Appendix: Nonparametric Smoothing

In this book we make use of two nonparametric smoothing techniques, namely, kernel density estimation and nonparametric regression for a single predictor. We discuss each of these in turn next.

A.1 Kernel Density Estimation

In this section we provide a brief practical description of density estimation based on kernel methods. We shall follow the approach taken by Sheather (2004).

Let $X_1, X_2, ..., X_n$ denote a sample of size n from a random variable with density function f. The kernel density estimate of f at the point x is given by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where the kernel, K satisfies $\int K(x)\,dx = 1$ and the smoothing parameter, h is known as the bandwidth. In practice, the kernel K is generally chosen to be a unimodal probability density symmetric about zero. In this case, K also satisfies the following condition

$$\int yK(y)dy = 0.$$

A popular choice for K which we shall adopt is the Gaussian kernel, namely,

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

Purely for illustration purposes we shall consider a small generated data set. The data consists of a random sample of size n = 20 from a normal mixture distribution made up of observations from a 50:50 mixture of $N\left(\mu = -1, \sigma^2 = 1/9\right)$ and $N\left(\mu = 1, \sigma^2 = 1/9\right)$. The data can be found on the book web site in the file bimodal.txt.

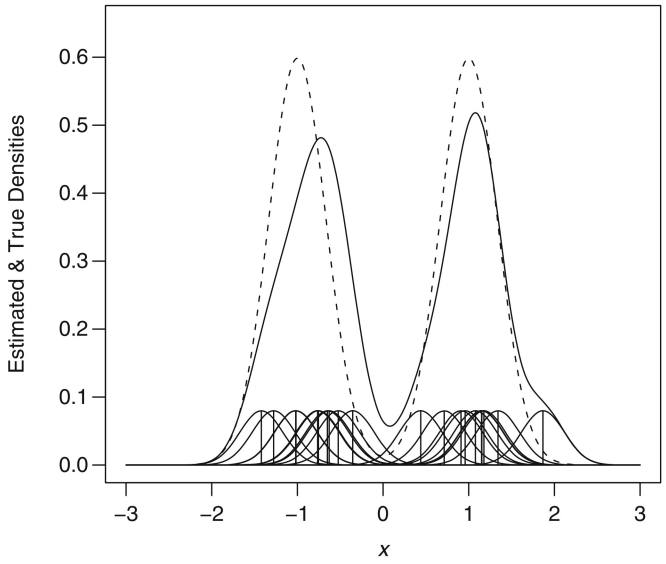

Figure A.1 True density (dashed curve) and estimated density with $h = 0.25$ (solid curve)

Figure A.1 shows a kernel density estimate for these data using the Gaussian kernel with bandwidth $h = 0.25$ (the solid curve) along with the true underlying density (the dashed curve). The 20 data points are marked by vertical lines above the horizontal axis. Centered at each data point is its contribution to the overall density estimate, namely, $\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)$ (i.e., $\frac{1}{n}$ times a normal density with mean X_i and standard deviation h). The density estimate (the solid curve) is the sum of these scaled normal densities. Increasing the value of h to 0.6 widens each normal curve producing a density estimate in which the two modes are less apparent (see Figure A.2).

Assuming that the underlying density is sufficiently smooth and that the kernel has finite fourth moment, it can be shown that the leading terms in an asymptotic expansion for the bias and variance of a kernel density estimate are given by

$$\text{Bias}_{\text{asy}}\left\{\hat{f}_h(x)\right\} = \frac{h^2}{2}\mu_2(K)^2 f''(x)$$

