Penalized Estimation of Free-Knot Splines [*] [†]

Mary J. Lindstrom

September, 1997

University of Wisconsin–Madison
Department of Biostatistics & Medical Informatics

Technical Report #122

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

[*]This research was supported in part by National Institutes of Health grants Nos. DC00820

[†]This technical report has appeared as: M. J. Lindstrom, Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, 8:333–352, 1999.

Abstract

Polynomial splines are often used in statistical regression models for smooth response functions. 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 popular regression functions in part because of their excellent approximation properties (de Boor, 1978). For our purposes, a spline on the interval $[a, b]$ with knots $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, $a \leq \gamma_1 \leq \gamma_2 \leq, \ldots, \leq \gamma_k \leq b$ 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 parametric regression 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 order of 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 from the data. Freeing the knots improves the spline's approximating power (Burchard, 1974) but this added flexibility 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 numerous local optima in the residual sum-of-squares surface, 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 but are usually dismissed as computationally intractable.

The second disadvantage of free knots (which is related to the first) is that the optimal knot vector often includes replicate knots which allow non-smooth behavior of the predicted curve. In the commonly used cubic splines ($m = 4$), duplicate knots allow a discontinuous second derivative, triplicate knots a discontinuous first derivative and 4 identical knots in one place allow 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 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 related models and 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 estimating the error variance in Section 4. Computational and statistical properties of the penalized estimator are presented in Section 5. In Section 6 we present a real data example and discuss our results in Section 7. Some computational details are given in an Appendix.

2 Background

2.1 Related Models

Free knot splines differ from classical smoothing splines (Eubank, 1988; Wahba, 1990) and other penalized regression splines (O'Sullivan, 1986; Eilers and Marx, 1996) in that the amount of smoothing is locally adaptive (knots move to areas where the function is less smooth). In contrast to adaptive regression splines, (Friedman and Silverman, 1989; Kooperberg *et al.*, 1995; Stone *et al.*, 1997), and hybrid smoothing splines (Luo and Wahba, 1997), the choice of knot locations in a free-knot spline is a parameter estimation, rather than a predictor subset selection, problem. Bayesian nonparametric regression (Smith *et al.*, 1998; Denison *et al.*, 1998) is probably closest in spirit to the approach taken here in that the knot locations are treated as parameters and given a prior distribution. However the distribution is over a finite set of possible locations (usually the design points) and additional hard constraints are usually imposed to keep knots some minimum distance apart.

2.2 Computational difficulties

There are a number of interrelated reasons why finding the locations of the knots is so difficult. First, excellent starting values are required to find the global optimum since there are usually many local optima in the residual sum-of-squares surface. The existence of multiple optima in the objective surface 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 surface 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.

While good starting values are essential they are are also particularly difficult to construct. In most nonlinear regression models, the choice of starting values is aided by the roles that the parameters play, e.g., asymptotes, rates, 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 characterized by a lack of convergence with one or more parameters moving 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 surface is continuous and the corresponding spline is well defined.

2.3 Existing approaches to optimizing knot locations

If approximate knot locations can be deduced from inspection of the data, standard non-linear least squares can be used to obtain the knot estimates (e.g. Gallant and Fuller, 1973). If, as

is usually the case, the knots cannot be visually identified, 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, and possibly only, contribution in the first category. He defines a transformation of the knots that ameliorates lethargy and removes the redundancy in the knot vector. This is accomplished 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. The "Jupp parameters" are $\log(h_{i+1}/h_i),\ i = 1, \ldots, k$, where $\boldsymbol{\gamma}$ 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 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, in the examples given in Jupp (1978), increases the chance of finding the global optimum. We have found the Jupp transformation quite helpful in regularizing the optimization problem and we use it in all of our calculations. However, false convergence and failure to progress in the neighborhood of knot vectors with replicate knots still occur frequently. Given these realities, many runs with different starting values are still required in order to be reasonably sure that the global optimum is found.

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 nearby, 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. These methods are unlikely to consistently find optimal knot vectors, however, because they have no way of moving away from undesirable, local optima. 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 estimating the function itself.

3 Penalized Estimation

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

Penalizing knot vectors with replicate and nearly replicate knots is intuitively appealing. If the underlying curve is assumed to be smooth, replicate knots are undesirable. However, the global optimum often contains replicate knots even when the data seem to indicate an underlying

smooth function. (See Figure 6 in Section 5.1 for examples of such non-smooth behavior.) When nearly replicate knots are needed to fit the data (the objective surface is sharply convex near the optimum) 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 surface is nearly flat near the optimum and the replicate knots are adding unneeded, 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.

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) and avoiding the need to estimate a smoothing parameter for each data set. The choice of a multiplicative penalty reflects our view of the penalty as a minimal, regularizing, factor which need not vary with the individual data problem. For models where a multiplicative penalty is difficult to apply (e.g. GLIM) an additive penalty would be an appropriate alternative.

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 near 1 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 $\boldsymbol{\gamma}$ and which depend only on the knots (not the spline coefficients). The following penalty, a normalized inverse of the product of the intervals between the augmented knots, has these properties. We define $J_1(\boldsymbol{\gamma}, \alpha, \beta) = \alpha[P(\boldsymbol{\gamma})^\beta - 1] + 1$ where $P(\boldsymbol{\gamma}) = (1/(k+1))^{(k+1)} / \prod_{i=1}^{k+1} h_i$ and 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.

