# Statistical Modeling with Spline Functions Methodology and Theory

Mark H. Hansen
University of California at Los Angeles

Jianhua Z. Huang
University of Pennsylvania

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Charles J. Stone
University of California at Berkeley

Young K. Truong
University of North Carolina at Chapel Hill



January 5, 2006

# 10
# Alternate Optimization Methods

In the previous chapters, we explored greedy, deterministic search algorithms for knot placement that are meant to approximately minimize a model selection criterion like AIC. Broadly, each procedure involves the use of stepwise knot addition to grow a model of dimension $M$, after which knots are deleted sequentially. Clearly, the number of models examined during this two-pass process is at most $2M$. In the previous chapters, we have taken $M$ to be of the form

$$M = Cn^{\alpha} \tag{10.0.1}$$

where $n$ denotes sample size and $\alpha$ is some constant in $(0, 1]$. This expression was derived largely through simulation and empirical observation, although it can be loosely justified via the theoretical work in Chapter 13.[1] For example, in Logspline and Hare we set the constants in (10.0.1) to be $\alpha = 0.2$ and $C = 4$. Therefore, given one million data points, the largest model we will fit for either of these schemes consists of at most $M = 65$ basis functions.

By way of comparison, consider estimating a function of a single variable, say a univariate density (as in Logspline) or a univariate regression function (as in the scatterplot smoothing methods of Chapter 3). Suppose we place $K$ candidate knots at equally spaced quantiles of the input data. Associating subsets of these points with knot sequences, we can construct

[1] Depending on the assumed smoothness of the unknown function, these results tell us roughly (the order of magnitude) how many knots we should use when fitting a model with fixed breakpoints.

$2^K$ different spline spaces. For even modest values of $K$, the number of models left *unexplored* by our stepwise algorithm is enormous. While many are likely to provide very poor fits to the data, it is not unreasonable to wonder about the models ignored by our greedy search.

The main motivation for stepwise methods is that they are efficient computationally. Clearly, by entertaining a very small subset of the candidate models, we incur significant savings. In addition, by introducing simple approximations to the final model selection criterion, we can greatly reduce the effort required to evaluate "nearby" fits; or more precisely, models that differ by the position, presence or absence of a single knot. As seen repeatedly in Chapters 4 through 8, these shortcuts typically involve simple Taylor's expansions of the log-likelihood. For example, the use of Rao and Wald statistics to evaluate the impact of adding or deleting a single knot makes Polyclass, Logspline, Hare, and Lspec feasible. Although discussed in connection with greedy, stepwise algorithms, these approximations and the corresponding computational savings can be applied in any procedure that iteratively adds and deletes knots.

In this chapter we explore whether the gains in estimation performance outweigh the computational expense of more exhaustive search procedures. A major source of inspiration for this study comes from the recent work on Bayesian model selection and the accompanying Markov chain Monte Carlo (McMC) schemes for identifying promising models. In the last decade, Bayesian approaches to model selection have advanced considerably, mainly through the development of convenient computational tools. As we have seen, the greedy methods like MARS and TURBO were constructed by borrowing ideas from traditional approaches to model selection in linear models. Recently, several Bayesian spline methods have also made use of the connection between variable selection and knot placement, and in so doing have brought new ideas from Bayesian computing into the practice of nonparametric estimation.

## 10.1 Normal linear regression revisited

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote $n$ independent observations from a (normal) regression model. To describe the conditional mean of $Y_i$ given $X_i$, we consider splines of order $k$. In the notation of Chapter 3, let $\boldsymbol{t} = (t_1, \ldots, t_m)$ denote a sequence of $m$ knots or breakpoints and let $\mathcal{S}_k(\boldsymbol{t})$ denote the space of splines (piecewise polynomials) of order $k$ having $k-1$ continuous derivatives across each point in $\boldsymbol{t}$. So defined, $\mathcal{S}_k(\boldsymbol{t})$ is a $(m+k)$-dimensional linear space with the collection of functions

$$1, x, \ldots, x^{k-1}, (x - t_1)_+^{k-1}, \ldots, (x - t_m)_+^{k-1} \tag{10.1.1}$$

forming a basis. An estimate of the unknown regression function is obtained by an ordinary least squares (OLS) projection into $\mathcal{S}_k(\boldsymbol{t})$. As we have seen,

the success of this simple scheme depends on how we arrange points in $\boldsymbol{t}$: We would like to place more knots in regions exhibiting strong features, while locating relatively fewer breakpoints in other areas. Unfortunately, we rarely know much about the function we are trying to estimate and must rely on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to help us determine both the number of knots as well as their placement. By examining how knot placement affected the statistical properties of an OLS projection into $\mathcal{S}_k(\boldsymbol{t})$, we derived constraints on $\boldsymbol{t}$ that related to (possibly realistic) prior assumptions about the smoothness of the unknown regression function. One such constraint involved the minimum number $M$ of input points $X_1, \ldots, X_n$ that must separate neighboring elements of $\boldsymbol{t}$. The so-called minimal span $M$ helps to reduce the influence of noise when using the data to find $\boldsymbol{t}$. In so doing, it acts as a smoothing parameter, implicitly restricting the shape of the fitted curve.

In addition to the span $M$, we also considered other aspects of the structure of $\mathcal{S}_k(\boldsymbol{t})$ that determine the statistical properties of the spline fit. For example, rules for extrapolating beyond the support of the data could help reduce excess variability near the boundaries. In the case of cubic splines ($k = 4$), the so-called natural boundary conditions forced each curve to blend smoothly (two continuous derivatives) to a line outside the range of the data. Recall from Chapter 3, that the set of functions in $\mathcal{S}_4(\boldsymbol{t})$ that satisfy this condition again forms a linear space with a basis given by

$$\phi_1(x)\,,\ \phi_2(x)\,, \quad \text{and} \quad R(x, t_l)\,,\ l = 1, \ldots, m \tag{10.1.2}$$

where $\phi_1(x) = 1$, $\phi_2(x) = k_1(x)$ and

$$R(x, x') = k_2(x)k_2(x') - k_4\,(\,|x - x'|\,)\ . \tag{10.1.3}$$

The functions $k_1$, $k_2$, and $k_4$ are constant multiples of Bernoulli polynomials and are given by

$$k_1(x) = x - 1/2\,, \qquad k_2(x) = (k_1^2(x) - 1/12)/2$$

and

$$k_4(x) = (k_1^4(x) - k_1^2(x)/2 + 7/240)/24\,.$$

In deriving these closed forms, we have assumed that the predictor variables $X_i$ all belong to the interval $[0, 1]$, perhaps as a result of rescaling.

In Chapter 3, we applied popular selection criteria like AIC to evaluate the relative performance of different knot sequences $\boldsymbol{t}$. Let $RSS(\boldsymbol{t})$ denote the residual sum of squares associated with the OLS projection into $\mathcal{S}_k(\boldsymbol{t})$. Recall that for normal regression, the family of generalized AIC criteria is given by

$$AIC(\boldsymbol{t}) = \frac{n}{2} \log RSS(\boldsymbol{t}) + \frac{\alpha}{2} J(\boldsymbol{t})\,, \tag{10.1.4}$$

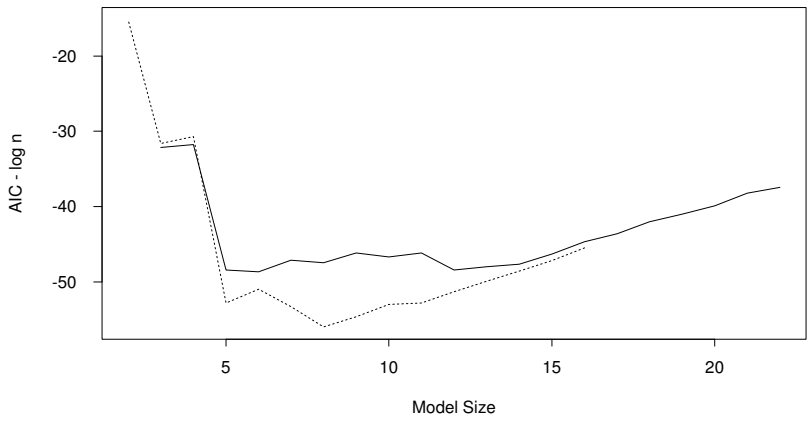


FIGURE 10.1. AIC (with $\alpha = \log n$) as a function of the model size for the $^{87}\delta\,\mathrm{Sr}$ data using a stepwise addition (solid) followed by deletion (dotted) algorithm.

where $J(\boldsymbol{t})$ denotes the dimension of the spline space $\mathcal{S}_k(\boldsymbol{t})$. Through the penalty $\alpha$, we explicitly trade fidelity to the data (as measured by the residual sum of squares of the OLS fit) with model complexity (the dimension of the spline space, or rather the degrees of freedom associated with the OLS fit). The fact that the flexibility of a spline space increases with knot count implies that $\alpha$ acts as a smoothing parameter: by choosing a large value for $\alpha$, we focus our interest on spline spaces with very few breakpoints.

So far, we have only studied simple greedy algorithms for identifying a the knot sequence $\boldsymbol{t}$. These techniques were applied to both the ordinary polynomial splines (10.1.1) as well as the space of natural splines (10.1.2). As their name suggests, these greedy schemes add and delete knots sequentially, at each stage performing an act that is optimal in terms of its effect on the selection criterion (10.1.4). When adding a knot, this means we select that single point which yields the greatest drop in AIC (or, equivalently, the residual sum of squares). Similarly, when removing a knot from an existing sequence, we choose the point that produces the smallest rise in AIC (or, equivalently, the residual sum of squares). Such schemes make it easy to incorporate restrictions on the knot sequence (say, enforcing a given minimal span), and are easily implemented using the bases in (10.1.1) or (10.1.2) and standard statistical computing environments.

### *10.1.1 Greedy methods and the $^{87}\delta$ Sr data*

Consider again data introduced in Chapter 3 recording a standardized ratio, $^{87}\delta\,\mathrm{Sr}$, of strontium isotopes $^{87}\mathrm{Sr}$ to $^{86}\mathrm{Sr}$ present in shells of marine organisms. Recall that interest lies in the shape of the regression function near the Cretaceous-Tertiary boundary (roughly 66 million years ago), referred to as the KTB. Starting from a simple linear relationship (consisting

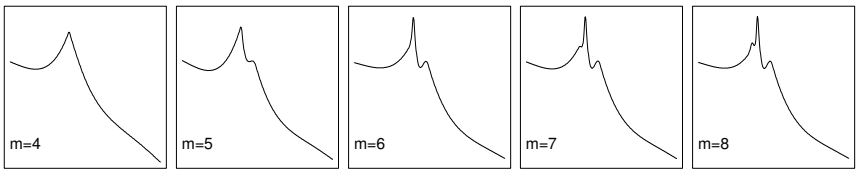


FIGURE 10.2. Curves found during backward deletion beginning with a model consisting of 14 knots. Here, $m$ refers to the number of knots in the model, and the dimension of the natural spline space is then $m + 2$.

only of the basis functions 1 and $x$), knots were added sequentially to a natural spline space. Initially, we set the minimum span $M = 0$, so that any knot could be entered providing it was not already in the model. At each step, we considered the set of unique input points $X_1, \ldots, X_n$ as candidates for addition. The AIC values associated with each spline model encountered during this process is plotted as a function of dimension in Figure 10.1 (upper curve). Knots were then removed, producing another series of models that were again evaluated via AIC. Starting from each model consisting of between 10 and 14 knots (models of size 12 to 16), the deletion process consistently found the same spline space of six knots. This range of starting points for stepwise deletion covers all the rules mentioned in Chapter 3. The AIC values for the sequence of models starting with the 14-knot fit are also plotted in Figure 10.1 (lower curve).

In Figure 10.2, we plot the best AIC fit (corresponding to 8 knots, or a 10-dimensional space) as well as the four nearest fits, in terms of this selection criterion, encountered during either knot addition or deletion. These models vary largely in the region around the KTB, exhibiting from between one and three modes! While the nearby curves consisting of 4, 5, 7 and 8 knots are close in terms of the selection process, their appearance can be quite different. In terms of AIC, the 6-knot model is deemed to be the best for penalties $\alpha$ ranging from 1 to 9 (keep in mind that traditional AIC would take $\alpha = 2$ and BIC selects $\alpha = \log 45 = 3.8$). For smaller values of $\alpha$, we can tip the criterion in favor of models with more knots ($m = 7$ or 8), and with a larger $\alpha$, we can force the selection of one of the smaller models ($m = 4$ or 5). This story does not change substantially if we enforce a minimum span $M$ larger than zero. In Figure 10.3, we plot the minimal AIC models setting $M$ equal to 1, 2 and 3. As with the $M = 0$ case, there was considerable agreement across starting points of the deletion process, where again our attempt was to cover all the suggested rules presented in Chapter 3. Here, we can see clearly that the span restriction has resulted in smoother selected models.

Throughout this text, we have emphasized the need to examine the curves obtained by varying parameters like the span $M$ or the penalty on dimension $\alpha$ used in defining AIC. This analysis helps us to assess the

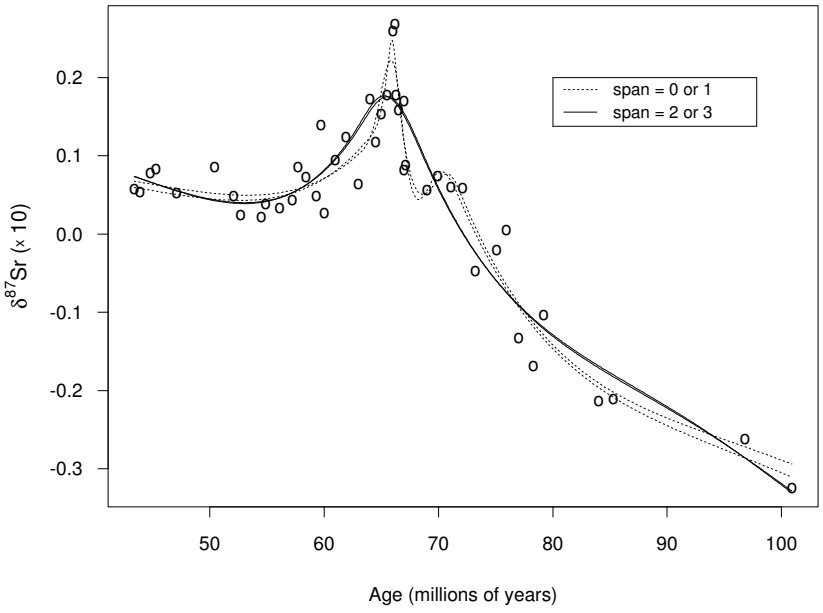


FIGURE 10.3. Several fits for the $^{87}\delta$Sr data.

believability of features evident in our final, selected curve. We have also introduced various simulation techniques to guide our judgment about the height of peaks, the depth of values or the number of modes. In each case, however, these tools make use of the same greedy search procedure for knot placement. While motivated largely for computational purposes, it is not unreasonable to question whether the fits in Figures 10.2 or 10.3 are representative of the entire population of models. How much different is the best (as measured by AIC) 6-knot model from the one found during our passes of stepwise knot addition and deletion? In the knot-selection context, we are dealing with potentially large sets of candidate predictor variables, perhaps one per data point. For such applications, greedy schemes are a compromise. While they can easily miss large numbers of "good fitting" models, it is well known that more exhaustive search procedures frequently identify spurious structures (a problem often referred to as selection bias). Therefore, we might naturally question whether, in addition to explicit parameters like $M$ or $\alpha$, the *search procedure itself* might have an effect on the statistical properties of the final selected model. Next, we will explore these difficulties in more detail using the $^{87}\delta$Sr data from Chapter 3.

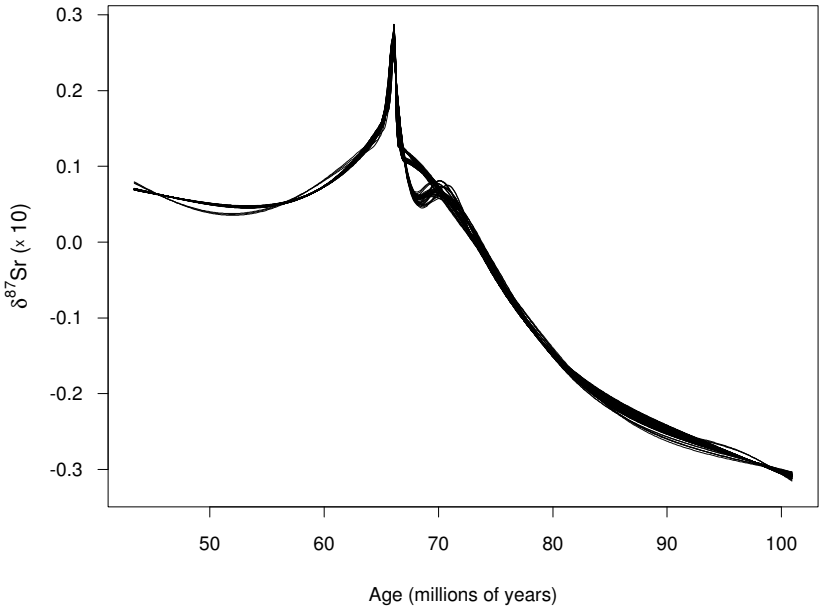


FIGURE 10.4. All models that achieve an AIC score within 1% of the best model for the $^{87}\delta\,\mathrm{Sr}$ data, as determined by an exhaustive search.

