UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #122

September 1997

PENALIZED ESTIMATION OF FREE-KNOT SPLINES

Mary J. Lindstrom, PhD

Department of Biostatistics
and Medical Informatics

MADISON, WISCONSIN

Penalized Estimation of Free-Knot Splines

Mary J. Lindstrom

September, 1997

University of Wisconsin–Madison

Department of Biostatistics

Technical Report #122.

Key words: adaptive smoothing, least-squares splines, local smoothing, nonparametric regression, regression splines

[1]This research was supported in part by National Institutes of Health grant No. DC00820

Abstract

Polynomial splines are often used in statistical models for smooth response functions and densities. When the number and location of the knots are optimized, the approximating power of the spline is improved and the model is nonparametric with locally determined smoothness. However, finding the optimal knot locations is an historically difficult problem. We present a new estimation approach that improves computational properties by penalizing coalescing knots. The resulting estimator is easier to compute than the unpenalized estimates of knot positions, eliminates unnecessary "corners" in the fitted curve, and in simulation studies, shows no increase in the loss. A number of GCV and AIC type criteria for choosing the number of knots are evaluated via simulation.

1 Introduction

Splines are piecewise polynomial functions whose excellent approximation properties make them attractive regression functions (de Boor and Rice, 1968). For our purposes, a spline on the interval $[a, b]$ with knots $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, $a \leq \gamma_1 \leq \gamma_2 \leq, \ldots, \leq \gamma_k \leq b$ is a polynomial of order m (degree $m-1$) between each pair of adjacent knots and in the intervals $[a, \gamma_1]$ and $[\gamma_k, b]$. If the knots are distinct, then the polynomial pieces are joined so that the first $m-2$ derivatives of the spline are continuous.

Splines are considered parametric models when the unknown function is assumed to be a spline with the same order and number of knots as the spline model. More commonly, the truth is assumed to be well approximated by a spline. In this latter case, Agarwal and Studden (1980) established that if the number of (suitably distributed) knots increases with the number of data points at a rate of $n = O(k^{2m+1})$, then the asymptotic integrated mean squared distance from the estimated spline to an unknown function with m continuous derivatives is $O(n^{-2m/(2m+1)})$ which is the best rate possible as defined in Stone (1982).

A free-knot spline is a spline where the knot locations are considered parameters to be estimated. Freeing the knots improves the spline's approximating power (Burchard, 1974). In contrast to classical smoothing splines (Eubank, 1988; Wahba, 1990) and other penalized regression splines (e.g. O'Sullivan, 1986; Eilers and Marx, 1996), free-knot splines are locally adaptive smoothers. In contrast to adaptive regression splines (e.g. Friedman and Silverman, 1989; Kooperberg *et al.*, 1995) and hybrid smoothing splines (Luo and Wahba, 1997) the choice of knot locations in free-knot splines is a parameter estimation, rather than a predictor subset selection, problem.

The advantages of free over fixed knots comes with at least two major disadvantages. The first is that finding the optimal locations for the knots is very difficult. The knots enter the model nonlinearly and there are many local optima in the residual sum-of-squares function, many of which correspond to knot vectors with replicate knots. If, in addition, the number of knots is estimated, the difficulties multiply. Free-knot splines are often mentioned in discussions of nonparametric regression approaches but they are usually quickly dismissed as computationally intractable.

The second disadvantage of free knots (which is related to the first) is that the optimal knot

vector often include replicate knots which allow non-smooth behavior of the predicted curve. Duplicate knots allow a discontinuous second derivative, triplicate knots a discontinuous first derivative and 4 identical knots in one place allows a discontinuity in the fitted curve itself. If an assumption of smoothness for the true underlying function is warranted, we might wish to exclude solutions with replicate knots, however, there is no guarantee that a solution without replicate knots exists and, even if one does, it may be quite poor compared to solutions with replicate knots.

In order to improve the computational and estimation properties of free-knot splines while retaining their adaptive smoothing properties, we propose an estimator for the knots and coefficients defined as the optimizer of a slightly penalized residual sum-of-squares. The proposed penalty is a measure of the closeness of the knots to one another and to the endpoints a and b. Knot vectors with replicate or nearly replicate knots are penalized most severely with the penalty decreasing as the knots move towards equal spacing. This scheme improves the computational properties of the optimization problem while allowing knots to move relatively freely as long as they don't get too close to one another. The penalty can also be viewed as adding a small amount of global smoothing to the locally adaptive free knot splines.

In Section 2 we discuss existing approaches to finding optimal knot locations. In Section 3 we present the penalized estimator of knot locations and choose a penalty function. We discuss methods for choosing the number of knots and the estimation of the error variance in Section 4. The statistical and computational properties of the penalized estimator are presented in Section 5 and in Section 6 we discuss the results. Some computational details are given in an Appendix.

2 Optimizing the Knot Locations

2.1 Computational difficulties

There are a number of interrelated reasons why finding the locations of the knots is so difficult. First, there are usually many local optima in the residual sum-of-squares function necessitating excellent starting values in order to find the global optimum. However, good starting values are particularly difficult to find. In most nonlinear regression models, the choice of starting values

is aided by the roles that the parameters play, e.g., asymptotes, intercepts, and inflection points. As long as there are sufficient data to estimate the parameters and the role of the parameters is well understood, it is usually possible to find data-based starting values that will converge to the global optimum. In contrast, while knot positions do play an identifiable role (they allow a jump discontinuity in the $m-1$st derivative), it is difficult or impossible to identify these discontinuities by inspection of the data.

Over-parameterization is also very difficult to diagnose in a free-knot spline. In a typical nonlinear model, over-parameterization is usually characterized by a lack of convergence with one or more parameters going off to infinity. In free-knot splines, over-parameterization (too many knots) produces no such red flags. Knots often coalesce, causing problems for the optimization algorithm, but this is not necessarily a sign that there are too many knots. It just means that the algorithm has either found a minimum (local or global) with replicate knots or is bogged down in a flat area. Knot vectors with replicate knots are not singularities in the parameter space. As long as the support of each spline basis function contains at least one design point, the residual sum-of-squares function is continuous and the corresponding spline is well defined. However, replicate knots often allow the fitted curve to follow unusual data points and may not be desirable in cases where the response is knows to be smooth.

The existence of multiple optima in the objective function is related to the symmetry induced by the exchangeability of the knot parameters (Jupp, 1978). For example, in a problem with two free-knots, the objective function is symmetric along any normal to the line defined by equal knots. Thus the derivative along the normal at the intersection to the equal-knot line is zero. This property (called "lethargy" by Jupp (1978)) leads to many stationary points and ridges along lines or planes in the parameter space where two or more knots coincide. See Section 5.1 for a formal definition of lethargy.

2.2 Existing approaches to optimizing knot locations

There are two general approaches to finding the optimal (or approximately optimal) knot placement. Either the problem is modified to improve the performance of general purpose, derivative

based, optimization algorithms or a special purpose optimization algorithm is developed to attack the problem.

The approach of Jupp (1978) is the major (possibly only) contribution in the first category. He defines a transformation of the knots that removes the redundancy in the knot vector by mapping the simplex of ordered knot vectors to all of R^k where the boundaries of the simplex, corresponding to replicate knots, are mapped to plus and minus infinity. This transformation ameliorates the lethargy problem describe above. The "Jupp parameters" are $\log(h_{i+1}/h_i)$, $i = 1, \ldots, k$, where γ is the sorted knot vector, $[a, b]$ is the estimation interval,

$$h_i = (\gamma_i - \gamma_{i-1})/(b - a) \quad i = 1, \ldots, k+1 \tag{1}$$

k is the number of knots, $\gamma_0 = a$, and $\gamma_{k+1} = b$. Jupp's definition of h_i does not include division by $b - a$ but this normalization of the intra-knot spacings so that they sum to one does not alter the transformation and will be useful to us. The Jupp transformation is related to the various log-ratio transformations described in Aitchison (1986).

When used with a Levenberg-Marquardt adjustment, the Jupp transformation improves the behavior of Newton-Raphson algorithms and increases the chance of finding the global optimum from any particular starting value. However, in order to insure that the global optimum is found, many runs with different starting values are still required. In addition, false convergence and failure to progress in the neighborhood of knot vectors with replicate knots still occur frequently. If plus and minus infinity are considered potential parameter values, the number of local optima is not reduced by this transformation.