The parameters α and β control the size and shape of the penalty function. Large β corresponds to a sharp corner in the penalty as a function of the minimum distance between knots, smaller

β gives a more gradual transition. However a change in β also changes the value of the penalty dramatically for a fixed $\boldsymbol{\gamma}$. 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, not equally spaced, knot vector $\boldsymbol{\gamma}^0(k)$ is equal to a constant p for all values of β. That is, we define

$$J_2(\boldsymbol{\gamma}, \beta, p) = J_1(\boldsymbol{\gamma}, \alpha_1(p, \boldsymbol{\gamma}^0(k), \beta), \beta) = \alpha_1(p, \boldsymbol{\gamma}^0(k), \beta)[P(\boldsymbol{\gamma})^\beta - 1] + 1$$

where $\alpha_1(p, \boldsymbol{\gamma}^0(k), \beta) = (p-1)/[J_1(\boldsymbol{\gamma}^0(k), 1, \beta) - 1] = (p-1)/[P(\boldsymbol{\gamma}^0(k))^\beta - 1]$ so that $J_2(\boldsymbol{\gamma}^0(k), \beta, p) = p$ 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:

$$J(\boldsymbol{\gamma}, p) = \lim_{\beta \to 0} J_2(\boldsymbol{\gamma}, p, \beta) = \alpha(p, \boldsymbol{\gamma}^0(k)) \log(P(\boldsymbol{\gamma})) + 1$$

where $\alpha(p, \boldsymbol{\gamma}^0(k)) = (p-1)/[\log(P[\boldsymbol{\gamma}^0(k)])]$ and $\log(P(\boldsymbol{\gamma})) = -(k+1)\log(k+1) - \sum_{i=1}^{k+1} \log(h_i) = -\sum_{i=1}^{k+1} \log((k+1)h_i)$.

Note that $P(\gamma)$ and $\log(P(\gamma))$ are equivalent to the popular inverse and logarithmic "barrier functions" respectively. The barrier function method (Gill *et al.*, 1989) is a technique for finding solutions to a constrained, nonlinear, minimization problem. Instead of minimizing the objective function directly, the sum of the objective function and a scalar (called the barrier parameter) times a barrier function is minimized where the barrier parameter decreases as the iterations progress. The fact that $\log(P(\boldsymbol{\gamma}))$ is equivalent to the well tested logarithmic barrier function provides some evidence that $J(\gamma, p)$ will perform well.

We could, of course, apply the barrier method to knot estimation directly by first constraining the knots to be some fixed minimum distance apart. However, the penalized estimator seems more desirable than a constrained solution since it enforces a positive minimum distance between the knots while allowing the knots to move arbitrarily close together (as long as they are not equal) when the data dictated a sharp change in the second derivative. A hybrid approach might be constructed where initially the penalty parameter would be set to a large value and then gradually reduced to a predetermined fixed value as the iterations progress. We do not pursued this approach further here.

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

If the parameter p is allowed to vary freely, fixing $\boldsymbol{\gamma}^0(k)$ does not limit the flexibility of $J(\boldsymbol{\gamma}, p)$. However, as discussed above, we hope to fix the parameters in the penalty and so the dependence of $\boldsymbol{\gamma}^0(k)$ on k should be carefully considered. We set $\boldsymbol{\gamma}^0(k)$ to be a minimum penalty knot vector with minimum spacing between knots $h_{[1]} = d/(k+1)$, $d < 1$. For penalties which depend on $\boldsymbol{\gamma}$ 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 $\log(J_1[\boldsymbol{\gamma}^0(k), 1, 1])$ as a function of k is relatively invariant to changes in d (we use $d = 0.04$). This choice of $\boldsymbol{\gamma}^0(k)$ implies that as k increases, J will also increase for knot vectors which stray from even spacing.

Once $\boldsymbol{\gamma}^0(k)$ 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). It is likely that the optimal value of p will vary depending on the order of the spline. We investigate only cubic splines here as they are the most popular regression splines due to their smooth appearance (when all knots are unique) and low order.

In order to evaluate the behavior of penalized estimation for different values of p we use variable-truth simulations (described in detail in Section A.2). In short, 200 different splines are created by simulating knot locations and spline coefficients. For each of these splines, pseudo-errors are added to create simulated data. Penalized, cubic free-knot splines are fit to the data using 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.

Figure 1 shows the percent change in the loss for a range of values of p compared to $p = 1$

(no penalty) for four combinations of k_G (the number of knots in the random spline defining the truth) and σ. Larger k_G corresponds to more complex, less smooth true functions. 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 increasing p improves the computational properties of the optimization problem, we have chosen $p = 1.1$, the largest p which does not increase the loss over $p = 1$. Thus the penalty has the functional form $\alpha \log(P(\boldsymbol{\gamma})) + 1$ where α is chosen so that the penalty evaluated at $\boldsymbol{\gamma}^0(k)$ is equal to 1.1 (i.e. a %10 penalty).

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.

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): $G_{SS}(\boldsymbol{\gamma}) = R(\boldsymbol{\gamma}, \hat{\boldsymbol{c}}(\boldsymbol{\gamma}))/[\text{trace}(\boldsymbol{I} - \boldsymbol{A}_{\boldsymbol{c}|\boldsymbol{\gamma}})]^2 = R(\boldsymbol{\gamma}, \hat{\boldsymbol{c}}(\boldsymbol{\gamma}))/(n - n_b)^2$ where $\boldsymbol{\gamma}$ 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 $\boldsymbol{A}_{\boldsymbol{c}|\boldsymbol{\gamma}}$ is the influence matrix for the linear least squares estimation of $\boldsymbol{c}$ given fixed $\boldsymbol{\gamma}$. Other authors including Stone *et al.* (1997) have proposed the corresponding $AIC_{SS}(\boldsymbol{\gamma}) = n \log(R(\boldsymbol{\theta})/n) + 2n_b$. Both of these criteria ignore the loss of error degrees of freedom due to the estimation of the knots and, in our

simulations, substantially over estimates the number of knots needed (see Figure 3). Stone *et al.* (1997) obtained similar results and suggests adding an additional multiplier to the $2n_b$ term to further penalized increasing n_b. Instead, we propose estimating the degrees of freedom using the trace of the influence matrix for the complete parameter vector (coefficients and knots).