### *10.1.2 Results from an exhaustive search*

At some level, a stepwise scheme for $^{87}\delta\,\mathrm{Sr}$ data is unnecessary. The sample size ($n = 45$) is small enough that it is possible to evaluate every knot configuration that can be formed from the unique input points $X_1, \ldots, X_n$. To make this process stable numerically, we considered the space of natural splines (10.1.2), mapping our support of the original data, $[44.3, 100.9]$, to the interval $[0, 1]$. In so doing, our fits will join smoothly to a line both before the first data point $X_1$ and after the last $X_n$. This leaves us with 43 candidate basis functions. However, as one of the input points near the KTB is repeated, we in fact have only 42 unique candidates. After standardizing the kernel functions (10.1.3) to have mean zero and variance one, we used a well-known exhaustive search procedure for regression modeling (also known as "regression by leaps and bounds," this method was first introduced by Furnival and Wilson (1974).

For the moment, we do not consider the entire set of $2^{42}$ models, but instead focus on all spline spaces consisting of 15 or fewer knots. In Figure 10.4, we plot all such models achieving an AIC score that is not more than 1% larger than the smallest value computed for this subset. There are roughly two regimes evident in this figure, a unimodal and a bimodal fit. After closer inspection, we see that the unimodal curves can be further classified into two groups, one dropping to lower values across the KTB

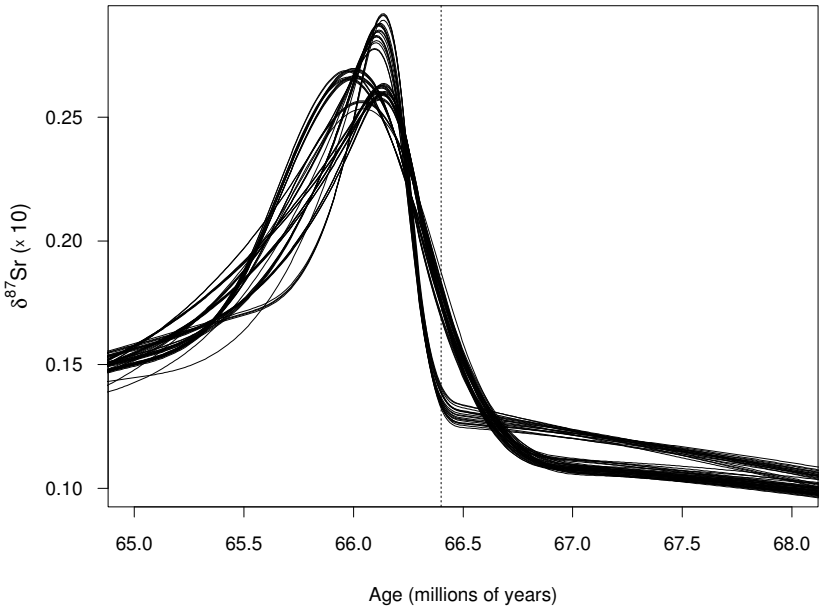


FIGURE 10.5. Close up of the drop across the KTB for the unimodal fits found by an exhaustive search.

than the other; see Figure 10.5. The bimodal and unimodal fits are mixed in terms of their AIC values, with the absolute minimum in AIC achieved by a unimodal curve. The models encountered during our greedy search are at least 2% larger than the minimum AIC value, and hence are not represented in this sample of curves. By examining more models, we observe variations that result in essentially equivalent fits (from the point of view of AIC). In so doing, we might question the reliability of features that are present in only a few models. In Figure 10.6 we present Quantile-Quantile plots of residuals computed under each of the bimodal and unimodal fits. Given that we have only 45 data points, these plots do not point to any deficiencies in either curve.

So far, we have restricted our attention to models that consist of at most 15 knots. We have done so mainly because this was the range suggested by our stepwise procedures. This kind of approach is frequently applied in regression analysis (see Miller 1990). By minimizing AIC over the entire range of knot configurations (with minimal span $M = 0$), however, we would choose a model consisting of 26 breakpoints. This situation is not uncommon when using selection criteria like AIC or GCV. As the number of candidate models increases, the potential for identifying spurious structures increases. In Figure 10.7, we present a curve of model dimension versus AIC. Because of the form of this criterion (10.1.4), the models represented

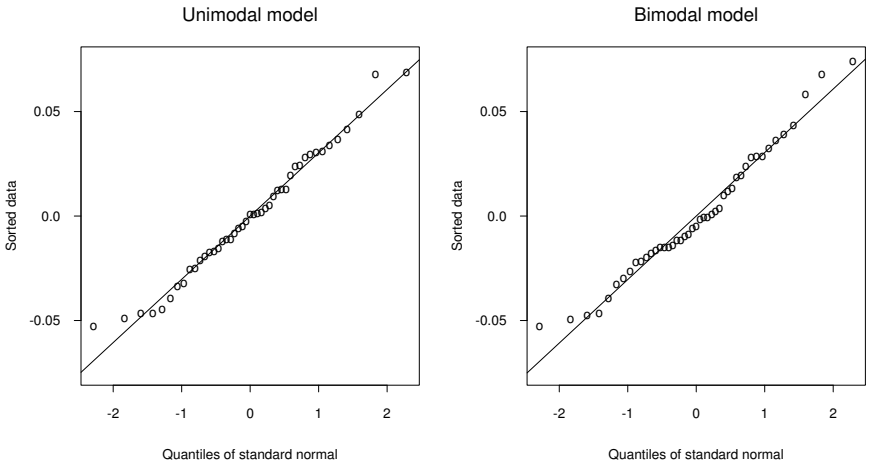


FIGURE 10.6. Quantile-Quantile plots of residuals computed under each of the bimodal and unimodal fits for the $^{87}\delta\,\mathrm{Sr}$ data.

by this curve achieve the lowest RSS among all models of their size. We plot the value of AIC for several choices of $\alpha$, each larger than the penalty assumed by BIC. The implication here is that we need roughly 1.2 times the BIC penalty (or $1.2J(\boldsymbol{t}) \log n/2$) to identify the 6-knot models as best. In a purely heuristic fashion, we can motivate this multiplier through the analysis in Section 5 of Chapter 3: the more extensive our search, the greater the degrees of freedom we need to charge for each additional basis function.

With the benefit of a complete search, we can see that several competing models seem to provide adequate descriptions of the data, and yet differ in their behavior near the KTB. Throughout this text, we have focused on deriving a single "best" model according to a given selection criterion. One can argue for this approach both in terms of computational savings and ease of interpretation. Indeed, as we saw in the previous section, the greedy schemes do not give an adequate reflection of the models in a neighborhood of the "best," so that it did not make sense to try to incorporate other fits encountered during the addition and deletion process. When more good-fitting models are available, however, it might make sense to combine the predictions from each rather than force a selection. Greedy schemes are notoriously susceptible to small changes in the input data, while average or ensemble models are less so. In Figure 10.8, we present both the single best fit as well as the average of the models within 1% of the best (both in terms of AIC). In this chapter, we will study the gains to be had by combining several good-fitting models in this way.

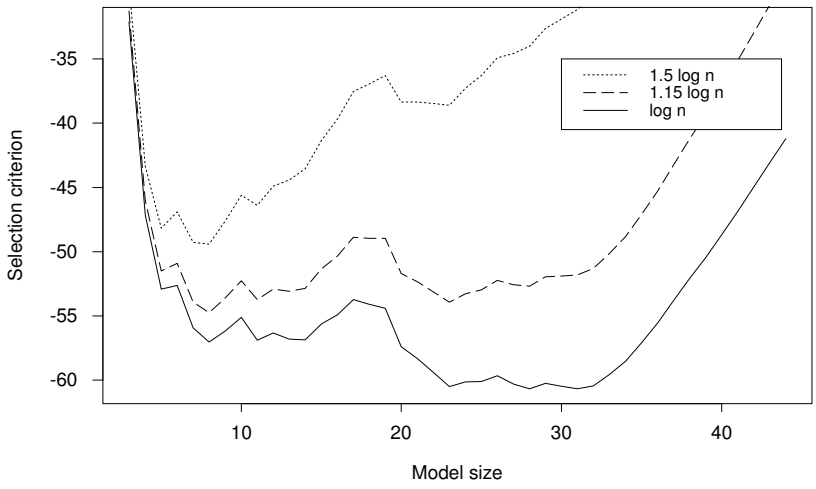


FIGURE 10.7. Several penalties for the selection criterion. Over the range $1.15 \log n = 4.4$ to $1.5 \log n = 5.7$, a model of dimension 8 (which has 6 knots) is preferred.

## 10.2 Bayesian Formulations

Our goal in this section is to present some well-known ideas from Bayesian statistics, providing enough notation and terminology to be able to motivate our application to function estimation. In the Bayesian approach to modeling, we find a convenient framework for considering both model selection and model combination, two topics that featured prominently in our previous discussion. Our connection to classical Bayesian statistics is again through the parametric nature of our estimates. That is, a spline space $\mathcal{G}$ is nothing more than a linear model. Well-known Bayesian methods for estimation and inference are now immediately applicable to the problem of curve and surface fitting. A significant component of our discussion will also cover supporting computational methods. In the last ten years, the field of Bayesian computation has become one of the most active areas of statistics research. We will present techniques for Markov chain Monte Carlo that make possible the construction of many "good fitting" spline models by varying both the number and placement of knots.

Broadly, these tools offer us a more complete picture of the variability among competing fits, as well as a scheme for combining them to improve predictive power. In a Bayesian parlance, we use this framework to address the *structural uncertainty* associated with the knot sequence. We begin this section with a general overview of Bayesian methods for linear models. Our treatment is necessarily limited and slanted toward the task of function estimation (i.e., the linear space in question should be thought of as an

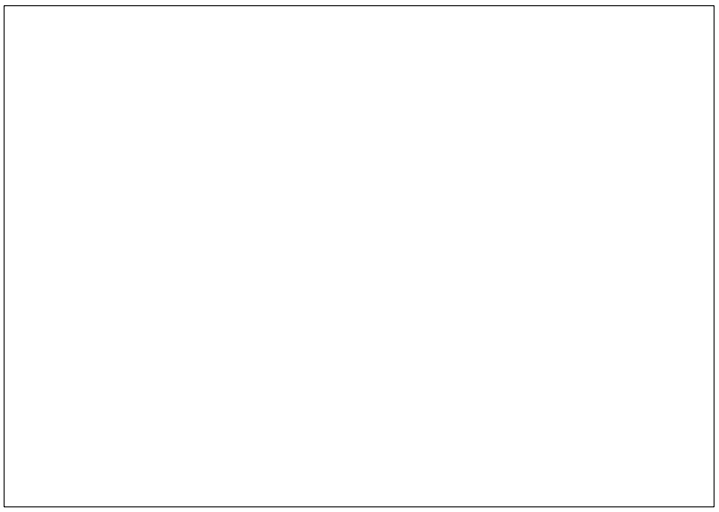

FIGURE 10.8. The single best fit based on AIC (solid), and the average of all fits that are within 1% of that AIC value (dashed) for the $^{87}\delta\,\mathrm{Sr}$ data.

approximation space). Readers wanting a more thorough presentation of Bayesian statistics are referred to O'Hagan (1994).

### *10.2.1 A single linear space*

In what follows, we let $\mathcal{D}$ denote a collection of observations from which we are to estimate an unknown function $f$, and we let $p(\mathcal{D}|f)$ denote the likelihood connecting the two. Let $\mathcal{G}$ be a $J$-dimensional linear space of functions, chosen because of its favorable approximation properties. Let $\{B_1, \ldots, B_J\}$ represent a basis for $\mathcal{G}$ so that we can write any function $g \in \mathcal{G}$ in the form

$$g(\boldsymbol{x}; \boldsymbol{\theta}) = \theta_1 B_1(\boldsymbol{x}) + \cdots \theta_J B_J(\boldsymbol{x})\,, \tag{10.2.1}$$

for some value of the coefficient vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$. As usual, we are interested in assessing various features of $f$ and introduce $\mathcal{G}$ as a device to explore them. The maximum likelihood estimate (MLE) of $f$ from $\mathcal{G}$ is defined as

$$\widehat{g} \;=\; \mathrm{argmax}_{g \in \mathcal{G}}\, p(\mathcal{D}|g)\,.$$

Because we are working with a linear model, we can write the likelihood in terms of the coefficient vector $\boldsymbol{\theta}$,

$$p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})\,, \tag{10.2.2}$$

where we have explicitly included the dependence on the space $\mathcal{G}$. Using this notation, the MLE can be expressed as

$$\widehat{g} = g(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}), \quad \text{where} \quad \widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G}).$$

Optimization methods like Newton-Raphson iterations are then applied to find $\widehat{\boldsymbol{\theta}}$. In all of the estimation contexts we have studied so far, $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})$ is a *concave function* of $\boldsymbol{\theta}$, greatly simplifying numerical schemes for finding $\widehat{\boldsymbol{\theta}}$ (see Chapter 2 for more on these topics).

In a Bayesian setup, we treat elements of $\mathcal{G}$ as random quantities. We define a *prior distribution* for functions $g \in \mathcal{G}$ that expresses our prior knowledge or ignorance about aspects of $f$. For example, given our understanding of the physical process that generated the data, we might place less weight on wiggly functions, and instead favor smooth representations. Using the fact that $\mathcal{G}$ is a linear model, such prior assignments are best made in terms of the coefficient vector $\boldsymbol{\theta}$. In this case, we will let $p(\boldsymbol{\theta}|\mathcal{G})$ denote the prior distribution on $\boldsymbol{\theta}$. Given the observations $\mathcal{D}$, the prior distribution of $\boldsymbol{\theta}$ is then updated using Bayes' rule

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{G}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})\, p(\boldsymbol{\theta}|\mathcal{G})}{p(\mathcal{D}|\mathcal{G})} \qquad (10.2.3) \\ &\propto \text{likelihood}\,(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G}) \times \text{prior}\,(\boldsymbol{\theta}|\mathcal{G}), \end{aligned}$$

where the quantity on the right is referred to as the *posterior distribution*. In the denominator of (10.2.3) we find the so-called *marginal* or *mixture distribution* of the data,

$$p(\mathcal{D}|\mathcal{G}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})\, p(\boldsymbol{\theta}|\mathcal{G})\, d\boldsymbol{\theta}.$$

Notice that as with (10.2.2), we have made explicit the dependence of $\boldsymbol{\theta}$ on the space $\mathcal{G}$. In the next section, we will consider several approximation spaces, and this notation will be important.

In a classical Bayesian linear model, the posterior distribution (10.2.3) incorporates all the information available in the data about $f$. In our setting, however, this is only really true if we assume that $f$ is a member of the approximating space $\mathcal{G}$. In deriving our prior distribution, we have implicitly assigned zero mass to functions outside of $\mathcal{G}$, and hence the posterior also has support on $\mathcal{G}$. Such an assumption is clearly difficult to justify, and as a result our inferences are limited by the properties of our chosen approximation space. Recall that we encountered similar difficulties with the likelihood-based approaches when we examined classical parametric confidence intervals (see Chapter 3). Assuming that $\mathcal{G}$ is sufficiently flexible to capture the major structures in $f$, we will overlook this technicality for the moment and instead explore what the Bayesian formalism offers us.

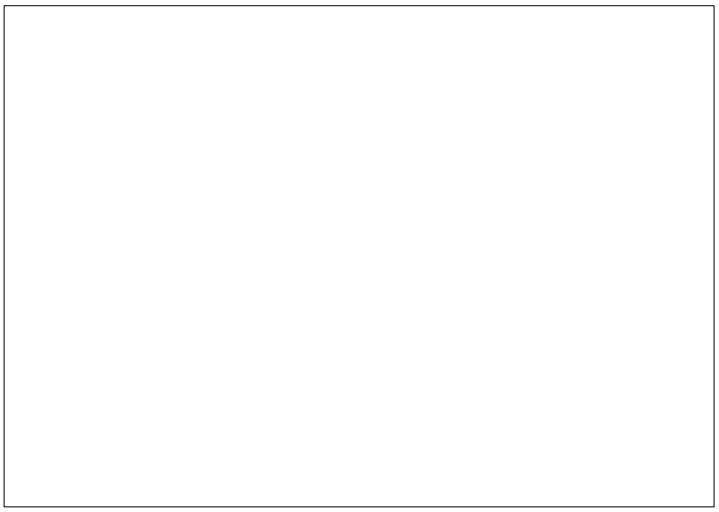

FIGURE 10.9. Assessing variability from the posterior using a single linear model $\mathcal{G}$. Fifty samples from the posterior are generated via (10.2.4).

