SEQUENTIAL ADAPTIVE NONPARAMETRIC REGRESSION VIA H-SPLINES

Ronaldo Dias

Universidade Estadual de Campinas

Departamento de Estatística, IMECC,

Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL

Key Words: penalized least squares; B-splines; smoothing splines; Hellinger distance; generalized cross validation; hybrid splines.

ABSTRACT

The hybrid spline method (H-spline) introduced by Dias (1994) is a hybrid method of curve estimation which combines ideas of regression spline and smoothing spline methods. In the context of nonparametric regression and by using basis functions (B-splines), this method is much faster than smoothing spline methods (e.g. (Wahba, 1990)). The H-spline algorithm is designed to compute a solution of the penalized least square problem, where the smoothing parameter is updated jointly with the number of basis functions in a performance-oriented iteration. The algorithm increases the number of basis functions by one until the partial affinity between two consecutive estimates satisfies a constant determined empirically.

1. INTRODUCTION

There are many applications where a unknown function g of one or more variables and a set of measurements are given such that:

$$y_i = L_i g + \epsilon_i$$

where $L_1, \ldots, L_n$ are linear functionals defined on some linear space $\mathcal{H}$ containing g, and $\epsilon_1, \ldots, \epsilon_n$ are measurement errors usually assumed to be independently identically normal distributed with mean zero and unknown variance σ^2. Typically, the L_i will be point evaluation of the function g.

Straight forward least square fitting is often appropriate but it produces a function which is not sufficiently smooth for some data fitting problems. In such cases, it may be better to look for a function which minimizes a criterion that involves a combination of goodness of fit and an appropriate measure of smoothness. Such criterion is the well known penalized least square problem defined as the following: Finding the minimizer of the penalized least square equation which is,

$$A_\lambda(g) = \sum_{i=1}^{n} (y_i - L_i g)^2 + \lambda J(g), \tag{0.1}$$

where $J(g)$ is the penalty term usually taken as $\int (g'')^2$ and λ is the smoothing parameter which controls the trade off between fidelity to the data and smoothness.

It is of interest to estimate the curve g. For this assume that the points $t_1 < t_2 < \ldots, < t_n$ are in the interval $[a, b]$ such that $L_i g = g(t_i)$ and the function $g \in \mathcal{W}_2^2[a, b] = \{g : g' \text{ abs. continous and } \int (g'')^2 < \infty\}$. Define $\hat{g}$ as the estimate of the curve g so that:

$$\hat{g} = \arg \min_{g \in \mathcal{W}_2^2[a,b]} A_\lambda(g).$$

It is well known that $\hat{g}$ is necessarily a natural cubic spline with knots at t_i (see, for example, Silverman and Green (1994), Wahba (1981) and Craven and Wahba (1979)). Note that the roughness penalty $\int_a^b (g''(t))^2 dt$ has the property of reducing the problem of choosing g from an infinite-dimensional class of functions to a finite class of functions since $\hat{g}$ can be written as linear combination of basis functions. Although this fact might lead someone to think that the nonparametric regression problem becomes a parametric problem, one notices that the number of parameters can be as large as the number of observations, and there may be difficulties in interpreting a curve or surface g. Moreover, if the number of observations is large, the system of linear equations for exact solution is too expensive to solve.

In regression splines, the idea is to approximate g by a finite dimensional subspace of $\mathcal{W}$ spanned by basis functions $B_1, \ldots, B_K$, $K \leq n$. That is,

$$g \approx g_K = \sum_{j=1}^{K} c_j B_j,$$

where the parameter K controls the flexibility of the fitting. A very common choice for basis functions is the set of cubic B-splines (de Boor, 1978). The B-splines basis functions provide numerically superior scheme of computation and have the main feature that each B_j has compact support. In practice, it means that we obtain a stable evaluation of the resulting matrix with entries $B_{i,j} = B_j(x_i)$, for $j = 1, \ldots, K$ and $i = 1, \ldots, n$ is banded.

Unfortunately, the main difficulty when working with regression splines is to select the number and the positions of a sequence of breakpoints called knots where the piecewise cubic polynomials are tied to enforce continuity and lower order continuous derivatives. (See Schumaker (1972) for details.)