$$\text{Var}_{\text{asy}}\left\{\hat{f}_h(x)\right\} = \frac{1}{nh}R(K)f(x)$$

where

$$R(K) = \int K^2(y)dy,\ \ \mu_2(K) = \int y^2 K(y)dy$$

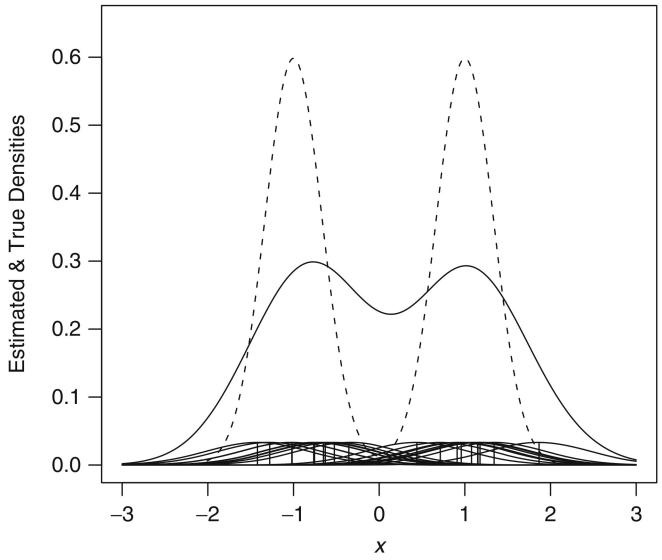

Figure A.2 True density (dashed curve) and estimated density with $h = 0.6$ (solid curve)

(e.g., Wand and Jones, 1995, pp. 20–21). In addition to the visual advantage of being a smooth curve, the kernel estimate has an advantage over the histogram in terms of bias. It can be shown that the bias of a histogram estimator with bandwidth h is of order h, compared to leading bias term for the kernel estimate, which is of order h^2. Centering the kernel at each data point and using a symmetric kernel makes the bias term of order h equal to zero for kernel estimates.

A widely used choice of an overall measure of the discrepancy between $\hat{f}_h$ and f is the mean integrated squared error (MISE), which is given by

$$\begin{aligned}\mathrm{MISE}(\hat{f}_h) &= \mathrm{E}\left\{\int \left(\hat{f}_h(y) - f(y)\right)^2 dy\right\} \\ &= \int \mathrm{Bias}\left(\hat{f}_h(y)\right)^2 dy + \int Var\left(\hat{f}_h(y)\right) dy\end{aligned}$$

Under an integrability assumption on f, the asymptotic mean integrated squared error (AMISE) is given by

$$\mathrm{AMISE}\left\{\hat{f}_h\right\} = \frac{1}{nh} R(K) + \frac{h^4}{4} \mu_2(K)^2 R(f'')$$

The value of the bandwidth that minimizes AMISE is given by

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{1/5} n^{-1/5}.$$

The functional $R(f'')$ is a measure of the underlying curvature. In particular, the larger the value of $R(f'')$ the larger the value of AMISE (i.e., the more difficult it is to estimate f) and the smaller the value of h_{AMISE} (i.e., the smaller the bandwidth needed in order to capture the curvature in f).

There are many competing methods for choosing a global value of the bandwidth h. For a recent overview of these methods see Sheather (2004).

A popular approach commonly called *plug-in methods* is to replace the unknown quantity $R(f'')$ in the expression for h_{AMISE} given above by an estimate. This method is commonly thought to date back to Woodroofe (1970) who proposed it for estimating the density at a given point. Estimating $R(f'')$ by $R(f_g'')$ requires the user to choose the bandwidth g for this estimate. There are many ways this can be done. We next describe the "solve-the-equation" plug-in approach developed by Sheather and Jones (1991), since this method is widely recommended (e.g., Simonoff, 1996, p. 77; Bowman and Azzalini, 1997, p. 34; Venables and Ripley, 2002, p. 129) and it is available in R, SAS and Stata.

Different versions of the plug-in approach depend on the exact form of the estimate of $R(f'')$. The Sheather and Jones (1991) approach is based on writing g, the bandwidth for the estimate $R(\hat{f}'')$, as a function of h, namely,

$$g(h) = C(K)[R(f'')/R(f''')]^{1/7} h^{5/7}$$

and estimating the resulting unknown functionals of f using kernel density estimates with bandwidths based on a normality assumption on f. In this situation, the only unknown in the following equation is h.

$$h = \left[\frac{R(K)}{\mu_2(K)^2 R(\hat{f}''_{g(h)})} \right]^{1/5} n^{-1/5}.$$

The Sheather–Jones plug-in bandwidth, h_{SJ} is the solution to this equation. For hard-to-estimate densities (i.e., ones for which $|f''(x)|$ varies widely due, for example, to the existence of many modes) the Sheather–Jones plug-in bandwidth tends to over-smooth and the method known as least squares cross-validation (Bowman and Azzalini, 1997, p. 32) can be recommended. However, in settings in which parametric regression models are appropriate, the Sheather–Jones plug-in bandwidth appears to perform well.