### Sampling from the posterior

Using the expansion (10.2.1), the posterior for $\boldsymbol{\theta}$ (10.2.3) gives rise to a posterior distribution for functions in $\mathcal{G}$. By drawing realizations from this distribution, we can see at least informally the kinds of curves that receive the most support from the data. The simple recipe for generating functions from the posterior is given by

$$g(\boldsymbol{x};\boldsymbol{\theta}^*)\,, \qquad \boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\mathcal{D})\,. \tag{10.2.4}$$

By sampling several values of $\boldsymbol{\theta}^*$ and plotting the associated curves, we are able to visually assess the variability captured by the posterior distribution. In Figure 10.12, we present the $^{87}\delta\,\mathrm{Sr}$ data and fifty curves generated from the posterior associated with the best-fitting unimodal model. The exact specification of the priors is given in the next section. The plot is reminiscent of one we obtained in Chapter 3 using bootstrap samples. The use of this technique in more general smoothing contexts and its relation to the bootstrap are discussed in Hastie and Tibshirani (1990).

This sampling technique can also be used to help us draw inference about possibly complicated *functionals* of $f$. For example, the slope of the $^{87}\delta\,\mathrm{Sr}$ curve at the KTB provides earth scientists with information about the events that accompanied the mass extinction 66 million years ago. Using the same setup leading to Figure 10.9, we sampled 1000 curves and numerically evaluated the average slope in a 200,000 year window just after the KTB. A histogram of this distribution is given in Figure 10.10. The mean slope is 0.26 with a standard deviation of 0.01. Other quantities of interest might

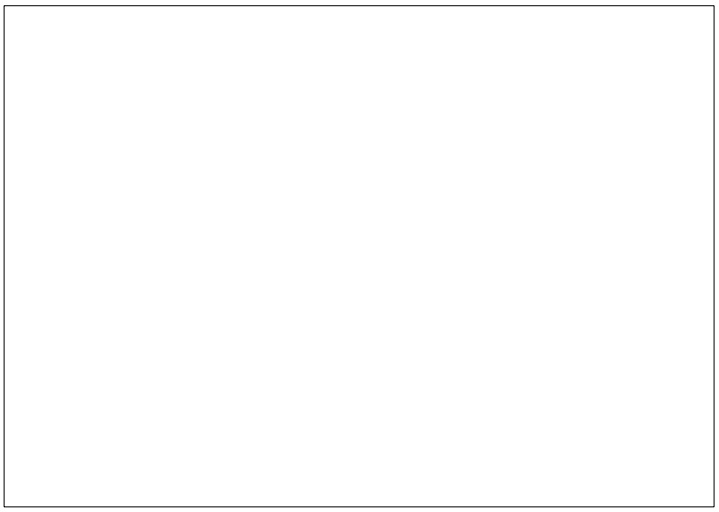

FIGURE 10.10. Histogram of the average slope in a 200,000 year window just after KTB for the 1000 curves displayed in 10.9.

include the number of peaks exhibited by a curve or its maximum in some region. In general, we can approximate the posterior distribution of each of these functionals via sampling. Through simple diagnostics like the plot in Figure 10.10, we obtain a picture of the features supported by the posterior. See Silverman (1985) for more examples of this technique.

**Estimation**

Next, we want to form an estimate of $f$. Following the approach given above, a reasonable choice is the posterior mean,

$$\bar{g}(\boldsymbol{x}) = g(\boldsymbol{x}; \bar{\boldsymbol{\theta}}) \,, \qquad \text{where} \qquad \bar{\boldsymbol{\theta}} \;=\; E(\boldsymbol{\theta} | \mathcal{D}, \mathcal{G}) \,, \tag{10.2.5}$$

and the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$. By a standard argument, we can show that this choice is optimal in a mean squared sense. To be more precise, let $\boldsymbol{x}$ be any point in $I$ and consider the squared loss

$$L(g, h) = [\, g(\boldsymbol{x}) - h(\boldsymbol{x}) \,]^2 \,,$$

where $g \in \mathcal{G}$ and $h$ is any other function defined on $I$. We then consider the expected loss

$$E\left( L(g, h) \,\middle|\, \mathcal{D}, \mathcal{G} \right) \tag{10.2.6}$$

where the expectation is taken with respect to the posterior distribution over $g(\boldsymbol{x})$. As was done above, we obtain the posterior distribution of $g(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\theta})$ from the posterior on $\boldsymbol{\theta}$ (10.2.3). Then, a minimizer of (10.2.6)

over all functions $h$ is $\bar{g}(\boldsymbol{x})$. [2] In the next section, we will see that for a certain assignment of priors $p(\boldsymbol{\theta}|\mathcal{G})$, the posterior mean (10.2.5) is a familiar smoothing spline fit constructed in Chapter 3.

### *10.2.2 Many spaces*

As in the previous section, let $\mathcal{D}$ denote a collection of observations from which we are to estimate an unknown function $f$, but now let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ represent a series of linear spaces, differing perhaps in their ability to resolve features of $f$. To make use of these spaces, we need to construct a prior for the collection $\mathcal{G} = \cup_m \mathcal{G}_m$. Repeating the recipe above, we construct a basis for each space

$$g_m(\boldsymbol{x}; \boldsymbol{\theta}_m) = \theta_{m1} B_{m1}(\boldsymbol{x}) + \cdots + \theta_{mJ_m} B_{mJ_m}(\boldsymbol{x}) \,,$$

where $\boldsymbol{\theta}_m = (\theta_{m1}, \ldots, \theta_{mJ_m})$; and assign a prior distribution to functions in $\mathcal{G}_m$ through the coefficient vector $\boldsymbol{\theta}_m$, which we denote $p(\boldsymbol{\theta}_m|\mathcal{G}_m)$. Next, we associate with $\mathcal{G}_m$ a prior probability $p(m)$ that reflects our knowledge about the function we are estimating. For example, suppose $\mathcal{G}_m$ is a space of natural splines with $m$ knots distributed in some way over the domain of $f$. If we expect $f$ to exhibit only very smooth behavior, our prior probabilities should decrease with $m$. Having constructed $p(m)$ and $p(\boldsymbol{\theta}_m|\mathcal{G}_m)$, we simulate curves from our prior in two steps given below. This scheme will produce functions from the collection $\mathcal{G} = \cup_m \mathcal{G}_m$.

Given the likelihood $p(\mathcal{D}|\boldsymbol{\theta}_m, \mathcal{G}_m)$, we can compute the marginal distribution of the data $\mathcal{D}$ given the linear space $\mathcal{G}_m$,

$$p(\mathcal{D}|\mathcal{G}_m) = \int p(\mathcal{D}|\boldsymbol{\theta}_m, \mathcal{G}_m)\, p(\boldsymbol{\theta}_m|\mathcal{G}_m)\, d\boldsymbol{\theta}_m \,.$$

Using the prior $p(m)$ to combine contributions from each space, the data $\mathcal{D}$ now have a mixture distribution

$$p(\mathcal{D}) = \sum_m p(\mathcal{D}|\mathcal{G}_m) p(m) \,.$$

---

[2] Many sensible loss functions produce the same result. For example, we could define

$$L(g, h) \;=\; \int_I [\, g(\boldsymbol{x}) - h(\boldsymbol{x}) \,]^2 \, d\boldsymbol{x} \,.$$

In the regression context, where the data $\mathcal{D}$ consist of the pairs $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, we could alternately define the loss function as

$$L(g, h) \;=\; \frac{1}{n} \sum_{i=1}^{n} [\, g(\boldsymbol{X}_i) - h(\boldsymbol{X}_i) \,]^2 \;.$$

If $\mathcal{G}$ is identifiable with respect to the input values $\{\boldsymbol{X}_i\}$, then the minimizer of (10.2.6) is still $\bar{g}$.

Simulating from the prior

Prior 1: First, we select a model by sampling an index $m^*$ from the distribution $p(m)$;

Prior 2: then, given $m^*$, we sample a coefficient vector $\boldsymbol{\theta}^*$ from the prior $p(\boldsymbol{\theta}_{m^*}|\mathcal{G}_{m^*})$ and form $g_{m^*}(\boldsymbol{x};\boldsymbol{\theta}^*)$.

Simulating from the posterior

Post 1: First, we select a model by sampling an index $m^*$ from the posterior distribution $p(m|\mathcal{D})$;

Post 2: then, given $m^*$, we sample a coefficient vector $\boldsymbol{\theta}^*$ from the posterior computed as in (10.2.3) for $\mathcal{G}_{m^*}$,

$$p(\boldsymbol{\theta}_{m^*}|\mathcal{D},\mathcal{G}_{m^*}) = p(\mathcal{D}|\boldsymbol{\theta}_{m^*},\mathcal{G}_{m^*})\, p(\boldsymbol{\theta}_{m^*}|\mathcal{G}) \,/\, p(\mathcal{D}|\mathcal{G}_m)\,,$$

and form $g_{m^*}(\boldsymbol{x};\boldsymbol{\theta}^*)$.

FIGURE 10.11. Two-stage or hierarchical model specification. Sampling from the prior and the posterior.

Given data $\mathcal{D}$, we apply Bayes' rule to update the probabilities $p(m)$ on each model space

$$p(m|\mathcal{D}) \;=\; \frac{p(\mathcal{D}|\mathcal{G}_m)\, p(m)}{p(\mathcal{D})}\,. \tag{10.2.7}$$

Models receiving high posterior probability $p(m|\mathcal{D})$ are favored by the data. In terms of the actual function spaces, the posterior distribution for elements in $\mathcal{G} = \cup_m \mathcal{G}_m$ is again specified in two steps. We now consider how introducing the different models changes our approach to Bayesian inference and estimation.

**Posterior inference**

As we mentioned above, the incorporation of several model spaces allows us to assess uncertainty in both the number and location of knots. In Figure 10.10, we present samples from the posterior obtained from all possible models. The prior is described in detail in the next section. For comparison purposes, the prior conditional on the best-AIC model from Figure 10.12 was the same. The samples here exhibit different variability across the interval. Similarly, the overall spread of these functions is wider, reflecting in

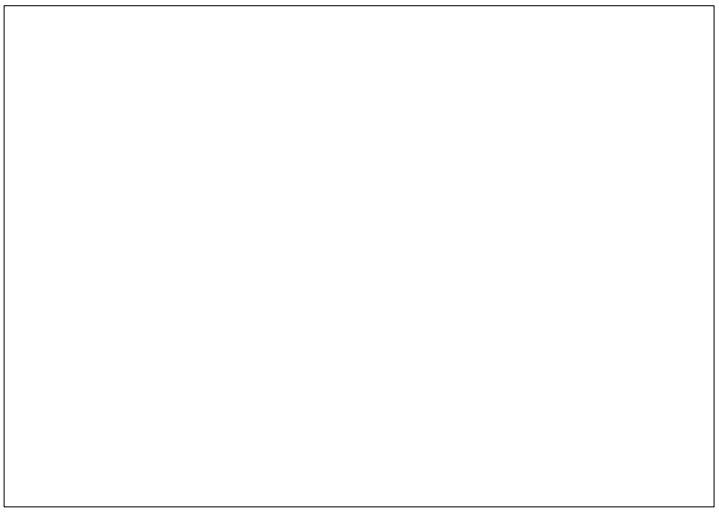

FIGURE 10.12. Assessing variability from the posterior using models with between 0 and 43 knots. Fifty samples from the posterior are generated via the prescription in Figure 10.11.

some sense the uncertainty coming from the knot sequence. The curves in Figure 10.2.3 do not incorporate this component of variability and hence tend to be overly optimistic.

We can make this idea precise by considering again estimates of a functional of $f$. In the previous section we considered the average slope in some neighborhood of the KTB, and other candidates might include the maximum of the curve in some region or the number of modes of $f$. Let $\psi(\cdot)$ denote any such functional. Given a linear space $\mathcal{G}_m$, we used samples $\theta_m^*$ from the posterior $p(\boldsymbol{\theta}_m|\mathcal{D}, \mathcal{G}_m)$ to generate random functions $g_m^*(\boldsymbol{x}) = g_m(\boldsymbol{x}; \boldsymbol{\theta}_m^*)$. We in turn used the distribution of $\psi(g_m^*)$ to draw conclusions about $\psi(f)$. Suppose that for each $m$, the variables $\psi(g_m^*)$ have mean $\mu_m$ and variance $\sigma_m^2$. Next, we remove the conditioning on the linear space $\mathcal{G}_m$ and consider the distribution of $\psi(g^*)$ where $g^*$ is constructed according to the posterior over $\mathcal{G} = \cup_m \mathcal{G}_m$ given in Figure 10.11.

Let $\pi_m = p(m|\mathcal{D})$ denote the posterior probability of the $m$th model. The mean of $\psi(g^*)$ is then given by

$$\mu = \sum_m \pi_m \mu_m$$

while the variance is given by

$$\begin{aligned}\sigma^2 &= \sum_m \pi_m \sigma_m^2 + \sum_m \pi_m (\mu_m - \mu)^2 \\ &= \begin{pmatrix} \text{Within} \\ \text{Model} \\ \text{Variance} \end{pmatrix} + \begin{pmatrix} \text{Between} \\ \text{Model} \\ \text{Variance} \end{pmatrix}.\end{aligned}$$

This final expression makes explicit the effect of including different models when drawing inferences about functionals of $f$. We can decompose the posterior variation in estimating $\psi(f)$ into two components. The first represents the average variation resulting from any particular choice of model, while the second records the variation arising purely from uncertainty in our choice of model.

**Estimation**

There are two approaches to estimating $f$ given the posterior distribution described in Figure 10.11. The first mimics the technique we employed to estimate functionals of $f$, namely via a posterior mean. Theoretical considerations indicate that this kind of averaging across spaces can offer improvements in predictive performance (Barron, Shervish, and Wasserman 1999; X. and L. 1998; Huang 2001). The downside, however, is that we are forced to consider many knot sequences to produce an estimate. A single, low-dimensional spline space is appealing for many applications. Therefore, the second estimation scheme chooses one model according to the posterior probabilities (10.2.7).

*Model averaging.* Applying the recipe above for functionals of $f$, we can form an estimate of $f$ via the pointwise average

$$\bar{g}(\boldsymbol{x}) = \sum_m \bar{g}_m(\boldsymbol{x}) \pi_m \,, \tag{10.2.8}$$

where $\pi_m = p(m|\mathcal{D})$ is the posterior probability (10.2.7), and $\bar{g}_m$ is the posterior mean (10.2.5) computed for the space $\mathcal{G}_m$. This estimate is also optimal in a mean squared sense applying the same loss functions and reasoning as in the single-model case presented above. In this case, we minimize the expected loss with respect to the mixture posterior (10.2.7). The estimate in expression (10.2.8) is often said to be the result of Bayesian model averaging. In the spline context, this approach dates back to an early paper by Halpern (1973). Unfortunately, given the restricted computational tools of his day, Halpern's estimator did not find substantial practical application for nearly two decades.

*Maximum a Posteriori Estimate.* Rather than average together a large number of models, we might instead choose to select a single model. Using

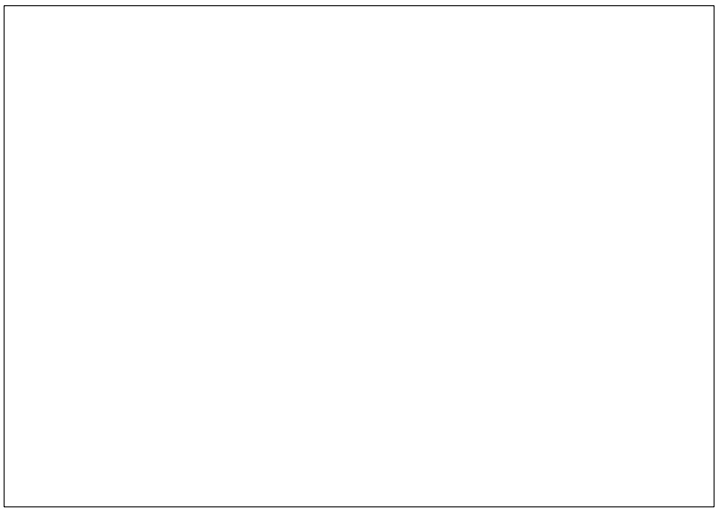

FIGURE 10.13. Whatever fig:resbma is.

the posterior model probabilities, we might consider

$$\widehat{m} = \text{argmax}_m p(m|\mathcal{D})$$

and then take as our estimate $\bar{g}_{\widehat{m}}(\boldsymbol{x})$. It can be shown that this estimate serves to minimize a different loss function, the so-called 0-1 loss that charges the value 0 if we get the model right, and 1 otherwise. The convenient thing about this estimate is that it involves only one model. Therefore, just like our greedy scheme, we are afforded some degree of simplicity in the representation of our estimate. We only require a singe space, rather than the (potentially) large suite of candidates.

