ON THE SMOOTHING SPLINE REGRESSION MODELS

Nicoleta Breaz and Mihaela Aldea

Abstract. In this paper, we discuss about a modern tool used in the regression models framework, namely the smoothing spline function. First, we present the smoothing problem versus the fitting one and show when a smoothing is appropriate for data. Then we present the smoothing spline regression model as a penalized least squares regression model. An application of the smoothing spline regression to the real data is also presented. Finally, we discuss about some extensions possibilities of this model as Lg-smoothing spline model or smoothing spline model in case of data from multiple sources.

2000 Mathematics Subject Classification: 65D07, 62G08, 93E14.

Keywords and phrases: spline, regression, smoothing.

1.Introduction

Used in an incipient form as early as in antiquity, at the calculus of areas and volumes and introduced under this titulature, (by I.J.Schoenberg, in 1946, in the paper [8]), as piecewise polynomials, joined at the breaks together with several of their derivatives, the spline functions still constitute a field with great diponibility for new. From the elementary definition, up to what can be meant today by a spline function, the way of scientifical researches has beared a lot of ramifications, as a result of multiple posibilities for generalizations and extensions provided by this notion. Thus, the spline function can be defined as a piecewise function (not necessarily polynomial) or a solution to a variational problem. There are also spline functions that satisfy both features, one related to piecewise nature and the other linked with a variational problem.

A large number of papers (over 700) related to the subject, give evidence about the interest of the mathematicians for the spline functions. We remind here [7], an work having bibliographical feature, useful in the synthesis of various subjects discussed in this domain, the result of an extensive research in spline functions, by the romanian mathematician, Gh. Micula. Beside its applicability as an approximation tool in numerical analysis problems, the spline function is also used in statistical framework. From this reasearch area, we referee here, the papers, [3]-[5], [9]-[11], with emphasis on [10], an excelent monography in this domain, written by the principal leader of the spline functions applied in statistics school, Grace Wahba.

The goal of the present paper is to expose in the data analysis framework, the variational theory together with some generalizations of spline functions. We begin with some elementary definitions of a spline and also of a reproducing kernel Hilbert space, then in section 2, we present the smoothing problem versus the fitting one showing when a smoothing is appropriate for data and discuss how a spline function can be a statistical tool in the regression analysis. In the section 3, we present the penalized least squares smoothing spline model as a solution to the smoothing problems and show how this spline model work, on real data. Finally, in section 4, we prsent some possibilities of extensions of the smoothing spline model, like Lg-smoothing spline model defined on a reproducing kernel Hilbert space or smoothing spline model in case of data from multiple sources.

Definition 1.1. *Let be the following partition of the real line:* $\Delta : -\infty \leq a < t_1 < t_2 < ... < t_N < b \leq \infty$*. The function* $s : [a, b] \to \mathbb{R}$*, is called spline function of* m *degree (order* $m + 1$*), with the breaks (knots of the function),* $t_1 < t_2 < ... < t_N$*, if the following conditions are satisfied:*
i)$s \in P^m, t \in [t_i, t_{i+1}], i = \overline{0, N}, t_0 = a, t_{N+1} = b,$
ii)$s \in C^{m+1}, t \in [a, b].$
where, P^m *is the class of polinomyals of degree* m *or less and* C^{m-1} *is the class of functions with* $m - 1$ *continuous derivatives.*

The following formula uniquely gives the reprezentation of an element from $\delta_m(\Delta)$(the space of spline functions of m degree with the breaks given by the partition Δ), by means of the truncated power functions basis:

$$s(t) = p(t) + \sum_{i=1}^{N} c_i (t - t_i)_+^m, t \in [a, b], p \in P^m, c_i \in \mathbb{R} \tag{1.1}$$

where $t_+^m = \begin{cases} 0, t \leq 0 \\ t^m, t > 0 \end{cases}$

Definition 1.2. *Let Δ be the partition from definition 1.1 and the space of odd degree $(2m-1)$ spline functions, $\delta_{2m-1}(\Delta)$, $m \geq 1$. It is called the natural spline function of $2m-1$ degree (order $2m$), an element s, from the space $\delta_{2m-1}(\Delta)$, which satisfies the condition: $s \in P^{m-1}, t \in [a, t_1] \cup [t_N, b]$.*

We denote by $S_{2m-1}(\Delta)$, the space of natural spline functions of $2m-1$ degree, related to the partition Δ.