There have been many proposals for special purpose algorithms to find nearly optimal knot locations. Most such algorithms are stepwise knot insertion and deletion algorithms and can also be viewed as methods for finding good starting values since the resulting knot vector can be fed into a derivative based optimization algorithm to find the near by, hopefully global, optimum. These methods come in two main varieties.

In the statistical literature, the fact that the addition or deletion of a knot can be viewed as the addition of a regressor in the "plus function" basis (Eubank, 1988, page 355) has been used

to develop adaptations of standard stepwise regression algorithms to find nearly optimal knots (Kooperberg *et al.*, 1995). Usually the set of possible knot locations is taken to be some subset of the design points and the addition or deletion of a knot is determined using an information or cross-validation criterion. There is an older tradition in the numerical analysis literature of knot insertion and deletion algorithms for cubic splines. Many of these attempt to place knots where the absolute value of the fourth derivative of the unknown function is largest (de Boor, 1978; Goldman and Lyche, 1993). This approach is motivated by the fact that cubic polynomial splines have piecewise constant 3rd derivative and more knots are necessary to follow a rapidly changing 3rd derivative (large 4th derivative) than a slowly changing one. It also has a basis in statistical asymptotic theory. Agarwal and Studden (1980) point out that minimum asymptotic integrated mean squared error in a fixed knot spline is achieved when the the density of the knots and design points are proportional to the absolute value of the fourth derivative of the true, unknown, function. While this approach may be practical when there is excellent information on the unknown function (typically assumed in the numerical analysis literature), in the statistical problem, the function is unknown and estimating high order derivatives from the data is usually much more difficult than the original problem of estimating the function itself.

3 Penalized Estimation

We propose a new estimate $\hat{\gamma}_J$ for the knot locations which improves the computational properties of free-knot spline estimation. This estimator is defined as the minimizer of the residual sum-of-squares times a penalty $J(\gamma)$. That is, $\hat{\gamma}_J$ minimizes $R(\gamma, \boldsymbol{c})J(\gamma)$ where $1 \leq J(\gamma) < \infty$, $J(\gamma)$ increases as two or more knots coalesce, and $R(\gamma, \boldsymbol{c})$ is the residual sum-of-squares for a spline with knots γ and spline coefficients $\boldsymbol{c}$. (The spline coefficients are the multipliers of the spline basis functions. See the Section A.1 in the Appendix for basis function details.) Equivalently, we can optimize $R(\gamma, \hat{\boldsymbol{c}}(\gamma))J(\gamma)$ where $\hat{\boldsymbol{c}}(\gamma)$ is the linear least squares estimate of the spline coefficients given the knot locations. For convenience, we will sometimes write $R(\gamma)$ for $R(\gamma, \hat{\boldsymbol{c}}(\gamma))$, $R(\boldsymbol{\theta})$ for $R(\gamma, \boldsymbol{c})$, and $J(\boldsymbol{\theta})$ for $J(\gamma)$ where $\boldsymbol{\theta} = [\gamma^T, \boldsymbol{c}^T]^T$.

Penalizing knot vectors with replicate and nearly replicate knots is intuitively appealing since

we know that (due to lethargy) many of the local optima have replicate knots. In addition, if the fitted spline is used to obtain estimates of the derivatives of the unknown function, replicate knots (which allow step discontinuities in the lower order derivatives) are undesirable. However, sometimes the global optimum may contain replicate knots. When nearly replicate knots are needed to fit the data (the objective function is sharply convex near the optimum) then a properly penalized solution will be close to the unpenalized solution. Likewise, when the replicate knots in the unpenalized solution are not needed (the objective function is nearly flat near the optimum or the replicate knots are adding unwanted non-smooth features to the fit then the penalized optimum will be further away from the unpenalized optimum but the fit should not be substantially degraded and may be improved. (See Figure 6 in Section 5.1 for examples of such non-smooth behavior.)

We use a multiplicative penalty rather than the more traditional additive penalty so that the penalty defines a percentage increase in $R(\boldsymbol{\theta})$. This scheme means that the penalty is scale-free, allowing it to be fixed *a priori* (possibly as a function of the order of the spline, the number of knots, and/or the number of observations). Also, since the penalty induces a percent increase in $R(\boldsymbol{\theta})$ rather than an absolute one, knot vectors with smaller residual sum-of-squares are penalized proportionally less.

3.1 The form of the penalty

As mentioned above, the goal in constructing a penalty is to penalize knots vectors that contain replicate and nearly replicate knots (including interior knots close to the ends of the estimation interval). Three nearly replicate knots should be penalized more than two; the penalty should be relatively small for knot vectors with no nearly replicate knots; and tuning parameters should be included to vary the abruptness and location of the switch from small to large penalty. For theoretical and practical reasons we only consider penalties with at least three continuous derivatives with respect to γ and which depend only on the knots (not the spline coefficients). Penalties that depend only on the knot locations do not affect the conditional linearity of the spline coefficients, an important property for efficient computation. The following penalty, a normalized inverse of

the product of the intervals between the augmented knots, has these properties. We define

$$J_1(\gamma, \alpha, \beta) = \alpha \left[\left(\frac{\left(\frac{1}{k+1}\right)^{(k+1)}}{\prod_{i=1}^{k+1} h_i} \right)^{\beta} - 1 \right] + 1$$

where h_i is defined in Equation 1. This penalty achieves its minimum of 1 when the knots are equally spaced and increases to infinity as two or more knots coalesce. If the spacing of the design points in the interval $[a, b]$ is very uneven, we might wish to define h_i as the proportion of data points between $\gamma_i - \gamma_{i-1}$ so that the minimum penalty would be achieved when the knots corresponded to the quantiles of the design points. However, this penalty would not be differentiable with respect to the knots unless a smooth density function for the locations of the observations were estimated and distance measured relative to this measure. We have not pursued this option further.

The shape of the penalty $J_1(\gamma, \alpha, \beta)$ as a function of γ is governed by β. However a change in β also changes the value of the penalty dramatically for a fixed γ. In order to separate out (as much as possible) the shape effects of β from the scaling effects, we redefine α so that penalty evaluated at a fixed knot vector $\gamma^0(k)$ is equal to a constant p for all values of β. That is, we define

$$J_2(\gamma, \beta, p) = J_1\left(\gamma, \frac{p-1}{J_1(\gamma^0(k), 1, \beta) - 1}, \beta\right) = \frac{p-1}{J_1(\gamma^0(k), 1, \beta) - 1} \left[J_1(\gamma, 1, \beta) - 1\right] + 1$$

so that $J_2(\gamma^0(k), \beta, p) = p$ for all β. In simulation studies (not shown) we have found that over a range of test functions, sample sizes, and error variances, smaller values of β typically performed better. Thus, we use the limiting value of J_2 as β goes to zero as our penalty:

$$J(\gamma, p) = \lim_{\beta \to 0} J_2(\gamma, p, \beta) = \frac{p-1}{\log\left(J_1(\gamma^0(k), 1, 1)\right)} \log\left(J_1(\gamma, 1, 1)\right) + 1$$

3.2 Choice of $\gamma^0(k)$ and p

If the parameter p is allowed to vary freely, fixing $\gamma^0(k)$ does not limit the flexibility of $J(\gamma, p)$. However, as discussed above, we hope to fix the parameters in the penalty and so the dependence

of $\gamma^0(k)$ on k should be carefully considered. Note that the knot vector γ^0 can also depend on other known quantities such as the number of observations but we only consider dependence on the number of knots.

We set $\gamma^0(k)$ to be a minimum penalty knot vector with minimum spacing between knots $h_{[1]} = d/(k+1)$, $d > 1$. For penalties which depend on γ only through $1/(\prod h_i)$, a knot vector with minimum inter-knot spacing of $d/(k+1)$ and minimum penalty must have all other inter-knot spacings equal to $(1 - d/(k+1))/k$. The choice of d is not critical since the shape of $\log(J_1(\gamma^0(k), 1, 1))$ as a function of k is relatively invariant to changes in d (we use $d = 0.04$). This choice of $\gamma^0(k)$ implies that as k increases, J will also increase for knot vectors which stray from even spacing.