### *10.2.3 Computation*

With the benefit of a complete search, we can see that several

When the number of candidate models is reasonable, the posterior-based methods mentioned above are sensible. For example, if we took the spaces $\mathcal{G}_m$ to be the cubic natural spline space with knots at $m$ equally spaced quantiles, we could entertain at most $n$ models. In Figure 10.13 we reproduce the simulation from Chapter 3 with 200 equally-spaced data points. We have added the curve posterior mean curve using the prior discussed in the next section. We have also plotted the equivalent kernel. The fit behaves like a fixed-bandwidth kernel estimate. To improve matters, we would prefer to include knots sequences that concentrate breakpoints in regions where they are needed. So, given $m$ knots, we might have many more sequences. How doe we compute with such a large number?

**Exhaustive search**

Given a search procedure like regression by leaps-and-bounds, we can sift through model space and keep only promising models. A variant of this idea, known as Occam's window, was suggested by Madigan and Raftery (1994). Broadly, these ideas make use of some kind of deterministic search through model space, reducing options based on the residual sum of squares or other heuristics associated with the fit. This is our approach to the $^{87}\delta\,\mathrm{Sr}$ data.

**Markov chain Monte Carlo**

Suppose we have a finite number of candidate models, denoted $\mathcal{G}_1, \ldots, \mathcal{G}_M$, for which we can compute the posterior model probabilities $p(m|\mathcal{D})$. The idea behind Markov chain Monte Carlo (McMC) is to construct a Markov chain $m(t)$, $t = 1, 2, \ldots$, on the indices $1, \ldots, M$ having the equilibrium distribution $p(m|\mathcal{D})$. Let $m(1), \ldots, m(N)$ be $N$ observations from this chain, and suppose we want to estimate a functional of $f$. For simplicity, we will consider the value of $f$ at the point $\boldsymbol{x}$. Then, by standard McMC results the average

$$\widehat{g}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \bar{g}_{m(i)}(\boldsymbol{x})\,.$$

tends to

$$\widehat{g}(\boldsymbol{x}) \to E\,(\,g(\boldsymbol{x})|\mathcal{D}\,)$$

almost surely as $N$ tends to infinity. Again, the expectation above is with respect to the posterior distribution. This technology has been developed mainly in the context of model averaging and through the connection with linear spaces, has found application in spline modeling. We describe two basic approaches that have appeared in the literature.

*Metropolis-Hastings.* The simplest method for constructing a Markov chain on model indices is via the Metropolis-Hastings algorithm. Here, we define a neighborhood of models $\mathrm{nbhd}(\mathcal{G}_m)$. We then construct a transition matrix $Q(i, j) = 0$ if $i \notin \mathrm{nbhd}(\mathcal{G}_j)$.

$$\min\left\{1, \frac{p(m'|\mathcal{D})}{p(m|\mathcal{D})}\right\}$$

> Finally! The simple Met-Hastings approach will allow us to talk about simulated annealing, etc. The disc starting from this point will appear shortly.

*Gibbs sampler.* When the number of candidate knots is fixed (say given by the input data points), we can introduce a binary vector $\gamma$ as a replacement for the index $m$. Here, each element of $\gamma$ corresponds to a knot location and is one if the knot is in the model and zero otherwise. The neighborhood of a

given model with index $\gamma$ is defined as those which differ by the placement, the presence or absence of a single knot.

The kernel obtained by conditioning on $Y$ produces $\gamma^{j+1}$ by sampling from the conditional distribution $\gamma|Y$. As a result, for any $y$, the conditional distribution of $\gamma|y$ is an invariant distribution for this kernel.

George and McCulloch (1993) used the Gibbs sampler to identify promising subsets of variables in linear regression models. Smith and Kohn (1996) proposed an alternative Gibbs sampler to that of George and McCulloch (1993) to generate iterates from $p(\gamma|\boldsymbol{Y})$. They generate the $\gamma_i$ one at a time from their conditional densities $p(\gamma_i|\boldsymbol{Y}, \gamma_{j\neq i})$, with both $\boldsymbol{\beta}$ and $\sigma^2$ integrated out to speed up convergence. A computationally efficient algorithm for implementing this is outlined in Smith and Kohn (1996).

*Reversible jump Markov chain Monte Carlo.* In many cases, we would prefer to treat the location of knots as continuous. This so-called free-knot spline approach means that we cannot discretize the space and generate a fixed set of alternatives. Instead, our sampler must draw locations as well.

This framework is quite general and in fact each of the previous McMC schemes can be interpreted in this way. This sampling scheme was introduced by Green (1995) and used by Denison, Mallick, and Smith (1998) and Holmes and K. (1998) for nonparametric regression. The sampler assumes that the number of active variables (which is $m_\gamma$, the number of indicator variables equal to 1) has a Poisson distribution with mean $\tau$ which we denote as $p_\tau(m_\gamma)$. The sampler is based on the repeated application of one of the following three elementary steps: (a) A *birth* in which a randomly chosen $\gamma_i$ which is currently 0 is turned into a 1; (b) A *death* in which a randomly chosen $\gamma_i$ which is currently 1 is turned into a 0 ; (c) A *move* in which a randomly chosen $\gamma_i$ which is currently 1 is turned into a 0 and simultaneously a randomly chosen $\gamma_j$ which is currently 0 is turned into a 1. If $m_\gamma = k$ currently, then the probability of a birth is $b_k = 0.4 \min\{1, p_\tau(k+1)/p_\tau(k)\}$, the probability of a death is $d_k = 0.4 \min\{1, p_\tau(k)/p_\tau(k+1)\}$ and the probability of a move is $1 - b_k - d_k$. Each of the birth, death and move steps is then either accepted or rejected based on Metropolis-Hastings step. Details are given by Green (1995).

The priors in Denison et al. (1998) and Holmes and K. (1998) are different than those in Section 10.2. First, as noted above, these authors follow Green (1995) and use a Poisson prior with mean $\tau$ for $m_\gamma$. Denison et al. (1998) assume a given value for $\tau$, whereas Holmes and K. (1998) place a gamma prior on $\tau$. Both papers assume an inverse gamma prior on $\sigma^2$, but do not put a prior on the regression coefficients. At each iteration, both papers estimate the regression coefficients of the active variables by least squares and use these estimates to generate $\gamma$ and $\sigma^2$. Strictly speaking, neither Denison et al. (1998) nor Holmes and K. (1998) provide genuine Markov chain Monte Carlo sampling schemes which would require generating the regression coefficients from their conditional distributions. One practical

consequence of not generating the regression coefficients in their sampling scheme is that the confidence intervals for the regression surface may be too narrow.

### *10.2.4 Connection with model selection criteria*

As we have seen in the Smith and Kohn (1996) procedure, it is possible to construct a prior so that the posterior agrees with a selection criterion like our generalized AIC. In fact, the Bayesian information criterion BIC ($\alpha = \log n$), is obtained directly by expanding the posterior distribution (10.3.7) in a quadratic Taylor's series around the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$. Therefore, under $\mathcal{G}_m$ we have

$$p(\boldsymbol{\beta}|\mathcal{G}_m, \mathcal{D}) \approx (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})p(\widehat{\boldsymbol{\beta}}|\mathcal{G}_m, \mathcal{D}) + p'(\widehat{\boldsymbol{\beta}}|\mathcal{G}_m, \mathcal{D}) - \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})I(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}),$$

where $p$ is simply the posterior and $I$ is the Fisher information. Because $\widehat{\boldsymbol{\beta}}$ is the MLE, the linear term drops from this expression leaving

$$p(\boldsymbol{\beta}|\mathcal{G}_m, \mathcal{D}) \approx (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})p(\widehat{\boldsymbol{\beta}}|\mathcal{G}_m, \mathcal{D}) - \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})I(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\,.$$

Therefore, we see that the posterior is essentially given by a Gaussian distribution with mean $\widehat{\boldsymbol{\beta}}$ and variance-covariance $I(\widehat{\boldsymbol{\beta}})^{-1}$. This is commonly called a Laplace expansion of the posterior.

Schwarz (1978) derives BIC as an approximate means of comparing models. The nice thing about this expansion is that it effectively eliminates the prior specification. Through judicious choice of prior, it is possible to construct a posterior that scales like any member of the AIC family. For regression, George and Foster (2000) make use of this. For generalized regression Clyde (2000) introduces a particular prior that gives rise to the same posterior. In each case, a tuning parameter on the prior for $\boldsymbol{\beta}$ is equivalent to setting $\alpha$. For this reason, Clyde (2000) calls this class of priors the Calibrated Information Criterion priors. They are part of a class of objective Bayesian procedures that are mean to be automatic.

### *10.2.5 Theoretical justification*

## 10.3 Normal linear regression

### *10.3.1 Prior specification*

Our treatment of Bayesian approaches to the normal linear regression model is necessarily brief. Readers requiring more motivation are referred to the text by O'Hagan (1994). The reader in need of more motivation than is given in these pages is referred to this text. In the normal linear regression

model, our data $\mathcal{D}$ consists of $n$ independent pairs $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ observed from the model

$$Y = f(\boldsymbol{X}) + \epsilon\,, \tag{10.3.1}$$

where $\epsilon$ has a Gaussian distribution with mean zero and variance $\sigma^2$. For simplicity, we will assume that $\sigma^2$ is known. Then, writing the likelihood in terms of $\boldsymbol{\beta}$ we find

$$p(\mathcal{D}|\boldsymbol{\beta}, \mathcal{G}) \;=\; \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}\left(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\beta})\right)^2}{2\sigma^2}\right]. \tag{10.3.2}$$

Define the matrix $\mathbf{B}$ to be the design matrix

$$[\,\mathbf{B}\,]_{i,j} \;=\; B_i(\boldsymbol{X}_j) \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, J\,,$$

corresponding to the basis in (10.2.1), and set $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. With this notation, the likelihood (10.3.2) becomes

$$p(\mathcal{D}|\boldsymbol{\beta}, \mathcal{G}) \;=\; \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})}{2\sigma^2}\right].$$

Motivated mainly by computational convenience, the prior on $\boldsymbol{\beta}$ is usually taken to be normal with some mean mean $\boldsymbol{b}$ and variance-covariance $\Sigma$.[3] With this choice, we use Bayes' rule (10.2.3) to derive the posterior distribution

$$p(\boldsymbol{\beta}|\mathcal{D}, \mathcal{G}) \;\propto\; \exp\left[-\frac{(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})}{2\sigma^2} - \frac{(\boldsymbol{\beta} - \boldsymbol{b})'\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{b})}{2}\right] \tag{10.3.3}$$

where multiplicative constant does not depend on $\boldsymbol{\beta}$. Then, rewriting the expression within brackets as a quadratic function of $\boldsymbol{\beta}$, it is easy to verify that the posterior distribution of $\boldsymbol{\beta}$ given $\mathcal{D}$ is normal but with mean and variance

$$\left(\Sigma^{-1} + \mathbf{B}'\mathbf{B}\right)^{-1}\left(\Sigma^{-1}\boldsymbol{b} + \mathbf{B}'\mathbf{Y}\right) \quad \text{and} \quad \left(\Sigma^{-1} + \mathbf{B}'\mathbf{B}\right)^{-1}, \tag{10.3.4}$$

respectively. From the quantity on the left, we can see immediately that if $\boldsymbol{b}$ is a vector of zeros, then the posterior mean has the form of a generalized *ridge regression*

$$\overline{\boldsymbol{\beta}} \;=\; \left(\Sigma^{-1} + \mathbf{B}'\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{Y}\,. \tag{10.3.5}$$

and we take as our estimate of $f$ the function $g(\boldsymbol{x}; \overline{\boldsymbol{\beta}})$.

[3] This choice is known as the *conjugate prior* because the both the prior and the posterior distributions belong to the same parametric family. In this case, both are Gaussian.

**Smoothing splines**

In Chapter 3, we found that the cubic smoothing spline produced an estimate the form (10.3.5). To make the connection with Bayesian methods, we recall some notation. First, we will focus on univariate smoothing splines based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in [0, 1]$. Let $\mathcal{G}$ be the space of natural splines corresponding to some knot sequence and let $B_1, \ldots, B_J$ be some basis for the space. A smoothing spline is a penalized least squares fit into $\mathcal{G}$, where the penalty is given by a *roughness measure* $S(g)$ for functions $g \in \mathcal{G}$. While $S$ is a property of the function $g$, given a basis and the expansion (10.2.1) we can express it in terms of the coefficient vector $\boldsymbol{\beta}$:

$$S(g) = \int (g'')^2 = \boldsymbol{\beta}' \mathbf{A} \boldsymbol{\beta},$$

where

$$[\,\mathbf{A}\,]_{ij} = \int g_i''(x) g_j''(x) dx \qquad \text{for } 1 \le i, j \le J.$$

Because $S$ has the property that any function of the form $\beta_0 + \beta_1 x$ is assigned roughness zero, the matrix $\mathbf{A}$ is only positive semidefinite.

With these definitions, the smoothing spline is given as the minimum of the penalized likelihood

$$\sum_{i=1}^{n} (Y_i - g(X_i))^2 + S(g)\,, \qquad g \in \mathcal{G}\,,$$

which we can rewrite this in terms of the coefficient vector $\boldsymbol{\beta}$,

$$\sum_{i=1}^{n} (\, Y_i - g(\boldsymbol{X}_i; \boldsymbol{\beta})\,)^2 + \lambda \boldsymbol{\beta}' \mathbf{A} \boldsymbol{\beta} \;=\; (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' \mathbf{A} \boldsymbol{\beta} \quad (10.3.6)$$

using the definitions of $\mathbf{Y}$ and $\mathbf{B}$ given in the previous section. Comparing this to (10.3.3), we see that the penalized likelihood is in fact (up to some additive constants that do not depend on $\boldsymbol{\beta}$) the (negative) log-posterior corresponding to the normal regression model (10.3.1) with a normal prior on $\boldsymbol{\beta}$. The noise variance $\sigma^2$ has been absorbed into the penalty parameter $\lambda$.

To obtain the smoothing spline criterion (10.3.6), the normal prior on $\boldsymbol{\beta}$ has mean zero and variance-covariance $\lambda^{-1}\mathbf{A}^{-1}$. Unfortunately, because $\mathbf{A}$ is derived from a roughness measure that gives each function of the form $\beta_0 + \beta_1 x$ the value zero, it is only positive semi-definite and hence not invertible. Therefore, the prior on $\boldsymbol{\beta}$ that produces a smoothing spline is *partially improper*: It concentrates mass on a hyperplane of dimension $J - 2$. From a Bayesian perspective, improper priors of this sort are acceptable providing the posteriors are well behaved. In this case, as long as the linear space spanned by $1, x$ is identifiable with respect to the input data $X_1, \ldots, X_n$,

the posterior is proper and normal with the mean and variance-covariance indicated in (10.3.4). When the identifiability condition does not hold, the matrix $\mathbf{A}+\mathbf{B}'\mathbf{B}$ is not invertible, the posterior is again improper, and as we found in Chapter 3, a single solution to the penalized least squares problem does not exist.

**Calibrated information criterion priors**

Many prior specifications have been suggested for the normal linear model, and several have been applied *as is* to smoothing via a linear space $\mathcal{G}$. One of the earliest modern treatments was considered by Smith (1996) and Smith and Kohn (1996). A normal prior with mean zero and variance-covariance $\lambda(\mathbf{X}'\mathbf{X})^{-1}$ is introduced. This means that the posterior is given by

$$BIC(\boldsymbol{t})$$

The idea behind the Smith–Kohn technique is to introduce a binary vector $\gamma = (\gamma_1, \ldots, \gamma_{M+4})$ that indexes the columns of the design matrix $X$ corresponding to the truncated power basis: $\gamma_i$ equals zero or one according as the coefficient $\beta_i$ of the $i$th basis function does or does not equal zero. The components of $\gamma$ are assumed to be a priori independent, with probability one-half of equaling zero. This corresponds to giving all possible subsets of the set of $M + 4$ variables the same prior probability. After also specifying prior distributions for $\beta = (\beta_1, \ldots, \beta_{M+4})|(\gamma, \sigma^2)$ and $\sigma^2|\gamma$, Smith and Kohn derive the posterior distribution of $\gamma$ given the vector of $n$ observations $y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$. Specifically, if we let $\beta_\gamma$ and $X_\gamma$ denote the coefficient vector and design matrix, respectively, corresponding to a model containing exactly those variables for which $\gamma_i$ equals one, then by setting

$$p(\beta_\gamma|\gamma, \sigma^2) = N(0, c(X_\gamma^T X_\gamma)^{-1}) \quad \text{and} \quad p(\sigma^2|\gamma) \sim 1/\sigma^2$$

we find that the posterior probability function of $\gamma$ is given by

$$p(\gamma|y) \sim (1+c)^{-q_\gamma/2}\Big(y^T y - \frac{c}{c+1} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y\Big)^{-n/2},$$

where $q_\gamma = \sum_i \gamma_i$ is the number of terms in the model and $c$ is a user-specified constant. Smith and Kohn apply the Gibbs sampler to simulate from the posterior distribution of $\gamma$ and either report the posterior mode of $\gamma$ or the posterior mean of $\beta$. The model has been specified so that the sampling procedure steps through many models with high posterior probability in a computationally efficient manner. After $K$ Gibbs iterations, two alternative approaches are applied to the samples $\gamma^{[k]}$, $k = 1, \ldots K$, to estimate $\beta$: (i) $\hat{\beta}$ is obtained by an OLS fit to those variables included in the model specified by the vector $\gamma^{[k]}$ that maximizes $p(\gamma^{[k]}|y)$; or (ii) the posterior mean $E(\beta|y)$ is estimated by the average value of $E(\beta|y, \gamma^{[k]})$,

where the indicated conditional expectation is computed exactly using the fact that $\beta|(y,\gamma)$ has a multivariate $t$ distribution.