Regression splines are attractive because of their computational scheme where standard linear model techniques can be applied. But smoothness of the estimate cannot easily be varied continuously as functions of a single smoothing parameter (Hastie and Tibshirani, 1990).

The objective of this work is to extend a simulation study (Dias, 1996) using a hybrid method introduced by Dias (1994) which combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively. A previous study (Dias, 1996) based on a small number of simulations suggested that a sequential method could produce an efficient procedure. Recently, Luo and Wahba (1997) proposed a method based on GCV criterion to select the number of basis (which are not B-splines, although they are equivalent in the univariate case) among q fixed basis for a particular reproducing kernel in Hilbert space. Note that the computations for all q models have to be made in order to choose the best model. In this work, a sequential approach is to be taken. The H-splines procedure chooses the number of basis functions and the smoothing parameter sequentially and iteratively. The number of basis functions is increased by one until it satisfies a stop criterion which based on a closed related Hellinger's distance between (affinity) two consecutives estimates. Simulations have shown that this sequential scheme provides a computational efficient procedure and it usually stops around $O(\log n)$, where n is the sample size.

2. HYBRID SPLINES METHOD

In the smoothing techniques the number of basis functions is chosen to be as large as the number of observations and then let the choice of the smoothing parameter contolling the smoothing (Bates and Wahba, 1982). Here a different approach is to be taken. The H-splines method introduced by Dias (1994) in the case of nonparametric density estimation, combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively according to a criterion that is described below. With the point evaluation functionals $L_i g = g(t_i)$ the equation (0.1) becomes,

$$A_\lambda(g) = \sum_{i=1}^{n} (y_i - g(t_i))^2 + \lambda \int (g'')^2. \tag{0.2}$$

Assume that $g \approx g_K = \sum_{i=1}^{K} c_i B_i = Xc$ so that $g_K \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the space of natural cubic splines (NCS) spanned by the basis functions $\{B_i\}_{i=1}^{K}$ and X is a $n \times K$ matrix with entries $X_{ij} = B_i(t_j)$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n$. Then, the numerical problem is to find a vector $c = (c_1, \ldots, c_K)^T$ that minimizes,

$$A_\lambda^*(c) = \|y - Xc\|_2^2 + \lambda c^T \Omega c,$$

where Ω is $K \times K$ matrix with entries $\Omega_{ij} = \int B_i''(t) B_j''(t) dt$ and y is the vector $(y_1, \ldots, y_n)^T$. Standard calculations (de Boor, 1978) provide c as a solution of the following linear system $(X^T X + \lambda\Omega)c_\lambda = X^T y$. Note that the linear system now involves $K \times K$ matrices instead of using $n \times n$ matrices which is the case of smoothing splines. Both K and λ controls the trade off between smoothness and fidelity to the data. In particular, when $\lambda = 0$ we have the regression spline case, where K is the parameter that controls the flexibility of the fitting. To exemplify the action of K on the estimated curve, let us consider an example by simulation with $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$ with $\epsilon \sim N(0, .05)$.