Since free-knot splines are nonlinear, an expression for the influence matrix for nonlinear regression is required. 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 $\boldsymbol{A}_{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the derivative of $\hat{\boldsymbol{y}}$ with respect to $\boldsymbol{y}$. That is, $\boldsymbol{A}_{i,j}$ 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:

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

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., $\boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}))$ where $\boldsymbol{X} = \partial\boldsymbol{\eta}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Then we have $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y})$ for a fixed point $\boldsymbol{\theta}_0$ near $\hat{\boldsymbol{\theta}}$ and

$$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \boldsymbol{y}} \;\approx\; \frac{\partial(\boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y}))}{\partial \boldsymbol{y}} = (\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T \quad \text{where} \quad \boldsymbol{X}_0 = \left.\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0} \tag{2}$$

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

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

where $\hat{\boldsymbol{X}} = \partial\boldsymbol{\eta}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}}$. This expression for $\boldsymbol{A}_{\boldsymbol{\theta}}$ 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, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ where $\boldsymbol{X}_1$ is the matrix of spline basis functions evaluated at $\boldsymbol{\gamma}$ and the design,

and $\boldsymbol{X}_2 = (\partial \boldsymbol{X}_1/\partial \boldsymbol{\gamma})\,\boldsymbol{c}$ where column i of $(\partial \boldsymbol{X}_1/\partial \boldsymbol{\gamma})\,\boldsymbol{c}$ is equal to $(\partial \boldsymbol{X}_1/\partial \gamma_i)\,\boldsymbol{c}$.

We can now construct two new "least squares" knot choice criteria by substituting $\boldsymbol{A}_{\boldsymbol{\theta}}$ for $\boldsymbol{A}_{\boldsymbol{c}|\boldsymbol{\gamma}}$ in G_{SS} and AIC_{SS} to obtain $G_{LS}(\boldsymbol{\theta}) = R(\boldsymbol{\theta})/[\text{trace}(\boldsymbol{I} - \boldsymbol{A}_{\boldsymbol{\theta}})]^2$ and $AIC_{LS} = n\log(R(\boldsymbol{\theta})/n) + 2\,\text{trace}(\boldsymbol{A}_{\boldsymbol{\theta}})$. We can also define two additional criteria by assuming the derivative matrix $\hat{\boldsymbol{X}}$ is of full rank: $AIC_{\text{full}} = n\log(R(\boldsymbol{\theta})/n) + 2(n_b + k)$ and $G_{\text{full}}(\boldsymbol{\theta}) = R(\boldsymbol{\theta})/(n - n_b - k)^2$. As shown in Figure 3, accounting for all the degrees of freedom in the model improves the performance of the criteria substantially. 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). BIC criteria of the form $\log(R(\boldsymbol{\theta})/n) + log(n) \times$ (degrees of freedom for the model) were also tried but did not improve on G_{full} (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 $\text{trace}(\boldsymbol{A}_{\boldsymbol{\theta}})$ (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 spacing, eliminating k degrees of freedom in the model. Alternatively, we can generalize the expression for the influence matrix 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 $\boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{y}) = -\boldsymbol{\Omega}^{-1}\boldsymbol{\omega}$ where $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}$ are the gradient and Hessian of the objective function with respect to $\boldsymbol{\theta}$. This change effects the approximation in Equation 2 which can be generalized to

$$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \boldsymbol{y}} \;\approx\; \frac{\partial(\boldsymbol{\theta}_0 + \boldsymbol{\delta}(\boldsymbol{\theta}_0, \boldsymbol{y}))}{\partial \boldsymbol{y}} = -\left.\frac{\partial \boldsymbol{\Omega}^{-1}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{\theta}_0} \boldsymbol{\omega}_0 - \boldsymbol{\Omega}_0^{-1}\left.\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{\theta}_0}$$

where $\boldsymbol{\Omega}_0$ and $\boldsymbol{\omega}_0$ are $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}$ evaluated at $\boldsymbol{\theta}_0$ and where, as before, $\boldsymbol{\theta}_0$ is close to $\hat{\boldsymbol{\theta}}$. We again define the influence matrix as the limit of the approximation as $\boldsymbol{\theta}_0$ moves toward $\hat{\boldsymbol{\theta}}$. That is,

$$\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]} = \lim_{\boldsymbol{\theta}_0 \to \hat{\boldsymbol{\theta}}} \hat{\boldsymbol{X}} \left(-\left.\frac{\partial \boldsymbol{\Omega}^{-1}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{\theta}_0} \boldsymbol{\omega}_0 - \boldsymbol{\Omega}_0^{-1} \left.\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{\theta}_0} \right) = -\hat{\boldsymbol{X}}\hat{\boldsymbol{\Omega}}^{-1} \left.\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{y}}\right|_{\hat{\boldsymbol{\theta}}}$$

where $\hat{\boldsymbol{\Omega}}$ is $\boldsymbol{\Omega}$ evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\omega}} = 0$ by definition. 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 $J\boldsymbol{r}^T\boldsymbol{r}$, $\partial\boldsymbol{\omega}/\partial\boldsymbol{y} = 2(-J\boldsymbol{X}^T + \partial J/\partial\boldsymbol{\theta}\ \boldsymbol{r}^T)$ and