Transforming the expression for the posterior probability function of $\gamma$ into something more familiar to the smoothing community, Foster and George (1996) have found that under certain conditions the value of $\gamma$ that maximizes this function also minimizes the quantity $\mathrm{RSS}(\gamma) + (1 + c^{-1}) \log(c+1) q_\gamma \hat{\sigma}^2$, where $\mathrm{RSS}(\gamma)$ is the residual sum of squares for the model specified by $\gamma$, and $\hat{\sigma}^2$ is estimated from the full model with all $M+4$ variables. By selecting $c$ properly, we can perform model selection with respect to Mallow's $C_p$, AIC, or BIC. Making this connection, we see that the Gibbs sampler of Smith and Kohn is in fact an alternative to our approach of minimizing BIC in a stepwise fashion. Unfortunately, because of the way the vector $\gamma$ treats all variables as candidates for inclusion or exclusion, we are left with the deficiencies described in the previous section.

The performance of this technique, however, is heavily dependent on the number of knots used to define the truncated power basis. An alternative approach followed by Denison, Mallick, and Smith (1998) involves defining prior distributions for the number and location of knots as well as the coefficients in a spline expansion. The resulting "automatic Bayesian curve fitting" procedure makes use of reversible jump Markov chain Monte Carlo methods (Green 1995) to compute the posterior distribution, this time over collections of models having different numbers and positions of knots. At each step in their sampling procedure, one of several possible transitions is chosen at random. These transitions include adding a new knot and either moving or deleting an existing knot, and they can in principle be made efficient through the use of Rao and Wald statistics.

### *10.3.2 Computation*

As in the previous section, let $\mathcal{D}$ denote a collection of observations from which we are to estimate an unknown function $f$, and let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ represent a series of linear spaces, differing perhaps in their ability to resolve features of $f$. Each space is assigned a prior probability $p(\mathcal{G}_m)$ that is consistent with our assumptions about the function we are estimating. For example, suppose $\mathcal{G}_m$ is a space of natural splines with $m$ knots distributed in some way over the domain of $f$. Then, if we expect $f$ to exhibit only very smooth behavior, our prior probabilities should decrease with $m$. If we let $p(\mathcal{D}|\mathcal{G}_m)$ denote the marginal probability of the data given the linear space $\mathcal{G}_m$, the data $\mathcal{D}$ have a mixture distribution

$$p(\mathcal{D}) = \sum_m p(\mathcal{D}|\mathcal{G}_m) p(\mathcal{G}_m)\,.$$

As we did in the case of a single linear model, the prior distribution on the collection of linear spaces is updated using Bayes' rule

$$p(\mathcal{G}_m|\mathcal{D}) \;=\; \frac{p(\mathcal{D}|\mathcal{G}_m)\,p(\mathcal{G}_m)}{p(\mathcal{D})}\,, \tag{10.3.7}$$

the quantity on the right being referred to as the posterior distribution for $\mathcal{G}_m$. Inference about $f$ is now made on the basis of (10.3.7).

## 10.4 Extended linear models

### *10.4.1 Logspline density estimation*

In Figure 10.14, we compare the greedy scheme with several versions of stochastic techniques. Here, we have applied a prior that controls the smoothness of the data. It is still possible to form a type of calibration with the model selection criterion and this was done in Hansen and Kooperberg (2000). What we find from this example is that there is a price to be paid for the more exhaustive search schemes. The best results were obtained with a penalty $\alpha = \log n$.

### *10.4.2 Triogram regression*

A major difficulty with the Triogram procedure was that the chain of models visited during addition tended to be very similar to those seen on the way down. This is because the placement of knots tends to be rather highly constrained: Splitting edges and adding vertices produce triangulations from which it is difficult to remove structures other than those that were added. In short, the process is much more constrained and the number of candidate models examined during the greedy search is even more limited than in the univariate examples given above.

In the previous section, a greedy search was employed to identify a single spline space $G$. From among the functions in $G$ we then chose an estimate $\widehat{g}$ via maximum likelihood. By restricting ourselves to the class of (concave) extended linear models, we were naturally led to stepwise procedures involving nested spaces. We now present a general Bayesian framework for model adaptation that will allow us to compare this greedy procedure with other more recent approaches to the problem. In this Bayesian setup, *model uncertainty* comes from both the structural aspects of the space $G$ (knot or vertex placement) as well as from our selection of members $g \in G$.

### *10.4.3 ELM Prior specification*

We begin with a simple hierarchical formulation. At the first level, we assign a prior distribution $p(G)$ to some set of candidate models $G$. In the parlance

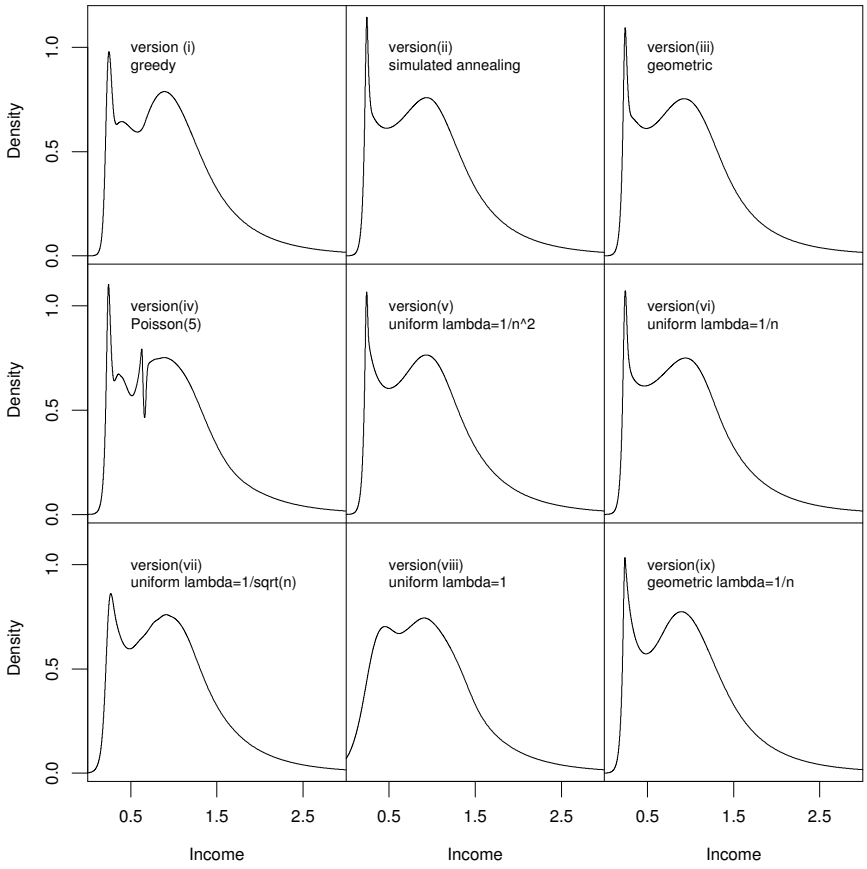


FIGURE 10.14. Several versions of Logspline density estimation for the income data.

of the previous section, those $G$ for which $p(G) > 0$ represent the collection of allowable spaces. For example, through $p(G)$ we can enforce properties like the minimum spacing between knots or the largest acceptable aspect ratio of triangles in a mesh. Next, given a space $G$, we generate elements $g$ (univariate or multivariate splines) according to the distribution $p(g|G)$. By selecting a basis for $G$ consisting of the functions $g_1, \ldots, g_J$, a natural prior for $g$ involves the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ in the expansion

$$g = \beta_1 g_1 + \cdots + \beta_J g_J \,. \tag{10.4.1}$$

We will use a partially improper, normal distribution for $\boldsymbol{\beta}$ frequently encountered in smoothing spline applications (Silverman 1985; Wahba 1990; Green and Silverman 1994). We are led to this prior because it is independent of how we select our basis for the expansion (10.4.1). Finally, the data enter through the likelihood specified by an extended linear model $p(W|g)$

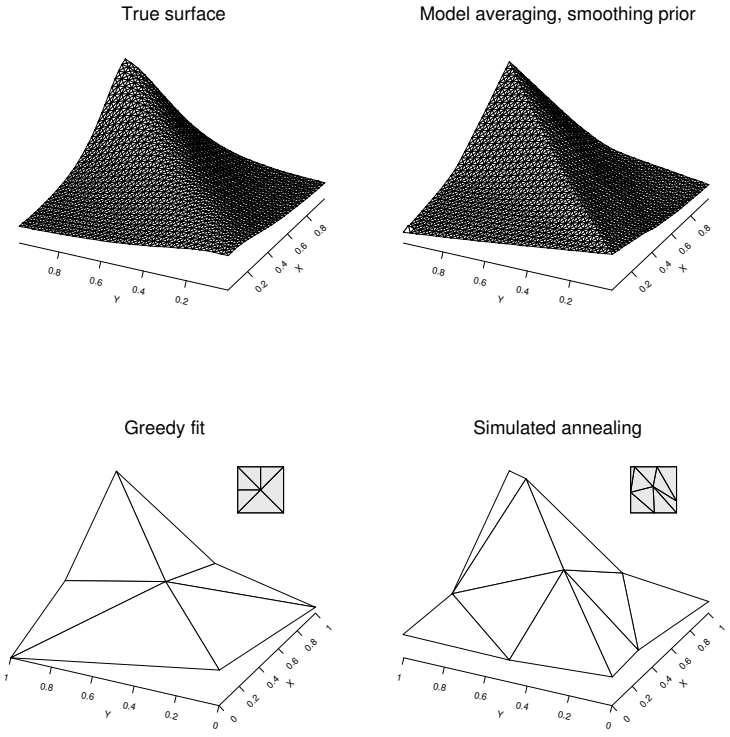


FIGURE 10.15. In the top row we have the true surface (left) and the fit resulting from model averaging (right). In the bottom row we have two isolated fits, each a "minimal" BIC model, the leftmost coming from a greedy search, and the rightmost produced by simulated annealing (the triangulations appear at the top of each panel).

defined in Chapter 4. The prior components of this general specification deserve further elaboration. Implementation issues associated with Bayesian versions of both Logspline and Triogram regression will be discussed at the end of this section.

### Priors on model space

In many applications, the most direct specification of $p(G)$ involves the collection of knots or vertices. For the moment, consider univariate spaces $G$ and let $\boldsymbol{\tau} = \{t_1, \ldots, t_K\}$ denote the set of breakpoints. Introducing another layer of conditioning, we first choose the number of knots $K$ (closely related to the dimension $J$ of $G$) according to $p(K)$, and then given $K$, we generate $\boldsymbol{\tau}$ from the distribution $p(\boldsymbol{\tau}|K)$. Regularity conditions on the

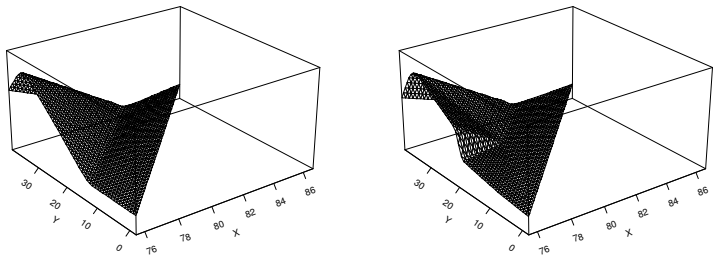


FIGURE 10.16. The volt data from Cleveland and Fuentes (1996).

structural aspects of the associated spline space $G$ can be imposed by restricting the placement of $t_1, \ldots, t_K$ through $p(\boldsymbol{\tau}|K)$. Both continuous and discrete specifications for knot sequences given $K$ have been suggested in the literature. In the context of fitting piecewise constant splines, for example, Green (1995) places knots continuously in some interval $[a, b]$ that covers the support of the input variables $\{x_1, \ldots, x_n\}$. They are "stochastically spaced" at order statistics of a collection of random variables, each uniformly distributed over $[a, b]$. Denison, Mallick, and Smith (1998) use this idea for univariate curve fitting, but discretize their choice of knot locations. Here, candidate breakpoints are located at order statistics of a uniform sample with state space $\{x_i; 1 \leq i \leq n\}$. To prevent knots from coalescing, Denison et al. (1998) define as allowable those spaces for which at least $K_{\text{sep}}$ data points fall in the intervals $(t_k, t_{k+1})$, $k = 1, \ldots, K-1$. Smith and Kohn (1996) also create a discrete space of candidate breakpoints, but consider only $\max(35, n)$ sites located at quantiles of the $\{x_i\}$. Any collection of points from this reduced set represents an allowable space. Similar ideas are discussed in Wong, Hansen, Kohn, and Smith (1998) where natural splines and spaces of radial basis functions are also considered. In what follows, we find that in some cases (Logspline density estimation), the discrete approach is sufficient, while in others (Triogram regression) a continuous choice of knots is required.

Broadly, each of these schemes places (essentially) uniform mass on sufficiently regular knot sequences. This leaves us with the task of specifying $p(K)$. To the extent that the number of knots also acts as a smoothing parameter, this distribution can have a considerable effect on the look of the final curves produced. We will explore several of the proposals that have appeared in the literature. The first is a simple Poisson distribution with mean $\gamma$ suggested initially by Green (1995). Denison, Mallick, and Smith (1998) take the same distribution for more general spline spaces and

argue that their results are somewhat insensitive to the value of $\gamma$. (In later attempts, this parameter is assigned a hyper-prior and sampled within a larger Markov chain Monte Carlo scheme; see Holmes and K. 1998). Just as the bounds $K_{\min} < K_{\max}$ helped restrict the greedy search in the previous section, this choice of $p(K)$ is usually truncated to have support on the set $\{K_{\min}, \ldots, K_{\max}\}$.

The next prior we will consider was suggested by Smith and Kohn (1996) and later used by Wong, Hansen, Kohn, and Smith (1998). Either by greatly reducing the number of candidate knots or by scaling the prior on the coefficients in (10.4.1), these authors suggest that $K$ be distributed uniformly on the set $K_{\min} \ldots, K_{\max}$.

The final proposal for $p(K)$ is somewhat more aggressive in enforcing small models. To properly motivate this distribution, we think of the model selection procedure as two stages: in the first we find the posterior average of all models with $k$ knots by integrating out $\boldsymbol{\tau}$ and $g$, to obtain, say $\bar{g}_k$ and its posterior probability $P(\bar{g}_k|W, K = k)$. Suppose that we consider $\bar{g}_k$ to have $k$ degrees of freedom (an admittedly questionable assumption). If we now were to use an AIC-like criterion to choose among the $\bar{g}_k$, we would select the model that minimized

$$-2 \log P(\bar{g}_k|W, K = k) + ak$$

(compare (3.2.29)). On the other hand, using the posterior to evaluate the best model suggests maximizing

$$P(\bar{g}_k|W, K = k)P(K = k).$$

If we take $P(K) \propto \exp(-a/2)$ these two approaches agree. Thus, taking a geometric distribution for $P(K)$ implies an AIC-like penalty on model dimension. In particular $a = \log n$ and $q = 1/\sqrt{n}$ imposes the same cost per knot as BIC. For reasonable settings of $K_{\min}$ and $K_{\max}$, however, the expected prior number of knots under this prior will tend to zero with $n$. While it is certainly intuitive that the prior probability of $K$ decreases monotonically with $K$, this drop may be at a faster rate than we would expect. If $a \geq 2$ then $P(K = k + 1)/P(K = k) \leq 1/e$.

