Bounded Optimal Knots for Regression Splines

Nicolas Molinari*, Jean-François Durand, Robert Sabatier [a,b,c]

[a] *Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique, 641 avenue Gaston Giraud, 34093 Montpellier, France*

[b] *Unité de Biométrie, ENSAM-INRA, 2, place Viala, 34060 Montpellier, France*

[c] *Laboratoire de Physique Moléculaire et Structurale, 15 av. Ch. Flahaut, 34060 Montpellier, France*

Abstract

Using a B-spline representation for splines with knots seen as free variables, the approximation to data by splines improves greatly. The main limitations are the presence of too many local optima in the univariate regression context, and it becomes even worse in multivariate additive modeling. When the number of knots is a priori fixed, we present a simple algorithm to select their location subject to box constraints for computing least-squares spline approximations. Despite its simplicity, or perhaps because of it, the method is comparable with other more sophisticated techniques and is very attractive for a small number of variables, as shown in the examples. In a complete algorithm, the BIC and AIC criteria are evaluated for choosing the number of knots as well as the degree of the splines.

Key words: additive models, bound constrained optimization, free knots selection, surface estimation, AIC, BIC.

1 Introduction

The approximation of functions by splines has long been known to improve dramatically if the knots are free parameters. A serious problem is the existence of many stationary points for the least-squares objective function, and the apparent impossibility of deciding when the global optimum has been

* Corresponding author: Nicolas Molinari, Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique, 641 avenue Gaston Giraud, 34093 Montpellier, France, fax: +33 4 67 54 27 31, email address: molinari@iurc.montp.inserm.fr

found. Moreover, another disadvantage of free knots is that the optimal knot vector often includes identical knots which yield a non smooth behavior of the predicted curve. For the commonly used splines of degree three, coalescing knots cause a discontinuous second derivative and 4 identical knots allow a discontinuity in the fitted curve itself. If an assumption of smoothness for the true underlying function is warranted, one may then have to exclude solutions with duplicate knots. Free knot splines have not been as popular as might be expected in part for these reasons. Another problem is that analytic expressions for optimal knot locations, or even for general characteristics of optimal knot distributions, are not easy to derive.

Computationally, things are different: there exist several algorithms to find knot locations. Adaptive regression splines methods consider a predictor subset selection problem. Friedman and Silverman (1989) developed a method called TURBO and, in the discussion, Hastie proposes an alternative method based on the ACE type backfitting. The Delete-Knot/Cross-Validation method (DKCV) of Breiman (1993) overcomes the difficulties encountered by ACE on small noisy data sets. Friedman (1991) introduced multivariate adaptive regression splines (MARS), which is a polynomial spline methodology for estimating the regression function that involves interactions. PolyMARS (Stone et al., 1997) allows multiresponse data sets. DKCV minimizes the least squares criterion by greedy backward deletion of knots; TURBO, MARS and PolyMARS applies stepwise addition and deletion to select a set of knots. The reduction to a prespecified candidate knot sites (data points or quantiles of the input data) is common, nevertheless it is possible to run these algorithms with the knots taken as continuous variables. However, knots located at the data points are not necessarily a good choice, especially in regions of little or no data. Figure 1 illustrates this fact with optimal knots falling in an empty area.

In contrast to adaptive regression splines, choosing knot locations in a free knot spline is a parameter estimation problem. Gallant and Fuller (1973) use an iterative algorithm based on the Gauss-Newton method to solve the problem. If no approximate knot location can be deduced from inspection of the data, penalized nonlinear least-squares can be used to obtain the knot estimates. Jupp (1978) introduces a transformation of the knots to avoid the "lethargy" phenomenon: this transformation pushes the knot set boundaries to infinity, making it impossible for the free knots to coalesce. With the same approach, Lindstrom (1999), Dierckx (1993) and Guertin (1992) penalize the distance from equidistant knots.
Using a Bayesian approach, Denison et al.(1998) compute a joint distribution over both the number and the position of the knots, thus allowing the computation of the posterior distribution, using a reversible jump Markov chain Monte Carlo method. The Bayesian Subset Selection method (BSS) needs prespecified candidate knot sites (usually the design points), and it appears to be robust for reasonable choices of the prior distribution parameters. Smith and

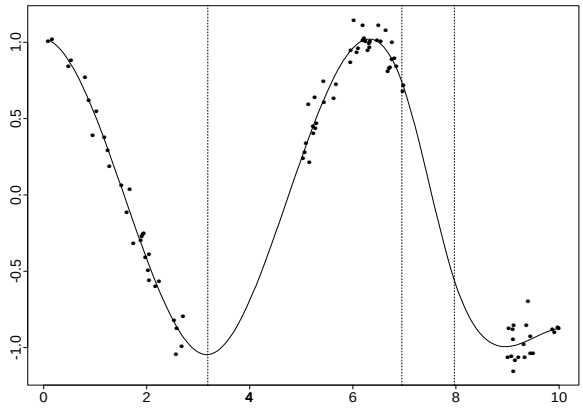

Fig. 1. Data presented in section 4, spline estimation with 3 optimal knots. Vertical lines indicate knot locations. Note that optimal knots are not on data points.

Kohn (1996) apply the Bayesian machinery to univariate curve fitting and additive modeling. Moreover in a bivariate context, the Markov Chain produced by Smith and Kohn (1997) is efficient and can search through a large number of models.

The present paper proposes a simple and rather computationally time efficient method for free knots B-spline regression models. To solve the difficult problem of optimal knot locations, we first assume that the number of knots is a priori fixed. We imagine that the knots belong to disjoint intervals. The box constrained minimizing algorithm leads to a computationally efficient method for exploring the local minima (knot locations) which in turn avoids the coalescing of free knots. The optimal spline model dimension (the number of knots) is determined through classical information criterion.
In section 2, we recall the free knots problem and the consequences of the "lethargy" theorem. Section 3 proposes a new algorithm to compute knot locations in simple regression by least-squares splines. Using this method, one has only to decide on the number of knots. A complete algorithm with an automatic number of knots and spline degree determination is proposed through a selection model with the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). Section 4 presents the multiple regression case. The method is generalized to the simple additive model. In this case, the number of variable increases the computation time. When the model contains interactions terms, the computation time increases significantly. The tensor product regression splines are presented and illustrated with classical examples.

2 Fixed and free knots for least-squares splines

Using spline functions in a simple or multiple regression model allows the investigation of nonlinear effects with continuous covariates. In particular, B-spline basis functions are appropriate in this case due to the fact that they are numerically well conditioned, and also because they achieve a local sensitivity to data. With fixed knots, the least-squares splines approximation is equivalent to a linear problem. On the other hand, for a fixed number of distinct knots whose location has to be optimized, one is usually faced with local optima located on multiple or coalescent knots which corresponds to a degenerate case.

2.1 B-spline functions

Let $(\xi_0 =)a < \xi_1 < \xi_2 < \cdots < \xi_K < b(= \xi_{K+1})$ be a subdivision of K distinct points on the interval $[a, b]$ on which the x variable is valued and denote these points the "knots". The spline function $s(x)$ used to transform the x variable is a polynomial of degree d (or order $d+1$) on any interval $[\xi_{i-1}, \xi_i]$, and has $d-1$ continuous derivatives on the open interval (a, b). For each fixed sequence of knots $\xi = (\xi_1, \xi_2, \cdots, \xi_K)'$, the set of such splines is a linear space of functions with $K + d + 1$ free parameters (de Boor, 1978). A useful basis $\{B_l(., \xi)\}_{l=1,\cdots,K+d+1}$, for this linear space is given by Schoenberg's B-splines, or *Basic*-splines (Curry and Schoenberg, 1966). De Boor (1978) gives an algorithm to compute B-splines of any degree from B-splines of lower degree.