Once $\gamma^0(k)$ is fixed the only parameter left in the penalty is p. We prefer to fix p rather than choose it using a data based criterion to avoid increasing the computational burden. A value of p that consistently minimizes the loss for a variety of "true" functions would be a good candidate for a fixed value of p (where the loss is defined to be the square root of the mean of the squared distance from the truth to the estimated curve at the design points).

In order to evaluate the behavior of penalized estimation for different values of p we use a variable-truth simulations (described in detail in Section A.2 of the Appendix). In short, 200 different splines are created by simulating knot locations and spline coefficients. For each of these splines representing the "truth", a penalized free-knot spline is fit with a range of knot vector lengths and a range of values of p. For each k and p, the knot vector which minimizes the penalized residual sum of squares is recorded. A variety of objective functions are then used to choose between the knot vectors of various lengths for each p. Here we only consider the knot vectors where k is chosen by minimizing the loss. Of course the loss is not a viable method for choosing k in real data problems since it is not available unless the truth is known. We use it here to avoid confounding the effects of the method for choosing k with the choice of p. Data driven methods for choosing k are discussed in Section 4. Four combinations of k_G (the number of knots in the random spline defining the truth) and σ are presented. Larger k_G corresponds to a more complex, less smooth true function.

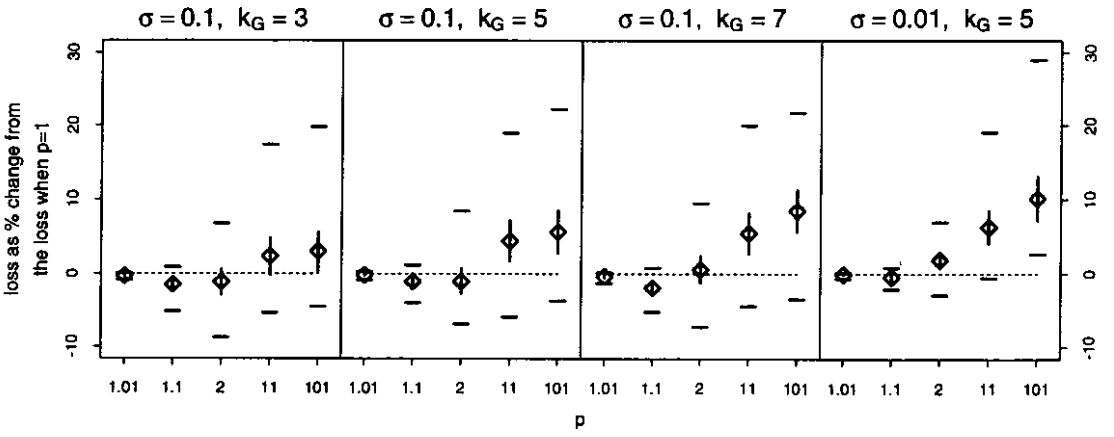

Figure 1: Choosing p: results from variable-truth simulations (see Section A.2 of the Appendix for details) for various values of the true σ and the generating k (k_G). The truth is a random spline with k_G knots and the errors are simulated from a $\mathcal{N}(0, \sigma^2)$ distribution. The y axis is the loss as the percent change from the loss when when $p = 1$, i.e., when there is no penalty. The diamonds are centered horizontally at the median values, the vertical lines are 95% confidence intervals for the medians, the short horizontal lines delineate the interquartile ranges.

Figure 1 shows the percent change in the loss for a variety of values of p compared to $p = 1$ (no penalty). At $p = 1.01$ the loss is essentially unchanged from $p = 1$. Increasing p to 1.1 decreases the loss slightly for all combinations of k_G and σ. At $p = 2$ the loss starts to increase for 3 out of 4 of the simulations and the largest increase of the loss occurs at $p = 101$, which corresponds to nearly fixed, equally-spaced knots. Since the larger p is the more tractable the optimization problem, we have chosen $p = 1.1$, the largest p which does not increase the loss over $p = 1$.

One obvious question about the effect of our choice of p is how much smoothing is induced when the underlying truth includes a sharp corner. In the top row of plots in Figure 2 both the penalized and unpenalized splines fit a sharp corner well when there is very little noise in the data. The bottom row of plots (where the signal to noise ratio is somewhat smaller) show some smoothing of the sharp corner in the penalized fit. Thus, when the data clearly indicate a sharp corner the penalized smoothing spline includes knots which are very close together. As soon as there is enough noise in the data to allow for a smooth interpretation the corner is smoothed over but the fit to the data is still excellent.

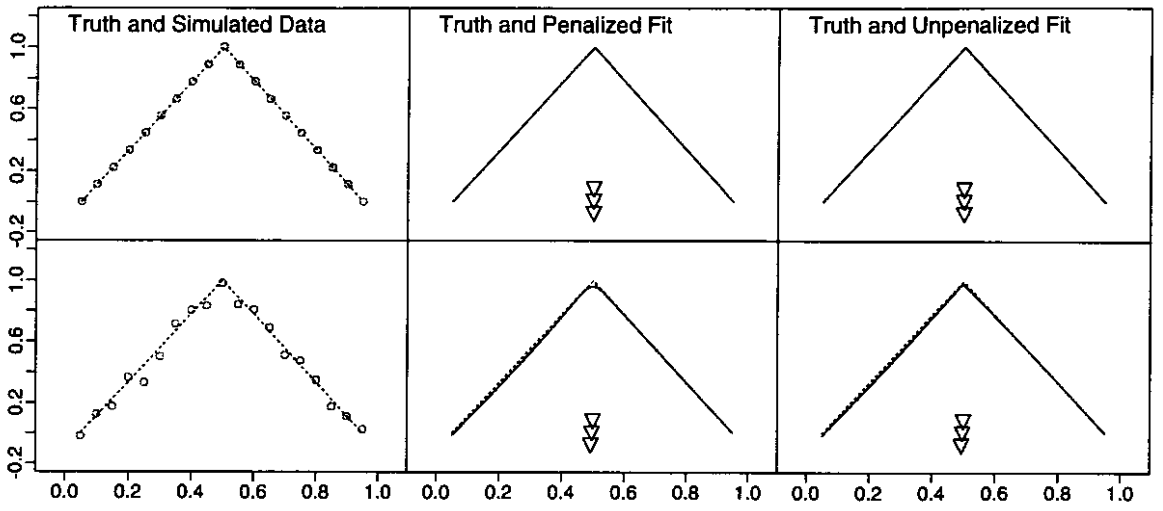

Figure 2: Simulated data with penalized ($p = 1.1$) and unpenalized free-knot spine fits. The dotted line in all panels is the true function. The solid lines are the fitted curves. The solid and dotted lines overlap for the top center and right panels. The errors for the top row of plots are simulated from a $\mathcal{N}(0, 0.001)$ distribution and for the bottom row, $\mathcal{N}(0, 0.05)$ errors are used. The inverted triangles indicate the locations of the estimated knots

4 Choosing k and Estimating σ^2

4.1 Choosing k: unpenalized estimator

We first consider the generalized cross validation (GCV) criterion for choosing the number of knots for unpenalized free-knot spline regression. Eubank (1988) proposes a GCV criterion for choosing the number and location of the knots in a free-knot spline based on a subset selection GCV criterion given by Wahba (1977):

$$G_{SS}(\gamma) = \frac{R(\gamma, \hat{c}(\gamma))}{[\text{trace}(I - A_{c|\gamma})]^2} = \frac{R(\gamma, \hat{c}(\gamma))}{(n - n_b)^2}$$

where γ is a knot vector of length k, n is the number of data points, n_b is the number of spline basis functions ($k+4$ for a standard spline and $k+2$ for a natural spline), and $A_{c|\gamma}$ is the influence matrix for the linear least squares estimation of c given fixed γ. This approach ignores the loss of error degrees of freedom due to the estimation of the knots and substantially over estimates the number of knots needed (see Figure 3).