**Priors on splines in a given space**

We parameterize $p(g|G)$ through the coefficients in the expansion (10.4.1). As the solution to a penalized maximum likelihood fit, smoothing splines (Wahba 1990) have a straightforward Bayesian interpretation (Silverman 1985). In univariate smoothing, for example, $G$ is a space of natural splines (given some knot sequence $\boldsymbol{\tau}$), and the "roughness" of any $g \in G$ is measured by the quantity $\int (g'')^2$. Expanding $g$ in a basis (10.4.1), it is not hard to see that

$$\int (g'')^2 = \boldsymbol{\beta}' A \boldsymbol{\beta}, \quad \text{where} \quad A_{ij} = \int g_i''(x) g_j''(x) dx \quad \text{for } 1 \leq i, j \leq J.$$

The traditional smoothing spline fit maximizes the penalized likelihood

$$\text{argmax}_{\boldsymbol{\beta}} \{ p(\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' A \boldsymbol{\beta} \},$$

for some parameter $\lambda$. Silverman (1985) observes that the solution to this problem can be viewed as a posterior mode, where $\boldsymbol{\beta}$ is assigned a partially improper, normal prior having mean $\mathbf{0}$ and variance-covariance matrix $(\lambda A)^{-1}$. (In regression problems, the posterior distribution for $\boldsymbol{\beta}$ is again normal, so the solution above is also the posterior mean.) This setup has the favorable property that it is invariant to our choice of basis. More precisely, given a positive constant $c$, *a priori* the probability that the roughness $\int (g'')^2 < c$ is a property of $g$ and not the underlying basis. When we move to more general spline spaces, we will continue to use some form of smoothing prior, also depending only characteristics of the function itself and not on the particular basis chosen to represent it.

By contrast, Smith and Kohn (1996) and Wong, Hansen, Kohn, and Smith (1998) borrow a prior distribution from Zellner (1986), taking $p(g|G) = p(\boldsymbol{\beta}|G)$ to be normal with mean $\mathbf{0}$ and variance-covariance $\lambda(X'X)^{-1}$. Here, $X$ is the design matrix corresponding to the basis functions $g_1, \ldots, g_J$. (For simplicity, we have left out details concerning nuisance parameters in the prior specification.) By considering a discrete set of candidate knots $\{t_1, \ldots, t_N\}$, Smith and Kohn (1996) specify spaces $G$ through a binary vector $\gamma$ that records the presence ($\gamma_i = 1$) or absence ($\gamma_i = 0$) of each breakpoint $t_i$. Given $G$, or equivalently $\gamma$, Zellner's prior yields a simple expression for the posterior of $\boldsymbol{\beta}$. Integrating with respect to $\boldsymbol{\beta}$, Smith and Kohn (1996) produce a closed-form expression for the (marginal) posterior distribution of $\gamma$, or equivalently $G$. The collection of knots (the value of $\gamma$) that receives the most support from the posterior is selected. George and Foster (2000) show that the hyper-parameter $\lambda$ appearing in Zellner's prior can be calibrated so that this posterior agrees with our generalized AIC criterion (3.2.29) in the sense that they both rank the candidate knot configurations similarly. As a final note, the function estimate produced by Smith and Kohn (1996) is obtained by shrinking the MLE $\widehat{g}$ computed in the single best space $G$. While Zellner's prior was chosen for purely computational convenience, it is not immediately clear why this shrinkage is reasonable for smoothing applications. Realistically, one hopes that its overall effect is minor. When a (model averaged) posterior mean is desired, however, the situation is somewhat murkier.

In any event, this class of priors can be thought of as a device for calibrating the resulting posterior with known model selection criteria. In the scheme of Denison, Mallick, and Smith (1998), on the other hand, no stochastic structure is assigned to the coefficients. Instead maximum likelihood is employed to make a deterministic choice of $\boldsymbol{\beta}$ given $G$. Computational efficiency also drives this specification. A reasonable alternative that also again choosing a prior for $\boldsymbol{\beta}$ is a simple BIC approximation motivated by an appropriate Laplace expansion (Schwarz 1978; Kass and Raftery

1995). Although motivated by simulation and other empirical studies, the default value of $a = \log n$ in the generalized AIC criterion (3.2.29) of Stone, Hansen, Kooperberg, and Truong (1997) is an application of this idea to spline modeling.

### *10.4.4 Computation*

In our greedy search algorithm, we generated a sequence of nested models, each differing from its predecessor by a single knot addition or deletion. One byproduct of taking a Bayesian approach is that a variety of McMC schemes exist for exploring the space of candidate knot configurations. In the case of Smith and Kohn (1996), for example, a Gibbs sampler is applied to the elements of the index vector $\gamma$. In short, the posterior distribution of each $\gamma_i$ given the remaining $\gamma_j$, $j \neq i$, is Bernoulli with success probability depending on the number of knots currently in the fit and the residual sum of squares (this method was derived in the context of normal regression). As the sampler progresses along the vector $\gamma$, we are again generating a sequence of nested spaces, each obtained by inserting or removing a knot in the previous model.

In situations like Triogram regression, however, it is somewhat unnatural to discretize the space of knots. In order to treat a variety of estimation problems simultaneously, we have chosen the reversible jump McMC scheme developed by Green (1995). Denison, Mallick, and Smith (1998) implement this technique in the context of general univariate and additive regression. At this point, we expect that details of the scheme are well known, and we instead focus on the type of moves that we need to implement the sampler. In general, we alternate (possibly at random) between the following moves.

- **Increase model dimension.** In this step, we introduce a new knot or vertex into an existing collection of breakpoints. Given the concavity properties of ELMs the change in the log-likelihood can either be computed exactly, or approximated using the appropriate Rao statistic. When knots are selected from a discrete set, candidates are chosen at uniformly from among the set that yields an allowable space.

- **Decrease model dimension.** As with the greedy scheme, knots are deleted by imposing a constraint on one or more coefficients in the spline expansion. Again, concavity suggests that we can either evaluate the drop in the log-likelihood exactly, or through the Wald statistics described in the previous section. For univariate models, any knot can be removed at any time (assuming we have more than $K_{\min}$ breakpoints to chose from). This is not true in the case of Triogram models. Only knots in one of the configurations described in Figure 9.10 are candidates for removal. To maintain the reversibility

of the Markov chain, more elaborate removal schemes would require more elaborate addition moves.

- **Make structural changes to $G$ that do not change dimension.** Unlike our standard greedy scheme, non-nested steps like moving a knot are now possible. Moving a knot from $t_k$ to $t_k^*$ technically involves deleting $t_k$ and then inserting a new breakpoint at $t_k^*$. In the context of non-linear models like Logspline, we were initially concerned that such a move would be computationally expensive. On the contrary, with smart initial conditions on the Newton-Raphson steps, we can calculate the change in the log-likelihood exactly and still maintain an efficient algorithm.

- **Update (possibly) $g$ and any nuisance parameters.** In a non-linear model like Logspline, we can either apply a suitable approximation to the posterior and integrate with respect to the coefficients $\boldsymbol{\beta}$, or we can fold sampling them into our Markov chain. Because of the concavity of ELMs and our use of normal priors on $\boldsymbol{\beta}$, our posterior is again concave in $\boldsymbol{\beta}$. We choose to generate them in a Metropolis step from a straightforward normal approximation. Similar in spirit is the rejection sampling scheme of Zeger and Karim (1991).

### *10.4.5 Logspline density estimation*

#### Priors and computation

In our implementation we consider those spaces $G$ for which all $K$ knots are located at data points (recall that $K$ knots produce a space with dimension $J = K - 1$). We require that there are at least $K_{\text{sep}}$ data points in between any two knots. The restriction that knots are located at data points is purely for convenience, but represents little loss of flexibility especially in the context of density estimation (where peaks in the underlying density naturally produce more candidate knots). Some restriction, however, to prevent consecutive knots from being to close together is needed for numerical stability.

Following Green (1995) and Denison, Mallick, and Smith (1998), we cycle between proposals for adding, deleting and moving knots, assigning these moves probabilities $b_J$, $d_J$ and $1 - b_J - d_J$ (see Denison et al. 1998). New knots can be positioned at any data point that is at least $K_{\text{sep}}$ data points removed from one of the current knots. Subject to this constraint, knot addition follows a simple two step procedure. First, we select one of the intervals $(L, t_1), (t_1, t_2), \ldots, (t_K, U)$ uniformly at random (where the $t_k$ are the current breakpoints). Within this interval, the candidate knot is then selected uniformly at random from one of the allowable data points. When moving a knot, we either propose a large move (in which a knot is first deleted, and then added using the addition scheme just described) or a

small move (in which the knot is only moved within the interval between its two neighbors). Each of these two proposals have probability $(1-d_J-b_J)/2$.

After each reversible jump step, we then update the coefficients $\boldsymbol{\beta}$. To do this, we use the fact that for a given set of knots, we have a parametric model, and that the posterior distribution of $\boldsymbol{\beta}$ given $G$, $K$ and the data is thus approximately multivariate normal with covariance matrix $\Sigma = (\lambda A + H)^{-1}$, and mean $\Sigma H\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ in $G$, and $H$ is the Hessian of the log-likelihood function at $\widehat{\boldsymbol{\beta}}$. An observation from this distribution is used as a proposal in a Metropolis step. Because we are using (partially improper) smoothing priors, the acceptance ratio for this proposal is formally undetermined (recall that the prior covariance matrices are degenerate). We solve this problem by "canceling" the zero eigenvalue in the numerator and the denominator (see also Besag and Higdon 1999).

**Application and Simulation**

To compare the performance of the various possible implementations of Logspline density model selection procedures, we carried out a simulation study. We generated data from three densities:

**normal** the standard normal density;

**slight bimodal** $f(y) = 0.5{*}f_Z(y; 1.25, 1){+}0.5{*}f_Z(y; -1.25, 1.1)$, where $f_Z(y; \mu, \sigma)$ is the normal density with mean $\mu$ and standard deviation $\sigma$;

**sharp peak** $f(y) = 0.8 * g(y) + 0.2 * f)Z(y; 2, .07)$, where $g(Y)$ is the density of the lognormal random variable $Y = \exp(Z/2)$ and $Z$ has a standard normal distribution.

These three densities are displayed in Figure 10.17. From each we generated 100 independent samples of size $n = 50$, 200, 1000, and 10000. We applied a variety of Logspline methods to these data sets: (i) the greedy, stepwise addition-deletion proposed by Stone, Hansen, Kooperberg, and Truong (1997); (ii) simulated annealing tuned to optimize BIC, termed SALSA for simulated annealing Logspline approximation (described below); and several Bayesian schemes outlined in the previous section. For all the Bayesian methods we estimated the posterior mean by a simple pointwise average of the McMC samples. Otherwise, the Bayesian approaches differ in two aspects:

- The prior on the model size: we used the geometric prior with parameter $p = 1 - 1/\sqrt{n}$, the Poisson prior with parameter 5, as proposed by Denison, Mallick, and Smith (1998), and a uniform prior; and
- Parameter estimates $\widehat{\boldsymbol{\beta}}$: we took either the maximum likelihood (ML) estimate, or we assigned a multivariate normal prior to $\boldsymbol{\beta}$ (for one of several choices for $\lambda$).

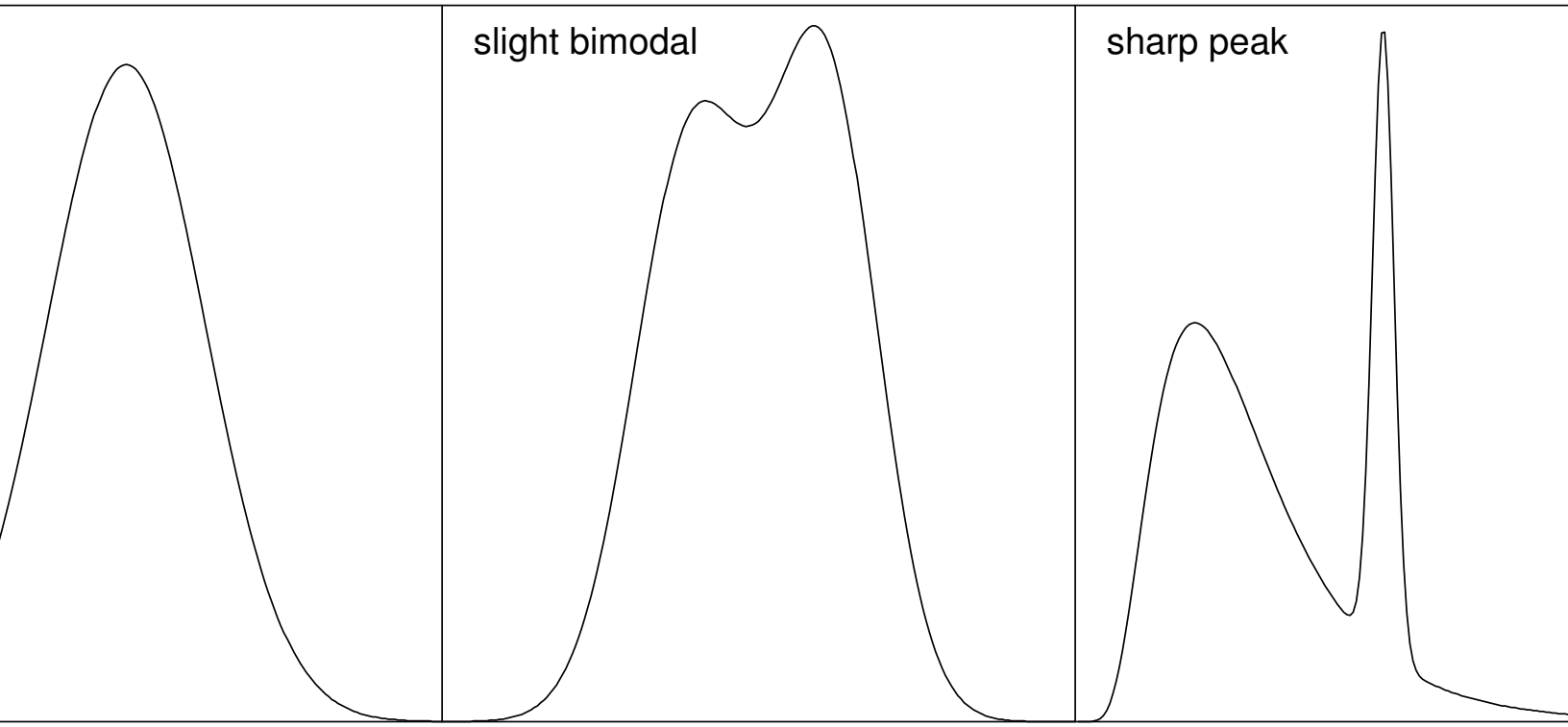


FIGURE 10.17. Densities used in the simulation study.

Table 10.1 summarizes the versions of Logspline which are reported here.

For simulated annealing (ii) we ran the same McMC iterations as for version (iii), but rather than selecting the mean of the sampled densities, we chose the density which minimizes BIC. As described in above, this is very similar taking the density with the largest a posteriori probability (the mode), except that we ignore the prior on knot locations given the number of knots, $K$. This would have changed the penalty in the BIC criterion from $K \log n$ to $K \log n + \frac{1}{2} \log \binom{n}{K}$. Since version (ii) begins with the fit obtained by the greedy search (i), it is guaranteed to improve as far as BIC is concerned. Version (iii) uses the same penalty structure as version (ii), but averages over McMC samples. Version (iv) is included since a Poisson (5) prior was proposed by Denison et al. (1998). It applies a considerably smaller penalty on model size. Versions (v)–(ix) experiment with penalties on the coefficients. (After trying several values for $\lambda$, we decided that $1/n$ seemed reasonable.) As argued above, generating the parameters using a

| | model size | parameters |
|---|---|---|
| (i) | greedy optimization of BIC | |
| (ii) | simulated annealing optimization of BIC | |
| (iii) | geometric | MLE |
| (iv) | Poisson (5) | MLE |
| (v) | uniform | $\lambda = 1/n^2$ |
| (vi) | uniform | $\lambda = 1/n$ |
| (vii) | uniform | $\lambda = 1/\sqrt{n}$ |
| (viii) | uniform | $\lambda = 1$ |
| (ix) | geometric | $\lambda = 1/n$ |

TABLE 10.1. Versions of Logspline Density Estimation used in the simulation study.

multivariate normal prior distribution implies smoothing with a BIC-like penalty. As such, we would expect that using $\lambda = 1/n$ with a uniform prior (version viii) may give reasonable results, but that using a geometric prior (version ix) would smooth too much. Choosing $\lambda$ too large, as in versions (vii)–(viii), leads to oversmoothing, while choosing $\lambda$ too small tends to produce overly wiggly fits.

For versions (iii) and (iv) we ran 600 McMC iterations, of which we discarded the first 100 as burn-in. Some simple diagnostics (not reported) suggest that after 100 iterations the chain is properly mixed. For versions (v)–(ix) each structural change was followed by an update of the coefficients $\boldsymbol{\beta}$.

In Table 10.2, we report ratios of integrated squared errors between the greedy scheme and the other methods outlined above. In addition, we feel that it is at least as important for a density estimate to provide the correct general "shape" of a density as to have a low integrated squared error. To capture the shape of our estimates, we counted the number of times that a scheme produced densities having too few, too many and the correct number of modes. These results are summarized in Tables 10.3 and 10.4. Table 10.5 calculates the "total" lines of Tables 10.3 and 10.4. Note that for simulations of a normal distribution it is not possible for an estimate to have too few modes.