We can now write a spline as :

$$s(x, \beta, \xi) = \sum_{l=1}^{K+d+1} \beta_l B_l(x, \xi),$$

where the vector $\beta = (\beta_1, \cdots, \beta_{K+d+1})'$ of coefficients and the vector ξ of knots are considered as tuning parameters.

2.2 The lethargy problem in simple regression

Let $\{x_i, y_i\}_{i=1,\cdots,n}$ be a set of n observations ranging over $[a, b] \times I\!R$. Denote $B(\xi) = \{B_l(x_i, \xi)\}_{i=1,\cdots,n}^{l=1,\cdots,K+d+1}$, the $n \times (K + d + 1)$ matrix of sampled basis functions, and $y = (y_1, \cdots, y_n)'$.

When the knots ξ are fixed, the spline fit to the data is accomplished via a straightforward linear least-squares problem

$$\hat{\beta}(\xi) = \arg\min_{\beta} \|y - B(\xi)\beta\|^2 = (B(\xi)^{'} B(\xi))^{+} B(\xi)^{'} y, \tag{1}$$

where $\|.\|$ is the Euclidean norm, and B^{+} is the Moore-Penrose inverse of B. Then, $s(x, \beta, \xi)$ may be estimated by $s(x, \hat{\beta}(\xi), \xi)$, the least-squares spline (LSS) estimator of the regression function. In this paper, we want to select the coefficient vector with minimum Euclidean norm. Note that it is common to use a penalty that has to do with the smoothness of the curve. With the classical smoothing spline, a parameter controls the trade-off between the fit to the data and the smoothness of the estimator.

For fixed K, when knots are free variables, the class of splines is no longer linear but form a mixture of linear and nonlinear parameters. Then, the non-parametric least-squares problem may be written as

$$\min_{\beta \ \xi \in [a,b]^K} \|y - B(\xi)\beta\|^2. \tag{2}$$

For each fixed ξ the problem (2) reduces to (1). In fact, Golub and Pereyra (1973) show that the solution to (2) is

$$\min_{\xi \in [a,b]^K} \|y - B(\xi)\hat{\beta}(\xi)\|^2, \tag{3}$$

where $\hat{\beta}(\xi)$ is the solution to the linear problem (1). Henceforth, from now on we denote the objective function

$$F(\xi) = \|y - B(\xi)\hat{\beta}(\xi)\|^2.$$

The following "lethargy" property (Jupp 1975) is intrinsic to free knots problems and affects the stability and effective computation of the optimal knots.

The lethargy theorem: *Denote* $S_K[a,b] = \{\xi \in I\!R^K; a < \xi_1 < \xi_2 < \cdots < \xi_K < b\}$ *the open simplex of knots, and* $S_K^{(p)}$ *the pth (open) main face of* $S_K[a,b]$. $S_K^{(p)}$ *is defined by the system* $(\xi_j - \xi_{j-1}) > 0$, *for* $j \neq p$, *and* $\xi_p = \xi_{p-1}$. *On the pth main face,* $S_K^{(p)}$,

$$n_p^{'} \nabla F(\xi) = 0 \text{ for } p = 2, \cdots, K,$$

where n_p *is the unit outward normal to* $S_K^{(p)}$, *and* $\nabla F(\xi)$ *the gradient of* $F(\xi)$.

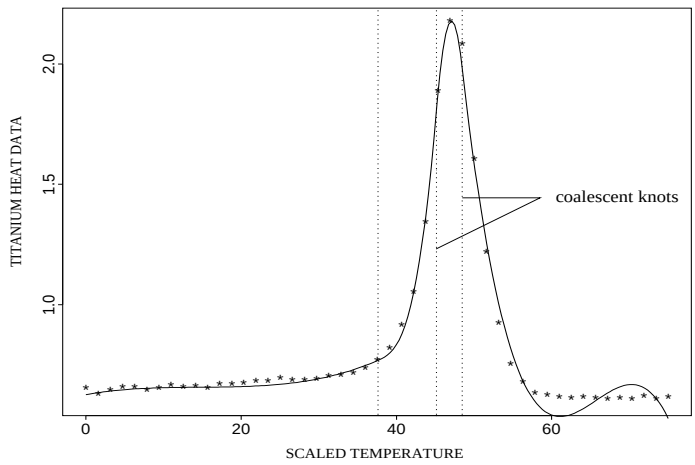

Fig. 2. Titanium heat data approximated with a cubic spline. A usual gradient method locates the 5 knots on only 3 distinct positions. Vertical lines indicate the location of knots.

The first consequence of this theorem is the existence of many stationary points of $F(\xi)$ on the faces $S_K^{(p)}$. The presence of many stationary points implies the poor convergence, or "lethargy" property, of algorithms that attempt to solve the free knots problem when they are near the boundaries $S_K^{(p)}$. The second consequence is about what replicate knots mean. It implies lower smoothness of the spline estimation (see de Boor, 1978 or Schumaker, 1981). Figure 3 shows $S_2[a,b]$ which is a triangle. The "lethargy" property is illustrated on the titanium heat data (de Boor and Rice , 1968) which include 49 measurements of a thermal property of titanium. Figure 2 shows the approximation by cubic splines corresponding to a local minimum with some coalescent knots. Note that with very few knots, the lethargy is also a problem. For example, when we estimate $f(x) = \sin(4\pi x)$, for $x \sim \mathcal{U}[0,1]$, by a spline of degree 1 with 2 knots, $\xi = (0.5, 0.5)$ corresponds to a local minimum.

3 Bounded Optimal Knots

To avoid the lethargy problem, we require the knot ξ_i, to lie within some window $[l_i, u_i]$ and we further impose the constraint that the windows are disconnected.

Let $l = (l_1, \cdots, l_K)$ and $u = (u_1, \cdots, u_K)$ be the vectors of lower and upper bounds. Disjoint windows implies that we take $l_{i+1} - u_i = \varepsilon > 0$ for $i = 1, \cdots, K-1$ and $l_1 - a = b - u_K = \varepsilon$. Note that $\lim_{\varepsilon \to 0} \bigcup_{i=1,K} [l_i, u_i] = [a, b]$.
When the windows are fixed, the *bounded optimal knot* problem is to find

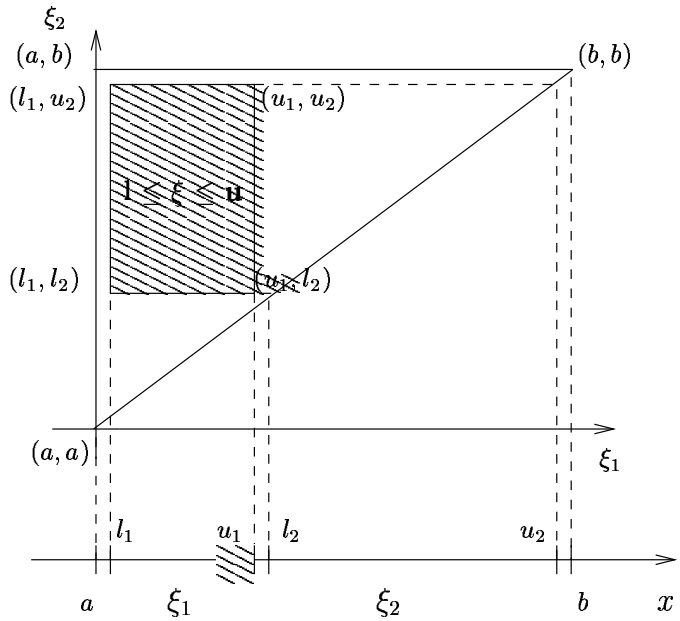

Fig. 3. The simplex $S_2[a,b]$ of variable knots $\xi = (\xi_1, \xi_2)$, and the shaded box of constrained knots associated with the windows for the variable x.

$$\hat{\xi}(l,u) = \arg \min_{l \leq \xi \leq u} F(\xi). \tag{4}$$

Clearly $\bigcup_{i=1,K}[l_i,u_i] \subset [a,b]$ for $i = 1,\cdots,K$, and (4) does not necessarily provide the global minimum to (3) because

$$\min_{l \leq \xi \leq u} F(\xi) \geq \min_{\xi \in ([a,b])^K} F(\xi).$$

However, problem (4) is very easy to solve by using classical fast algorithms based on the Fortran functions *dmnfb*, *dmngb* and *dmnhb* (Gay, 1983, 1984, A T & T, 1984) from NETLIB (Dongarra and Grosse, 1987). The visual examination of the spline approximation allows the user to experiment different choices for selecting the windows. Figure 3 shows a two-knot example of the space where (4) is to be solved.

