Computational Statistics (2000) 15:443–462

Computational Statistics

A Comparison of Regression Spline Smoothing Procedures

M. P. Wand

Department of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, MA 02115, USA

Summary

Regression spline smoothing involves modelling a regression function as a piecewise polynomial with a high number of pieces relative to the sample size. Because the number of possible models is so large, efficient strategies for choosing among them are required. In this paper we review approaches to this problem and compare them through a simulation study. For simplicity and conciseness we restrict attention to the univariate smoothing setting with Gaussian noise and the truncated polynomial regression spline basis.

Keywords: Bayesian variable selection, B-spline, Gibbs sampling, Nonparametric regression, Polynomial spline, Roughness penalty, Stepwise regression.

1 Introduction

The use of piecewise polynomials to model regression functions has a long history. See, for example, Hudson (1966), Fuller (1969), Studden and Van Arman (1969), Gallant and Fuller (1973), Wold (1971, 1974) and Smith (1979). However, their use in nonparametric regression, or *smoothing*, is relatively young. In this context the number of polynomial pieces and the locations of the join points, or *knots*, are arbitrary which permits a very large class of possible fits. The cost of this flexibility is a challenging model selection problem because of the high number of candidate models. This can be appreciated through the consideration of Figure 1. Figure 1a is a scatterplot of simulated regression data while Figure 1b shows a set of possible functions of the independent variable for use as predictors. If the response variable Y is regressed on all of these functions of the independent variable, x, using least squares then the resulting estimate is the one given in Figure 1c. Mathematically, this corresponds to fitting the model

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.1)_+ + \ldots + \beta_{19}(x_i - 0.9)_+$$

where $u_+ = \max\{0, u\}$. This fit is somewhat unsatisfactory because of the high degree of variability on the right hand side. A visually more pleasing answer is obtained by fitting the reduced model

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.1)_+ + \ldots + \beta_{15}(x_i - 0.5)_+ \qquad (1.1)$$

which leads to the fit shown in Figure 1d. Smoother fits could be obtained by using higher degree, such as those of the form $(x_i - \kappa_k)_+^p$ for p equal to 2 or 3. Other bases, such as B-spline bases (e.g. de Boor 1978), can be used instead. In nonparametric regression contexts this type of modelling is often called *regression spline smoothing*, and is also known as polynomial spline smoothing and B-spline smoothing.

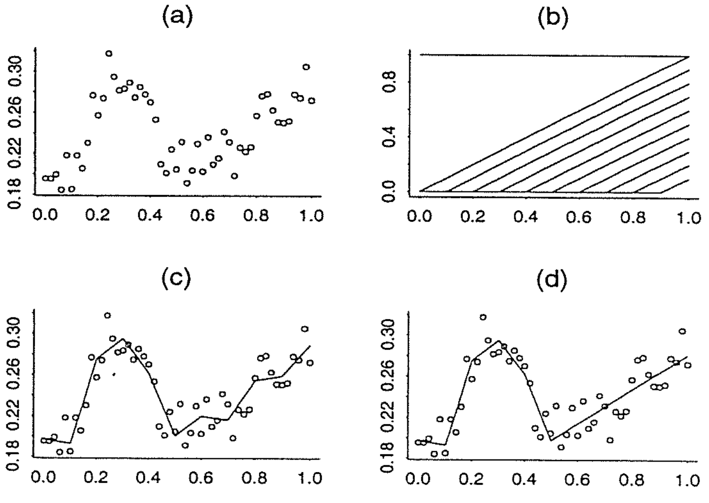

Figure 1: Portrayal of regression spline smoothing. (a) fictional regression data set, (b) set of possible basis functions, (c) fit based on all basis functions (d) fit based on basis functions with knots less than or equal to 0.5.

In the above example it is not overly difficult to arrive at a model close to (1.1) since it is clear that there is a lot more curvature on the left than on the right, and therefore there should be more polynomial pieces on the left hand side to capture this curvature. However, the automation of this idea for the smoothing of general scatterplots is a non-trivial problem. The main underlying reason for this is that a model with K knots has 2^{K+2} sub-models. Even for moderate values of K this results in a prohibitively large number of models to have to fit and compare through, say, a model selection criterion such as AIC or GCV. This has led to the recent development of strategies for simplifying this model selection problem. These strategies are now quite numerous and varied so a review and comparison seems in order.

The main purpose of this article is to compare current regression spline smoothing strategies. Theoretical comparison is very difficult for adaptive regression spline smoothing estimators because of their complicated implicit nature. Therefore, the comparison presented here will be done entirely through performance in a large simulation study. This necessarily entails restriction to a limited class of settings. However, we alleviate this restriction through the use of "families" of settings, where one factor is varied at a time. This format allows patterns to be more easily detected. Another shortcoming of a simulation-based comparison is that, for simplicity, quantification of the performance of a smoother is usually reduced to a single number corresponding to a conveniently chosen measure of error. While this tells us something about the comparative statistical performance of a set of smoothers, it ignores other attributes of a smoother such as simplicity and interpretability. Nevertheless, comparison of the statistical accuracies of regression smoothers is still of considerable interest. For conciseness, this study is restricted to the single predictor setting with independent Gaussian noise. This is only one special case of the very high number of settings to which regression spline smoothers have been extended. However, this setting is also the most fundamental, so a particular algorithm should perform well here to be a serious contender as a general principle for extension to other settings. Finally, for simplicity, we will concentrate on the truncated polynomial basis. It is expected that the results presented here are not very sensitive to the choice of basis.

