions simultaneously lie within c standard errors of $\overline{f}$, that is,

$$\bigcap_{j=1}^{m} \left\{ \left| \frac{\widehat{f}(x_j) - \overline{f}(x_j)}{\sigma ||S(x_j)||} \right| \leq c \right\},$$

where c is chosen to correspond to the required $1 - \alpha$ level of the confidence statement. Then

$$\Pr\left(\bigcap_{j=1}^{m} \left\{ \left| \frac{\widehat{f}(x_j) - \overline{f}(x_j)}{\sigma ||S(x_j)||} \right| \leq c \right\} \right) = \Pr\left(\max_{x_1,\ldots,x_m} \left| \frac{\widehat{f}(x_j) - \overline{f}(x_j)}{\sigma ||S(x_j)||} \right| \leq c \right). \tag{11.18}$$

Now suppose that $m \to \infty$ to give the limiting expression for (11.18) as

$$\Pr\left(\sup_{x \in [a,b]} \left| \frac{\widehat{f}(x) - \overline{f}(x)}{\sigma ||S(x)||} \right| \leq c \right) = \Pr(M \leq c).$$

Sun and Loader (1994), following Knafl et al. (1985), considered approximating this probability in the present context. Let $T(x) = S(x)/||S(x)||$. Based on the theory of Gaussian processes,

$$\Pr(M \geq c) \approx 2\left[1 - \Phi(c)\right] + \frac{\kappa_0}{\pi} \exp(-c^2/2),$$

where

$$\kappa_0 = \int_a^b ||T'(x)||\, dx,$$

[5] With parametric models, we are often interested in simple models with a fixed number of parameters, even if we know they are not "true". For example, when we carry out linear regression, we do not usually believe that the "true" underlying function is linear; rather, we simply wish to estimate the linear association.

$T'(x) = [T_1'(x), \ldots, T_n'(x)]^{\mathrm{T}}$ and $T_i'(x) = \partial T_i(x)/\partial x$ for $i = 1, \ldots, n$. We choose c to solve

$$\alpha = 2\left[1 - \Phi(c)\right] + \frac{\kappa_0}{\pi}\exp(-c^2/2), \tag{11.19}$$

and κ_0 may be evaluated using numerical integration over a grid of x values. To summarize, once an α level is chosen, we obtain κ_0 and c and then form bands $\widehat{f}(x) \pm c\sigma||\boldsymbol{S}(x)||$.

In the case of nonconstant variance, we replace σ by $\sigma(x)$. Section 11.4 contains details on estimation of the error variance. Throughout this section, we have conditioned upon a λ value, which is usually estimated from the data. Hence, in practice, the uncertainty in λ is not accounted for in the construction of interval estimates. A Bayesian mixed model approach (Sect. 11.2.9) treats λ as a parameter, assigns a prior, and then averages over the uncertainty in λ in subsequent inference.

In some contexts, interest may focus on testing the adequacy of a parametric model, comparing nested smoothing models, or testing whether the relationship between the expected response and x is flat. In each of these cases, likelihood ratio or F tests can be performed (see, e.g., Wood 2006, Sect. 4.8.5), though the nonstandard context suggests that the significance of test statistics should be judged via simulation.

Example: Light Detection and Ranging

We fit a cubic penalized regression spline, with penalization $\lambda\sum_{l=1}^{L} b_l^2$ and λ estimated using generalized cross-validation. Figure 11.8(a) gives pointwise confidence intervals and simultaneous confidence bands under the assumption of constant variance. Figure 11.8(b) presents the more appropriate intervals with allowance for nonconstant variance (for details on how $\sigma(x)$ is estimated, see the example at the end of Sect. 11.4). The coverage probability is 0.95, and the critical value for c is 1.96 for the pointwise intervals and 3.11 for the simultaneous intervals, as calculated from (11.19), with κ_0 estimated as 15.4. Under a nonconstant variance assumption, the intervals are very tight for low ranges and increase in width as the range increases.

11.2.8 Linear Mixed Model Spline Representation: Likelihood Inference

In this section we describe an alternative mixed model framework for the representation of regression spline models. A benefit of this framework is that the smoothing parameter may be estimated using standard inference (e.g., likelihood or Bayesian) techniques. It is also possible to build complex mixed models that can model dependencies within the data using random effects, in addition to performing

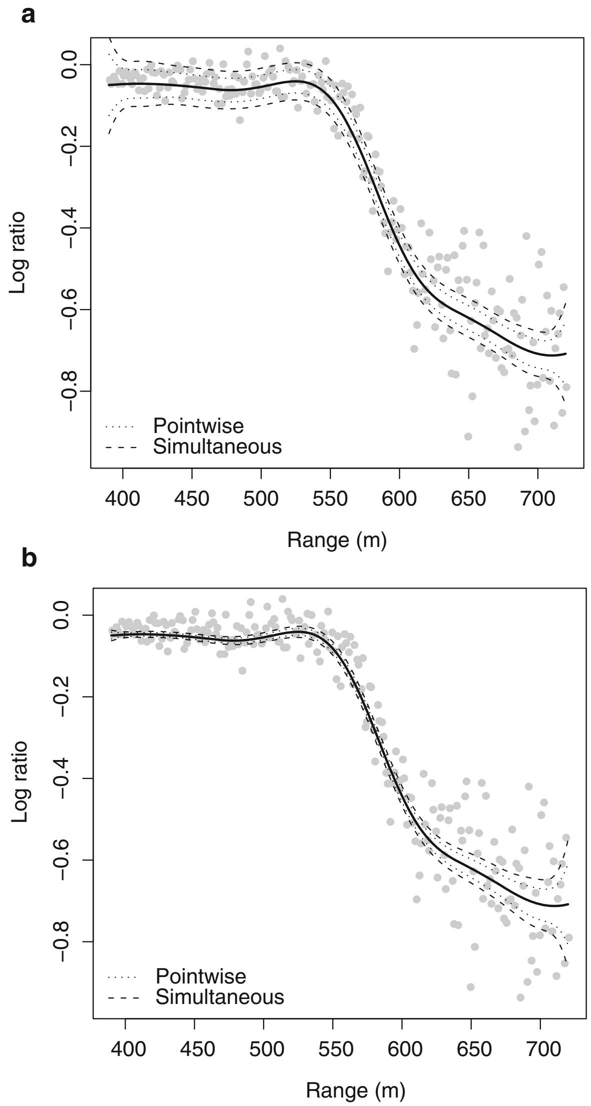

Fig. 11.8 Pointwise confidence intervals and simultaneous confidence bands for $\overline{f}(x)$ for the LIDAR data under the assumption of (**a**) homoscedastic errors and (**b**) heteroscedastic errors

the required smoothing. In the following, we lean heavily on the material on linear random effects modeling contained in Chap. 8. Consider the $(p+1)$th-order (degree p polynomial) penalized regression spline with L knots, that is,

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{l=1}^{L} b_l (x - \xi_l)_+^p.$$

A penalized least squares approach with L_2 penalization of the L truncated cubic coefficients leads to minimization of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b})^2 + \lambda\sum_{l=1}^{L} b_l^2, \tag{11.20}$$

where

$$\boldsymbol{x}_i = [1, x_i, \dots, x_i^p], \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{z}_i = [\,(x_i - \xi_1)_+^p, \dots, (x_i - \xi_L)_+^p\,], \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_L \end{bmatrix}.$$

Let $\boldsymbol{D} = \text{diag}(\boldsymbol{0}_{p+1}, \boldsymbol{1}_L)$ and $\boldsymbol{c}$ be the $n \times (p + 1 + L)$ matrix with ith row $\boldsymbol{c}_i = [1, x_i, \dots, x_i^p, (x_i - \xi_1)_+^p, \dots, (x_i - \xi_L)_+^p]$, so that $\boldsymbol{c} = [\boldsymbol{x}, \boldsymbol{z}]$, where $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^{\mathrm{T}}$ and $\boldsymbol{z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_n]^{\mathrm{T}}$. The penalized sum of squares (11.20) can be written as

$$(\boldsymbol{y} - \boldsymbol{c}\boldsymbol{\gamma})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{c}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\gamma}, \tag{11.21}$$

where $\boldsymbol{\gamma} = [\,\boldsymbol{\beta}, \boldsymbol{b}\,]^{\mathrm{T}}$.

We now reframe this approach in mixed model form with mean model

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i \\ &= \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b} + \epsilon_i, \end{aligned}$$

and covariance structure and distributional form determined by $\epsilon_i \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ and $b_l \mid \sigma_b^2 \sim_{iid} \text{N}(0, \sigma_b^2)$ with ϵ_i and b_l independent, $i = 1, \dots, n$, $l = 1, \dots, L$. This formulation sheds some light on the nature of the penalization. Since the distribution of b_l is independent of $b_{l'}$ for $l \neq l'$, we are assuming that the size of the contribution due to the lth basis is not influenced by any other contributions, in particular, the closest (in terms of x) basis. For example, knowing the sign of b_{l-1} does not imply we believe that b_l is of the same sign. This is in contrast to the P-spline difference penalty described in Sect. 11.2.5.

Minimization of (11.21) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ is then equivalent to minimization of

$$\frac{1}{\sigma_\epsilon^2}\left[(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{b}) + \frac{\sigma_\epsilon^2}{\sigma_b^2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right]$$

so that $\lambda = \sigma_\epsilon^2/\sigma_b^2$. We summarize likelihood-based inference for this linear mixed model; Sect. 8.5 contains background details. The maximum likelihood (ML) estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y} \tag{11.22}$$

where $\boldsymbol{V} = \sigma_b^2\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} + \sigma_\epsilon^2\mathbf{I}_n$, and the best linear unbiased predictor (BLUP) estimator/predictor of $\boldsymbol{b}$ is

$$\widehat{\boldsymbol{b}} = \sigma_b^2\boldsymbol{z}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) \tag{11.23}$$

Let $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_b^2$ be the restricted maximum likelihood (REML) estimators (see Sect. 8.5.3) of σ_ϵ^2 and σ_b^2 so that

$$\widehat{\lambda} = \left(\frac{\widehat{\sigma}_\epsilon^2}{\widehat{\sigma}_b^2}\right).$$

In practice, we use

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\boldsymbol{x}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{Y}\\ \widehat{\boldsymbol{b}} &= \widehat{\sigma}_b^2\boldsymbol{z}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}).\end{aligned}$$

The (penalized) estimator of $\boldsymbol{\gamma} = [\,\boldsymbol{\beta}, \boldsymbol{b}\,]^{\mathrm{T}}$ can be written as

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Y} \tag{11.24}$$

(Exercise 11.2). Hence, we can write the fitted values as the linear smoother:

$$\begin{aligned}\widehat{\boldsymbol{f}} &= \boldsymbol{c}\widehat{\boldsymbol{\gamma}} = \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}\\ &= \boldsymbol{c}(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Y}.\end{aligned}$$

The degrees of freedom of the model is defined as

$$\begin{aligned}\mathrm{df}(\lambda) &= \mathrm{tr}\left(\boldsymbol{S}^{(\lambda)}\right)\\ &= \mathrm{tr}\left[\boldsymbol{c}(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}^{\mathrm{T}}\right].\end{aligned} \tag{11.25}$$

We consider inference for a particular value x:

$$\begin{aligned}\widehat{f}(x) &= \boldsymbol{x}(x)\widehat{\boldsymbol{\beta}} + \boldsymbol{z}(x)\widehat{\boldsymbol{b}}\\ &= \boldsymbol{c}(x)\widehat{\boldsymbol{\gamma}}\\ &= \boldsymbol{c}(x)(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Y}\end{aligned}$$

where $\boldsymbol{x}(x) = [\,1, x, \ldots, x^p\,]$, $\boldsymbol{z}(x) = [\,(x - \xi_1)^p, \ldots, (x - \xi_L)^p\,]$ and $\boldsymbol{c}(x) = [\,\boldsymbol{x}(x), \boldsymbol{z}(x)\,]$.