A.2 Nonparametric Regression for a Single Predictor

In this section we provide a brief practical description of nonparametric regression for a single predictor, which is sometimes called scatter plot smoothing. In this section we are interested in nonparametric estimates of the regression function, $m(.)$ under the assumption of iid errors with constant variance. Thus, in symbols, we assume the following model for $i = 1, \ldots, n$

$$Y_i = m(x_i) + e_i = \mathrm{E}(Y \mid X = x_i) + e_i.$$

We shall consider two classes of estimators, namely, local polynomial kernel estimators and penalized linear regression splines.

A.2.1 Local Polynomial Kernel Methods

Local polynomial kernel methods (Stone, 1977; Cleveland, 1979) are based on the idea of approximating $m(x)$ by a low-order polynomial putting highest weight on the values of y corresponding to x_i's closest to x. According to Cleveland (1979), the idea of local fitting of polynomials to smooth scatter plots of time series, measured at equally spaced time points, dates back to at least the 1930s. The local polynomial estimator $\hat{m}_p(x)$ is the value of b_0 that minimizes

$$\sum_{i=1}^{n} \left\{ y_i - b_0 - b_1 (x_i - x) - b_2 (x_i - x)^2 - \cdots b_p (x_i - x)^p \right\}^2 \frac{1}{h} K\left(\frac{x_i - x}{h} \right)$$

where once again the kernel, K satisfies $\int K(x)dx = 1$ and the smoothing parameter, h is known as the bandwidth.

The local constant estimator is obtained by setting $p = 0$ in the last equation. Thus, in this case we seek to minimize

$$\sum_{i=1}^{n} \{y_i - b_0\}^2 \frac{1}{h} K\left(\frac{x_i - x}{h} \right)$$

Differentiating with respect to b_0 and setting the result to zero gives

$$-2 \sum_{i=1}^{n} \{y_i - b_0\} \frac{1}{h} K\left(\frac{x_i - x}{h} \right) = 0$$

Solving this equation for b_0 gives the local constant estimator $\hat{m}_0(x)$ where

$$\hat{m}_0(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h} \right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h} \right)}$$

This estimator is also known as the Nadaraya-Watson estimator, as they were the first to propose its use (Nadaraya, 1964; Watson, 1964). It is also possible to derive an explicit regression for the local linear estimator $\hat{m}_1(x)$ (see, e.g., Wand and Jones, 1997, pp. 119, 144).

Choosing a higher degree polynomial leads in principle to a better approximation to the underlying curve and hence less bias. However, it also leads to greater variability in the resulting estimate. Loader (1999, p. 22) provides the following advice:

> It often suffices to choose a low degree polynomial and concentrate on choosing the bandwidth to obtain a satisfactory fit. The most common choices are local linear and local quadratic. … a local constant fit is susceptible to bias and is rarely adequate. A local linear estimate usually performs better, especially at boundaries. A local quadratic estimate reduces bias further, but increased variance can be a problem, especially at boundaries. Fitting local cubic and higher orders rarely produces much benefit.

Based on their experience, Ruppert, Wand and Carroll (2003, p. 85) recommend $p = 1$ if the regression function is monotonically increasing (or decreasing) and $p = 2$ otherwise.

For illustration purposes we shall consider a generated data set. The data consists of $n = 150$ pairs of points (x_i, y_i) where $y_i = m(x_i) + e_i$ with x_i equally spaced from 0 to 1, $e_i \sim N(0, \sigma^2 = 4)$ and

$$m(x_i) = 15\left(1 + x_i \cos(4\pi x_i)\right)$$

The data can be found on the book web site in the file curve.txt.

Figure A.3 shows a local linear regression estimate for these data using the Gaussian kernel with bandwidth $h = 0.026$ (the solid curve) along with the true underlying curve (the dashed curve). The value of the bandwidth was chosen using the plug-in bandwidth selector of Ruppert, Sheather and Wand (2005). Marked as a dashed curve on Figure A.3 is the weight function for each x_i used to estimate the curve at $x = 0.5$, namely, $\frac{1}{h} K\left(\frac{x_i - 0.5}{h}\right)$ (i.e., a normal density with mean 0.5 and standard deviation h).