In our algorithm, ε represents the minimal distance between two knots. When $\varepsilon > 0$, we avoid the lethargy phenomenon. If the user chooses $\varepsilon = 0$, the algorithm may identify replicate knots. Small values of ε are used to obtain compact constrained sets in the optimization problems without eliminating significant regions of the simplex domain. We have varied this parameter in the course of preliminary experiments. On the presented simulations, we did not encounter any effect on mean squared error. We heuristically used the minimal distance between two successive points divided by two for no replicate data, and the range of $[a,b]$ divided by 10^3 if not. Moreover, ε allows to introduce a minimum span to avoid that multiple knots (or ε-spaced knots) occur between two data points. For example with $\varepsilon > \min_{i,j} |x_i - x_j|$, multi-

ple ε-spaced knots cannot fall between two data points. TURBO, MARS and PolyMARS also work with a minimum span.

3.1 A non deterministic algorithm

To explore most of the local optima, our strategy presented below provides an automatic selection of the windows. We construct a sequence $\{\hat{\xi}(l^{(i)}, u^{(i)})\}_i$ of N solutions to (4) based on N independent uniformly distributed partitions $\{l_1^{(i)}, u_1^{(i)}, \cdots, l_K^{(i)}, u_K^{(i)}\}_{i=1,N}$. In fact, because upper and lower bounds are explicitly linked ($u_j = l_{j+1} - \varepsilon$), we need only to generate a sequence $\{l_2^{(i)}, \cdots, l_K^{(i)}\}_{i=1,N}$ of $K-1$ uniformly drawn lower bounds. For sufficiently large N, we expect that

$$\hat{\xi} = \arg \min_{i=1,N} F(\hat{\xi}(l^{(i)}, u^{(i)}))$$

provides a good approximation to the optimal knot locations. Clearly, the number N of experiments is unknown, and we heuristically use $N = 100$. The preceding algorithm, BOK (for *Bounded Optimal Knots*, Algorithm 1), constructs a sequence of N sets of windows uniformly distributed on $[a, b]$. The two following heuristics are added to this algorithm to reduce the number N of trials.

Algorithm 1: *Bounded Optimal Knots Algorithm*

Inputs: $X, Y, d, K, N, \varepsilon = 10^{-3}$

for $i = 1$ to N **do**
  $l^{(i)} \sim (\mathcal{U}[a, b])^{K-1}$
  compute $u^{(i)}$
  bounded minimization: $\xi^{(i)} \leftarrow \arg \min_{l^{(i)} \leq \xi \leq u^{(i)}} F(\xi)$
  compute $F(\xi^{(i)})$
end for

$F(\hat{\xi}) \leftarrow \min\{F(\xi^{(1)}), \cdots, F(\xi^{(N)})\}$

3.2 Two heuristics to reduce the computational cost

Two sets of nearly identical windows will clearly provide the same set of knots. Moreover, one is often faced with knots located at the boundaries of the win-

dows, thus indicating that the concerned window is not well adapted and must be relaxed.
The following procedure referred to as the "Evolutionary Bounded Optimal Knots" (EBOK) algorithm, presented in Algorithm 2 (see Appendix A), addresses the two preceding points.
Concerning the first point, the ith candidate $l^{(i)} = (l_2^{(i)}, \cdots, l_K^{(i)})$ will be discarded if it is too close to one of the previous vectors. More precisely, if there exits an $l^{(j)}$ in $\{l^{(1)}, \cdots, l^{(i-1)}\}$ such that

$$d_\infty(l^{(i)}, l^{(j)}) < \rho, \tag{5}$$

where $d_\infty(l^{(i)}, l^{(j)}) = \max_{l \in \{2,\cdots,K\}} |l_l^{(i)} - l_l^{(j)}|$ is the sup-distance, then $l^{(i)}$ is discarded if $l^{(i)} \in \bigcup_{j=1,i-1} \mathcal{B}_\infty(l^{(j)}, \rho)$, where $\mathcal{B}_\infty(l, \rho)$ denotes the ball of radius ρ and centered at l. On the contrary, the ith candidate $l^{(i)}$ will be accepted if it differs significantly from all the preceding ones. Because equidistant knots are commonly used without a priori information, the first candidate $l^{(1)}$ involves equidistant coordinates. Our experience with the value of ρ suggests that ρ is not a preponderant tuning parameter; its only role is to limit the number of potential candidates for $l^{(i)}$. When ρ is large enough ($\geq \frac{b-a}{2}$) the only accepted vector is $l^{(1)}$. Small values of ρ is preferable, since $\rho = 0$ implies that all candidates are accepted, thus corresponding to the BOK procedure. Clearly, ρ should slightly increase with both values of $b - a$ and K. The heuristic used with $K < 10$ is $\rho = \frac{K-1}{20}(b - a)$.

Concerning the second point, if a minimization stops at a knot confounded with a bound, the algorithm modifies this bound to carry on minimizing. The candidate window is enlarged when possible, by locating the concerned bound half way between its preceding position and the adjacent knot. Then we restart the minimization and repeat this procedure until all the knots stabilize inside the windows. If the algorithm does not produce a knot vector with knots interior to the windows, the procedure is stopped if 100 iterations ($count = 100$) fail to stabilize knots inside the windows and other window candidates are considered. Figure 4 illustrates the use of the heuristic for modifying the box of constrained knots when one knot is located on the edge of the box.
The program is sometimes unable to find points inside the windows. It may indicate that the global minimum of $F(\xi)$ actually involves duplicated knots. Because our approach can only place ε-separated knots, it will not converge to the optimal solution. However, even if, at a given step, the procedure has not converged in 100 iterations and is restarted with other window candidates, the corresponding suboptimal solution is kept to be compared to the solutions obtained by the other trials.