The variance, conditional on $\boldsymbol{b}$, is

$$\mathrm{var}\left(\widehat{f}(x) \mid \boldsymbol{b}\right) = \sigma_\epsilon^2\boldsymbol{c}(x)(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} + \lambda\boldsymbol{D})^{-1}\boldsymbol{c}(x)^{\mathrm{T}},$$

which is identical to the variance obtained from ridge regression (10.30). Ruppert et al. (2003, Sect. 6.4) argue for conditioning on $\boldsymbol{b}$ to give the appropriate measure of variability. Specifically, they state (in the notation used here): "Randomness of $\boldsymbol{b}$ is a device used to model curvature, while ϵ accounts for variability about the curve." Asymptotic 95% pointwise confidence intervals for $f(x)$ are

$$\widehat{f}(x) \pm 1.96 \times \sqrt{\mathrm{var}\left(\widehat{f}(x)\right)}.$$

Approximate or fully Bayesian approaches to confidence interval construction for the complete curve have been recently advocated and have shown to be accurate in simulation studies; see Chap. 17 of Ruppert et al. (2003) and the detailed account of Marra and Wood (2012). These accounts build upon the work of Wabha (1983); Silverman (1985), and Nychka (1988). The latter showed, for univariate x, that a Bayesian interval estimate of the curve, constructed using a cubic smoothing spline, has good frequentist coverage probabilities when the bias in curve estimation is a small contributor to the overall mean squared error. In this case, the average posterior variance is a good approximation to the mean squared error of the collection of predictions. Marra and Wood (2012) provide a far-ranging discussion of Bayesian confidence interval construction, in the context of generalized additive models, as described in Sect. 12.2; included is a discussion of when the coverage probability of the interval is likely to be poor, one instance being when a relatively large amount of bias occurs, for example, when one over-smooths.

Tests of the adequacy of a parametric model or of a null association via likelihood ratio and F tests are described in Ruppert et al. (2003, Sects. 6.6 and 6.7). We illustrate confidence interval construction with an example.

Example: Light Detection and Ranging

We fit a cubic spline with 20 equally spaced knots (so that we have 4 fixed effects and 20 random effects) with REML estimation of the smoothing parameter. The resultant fit is shown in Fig. 11.6 as a dashed line. The variance components are estimated as $\widehat{\sigma}_\epsilon^2 = 0.079^2$ and $\widehat{\sigma}_b^2 = 0.012^2$, to give smoothing parameter $\widehat{\lambda} = 45.8$, which equates to an effective degrees of freedom of 8.5. This is quite similar to the effective degrees of freedom of 9.4 that was chosen by GCV for the natural cubic spline fit, which is also shown in Fig. 11.6. The fits are virtually indistinguishable, which is reassuring. Again we point out that this analysis ignores the clear heteroscedasticity in these data. Within the linear mixed model framework, it would be natural to assume a parametric or nonparametric model for σ_ϵ^2 as a function of x.

In Fig. 11.9, we display the contributions $\widehat{b}_l(x - \xi_l)_+^3$ from the $l = 1, \ldots, 20$, truncated cubic segments. The contribution from the fixed effect cubic, $\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3$, is shown as the solid line in each of the plots in this figure. The 1st and 16th–20th cubic segments offer virtually no contribution to the fit, while the contribution of the 4th–14th segments is considerable, which reflects the strong rate of change in the response between ranges of 550 m and 650 m.

11.2.9 Linear Mixed Model Spline Representation: Bayesian Inference

We now discuss a Bayesian mixed model approach. The model is the same as in the last section, with carefully chosen priors. We will not discuss implementation in detail, but lean on the INLA method described in Sect. 3.7.4.

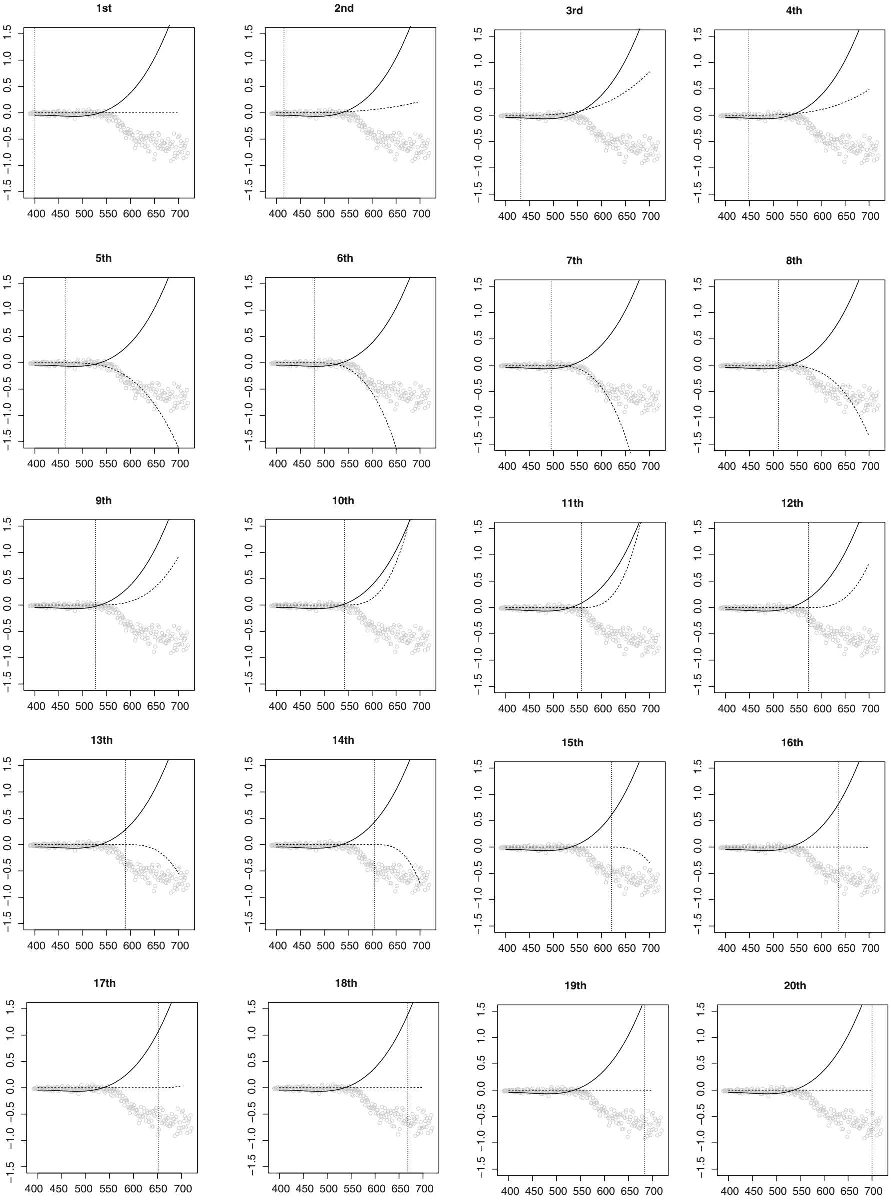

Fig. 11.9 Contributions of the 20 spline bases to the linear mixed model fit to the LIDAR data. The cubic fixed effects fitted line is drawn as the *solid line* on each plot, and the 20 contributions from each of the truncated cubic segments are drawn as *dotted lines* on each plot. The *dotted vertical line* on each plot indicates the knot location associated with the truncated line segment displayed in that plot

Prior distributions on smoothing parameters have the potential to increase the stability of the fit, if the priors are carefully specified. An approach suggested by Fong et al. (2010) is to place a prior on σ_b^2 and examine the induced prior on the effective degrees of freedom, a more easily interpretable quantity. The idea is to experiment with prior choices on σ_b^2 until one settles on a prior on the effective degrees of freedom that one is comfortable with. The effective degrees of freedom is given by (11.25) and can be rewritten as

$$\text{df}(\lambda) = \text{tr}[(\boldsymbol{c}^{\text{T}}\boldsymbol{c} + \lambda \boldsymbol{D})^{-1}\boldsymbol{c}^{\text{T}}\boldsymbol{c}].$$

The total degrees of freedom can be decomposed into the degrees of freedom associated with $\boldsymbol{\beta}$ and $\boldsymbol{b}$. This decomposition can be extended easily to situations in which we have additional random effects beyond those associated with the spline basis. In each of these situations, the degrees of freedom associated with the respective parameter are obtained by summing the appropriate diagonal elements of $(\boldsymbol{c}^{\text{T}}\boldsymbol{c} + \lambda \boldsymbol{D})^{-1}\boldsymbol{c}^{\text{T}}\boldsymbol{c}$. Specifically, for d sets of parameters, let $\boldsymbol{E}_j$ be the $(p+1+L) \times (p+1+L)$ diagonal matrix with ones in the diagonal positions corresponding to set j, $j = 1, \ldots, d$. Then, the degrees of freedom associated with this set are

$$\text{df}_j(\lambda) = \text{tr}[\boldsymbol{E}_j(\boldsymbol{c}^{\text{T}}\boldsymbol{c} + \lambda \boldsymbol{D})^{-1}\boldsymbol{c}^{\text{T}}\boldsymbol{c}].$$

Note that the effective degrees of freedom change as a function of L, as expected. To evaluate λ, σ_ϵ^2 is required; Fong et al. (2010) recommend the substitution of an estimate of σ_ϵ^2. For example, one may use an estimate obtained from the fitting of a spline model in a likelihood implementation. For further discussion of prior choice for σ_b^2 in a spline context, see Crainiceanu et al. (2005). We first illustrate the steps in prior construction in a toy example, before presenting a more complex example.

Example: One-Way ANOVA Model

As a simple non-spline demonstration of the derived effective degrees of freedom, consider the one-way ANOVA model:

$$Y_{ij} = \beta_0 + b_i + \epsilon_{ij},$$

with $b_i \mid \sigma_b^2 \sim_{iid} \text{N}(0, \sigma_b^2)$ and $\epsilon_{ij} \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ for $i = 1, \ldots, m$ groups and $j = 1, \ldots, n$ observations per group. This model may be written as $\boldsymbol{y} = \boldsymbol{c}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\boldsymbol{c}$ is the $nm \times (m+1)$ design matrix

$$\boldsymbol{c} = \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{1}_n & \mathbf{0}_n & \mathbf{1}_n & \cdots & \mathbf{0}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_n & \mathbf{0}_n & \mathbf{0}_n & \cdots & \mathbf{1}_n \end{bmatrix},$$

and $\boldsymbol{\gamma} = [\beta_0, b_1, \ldots, b_m]^{\text{T}}$. The effective degrees of freedom are given by (11.25), with $\lambda = \sigma_\epsilon^2/\sigma_b^2$ and $\boldsymbol{D}$ a diagonal matrix with a single zero followed by m ones.