Decreasing the value of h fivefold to 0.005, shrinks each normal curve so that each straight line is effectively fit over a very small interval. This produces a curve estimate which is much too wiggly (see the top panel of Figure A.4). On the other hand, increasing the value of h fivefold to 0.132 widens each normal curve so that each straight line is effectively fit over a very large interval. This produces a curve estimate which is clearly over-smoothed, missing the bottom or the top of the peaks in the underlying curve (see the bottom panel of Figure A.4). As the bandwidth h approaches infinity the local linear regression estimate will approach a straight line.

Thus far, in this section we have considered an example based on equally spaced x's. In settings in which parametric regression models are generally appropriate it is common for the x's not to be equally spaced. In particular, outliers, and sparse regions in the x values are common when the distribution of x is skewed. In such situations using a fixed value of the bandwidth h can be problematic, since there may be very few (sometime even no) points in certain regions of the x-axis so that it is not possible to fit a local polynomial for certain values of x. One way of solving this problem is to adjust the bandwidth with the value of x so that the number of points used to estimate $m(x)$ effectively remains the same for all values of x. This is achieved using the concept of the *nearest neighbor bandwidth*.

For $i = 1, 2, \ldots, n$, let $d_i(x)$ denote the distance x_i is away from x, then

$$d_i(x) = |x - x_i|$$

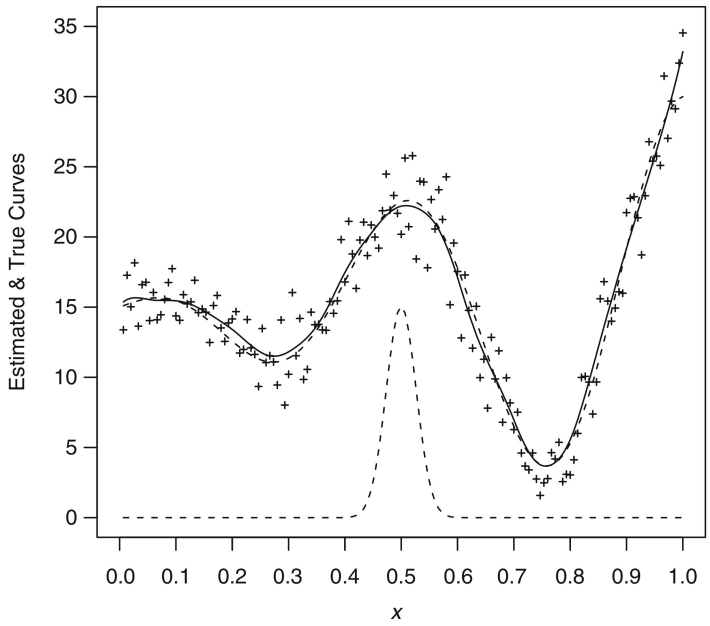

Figure A.3 True curve (dashed) and estimated curve with $h = 0.026$ (solid)

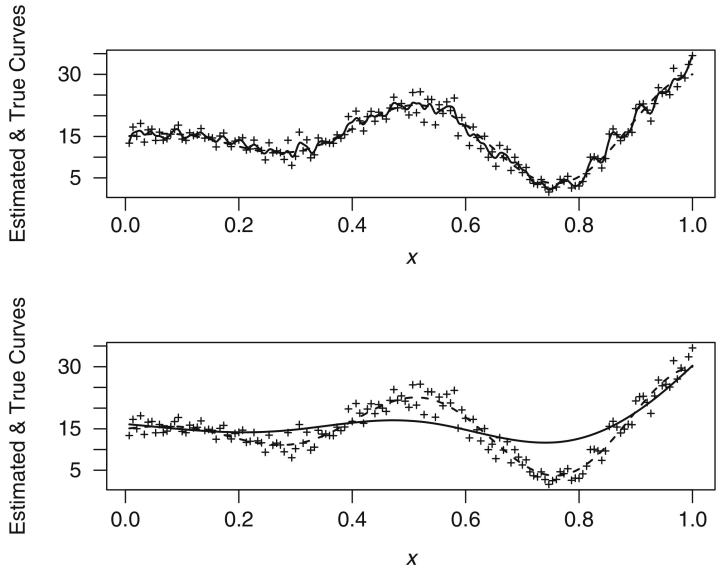

Figure A.4 True curve (dashed) and estimated curves (solid) with $h = 0.005$ (upper panel) and $h = 0.132$ (lower panel)

The *nearest neighbor bandwidth*, $h(x)$ is defined to be the kth smallest $d_i(x)$. In practice, the choice of k is based on what is commonly called the *span* α, namely,

$$k = \lfloor n\alpha \rfloor.$$

Thus, the span plays the role of a smoothing parameter in nearest neighbor bandwidths.