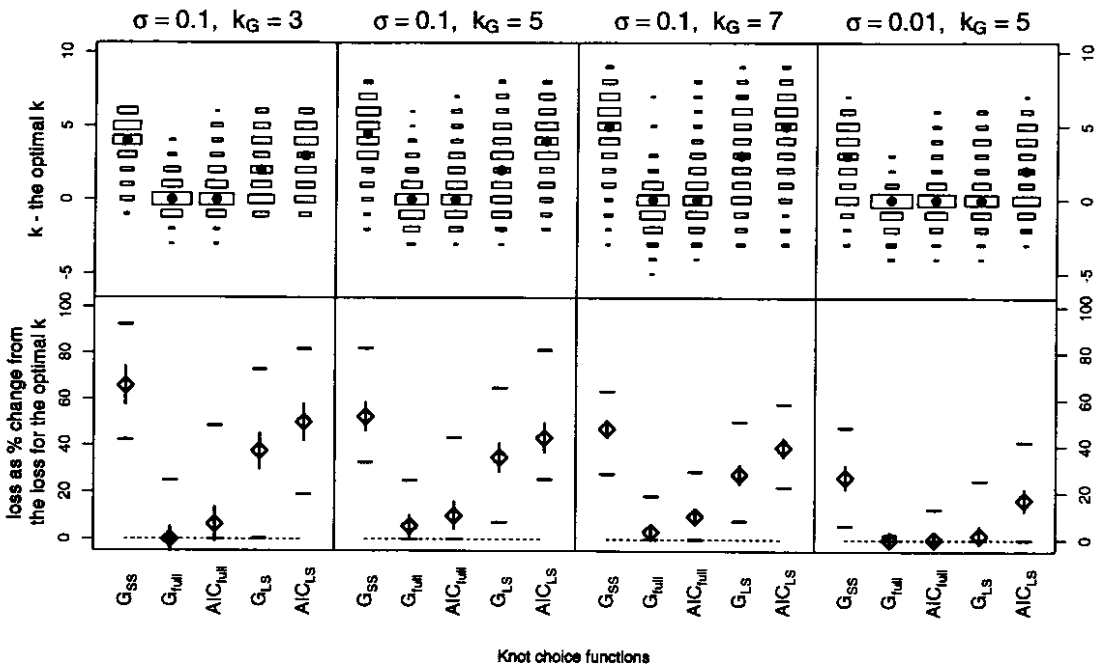

Figure 3: Choosing k, no penalty: Results from variable-truth simulations (see Section A.2 of the Appendix for simulation details). The **top panels** show the difference between the k chosen by minimizing the objective functions shown on the x-axis and the optimal k chosen by minimizing the loss. The area of the boxes is proportional to the frequency of occurrence in the 200 simulation replications. A dot indicates the median difference. The **bottom panels** show the loss when k is chosen using the objective functions on the x-axis as the percent change from the loss when using the optimal k chosen by minimizing the loss. See the legend for Figure 1 for an explanation of the plotting symbols.

An alternative approach to defining a GCV criteria is to generalize the one used for least squares regression:

$$G_{LS}(\boldsymbol{\theta}) = \frac{R(\boldsymbol{\theta})}{[\text{trace}(\boldsymbol{I} - \boldsymbol{A}_{\boldsymbol{\theta}})]^2}$$

where $\boldsymbol{A}_{\boldsymbol{\theta}}$ is the influence matrix and $\text{trace}(\boldsymbol{I} - \boldsymbol{A}_{\boldsymbol{\theta}})$ is the number of degrees of freedom for error. For linear least squares estimation $\boldsymbol{A}_{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Since free-knot splines are nonlinear, a GCV criterion which takes into account the estimation of the knots requires an expression for the influence matrix in nonlinear regression.

We derive such a matrix using a method that generalizes to the penalized estimation case (see Section 4.2). This derivation is based on the fact that the influence matrix in linear regression is

the derivative of $\hat{\boldsymbol{y}}$ with respect to $\boldsymbol{y}$. That is, $A_{i,j}$ is the change in $\hat{y}_i$ due to a change in y_j. For nonlinear regression of the form $\boldsymbol{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \text{error}$, we use the chain rule to write the influence matrix as:

$$\boldsymbol{A}_{\boldsymbol{\theta}} = \frac{\partial \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{y}} = \frac{\partial \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \frac{\partial \hat{\boldsymbol{\theta}}}{\partial \boldsymbol{y}} \tag{2}$$

where $\hat{\boldsymbol{\theta}}$ is the nonlinear least squares estimate of $\boldsymbol{\theta}$. If $\boldsymbol{\eta}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ this expression reduces to the linear least squares influence matrix described above. If $\boldsymbol{\eta}(\boldsymbol{\theta})$ is nonlinear in $\boldsymbol{\theta}$ the expression for $\boldsymbol{A}_{\boldsymbol{\theta}}$ cannot be evaluated directly since there is no closed form expression for $\hat{\boldsymbol{\theta}}$ as a function of $\boldsymbol{y}$. However, following a devise from Pregibon (1981), we can derive an approximate influence matrix. Let $\boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{y})$ be the Gauss-Newton increment for the nonlinear least squares estimation of $\boldsymbol{\theta}$, i.e., $\boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{y}) = (\hat{\boldsymbol{X}}_{\boldsymbol{\theta}}^T \hat{\boldsymbol{X}}_{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{X}}_{\boldsymbol{\theta}}^T (\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}))$ where $\hat{\boldsymbol{X}} = \partial \boldsymbol{\eta}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \big|_{\hat{\boldsymbol{\theta}}}$. Then, for a fixed point $\boldsymbol{\theta}_0$ near $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y})$ and

$$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \boldsymbol{y}} \approx \frac{\partial (\boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y}))}{\partial \boldsymbol{y}} = (\boldsymbol{X}_0^T \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^T \quad \text{where} \quad \boldsymbol{X}_0 = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_0}$$

As $\boldsymbol{\theta}_0$ moves toward $\hat{\boldsymbol{\theta}}$, the approximation becomes more accurate. Thus we define

$$\boldsymbol{A}_{\boldsymbol{\theta}} = \lim_{\boldsymbol{\theta}_0 \to \hat{\boldsymbol{\theta}}} \hat{\boldsymbol{X}} (\boldsymbol{X}_0^T \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^T = \hat{\boldsymbol{X}} (\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^T \tag{3}$$

This expression for $A_{\boldsymbol{\theta}}$ agrees with the result of Pregibon (1981), has trace equal to the number of parameters in the model when $\hat{\boldsymbol{X}}$ is of full rank, and reduces to the linear least squares influence matrix in the linear case.

In free-knot spline nonlinear regression, $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2]$ where $\hat{\boldsymbol{X}}_1$ is the matrix of spline basis functions evaluated at $\hat{\gamma}$ and the design, and $\hat{\boldsymbol{X}}_2 = (\partial \hat{\boldsymbol{X}}_1 / \partial \hat{\gamma})\, \boldsymbol{c}$ where column i of $(\partial \hat{\boldsymbol{X}}_1 / \partial \hat{\gamma})\, \boldsymbol{c}$ is equal to $(\partial \hat{\boldsymbol{X}}_1 / \partial \hat{\gamma}_i)\, \boldsymbol{c}$. If $\hat{X}$ is of full rank then $\text{trace}(\boldsymbol{I} - \boldsymbol{A}_{\boldsymbol{\theta}}) = n - (n_b + k)$. Thus, a second option for defining a nonlinear GCV criterion is

$$G_{\text{full}}(\boldsymbol{\theta}) = R(\boldsymbol{\theta}) / (n - n_b - k)^2 \qquad .$$

Figure 3 shows the results for various criteria of choosing k compared to choosing k by minimizing

the (normally unknown) loss. These include G_{SS}, G_{full}, $AIC_{\text{full}} = R(\boldsymbol{\theta})(n+(n_b+k))/(n-(n_b+k)-1)$, G_{LS}, and $AIC_{LS} = R(\boldsymbol{\theta})(n+\text{trace}(\boldsymbol{A}_{\boldsymbol{\theta}}))/(n-\text{trace}(\boldsymbol{A}_{\boldsymbol{\theta}})-1)$. The values of k chosen using G_{full} and G_{LS} consistently come the closest to the k chosen by minimizing the loss. Correspondingly the loss obtained using these criteria is closest to the minimum loss. We recommend choosing k via G_{full} in the unpenalized case because it results in more accurate estimates of σ^2 than does G_{LS} (results not shown).