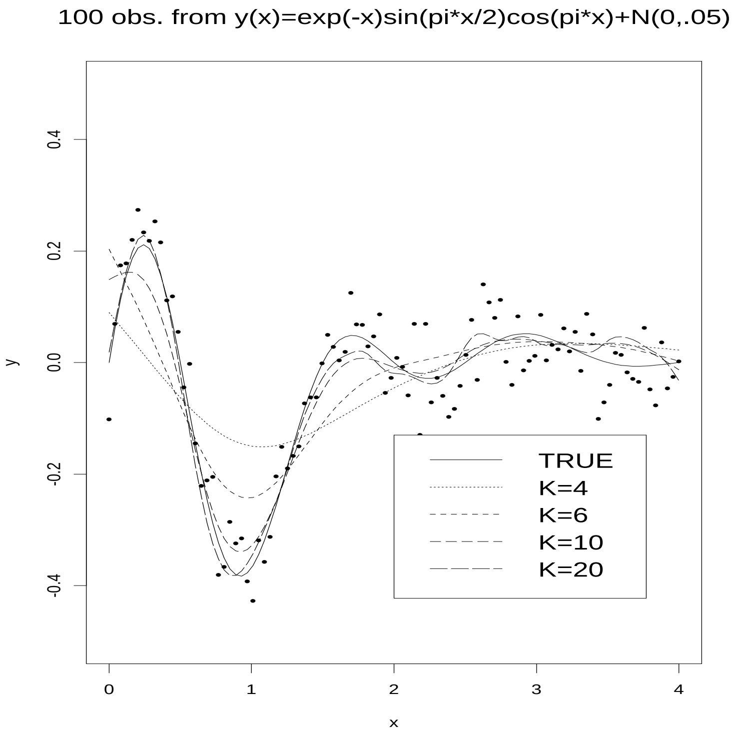

Figure 0.1: Five replications of $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$.

Figure 2.1 shows the effect of varying the number of basis functions on the estimation of the true curve. Note that the number of basis functions is the same as the number of knots since it is assumed that we are dealing with natural cubic splines space. Observe that small values of K make smoother the estimate and hence oversmoothing may occur. Large values of K may cause undersmoothing.

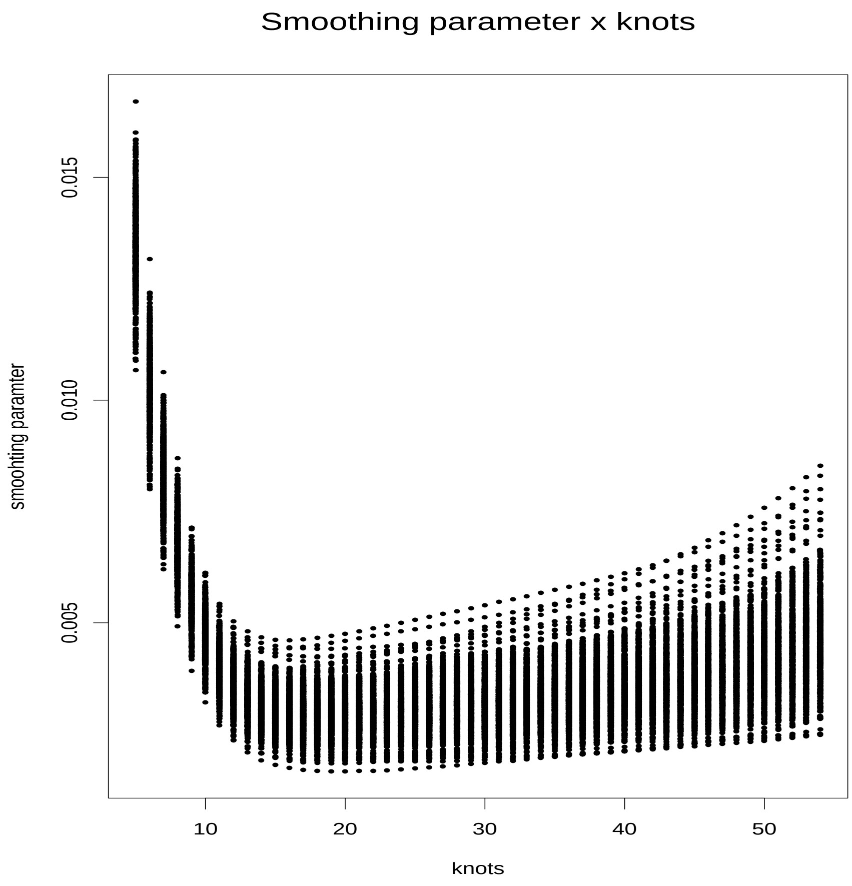

Figure 0.2: Five thousand replicates of $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$.

Figure 2.2 shows, for $\lambda > 0$, an example of the relationship between K and λ. Note that when the number of basis functions increases, the smoothing parameter decreases to a point and then it increases with K. That is, for large values of K, the smoothing parameter λ becomes larger in order to enforce smoothness.

Based on the facts described previously, the idea is to provide a procedure that estimates the smoothing parameter and the number of basis functions iteratively. Consider the following algorithm.

Algorithm 0.1

(1) Let K_0 be the initial number of basis functions and fix λ_0.