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

If J does not depend on the spline coefficients then the Hessian simplifies to

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

where $\boldsymbol{S} = -2(\boldsymbol{X}_2^T\boldsymbol{r}\partial J/\partial\boldsymbol{\gamma}^T + \partial J/\partial\boldsymbol{\gamma}^T\boldsymbol{r}^T\boldsymbol{X}_2) + \boldsymbol{r}^T\boldsymbol{r}\partial^2 J/\partial\boldsymbol{\gamma}^2$. 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}. The penalized versions do somewhat worse than the other critera 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 (results not shown). This mixed result motivates the definition of a combined criterion dependent on the value of p:

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

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 $\text{trace}(\boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]})$ but this is not essential since it would typically be calculated only once for each value of k considered.

4.3 Estimating $\boldsymbol{\sigma}$

The choice of estimator for 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, $\hat{\sigma}^2_{\text{pen}} = R(\boldsymbol{\theta})/\text{trace}(I - \boldsymbol{A}_{\boldsymbol{\theta}}^{[\text{pen}]})$ performs better than $\hat{\sigma}^2 = R(\boldsymbol{\theta})/\text{trace}(I - \boldsymbol{A}_{\boldsymbol{\theta}})$.

5 Properties of Penalized Estimation of Knot Locations

5.1 Lethargy

Under mild assumptions, penalized estimation of knot locations 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 $i \in 2, \ldots, k$. The following theorem can be easily extended to the case of a knot coalescing to an end point of the estimation interval or to three or more coalescing knots.

Theorem. If (1) $J(\boldsymbol{\gamma}_{i,\delta})$ is monotone and continuously differentiable with respect to δ for $0 < \delta < \min(c-\gamma_{i+2}, \gamma_{i+1}-c)$, $i \in 2, \ldots, k$; if (2) $\lim_{\delta\to 0} J(\boldsymbol{\gamma}_{i,\delta}) = \infty$ and $\lim_{\delta\to 0} \partial J(\boldsymbol{\gamma}_{i,\delta})/\partial\delta = \infty$ and if (3) for some fixed $d > 0$, $R(\boldsymbol{\gamma}_{i,\delta}) > 0$ for all $\delta < d$; then the minimization of $R(\boldsymbol{\gamma}, \boldsymbol{c})J(\boldsymbol{\gamma})$ does not suffer from lethargy; i.e.

$$\lim_{\delta\to 0} \frac{\partial(R(\boldsymbol{\gamma}_{i,\delta}, \boldsymbol{c})J(\boldsymbol{\gamma}_{i,\delta}))}{\partial\delta} \neq 0, \quad i \in 2, \ldots, k$$

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

We investigate 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 the simulated data set shown in Figure 6 and to describe the behavior of the optimization algorithm over a variety of starting values. To begin, 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 functions were explored along with the the 5 knot penalized objective function.

The simulation proceeds by simulating 600 starting knot vectors of length 6 and 600 of length 5 from a uniform distribution on (0,1). For each starting knot vector, for each of the appropriate objective functions (6 knot penalized and unpenalized or 5 knot unpenalized) 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. Next, the knots were categorized into equivalence classes representing unique stationary

Table 1: Härdle example: results from the starting-value simulation.

	Unpenalized		Penalized
	$k = 5$	$k = 6$	$k = 6$, $p = 1.1$
Number of unique stationary points found	54	103	5
Median iterations to find a stationary point	62	72	26
Median function evaluations to find a stationary point	161	191	63
Percent of the starting values which lead to convergence	58%	62%	100%
Percent of the stationary points equal to the found minimum	17.3%	3.2%	52.7%
$R(\boldsymbol{\theta})$ at found minimum	19.69	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

points. For the penalized objective function this is simple because the knot vectors were either identical to 4 significant figures 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 either (1) the Euclidean distance between them was less than $\sqrt{k}\times$(range of the predictor variable)/200 or (2) no maximum in the objective function was found along a line connecting the two knot vectors. 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 most striking difference is that the penalized objective function has many fewer stationary points (5 v.s. 54 and 103). In addition, the optimization algorithm required fewer function evaluations to find the nearest stationary point and convergence was declared a much higher percent of the time. Most importantly, the found global minimum was obtained much more frequently for the penalized problem (53% v.s. 17% and 3%) 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: five stationary points are found whether 100 or 600 starting values are used. For the unpenalized objective function, the number of found stationary points increases with the number of starting values. The last starting values to results in a new stationary point were numbers 598 and 597 out of 600 for

the 5 and 6 knot problems respectively, indicating that that 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 found is not necessarily better, and may be considerably worse, than the penalized fit.

5.2 Small sample variance and bias

To explore the bias and variance properties of free-knot spline estimator in small samples we simulated 250 data sets consisting of 75 evenly spaced true values from the Härdle test function $(\sin(2\pi x^3))^3$ (Figure 6) plus $\mathcal{N}(0, 0.3^2)$ random noise. For each of these simulated data sets, the optimal knot vectors were found using the methods described in section A.3. Figure 7 shows the knots chosen by minimizing G_{comb} for each of the data sets. Note the effect of increasing the penalty on the locations of the knots. Figure 8 shows the pointwise standard deviation and mean bias of the 250 estimated curves. 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 does not increase the bias substantially.