4.2 Choosing k: penalized estimator

Just as in the unpenalized case, we start by developing an estimate for the degrees of freedom for signal. One option is to use $\text{trace}(A_{\boldsymbol{\theta}})$ (Equation 3). However, this estimate does not take into account the loss in degrees of freedom for signal due to the penalty. In the extreme case as $p \to \infty$ the degrees of freedom should approach n_b since large p forces the knots toward fixed and equal spacing, eliminating k degrees of freedom in the model. Alternatively, we can generalize the expression for the influence matrix (Equation 2) to penalized free-knot spline estimation by generalizing the nonlinear least squares case to general Newton-Raphson minimization. The form of the general Newton-Raphson increment is $\delta = -\boldsymbol{\Omega}^{-1}\boldsymbol{\omega}$ where $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}$ are the gradient and Hessian of the objective function with respect to c and γ. This change effects the $\partial\hat{\boldsymbol{\theta}}/\partial\boldsymbol{y}$ term in Equation 2 which can be expanded as

$$\frac{\partial\hat{\boldsymbol{\theta}}}{\partial\boldsymbol{y}} \approx \frac{\partial(\boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y}))}{\partial\boldsymbol{y}} = -\frac{\partial\boldsymbol{\Omega}^{-1}}{\partial\boldsymbol{y}}\boldsymbol{\omega} - \boldsymbol{\Omega}^{-1}\frac{\partial\boldsymbol{\omega}}{\partial\boldsymbol{y}} = -\boldsymbol{\Omega}^{-1}\frac{\partial\boldsymbol{\omega}}{\partial\boldsymbol{y}}$$

since $\boldsymbol{\omega}$ is by definition equal to zero at $\hat{\boldsymbol{\theta}}$. If we define $\boldsymbol{r}(\boldsymbol{y}, \boldsymbol{\theta}) = \boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})$ then the penalized objective function (suppressing the dependence on $\boldsymbol{y}$ and $\boldsymbol{\theta}$) is $J\boldsymbol{r}^T\boldsymbol{r}$ and we have $\boldsymbol{\omega} = -2J\hat{\boldsymbol{X}}^T\boldsymbol{r} + \boldsymbol{r}^T\boldsymbol{r}(\partial J/\partial\boldsymbol{\theta})$ and

$$\boldsymbol{\Omega} = 2J\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} - 2\hat{\boldsymbol{X}}^T\boldsymbol{r}\frac{\partial J^T}{\partial\boldsymbol{\theta}} - 2\frac{\partial J}{\partial\boldsymbol{\theta}}\boldsymbol{r}^T\hat{\boldsymbol{X}} + \boldsymbol{r}^T\boldsymbol{r}\frac{\partial^2 J}{\partial\boldsymbol{\theta}^2}$$

The penalized influence matrix is then

$$\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]} = 2\hat{\boldsymbol{X}}\boldsymbol{\Omega}^{-1}\left(J\hat{\boldsymbol{X}}^T - \frac{\partial J}{\partial\boldsymbol{\theta}}\boldsymbol{r}^T\right)$$

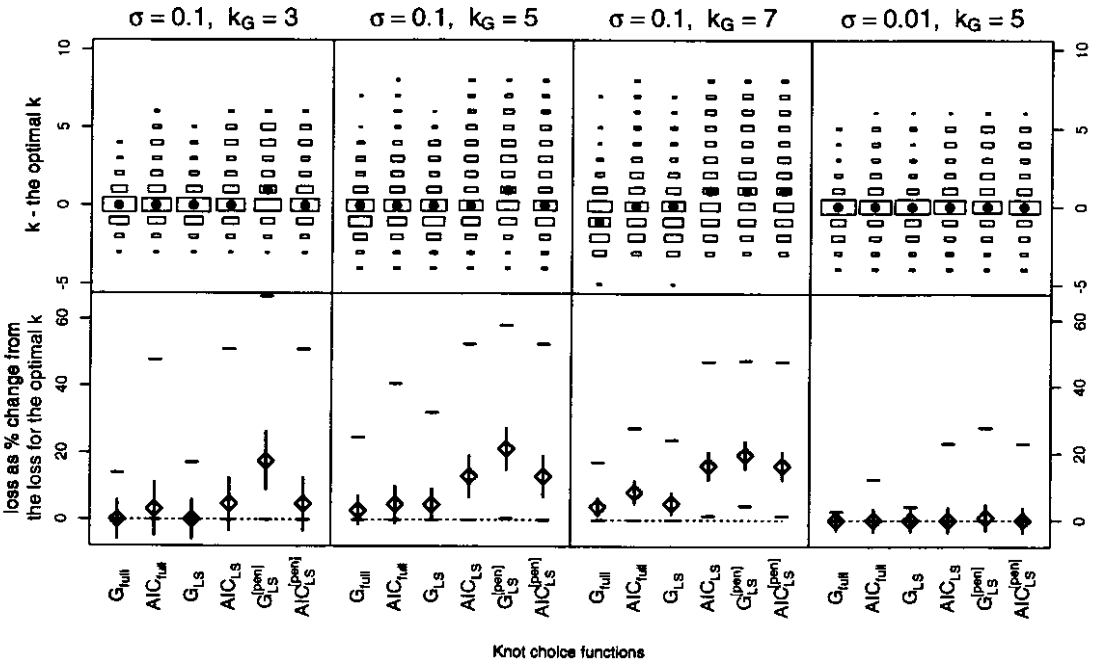

Figure 4: Choosing k, $p = 1.1$: results from variable-truth simulations (See Section A.2 of the Appendix for simulation details). See the legend for Figure 3 for a description of the plots.

Note that if J does not depend on the spline coefficients then the Hessian simplifies to

$$\boldsymbol{\Omega} = 2J\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S} \end{bmatrix}$$

where

$$\boldsymbol{S} = -2\left(\hat{\boldsymbol{X}}_2^T\boldsymbol{r}\frac{\partial J^T}{\partial \gamma} + \frac{\partial J^T}{\partial \gamma}\boldsymbol{r}^T\hat{\boldsymbol{X}}_2\right) + \boldsymbol{r}^T\boldsymbol{r}\frac{\partial^2 J}{\partial \gamma^2}$$

Also, if $J \equiv 1$ then $\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]} = \boldsymbol{A}_{\boldsymbol{\theta}}$.

Figure 4 shows a comparison of a number of k-choice criteria including $G_{LS}^{[\text{pen}]}$ and $AIC_{LS}^{[\text{pen}]}$ created by substituting $\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]}$ for $\boldsymbol{A}_{\boldsymbol{\theta}}$ in G_{LS} and AIC_{LS}. These two do somewhat worse than the unmodified criterion in choosing the number of knots when $p = 1.1$. However, for larger values of p, $G_{LS}^{[\text{pen}]}$ does better because it accounts for the loss in degrees of freedom due to the larger penalty. This mixed result motivates the definition of a combined criterion dependent on the value

of p:

$$G_{\text{comb}}(\boldsymbol{\theta}, p) = \begin{cases} G_{\text{full}}(\boldsymbol{\theta}) & p \leq 1.1 \\ G_{LS}^{[\text{pen}]}(\boldsymbol{\theta}) & p > 1.1 \end{cases} \tag{4}$$

which works well as a k choice criterion for all values of p considered here. Note that matrix decompositions can be used to speed calculation of trace($\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]}$) but this is not essential since it would typically be calculated only once for each value of k considered.

4.3 Estimating σ

Estimating the degrees of freedom for signal (and thus the degrees of freedom for error) is also an important issue when estimating σ. Figure 5 shows that, over a range of values of p, $\hat{\sigma}^2 = R(\boldsymbol{\theta})/\text{trace}(I - \boldsymbol{A}_{\boldsymbol{\theta}})$ does not do as well as well as $\hat{\sigma}^2_{\text{pen}} = R(\boldsymbol{\theta})/\text{trace}(I - \boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]})$.

5 Properties of Penalized Estimation of Knot Locations

5.1 Lethargy

Penalized estimation of knots (as described above) does not suffer from lethargy. If we define the ordered knot vector $\boldsymbol{\gamma}_{i,\delta} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-2}, c - \delta, c + \delta, \gamma_{i+1}, \ldots, \gamma_k)$ then the lethargy property for two coalesced knots can be stated as: $\lim_{\delta \to 0} \partial R(\boldsymbol{\gamma}_{i,\delta})/\partial \delta = 0$ for all $i \in 1, \ldots, k+1$. The following theorem can be easily extended to three or more coalescing knots.