(2) Compute c_{λ_0} by solving $(X^TX + \lambda_0\Omega)c_{\lambda_0} = X^Ty$.

(3) Find $\hat{\lambda}$ which minimizes,

$$GCV(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - g_{K_0}(t_i))^2}{1 - n^{-1}tr(A(\lambda))},$$

where $A(\lambda) = X(X^TX + \lambda\Omega)^{-1}X^T$.

(4) Compute $g_{K_0,\hat{\lambda}} = A(\hat{\lambda})y$.

(5) Increment the number of basis functions by one and repeat steps (2) to (4) in order to get $g_{K_0+1,\hat{\lambda}}$.

(6) For a real number $\delta > 0$, if a distance $d(g_{K_0,\hat{\lambda}}, g_{K_0+1,\hat{\lambda}}) < \delta$, stop the procedure. The number δ can be determined empirically according to the particular distance $d(\cdot,\cdot)$.

Note that each time the number of basis functions K is incremented by one the numerator of GCV changes and hence this procedure provides an optimal smoothing parameter λ for the estimate g_K based on K basis functions.

The aim is to find a criterion able to tell when to stop increasing the number of basis functions. That is, to find the dimension of the natural cubic spline space where one is looking for the approximation of the solution of (0.1). For this, let us define the following transformation. Given any function in $\mathcal{W}_2^2[a,b]$, take

$$t_g = \frac{g^2}{\int g^2},$$

then $t_g \geq 0$ and $\int t_g = 1$. For any functions $f, g \in \mathcal{W}_2^2[a,b]$, define a pseudo distance closely related to the square of the Hellinger distance,

$$d^2(f,g) = \int (\sqrt{t_f} - \sqrt{t_g})^2 = 2(1 - \rho(f,g)),$$

where

$$\rho(f,g) = \int \sqrt{t_f t_g} = \int \sqrt{\frac{f^2 g^2}{\int f^2 \int g^2}} = \int \frac{|fg|}{\sqrt{\int f^2 \int g^2}},$$

is the affinity between f and g. It is not difficult to see that $0 \leq \rho(f,g) \leq 1$, $\forall f, g \in \mathcal{W}_2^2[a,b]$. Note that $d^2(f,g)$ is minimum when $\rho(f,g) = 1$, i.e., $(\int f^2 g^2)^{1/2} = \int |fg|$ only if $\alpha|f| + |g| = 0$ for some α.

Increasing the number of basis functions K by one, the procedure will stop when $g_{K,\hat{\lambda}} \approx g_{K+1,\hat{\lambda}}$ in the sense of the partial affinity,

$$\rho(g_{K,\hat{\lambda}}, g_{K+1,\hat{\lambda}}) = \frac{\int |g_{K,\hat{\lambda}} g_{K+1,\hat{\lambda}}|}{\sqrt{\int g^2_{K,\hat{\lambda}} g^2_{K+1,\hat{\lambda}}}} \approx 1,$$

where the dependence of λ on K is omitted for the sake of simplicity.

Simulations were performed in order to verify the behavior of the affinity and the partial affinity. Figure 2.3 shows a typical example given by the underlying function $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$.