6 Titanium Heat Example

The titanium heat data set (Figure 9), introduced by de Boor and Rice (1968a), has been used as a test case for evaluating knot finding algorithms (de Boor and Rice, 1968b; Jupp, 1978). The x-axis is temperature and the response is identified only as a "thermal property of titanium". The existence of a 5 knot solution with no replicate knots (shown in the first row of plots in Figure 9) has contributed to the expectation that hard-to-find, distinct-knot solutions usually exist. In the author's experience it is more common to have no such solution. In fact, even for the titanium data, the distinct knot solution is not a very good fit to the data (see the residual plot in Figure 9). Note that the 5 knot solution presented here, [830.88, 877.51, 898.21, 916.12, 974.48] is slightly different from that presented in de Boor and Rice (1968b) and Jupp (1978) because they minimized

Table 2: Titanium Heat Example: results from the starting-value simulation.

	Unpenalized $k = 7$	Penalized $k = 7$, $p = 1.1$
Number of unique stationary points found	106	12
Median iterations to find a stationary point	166	156
Median function evaluations to find a stationary point	401	347
Percent of the starting values which lead to convergence	64.8%	99.8%
Percent of the stationary points equal to the found minimum	3.5%	9.7%
$R(\boldsymbol{\theta})$ at found minimum	0.001363	0.001383
$J(\boldsymbol{\theta})$ at found minimum	1.00	1.142

a weighted residual sum of squares to reduced the influence of the end data points whereas we are using natural splines to the same end.

The optimal penalized and unpenalized solutions (using the G_{full} criterion) have 7 knots. The solutions are [800.85, 860.37, 876.40, 881.82, 907.40, 912.83, 974.83] and [800.07, 858.74, 879.75, 879.75, 910.37, 910.37, 974.90] respectively. As show in Figure 9, the penalized and unpenalized fits are very similar. The advantage of the penalized solutions that it has a continuous second derivative whereas the unpenalized fit does not because of the two sets of duplicate knots. Also, as in the Härdle example, the objective surfaces, and thus the optimization problems, are quite different. Table 2 summarizes a second starting-value simulation which was conducted in the same way as that described in Section 5.1 for the Härdle example. The number of stationary points is reduced dramatically from 106 to 12. The number of iterations and function evaluations necessary to find the optima are not substantially reduced but the percent which result in convergence is improved from 64.8% to 99.8%. The percent of starting values that converged to the found global minimum increased from 3.5% to 9.7%. While both values are low (indicating a very difficult optimization problem) the relative increase is substantial.

7 Discussion

We have demonstrated the computational advantages of penalizing the knot locations in a free-knot spline. Under mild conditions, lethargy is mathematically eliminated. In our examples, penalizing reduces the number of local optima in the objective function dramatically and increases the chance of converging to the global minimum.

In addition, the statistical properties of the penalized free-knot spline estimator look promising. In noisy data examples such as the simulated Härdle test function example, optimal unpenalized

free-knot splines tend to include non-smooth features which reduce the residual sum of squares by following a few data points. 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 successful in problems with a high signal to noise ratio (their traditional application in the numerical analysis literature). However, penalized free-knot splines also do well in this situation and they have improved computational properties as demonstrated in the titanium heat example.

There are many remaining questions about penalized estimation of free-knot splines 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 their improved behavior at the endpoints of the interval of estimation. 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 derivatives of these basis functions with respect to the knots (required for the Newton-Raphson algorithm) are readily calculated using the simple form of the derivative of a divided difference (Jupp, 1978). Note the following typographical errors: in Eubank (1988) on page 269 the knots in the first divided difference should run from 1 to $m + j$ rather than from 1 to $m + j - 1$; in Greville (1969) the knots should run from 1 to $k + i$ rather than from 1 to $k + 1$.

A.2 Variable-truth simulations

For each combination of k_G and σ, 200 replications of the following steps are performed.

1. Create the "truth" by simulating k_G generating knots from a uniform distribution on $(0, 1)$ and $k_G + 2$ generating spline coefficients from a $N(0, 1)$ distribution. Scale the spline coefficients so that the corresponding true y values range from 0 to 1 and evaluate the spline at 40 equally spaced points in $(0, 1)$.

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

3. 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 $J(\boldsymbol{\theta}, p)R(\boldsymbol{\theta})$ (See Section A.3 for details).

   (b) Record the knot vector that minimizes the knot choice objective functions of interest over the knot vectors (of various lengths) obtained in Step 3a.

The raw result of one simulation replication is 54 (= 6 values of p × 9 knot-choice objective functions) knot vectors of varying lengths. An attempt was made to choose k_{max} larger than the length of the longest knot vector chosen in Step 3b. This was successful in all but a few iterations.

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 functions of similar complexity.

A.3 Minimizing the objective function

It is not know what the best way is to find the knot vector that minimizes either the penalized or unpenalized objective function. We are currently comparing a number of different methods. 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 and a model selection criterion such as G_{full}, 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 with the minimum objective function value.

3. Use the knot vector from Step 2 as the starting value for a Newton Raphson optimization of $R(\boldsymbol{\gamma})J(\boldsymbol{\gamma}, p)$ (in the Jupp parameter space). Iterate to convergence, failure to converge, 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 the coverage of the parameter space by the random knot vectors is better for smaller k. Since the simulations presented here compare different ways to use the same starting values, a small bias toward shorter knot vectors shouldn't invalidate the results.

Acknowledgments

The author thanks three referees and an associate editor for many helpful suggestions. This work was supported in part by National Institutes of Health grants Nos. DC00820 and CA75097.

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". p. 416.

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

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998). "Automatic Bayesian curve fitting". *Journal of the Royal Statistical Society – Series B*, 60 333–350.