The following variational property of the natural spline functions is a previous step for the introduction of the smoothing spline function, used in statistics:

Theorem 1.3. *Let $a \leq x_1 < x_2 < ... < x_n \leq b$, be a partition of $[a, b]$, $y_i, i = \overline{1, n}$, $n \geq m$, real numbers and the set*

$$J(y) = \left\{ f \in H^{m,2}[a, b] : f(x_i) = y_i, \forall i = \overline{1, n} \right\}.$$

Then there exists a unique $s \in J(y)$, such that,

$$\int_a^b \left[s^{(m)}(x) \right]^2 dx = \min \left\{ \int_a^b \left[f^{(m)}(x) \right]^2 dx, f \in J(y) \right\}.$$

Moreover, the following statements hold:
i)$s \in C^{2m-2}([a, b])$,
ii)$s \left|_{[x_i, x_{i+1}]} \right. \in P^{2m-1}, \forall i = \overline{1, n-1}$,
iii)$s \left|_{[a, x_1]} \right. \in P^{m-1}$ and $s \left|_{[x_n, b]} \right. \in P^{m-1}$.

We denoted by $H^{m,2}[a, b]$ the following functions space:

$$\begin{aligned} H^{m,2}[a, b] = \{ f : [a, b] \to \mathbb{R} | \ & f, f', ..., f^{(m-1)}, \\ & \text{absolutely continuous}, f^{(m)} \in L_2[a, b] \}. \end{aligned} \tag{1.2}$$

An important particular spline function, from computational point of view, is B-spline function.

Definition 1.4. *Let $\Delta : a < t_1 < t_2 < ... < t_N < b$, be a partition of the interval $[a, b]$ and*

$$\Omega : t_{-m} = t_{-m+1} = ... = t_{-1} = t_0 = a < t_1 < t_2 < ... < t_N < b = t_{N+1} = ... =$$
$$= t_{N+m+1},$$

an extension of it, in which, $t_1, ..., t_N$ are called interior knots and a, b are called end knots, each one of two having $m+1$ multiplicity (without continuity conditions). It is called B-spline function of order $m+1$ (degree m), the function $M_{i,m}$, defined by means of divided difference (see [7]),

$$M_{i,m}(t) = \left[t_i, ..., t_{i+m+1}; (x-t)_+^m\right].$$

Also, it is called the normalized B-spline function, the function given by

$$N_{i,m}(t) = (t_{i+m+1} - t_i) \cdot M_{i,m}(t).$$

These functions form a basis for the liniar space $\delta_m(\Delta)$. The computational advantage of the basis formed with B-spline functions is due to their property that having compact support that is the function is zero outside the knots.

Begining with B-spline, a similar basis can be constructed for the space of natural spline function, $S_{2m-1}(\Delta)$, too.

Notations 1.5. We consider the partition $\Delta : a < t_1 < t_2 < ... < t_N < b, N \geq 2m+1, m \geq 1$ and we introduce the following notations

$$M_{i,2m-1}(t) = \left[t_i, t_{i+1}, ..., t_{i+2m}, (x-t)_+^{2m-1}\right],$$

$$M_{i,2m-1}^j(t) = \left[t_i, t_{i+1}, ..., t_{i+j}, (x-t)_+^{2m-1}\right], 1 \leq j \leq 2m,$$

$$\widetilde{M_{i,2m-1}^j}(t) = \left[t_i, t_{i+1}, ..., t_{i+j}, (t-x)_+^{2m-1}\right], 1 \leq j \leq 2m.$$

Theorem 1.6. *If $N \geq 2m+1$, the following N functions,*

$$M_{1,2m-1}^i, i = m, m+1, ..., 2m-1,$$

$$N_{i,2m-1}, i = \overline{1, N-2m} \text{ (normalized function of } M_{i,2m-1}),$$

$$\widetilde{M}^{i}_{N-i,2m-1}, i = m, m+1, ..., 2m-1,$$

form a basis for the space of natural, $2m-1$ *degree polynomial spline function,* $S_{2m-1}(\Delta)$*.*

Remark 1.7. From these N functions, only $N_{i,2m-1}, i = \overline{1, N-2m}$ are B-spline (having a compact support).

In the last section we will present an application of spline function in statistics that need the definition of a reproducing kernel Hilbert space.