The titanium heat data illustrates for the simple regression context that the

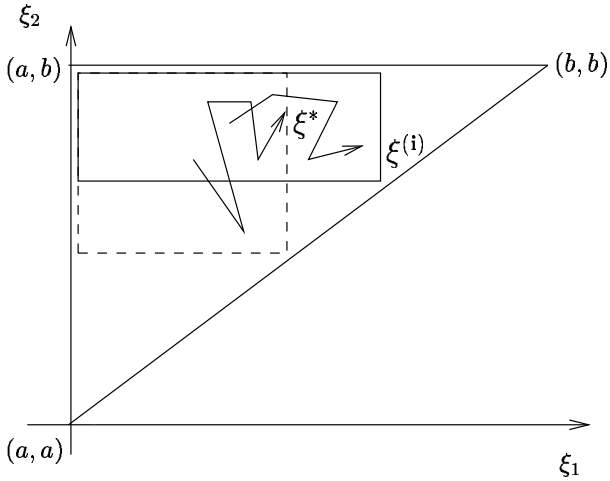

Fig. 4. The $S_2[a,b]$ simplex with 2 successive boxes dashed rectangle shows the initial bounds that lead to ξ^* located on the edge; the modified solid box provides interior knots $\xi^{(i)}$.

algorithm is effective in finding the 5 knots global optimum already explored in Jupp (1978). This example was implemented on an Ultra-Sparc station, through the *BOK* and *EBOK* (for *Evolutive Bounded Optimal Knots*) functions in S-Plus© (MathSoft, 1996) that use the native function *nlminb* applied successively on N sets of windows. By default, initial knots are located at the center of the windows and your method requires the user to specify only one input (the number of experiments, N, the number of knots, K, the degree, d, of the spline polynomials) and call the function $BOK(x,\ y,\ N,\ K,\ d)$ or $EBOK(x,\ y,\ N,\ K,\ d)$. The error measure we use here is

$$\|e\|^2 = (\frac{1}{n-1}\sum_{i=1}^{n} w_i|e_i|^2)^{\frac{1}{2}}$$

(where $w_1 = w_n = \frac{1}{2}$, and $w_i = 1$ otherwise). Using *BOK* with $(N,K,d) = (500,5,3)$, the global optimum of Jupp (1978) is obtained after 8400 seconds of cpu time, but is obtained after 300 seconds of cpu time with $EBOK(x,y,5,5,3)$. The fitting curve is presented in Figure 5 and a comparison of results of the different methods applied to the data is summarized in Table 1. Selecting the initial knots located at the center of the windows did not affect the convergence of the procedure in all the examples we tried.

3.3 Model selection

The method presented in the preceding section assumes that the number K of knots is fixed. Several methods for choosing K have appeared in the literature. We propose to compute spline models with different numbers of optimized

Table 1
Comparison between the Jupp algorithm, BOK and EBOK on titanium heat data.

	Jupp	BOK	EBOK
Final point	$\hat{\xi}$	$\hat{\xi}$	$\hat{\xi}$
where $\hat{\xi} = (37.6, 43.9, 47.4, 50.2, 59.2)$			
Residual error	0.1249	0.1249	0.1249
N	.	500	10
Need well located initial points	Yes	No	No

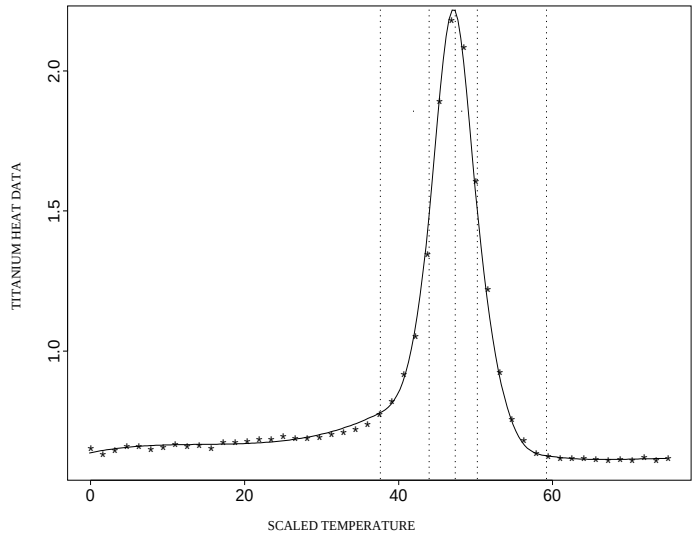

Fig. 5. Titanium heat data approximated by using a cubic spline with knots located by the *BOK* algorithm. The location of optimal knots is indicated by vertical lines.

knots and to select the model which minimizes the BIC (Schwarz, 1978) or the AIC (Akaike, 1974) criterion. Let K denotes the largest number of knots used. The procedure to determine the appropriate model is summarized in the following algorithm:

For $k = 1$ to K do
$\quad F(\xi_{(k)}) \leftarrow EBOK(X, Y, d, k, N, \varepsilon, \rho)$
end for
$F(\hat{\xi}) \leftarrow \arg\min_{k=1,\ldots,K}\{AIC(F(\xi_{(k)}))\}$

In this algorithm, we use the AIC criterion. Clearly a BIC or any other criterion can be used. The number of parameters for free knot spline functions is a much debated question. In his paper, Owen (1991) presented a summary on this subject. The cost one charge per knot depend on the smoothness of the space and the extent of the search. For piecewise linear functions, the cost is between 2 and 3 degrees of freedom. For a smooth model, say cubic splines, one knot is charged 2 degrees of freedom, providing we are not in a degenerate

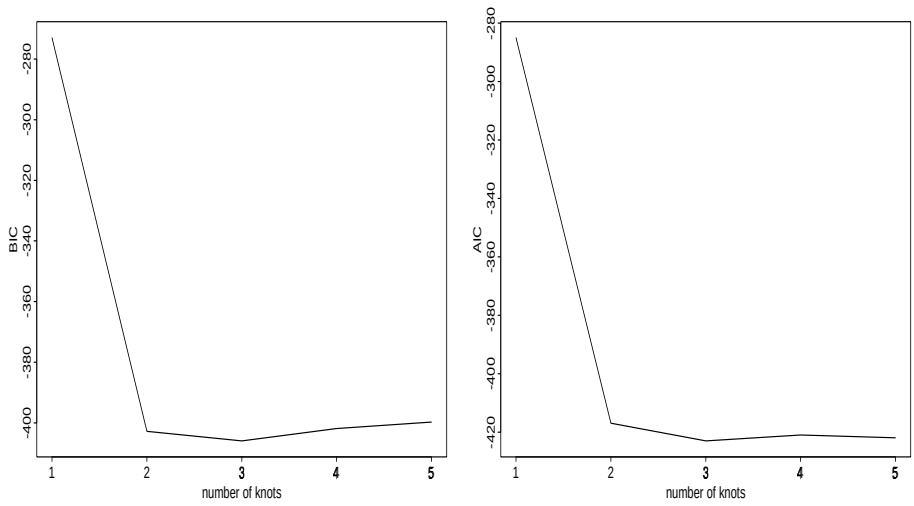

Fig. 6. BIC and AIC values computed for different numbers of knots with the example presented in section 3.3. Note that 3 knots minimize the criteria.

case of overlapping knots. According to Feder (1967), we define the number of parameters of the spline regression as $2K+d$ (K knots and $K+d$ coefficients), each knot worth 2: one for its position and one for the associated coefficient. Thus, we evaluate each model with

$$F(\xi_{(k)}) + \alpha * (2K + d),$$

where the penalization coefficient is $\alpha = 2$ for the AIC, $\alpha = \log n$ for the BIC.

With the algorithm presented herein, the user has to determine only the maximum number of knots. We illustrate this procedure with a relatively simple example based on the smooth signal function $f(x) = cos(x)$. Simulated data are created as follows: the function is evaluated at 90 points along $[0, 10]$, 30 points are uniformly generated from a $\mathcal{U}[0, 3]$, 40 from a $\mathcal{U}[5, 7]$ and 20 from a $\mathcal{U}[9, 10]$. Zero-mean normal noise is added with $\sigma = 0.1$ and data points are presented in Figure 1.

The preceding algorithms have been applied with splines of degree 3 and with a number of knots ranging form 1 to 5. For each model, values of the BIC and AIC criteria are presented in Figure 6. The three knots spline model has been selected because it minimizes both the BIC and AIC. The corresponding estimation is shown in Figure 1. Note again that the optimal knots are located on empty regions.

To determine the degree of the spline, the same algorithm can be used by adding a loop on d. Define a largest degree (in our experiment we choose 3) and the algorithm provides the degree and the number of knots which min-

imizes the criterion. By using this procedure with the simulation presented above, the degree 3 is adopted.