For illustration, assume $m = 10$ and $\sigma_b^{-2} \sim \text{Ga}(0.5, 0.005)$. Figure 11.10 displays the prior distribution for σ_b, the implied prior distribution on the effective

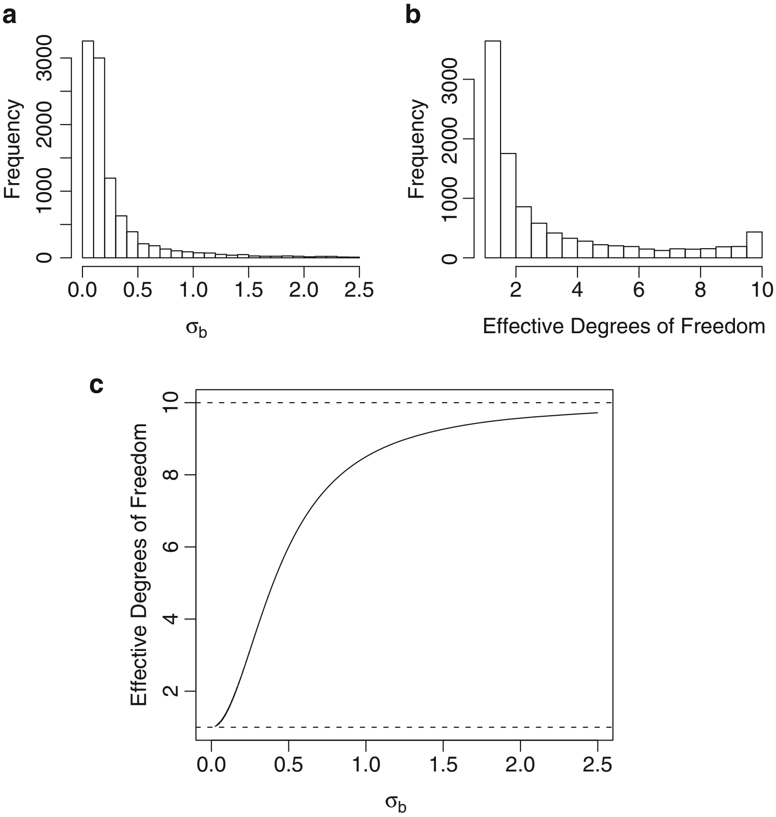

Fig. 11.10 Gamma prior for σ_b^{-2} with parameters 0.5 and 0.005, for the one-way ANOVA example. (**a**) Implied prior for σ_b, (**b**) implied prior for the effective degrees of freedom, and (**c**) effective degrees of freedom versus σ_b

degrees of freedom, and the bivariate plot of these quantities. For clarity, values of σ_b greater than 2.5 (corresponding to 4% of points) are excluded from the plots. In panel (c), we have placed horizontal lines at effective degrees of freedom equal to 1 (complete smoothing) and 10 (no smoothing). We also highlight the strong nonlinearity. From panel (b), we conclude that this prior choice favors quite strong smoothing.

Example: Spinal Bone Marrow Density

We demonstrate the use of the mixed model for nonparametric smoothing using O'Sullivan splines, which, as described in Sect. 11.2.5, are based on a B-spline basis, and using data introduced in Sect. 1.3.6. Recall that these data concern

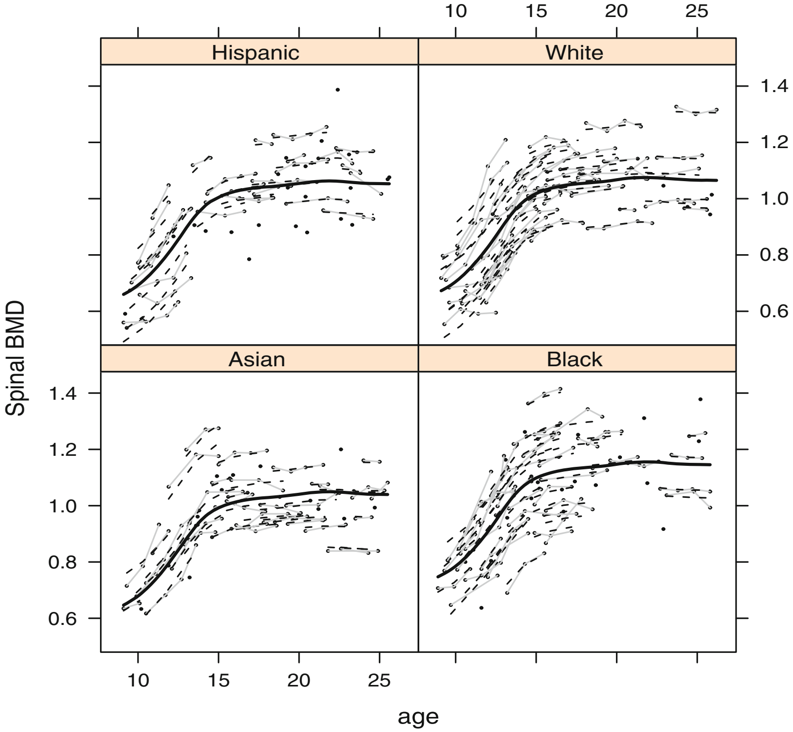

Fig. 11.11 Spinal bone mineral density measurements versus age by ethnicity. Measurements on the same woman are joined with *gray lines*. The *bold solid curve* corresponds to the fitted spline, and the *dashed lines* to the individual fits

longitudinal measurements of spinal bone mineral density (SBMD) on 230 female subjects aged between 8 and 27 years and of one of four ethnic groups: Asian, Black, Hispanic, and White. Let y_{ij} denote the SBMD measure for subject i at occasion j, for $i = 1, \dots, m = 230$ and $j = 1, \dots, n_i$ and with n_i ranging between 1 and 4. Let $N = \sum_{i=1}^{m} n_i$. Figure 11.11 shows these data with joined points indicating measurements on the same woman. For these data, we would like a model in which the response is a smooth function of age and in which between-woman variability in response is acknowledged. We therefore assume the model:

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta}_1 + \text{age}_{ij} \times \beta_2 + \sum_{l=1}^{L} z_{ijl} b_{1l} + b_{2i} + \epsilon_{ij}$$

where $\boldsymbol{x}_{ij}$ is a 1×4 vector containing an indicator for the ethnicity of individual i, with $\boldsymbol{\beta}_1$ the associated 4×1 vector of fixed effects, z_{ijl} is the lth basis associated with age, with associated parameters $b_{1l} \mid \sigma_1^2 \sim \text{N}(0, \sigma_1^2)$ and $b_{2i} \mid \sigma_2^2 \sim \text{N}(0, \sigma_2^2)$ are the woman-specific random effects, and $\epsilon_{ij} \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ represent the residual errors. All random terms are assumed independent. Note that the spline model is assumed common to all ethnic groups and all women, though it would be straightforward to allow, for example, a different spline for each ethnicity. Let $\boldsymbol{\beta} = [\,\boldsymbol{\beta}_1, \beta_2\,]^{\mathrm{T}}$ and $\boldsymbol{x}_i$ be the $n_i \times 5$ fixed effect design matrix with j-th row $[\boldsymbol{x}_{ij}, \text{age}_{ij}]$, $j = 1, \ldots, n_i$ (each row is identical since age_{ij} is the initial age). Also, let $\boldsymbol{z}_{1i}$ be the $n_i \times L$ matrix of age basis functions, $\boldsymbol{b}_1 = [b_1, \ldots, b_L]^{\mathrm{T}}$ be the vector of associated coefficients, $\boldsymbol{z}_{2i}$ represent the $n_i \times 1$ vector of ones, and $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^{\mathrm{T}}$. Then

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_{1i}\boldsymbol{b}_1 + \boldsymbol{z}_{2i}b_i + \boldsymbol{\epsilon}_i$$

and we may write:

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}_1\boldsymbol{b}_1 + \boldsymbol{z}_2\boldsymbol{b}_2 + \boldsymbol{\epsilon}\\ &= \boldsymbol{c}\boldsymbol{\gamma} + \boldsymbol{\epsilon},\end{aligned}$$

where $\boldsymbol{y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m]^{\mathrm{T}}$, $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^{\mathrm{T}}$, $\boldsymbol{z}_1 = [\boldsymbol{z}_{11}, \ldots, \boldsymbol{z}_{1m}]^{\mathrm{T}}$, $\boldsymbol{z}_2 = [\boldsymbol{z}_{21}, \ldots, \boldsymbol{z}_{2m}]^{\mathrm{T}}$, and $\boldsymbol{b}_2 = [b_{21}, \ldots, b_{2m}]^{\mathrm{T}}$.

We examine two approaches to inference, one based on REML (Sect. 8.5.3) and the other Bayesian, using INLA for computation. In each case, to fit the model, we first construct the basis functions and from these, the required design matrices. Running the REML version of the model, we obtain $\widehat{\sigma}_\epsilon = 0.033$, which we use to evaluate the effective degrees of freedom associated with the priors for each of σ_1^2 and σ_2^2. We assume the usual improper prior, $\pi(\sigma_\epsilon^2) \propto 1/\sigma_\epsilon^2$ for σ_ϵ^2. After some experimentation, we settled on the prior $\sigma_1^{-2} \sim \text{Ga}(0.5, 5\times10^{-6})$. For σ_2^2, we desire a 90% interval for b_{2i} of ±0.3 which, with 1 degree of freedom for the marginal distribution, leads to $\sigma_2^{-2} \sim \text{Ga}(0.5, 0.00113)$. See Sect. 8.6.2 for details on the rationale for this approach. Figures 11.12(a) and (d) shows the priors for σ_1 and σ_2, with the priors on the implied effective degrees of freedom displayed in panels (b) and (e). For the spline component, the 90% prior interval on the effective degrees of freedom is $[\,2.4, 10\,]$. Figures 11.12(c) and (f) shows the relationship between the standard deviations and the effective degrees of freedom.

Table 11.1 compares estimates from REML and INLA implementations of the model, and we see close correspondence between the two. Figures 11.12(a) and (d) show the posterior medians for σ_1 and σ_2, which correspond to effective degrees of freedom of 8 and 214 for the spline model and random intercepts, respectively, as displayed on panels (b) and (e). The effective degrees of freedom of 214 associated with the random intercepts show that there is considerable variability between the 230 women here. This is confirmed in Fig. 11.11, where we observe large vertical differences between the profiles. This figure also shows the fitted spline, which appears to mimic the age trend in the data well.