Cleveland (1979) proposed the use of local linear regression estimators based on nearest neighbor bandwidths with the tricube kernel function

$$K(y) = \left(1 - |y|^3\right) I\left(|y| < 1\right).$$

Cleveland (1979) also incorporated a robustness step in which large residuals were down weighted. This estimator is typically referred to as *lowess*. Cleveland and Devlin (1988) studied the properties of local linear regression estimators based on nearest neighbor bandwidths with the tricube kernel without a robustness step. This estimator is typically referred to as *loess*.

Figure A.5 shows the loess estimate based on $p = 2$ (i.e., local quadratic) with span $\alpha = 1/3$ (the solid curve), along with the true underlying curve (the dashed curve). This value of the span was chosen by eye as the value that gave a curve that seemed to best match the data.

The loess estimate with span $\alpha = 1/3$ in Figure A.5 fits the data well. Increasing the span to $\alpha = 2/3$, produces a curve estimate which is slightly over-smoothed, missing

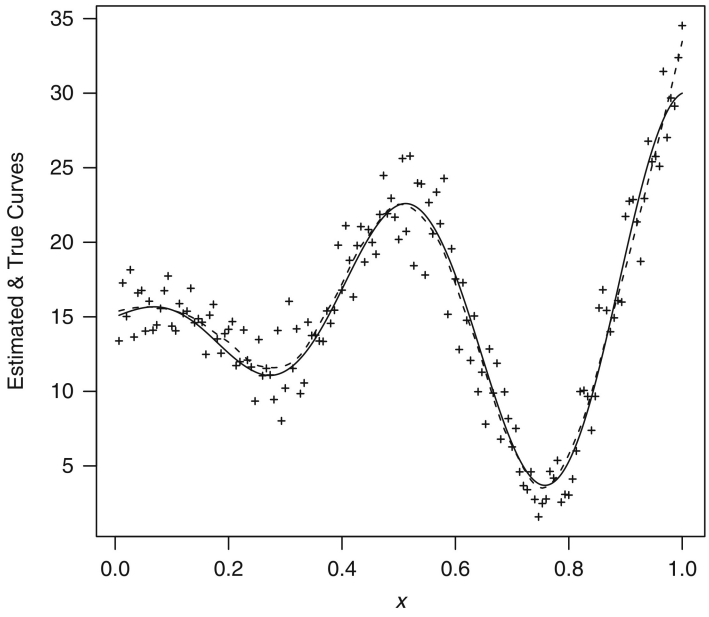

Figure A.5 True curve (dashed) and estimated curve (solid) with span = 1/3

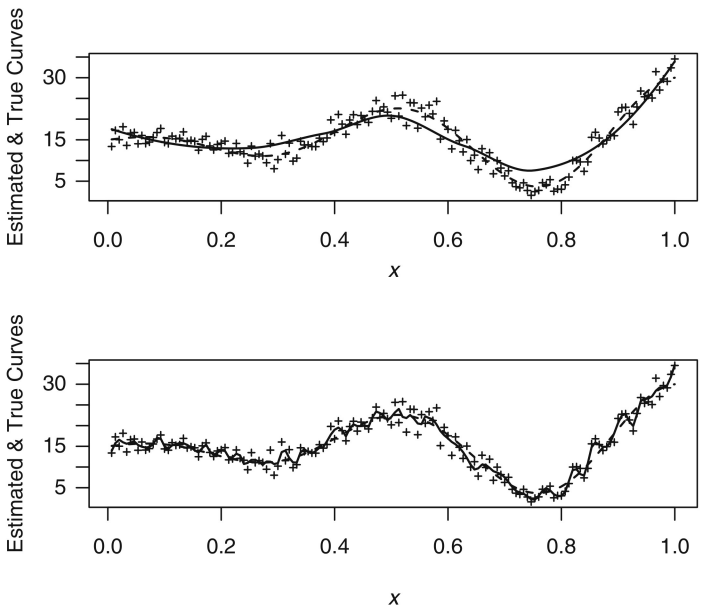

Figure A.6 True curve (dashed) and estimated curves (solid) with span = 2/3 (upper panel) and span = 0.05 (lower panel)

the bottom or the top of the peaks in the underlying curve (see the top panel of Figure A.6). On the other hand, decreasing the value of the span to $\alpha = 0.05$ produces a curve estimate which is much too wiggly (see the bottom panel of Figure A.6).