4 Multivariate Bounded Optimal Knots

With p covariates $(X_1, \cdots, X_p)$, and a random response variable Y, all measured on n observations gathered in the matrix X and the vector Y, the problem becomes the estimation of the conditional expectation or regression function

$$f(x_1, \cdots, x_p) = E(Y/X_1 = x_1, \cdots, X_p = x_p).$$

One could use a multivariate kernel estimator for this purpose. However, difficulties arise when p is large due to the scarcity of data points. The following example illustrates this "curse of dimensionality" (Scott, 1992, chapter 7) that uses the multivariate "running mean" estimator, which corresponds to the uniform kernel. If we consider as reasonable conditions for fitting the data that 5 observations belong to each multivariate rectangular window whose area is 10% of the global rectangular data area, then 5×10^p observations are needed for p predictors.

A natural extension of the univariate procedure is to solve the following multivariate optimization problem which is analogous to (4)

$$\hat{\xi}(l, u) = \arg \min_{\substack{l^1 \le \xi^1 \le u^1 \\ l^2 \le \xi^2 \le u^2 \\ \vdots \\ l^p \le \xi^p \le u^p}} F(\xi), \tag{6}$$

where the super vector of knots $\xi = (\xi^1, \cdots, \xi^p)$ is constrained to lie within the super vectors of lower and upper bounds. Once $F(\xi)$ is specified, equivalent algorithms as previously discussed can also be used. We define the $MBOK$ (for Multivariate Bounded Optimal Knots) procedure which is the analogue of $EBOK$ for the multivariate case. For each of the p predictors, we need the number of knots k_i and the polynomial degree d_i. Then, the $MBOK$ procedure solves (6). Moreover, to avoid a choice for the k_i's, we propose a generalization of the algorithm proposed in the univariate context. Here, K denotes the total number of knots used for the predictors. For a fixed K, the algorithm presented below determines the optimal k_i's and each knot location.

For $k_1 = 0$ to K do

for $k_2 = 0$ to $K - k_1$ do

...

for $k_p = 0$ to $K - k_1 - k_2 - ... - k_{p-1}$ do

$F(\xi_{(k_1,...,k_p)}) \leftarrow MBOK(X, Y, d_1, ..., d_p, k_1, ..., k_p, N, \varepsilon, \rho)$

end for

...

end for

end for

$F(\hat{\xi}) \leftarrow \arg\min_{k_1,...,k_p}\{AIC(F(\xi_{(k_1,...,k_p)}))\}$.

In the following subsection, we will define F for a couple of multivariate contexts, but the character of the optimization is the same. The function $MBOK$ evaluates the objective function with constrained knots on each variable. To determine the optimal K and also the degree for each predictor, the algorithm can be completed with loops on each parameter which can vary between a lower and an upper value. MARS and POLYMARS allow the possibility that a variable has no effect at all. BOK could not assume that each variable is in the model, at least linearly, requiring adding a choice of d_i in the main loop. In the applications, we use $d_i = 0, 1, 2, 3$ for the degrees, and we start with $K = 0$, then we add one knot until the criterion is increased. In the next section, we use our algorithm on the simple additive model.

Note that with a great number of variables or knots, the iterative algorithm implies a large computational time. This drawback comes from the fact that there are $\binom{K-1}{p+K-1}$ possibilities to divide K knots into p variables. For $p > 5$ and $K > 10$, the execution time could be a serious limitation. However, the computational time does not increase much with the number of data n. With our continuous approach to select knot locations, n only increases the dimension of the matrix B, whose size does not have much effect in the minimization procedure which concerns the vector ξ. In fact, the total number of knots K and the number of variables p are the important parameters for the minimization computational time. For p, *a priori* information on predictors allows to reduce the computation time. For example, if the user knows that the predictor X_i has a linear effect, he can impose $d_i = 1$ and $k_i = 0$ to reduce the number of iterations.

4.1 Additive model

An alternative to the use of multivariate smoothers is based on estimation of an additive approximation

$$y = f(x_1, \cdots, x_p) = \alpha + \sum_{j=1}^{p} f_j(x_j),$$

with $E(f_j(X_j)) = 0, j = 1, \cdots, p$, to ensure identifiability. The intercept $\alpha = E(Y)$ is typically estimated by $\bar{y} = \frac{1}{n}\sum_i y_i$. For simplicity, henceforth we use $\alpha = 0$.

To construct an estimator $\hat{f}$ of f defined by

$$\hat{f}(x_1, \cdots, x_p) = \sum_{j=1}^{p} \hat{f}_j(x_j),$$

where $\sum_{i=1}^{n} \hat{f}_j(X_{ij}) = 0, j = 1, \cdots, p$, we use a multivariate extension of the regression by least-squares splines. Given $\xi = (\xi^1, \cdots, \xi^p)'$, a set of knots for each predictor, *i.e.*, given
$\{\{B_{cl}^{j}(., \xi^j)\}_{l=1,\cdots,K_j+d_j+1} | j = 1, \cdots, p\}$, $\hat{f}$ is defined by

$$\hat{f}(x_1, \cdots, x_p) = \sum_{j=1}^{p} \sum_{l=1}^{K_j+d_j+1} \hat{\beta}_l^j(\xi) B_{cl}^{j}(x_j, \xi^j),$$

where $\hat{\beta}(\xi) = (\hat{\beta}_1^1(\xi), \cdots, \hat{\beta}_{K_1+d_1+1}^1(\xi), \cdots, \hat{\beta}_1^p(\xi), \cdots, \hat{\beta}_{K_p+d_p+1}^p(\xi))'$ is a solution to the least-squares problem

$$\hat{\beta}(\xi) = \arg \min_{(\beta_l^j)_{l=1,K_j+d_j+1}^{j=1,p}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \sum_{l=1}^{K_j+d_j+1} \beta_l^j B_{cl}^{j}(X_{ij}, \xi^j))^2, \tag{7}$$

where $\frac{1}{n}\sum_{i=1}^{n} B_{cl}^{j}(X_{ij}, \xi^j) = 0$ for any j, l. As the B-spline basis contains the constant (the sum of all the splines for a single variable is 1), this zero sum condition is necessary to avoid identifiability problems.
The least-squares spline coefficients defined in (4) may be written as

$$\hat{\beta}(\xi) = \arg\min_{\beta} \|Y - B(\xi)\beta\|^2 = (B'(\xi)B(\xi))^{+}B'(\xi)Y,$$

where $B(\xi) = [B^1(\xi^1)|\cdots|B^p(\xi^p)]$ is the $n \times \sum_j (K_j + d_j + 1)$ column centered super coding matrix. As in the univariate case, when knot locations are free variables, Golub and Pereyra's (1973) result also holds, and we need only minimize the objective function $F(\xi) = \|Y - B(\xi)\hat{\beta}(\xi)\|^2$ with respect to ξ.

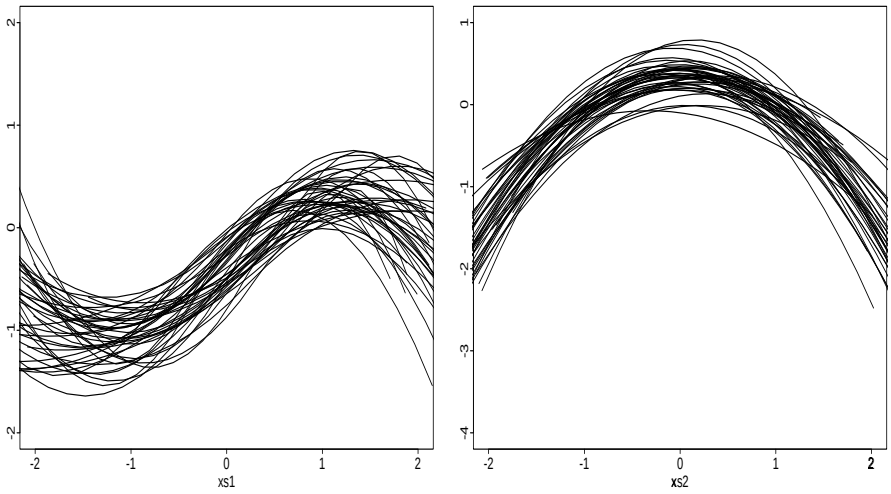

Fig. 7. Transformation estimations on the 50 simulations of the Hastie's data sets.