11.3 Kernel Methods

We now turn to another class of smoothers that are based on kernels. Kernel methods are used in both density estimation and nonparametric regression, and it is the latter on which we concentrate (though we touch on the former in Sect. 11.3.2). The basic

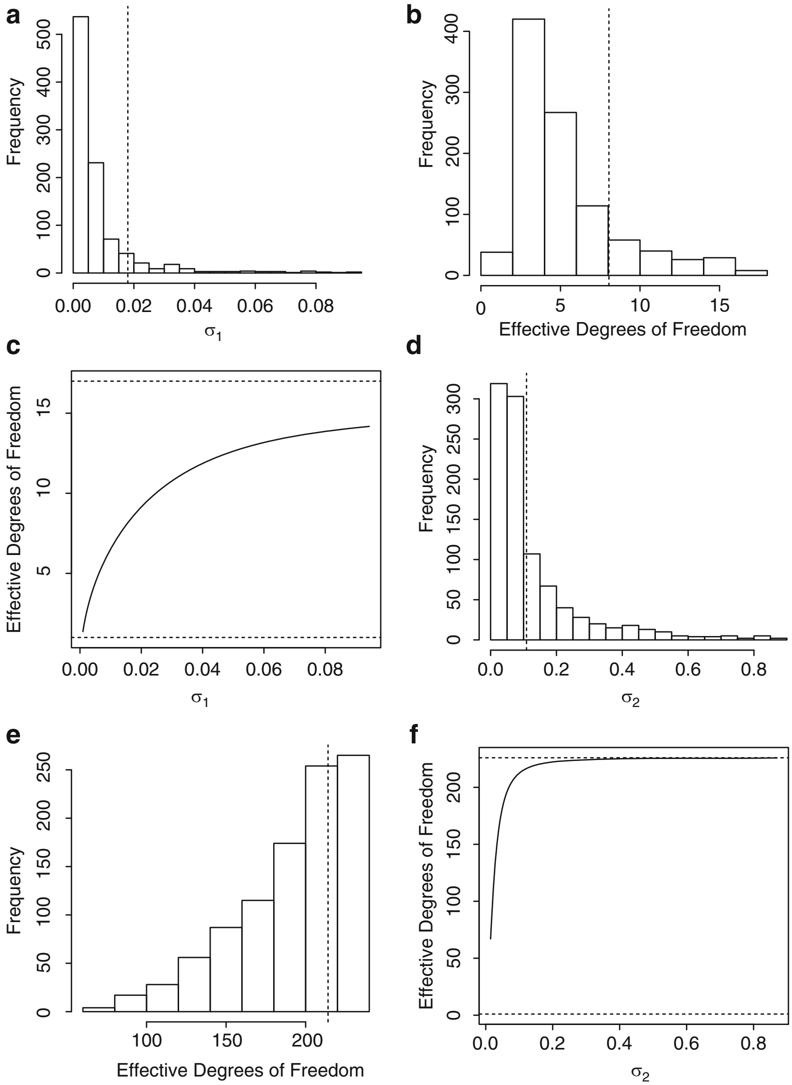

Fig. 11.12 Prior summaries for the spinal bone mineral density data. (**a**) σ_1, the standard deviation of the spline coefficients; (**b**) effective degrees of freedom associated with the prior for the spline coefficients; (**c**) effective degrees of freedom versus σ_1; (**d**) σ_2, the standard deviation of the between-individual random effects; (**e**) effective degrees of freedom associated with the individual random effects; and (**f**) effective degrees of freedom versus σ_2. The *lower and upper dashed horizontal lines* in panels (**c**) and (**f**) are the minimum and maximum attainable degrees of freedom, respectively. The *vertical dashed lines* on panels (**a**), (**b**), (**d**), and (**e**) correspond to the posterior medians

Table 11.1 REML and INLA summaries for the spinal bone data. The intercept corresponds to the Asian group. For the entries marked with a *, standard errors were unavailable

Variable	REML	INLA
Intercept	0.560 ± 0.029	0.563 ± 0.031
Black	0.106 ± 0.021	0.106 ± 0.021
Hispanic	0.013 ± 0.022	0.013 ± 0.022
White	0.026 ± 0.022	0.026 ± 0.022
Age	0.021 ± 0.002	0.021 ± 0.002
σ_1	0.018*	0.024 ± 0.006
σ_2	0.109*	0.109 ± 0.006
σ_ϵ	0.033*	0.033 ± 0.002

idea underlying kernel methods is to estimate the density/regression function *locally* with the kernel function weighting the data in an appropriate fashion. We begin by briefly defining, and giving examples of, kernels.

11.3.1 Kernels

A *kernel* is a smooth function $K(\cdot)$ such that $K(x) \geq 0$, with

$$\int K(u)\,du = 1, \quad \int uK(u)\,du = 0, \quad \sigma_K^2 = \int u^2 K(u)\,du < \infty. \tag{11.26}$$

In practice, a kernel is applied to a standardized variable, and so, in what follows, we do not include a scale parameter since the standardization has removed the dependence on scale.

We describe four common examples of kernel functions. The *Gaussian* kernel is

$$K(x) = (2\pi)^{-1/2} \exp\left(-\frac{x^2}{2}\right)$$

and is nonzero for all x, which makes this kernel relatively computationally expensive to work with since all points must be considered in calculations for a single x. We describe three alternatives but first define

$$I(x) = \begin{cases} 1 \text{ if } |x| \leq 1 \\ 0 \text{ if } |x| > 1. \end{cases}$$

The *Epanechnikov* kernel has the form

$$K(x) = \frac{3}{4}(1 - x^2)I(x), \tag{11.27}$$

while the *tricube* kernel is

$$K(x) = \frac{70}{81}\left(1 - |x|^3\right)^3 I(x). \tag{11.28}$$

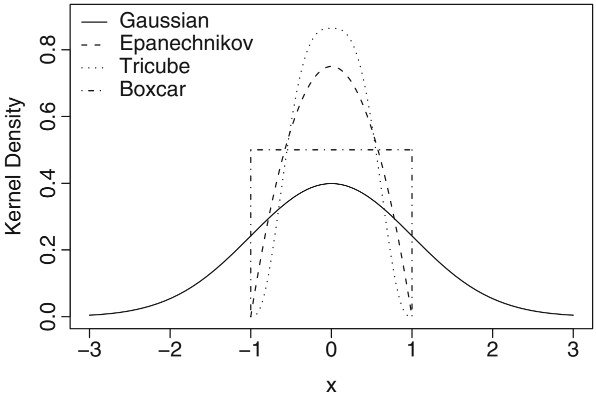

Fig. 11.13 Pictorial representation of four commonly used kernels

Finally, the *boxcar* kernel is

$$K(x) = \frac{1}{2} I(x). \tag{11.29}$$

All four kernels are displayed in Fig. 11.13. We first describe kernel density estimation, which is a simple technique used in a classification context (as described in Sect. 12.8.3).

11.3.2 Kernel Density Estimation

Consider a *random univariate sample* $x_1, \ldots, x_n$ from a density $p(\cdot)$. The kernel density estimate (KDE) of the unknown density, given a smoothing parameter λ, is

$$\widehat{p}^{(\lambda)}(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda} \right), \tag{11.30}$$

so that the estimate of the density at x is potentially built upon contributions from all n observed values, though for the finite range kernels (11.27)–(11.29), the sum will typically be over far fewer points. Choosing $K(\cdot)$ as a probability density function ensures that $\widehat{p}^{(\lambda)}(x)$ is also a density. We write $K_\lambda(u) = \lambda^{-1} K(u/\lambda)$ for a slightly more compact notation.

We now informally state a number of properties of the kernel density estimator. A number of regularity conditions are required, the most important of which is that the second derivative $p''(x)$ is absolutely continuous; Wand and Jones (1995, Chap. 2) contains more details. We also assume the conditions on $K(\cdot)$ given in (11.26).

Since $x_1, \ldots, x_n$ are a random sample from $p(\cdot)$, the expectation of the density estimator can be written as

$$\begin{aligned}
\mathrm{E}\left[\widehat{p}^{(\lambda)}(x)\right] &= \frac{1}{n\lambda}\sum_{i=1}^{n}\mathrm{E}_{X_i}\left[K\left(\frac{x-X_i}{\lambda}\right)\right] \\
&= \mathrm{E}_{T}\left[K_{\lambda}\left(x-T\right)\right] \\
&= \int K_{\lambda}(x-t)p(t)\,dt, \qquad (11.31)
\end{aligned}$$

which is a convolution of the true density with the kernel. Smoothing has, therefore, produced a biased estimator whose mean is a smoothed version of the true density. Clearly, we wish to have $\lambda \to 0$ as $n \to \infty$, so that the kernel concentrates more and more on x with increasing n, ensuring that the bias goes to zero.

We write λ_n to emphasize the dependence on n. It is straightforward to show that, as $n \to \infty$, with $\lambda_n \to 0$ and $n\lambda_n \to \infty$:

$$\mathrm{E}\left[\widehat{p}^{(\lambda_n)}(x)\right] = p(x) + \frac{1}{2}\lambda_n^2 p''(x)\sigma_K^2 + o(\lambda_n^2)$$

so that the estimator is *asymptotically unbiased.*

Proof. With $\widehat{p}^{(\lambda_n)}(x)$ given by (11.30),

$$\begin{aligned}
\mathrm{E}[\widehat{p}^{(\lambda_n)}(x)] &= \int K_{\lambda_n}(x-t)p(t)dt \\
&= \int K(u)p(x-\lambda_n u)du \\
&= \int K(u)\left[p(x) - \lambda_n u p'(x) + \frac{\lambda_n^2 u^2}{2}p''(x) + \ldots\right]du \\
&= p(x) + \frac{\lambda_n^2}{2}p''(x)\sigma_K^2 + o(\lambda_n^2). \qquad \square
\end{aligned}$$

The bias is large whenever the absolute value of the second derivative is large. In peaks, $p''(x) < 0$, and the bias is negative since $\widehat{p}^{(\lambda_n)}(x)$ underestimates $p(x)$, and in troughs, the bias is positive as $\widehat{p}^{(\lambda_n)}(x)$ overestimates $p(x)$.

Via a similar calculation,

$$\mathrm{var}\left[\widehat{p}^{(\lambda_n)}(x)\right] = \frac{1}{n\lambda_n}p(x)K_2 + o\left(\frac{1}{n\lambda_n}\right),$$

where $K_2 = \int K(u)^2\,du$ and $n\lambda_n$ is a "local sample size" (so that larger λ_n gives a larger effective sample size). The variance is also proportional to the height of the density. Overall, as λ_n decreases to zero, the bias diminishes, while the variance increases, with the opposite behavior occurring as λ_n increases. The combined effect is that, in order to obtain an estimator which converges to the true density, we require both λ_n and $1/n\lambda_n$ to decrease as sample size increases.

As discussed in Sect. 10.4, the accuracy of an estimator may be assessed by evaluating the mean squared error (MSE). For $\widehat{p}^{(\lambda_n)}(x)$,

$$\begin{aligned}\text{MSE}\left[\widehat{p}^{(\lambda_n)}(x)\right] &= \text{E}\left[\left(\widehat{p}^{(\lambda_n)}(x) - p(x)\right)^2\right]\\ &= \text{bias}\left[\widehat{p}^{(\lambda_n)}(x)\right]^2 + \text{var}\left[\widehat{p}^{(\lambda_n)}(x)\right]\\ &\approx \frac{\lambda_n^4}{4}p''(x)^2\sigma_K^4 + \frac{1}{n\lambda_n}p(x)K_2, \end{aligned} \tag{11.32}$$

where the expectation in (11.32) is over the uncertainty in $\widehat{p}^{(\lambda_n)}(x)$, that is, over the sampling distribution of $X_1, \dots, X_n$.