Nearest neighbor bandwidths do not perform well if the x-space is sparse when the curve is wiggly and/or the x-space is dense when the curve approximates a straight line. Fortunately, this is a highly unusual situation.

The marginal model plot method, proposed by Cook and Weisberg (1997) and described in Chapters 6 and 8, is based on loess fits. This is a natural choice for regression with continuous predictor and outcome variables due to the ability of loess to cope with sparse regions in the x-space. However, its use for binary outcome variables can be questioned, since it seems that no account is taken of the fact that binary data naturally have nonconstant variance. In this situation one could consider a local likelihood estimator, which takes account of the binomial nature of the data (see, e.g., Bowman and Azzalini, 1997, p. 55).

A.2.2 Penalized Linear Regression Splines

Another increasingly popular method for scatter plot smoothing is called penalized linear regression splines, which we discuss in this section. However, we begin by discussing linear regression splines.

Linear regression splines are based on the inclusion of the following term as a predictor

$$(x-c)_+ = \begin{cases} x-c & \text{if } x>c \\ 0 & \text{if } x\le c \end{cases}$$

The inclusion of $(x - c)_+$ as a predictor produces a fitted model which resembles a broken stick, with the break at c, which is commonly referred to as a knot. Thus, this predictor allows the slope of the line to change at c. (See Figure 10.16 for details.) In order to make the model as flexible as possible, we shall add a large number of knots c_1, ….,c_K and hence consider the following model

$$y = \beta_0 + \beta_1 x + \sum_{i=1}^{K} b_{1i}(x-c_i)_+ + e \tag{A.1}$$

We shall see that two approaches are possible for choosing the knots, corresponding to whether the coefficients b_{1i} in (A.1) are treated as fixed or random effects. If the coefficients are treated as fixed effects, then a number of knots can be removed leaving only those necessary to approximate the function. As demonstrated in Chapter 10, this is feasible if there are a relatively small number of potential knots. However, if there are a large number of potential knots, removing unnecessary knots is a "highly computationally intensive" variable selection problem (Ruppert, Wand and Carroll, 2003, p. 64).

We next investigate what happens if the coefficients b_{1i} in (A.1) are treated as random effects. In order to do this we consider the concept of penalized regression splines.

An alternative to removing knots is to add a penalty function which constrains their influence so that the resulting fit is not overfit (i.e., too wiggly). A popular penalty is to ensure that the b_{1i} in (A.1) satisfy $\sum_{i=1}^{K} b_{1i}^2 < C$, for some constant C, which has to be chosen. The resulting estimator is called a *penalized linear regression spline*. As explained by Ruppert, Wand and Carroll (2003, p. 66) adding this penalty is equivalent to choosing $\beta_0 \beta_1, b_{11}, b_{12}$, ... b_{1K} to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 + \lambda\sum_{i=1}^{K} b_{1i}^2 \tag{A.2}$$

for some number $\lambda \ge 0$, which determines the amount of smoothness of the resulting fit. The second term in (A.2) is known as a roughness penalty because it penalizes fits which are too wiggly (i.e., too rough). Thus, minimizing (A.2) shrinks all the b_{1i} toward zero. Contrast this with treating the b_{1i} as fixed effects and removing unnecessary knots, which reduces some of the b_{1i} to zero.

The concept of random effects and shrinkage is discussed in Section 10.1. In view of the connection between random effects and shrinkage, it is not too surprising that there is a connection between penalized regression splines and mixed models. Put briefly, the connection is that fitting model (A.1) with β_0 and β_1 treated as fixed effects and $b_{11,} b_{12}$, ... b_{1K} treated as random effects is equivalent to minimizing the penalized linear spline criterion (A.2) (see Ruppert, Wand and Carroll, 2003; Section 4.9 for further details).

Speed (1991) explicitly made the connection between smoothing splines and mixed models (although it seems that this was known earlier by a number of the

proponents of spline smoothing). Brumback, Ruppert and Wand (1999) made explicit the connection between penalized regression splines and mixed models.