4.1.1 Hastie's data set

In the discussion of (Friedman and Silverman, 1989), Trevor Hastie generated 50 data sets of sample size 100 from the model:

$$y = 0.667 \sin(1.3x_1) - 0.465x_2^2 + \varepsilon,$$

where ε, x_1, x_2 are $\mathcal{N}(0,1)$ and x_1, x_2 have correlation 0.4. In the discussion of MARS, Leo Breiman (1991) re-analyzed this data set. On each data set, we use additive splines of degree 2 with a total number of knots of 0, 1, 2, 3 or 4. The model with only one knot on the first variable minimizes the AIC for all the simulations, and the resulting transformations are given in Figure 7. This model is the best one for all the 50 simulations. Note that the variability is larger than the one obtained with MARS, but lower than with TURBO. The estimation on one sample is presented in Figure 8.

4.1.2 Multicollinearity and scarcity of data

The simulated data consists of 50 samples of $n = 200$ observations and $p = 5$ predictors. The exploratory variables are strongly correlated and generated as follows: X_1 is uniform on $[-1, 1]$, $X_2 = 0.9X_1 + \varepsilon$, $X_3 = -1.1X_2 + \varepsilon$, $X_4 = 0.9X_3 + \varepsilon$ and $X_5 = -1.1X_4 + \varepsilon$ where ε are normal $(0, 0.1)$, The response is generated by $Y_i = f(X_i) + \rho_i$ for $i = 1, \cdots, 200$, with the ρ_i independently drawn from a $\mathcal{N}(0,1)$ distribution. The function f is taken to be

$$f(x_1, \cdots, x_5) = 2\sin(\pi x_1) - 6x_2^3 + 3x_3 - 2x_4 + x_5.$$

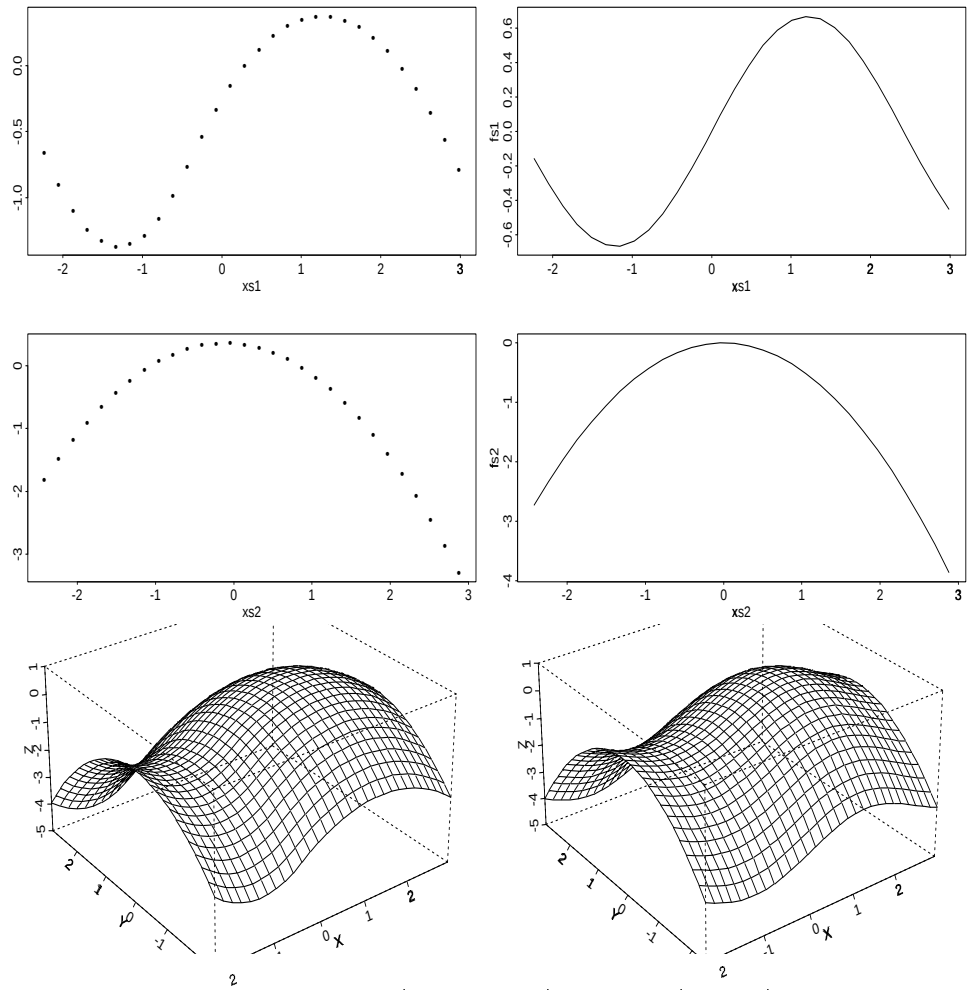

Fig. 8. Estimated transformation (dotted line), surface (on left) and true transformations (on right) on one sample of the Hastie's data.

The complete algorithm with selection of the number of knots and the spline degree for each predictor has been performed with the AIC criterion. The resulting transformations are given in Figure 9.

4.1.3 Simulations

In their article, Smith and Kohn (1997) proposed a data set on which they examine the properties of the principal regression spline methods. We applied our algorithm on the same example. The two predictors x_1 and x_2 are independent normal with mean 0.5 and variance 0.1, and $f(x_1, x_2) = 1/5 \exp(-8x_1^2) + 3/5 \exp(-8x_2^2)$, the additive model used by Gu *et al.* (1989). We generated $n = 300$ observations of x_1, x_2 and y with $\varepsilon \sim \mathcal{N}(0, (\text{range}(f)/4)^2)$. We carried out 100 replications of this simulation. The performance of our estimator was measured using an approximated integrated squared error (AISE) given by $\text{AISE}(\hat{f}) = 1/n \sum_{i=1}^{n} \{(f_i - \hat{f}_i)^2\}$. Here $\{f_i\}_{i=1}^n$ and $\{\hat{f}_i\}_{i=1}^n$ are the true and