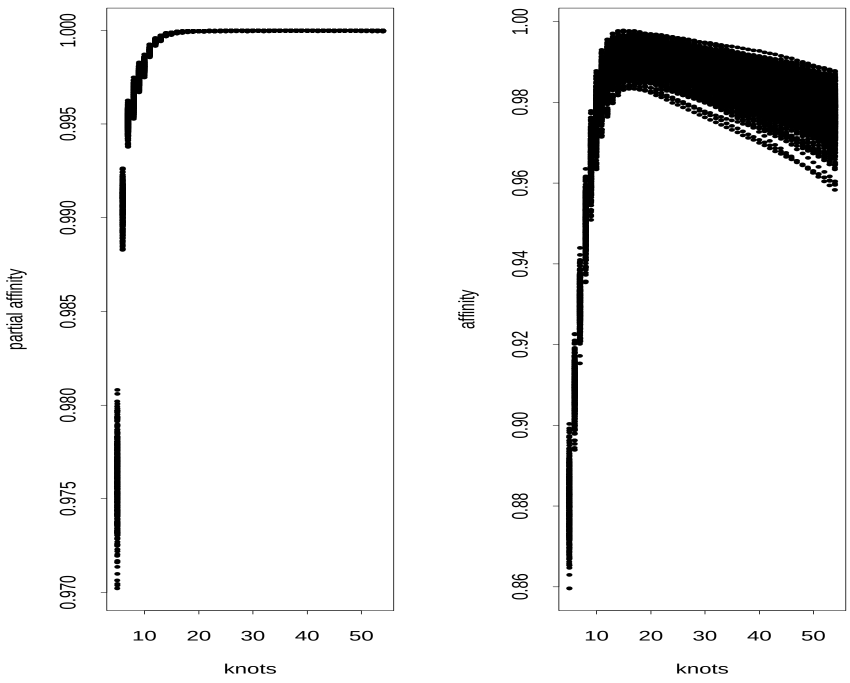

Figure 0.3: Five thousand replicates of the affinity and the partial affinity for adaptive nonparametric regression using H-splines with the true curve.

One may notice that the affinity is a concave function of the number of basis functions (knots) and the partial affinity approaches one quickly. Moreover, numerical experiments have shown that the maximum of the affinity and the stabilization of the partial affinity coincide. That means, increasing the K arbitrarily not only increases the computational cost but also does not provide the best fitted curve (in the pseudo Hellinger norm).

It would be useful to have the distribution of the affinity between the true curve and the estimate produced by the adaptive H-splines method. A previous study (Dias, 1996) showed an empirical unimodal density with support on $[0, 1]$ skewed to the left suggesting a beta model. To illustrate, five thousand replicates with sample size 20,100,200 and 500 were taken from a test function $y_i = x_i^3 + \epsilon_i$, where and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, 5)$.

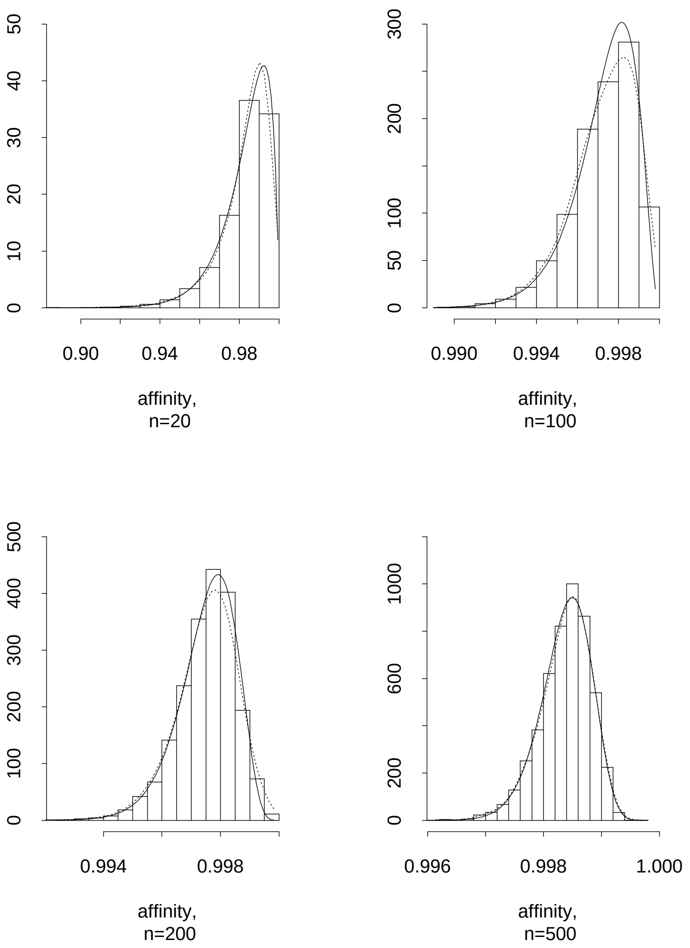

Figure 0.4: Density estimates of the affinity based on five thousand replicates of the curve $y_i = x_i^3 + \epsilon_i$ with $\epsilon_i \sim N(0, .5)$. Solid line is a density estimate using beta model and dotted line is a nonparametric density estimate.