In Section 2 we review the various regression spline procedures. Section 3 compares these procedures through a simulation study. Conclusions are given in Section 4.

2 Review of Methodology

While regression spline methodology has been extended in several directions (e.g. Stone, Hanson, Kooperberg and Truong 1997; Ruppert and Carroll, 1998), in this study we will concentrate on the one-dimensional nonparametric regression setting

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i$$

where m is a univariate function, v is a positive univariate function, the X_i are either deterministic real numbers or a random sample from a univariate distribution and the ε_i are independent $N(0,1)$ random variables.

In this normal errors context, a useful class of regression spline estimates of $m(x)$ is those of the form:

$$\hat{m}(x) = \sum_{j=0}^{p} b_j x^j + \sum_{k=1}^{K} b_{jk}(x - \kappa_k)_+^p. \tag{2.1}$$

Clearly $\hat{m}(x)$ is a linear combination of the set of functions,

$$1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_k)_+^p, \tag{2.2}$$

known as the *truncated power basis of degree p*. The κ_k lie within the range of the x_i and are called knots because they correspond to the join points of the piecewise polynomials that result from a linear combination of this basis. While this basis has the virtue of conceptual simplicity it does have shortcomings from a numerical viewpoint because of its near-multicollinear nature. For this reason, other bases, such as those based on B-splines (see e.g. de Boor 1978, Eilers and Marx 1996) and natural splines (see e.g. Green and Silverman 1994, pp.12–13), are often recommended instead. But because of its intuitiveness we will explain the methodology in terms of the truncated power basis.

In recent years three main types of approaches to fitting nonparametric regression splines have emerged: (1) stepwise selection, (2) Bayesian selection and (3) penalised shrinkage. The next three subsections describe these general approaches.

2.1 Stepwise selection

This idea appears to have been first proposed by Smith (1982) and has undergone refinement by C.J. Stone and coauthors since then. A comprehensive summary of their work can be found in Stone, Hanson, Kooperberg and Truong (1997). Their algorithm POLYMARS has one-dimensional nonparametric regression as a special case. It takes the following form:

1. Start with a subset of the full basis, called the *minimal basis.* Set this to be the current basis.

2. *Stepwise addition*

   Repeat until the current basis becomes the full basis:

   (a) Add to the current basis the basis function having the largest absolute Rao statistic among all those not in the current basis.

   (b) Fit the model with the new basis using least squares and record the GCV value of the fit.

3. *Stepwise deletion*

Repeat until the current basis becomes the minimal basis:

(a) Delete from the current basis the basis function having the smallest absolute Wald statistic among all those in the current basis.

(b) Fit the model with the new basis using least squares and record the GCV value of the fit.

The final estimate is that having the lowest GCV out of all models fit in the above process.

It remains to define the Rao and Wald statistics and the GCV criterion. It is easiest to do this in a general multiple linear regression context. Consider the model

$$Y = \beta_1 x_1 + \ldots + \beta_r x_r + \varepsilon$$

where Y is the $n \times 1$ vector of responses, each x_j is an $n \times 1$ vector containing values of the jth predictor variable (which could be an intercept) and the $n \times 1$ vector ε contains independent $N(0, \sigma^2)$ noise. Suppose that the current model is

$$Y = X_c \beta_c + \varepsilon$$

where the columns of X_c are a subset of the x_j. Then the Rao statistic corresponding to an x_k that is not in X_c is

$$R_k = \frac{x_k^T (I - H_c) Y}{\sqrt{x_k^T (I - H_c) x_k}}$$

where $H_c = X_c (X_c^T X_c)^{-1} X_c^T$ is the hat matrix associated with the current model. This criterion can be shown to correspond to a score-based hypothesis test and has the computational advantage that it does not require the model corresponding to x_k to be fitted. Such statistics are also used to construct added variable plots in regression diagnostics (see e.g. Weisberg, 1985).

The Wald statistic for deletion of the jth column of X_c is

$$W_j = \frac{\{(X_c^T X_c)^{-1} X_c^T Y\}_j}{\sqrt{\{(X_c^T X_c)^{-1}\}_{jj}}}$$

and is equivalent to the t-statistic attached to the least squares estimate of β_j.

Note that the variable selection nature of this method does not apply to the B-spline bases. If B-spline bases are used then deletion of a variable requires a new X_c matrix to be computed, corresponding to the vector space induced by the deleted knot.

The GCV criterion used by Stone *et al.* (1997) is of the form

$$\text{GCV} = n^{-1}\text{RSS} \Big/ \{1 - a(J-1)/n\}^2 \qquad (2.3)$$

where RSS is the residual sum of squares and J is the number of terms in the model and a is a parameter which these authors "typically set equal to 2.5".