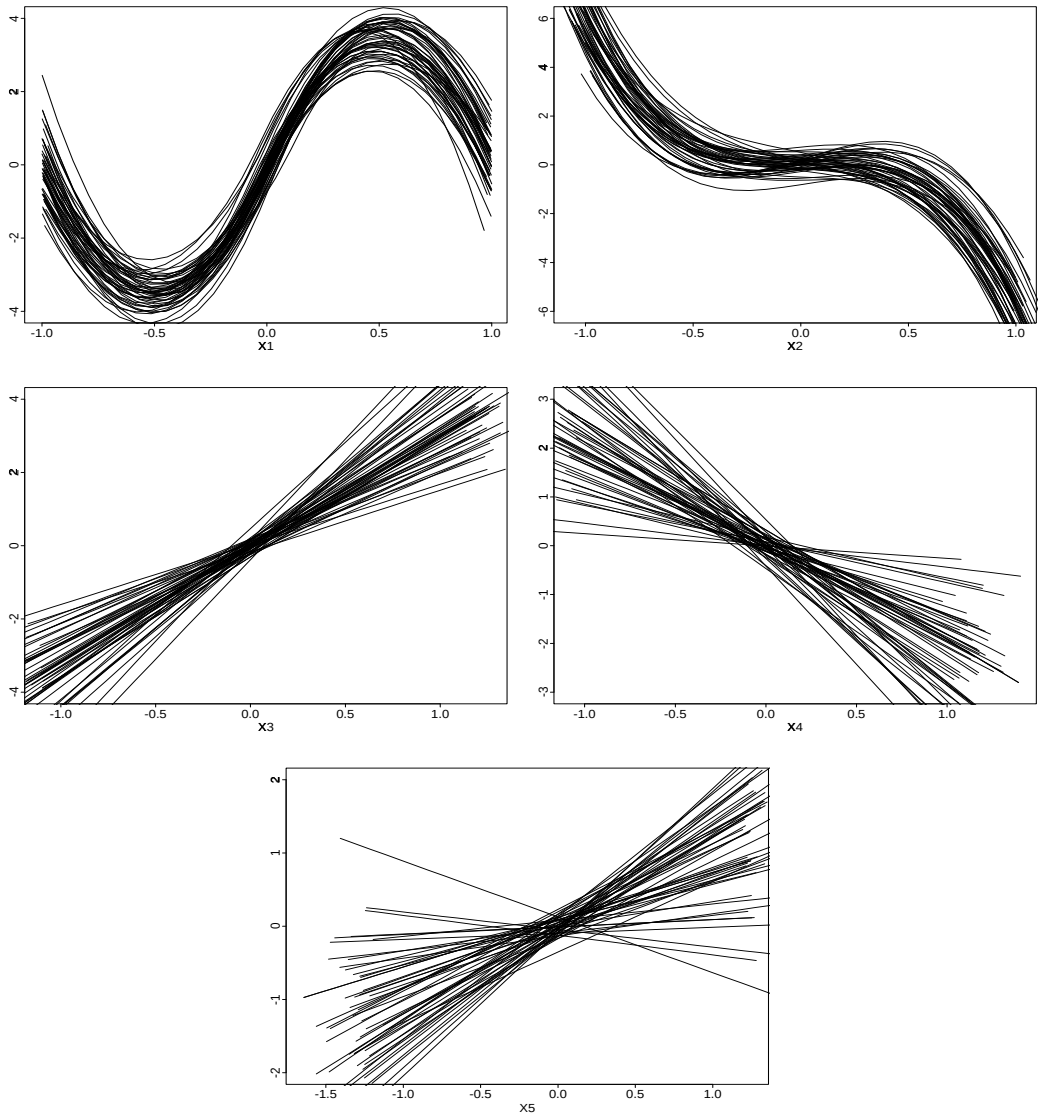

Fig. 9. Predictor estimations for the 50 samples of the data with multi collinearity.

estimated function values. Figure 10 provides boxplots of the results. For most of the samples, two knots for the first predictor are used and none for the second and a spline of degree two optimizes the AIC for both.
The Bayesian method of Smith and Kohn (1996, 1997) constructs a posterior distribution on knot sequences. MCMC is then employed to explore this distribution, generating a large number of different knot configurations and ultimately choosing the one that has the largest posterior mass. By cleverly picking their prior distributions, the posterior used by the authors is (essentially) BIC. Therefore, Smith and Kohn randomize knot locations to approximately minimize BIC, while the method presented in this paper randomizes knot barriers to approximately minimize the criterion.

Note that we also used the proposed method on several other simulations.

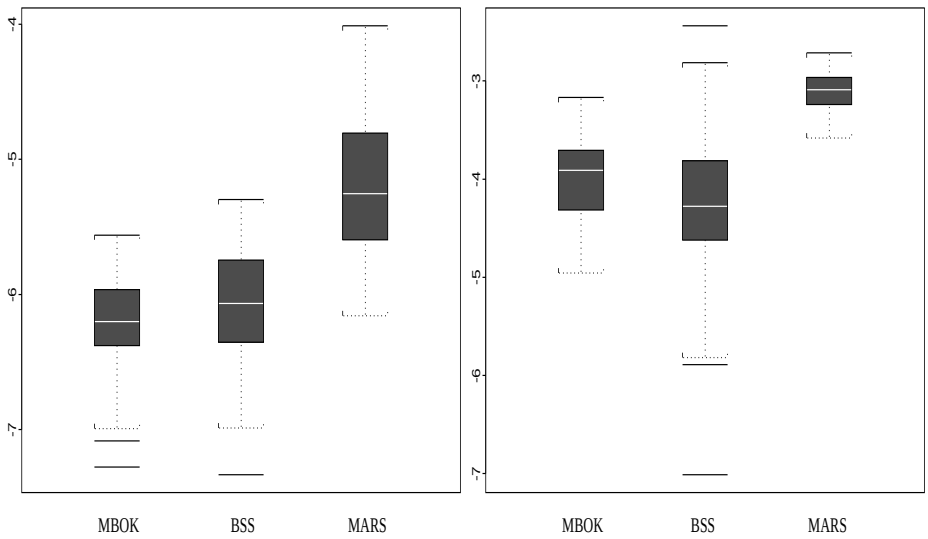

Fig. 10. Boxplots of log(AISE) for the simulated data with MBOK, BSS and MARS respectively. The additive example on left and the tensor product example on the right.

For $y = x_1^2 + x_2 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.2)$ and $n = 30$, the procedure determines that no knots are useful and provides a simple polynomial regression in x_1 and x_2 with degrees 2 and 1, respectively. For $y = 5\sin(\pi x_1) + x_2^2 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 1)$ and only 20 observations, the complete procedure yields splines of degree two for both predictors and one knot for the first. For these examples, a comparison with MARS or TURBO shows that BOK provides similar results.

4.2 Tensor product regression splines

A multivariate surface can be modeled using a tensor product of univariate functional bases. This section discusses how to estimate the surface f by modeling it as a linear combination of basis functions so that

$$f(x_1, ..., x_p) = \sum_i \beta_i B_i(x_1, ..., x_p), \tag{8}$$

where the choice for B_i is a tensor products of univariate B splines.

Smith and Kohn (1997) presented the bivariate context. In this case, the basis can be decomposed into main effects f_1 and f_2 in x_1 and x_2, along with an interaction part f_{12}, so that $f_1(x_1) = \sum_{i=1}^{k_1+d_1+1} \beta_i^1 B_i^1(x_1)$, $f_2(x_2) = \sum_{j=1}^{k_2+d_2+1} \beta_j^2 B_j^2(x_2)$ and $f_{12}(x_1, x_2) = \sum_{i=1}^{k_1+d_1+1} \sum_{j=1}^{k_2+d_2+1} \beta_{i,j}^{1,2} B_i^1(x_1) B_j^2(x_2)$.

Let x_1 and x_2 be independent uniforms on $[0,1]$, and $f(x_1,x_2) = x_1 \sin(4\pi x_2)$. We generated 300 observations of x_1, x_2 and y with $\varepsilon \sim \mathcal{N}(0, 1/4 \times \text{range}(f))$. We applied our algorithm on this sample with d_1 and d_2 in $\{0,1,2,3\}$ and a total number of knots equal to 1,2,3 or 4. For the two criteria, the minimum is obtained for $d_1 = 1, d_2 = 2, k_1 = 2$ and $k_2 = 0$.

To compare our results with those obtained with existing methods, we carried out 100 replications of this simulation. The performance was measured using the AISE. Figure 10 presents boxplots of the results obtained. The proposed model gives better results than MARS but slightly less accurate than those obtained with the Bayesian approach.
Note that the tensor product model takes more computational time than the additive model because each adjustment modifies more columns in the B matrix. To summarize, we can say that the most important parameter for the computational time is K, n does not have a very important effect.