Averaging the MSE over x gives the *integrated mean squared error*

$$\begin{aligned}\text{IMSE}\left[\widehat{p}^{(\lambda_n)}(x)\right] &= \int \text{MSE}\left[\widehat{p}^{(\lambda_n)}(x)\right]\,dx\\ &\approx \frac{1}{4}\lambda_n^4\sigma_K^4\int p''(x)^2dx + \frac{1}{n\lambda_n}K_2. \end{aligned} \tag{11.33}$$

If we differentiate (11.33) with respect to λ_n and set equal to zero, we obtain an asymptotic optimal bandwidth of

$$\lambda_n^\star = \left(\frac{K_2}{n\sigma_K^4\int p''(x)^2dx}\right)^{1/5}. \tag{11.34}$$

This formula is useful since it informs us that the optimal bandwidth decreases at rate $n^{-1/5}$. Then, substitution in (11.33) shows that the IMSE is of $O(n^{-4/5})$. It can be shown that there does not exist any estimator that converges faster than this rate, assuming only the existence of second derivatives, p''; for more details, see Chap. 24 of van der Vaart (1998).[6]

We turn now to a discussion of estimation of the amount of smoothing to carry out, that is, how to estimate the optimal λ_n. So-called "plug-in" estimators substitute estimates for unknown quantities (here the integrated squared second derivative in the denominator) in order to evaluate $\lambda_n^\star$. If we assume that $p(\cdot)$ is normal in (11.34), we obtain

$$\lambda_n^\star = (4/3)^{1/5} \times \sigma n^{-1/5}, \tag{11.35}$$

where σ is the standard deviation of the normal.

[6] The histogram estimator converges at rate $O(n^{-2/3})$; see, for example, Wand and Jones (1995, Sect. 2.5).

Leave-one-out cross-validation may be used to choose λ_n in order to minimize a measure of estimation accuracy. One convenient quantity that may be minimized is the *integrated squared error* (ISE), defined as

$$\begin{aligned} \text{ISE}\left[\,\widehat{p}^{(\lambda_n)}(x)\right] &= \int \left[\widehat{p}^{(\lambda_n)}(x) - p(x)\right]^2 \, dx \\ &= \int \widehat{p}^{(\lambda_n)}(x)^2 \, dx - 2\int p(x)\widehat{p}^{(\lambda_n)}(x) \, dx + \int p(x)^2 \, dx. \end{aligned}$$

The last term does not involve λ_n, and the other terms can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\int \widehat{p}_{-i}^{(\lambda_n)}(x)^2 \, dx\right) - \frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{-i}^{(\lambda_n)}(x_i),$$

where $\widehat{p}_{-i}^{(\lambda_n)}(x)$ is the estimator constructed from the data without observation x_i. The use of normal kernels gives a very convenient form for estimation, as described by Bowman and Azzalini (1997, p. 37).

11.3.3 The Nadaraya–Watson Kernel Estimator

We now turn to nonparametric regression and estimation of

$$\begin{aligned} f(x) &= \mathrm{E}[Y \mid x] \\ &= \int yp(y \mid x) \, dy \\ &= \frac{1}{p(x)}\int yp(x,y) \, dy. \end{aligned} \tag{11.36}$$

Suppose we estimate $p(x,y)$ by the product kernel

$$\widehat{p}^{(\lambda_x,\lambda_y)}(x,y) = \frac{1}{n\lambda_x\lambda_y}\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{\lambda_x}\right) K_y\left(\frac{y-y_i}{\lambda_y}\right),$$

and $p(x)$ by

$$\widehat{p}^{(\lambda_x)}(x) = \frac{1}{n\lambda_x}\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{\lambda_x}\right).$$

Substitution of these estimates in (11.36) gives the *Nadaraya–Watson* kernel regression estimator (Nadaraya 1964; Watson 1964):

$$\begin{aligned}
\widehat{f}(x) &= \frac{\frac{1}{n\lambda_x\lambda_y}\sum_{i=1}^{n}\int yK_x\left(\frac{x-x_i}{\lambda_x}\right)K_y\left(\frac{y-y_i}{\lambda_y}\right)\,dy}{\frac{1}{n\lambda_x}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{\lambda_x}\right)} \\
&= \frac{\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{\lambda_x}\right)\int(y_i+u\lambda_y)K_y(u)\,du}{\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{\lambda_x}\right)} \\
&= \frac{\sum_{i=1}^{n}K\left(\frac{x-x_i}{\lambda}\right)y_i}{\sum_{i=1}^{n}K\left(\frac{x-x_i}{\lambda}\right)} \qquad (11.37)
\end{aligned}$$

where we have used $\int K_y(u)\,du = 1$ and $\int uK_y(u)\,du = 0$. We also write $\lambda = \lambda_x$ and $K_x = K$ in the final line. This estimator may be written as the linear smoother:

$$\widehat{f}^{(\lambda)}(x) = \sum_{i=1}^{n} S_i^{(\lambda)}(x)Y_i,$$

where the weights $S_i^{(\lambda)}(x)$ are defined as

$$S_i^{(\lambda)}(x) = \frac{K\left(\frac{x-x_i}{\lambda}\right)}{\sum_{i=1}^{n}K\left(\frac{x-x_i}{\lambda}\right)}.$$

As a special case, a rectangular window (i.e., the boxcar kernel) produces a smoother that is a simple moving average. As with spline models, the choice of the smoothing parameter λ is crucial for reasonable behavior of the estimator.

We now examine the asymptotic IMSE which, as usual, can be decomposed into contributions due to squared bias and variance. An advantage of local polynomial regression estimators is that the form of the bias and variance is relatively simple, thus enabling analytic study. For the subsequent calculations, and those that appear later in this chapter, we state results without regularity conditions. See Fan (1992, 1993) for a more rigorous treatment.

As $\lambda_n \rightarrow 0$ and $n\lambda_n \rightarrow \infty$, the bias of the Nadaraya–Watson estimator at the point x is

$$\text{bias}\left[\widehat{f}^{(\lambda_n)}(x)\right] \approx \frac{\lambda_n^2\sigma_K^2}{2}\left(f''(x) + 2f'(x)\frac{p'(x)}{p(x)}\right), \qquad (11.38)$$

where $p(x)$ is the true but unknown density of x. The bias increases with increasing λ_n as we would expect. The bias also increases at points at which $f(\cdot)$ increases in "wiggliness" (i.e., large $f''(x)$) and where the derivative of the "design density," $p'(x)$, is large. The so-called *design bias* is defined as $2f'(x)p'(x)/p(x)$ and, as we will see in Sect. 11.3.4, may be removed if locally *linear* polynomial models are used.

The variance at the point x is

$$\text{var}\left[\widehat{f}^{(\lambda_n)}(x)\right] \approx \frac{K_2\sigma^2}{n\lambda_n}\frac{1}{p(x)}, \tag{11.39}$$

where we have assumed, for simplicity, that the variance $\sigma^2 = \text{var}(Y \mid x)$ is constant. The variance of the estimator decreases with decreasing measurement error, increasing density of x values, and increasing local sample size $n\lambda_n$. Consequently, we see the "usual" trade-off with small λ reducing the bias but increasing the variance. Combining the squared bias and variance and integrating over x gives the IMSE:

$$\text{IMSE}\left(\widehat{f}^{(\lambda_n)}\right) \approx \frac{\lambda_n^4\sigma_K^4}{4}\int\left(f''(x)+2f'(x)\frac{p'(x)}{p(x)}\right)^2 dx + \frac{K_2\sigma^2}{n\lambda_n}\int\frac{1}{p(x)}\,dx. \tag{11.40}$$

If we differentiate this expression and set equal to zero, we obtain the optimal bandwidth as

$$\lambda_n^\star = \left(\frac{1}{n}\right)^{1/5}\left(\frac{\sigma^2K_2\int p(x)^{-1}\,dx}{\sigma_K^4\int(f''(x)+2f'(x)p'(x)/p(x))^2\,dx}\right)^{1/5} \tag{11.41}$$

so that $\lambda^\star = O(n^{-1/5})$. Plugging this expression into (11.40) shows that the IMSE is $O(n^{-4/5})$, which holds for many nonparametric estimators and is in contrast to most parametric estimators whose variance is $O(n^{-1})$. The loss in efficiency is the cost of the flexibility offered by nonparametric methods. Expression (11.41) depends on many unknown quantities, and while there are "plug-in" methods for estimating these terms, a popular approach is cross-validation.

11.3.4 Local Polynomial Regression

We now describe a generalization of the Nadaraya–Watson kernel estimator, *local polynomial regression*, with improved theoretical properties. Let $w_i(x) = K\left[(x_i - x)/\lambda\right]$ be a weight function and choose $\beta_{0x} = f(x)$ to minimize the weighted sum of squares

$$\sum_{i=1}^n w_i(x)\left(Y_i - \beta_{0x}\right)^2$$

with solution

$$\widehat{f}(x) = \widehat{\beta}_{0x} = \frac{\sum_{i=1}^n w_i(x)Y_i}{\sum_{i=1}^n w_i(x)},$$

showing that the Nadaraya–Watson kernel regression estimator (11.37) corresponds to a locally constant model, estimated using weighted least squares. For notational simplicity, we have not acknowledged that the weight $w_i(x)$ depends on the smoothing parameter λ. We emphasize that we carry out a separate weighted least squares fit for each prediction that we wish to obtain.

This formulation suggests an extension in which a local polynomial replaces the locally constant model of the Nadaraya–Watson kernel estimator. For values of u in a neighborhood of a fixed x, define the polynomial:

$$P_x(u;\boldsymbol{\beta}_x) = \beta_{0x} + \beta_{1x}(u-x) + \frac{\beta_{2x}}{2!}(u-x)^2 + \ldots + \frac{\beta_{px}}{p!}(u-x)^p,$$

with $\boldsymbol{\beta}_x = [\,\beta_{0x}, \ldots, \beta_{px}\,]$. The idea is to approximate f in a neighborhood of x by the polynomial $P_x(u;\boldsymbol{\beta}_x)$.[7] The parameter $\widehat{\boldsymbol{\beta}}_x$ is chosen to minimize the locally weighted sum of squares:

$$\sum_{i=1}^{n} w_i(x)\left[Y_i - P_x(x_i;\boldsymbol{\beta}_x)\right]^2. \tag{11.42}$$

The ensuing local estimate of f at u is

$$\widehat{f}(u) = P_x(u;\widehat{\boldsymbol{\beta}}_x).$$

We could use this estimate in a local neighborhood of x, but instead, we fit a new local polynomial for *every* target x value. At a target value $u = x$,

$$\widehat{f}(x) = P_x(x;\widehat{\boldsymbol{\beta}}_x) = \widehat{\beta}_{0x}.$$

The weight function is $w(x_i) = K[(x_i - x)/\lambda]$, so that the level of smoothing is controlled by the smoothing parameter λ, with $\lambda = 0$ resulting in $\widehat{f}(x_i) = y_i$ and $\lambda = \infty$ being equivalent to the fitting of a linear model. It is important to emphasize that $\widehat{f}(x)$ only depends on the intercept $\widehat{\beta}_{0x}$ of a local polynomial model, but should not be confused with the fitting of a locally constant model.