From Table 10.2 we note that most methods show a moderate overall improvement over the greedy version of Logspline, except for (viii). This scheme oversmooths the data, so that the details (like the mode in the sharp peaked distribution) are frequently missed. We note that version (iii), choosing the mode of a Bayesian approach, is the only version that outperforms the greedy version for all 12 simulation set-ups. Otherwise, the difference between versions (ii), (iii), (v) and (ix) seems to be minimal. In particular, if we had chosen another set of results than those for (i) to

| version | (i) | (ii) | (iii) | (iv) | (v) | (vi) | (vii) | (viii) | (ix) |
|---|---|---|---|---|---|---|---|---|---|
| n | MISE | ratio of MISE over MISE of the greedy version (i) | | | | | | | |
| normal distribution | | | | | | | | | |
| 50 | 0.0279 | 0.73 | 1.52 | 1.84 | 1.27 | 0.66 | 0.40 | 0.26 | 0.67 |
| 200 | 0.0107 | 0.49 | 0.60 | 1.23 | 0.69 | 0.79 | 0.50 | 0.24 | 0.66 |
| 1000 | 0.0021 | 0.59 | 0.58 | 1.33 | 0.64 | 0.87 | 0.90 | 0.42 | 0.73 |
| 10000 | 0.0002 | 0.33 | 0.49 | 1.45 | 0.69 | 1.35 | 1.10 | 0.80 | 0.87 |
| slightly bimodal density | | | | | | | | | |
| 50 | 0.0250 | 0.88 | 1.09 | 1.34 | 0.97 | 0.48 | 0.36 | 0.36 | 0.50 |
| 200 | 0.0077 | 0.80 | 0.61 | 1.14 | 0.88 | 0.70 | 0.38 | 0.46 | 0.61 |
| 1000 | 0.0016 | 0.57 | 0.60 | 1.13 | 0.87 | 0.89 | 0.66 | 0.40 | 0.77 |
| 10000 | 0.0002 | 0.77 | 0.61 | 0.88 | 0.79 | 0.71 | 0.82 | 0.51 | 0.84 |
| density with sharp peak | | | | | | | | | |
| 50 | 0.1523 | 0.97 | 0.78 | 0.81 | 0.66 | 0.68 | 0.90 | 1.12 | 0.72 |
| 200 | 0.0370 | 0.89 | 0.75 | 0.94 | 1.04 | 0.93 | 2.02 | 3.62 | 1.13 |
| 1000 | 0.0097 | 0.81 | 0.67 | 0.81 | 1.12 | 0.67 | 2.01 | 8.90 | 0.74 |
| 10000 | 0.0015 | 0.72 | 0.57 | 0.57 | 1.05 | 0.64 | 0.58 | 21.43 | 0.76 |
| average | 1.00 | 0.71 | 0.74 | 1.12 | 0.89 | 0.78 | 0.89 | 3.21 | 0.75 |

TABLE 10.2. Mean Integrated Squared Error (MISE) for the simulation study.

normalize by, the order of the average MISE for these four methods was often changed.

From Table 10.3 we note that version (viii), and to a lesser extent (ii) and (vii), have trouble with the slight bimodal density, preferring a model with just one peak. Versions (vii) and (viii) find too few modes, leading us to conclude that $\lambda$ should be chosen smaller than $1/\sqrt{n}$ when using a uniform prior on model size. On the other hand, the Poisson prior leads to models exhibiting too many peaks, as do versions (iii), (v) and (vi).

Overall, it appears that the greedy, stepwise search is not too bad. It is several orders of magnitude faster than any of the other methods. The greedy approach, as well as SALSA have the advantage that the final model is again a Logspline density, which can be stored for later use. For the other methods, we must record the posterior mean at a number points. This has the potential of complicating later uses of our estimate. Among the Bayesian versions that employ ML estimates, version (iii) seems to perform best overall, while among those that put a prior on the coefficient vector, versions (vi) and (ix) (both of which set $\lambda = 1/n$) are best. It is somewhat surprising that version (ix) performs so well, since it effectively imposes twice the BIC penalty on model size: one coming from the geometric prior, and one from the normal prior on the parameters. Kooperberg and Stone (1992) argue that the Logspline method is not very sensitive to the exact value of the parameter, possibly explaining the behavior of version (ix).

| Version n | (i) | (ii) | (iii) | (iv) | (v) | (vi) | (vii) | (viii) | (ix) |
|---|---|---|---|---|---|---|---|---|---|
| slightly bimodal density | | | | | | | | | |
| 50 | 45 | 52 | 4 | 0 | 0 | 21 | 74 | 99 | 31 |
| 200 | 6 | 22 | 13 | 0 | 16 | 1 | 18 | 96 | 19 |
| 1000 | 5 | 17 | 19 | 0 | 12 | 7 | 6 | 45 | 16 |
| 10000 | 4 | 12 | 4 | 1 | 7 | 3 | 4 | 2 | 10 |
| density with sharp peak | | | | | | | | | |
| 50 | 24 | 38 | 1 | 0 | 1 | 9 | 56 | 99 | 13 |
| 200 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 89 | 1 |
| 1000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| total | 84 | 142 | 41 | 1 | 38 | 41 | 158 | 430 | 90 |

TABLE 10.3. Number of times out of 100 simulations that a Logspline density estimate had too few modes.

We applied the nine versions of Logspline used for the simulation study to the income data discussed in Stone, Hansen, Kooperberg, and Truong (1997). the results are displayed in Figure 10.14. For the computations on the income data we ran the McMC chain for 5000 iterations in which a new model was proposed, after discarding the first 500 iterations for burn-in. For the versions with priors on the parameters we alternated these iterations with updates of the parameters. In Kooperberg and Stone (1992) it was argued that the height of the peak should be at least about 1. Thus, it appears that versions (vii) and (viii) have oversmoothed the peak. On the other hand, version (iv) seems to have too many small peaks.

It is interesting to compare the number of knots for the various schemes. The greedy estimate (version i) has 8 knots, and the simulated annealing estimate (version ii) has 7 knots. The Bayesian versions (iii), (v), (vi) and (ix) have an average number of knots between 5 and 8, while the three versions that produced unsatisfactory results (iv, vii, and viii) have an average number of knots between 14 and 17.

The McMC iterations can also give us information about the uncertainty in the knot locations. To study this further, we ran a chain for version (iii) with 500,000 iterations. Since the knots are highly correlated from one iteration to the next (at most one knot moves at each step), we only considered every 250th iteration. The autocorrelation function of the fitted log-likelihood suggested that this was well beyond the time over which iterations are correlated. This yielded 2000 sets of knot locations: 1128 with five knots, 783 with six knots, 84 with seven knots, and 5 with eight knots. When there were five knots, the first three were always located close to the mode, the fourth one was virtually always between 0.5 and 1.25, and the last knot between 1 and 2. Figure 10.18 displays Logspline density

| Version n | (i) | (ii) | (iii) | (iv) | (v) | (vi) | (vii) | (viii) | (ix) |
|---|---|---|---|---|---|---|---|---|---|
| slightly bimodal density | | | | | | | | | |
| 50 | 18 | 11 | 94 | 100 | 99 | 49 | 5 | 0 | 28 |
| 200 | 34 | 9 | 38 | 100 | 43 | 81 | 21 | 0 | 24 |
| 1000 | 26 | 4 | 15 | 91 | 13 | 68 | 54 | 32 | 32 |
| 10000 | 4 | 1 | 7 | 61 | 10 | 31 | 29 | 1 | 17 |
| slightly bimodal density | | | | | | | | | |
| 50 | 4 | 1 | 84 | 99 | 74 | 6 | 0 | 0 | 4 |
| 200 | 16 | 1 | 19 | 99 | 31 | 55 | 4 | 0 | 5 |
| 1000 | 15 | 1 | 13 | 93 | 5 | 51 | 31 | 1 | 17 |
| 10000 | 6 | 1 | 8 | 68 | 1 | 33 | 39 | 0 | 6 |
| density with sharp peak | | | | | | | | | |
| 50 | 15 | 8 | 90 | 93 | 66 | 3 | 1 | 0 | 2 |
| 200 | 36 | 19 | 46 | 94 | 30 | 43 | 5 | 0 | 5 |
| 1000 | 28 | 14 | 30 | 77 | 20 | 32 | 12 | 1 | 9 |
| 10000 | 25 | 12 | 15 | 31 | 16 | 20 | 30 | 11 | 7 |
| total | 227 | 82 | 459 | 1006 | 408 | 472 | 231 | 46 | 156 |

TABLE 10.4. Number of times out of 100 simulations that a Logspline density estimate had too many modes.

| | (i) | (ii) | (iii) | (iv) | (v) | (vi) | (vii) | (viii) | (ix) |
|---|---|---|---|---|---|---|---|---|---|
| too few | 84 | 142 | 41 | 1 | 38 | 41 | 158 | 430 | 90 |
| too many | 227 | 82 | 459 | 1006 | 408 | 472 | 231 | 46 | 156 |
| total | 311 | 224 | 500 | 1007 | 446 | 513 | 389 | 476 | 246 |

TABLE 10.5. Number of times out of 100 simulations that a Logspline density estimate had an incorrect number of modes.

estimates (version i) for the locations of the first three knots. The locations of these three knots overlap considerably. The peakiness in these density estimates is partly caused by discreteness in the data, and partly by the requirement that two knots cannot be too close, sometimes restricting the possible knot locations.

When there are six knots, the extra knot can either be a fourth knot in the peak, or it is beyond the fifth knot. This becomes apparent when we examine Figure 10.19, which contains plots of the location of knot 4 versus the location of knot 5 and of the location of knot 6 versus the location of knot 5 within the same panel. If knot 5 is below about 1.2, knot 4 is usually in the peak, and the 6 knot is below 2, but when knot 5 is above 1.2, knot 4 is between 0.5 and 1.2, and knot 6 is above 2. Note that in this plot knot 4 has to be below the diagonal line and knot 6 has to be above this line.

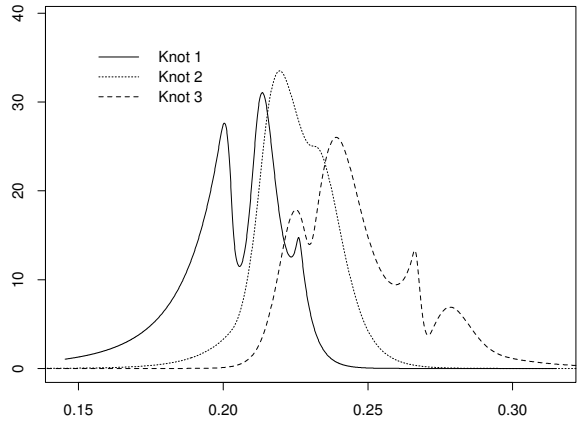


FIGURE 10.18. Density estimates of the location of the first three knots of Logspline density estimates for the income data with five knots.

We have no explanation for the apparent negative correlation between the location of knots 4 and 5 when the location or knot 5 is above 1.2.

### *10.4.6 Triogram regression*

**Priors and computational methods**

As with univariate spline models, a prior on the space of Triograms is most easily specified by first considering the structure of the approximation space, which in this case is a triangulation $\triangle$. Several authors have discussed the use of the Delaunay triangulation for estimating an unknown function. In this case, $\triangle$ is completely determined by a collection of vertices $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$, and the prior specification for $\triangle$ reduces to a point process for locating vertices. This is the approach followed in Green (1995). Besides simplifying the prior on $\triangle$, Rippa (1992) shows that for any set of coefficients $\beta_1, \ldots, \beta_K$, the Delaunay triangulation produces the surface having minimal roughness in the sense that the norm of $g = \sum \beta_k B_k$,

$$|g| = \sum_{\delta \in \triangle} \int_\delta \left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2 dxdy \tag{10.4.2}$$

is smallest among all triangulations $\triangle$ of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$. This remarkable fact is independent of the coefficients being interpolated. Recently, alternate smoothness measures have been proposed in the numerical analysis literature for which the Delaunay triangulation is not optimal (we will return to this subject shortly).

In general, the structure $\triangle$ is not determined solely by a collection of vertices or knots, but instead many triangulations connect a given set of

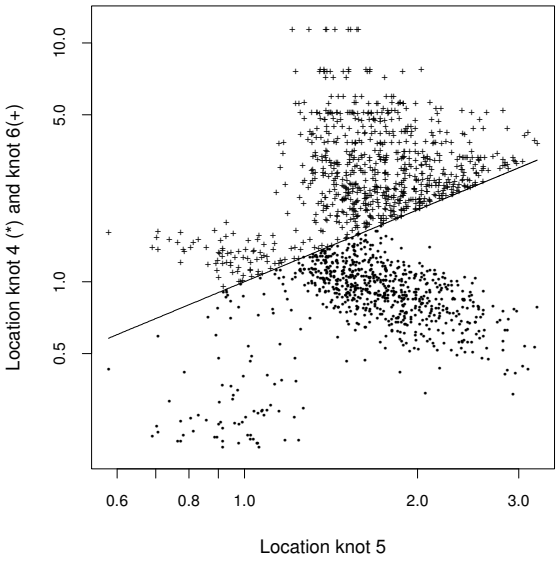


FIGURE 10.19. Locations of knots four through six of Logspline density estimates for the income data with six knots.

vertices (and the fixed polygonal boundary of $\mathcal{U}$). Unfortunately, a closed-form expression for the number of such triangulations does not exist, and estimating it for even moderately sized configurations is prohibitively expensive. To see how this complicates matters, suppose we follow the strategy for Logspline and consider a hierarchical prior of the form

$$p(\triangle|\boldsymbol{V}_K, K)\, p(\boldsymbol{V}_K|K)\, p(K), \tag{10.4.3}$$

where $\triangle$ is a triangulation of the vertices $\boldsymbol{V}_K = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$. Assigning any proper distribution to $\triangle$ given $\boldsymbol{V}_K$ introduces an unknown normalizing constant (obtainable only by enumerating the number of ways one can triangulate the set $\boldsymbol{V}_K$) that does not cancel between different configurations. With this setup, calculating acceptance probabilities in any McMC scheme for moves between models with different vertex sets poses a difficult computational problem.

Before we can describe a solution, we need some notation. For simplicity, let $p(\boldsymbol{V}_K, K)$ represent a Markov point-process on the interior of $\mathcal{U}$. Following Kelly and Ripley (1976) these processes are absolutely continuous with respect to the distribution of a Poisson process on $\mathcal{U}$ having unit intensity. Each such point-process on $\mathcal{U}$ is defined by a sequence of intensity functions $p_k(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ that do not depend on the order of the vertices

$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. The sequence must satisfy $\sum_k p_k = 1$, where

$$p_k = \begin{cases} \exp^{-\mu(\mathcal{U})}, & k = 0 \\ \frac{\exp^{-\mu(\mathcal{U})}}{k!} \int_{\mathcal{U}^k} p_k(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)\mu(\boldsymbol{v}_1)\cdots\mu(\boldsymbol{v}_k) \end{cases}$$

and $\mu$ is Lebesgue measure on $\mathcal{U}$. The values $p_k$, $k = K_{\min}, \ldots, K_{\max}$, are the probabilities of observing $k$ vertices. Kelly and Ripley (1976) introduce functions $p_k(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ proportional to

$$b^k \; c^{\mathcal{N}_k} \; \psi(k),$$

where $\mathcal{N}_k$ counts neighboring vertices (points separated by a distance $\rho$, say); and $c$ can be chosen to force inhibition ($c < 1$) or clustering ($c > 1$). (The formula for $\psi$ is given in Kelly and Ripley 1976). If $c = 1$, then we have a homogeneous Poisson process with intensity $b$. We have chosen this class of priors because it allows us to "stochastically space" vertices. Furthermore, by scaling the intensity functions $p_k$, we can mimic the behavior outlined for Logspline in terms of the prior number of knots. (In the next section, we will consider analogs of the Poisson, geometric and uniform priors described above.)

Returning to the distribution on triangulations, given the set $\boldsymbol{V}_K = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$, let $\Gamma_{\mathcal{U}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ denote the collection of all triangles that can be formed by joining these points. Then, following Nicholls (1998), the joint distribution $p(\triangle, \boldsymbol{V}_K, K) = p(\triangle, \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}, K)$ is $C^{-1}p(\boldsymbol{V}_K, K)$ where the normalizing constant $C$ is given by

$$C \;=\; \sum_{k=K_{\min}}^{K_{\max}} \int \; |\Gamma_{\mathcal{U}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)| \; p_k(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)\mu(\boldsymbol{v}_1)\cdots\mu(\boldsymbol{v}_k).$$

This construction is possible, because the number of triangulations connecting $\boldsymbol{V}_K$ is finite and bounded by

$$|\Gamma_{\mathcal{U}}(\boldsymbol{V}_K)| \leq \frac{2[4K + 2l - 5]! \; [2l - 3]!}{[3K + 2l - 3]! \; K! \; [l - 1]! \; [l - 3]!}, \tag{10.4.4}$$

where $l$ is the number of vertices specifying the polygonal boundary of $\mathcal{U}$ (Nicholls 1998). From here, we can also define the "volume" of a set of triangles $\triangle$ according to

$$\int_{\triangle} d\delta = C^{-1} \sum_{k=K_{\min}}^{K_{\max}} \int_{\mathcal{U}^k} \mu(\boldsymbol{v}_1)\cdots\mu(\boldsymbol{v}_k) \sum_{\delta \in \Gamma_{\mathcal{U}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)} I(\delta \in A)$$

where $\Gamma_{\mathcal{U}}$ denotes the collection of all triangulations on $\mathcal{U}$. At the point $\delta \in \Gamma_{\mathcal{U}}$, the element of measure $d\delta$ is simply the set of triangulations that can be obtained from $\delta$ by moving the vertices of $\delta$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, within the volume $\mu(\boldsymbol{v}_1)\cdots\mu(\boldsymbol{v}_k)$.