He and Ng (1996) develop a stepwise knot selection algorithm in the quantile regression context. This can be viewed as a variation of the algorithm of Stone *et al.* (1997).

2.2 Bayesian selection

Smith and Kohn (1996) develop a Bayesian variable selection approach to choosing the regression spline knots. Once again, it is easiest to describe their approach in the general linear multiple regression context. Consider the model

$$Y = X\beta + \varepsilon$$

where Y is an $n \times 1$ vector of observations, X is an $n \times r$ design matrix, β is an $r \times 1$ vector of coefficients and $\varepsilon \sim N(0, \sigma^2 I)$. Define γ to be the $r \times 1$ vector of indicator variables with ith entry given by

$$\gamma_i = \begin{cases} 0, & \text{if } \beta_i = 0 \\ 1, & \text{if } \beta_i \neq 0 \end{cases}$$

For a given γ, β_γ is defined to be the vector of non-zero β_i's and X_γ is the matrix containing only those columns of X with a corresponding γ_i equal to one. In their empirical work, Smith and Kohn (1996) place the following priors

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

on the parameters and then show that the posterior distribution of γ given the data is

$$p(\gamma|Y) \propto (1+c)^{-\frac{1}{2}\sum_{i=1}^{r}\gamma_i} \left\{ Y^T Y - \frac{c}{c+1} Y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T Y \right\}^{-n/2}.$$

With respect to squared error loss, the Bayes estimate of m is the posterior mean

$$\begin{aligned} \hat{m}_{\text{Bayes}} &= XE(\beta|Y) = \sum_{\gamma \in \{0,1\}^r} X_\gamma E(\beta_\gamma|\gamma, Y) p(\gamma|Y) \\ &= \frac{c}{c+1} \sum_{\gamma \in \{0,1\}^r} X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T Y p(\gamma|Y). \end{aligned}$$

However, this estimator is impractical for general use because it involves computation of 2^r terms. Smith and Kohn (1996) propose the approximation

$$\hat{m}_{\text{Gibbs}} = \frac{c}{c+1} G^{-1} \sum_{g=1}^{G} X_{\gamma[g]} (X_{\gamma[g]}^T X_{\gamma[g]})^{-1} X_{\gamma[g]}^T Y$$

where $\gamma^{[1]}, \ldots, \gamma^{[G]}$ are a sample from $p(\gamma|Y)$ obtained using Gibbs sampling (Gelfand and Smith, 1990). The algorithm for accomplishing this task is:

```
1. Start with an initial value $\gamma^{[0]}$.
2. Warm-up period
   (a) Set $\gamma^{\text{current}} = \gamma^{[0]}$.
   (b) Repeat $W$ times:
          For $i = 1, \ldots, r$:
             Determine $p(\gamma_i|Y, \gamma_{j\neq i}^{\text{current}})$ and reassign $\gamma_i^{\text{current}}$ to be a value
             randomly generated from this distribution.
3. Generation of Gibbs iterates
       For $g = 1, \ldots, G$:
        i. For $i = 1, \ldots, r$:
             Determine $p(\gamma_i|Y, \gamma_{j\neq i}^{\text{current}})$ and reassign $\gamma_i^{\text{current}}$ to be a value
             randomly generated from this distribution.
       ii. Set $\gamma^{[g]} = \gamma^{\text{current}}$.
```

Smith and Kohn (1996) take $\gamma^{[0]}$ to be the K-tuple $(1, 0, 1, 0, 1, \ldots)$, $W = 100$ and $G = 1500$. The parameter c controls the diffuseness of the prior on the coefficients. Smith and Kohn (1996) report that the estimator is not very sensitive to the choice of this parameter. The default value for c is 100, assuming that all columns of X are standardised. Smith and Kohn (1996) also develop an efficient procedure for determining the required conditional distributions.

More recently Denison, Mallick and Smith (1998) have developed an alternative Bayesian approach to regression spline fitting. The main differences between their approach and the approach of Smith and Kohn (1996) is that the number of knots, and their locations, are not fixed in advance and instead are considered random components of a Bayesian model. The Markov Chain Monte Carlo (MCMC) strategy for selecting the model involves knots being added and deleted, and therefore a change in the dimension of the model. This leads to the deployment of reversible jump MCMC methods (Green, 1995).

2.3 Penalised shrinkage

Consider the full truncated polynomial regression spline model

$$m(x) = \sum_{j=0}^{p} \beta_j x^j + \sum_{k=1}^{K} \beta_{pk}(x - \kappa_k)_+^p.$$

The least squares criterion for fitting the full model is

$$S = \sum_{i=1}^{n} \{Y_i - m(x_i)\}^2.$$

For a large number of knots this fit tends to be too noisy so a simple remedy is to borrow an idea from classical spline smoothing methodology (Reinsch 1967) and replace S by a penalised criterion:

$$S + \alpha \sum_{k=1}^{K} \beta_{pk}^2. \tag{2.4}$$

This has the effect of shrinking the coefficients towards a smoother estimator, the extent to which is controlled by $\alpha \geq 0$. Roughness penalties based on other functions of the coefficients of m may also be used, but the one used in (2.4) leads to a particularly simple estimate.