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.

Gallant, A. R. and W. A. Fuller (1973). "Fitting segmented polynomial regression models whose join points have to be estimated". *Journal of the American Statistical Association*, 68(341) 144–147.

Gill, P. E., W. Murray, M. A. Saunders, and M. H. Wright (1989). "Constrained nonlinear programming". In Nemhauser, G. L., A. H. G. Rinnoony Kan, and M. J. Todd, eds., *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*. Elsevier Science Publishers, North-Holland.

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.

Härdle, W. (1990). *Smoothing Techniques With Implementation in S*. Springer-Verlag, 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. Truong (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.

de Boor, C. and J. R. Rice (1968). "Least squares cubic spline approximation I – fixed knots". Technical Report 20, Computer Sciences Department, Purdue.

—— (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.

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.

Smith, M., C.-M. Wong, and R. Kohn (1998). "Additive nonparametric regression with autocorrelated errors". *Journal of the Royal Statistical Society – Series B*, 60 311–331.

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

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.

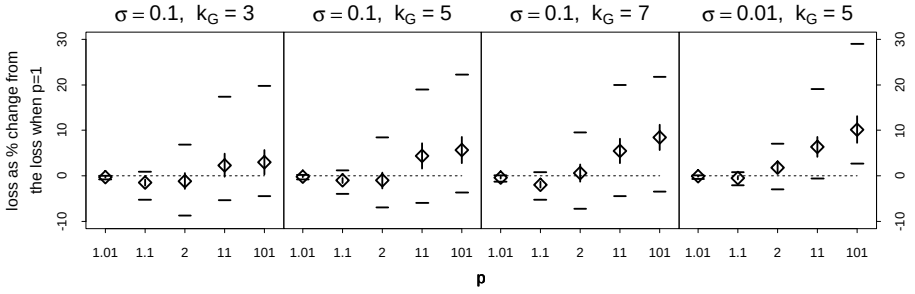

Figure 1: Choosing p: results from variable-truth simulations (see Section A.2 for details). The truth is a random spline with k_G knots and the errors are simulated from a $\mathcal{N}(0, \sigma^2)$ distribution. The y axis is the loss as the percent change from the loss when when $p = 1$, i.e., when 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.

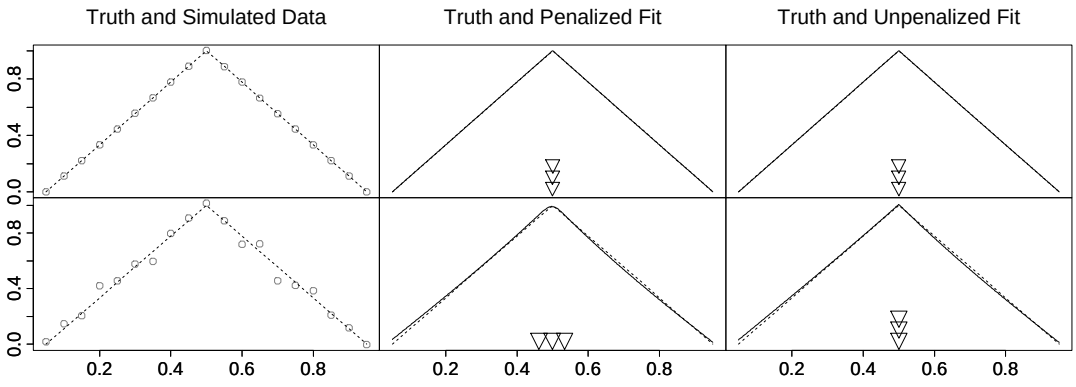

Figure 2: Simulated data with penalized ($p = 1.1$) and unpenalized free-knot spine fits. The dotted line in all panels is the true function. The solid lines are the fitted curves. The solid and dotted lines overlap for the top center and right panels. The errors for the top row of plots are simulated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

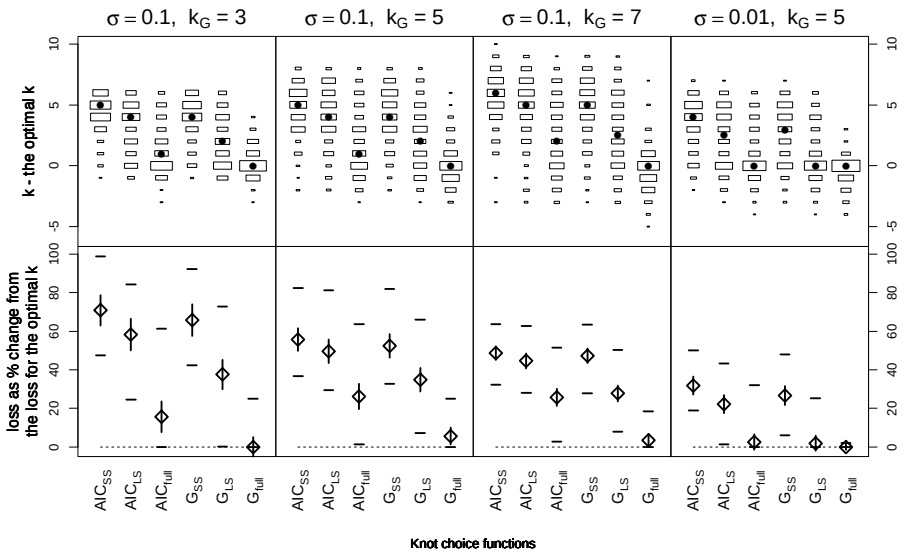

Figure 3: Choosing k, no penalty: Results from variable-truth simulation (see Section A.2 for the definition of k_G and for other simulation details). The **top panels** show the difference between the k chosen by minimizing the objective functions shown on the x-axis and the optimal k chosen by minimizing 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.