Theorem. If $J(\boldsymbol{\gamma}_{i,\delta})$ is monotone and continuously differentiable with respect to c in $(-d_i, 0)$ and $(0, e_i)$ for $d_i \leq c - \gamma_{i-2}$ and $e_i \leq \gamma_{i+1} - a$, $i \in 1, \ldots, k+1$, and if

$$\lim_{\delta \to 0} J(\boldsymbol{\gamma}_{i,\delta}) \neq 0 \quad \text{and} \quad \lim_{\delta \to 0} \frac{\partial J(\boldsymbol{\gamma}_{i,\delta})}{\partial \delta} \neq 0$$

then the minimization of $R(\boldsymbol{\gamma}, \boldsymbol{c})J(\boldsymbol{\gamma})$ does not suffer from lethargy; i.e.

$$\lim_{\delta \to 0} \frac{\partial (R(\boldsymbol{\gamma}_{i,\delta}, \boldsymbol{c})J(\boldsymbol{\gamma}_{i,\delta}))}{\partial \delta} \neq 0, \quad i \in 1, \ldots, k+1$$

Proof. The proof of the Theorem follows directly from the chain rule and the following lemma.

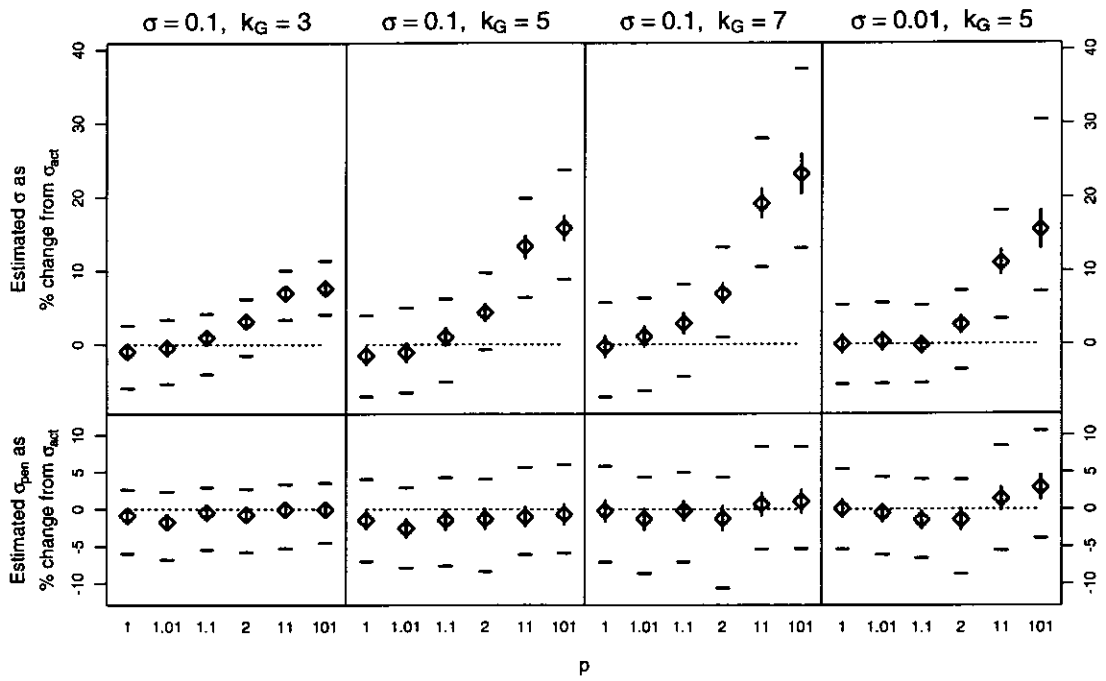

Figure 5: Estimating σ: Results from variable-truth simulations for various values of p (see Section A.2 of the Appendix for simulation details). The length of the knot vector k is chosen separately for each simulation replication as the minimizer of G_{comb}. Estimates are displayed as percent change from σ_{act} (defined as the square root of the mean squared error of the simulated errors) calculated separately for each simulation replication. The **top panels** show the difference between $\hat{\sigma}$ and σ_{act} and the **bottom panels** show the difference between $\hat{\sigma}_{\text{pen}}$ and σ_{act}. See Figure 1 for description of the plotting symbols.

Lemma. If $f(x)$ is monotone and has one continuous derivative in the interval $(0, e)$, $e > 0$ and if $\lim_{x \downarrow 0} f(x) = \infty$, then $\lim_{x \downarrow 0} f'(x) = \infty$.

We investigated the practical consequences of the elimination of lethargy in penalized estimation using a starting-value simulation. The goal of this simulation is to catalog the stationary points in the penalized ($p = 1.1$) and unpenalized objective functions for a particular example and to describe the behavior of the optimization algorithm over a variety of starting values. For this simulation, one test example data set was simulated (see Figure 6). As far as possible, the knot vectors which minimize G_{comb} for both penalized and unpenalized estimation were identified. The found optimum for the penalized problem has 6 knots and the found optimum for the unpenalized problem has 5. In order to provide a fair comparison, both the 6 and 5 knot unpenalized objective

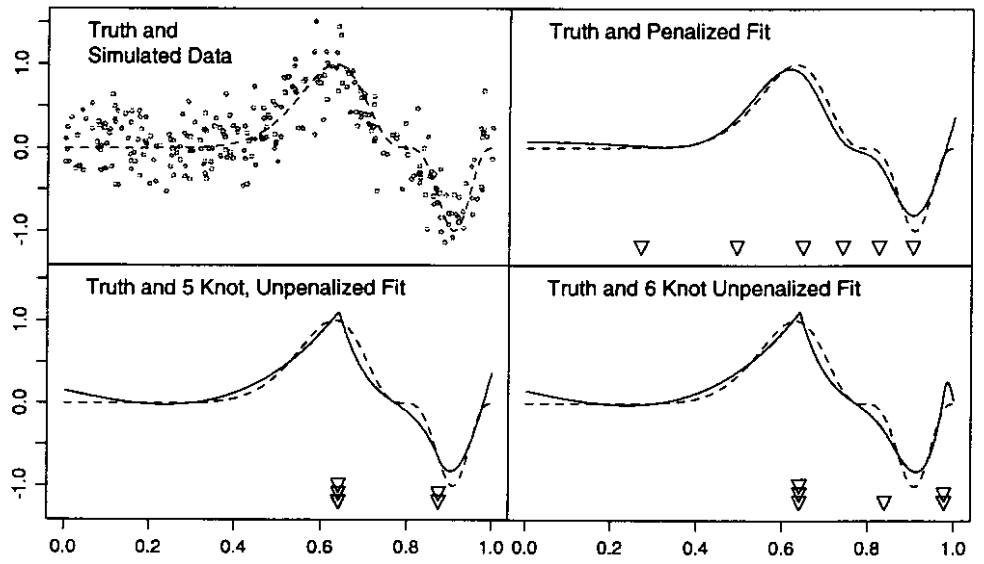

Figure 6: The test function $(\sin(2\pi x^3))^3$ suggested by Härdle (1990) with 256 simulated observations with independent, $\mathcal{N}(0, 0.3^2)$ errors. In each panel the dotted line is the truth. The circles in the upper left panel are the simulated data and the solid lines in the other panels are the indicated fits. The inverted triangles denote the locations of the fitted knots.

functions were explored.

The simulation proceeds by generating 600 starting knot vectors of length 6 and 600 of length 5 by simulating knots from a uniform distribution on (0,1). For each starting knot vector, for each of the three objective functions (6 knot unpenalized plus 5 and 6 knot penalized) a Newton-Raphson/Levenberg-Marquardt algorithm was allowed to run until either convergence or failure to progress was declared. The result of this exercise is a list of 600 final knot vectors for each objective function. The next step is to categorize these knot vectors into equivalence classes representing unique stationary points. For the penalized objective function this is simple because the knot vectors were either identical to 4 decimal places or quite different. For the unpenalized objective function there was no clear cutoff point for deciding that two knot vectors represented the same stationary point. Arbitrarily, two knot vectors were deemed to represent the same stationary point if the Euclidean distance between them was less than 0.01. This is quite conservative (will result in fewer stationary points) in that the Newton-Raphson algorithm should be able to estimate a knot vectors to much greater accuracy. The results are summarized in Table 1.