The resulting family of estimates of $m = [m(x_1), \ldots, m(x_n)]^T$ can be easily shown to be

$$\hat{m}(\alpha) = X(X^T X + \alpha D)^{-1} X^T Y$$

where X is the design matrix for the full basis and D is the $(K + p + 1) \times (K + p + 1)$ diagonal matrix with zeroes in the first $p + 1$ diagonal positions and ones in the remaining K diagonal positions.

This approach to regression spline smoothing was introduced by O'Sullivan (1986, 1988). It has since been extended by Eilers and Marx (1996) and Ruppert and Carroll (1998). The first three of these references work with the B-spline basis rather than the truncated polynomial basis.

The advantage of the penalised shrinkage approach is that the model selection problem reduces to the choice of a single parameter, α. Any common model selection criterion can be used to choose α. Eilers and Marx (1996) use GCV while, in their examples, Ruppert and Carroll (1996) use Mallows' C_p. In this context C_p can be expressed as

$$C_p(\alpha) = \text{RSS}(\alpha) + 2\hat{\sigma}^2 \text{tr}\{X^T X (X^T X + \alpha D)^{-1}\}$$

where $\text{RSS}(\alpha)$ is the residual sum of squares of $\hat{m}(\alpha)$. An appropriate choice for $\hat{\sigma}^2$ is $\text{RSS}(0)/(n - K - p - 1)$, the estimated residual variance based on fitting the full model.

As a referee has pointed out, regression spline smoothing procedures fall into two categories:

1. Select a subset of basis functions and apply ordinary least squares to obtain the estimate.

2. Use all the basis functions, but do not use ordinary least squares to estimate the coefficients.

Stepwise selection is in the first category, while the Bayesian technique and penalized shrinkage is in the second category.

2.4 Other approaches

Several other proposals for knot selection exist. One of the earliest is an unpublished algorithm of Agarwal and Studden (1978) and discussed in Agarwal and Studden (1980). The algorithm is based on their asymptotic results and is somewhat complicated. Moreover, their numerical results suggest that it it not competitive with the more recent algorithms described in Sections 2.1–2.3. The TURBO algorithm of Friedman and Silverman (1989) and its subsequent generalisation, the MARS algorithm (Friedman, 1991), include knot selection for univariate scatterplot smoothing as a special case. However, TURBO and MARS are tailored for multivariate smoothing and their computational overhead requires restriction to piecewise linear basis functions for practical implementation. Luo and Wahba (1997) propose a procedure that can be considered as a combination of stepwise selection and penalized shrinkage. It involves performing forward stepwise regression to select basis functions, with a version of GCV used as a stopping criterion. A penalized regression is then performed on this basis.

Other approaches, not explicitly mentioned in the literature, might be considered for regression spline smoothing. Examples include all subsets regression approximations such as "Leaps and Bounds" (Furnival and Wilson, 1974) and the "Least Absolute Shrinkage and Selection Operator (LASSO)" of Tibshirani (1996).

3 Simulation Study

As we saw in the previous section, nonparametric approaches to regression spline fitting are now quite numerous and varied. There is clearly a strong imperative to compare their practical performance. In this section we describe a simulation study which aims to perform an objective comparison.

The knots for the largest possible model were chosen according to the rule

$$\kappa_k = (x_{(dk)} - x_{(dk+1)})/2, \quad k = 1, \dots, K$$

where $K = \lfloor n/d - 1 \rfloor$ and $d = \max\{4, \lfloor n/35 \rfloor\}$. This assignment ensures that there are at least d observations between each knot.

The truncated cubic basis ((2.2) with $p = 3$) was used throughout. Truncated polynomial bases are used in Smith and Kohn (1996) and Ruppert and Carroll (1998). Eilers and Marx (1996) work with B-spline bases, but for penalised shrinkage, the estimators are equivalent to those obtained with truncated polynomial bases of the same degree. Stone *et al.* (1997) recommend the natural cubic B-spline basis. As we mentioned in Section 2.1, stepwise selection estimators are different for B-spline bases. Nevertheless, we have stuck with the truncated cubic basis in this study so that all methods are directly comparable. Thus we aim to answer the question: given a large set of regression spline basis functions, how do each of the selection methods compare in terms of estimating the underlying regression function?

A fully automatic procedure was chosen from each of the three types of methods described in Section 2. Specifically, the procedures used in the study were:

1. The stepwise procedure described in Section 2.1 with minimal basis $1, x, x^2, x^3$ and Stone et al.'s GCV as given by (2.3) with their default of $a = 2.5$ (further justification for use of this default is given below).

2. The Bayesian variable selection procedure of Smith and Kohn with defaults as described in the second last paragraph of Section 2.2.