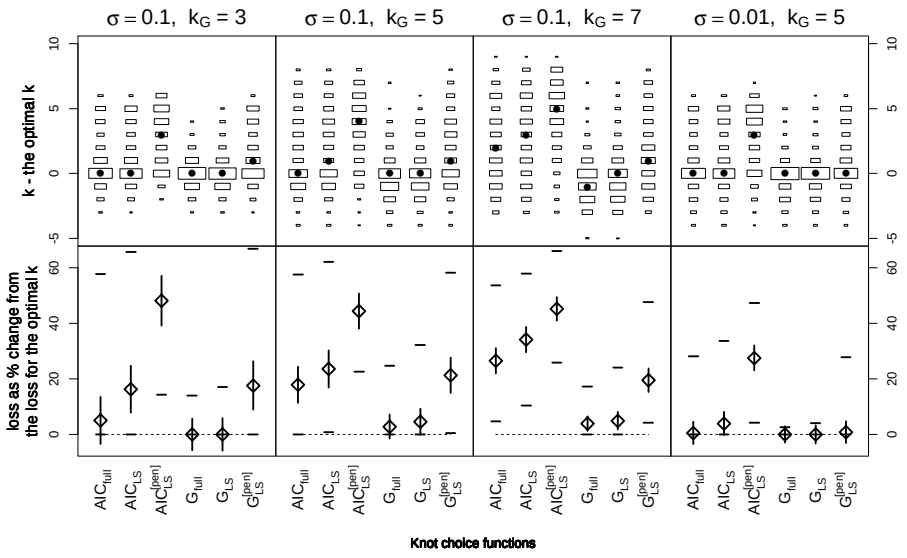

Figure 4: Choosing k, $p = 1.1$: results from variable-truth simulations (See Section A.2 for simulation details). See the legend for Figure 3 for a description of the plots.

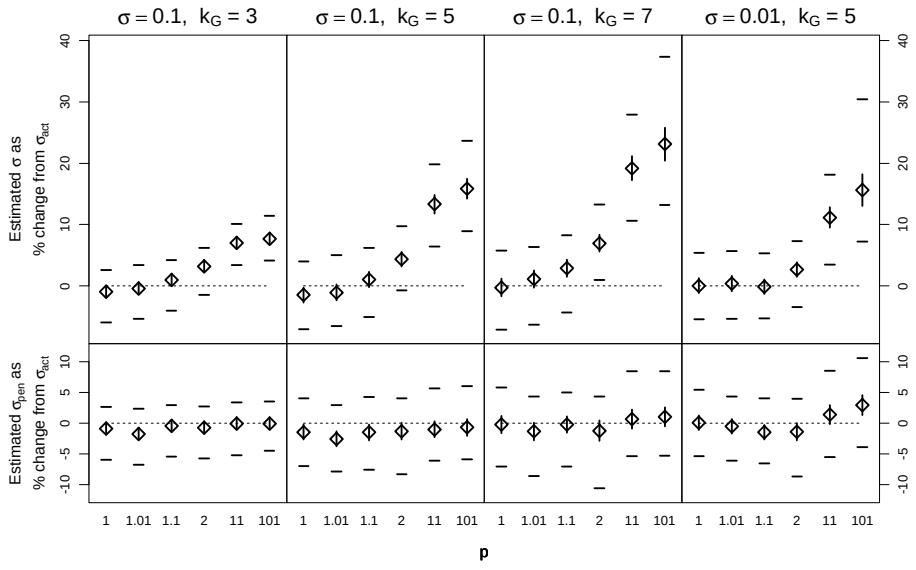

Figure 5: Estimating σ: Results from variable-truth simulations for various values of p (see Section A.2 for simulation details). The length of the knot vector k is chosen separately for each simulation replication as the minimizer of G_{comb}. Estimates are displayed as percent change from σ_{act} (defined as the square root of the mean squared 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 definitions of the plotting symbols.

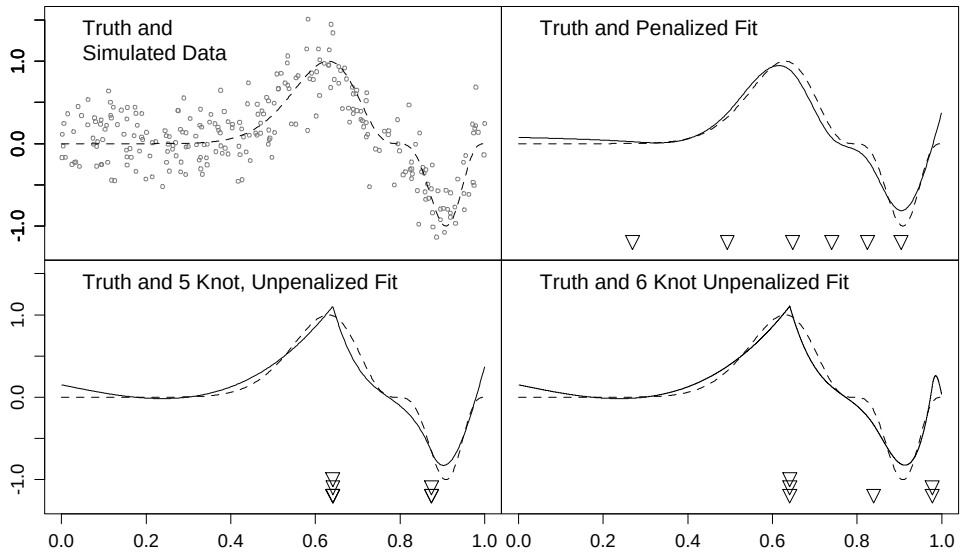

Figure 6: The test function $(\sin(2\pi x^3))^3$ suggested by Härdle (1990) with 256 simulated observations with independent, $\mathcal{N}(0, 0.3^2)$ errors. In each panel the dotted line is the truth. The circles in the upper left panel are the simulated data and the solid lines in the other panels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