While this approach produces a valid distribution on triangulations, our prior on model size is no longer $p_k$ as it would have been if we had specified a uniform distribution for $p(\triangle|\boldsymbol{V}_K, K)$ in (10.4.3). Therefore, we propose "normalizing" by a factor $h(K)$, so that $p(\triangle, V_K, K) \propto p(\boldsymbol{V}_K, K)h(K)$, for some function $h$. Shortly, we experiment with $h(K) = 1$ and $h(K) = K!$. Simple calculations with (10.4.4) assuming $\mathcal{U}$ is a triangle ($l = 3$) suggest that the latter choice will perform well for models having as many as 15 internal vertices. So far, we have considered *all possible* triangulations of a point set, $\Gamma_{\mathcal{U}}$. Frequently, we would like to restrict out set of meshes, perhaps insisting that each triangle be larger than a minimum size or have an aspect ratio larger than some bound. Each of these conditions will lower the count $|\Gamma_{\mathcal{U}}(\boldsymbol{V}_K)|$, complicating our choice of weight functions $h(k)$. As an alternative in these more complicated settings, we also consider a simple uniform distribution for $p(\triangle, V_K, K)$, embedding a penalty on model size in the prior on coefficients.

Now, unlike the Logspline example, we do not have a single obvious choice for the smoothing prior on the coefficients given a fixed triangulation. As mentioned above, the Sobolev semi-norm (10.4.2) and its connection with the Delaunay triangulation led to the creation of several techniques for measuring the smoothness of a continuous, piecewise-linear surface (Dyn, Levin, and Rippa 1990b; Dyn, Levin, and Rippa 1990a). One class of criteria measures how near the fit is to a plane. Typically, these are edge-based, compiling a roughness penalty along edges in $\triangle$. For example, (Dyn et al. 1990b; Dyn et al. 1990a) accumulate the jump in the normal derivative across each edge. Measuring the squared differences yields a quadratic penalty on the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ which can be written $\boldsymbol{\beta}^t A \boldsymbol{\beta}$ for a positive-semidefinite matrix $A$. As constant and linear functions have zero roughness by this measure, $A$ has two zero eigenvalues. As was done for Logspline, we use $A$ to generate a partially improper normal prior on $\boldsymbol{\beta}$ (with prior variance $\lambda\sigma^2$, where $\sigma^2$ is the error variance). This assignment is similar in spirit to that taken by Nicholls (1998) who looked at differences between neighboring triangles in piecewise constant fits. Because Triogram regression is a simple linear model, we are able to remove $\boldsymbol{\beta}$ entirely by integration. This approach allows us to focus on structural changes and was used by Smith and Kohn (1996) and Wong, Hansen, Kohn, and Smith (1998) for univariate and multivariate regression, respectively.

Following Denison, Mallick, and Smith (1998), we assign a proper, inverse-gamma distribution to $\sigma$, and update its value after each structural (reversible jump) move. An alternative approach would be to integrate out $\sigma^2$ completely, as is done in Wong, Hansen, Kohn, and Smith (1998). Because of the simplicity of the Triogram model, we can also consider assigning a prior to the smoothing parameter $\lambda$, and either sample it along with $\sigma^2$ or integrate it out. We have instead chosen to compare a number of fixed choices for $\lambda$ that depend on the sample size $n$.

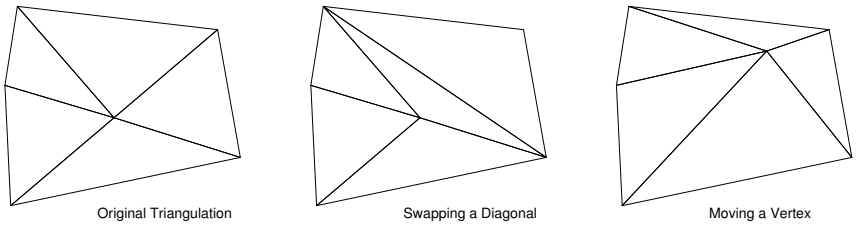


FIGURE 10.20. Additional structural moves for the reversible jump McMC scheme. Note that these two proposals result in a non-nested sequence of spaces.

Finally, we must augment our set of structural changes to $\triangle$. To implement the reversible jump McMC sampler outlined at the beginning of this section, we have included two non-nested moves that maintain the dimension of the space $G$, but change its structure. In Figure 10.20, the middle panel illustrates moving a vertex inside the union of triangles that contain it; while in the final panel, we demonstrate "swapping" an edge. It can be shown that all triangulations of a given point set $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$ can be obtained by this operation. For Triograms, the notion of an allowable space can appear through size or aspect ratio restrictions on the triangulations, and serves to limit the region in which we can place new vertices or to which we can move existing vertices. For example, given a triangle, the set into which we can insert a new vertex and still maintain a minimum area condition is a subtriangle, easily computable in terms of barycentric coordinates (see Chapter 9)). Next, we explore how these conditions impact the final mean squared error results.

**Simulation results**

In Figure 10.15, we present a series of three fits to a simulated surface plotted in the upper lefthand corner. A data set consisting of 100 observations was generated by first sampling 100 design points uniformly in the unit square. The actual surface is described by the function

$$\frac{40\exp\{8[(x_1-0.5)^2+(x_2-0.5)^2]\}}{\exp\{8[(x_1-0.2)^2+(x_2-0.7)^2]\}+\exp\{8[(x_1-0.7)^2+(x_2-0.2)^2]\}},$$

to which we add standard Gaussian errors. This function first appeared in Gu, Bates, Chen, and Wahba (1990), and it will be hereafter referred to as simply GBCW. The signal-to-noise ratio in this setup is about 3. In the lower lefthand panel in Figure 10.15, we present the result of applying the greedy, Triogram algorithm. As is typical, the procedure has found a fairly regular, low-dimensional mesh describing the surface (the MISE is 0.31).

| | model size | parameters |
|---|---|---|
| (i) | greedy optimization of BIC | |
| (ii) | simulated annealing optimization of BIC | |
| (iii) | Poisson (5) | MLE |
| (iv) | geometric | MLE |
| (v) | uniform | $\lambda = 1/n$ |

TABLE 10.6. Versions of Triogram used in the simulation study.

For the fit plotted in the lower righthand panel, we employed a simulated annealing scheme similar to that described for Logspline. The geometric prior is used to guide the sampler through triangulations $\triangle$, and in each corresponding spline space $G$ we consider $\widehat{g}$, the MLE (or in this case the ordinary leastsquares fit). In this way, the objective function matches that of the greedy search, the generalized AIC criterion (3.2.29). The scheme alternates between (randomly selected) structural changes (edge swaps and vertex moves, additions and deletions) and updating the estimate $\widehat{\sigma}^2$ of the noise variance. After 6,000 iterations, the sampler has managed to find a less regular, and marginally poorer-fitting model (the MISE is 0.32). This effect is typical across the experiments conducted in this section. In the context of triangulations, the greedy search is subject to a certain regularity that prevents configurations like the one in the Figure 10.15. We can recapture this either by placing restrictions on the triangulations in each mesh (say, imposing a smallest allowable size or aspect ratio) or by increasing the penalty on dimension, specified through our geometric prior.

In the last panel, we present the result of model averaging using a uniform prior on model size and a smoothing prior on the coefficients ($\lambda = 1/n$). The sampler is run for a total of 6,000 iterations, of which 1,000 are discarded as burn-in. We then estimate the posterior mode as a pointwise average of the sampled surfaces. The final fit is smoother in part because we are combining many piecewise-planar surfaces. We still see sharp effects, however, where features like the central ridge are present. The model in the lower righthand panel is not unlike the surfaces visited by this chain. As spaces $G$ are generated, the central spine (along the line $y = x$) of this surface is always present. The same is true for the hinged portions of the surface along the lines $x = 0$ and $y = 0$. With these caveats in mind, the MISE of the averaged surface is about half of the other two estimates (0.15).

We repeated these simulations for several sample sizes, taking $n = 100$, 500 and 1000 (100 repetitions for each value of $n$). In Table 10.6, we present several variations in the prior specification and search procedure. In addition to GBCW, we also borrow a test function from Breiman (1991), which we will refer to as Exp. Here, points $\boldsymbol{u} = (u_1, u_2)$ are selected uniformly from the square $[-1, 1]^2$. The response is given by $\exp(u_1 \sin(\pi u_2))$ to which

| distribution | version n | (i) MISE | (ii) | (iii) | (iv) | (v) |
|---|---|---|---|---|---|---|
| | | | ratio of MISE over (i) | | | |
| GBCW (high snr) | 100 | 0.31 | 1.35 | 0.85 | 0.78 | 0.77 |
| GBCW (high snr) | 500 | 0.10 | 1.00 | 0.64 | 0.76 | 0.80 |
| GBCW (high snr) | 1000 | 0.08 | 0.91 | 0.82 | 0.94 | 0.79 |
| Exp (low snr) | 100 | 0.15 | 0.90 | 0.52 | 0.51 | 0.49 |
| Exp (low snr) | 500 | 0.04 | 0.85 | 0.46 | 0.50 | 0.47 |
| Exp (low snr) | 1000 | 0.03 | 0.51 | 0.32 | 0.40 | 0.46 |
| | average | 1.00 | 0.92 | 0.60 | 0.65 | 0.63 |

TABLE 10.7. Mean Integrated Squared Error (MISE) Two Smooth Test Functions.

normal noise is added ($\sigma = 0.5$). The signal-to-noise ratio in this setup is much lower, 0.9. The results are presented in Table 10.7. It seems reasonably clear that the simulated annealing approach can go very wrong, especially when the sample size is small. Again, this argues for the use of greater constraints in terms of allowable spaces when $n$ is moderate. It seems that model averaging with the smoothing prior ($\lambda = 1/n$) and the Poisson prior of Denison, Mallick, and Smith (1998) perform the best. A closer examination of the fitted surfaces reveals the same kinds of secondary structure as we saw in Figure 10.15. To be sure, smoother basis functions would eliminate this behavior. It is not clear at present, however, if a different smoothing prior on the coefficients might serve to "unkink" these fits.

Unlike Logspline (or we believe any univariate, extended linear models), the behavior of the Triogram procedure is extremely sensitive to our choice of "hyperparameters." For example, in the simulations reported above, we took as our prior on vertices a point process introduced by Kelly and Ripley (1976) and described above. The priors on model size were selected to yield either the Poisson (5) distribution or the geometric with parameter $1 - 1/\sqrt{n}$. In each case, we scaled these distributions by $h(K) = K!$. Without at least this scaling, the samplers quickly drifted into very high-dimensional, poor fitting models. By imposing aggressive limits on the minimum size for each triangle in the mesh, or perhaps the maximum aspect ratio, we can remove the need for this scaling and return the sampler to a better portion of model space. Unfortunately, these bounds are difficult to set a priori, and can yield 20-25% variations in mean squared error. As an automated procedure, it seems more reasonable to sidestep the issue of model size entirely and restrict our attention to priors on coefficient vectors. Finally, we comment on the somewhat surprising performance of the Poisson (5) distribution. While for Logspline this choice led to undersmoothed densities, it would appear that the Triogram scheme benefits from slightly larger models. We believe that this is because of the bias in-

volved in estimating a smooth function by a piecewise-linear surface. In general, these experiments indicate that tuning the Bayesian schemes in the context of a Triogram model is much more difficult than univariate set-ups. One comforting conclusion, however, is that essentially each of the schemes considered outperform the simple greedy search.

In Chapter 9, we studied the performance of the greedy Triogram procedure on a bivariate data set known to exhibit a simple hinge, not aligned with either of the coordinate axes (Cleveland and Fuentes 1996). While we have seen that a certain amount of smoothing is possible with the Bayesian estimator when the underlying target function is smooth, in this case, we hope that sampler will spend time in very simple, "nearby" ridge models. This would allow the non-greedy schemes to still capture ridges effectively. In Figure 10.15 we present two surfaces, one from simulation set-up (v) and one from (iii). It is clear from this figure that Poisson prior yields a chain that spends too much time in overly-complex models. The surface obtained by prior specification (v), on the other hand, is an improvement over the greedy scheme. While it is difficult to tell from the perspective plot, the ridge or central hinge more closely follows the line found by (Cleveland and Fuentes 1996).

As a final test, we repeated the simulations from Chapter 9. We took as our test functions two piecewise planar surfaces, one that the greedy scheme can jump to in a single move (Model 1), and one that requires several moves (Model 3). The results are summarized in Table 10.8. In this case, the model averaged fits (iv) were better than both simulated annealing and the greedy procedure. The estimate built from the Poisson prior tends to spend too much time in larger models, leading to its slightly poorer MISE results, while the geometric prior extracts a heavy price for stepping off of the "true" model. (Unlike the smooth cases examined above, the extra degrees of freedom do not help the Poisson scheme.) One message from this suite of simulations, therefore, is that a posterior mean does not over-smooth edges, and in fact preserves them better than the greedy alternatives.

Fibonacci search. This finds the maximum of a unimodal function on an interval, $[a, b]$, by evaluating points placed according to a Fibonacci sequence, $F_N$. If there are $F_N$ points in the interval, only $N$ evaluations are needed. In the continuous case, we begin with some interval of uncertainty, $[a, b]$, and we reduce its length to $(b-a)/F_N$. The ratio, $g_n = F_(n-1)/F_n$, is the key to the placements.

Here is the method for the continuous case: 1.Initialization. Let $x = a + (1 - g_N)(b - a)$ and $y = a + g_N(b - a)$. Evaluate $f(x)$ and $f(y)$ and set $n = N$. 2.Iteration. If $f(x) < f(y)$, reduce the interval to $(x, b]$ (i.e., set $a = x$), decrement $n$ to $n - 1$, and set $x = y$ and $y = a + g_n(b - a)$. If $f(x) >= f(y)$, reduce the interval to $[a, y)$ (i.e., set $b = y$), decrement $n$ to $n - 1$, and set $y = x$ and $x = a + (1 - g_n)(b - a)$.

The Fibonacci search method minimizes the maximum number of evaluations needed to reduce the interval of uncertainty to within the prescribed

| distribution | version n | (i) MISE | (ii) | (iii) | (iv) | (v) |
|---|---|---|---|---|---|---|
| | | | ratio of MISE over (i) | | | |
| model 1 | 50 | 0.16 | 0.97 | 0.70 | 0.35 | 0.80 |
| model 1 | 200 | 0.04 | 0.82 | 0.95 | 0.52 | 0.62 |
| model 1 | 1000 | 0.01 | 0.63 | 0.72 | 0.76 | 0.40 |
| model 3 | 50 | 0.70 | 1.40 | 0.86 | 0.51 | 0.50 |
| model 3 | 200 | 0.17 | 0.85 | 0.63 | 0.27 | 0.30 |
| model 3 | 1000 | 0.03 | 0.34 | 0.45 | 0.21 | 0.20 |
| | average | 1.00 | 0.83 | 0.72 | 0.44 | 0.47 |

TABLE 10.8. Mean Integrated Squared Error (MISE) for two piecewise-planar test functions.

length. For example, it will reduce the length of a unit interval [0,1] to 1/10946 (= .00009136) with only 20 evaluations. In the case of a finite set, Fibonacci search finds the maximum value of a unimodal function on 10,946 points with only 20 evaluations, but this can be improved – see lattice search.

For very large $N$, the placement ratio ($g_N$) approaches the golden mean, and the method approaches the golden section search. Here is a comparison of interval reduction lengths for Fibonacci, golden section and dichotomous search methods. In each case N is the number of evaluations needed to reduce length of the interval of uncertainty to $1/F_N$. For example, with 20 evaluations dichotomous search reduces the interval of uncertainty to .0009765 of its original length (with separation value near 0). The most reduction comes from Fibonacci search, which is more than an order of magnitude better, at .0000914. Golden section is close (and gets closer as $N$ gets larger).

## 10.5 Other optimization methods

Charles. This material is not totally empty. I have notes that I have been taking from two texts and also Pitt man's thesis. It will be added before you get back from Hawaii.

### *10.5.1 Simulated annealing*

### *10.5.2 Genetic algorithms*

### *10.5.3 Gradient descent machines*

## 10.6 Combining models

Discuss a full Bayesian approach and illustrate the improvements with model averaging.