In comparing the penalized and unpenalized found local optima, the penalized objective func-

Table 1: Results from the starting-value simulation.

	Unpenalized		Penalized
	$k = 5$	$k = 6$	$k = 6$, $p = 1.1$
Number of unique stationary points found	83	120	5
Median iterations to find a stationary point	61	72	26
Median function evaluations to find a stationary point	160	189	59
Percent of the starting values which lead to convergence	60%	41%	100%
Percent of the stationary points equal to the found minimum	14.5%	2%	57%
$R(\boldsymbol{\theta})$ at found minimum	19.70	19.39	19.50
$J(\boldsymbol{\theta})$ at found minimum	1.00	1.00	1.02
Loss at found minimum	0.095	0.096	0.087

tion has many fewer stationary points (5 v.s. 83 and 120), the optimization algorithm required fewer function evaluations to find the nearest stationary point and convergence was declared a higher percent of the time. The number of function evaluations is a more accurate than the number of iterations as an indication of how long the algorithm takes to find the stationary point since it accounts for calculations involved in step size halving. The calculation of the penalty is not included in these figures. However, even if the number of function evaluations is doubled to account for the penalty calculation, penalized estimation is still faster. Most importantly, the found global minimum was obtained much more frequently for the penalized problem (57% v.s. 14.5% and 2%) and has smaller loss.

The fitted curves corresponding to the found optimal penalized and unpenalized knot vectors are shown in Figure 6. The unpenalized fits include a sharp point and, in general, do not do as well as the penalized fit in capturing the true function. See Section 5.2 for simulation results on the bias and variance of estimates for this test function using multiple simulated response vectors.

The number of stationary points for the penalized objective function is very stable over increasing number of starting values. The same five stationary points are found whether 100 or 600 starting values are used making it is likely that the found minimum is in fact the global minimum of the objective function. The situation for the unpenalized objective function is much different. The number of found stationary points increases with the number of starting values used. For 5 knots there were 44 unique stationary points found in the first 200 starting values, 67 in the first

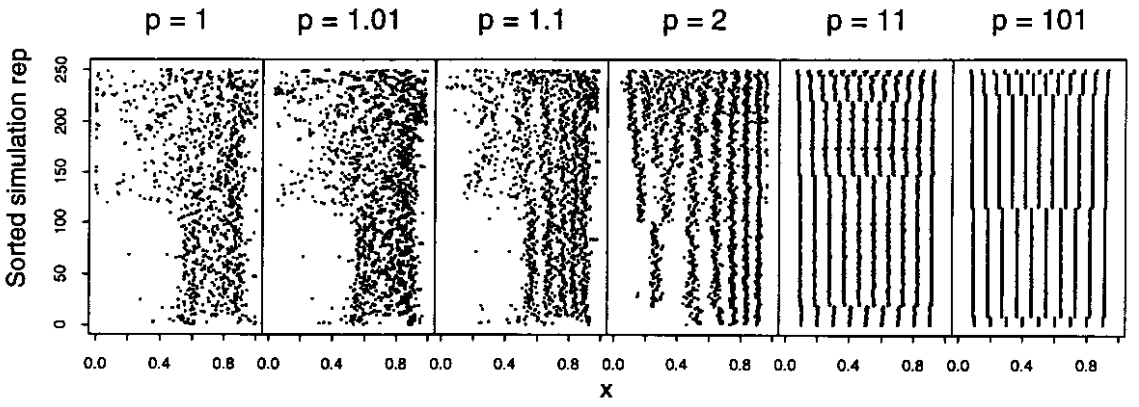

Figure 7: Knot locations from the fixed-truth simulation (see Section A.2 of the Appendix for details). The y axis is the simulation replication sorted from smallest to largest estimated k. The sorting is done separately for each sub-plot.

400 and 83 using all 600. For 6 knots the numbers were 63, 93, and 120. This pattern indicates that that there are likely more stationary points and the true global minimum may not have been obtained. However, this does not invalidate the important conclusions of the simulation, i.e., that the global minimum for the unpenalized problem is very difficult to obtain and that the best unpenalized fit that can be found is not necessarily a better fit, and may be considerably worse than the penalized fit.

5.2 Small sample variance and bias

We use another simulation to explore the bias and variance properties of free-knot spline estimator in small samples. This simulation is a fixed-truth simulation (described in Section A.2 of the Appendix) and is based on the Härdle test function shown in Figure 6. Figure 7 shows the knots chosen by minimizing G_{comb} for each of 250 simulated data sets of size 75. Note the effect of the increasing penalty on the locations of the knots. Figure 8 shows the means, variance, and bias from the fixed-truth simulation. As expected, the bias is smallest for the smaller values of p (less penalty), however, the variance is not uniformly smaller for the larger values of p. Instead, the variance is more constant over the range of the data for the fixed and nearly fixed knots rather than rising for the more difficult to estimate right half of the x range. Note that $p = 1.1$ (the value chosen to minimize the loss in the variable-truth simulation (Section 3.2)), is the largest p which

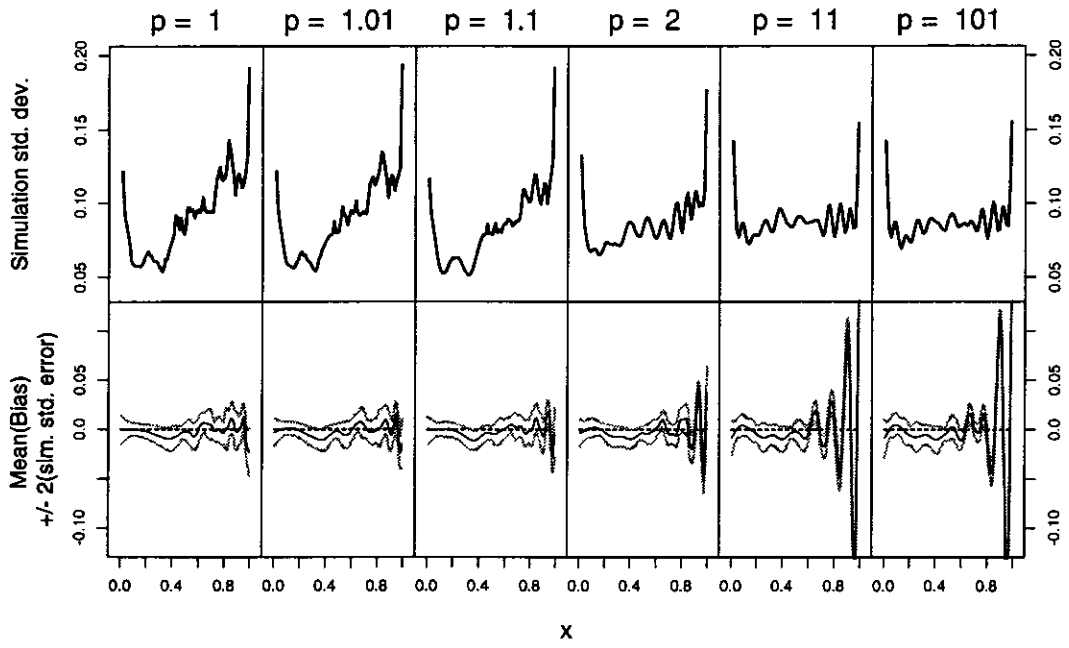

Figure 8: Variance and bias from the fixed-truth simulation (see Section A.2 of the Appendix for details). The **top panels** show the standard errors of the estimated curves over the 250 simulation replications. The **bottom panels** show the mean bias (dark lines) over the simulation replications and pointwise 95% confidence intervals (light lines) based on the simulation standard errors.

does not increase the bias substantially.

6 Discussion

We have demonstrated the computational advantages of penalizing the knot locations in a free-knot spline. Lethargy is eliminated and, in at least in one noisy data example, the number of local optima in the objective function are dramatically reduced and the chance of converging to the global minima is greatly increased.