Figure 2.4 shows that the empirical affinity distribution (unimodal, skewed to the left with range between 0 and 1), a nonparametric density estimate using kernel method and a parametric one using a beta model whose parameters were estimated using method of the moments. Similar results were obtained for several other test functions and some of them are exhibited on Figure 2.5 which brings more evidences to support a beta model.

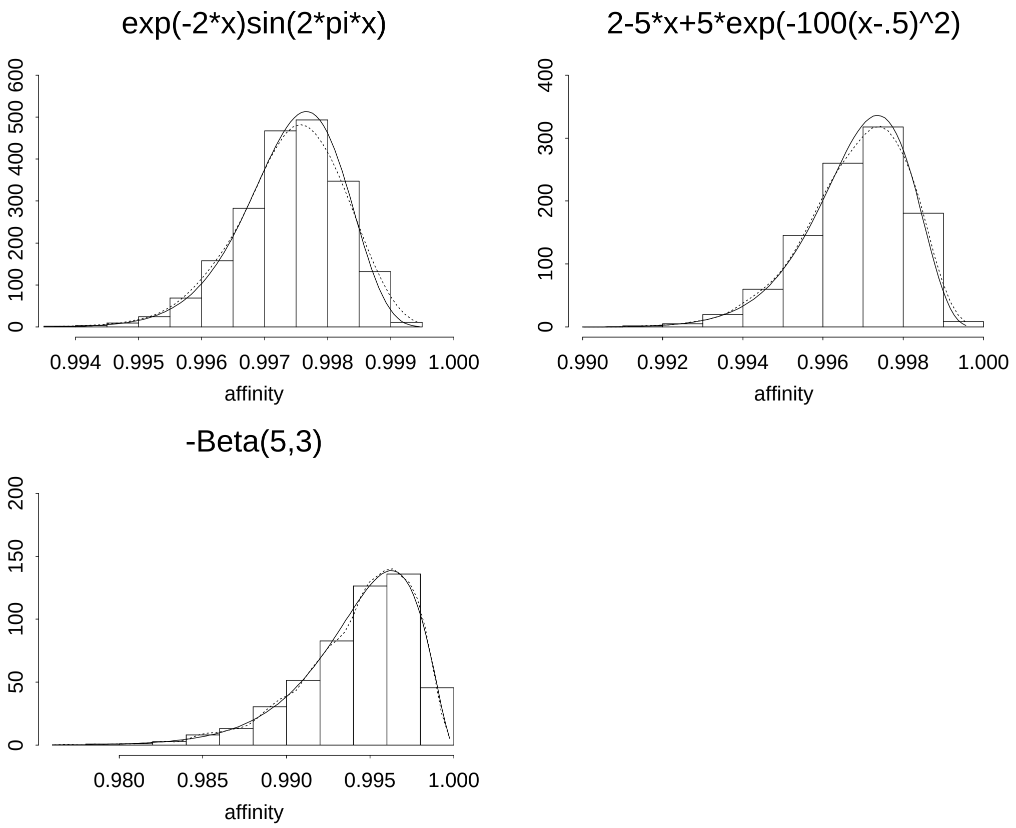

Figure 0.5: Five thousand replicates of the affinity using different test functions and beta models adjusted with parameters estimated by method of moments.

3. EXAMPLES AND COMPARISONS

In this section examples and comparisons are presented in order to verify, empirically, the performance of the sequential adaptive H-splines method applied to the nonparametric regression models. Several simulations were done and typical examples of the simulated regression curves are shown below. The basic idea is to show that H-splines approach with sequential method of selecting the number of basis functions can at least estimate regression curves as good as regression splines and smoothing splines methods under different degrees of difficulties of the proposed regression models.