3. The penalised spline method with α chosen using C_p, as described in Section 2.3.

The settings for the simulation were devised in a family-wise fashion where, for each family, a different factor was tweaked. The factors are (1) noise level, (2) distribution of the design variable, (3) degree of spatial variation and (4) the variance function. Table 1 summarises the settings. These are based on settings developed by Professor Steve Marron and most of the credit for their development belongs to him. The last group of settings involve heteroskedastic errors despite the fact that each of selection methods are each based on the homoskedasticity assumption. Nevertheless, it is of interest to see how the methods perform when this assumption breaks down. The simulation involved 250 replications.

The error criterion of an estimate $\hat{m}$ of m was taken to be the root mean squared error:

$$\text{RMSE} = \sqrt{n^{-1}\sum_{i=1}^{n}\{\hat{m}(x_i) - m(x_i)\}^2}.$$

As mentioned above, the value of a used in the GCV measure for stepwise selection was the Stone *et al.* (1997) default of 2.5. However, as pointed out by Charles Kooperberg in private communication, this value was chosen because it gave reasonable answers for multivariate linear splines and without any consideration of RMSE performance. Since this study involves univariate cubic splines a preliminary investigation into the effect of a on RMSE was carried out. One hundred replications of each of the simulation settings were run with a values of 2, 2.5, 3 and 4. In terms of RMSE performance, none of these values was found to be dominant over the others although $a = 2.5$

had the best average ranking. For this reason use of the Stone *et al.* (1997) default of $a = 2.5$ is justified for this study.

factor	generic form	particular choices
noise level	$Y_{ij} = m(x_i) + \sigma_j \varepsilon_i$	$n = 200$, $m = m_{\rm NM}$, $\sigma_j = 0.02 + 0.04(j-1)^2$
random design	$Y_{ij} = m(X_{ji}) + \sigma \varepsilon_i$	$n = 200$, $m = m_{\rm NM}$, $\sigma = 0.1$, $X_{ji} = F_j^{-1}(X_i)$
spatial variation	$Y_{ij} = m_j(x_i) + \sigma \varepsilon_i$	$n = 400$, $\sigma = 0.2$, $m_j(x) = \sqrt{x(1-x)}$ $\times \sin\left\{\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right\}$
variance function	$Y_{ij} = m(x_i) + v_j(x_i)^{1/2}\varepsilon_i$	$n = 200$, $m = m_{\rm NM}$, $v_j(x) = [0.15\{1 + 0.4$ $\times(2j-7)(x-0.5)\}]^2$

$j = 1, \ldots, 6$

$x_i = (i - \frac{1}{2})/n$, X_i i.i.d. Uniform$[0,1]$, ε_i i.i.d. $N(0,1)$

$m_{\rm NM}(x) = 1.5\phi\left(\frac{x-0.35}{0.15}\right) - \phi\left(\frac{x-0.8}{0.04}\right)$,

$\phi(u) = (2\pi)^{-1/2}e^{-u^2/2}$, F_j is the Beta$\left(\frac{j+4}{5}, \frac{11-j}{5}\right)$ c.d.f.

Table 1: Settings used in simulation study

Figures 2–5 provide graphical summary of the results. Each pair of panels corresponds to (1) one replication of data and the true mean function for the setting and (2) boxplots of the $\log_{10}$(RMSE) for each method. Paired Wilcoxon tests were performed to determine whether the median RMSE's were significantly different. Procedures shared the same RMSE ranking when the test showed no difference at the (5/3)% level. Otherwise, separate rankings were assigned with "1" signifying the best performer and "3" the worst. These rankings are listed at the base of each set of boxplots.

The raw RMSE data and `S-PLUS/Fortran` code for computation of each of the estimators are available on request from the author (current e-mail address: `mwand@harvard.hsph.edu`).

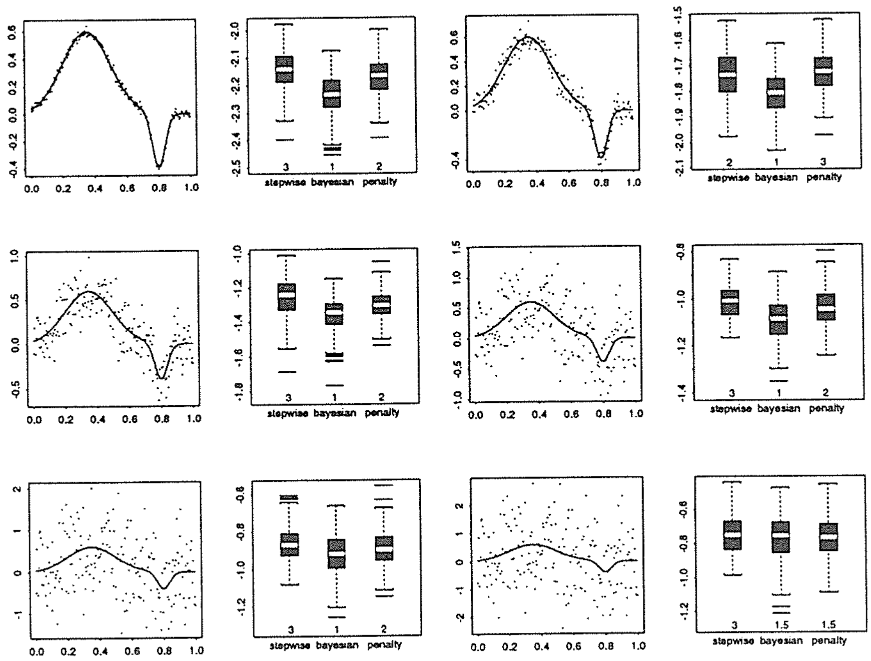

Figure 2: Varying the noise level. The left half of each pair of panels corresponds to one replication of data from the simulation study and the true mean function. The right half are boxplots of $\log_{10}$(RMSE) for each method along with paired Wilcoxon test rankings.