In addition, the statistical properties of the penalized free-knot spline estimator look promising. While no generalizations can be made at this point, in a number of noisy data examples that we have examined, the unpenalized free-knot spline solution includes non-smooth features which

reduce the residual sum of squares by following a few data point. This tendency degrades the average performance of the spline as measured by the loss. Penalizing the knots seems to keep the solution from these irregular features without introducing excessive smoothing. Unpenalized free-knot splines are more stable in problems with a high signal to noise ratio (their traditional application in the numerical analysis literature). However, if the true function is smooth, penalized free-knot splines also do well in this situation and they have better computational properties.

This description of penalized free-knot spline estimation is not intended to be the last word on the subject. There are many open questions including the efficient generation of good starting values, the asymptotic behavior of the function estimate when the number of knots is chosen using a data based criterion, the benefit of the multiplicative penalty over a more traditional additive penalty, and the wisdom of fixing p rather than choosing it (possibly from a finite set of possibilities) using a GCV or AIC criterion.

Appendix: Computational Details

A.1 Natural spline basis

We have chosen to use natural splines because of there improved behavior at the endpoints of the interval of estimation, $[a, b]$. From among the numerous possible bases for the space of natural splines (Eubank, 1988), we have chosen to use one described by Greville (1969) which has basis functions with local support. In fact, the central $k - m + 2$ basis functions (ordered by their intervals of positive support) correspond to the central $k-m+2$ b-splines in the standard b-spline basis (de Boor, 1978). More importantly for us, the derivative of the elements of the Greville basis with respect to the knots is readily calculated using the simple form of the derivative of a divided difference (Jupp, 1978). These derivatives are required for the Newton-Raphson algorithm and for the calculation of the asymptotic variance of the estimated knots. Note that both the description of the Greville basis in Eubank (1988) (page 269) and the original contain small errors. In Eubank's description, the knots in the first divided difference should run from 1 to $m + j$ rather than from 1 to $m + j - 1$. In Greville's description, the knots should run from 1 to $k + i$ rather than to $k + 1$ where, in his notation, k is the order of the spline.

A.2 Simulations

Variable-truth simulations

Each simulation includes 200 simulation replications. Each simulation replication consisted of the following steps

1. Simulate k_G generating knots from a uniform distribution on $(0, 1)$ and $k_G + 2$ generating spline coefficients from a $N(0, 1)$ distribution

2. Scale the spline coefficients so that the corresponding true y values range from 0 to 1.

3. Create the "truth" by evaluating the spline defined by the generating knots and the generating coefficients at 40 equally spaced points in $(0, 1)$.

4. Create the simulated data vector by adding independent $\mathcal{N}(0, \sigma^2)$ errors to the truth.

5. For $p = 1$, 1.01, 1.1, 2, 11, and 101

   (a) For $k = 2$ through $k_{\max}$ find and record the knot vector of length k that minimizes the penalized sum-of-squares $J(\boldsymbol{\theta}, p)R(\boldsymbol{\theta})$. This is accomplished by finding a suitable starting knot vector and letting a Newton-Raphson/Levenberg-Marquardt algorithm run until either convergence or failure to progress is declared. The method for finding the starting knot vector is discussed in Section A.3.)

   (b) For the 8 objective functions (loss, G_{SS}, G_{full}, AIC_{full}, G_{LS}, AIC_{LS}, $G_{LS}^{[pen]}$, and $AIC_{LS}^{[pen]}$) find and record the knot vector that minimizes the objective function over the knot vectors (of various lengths) obtained in Step 5a.

The raw result of one simulation replication is 48 (= 6 values of $p \times 8$ knot-choice objective functions) knot vectors of varying lengths.

We use splines for the true function in each simulation replication because of the ease of creating splines with a large variety of shapes. However, the goal in fitting these simulated data is not to estimate the generating knots but to estimate the true function. The result of this simulation should generalize to all functions of similar complexity.

For each simulation, k_G, σ and k_{max} must be specified. An attempt was made to choose k_{max} larger than the length of the longest knot vector chosen in Step 5b. This was successful in all but a few iterations.

Fixed-truth simulation

This simulation consists of 250 simulation replications and is similar to the variable-truth simulations above with Steps 1 – 3 replaced by

1′. Create the "truth" by evaluating the Härdle test function $(\sin(2\pi x^3))^3$ at 75 equally space x values in $[0, 1]$.

and Step 4 replaced by

4′. Create the simulated data vector by adding independent $\mathcal{N}(0, 0.3^2)$ errors to the truth.

A.3 Starting Values

There are many potential approaches to finding starting values for the knot locations in a free-knot spline and we are currently comparing a number of different methods, most based on stepwise insertion or deletion algorithms. For the purposes of the simulations described above we use the following simple method:

Given $k_{\max}$, the length of the longest knot vector to be checked; penalty parameters p, and $\gamma^0(k)$; and a model selection criterion such as AIC or GCV, the following steps are performed for each integer k between 2 and $k_{\max}$ inclusive:

1. Simulate $100 \times k$ knot vectors, each of length k.

2. Select the knot vector which evaluates to the minimum $R(\hat{b}(\gamma), \gamma)J(\gamma, p)$.

3. Use the knot vector from step 2 as the starting value for a Newton Raphson optimization of $R(\gamma)J(\gamma, p)$. Iterate to convergence or for 150 iterations whichever comes first.

4. Record the resulting "candidate knot vector" of length k.

Select from among the candidate knot vectors of various lengths using the model selection criterion.

The main advantage of this approach is that it is not biased for fixed k. That is, no knot vector has any greater chance of being found than any other. However, this approach is quite slow and may be biased in favor of fewer knots because of the "curse of dimensionality". To provide similar coverage in all dimensions, $C^k/k!$ random knot vectors would have to be used in Step 1 above for some constant C. For our simulations where candidate k values range from 2 to 13 $C > 2$ is infeasible. Instead we use $100 \times k$ random knot vectors in Step 1. The simulations presented here compare different ways to use the same starting values so a small bias toward shorter knot vectors shouldn't invalidate the results.

References

Agarwal, G. G. and W. J. Studden (1980). "Asymptotic integrated mean square error using least squares and bias minimizing splines". *The Annals of Statistics*, 8(6) 1307–1325.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, New York.

Burchard, H. G. (1974). "Splines (with optimal knots) are better". *Applicable Analysis*, 3 309–319.

de Boor, C. and J. R. Rice (1968). "Least squares cubic spline approximation II - variable knots". Technical Report 21, Computer Sciences Department, Purdue.

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

Eilers, P. H. C. and B. D. Marx (1996). "Flexible smoothing with B-splines and penalties (with discussion)". *Statistical Science*, 11(2) 89–102.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, Inc., New York.

Friedman, J. H. and B. W. Silverman (1989). "Reply to comments on "flexible parsimonious smoothing and additive modeling"". *Technometrics*, 31 35–39.

Goldman, R. N. and T. Lyche (1993). *Knot insertion and deletion algorithms for b-spline curves and surfaces.* SIAM, Philadelphia.

Greville, T. N. E. (1969). "Introduction to spline functions". In *Theory and Applications of Spline Functions (Proceedings of Seminar, Math. Research Center, Univ. of Wis., Madison, Wis., 1968)*, pp. 1–35. Academic Press, New York.

Jupp, D. L. B. (1978). "Approximation to data by splines with free knots". *SIAM Journal of Numerical Analysis*, 15(2) 328–343.

Kooperberg, C., C. J. Stone, and Y. K. Troung (1995). "Hazard regression". *Journal of the American Statistical Association*, 90 78–94.

Luo, Z. and G. Wahba (1997). "Hybrid adaptive splines". *Journal of the American Statistical Association*, 92 107–116.

O'Sullivan, F. (1986). "A statistical perspective on ill-posed inverse problems (with discussion)". *Statistical Science*, 1 502–527.

Pregibon, D. (1981). "Logistic regression diagnostics". *The Annals of Statistics*, 9 705–724.

Stone, C. J. (1982). "Optimal global rates of convergence for nonparametric regression". *The Annals of Statistics*, 10 1040–1053.

Wahba, G. (1977). "A survey of some smoothing problems and the method of generalized cross-validation for solving them". In *Applications of Statistics*, pp. 507–524.

—— (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.