Figure 3.1 shows that, in general, H-splines method has similar performance

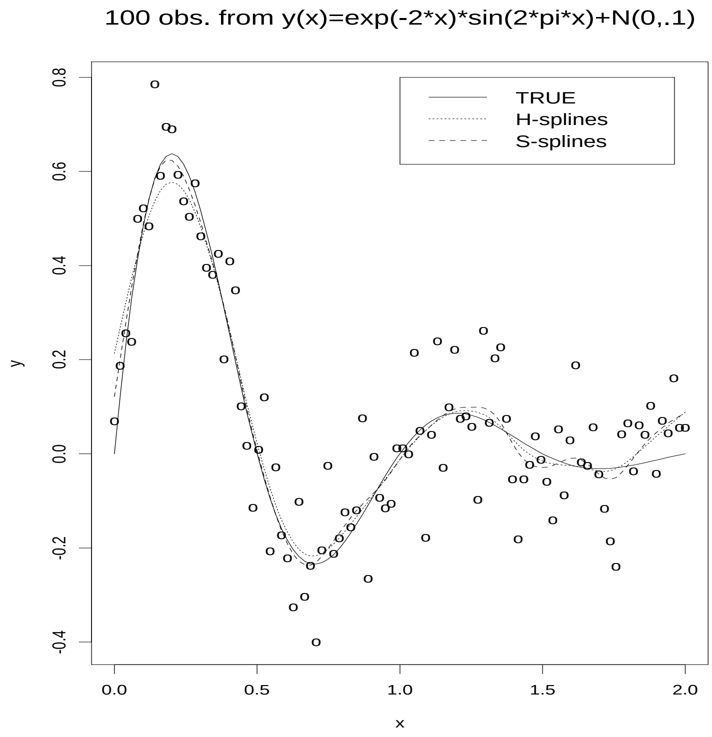

Figure 0.6: A comparison between smoothing splines (S-splines) and hybrid splines (H-splines) methods.

as smoothing splines. But as mentioned before the H-splines approach solves a linear system of order K while smoothing splines must have to solve a linear system of order $n \geq K$.

Visually, one may observe in figure 3.2 that H-splines has a better performance than smoothing splines approach, in the sense that, H-splines produces an estimate curve closer to the true one than smoothing splines does.

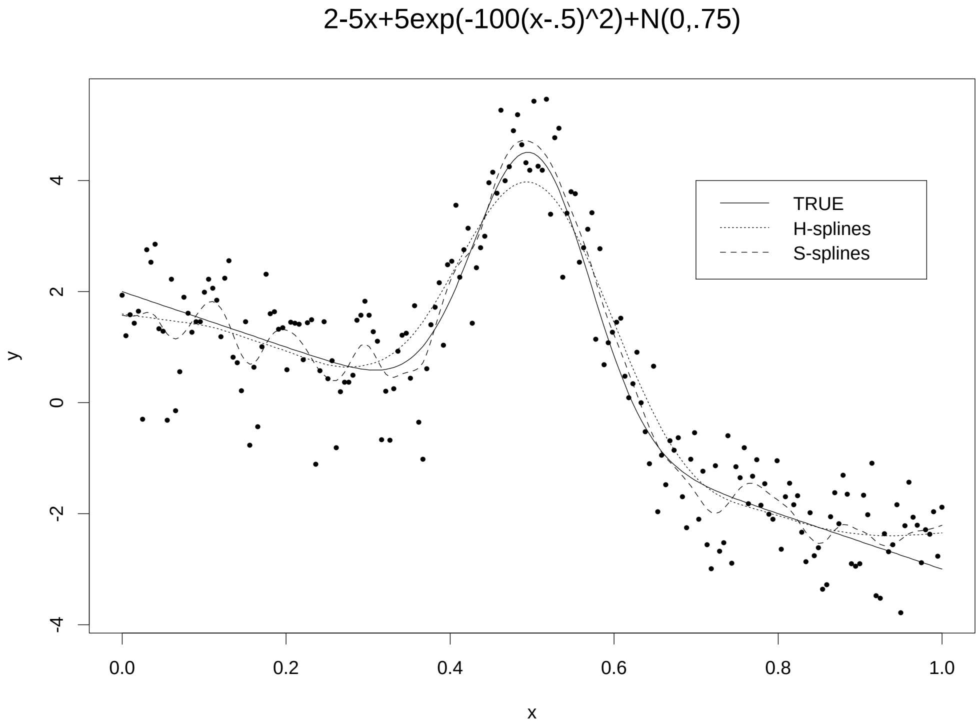

Figure 0.7: A comparison between smoothing splines (S-splines) and hybrid splines (H-splines) methods.