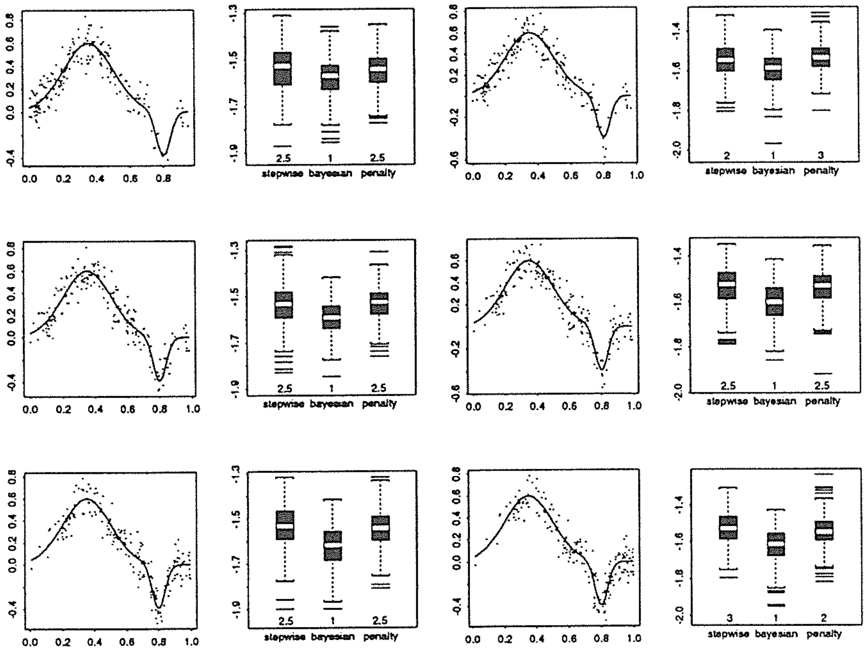

Figure 3: Varying a random design. The left half of each pair of panels corresponds to one replication of data from the simulation study and the true mean function. The right half are boxplots of $\log_{10}(\text{RMSE})$ for each method along with paired Wilcoxon test rankings.

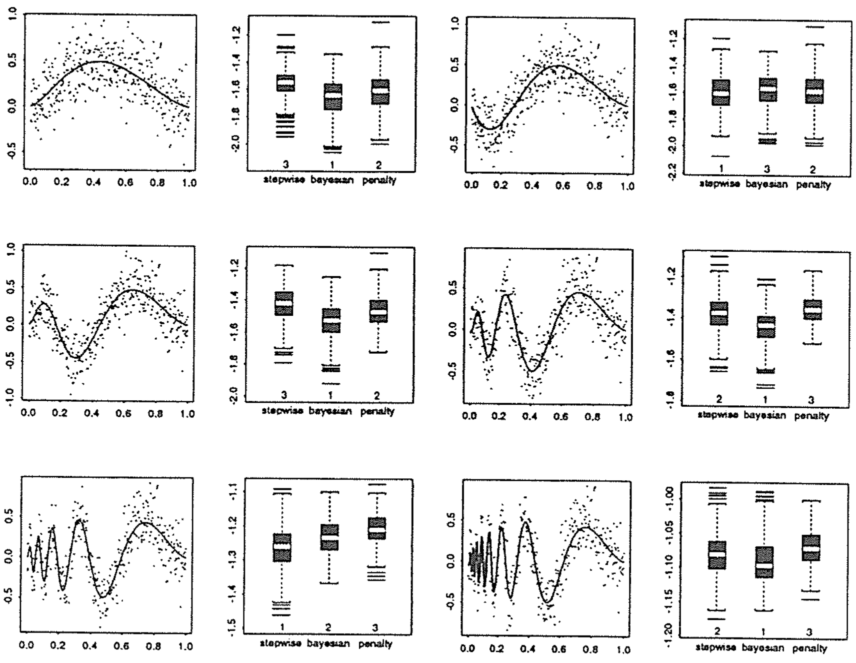

Figure 4: Varying the degree of spatial variation. The left half of each pair of panels corresponds to one replication of data from the simulation study and the true mean function. The right half are boxplots of $\log_{10}(\mathrm{RMSE})$ for each method along with paired Wilcoxon test rankings.