An important advantage of treating (A.1) as a mixed model is that we can then use the likelihood methods described in Sect. 10.1 to obtain a penalized linear regression spline fit.

Finally, one has to choose the initial set of knots. Ruppert, Wand and Carroll (2003, p. 126) recommend that the knots be chosen at values corresponding to quantiles of x_i, while other authors prefer equally spaced knots. Ruppert, Wand and Carroll (2003, p. 126) have found that the following default choice for the total number of knots K "usually works well":

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

Figure A.7 shows a penalized linear regression spline fit obtained by fitting (A.1) using restricted maximum likelihood or REML (the solid curve) along with the true underlying curve (the dashed curve). The equally spaced knots, which are 0.02 apart, are marked by vertical lines on the horizontal axis. Notice this is many more knots than is suggested by the rule above and it does not have any adverse effects on the fit. Increasing the spacing of the knots to 0.15 produces a curve estimate which is jagged, missing the bottom or the top of the peaks in the underlying curve and thus illustrating the problems associated with choosing too few knots (see Figure A.8).

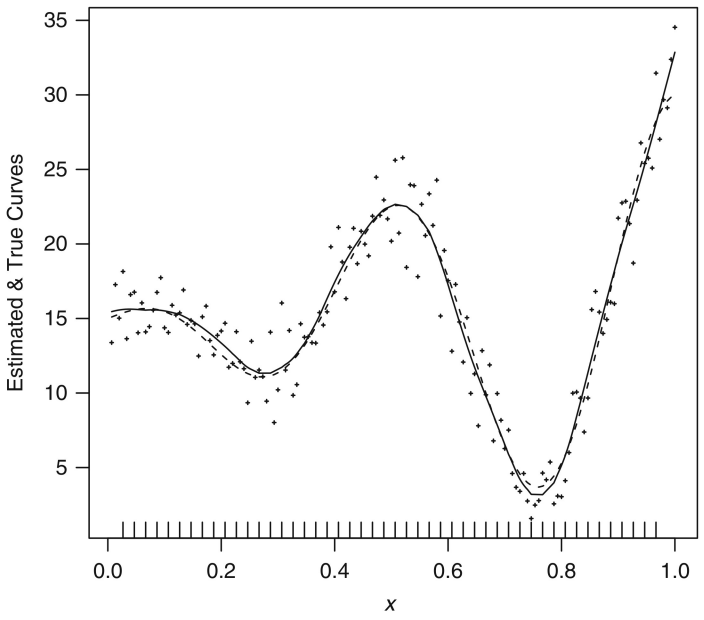

Figure A.7 True curve (dashed) and estimated curve (solid) with knots 0.02 apart

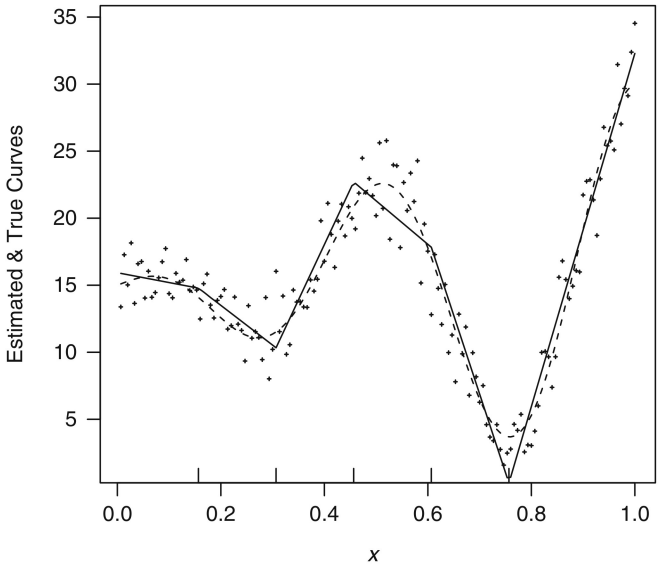

Figure A.8 True curve (dashed) and estimated curve (solid) with knots 0.15 apart

Recently, Krivobokova and Kauermann (2007) studied the properties of penalized splines when the errors are correlated. They found that REML-based fits are more robust to misspecifying the correlation structure than fits based on generalized cross-validation or AIC. They also demonstrated the simplicity of obtaining the REML-based fits using R.