Figure 3.3 exhibits an example comparing smoothing splines, H-splines and regression splines using B-splines as basis functions. For regression splines, the same criterion of H-splines was applied in order to determine the number of basis functions to be in the fitting. One may notice that this criterion can also be used to fit nonparametric regression models to the data without fixing the total number of knots a priori.

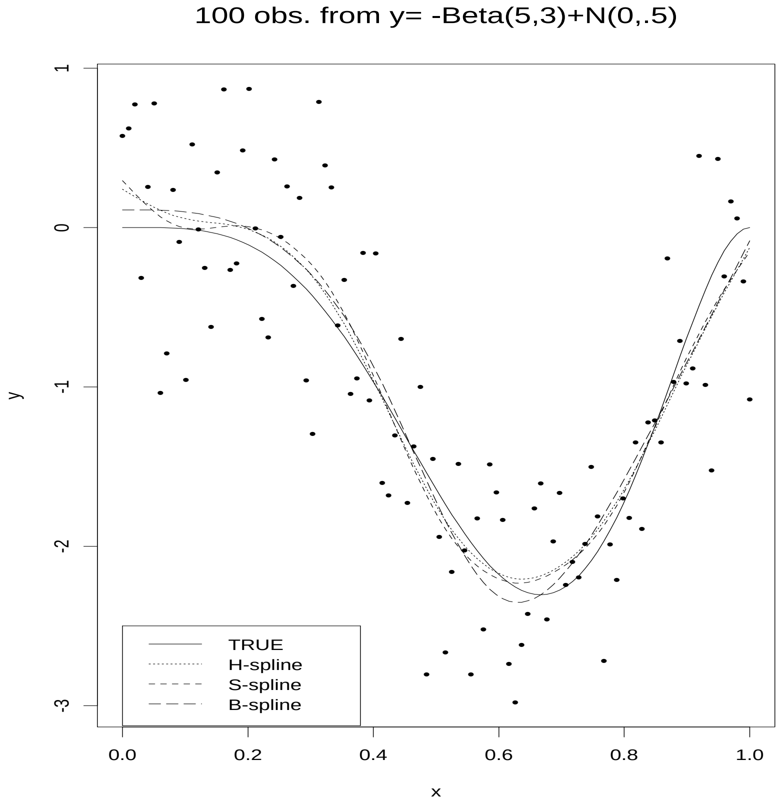

Figure 0.8: Comparing Smoothing splines (S-splines), hybrid splines (H-splines) and regression splines (B-splines) for $y(x) = -Beta(5,3)x + \epsilon$, where $Beta(5,3)$ the density of a random variable Beta with parameters 5 and 3 respectively.

4. CONCLUSION

Simulations have shown that the sequential adaptive H-splines method for nonparametric regression has, at least, similar performance of the smoothing splines method for large data sets with the advantage of solving a $K \times K$ linear system, instead of $n \times n$ system, where $K \ll n$. For small data sets and the underlying regression curve with more structure (several bumps) it seems to have a better performance than smoothing algorithms that uses all data points for the fitting. The use of an objective criterion based on partial affinity gives more flexibility to adjust curves to the data. This criterion can be used also under the context of regression splines and in the generalized additive models to select the number of knots. Moreover, promising results on the distributional properties of the affinity would give the theoretical properties of the H-splines approach for nonparametric regression needed for inference purposes.

ACKNOWLEDGEMENT

I would like to thank the Associate editor and an anonymous referee for the comments and suggestions that made this work better and clearer. This research was partially support by FAPESP grant no. 95/4996-3.

References

Bates, D. and Wahba, G. (1982). *Computational Methods for Generalized Cross-Validation with large data sets*, Academic Press, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

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

Dias, R. (1994). Density estimation via h-splines, *University of Wisconsin-Madison.* Ph.D. dissertation.

Dias, R. (1996). Sequential adaptive nonparametric regression via h-splines. Technical Report RP 43/96, University of Campinas, June 1996. Submitted.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall.

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

Schumaker, L. L. (1972). *Spline Functions and Aproximation theory*, Birkhauser.

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

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Ann. of Statistics* **9**: 146–156.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM:PA.