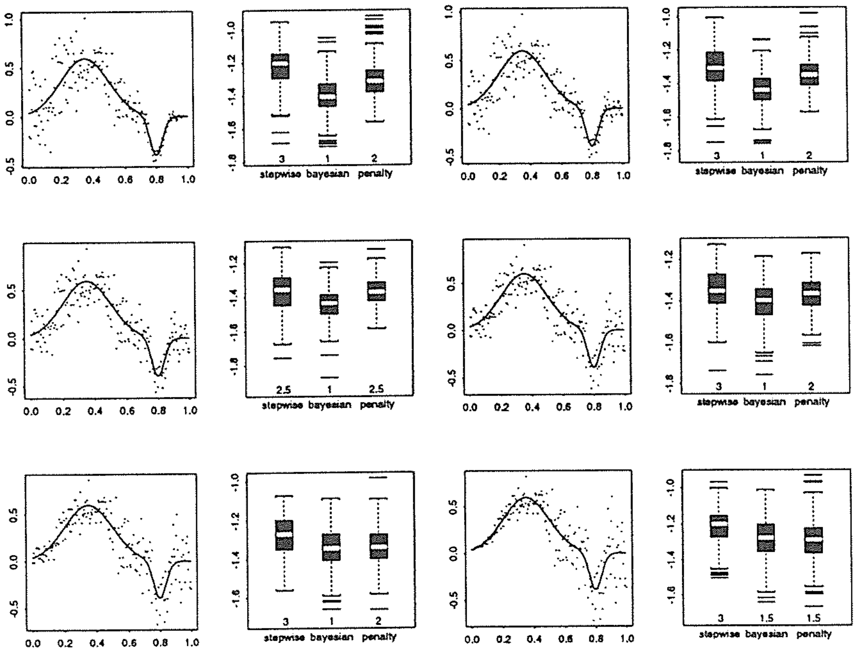

Figure 5: Varying the variance function. The left half of each pair of panels corresponds to one replication of data from the simulation study and the true mean function. The right half are boxplots of $\log_{10}$(RMSE) for each method along with paired Wilcoxon test rankings.

4 Conclusions

The rankings from the simulation study are summarised in Table 2.

	noise	design	spatial var.	variance func.	average
Stepwise	2.83	2.5	2.0	2.92	2.56
Bayesian	1.08	1.0	1.5	1.08	1.17
Penalty	2.08	2.5	2.5	2.00	2.27

Table 2: Average rankings from simulation study.

For the 24 different settings considered in the simulation study the average rankings were 2.56 for stepwise selection, 1.17 for Bayesian selection and 2.27 for penalised shrinkage. So by this measure the Bayesian procedure is clearly the best performer, with the other two performing roughly equally. Some patterns are apparent from Figures 2–5. In the case where the noise level was varied (Figure 2) we see that Bayesian selection is dominant for lower noise levels. But for higher the noise the methods each perform about the same. Penalised shrinkage performs relatively well for the normal mixture mean curve used in most of the settings but, as depicted in Figure 4, begins to suffer when there is more spatial variation.

The computational times for computation of each type of estimate are roughly equal, with none taking more than 3 seconds elapsed time throughout the whole study. So, in terms of this measure of performance the methods seem to be about the same.

To gain some appreciation for the types of problems that the methods can run into plots for the data and estimates at the 90th percentile of the RMSE distribution were obtained. Space restrictions do not permit showing these for all 24 settings, so a selection of 3 of the more revealing ones were chosen and are displayed in Figure 6. The settings correspond to the rows of Figure 6 and are (refer Table 1) (1) factor is design with $j = 6$, (2) factor is spatial variation with $j = 5$ and (3) factor is variance function with $j = 2$.

In the first and third rows of Figure 6 we see that stepwise selection leads to some spurious wiggles. For the first setting plots of the other two methods for the same data (not shown) lead to wiggles in the same place, but not quite as accentuated. For the third setting the other two methods are much more well-behaved near the left boundary, with the Bayesian selection estimator having no wiggles at all. The middle row shows a situation where stepwise selection performs very well, with Bayesian selection having trouble resolving the structure near the left boundary. The Achilles' heel of penalised shrinkage: not being able to adapt to the spatial variation in curvature, is apparent from the estimate depicted here in the third column.

It is possible that the problem of wiggliness in the tails, exhibited mainly by the stepwise approach, could be alleviated by using the *natural spline*