For estimating the function f at the point x, local regression is equivalent to applying weighted least squares to the model:

$$\boldsymbol{Y} = \boldsymbol{x}_x\boldsymbol{\beta}_x + \boldsymbol{\epsilon}_x, \tag{11.43}$$

with $\mathrm{E}[\boldsymbol{\epsilon}_x] = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\epsilon}_x) = \sigma^2\boldsymbol{W}_x^{-1}$,

$$\boldsymbol{x}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & \frac{(x_1-x)^p}{p!} \\ 1 & x_2 - x & \cdots & \frac{(x_2-x)^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & \frac{(x_n-x)^p}{p!} \end{bmatrix}$$

representing the $n \times (p+1)$ design matrix and $\boldsymbol{W}_x$ the $n \times n$ diagonal matrix with elements $w_i(x)$, $i = 1, \ldots, n$. Large values of w_i correspond to $x - x_i$ being small,

[7] This approximation may be formally motivated via a Taylor series approximation argument.

so that data points x_i close to x are most influential. With the finite range kernels described in Sect. 11.3.1, some of the $w_i(x)$ elements will be zero, in which case we would only consider the data with nonzero elements within (11.43). Note that $\boldsymbol{W}_x$ depends on the kernel function, $K(\cdot)$, and therefore upon the bandwidth, λ. Minimization of

$$(\boldsymbol{Y} - \boldsymbol{x}_x\boldsymbol{\beta}_x)^{\mathrm{T}}\boldsymbol{W}_x(\boldsymbol{Y} - \boldsymbol{x}_x\boldsymbol{\beta}_x)$$

gives

$$\widehat{\boldsymbol{\beta}}_x = (\boldsymbol{x}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{x}_x)^{-1}\boldsymbol{x}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{Y}. \tag{11.44}$$

Taking the inner product of the first row of $(\boldsymbol{x}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{x}_x)^{-1}\boldsymbol{x}_x^{\mathrm{T}}\boldsymbol{W}_x$ with $\boldsymbol{Y}$ gives $\widehat{f}(x) = \widehat{\beta}_{0x}$.

From (11.44), it is clear that this estimator is linear in the data:

$$\widehat{f}(x) = \sum_{i=1}^{n} S_i^{(\lambda)}(x)Y_i.$$

This estimator has mean

$$\mathrm{E}[\widehat{f}(x)] = \sum_{i=1}^{n} S_i^{(\lambda)}(x)f(x_i)$$

and variance

$$\mathrm{var}\left[\widehat{f}(x)\right] = \sigma^2\sum_{i=1}^{n} S_i^{(\lambda)}(x)^2 = \sigma^2||\boldsymbol{S}^{(\lambda)}(x)||^2,$$

where we have again assumed the error variance is constant and that the observations are uncorrelated. The effective degrees of freedom can be defined as $p^{(\lambda)} = \mathrm{tr}(\boldsymbol{S}^{(\lambda)})$ where $\boldsymbol{S}^{(\lambda)}$ is the "hat" matrix determined from $\widehat{\boldsymbol{Y}} = \boldsymbol{S}^{(\lambda)}\boldsymbol{Y}$.

Asymptotic analysis suggests that local polynomials of odd degree dominate those of even degree (Fan and Gijbels 1996), though Wand and Jones (1995) emphasize that the practical implications of this result should not be overinterpreted. Often $p = 1$ will be sufficient for estimating $f(\cdot)$. It can also be shown (Exercise 11.6) that with a linear local polynomial, we obtain

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} w_i(x)Y_i}{\sum_{i=1}^{n} w_i(x)} + (x - \overline{x}_w)\frac{\sum_{i=1}^{n} w_i(x)(x_i - \overline{x}_w)Y_i}{\sum_{i=1}^{n} w_i(x)(x_i - \overline{x}_w)^2},$$

where $\overline{x}_w = \sum_{i=1}^{n} w_i(x)x_i / \sum_{i=1}^{n} w_i(x)$ and $w_i(x) = K((x - x_i)/\lambda)$. Therefore, the estimator is the locally constant (Nadaraya–Watson) estimator plus a term that corrects for the local slope and skewness of the x_i.

For the linear local polynomial model, we have

$$\mathrm{E}\left[\widehat{f}^{(\lambda_n)}(x)\right] \approx f(x) + \frac{1}{2}\lambda_n^2 f''(x)\sigma_K^2$$

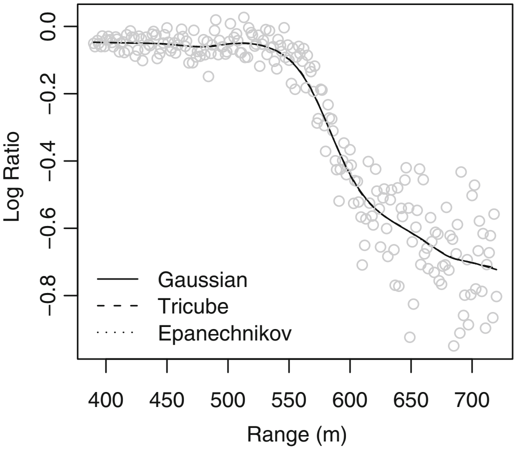

Fig. 11.14 Local linear polynomial fits to the LIDAR data, with three different kernels. The fits are indistinguishable

and

$$\text{var}\left[\widehat{f}^{(\lambda_n)}(x)\right] \approx \frac{1}{n\lambda_n} K_2 \sigma^2 \frac{1}{p(x)}.$$

Proofs of these expressions may be found in Wand and Jones (1995, Sect. 5.3). Notice that the bias is dominated by the second derivative, which is reflecting the error in the linear approximation. If f is linear in x, then $\widehat{f}$ is exactly unbiased.

For the local linear polynomial estimator,

$$\begin{aligned}\text{IMSE}\left(\widehat{f}^{(\lambda_n)}\right) &= \text{bias}\left[\widehat{f}^{(\lambda_n)}\right]^2 + \text{var}\left[\widehat{f}^{(\lambda_n)}\right] \\ &\approx \frac{\lambda_n^4 \sigma_K^4}{4}\left[\int f''(x)^2\, dx\right] + \frac{K_2 \sigma^2}{n\lambda_n}\int \frac{1}{p(x)}\, dx.\end{aligned}$$

In comparison with (11.40), the design bias is zero, showing a clear advantage of the linear polynomial over the Nadaraya–Watson estimator. The optimal λ is therefore

$$\lambda_n^\star = \left(\frac{1}{n}\right)^{1/5} \left(\frac{\sigma^2 K_2 \int p(x)^{-1} dx}{\sigma_K^4 \int f''(x)^2\, dx}\right)^{1/5}. \qquad (11.45)$$

Each of the terms in expression (11.45) can be estimated to give a “plug-in” estimator of λ_n, or cross-validation may be used. Since the local polynomial regression estimator is a linear smoother, inference for this model follows as in Sect. 11.2.7.

Example: Light Detection and Ranging

Figure 11.14 shows scatterplot smoothing of the LIDAR data using local linear polynomials and Gaussian, tricube and Epanechnikov kernels. In each case the smoothing parameter is chosen via generalized cross-validation, as described in Sect. 10.6.3. The choice of kernel is clearly unimportant in this example.

11.4 Variance Estimation

Accurate inference, for example, confidence intervals for $f(x)$ at a particular x, depends on accurate estimation of the error variance, which may be nonconstant.

We begin by assuming that the model is

$$y_i = \mathrm{E}[Y_i \mid x_i] + \epsilon_i = f(x_i) + \epsilon_i,$$

with $\mathrm{var}(\epsilon_i \mid x_i) = \sigma^2$ and $\mathrm{cov}(\epsilon_i, \epsilon_j \mid x_i, x_j) = 0$. We have made the crucial, and strong, assumption that the errors have constant variance (i.e., are *homoscedastic*) and are uncorrelated. We assume a linear smoother so that $\widehat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{Y}$ with $p = \mathrm{tr}(\boldsymbol{S})$ the effective degrees of freedom and suppressing the dependence on the smoothing parameter.

The expectation of the residual sum of squares is

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{Y} - \widehat{\boldsymbol{f}})^{\mathrm{T}}(\boldsymbol{Y} - \widehat{\boldsymbol{f}})] &= \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{S}\boldsymbol{Y})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{S}\boldsymbol{Y})] \\
&= \mathrm{E}[\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y}] \\
&= \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f} + \mathrm{tr}\left[(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{I}\sigma^2\right] \\
&\quad \text{using identity (B.4) from Appendix B} \\
&= \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f} + \sigma^2\mathrm{tr}(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}} - \boldsymbol{S} + \boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}) \\
&= \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f} + \sigma^2(n - 2p + \widetilde{p})
\end{aligned}$$

where

$$\widetilde{p} = \mathrm{tr}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}).$$

The bias is

$$\boldsymbol{f} - \mathrm{E}[\widehat{\boldsymbol{f}}] = \boldsymbol{f} - \boldsymbol{S}\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{f} - \boldsymbol{S}\boldsymbol{f} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f}.$$

Therefore,

$$\mathrm{E}\left[\frac{\mathrm{RSS}}{n - 2p + \widetilde{p}}\right] = \sigma^2 + \frac{\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f}}{n - 2p + \widetilde{p}}$$

with the second term being the sum of squared bias terms divided by a particular form of degrees of freedom. If the second term is small, it may be ignored to give the estimator:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n \left(Y_i - \widehat{f}(x_i)\right)^2}{n - 2p + \widetilde{p}}. \tag{11.46}$$

Notice that for idempotent $\boldsymbol{S}$, we have $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S} = \boldsymbol{S}$, $p = \widetilde{p}$, and (11.46) results in an estimator with a more familiar form, that is, with denominator $n - p$ with p the effective degrees of freedom.

We now derive an alternative *local differencing* (method of moments) estimator (Rice 1984). We begin by considering the expected differences:

$$\begin{aligned} \mathrm{E}\left[(Y_{i+1} - Y_i)^2\right] &= \mathrm{E}\left[(f_{i+1} + \epsilon_{i+1} - f_i - \epsilon_i)^2\right] \\ &= (f_{i+1} - f_i)^2 + \mathrm{E}\left[(\epsilon_{i+1} - \epsilon_i)^2\right] \qquad (11.47) \\ &= (f_{i+1} - f_i)^2 + 2\sigma^2 \qquad (11.48) \end{aligned}$$

for $i = 1, \ldots, n-1$. If $f_{i+1} \approx f_i$, then $\mathrm{E}[(Y_{i+1} - Y_i)^2] \approx 2\sigma^2$, leading to

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2. \tag{11.49}$$

This estimator will be inflated, as is clear from (11.48). An improved method of moments estimator, proposed by Gasser et al. (1986), is based on weighted second differences of the data. Specifically, first consider the line joining the points $[x_{i-1}, y_{i-1}]$ and $[x_{i+1}, y_{i+1}]$. This line is obtained by solving

$$\begin{aligned} y_{i+1} &= \alpha_i + \beta_i x_{i+1} \\ y_{i-1} &= \alpha_i + \beta_i x_{i-1}, \end{aligned}$$

to give

$$\begin{aligned} \widehat{\alpha}_i &= \frac{y_{i-1}x_{i+1} - y_{i+1}x_{i-1}}{x_{i+1} - x_{i-1}} \\ \widehat{\beta}_i &= \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}. \end{aligned}$$

Define a pseudo-residual as

$$\begin{aligned} \widetilde{\epsilon}_i &= \widehat{\alpha}_i + \widehat{\beta}_i x_i - y_i \\ &= a_i y_{i-1} + b_i y_{i+1} - y_i, \end{aligned}$$

where

$$a_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}$$

$$b_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}.$$

Gasser et al. (1986) show that var$(\widetilde{\epsilon}_i) = [a_i^2 + b_i^2 + 1]\sigma^2 + O(n^{-2})$ (the final term here is required because the pseudo-residuals do not have mean zero). We are therefore led to the estimator:

$$\widetilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} c_i^2 \, \widetilde{\epsilon}_i^2 \tag{11.50}$$

where $c_i^2 = (a_i^2 + b_i^2 + 1)^{-1}$, for $i = 2, \ldots, n$. Note that the variance estimators (11.49) and (11.50) depend only on (y_i, x_i), $i = 1, \ldots, n$ and not on the model that is fitted.