5 Conclusion

The presented method applied in both univariate and multivariate contexts is based on the use of bounded optimal knots to avoid coalescent knots. It constructs a non deterministic algorithm that tends to guard the fitted regression spline against the problems of scarcity of observations and multicollinearity of predictors. The complete algorithm determines the number of knots and the degree of splines through a model selection procedure. In this paper, we use the AIC and the BIC, although another penalization coefficient α can also be used in the multivariate context. The choice of the criterion is not debated in this paper.
The method BOK is particularly well adapted for additive structure models. Clearly, a truly additive regression function is rare, however, the additive model is a useful approximation. The tensor product regression is one way to estimate interaction terms. The method is very attractive for small problems with a small number of variables, parallel computing could possibly employed to tackle large-scale problems (Kontoghiorghes, 2000), (Hegland, 1999). One interesting feature of the method is that it can been adapted to different contexts : a simple additive model, a tensor product regression or an additive model with a multiplicative term to modelize interactions $F(\xi) = f_1(x_1) + f_2(x_2) + f_3(x_1 x_2)$. Moreover, the algorithms presented herein can also be used in other statistical methods such as Additive Splines Partial Least Squares, or Principal Component Analysis spline, where only the objective function has to be modified.

A The EBOK algorithm

Algorithm 2: *EBOK algorithm*

Inputs: $X, Y, d, K, N, \varepsilon = 10^{-3}, \rho = (b-a)/10$

$l^{(1)} \leftarrow$ equidistant bounds
Compute less than N ε-different candidates
    $i \leftarrow 2$
    $count \leftarrow 0$
    while $(i \leq N \ \ \& \ \ count < 100)$ **do**
        $l \sim (\mathcal{U}[a,b])^{K-1}$
        while $l \in \bigcup_{j=1,i-1} \mathcal{B}_\infty(l^{(j)}, \rho)$ **do**
            $count \leftarrow count + 1$
            $l \sim (\mathcal{U}[a,b])^{K-1}$
        end while
        $l^{(i)} \leftarrow l$
        $i \leftarrow i + 1$
    end while
    $N \leftarrow i$

Minimizations
    for $i = 1$ to N
        compute $u^{(i)}$
        bounded minimization: $\xi^* \leftarrow \arg \min_{l^{(i)} \leq \xi \leq u^{(i)}} F(\xi)$
        while ($\exists k$ such that $\xi_k^* = l_k^{(i)}$ or $\xi_k^* = u_k^{(i)}$)
            for all j such that $\xi_j^* = l_j^{(i)}$ **do**
                $l_j^{(i)} \leftarrow \frac{1}{2}(\xi_{j-1}^* + l_j^{(i)})$
            for all j such that $\xi_j^* = u_j^{(i)}$ **do**
                $u_j^{(i)} \leftarrow \frac{1}{2}(u_j^{(i)} + \xi_{j+1}^*)$
            bounded minimization: $\xi^* \leftarrow \arg \min_{l^{(i)} \leq \xi \leq u^{(i)}} F(\xi)$
        end while
        $\xi^{(i)} \leftarrow \xi^*$
        compute $F(\xi^{(i)})$
    end for

$F(\hat{\xi}) \leftarrow \min\{F(\xi^{(1)}), \cdots, F(\xi^{(N)})\}$

References

Akaike, H. (1974), *Information theory and an extension of the maximum likelihood principle.* In: 2nd International Symposium on Information Theory. Eds. B. N. Petrov and F. Csaki. Akademiai Kiado Budapest, 267-281.

A. T. & T. Bell Laboratories (1984), PORT Mathematical Subroutine Library Manual.

Breiman, L. (1993), Fitting Additive Models to Regression Data, *Computational Statistics and Data Analysis*, 15, 13-46.

Breiman, L. and Friedman, J. H. (1985), Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion), *Journal of the American Statistical Association*, 80, 580-619.

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

de Boor, C. and Rice, J. R. (1968), *Least-squares cubic spline approximation. II: Variable knots*, CSD Technical Report 21, Purdue University, IN.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. (1998), Automatic bayesian curve fitting, *J. R. Statist. Soc. B*, 60, 333-350.

Dierckx, P. (1993), *Curve and Surface Fitting with Splines*, Oxford University Press, Oxford.

Dongarra, J. J. and Grosse, E. (1987), Distribution of mathematical software via electronic mail, Communications of the ACM 30, pp. 403-407.

Feder, P. I. (1967), *On the likelihood ratio statistic with applications to broken line regression.* Ph.D. dissertation, Dep. Statist., Satanford Univ.

Friedman, J. H. (1991), Multivariate adaptive regression splines (with discussion), *Annals of Statistics*, 19, 1-141.

Friedman, J. H. and Silverman, B. W. (1989), Flexible Parsimonious Smoothing and Additive Modeling (with discussion), *Technometrics*, 31, 3-39.

Gallant, A. R. and Fuller, W. A. (1973), Fitting Segmented Polynomial Regression Models whose Join Points have to be estimated, *Journal of the American Statistical Association*, vol. 68, 341, 144-147.

Gay, D. M. (1983), Algorithm 611. Subroutines for Unconstrained Minimization using a Model/ Trust-Region Approach. *ACM Transactions on Mathematical Software*, 9, 503-524.

Gay, D. M. (1984), *A trust region approach to linearly constrained optimization in Numerical Analysis*, Proceedings, Dundee 1983, F. A Lootsma (ed.), Springer, Berlin, 171-189.

Golub, G. H. and Pereyra, V. (1973), The differentiation of pseudo-inverses and nonlinear least-squares problems whose variables separate, *SIAM Journal of Numerical Analysis*, 10, 33-45.

Guertin, M. C. (1992), *Sur les splines de régression noeuds variables*, mmoire de Maîtrise es Sciences, Université de Montréal.

Hegland M., McIntosh I. and Berwin A. Turlach (1999), A parallel solver for generalised additive models, *Computational Statistics and Data Analysis*, 31, 377-396.

Jupp, D. L. B. (1975), The Lethargy Theorem, a Property of Approximation

by γ-Polynomials, *Journal of Approximation Theory*, 14, 204-217.

Jupp, D. L. B. (1978), Approximation to Data by Splines with Free Knots, *SIAM Journal of Numerical Analysis*, 15, 328-343.

Kontoghiorghes, E. J. (2000), *Parallel Algorithms for Linear Models: Numerical Methods and Estimation Problems*, Kluwer Academic Publishers, Boston, MA.

Lindstrom, M. J. (1999), Penalized estimation of free-knot splines, *J. Comp. Graph. Statis.*, 8, 333-352.

MathSoft (1996), *S-Plus version 3.4 for Unix Supplement*, Data Analysis Products Division, MathSoft, Seattle.

Owen, A. (1991), Discussion about Multivariate Adaptative Regression Splines. *The Annals of Statistics*, 19:102-112.

Schoenberg, I. J. and Whitney A. (1953), On Polya frequency functions. III: The positivity of translation determinants with an application to the interpolation problem by spline curves, *Trans. Amer. Math. Soc.*, 74, 246-259.

Schumaker, L. L. (1981), *Spline Functions: Basic Theory*, Wiley Interscience.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.

Scott, D. W. (1992), *Multivariate Density Estimation*, Wiley Interscience.

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

Smith, M. and Kohn, R. (1997), A bayesian approach to nonparametric bivariate regression, *Journal of the American Statistical Association*, 92, 1522-1535.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997), Polynomial splines and their tensor products in extended linear modeling (with discussion), *Annals of Statistics*, 25, 1371-1470.