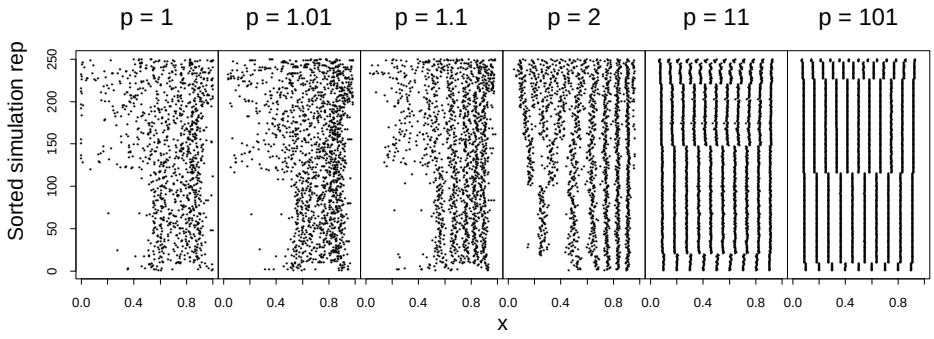

Figure 7: Knot locations from the Härdle function simulation The y axis is the simulation replication sorted from smallest to largest estimated k. The sorting is done separately for each sub-plot.

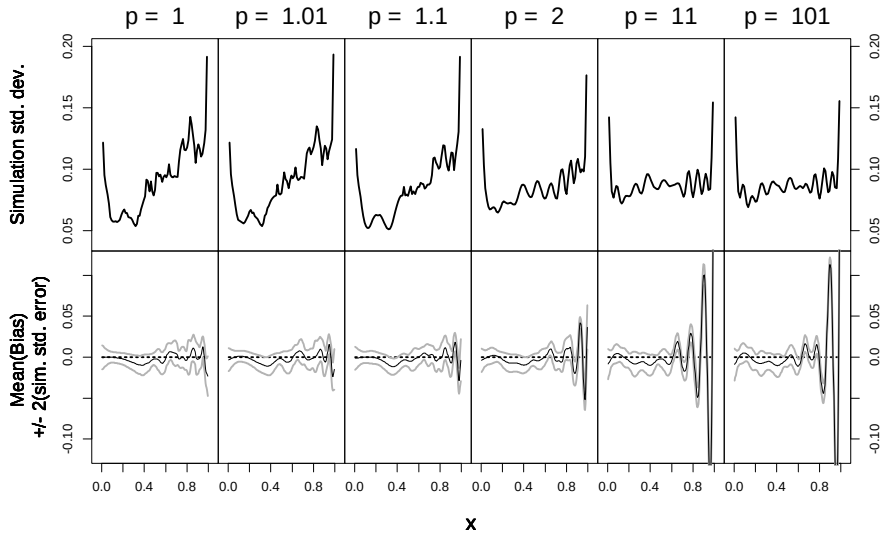

Figure 8: Standard deviation and bias. The **top panels** show the pointwise standard deviation of the 250 estimated curves. The **bottom panels** show the mean bias (dark lines) over the simulation replications and pointwise 95% confidence intervals (light lines) based on the simulation standard errors.

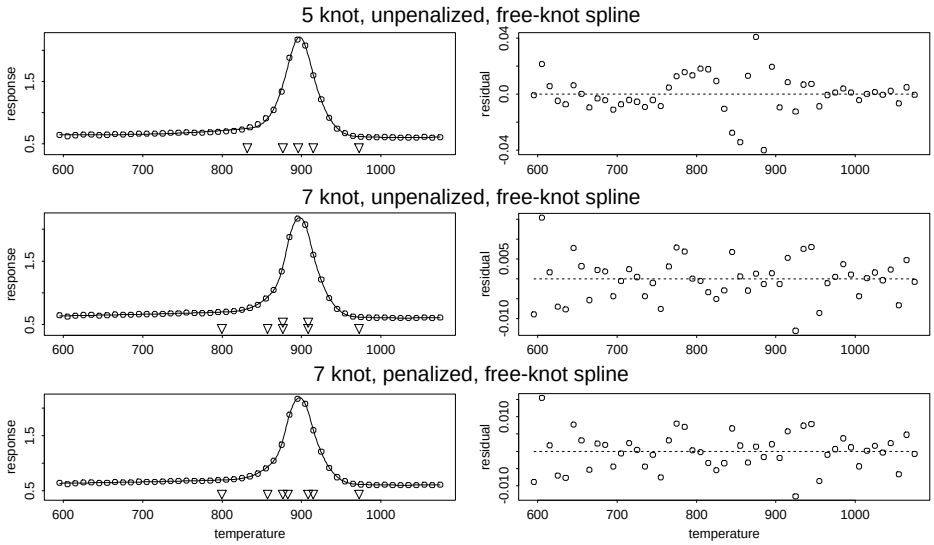

Figure 9: Penalized and unpenalized fits of the titanium heat data. The left panels show the data and the fitted splines with inverted triangles indicating the knot locations. The residuals from the fits are shown in the right panels.