Now suppose we believe the data exhibit nonconstant variance (*heteroscedasticity*). If the variance depends on $f(x)$ via some known form, for example, $\sigma^2(x) = \sigma^2 f(x)$, then quasi-likelihood (Sect. 2.5) may be used. Otherwise, consider the model:

$$Y_i = f(x_i) + \sigma(x_i)\epsilon_i,$$

with E$[\epsilon_i] = 0$ and var$(\epsilon_i) = 1$. Since the variance must be positive, a natural model to consider is

$$\begin{aligned} Z_i &= \log \left[(Y_i - f(x_i))^2 \right] = \log \left[\sigma^2(x_i) \right] + \log(\epsilon_i^2) \\ &= g(x_i) + \delta_i, \end{aligned} \tag{11.51}$$

where $\delta_i = \log(\epsilon_i^2)$. A simple approach to implementation is to first estimate $f(\cdot)$ under the assumption of constant variance, obtain fitted values, and then form residuals. One may then estimate $g(\cdot)$ using a nonparametric estimator to produce $\widehat{\sigma}(x)^2 = \exp\left[\,\widehat{g}(x)\,\right]$, for $i = 1, \ldots, n$. Subsequently, confidence intervals may be constructed based on $\widehat{\sigma}(x)$. For further details, see Yu and Jones (2004). A more statistically rigorous approach would simultaneously estimate $f(\cdot)$ and $g(\cdot)$.

Example: Light Detection and Ranging

Using the natural cubic spline fit, the variance estimate based on the residual sum of squares (11.46) is 0.080^2. The estimates based on the first differences (11.49) and second differences (11.50) are 0.082^2 and 0.083^2, respectively. In this example, therefore, the estimates are very similar though of course for these data, the variance of the error terms is clearly nonconstant. In Fig. 11.15(a), we plot the residuals (from a natural cubic spline fit) versus the range. To address the nonconstant error variance, we assume a model of the form (11.51). Figure 11.15(b) plots the log squared residuals z_i, as defined in (11.51), versus the range. Experimentation with smoothing models for z_i indicates that a simple linear model

$$\text{E}[Z_i \mid x_i] = \alpha_0 + \alpha_1 x_i$$

is adequate, and this is added to the plot. Figure 11.15(c) plots the estimated standard deviation, $\widehat{\sigma}(x) = \sqrt{\exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 x)}$, versus x, and Fig. 11.15(d) shows the standardized residuals:

$$\frac{y_i - \widehat{f}(x_i)}{\widehat{\sigma}(x_i)}$$

versus x_i. We see that the spread is constant across the range of x_i, suggesting that the error variance model is adequate.

11.5 Spline and Kernel Methods for Generalized Linear Models

So far we have considered models of the form, $Y = f(x) + \epsilon$, with independent and uncorrelated constant variance errors ϵ. We outline the extension to the situation in which generalized linear models (GLMs, Sect. 6.3) are appropriate in a parametric framework. To carry out flexible modeling, penalty terms or weighting may be applied to the log-likelihood and smoothing models (e.g., based on splines or kernels) may be used on the linear predictor scale.

Recall that, for a GLM, $\text{E}[Y_i \mid \theta_i, \alpha] = b'(\theta_i) = \mu_i$, with a link function $g(\mu_i)$ and a variance function $\text{var}(Y_i \mid \theta_i, \alpha) = \alpha b''(\theta_i) = \alpha V_i$. In a smoothing context, we may relax the linearity assumption and connect the mean to the smoother via $g(\mu_i) = f(x_i)$. The log-likelihood for a GLM is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha). \tag{11.52}$$

11.5.1 Generalized Linear Models with Penalized Regression Splines

Let $l(\boldsymbol{f})$ denote the log-likelihood corresponding to the smoother $f(x_i)$, $i = 1, \dots, n$. Maximizing over all smooth functions $f(\cdot)$ is not useful since there are an infinite number of ways to interpolate the data. Consider a regression spline model on the scale of the canonical link:

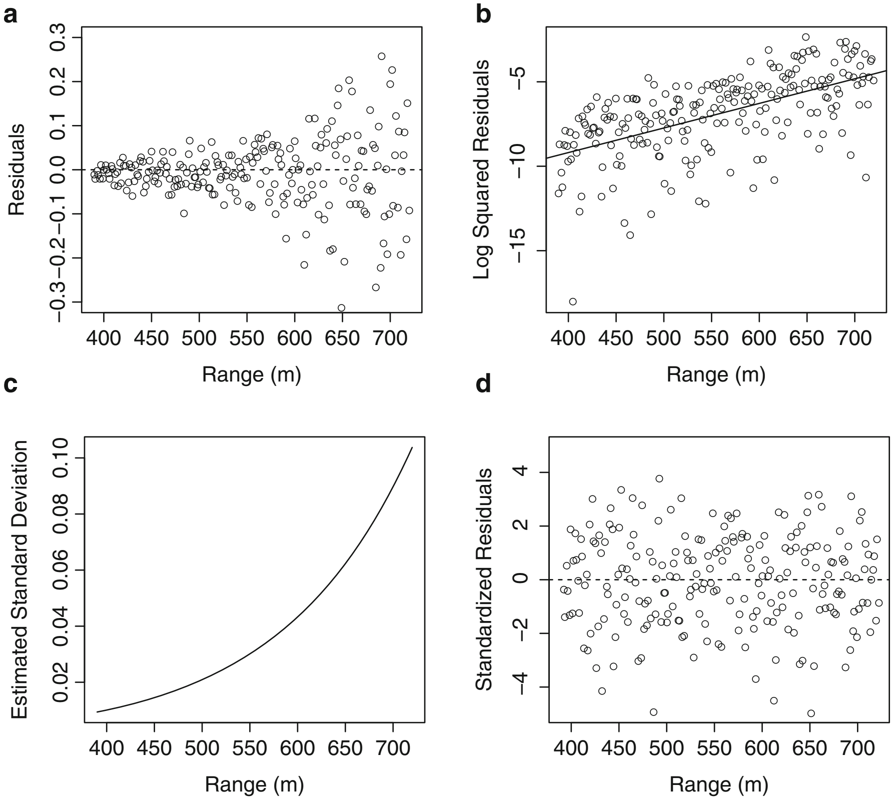

Fig. 11.15 Examination of heteroscedasticity for the LIDAR data. In all plots, the range is plotted on the x-axis, and on the y-axis we have (**a**) residuals from a natural cubic spline fit to the response data; (**b**) log squared residuals, with a linear fit; (**c**) the estimated standard deviation $\widehat{\sigma}(x)$; and (**d**) standardized residuals

$$\begin{aligned}\theta_i = f(x_i) &= \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \sum_{l=1}^{L} b_l (x_i - \xi_l)_+^p \\ &= \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b},\end{aligned}$$

with penalty term $\lambda \boldsymbol{b}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{b}$, where $\boldsymbol{D}$ denotes a known matrix that determines the nature of the penalization, as in Sect. 11.2.5. For example, an obvious form is $\lambda \int f''(t)^2 \, dt$. As in Sect. 11.2.8, we may write $f(x_i) = \boldsymbol{c}\boldsymbol{\gamma}$ with $\boldsymbol{c} = [\boldsymbol{x}, \boldsymbol{z}]$ and $\boldsymbol{\gamma} = [\boldsymbol{\beta}, \boldsymbol{b}]^{\mathrm{T}}$, and $\boldsymbol{D} = \mathrm{diag}(\mathbf{0}_{p+1}, \mathbf{1}_L)$ to give penalty $\lambda \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\gamma}$. To extend the penalized sum of squares given by (11.21), consider the penalized log-likelihood which adds a penalty to (11.52) to give

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \lambda \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\gamma}, \tag{11.53}$$

where

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha), \tag{11.54}$$

and $\theta_i = \boldsymbol{c}\boldsymbol{\gamma}$.

For known λ, the parameters $\boldsymbol{\gamma}$ can be estimated as the solution to

$$\frac{\partial l_p}{\partial \gamma_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \gamma_j} \frac{y_i - \mu_i}{\alpha V_i} - 2\lambda \boldsymbol{D} \gamma_j = 0.$$

To find a solution, a hybrid of IRLS (as described in Sect. 6.5.2) termed the penalized IRLS (P-IRLS) algorithm can be used. At the tth iteration, we minimize a penalized version of (6.16):

$$(\boldsymbol{z}^{(t)} - \boldsymbol{x}\boldsymbol{\gamma})^{\mathrm{T}} \boldsymbol{W}^{(t)} (\boldsymbol{z}^{(t)} - \boldsymbol{x}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\gamma}, \tag{11.55}$$

where, as in the original algorithm, $\boldsymbol{z}^{(t)}$ is the vector of pseudo-data with

$$z_i^{(t)} = \boldsymbol{x}_i \boldsymbol{\gamma}^{(t)} + (Y_i - \mu_i^{(t)}) \left. \frac{d\eta_i}{d\mu_i} \right|_{\boldsymbol{\gamma}^{(t)}}$$

and $\boldsymbol{W}^{(t)}$ is a diagonal matrix with elements:

$$w_i = \frac{\left(d\mu_i / d\eta_i|_{\boldsymbol{\gamma}^{(t)}} \right)^2}{\alpha V_i}.$$

The iterative strategy therefore solves (11.55) using the current versions of $\boldsymbol{z}$ and $\boldsymbol{W}$.

We define an influence matrix for the working penalized least squares problem at the final step of the algorithm as $\boldsymbol{S}^{(\lambda)} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{x} + \lambda \boldsymbol{D})^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}$. The effective degrees of freedom is then defined as $p^{(\lambda)} = \mathrm{tr}\left(\boldsymbol{S}^{(\lambda)}\right)$.

So far as inference is concerned, $\widehat{\boldsymbol{\gamma}}$ is asymptotically normal with mean $\mathrm{E}\left[\widehat{\boldsymbol{\gamma}}\right]$ and variance–covariance matrix:

$$\alpha(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{x} + \lambda \boldsymbol{D})^{-1} \boldsymbol{x} \boldsymbol{W} \boldsymbol{x} (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{x} + \lambda \boldsymbol{D})^{-1}.$$

For more details, see Wood (2006, Sect. 4.8).

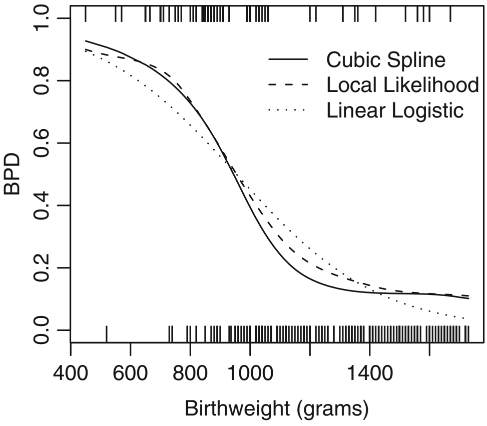

Fig. 11.16 Penalized cubic spline and local likelihood fits to the BPD/birthweight data, with linear logistic fit for comparison

Example: Bronchopulmonary Dysplasia

We illustrate GLM smoothing using the data introduced in Sect. 7.2.3, which consist of binary responses (BPD) Y_i along with birthweights x_i. We consider a logistic regression model:

$$Y_i \mid p(x_i) \sim_{ind} \text{Binomial}\,[\,n_i, p(x_i)\,]\,, \tag{11.56}$$

with

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = f(x_i). \tag{11.57}$$

The log-likelihood is

$$l\,(\boldsymbol{f}) = y_i f(x_i) - n_i \log\{\,1 + \exp\,[f(x_i)]\,\}\,.$$

A penalized spline model assumes

$$\begin{aligned} f(x_i) &= \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \sum_{l=1}^{L} b_l (x_i - \xi_l)_+^p \\ &= \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}. \end{aligned}$$

The predicted probabilities are therefore

$$p(x) = \frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b})}{1 + \exp(\boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b})}.$$

Figure 11.16 displays the data along with three fitted curves. The linear logistic model is symmetric in the tails, which appears overly restrictive for these data. We fit a penalized cubic spline model (11.53) with $L = 10$ knots using P-IRLS and pick the smoothing parameter using AIC. For this model,

$$\mathrm{AIC}^{(\lambda)} = -2l\left(\boldsymbol{f}^{(\lambda)}\right) + 2p^{(\lambda)},$$

where we have now explicitly written $f(x)$ as a function of the smoothing parameter λ and $p^{(\lambda)}$ is the effective degrees of freedom. Figure 11.16 gives the resultant fit, which has an effective degrees of freedom of 3.0. It is difficult to determine the adequacy of the fit with binary data, but in terms of smoothness and monotonicity, the curve appears reasonable. Notice that the behavior for high birthweights is quite different from the linear logistic model.