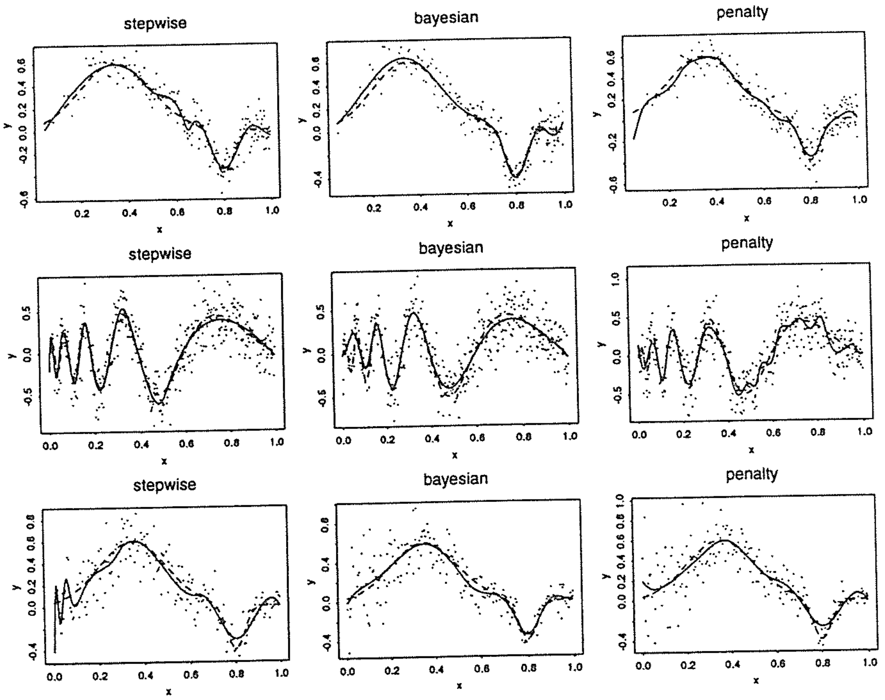

Figure 6: Data and estimates corresponding to the 90th percentile of the RMSE distribution from the simulation study. The dashed curve is the true mean while the solid curve is the estimate.The settings correspond to the rows.

basis rather than ordinary polynomial splines. This is because natural splines impose linearity constraints at the boundaries. Simulation results given in the rejoinder of Stone *et al.* (1997) show that the use of natural splines leads to substantial improvements. It would be interesting to see if substitution with the natural spline basis would change the rankings substantially.

The inferior performance of penalised shrinkage is due to the restrictiveness that comes with having the model selection controlled by a single parameter and comes as no surprise. The reason inferior performance of stepwise selection compared with Bayesian selection requires some deeper investigation. However, it is conjectured that this is due to (1) the Gibbs sampler traverses the model space in a more effective way that the stepwise procedure based on Rao and Wald statistics and (2) the Bayesian estimator is a weighted average of several regression spline fits, while the stepwise estimator is a single regression spline fit. On the other hand, Stone *et al.* have shown that stepwise selection extends naturally to a wide array of settings, particularly those of a non-Gaussian nature such as binary response data.

The extent to which Bayesian selection can be adapted successfully to these settings is unclear at this stage.

Acknowledgments

I am grateful to Mark Hansen, Xuming He, Roger Koenker, Robert Kohn, Charles Kooperberg, Steve Portnoy, Doug Simpson and Mike Smith for helpful comments. The final version of this paper also benefited from the helpful comments of two referees and an associate editor. Mike Smith's Bayesian regression code br was used in this study. The simulation settings are based on those originally developed by Steve Marron. Part of this research was carried out while the author enjoyed the hospitality of the Department of Statistics in the University of Illinois at Urbana-Champaign.

References

Agarwal, G.G. and Studden, W.J. (1978). An algorithm for selection of design and knots in the response curve estimation by spline functions. Technical Report No. 78–15, Purdue University.

Agarwal, G.G. and Studden, W.J. (1980). Asymptotic integrated mean square error using least squares and bias minimising splines. *Ann. Statist.*, **8**, 1307–1325.

de Boor, C. (1978) *A Practical Guide to Splines.* Springer, New York.

Denison, D.G.T, Mallick, B.K. and Smith, A.F.M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc., B*, **60**, 333–350.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.*, **89**, 89–121.

Friedman, J.H. and Silverman, B.W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**, 3–39.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Fuller, W.A. (1969). Grafting polynomials as approximating functions. *Austr. J. Agric. Econ.*, **13**, 35–46.

Furnival, G.M and Wilson. R.W. Jr. (1974). Regressions by leaps and bounds. *Technometrics*, **16**, 499–511.

Gallant, A.R. and Fuller, W.A. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *J. Amer. Statist. Assoc.*, **68**, 144–147.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, **85**, 398–409.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

He, X. and Ng, P. (1996). COBS: constrained smoothing made easy. Unpublished manuscript.

Hudson, D.J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Amer. Statist. Assoc.*, **61**, 1097–1129.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.*, **92**, 107–116.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci*, **1**, 505–527.

O'Sullivan, F. (1988). Fast computations of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.*, **9**, 363–379.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

Ruppert, D. and Carroll, R.J. (1998). Penalized regression splines. Unpublished manuscript.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 317–344.

Smith, P.L. (1979). Splines as a useful and convenient statistical tool. *Amer. Statistician.*, **33**, 57–62.

Smith, P.L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Ann. Statist.*, **25**, 1371–1425.

Studden, W.J. and Van Arman, D.J. (1969). Admissible designs for polynomial spline regression. *Ann. Math. Statist.*, **40**, 1557–1569.

Tibshirani, R. (1996). Regression shrinkage via the lasso. *J. Royal Statist. Soc., Ser. B*, **58**, 267–288.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

Wold, S. (1971). Analysis of kinetic data by means of spline functions. *Chemica Scripta*, **1**, 97–102.

Wold, S. (1974). Spline functions in data analysis. *Technometrics*, **16**, 1–11.