11.5.2 A Generalized Linear Mixed Model Spline Representation

The regression spline model described in Sect. 11.5.1 has an equivalent specification as a *generalized linear mixed model* (Sect. 9.3) with the assumption that $b_l \mid \sigma_b^2 \sim_{iid} \mathrm{N}(0, \sigma_b^2)$, $l = 1, \ldots, L$. The latter random effects distribution penalizes the truncated basis coefficients.

For a GLM with canonical link, maximization of the penalized log-likelihood (11.53) is then equivalent to maximization of

$$\frac{1}{\alpha}\left[\sum_{i=1}^{n}\{y_i(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}) - b(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}) + \alpha \times c(y_i, \alpha)\} - \frac{\alpha}{2\sigma_b^2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right] \tag{11.58}$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$, for fixed α, σ_b^2. In practice, estimates of α, σ_b^2 will also be required and will determine the level of smoothing. As discussed in Chap. 9, rather than maximize (11.58) as a function of both $\boldsymbol{\beta}$ and $\boldsymbol{b}$, an alternative is to integrate the random effects $\boldsymbol{b}$ from the model and then maximize the resultant likelihood. This approach is outlined for the case of a binomial model.

The likelihood as a function of $\boldsymbol{\beta}$ and σ_b^2 is calculated via an L-dimensional integral over the random effects $\boldsymbol{b}$:

$$L(\boldsymbol{\beta}, \sigma_b^2) = \prod_{i=1}^{n}\binom{n_i}{y_i}\int_{\boldsymbol{b}} \exp\left\{y_i\left(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}\right) - n_i \log\left[1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b})\right]\right\}$$
$$\times\ (2\pi\sigma_b^2)^{-L/2}\exp\left(-\frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}}{2\sigma_b^2}\right)\, d\boldsymbol{b}$$

and may be maximized to find $\boldsymbol{\beta}$ and σ_b^2. For implementation, some form of approximate integration strategy must be used; various approaches are described in Chap. 9. The latter also contains details on how the random effects b_l may be estimated (as required to produce the fitted curve), as well as Bayesian approaches to estimation. Under the mixed model formulation, smoothing parameter estimation is carried out via estimation of σ_b^2. Maximizing jointly for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ is formally equivalent to penalized quasi-likelihood (Breslow and Clayton 1993); see Chaps. 10

and 11 of Ruppert et al. (2003) for the application of penalized quasi-likelihood to spline modeling.

Inference from a likelihood perspective may build on mixed model theory, as described in Chap. 9 (see also, Ruppert et al. 2003, Chap. 11). A Bayesian approach can be implemented using either INLA or MCMC, both of which are described in Chap. 3.

11.5.3 Generalized Linear Models with Local Polynomials

The extension of the local polynomial approach of Sect. 11.3.4 to GLMs is relatively straightforward with a locally weighted log-likelihood replacing the locally weighted sum of squares (11.42). Recall that for the ith data point, the canonical parameter is $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ (Sect. 6.3). The local polynomial replaces the linear model in θ_i so that we have a local polynomial on the linear predictor scale. We write the log-likelihood for $\boldsymbol{\beta}$ as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l\left[y_i, \theta_i(\boldsymbol{\beta})\right].$$

To obtain the fit at the point x under a local polynomial model, we maximize the locally weighted log-likelihood:

$$l_x(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(x)\, l_x\left[y_i, P_x(x_i;\boldsymbol{\beta})\right],$$

where $w_i(x) = K[(x_i - x)/\lambda]$ and $P_x(x_i;\boldsymbol{\beta})$ is the local polynomial with parameters $\boldsymbol{\beta}$. Our notation also emphasizes that the likelihood is constructed for each point x at which a prediction is desired. The local likelihood score equations are therefore

$$\sum_{i=1}^{n} w_i(x) \frac{\partial}{\partial \boldsymbol{\beta}} l_x\left[y_i, P_x(x_i;\boldsymbol{\beta})\right].$$

Once we have performed estimation, the estimate (on the transformed scale) for x is evaluated as $\widehat{\beta}_0$. This method is often referred to as *local likelihood*. The existence and uniqueness of estimates are discussed in Chap. 4 of Loader (1999). For a GLM, an iterative algorithm is required; Chap. 11 of Loader (1999) gives details based on the Newton–Raphson method. We stress that the equations are solved at all locations x for which we wish to obtain the fit. The smoothing parameter may again be chosen in a variety of ways, with cross-validation being an obvious approach.

Example: Bronchopulmonary Dysplasia

Returning to the BPD/birthweight example, local log-likelihood fitting at the point x is based on

$$l_x(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(x) n_i \left(\frac{y_i}{n_i} P_x(\boldsymbol{x}_i; \boldsymbol{\beta}) - \log \left\{ 1 + \exp \left[P_x(\boldsymbol{x}_i; \boldsymbol{\beta}) \right] \right\} \right), \quad (11.59)$$

with $w_i(x) = K[(x_i - x)/h]$. Writing the likelihood in this form emphasizes that $w_i(x)n_i$ is acting as a local weight.

Figure 11.16 shows the local linear likelihood fit with a tricube kernel and smoothing parameter chosen by minimizing the AIC. The latter produces a model with effective degrees of freedom of 4.1. The local likelihood cubic curve bears more resemblance to the penalized cubic spline curve than to the linear logistic model, but there are some differences between the former two approaches, particularly for birthweights in the 900–1,500-gram range.

11.6 Concluding Comments

In this chapter we have described smoothing methods for general data types based on spline models and kernel-based methods. A variety of spline models are available, but we emphasize that the choice of smoothing parameter will often be far more important than the specific model chosen. For simple scatterplot smoothing, the spline and kernel techniques of Sects. 11.2 and 11.3 will frequently produce very similar results. If inference is required, penalized regression splines are a class for which the theory is well developed and for which much practical experience has been gathered. To obtain confidence intervals for the complete curve, a Bayesian solution is recommended; see Marra and Wood (2012). For inference about a curve, including confidence bands, care must be taken in variance estimation, as described in Sect. 11.4. In terms of smoothing parameter choice, there will often be no clear optimal choice, and a visual examination of the resultant fit is always recommended.

Kernel-based methods are very convenient analytically, and we have seen that expressions for the bias and variance are available in closed form which allows insight into when they might preform well. Spline models are not so conducive to such analysis though penalized regression splines have the great advantage of having a mixed model representation which allows the incorporation of random effects and the estimation of smoothing parameters using conventional estimation techniques.

11.7 Bibliographic Notes

Book-length treatments on spline methods include Wabha (1990) and Gu (2002). A key early reference on spline smoothing is Reinsch (1967). The book of Wand and Jones (1995) is an excellent introduction to kernel methods. Local polynomial methods are described in detail in Fan and Gijbels (1996) and Loader (1999). Bowman and Azzalini (1997) provides a more applied slant. The work of Ruppert et al. (2003) is a readable account of smoothing methods, with an emphasis on the mixed model representation of penalized regression splines.

11.8 Exercises

11.1 Based on (11.7) and (11.8), write code, for example, within `R`, to produce plots of the B-spline basis functions of order $M = 1, 2, 3, 4$, with $L = 9$ knots and for $x \in [0, 1]$.

11.2 Prove that (11.22) and (11.23) are equivalent to (11.24).

11.3 Show that an alternative basis for the natural cubic spline given by (11.3), with constraints (11.4) and (11.5), is

$$h_1(x) = 1, \;\; h_2(x) = x, \;\; h_{l+2}(x) = d_l(x) - d_{L-1}(x),$$

where

$$d_l(x) = \frac{(x - \xi_l)^3_+ - (x - \xi_L)^3_+}{\xi_L - \xi_l}.$$

11.4 In this question, various models will be fit to the fossil data of Chaudhuri and Marron (1999). These data consist of 106 measurements of ratios of strontium isotopes found in fossil shells and their age. These data are available in the `R` package `SemiPar` and are named `fossil`. Fit the following models to these data:

(a) A natural cubic spline (this model has n knots), using ordinary cross-validation to select the smoothing parameter.
(b) A natural cubic spline (this model has n knots), using generalized cross-validation to select the smoothing parameter.
(c) A penalized cubic regression spline with $L = 20$ equally spaced knots, using ordinary cross-validation to select the smoothing parameter.
(d) A penalized cubic regression spline with $L = 20$ equally spaced knots, using generalized cross-validation to select the smoothing parameter.
(e) A penalized cubic regression spline with $L = 20$ equally spaced knots, using a mixed model representation to select the smoothing parameter.

In each case report $\widehat{f}(x)$, along with an asymptotic 95% confidence interval, for the (smoothed) function, at $x = 95$ and $x = 115$ years.

11.5 In this question a dataset that concerns cosmic microwave background (CMB) will be analyzed. These data are available at the book website; the first column is the wave number (the x variable), while the second column is the estimated spectrum (the y variable):

(a) Fit a penalized cubic regression spline using, for example, the `R` package `mgcv`.
(b) Fit a Nadaraya–Watson locally constant estimator.
(c) Fit a locally linear polynomial model.
(d) Which of the three models appears to give the best fit to these data?
(e) Obtain residuals from the fit in part (c) and form the log of the squared residuals. Model the latter as a function of x.

(f) Compare the model for the fitted standard deviation with the estimated standard error (which is the third column of the data).

(g) Reestimate the linear polynomial model, weighting the observations by the reciprocal of the variance, where the latter is the square of the estimated standard errors (column three of the data). Repeat using your estimated variance function.

(h) Does the fit appear improved when compared with constant weighting?

At each stage provide a careful description of how the models were fitted. For example, in (a), how were the knots chosen, and in (b) and (c), what kernels and smoothing parameters were used and why?

11.6 For the locally linear polynomial fit described in Sect. 11.3.4, show that

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{i=1}^{n} w_i(x)} + (x - \overline{x}_w) \frac{\sum_{i=1}^{n} w_i(x)(x_i - \overline{x}_w) Y_i}{\sum_{i=1}^{n} w_i(x)(x_i - \overline{x}_w)^2}$$

where $\overline{x}_w = \sum_{i=1}^{n} w_i(x) x_i / \sum_{i=1}^{n} w_i(x)$ and $w_i(x) = K[(x - x_i)/\lambda]$ is a kernel.