Definition 1.8. *It is called the reproducing kernel Hilbert space (r.k.h.s.), the Hilbert space* $H_{\mathcal{R}}$*, of the functions* f*,* $f : \mathcal{I} \to \mathbb{R}$*, which has the property that for each* $x \in \mathcal{I}$*, the evaluation functional* L_x*, which associates* f *with* $f(x)$*,* $L_x f \to f(x)$*, is a bounded linear functional. The corresponding reproducing kernel is a positive definite function,* $\mathcal{R} : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$*, given by* $\mathcal{R}(x,t) = (\mathcal{R}_x, \mathcal{R}_t), \forall x, t \in \mathcal{I}$*, where* $\mathcal{R}_x, \mathcal{R}_t$ *are the representers of the functionals* L_x*, respectively* $L_t (f(x) = L_x(f) = (R_x, f), f \in H)$*.*

Here, the boundeness of evaluation linear functional, L_x, has to be understood as:

$$\exists M = M_x, \text{such that} \left|L_x(f)\right| = \left|f(x)\right| \leq M \left\|f\right\|, f \in H,$$

with $\|\cdot\|$, the norm of the space $H_{\mathcal{R}}$. An example of r.k.h.s. is the space $H^{m,2}$ from (1.2).

Propozition 1.9. *i) It holds that* $H^{m,2} = H_0 \oplus H_1$*, where* H_0*, is the* m*-dimensional space of the polynomials with* $m-1$ *degree at most and* $H_1 = \left\{f \in H^{m,2} \middle| \, f^k(0) = 0, k = \overline{0, m-1}\right\}$*.*
ii) Together with the norm $\|f\|^2 = \sum\limits_{k=0}^{m-1} \left(f^{(k)}(0)\right)^2 + \int\limits_0^1 \left[f^{(m)}(u)\right]^2 du$*, the space* $H^{m,2}$ *is a reproducing kernel Hilbert space.*

Remark 1.10. The smoothness penalty functional, $J_m = \int\limits_0^1 \left(f^{(m)}(u)\right)^2 du$, which is in fact a seminorm on $H^{m,2}$, can be written as

$$J_m(f) = \|P_1 f\|^2_{H^{m,2}} \tag{1.3}$$

where P_1 is the orthogonal projector on H_1, in $H^{m,2}$. An important matter related to (1.3), is to remain valid in any reproducing kernel Hilbert space, fact that is useful in the approach of the spline smoothing problem, in such spaces.

2.Fitting and smoothing in regression framework

In the proccesing data setting we can fit or smooth the data, the approach depending on what kind of closeness we want between $f(x_i)$ and y_i, $i = \overline{1,n}$.

Definition 2.1. *i)It is called a fitting problem related to the data* (x_i, y_i), $i = \overline{1,n}$, *the problem which consists of the determination of a function* $f : I \to \mathbb{R}, I \subset \mathbb{R}$, $x_i \in I, \forall i = \overline{1,n}$, *whose values at the data sites* x_i, *"come close" to data* y_i, *as much as possible, whithout leading necessarily to equality,* $f(x_i) \cong y_i, \forall i = \overline{1,n}$. *We say that a fitting problem is the better closeness to data fitting problem with respect to the criteria* E, *if it consists of the determination of a function for which* $E(f)$ *is minimum/maximum. The criteria* E *is chosen such that its minimization/maximization coresponds to the closeness to data. It is called the least squares problem related to the data,* $x_i, y_i, i = \overline{1,n}$, *the problem which consists of the determination of a function (from a settled functions space),* $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$, $x_i \in I, i = \overline{1,n}$, *that is the solution to the minimization problem:*

$$E(F) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - F(x_i)]^2 = \min .$$

ii) We define the smoothing problem related to the data $(x_i, y_i), i = \overline{1,n}$, *as a problem which consists of the determination of a function* $f : I \to \mathbb{R}, I \subset \mathbb{R}$, $x_i \in I$, $i = \overline{1,n}$, *whose values at the data sites* x_i, *"come close" to data* y_i, *so much that the function remains smooth. In other words, we are searching for f as a solution to the following minimum/maximum criteria:*

$$E(f) + \lambda J(f) \to \min / \max$$

where, $E(f)$ *is a functional that reflects, by minimizing/maximizing, the closeness to data (fitting),* $J(f)$ *is related to the smoothing condition, the minimization/maximization of this functional leading to a function with some smoothness properties and the parameter* λ, *takeing values in the interval* $(0, \infty)$, *is called smoothing parameter.*

Corresponding to these data approaches, we can define fitting and smoothing spline functions.

Definition 2.2. *i)An element from the $\delta_m(\Delta)$ is called (least squares) fitting spline function if it is a solution to the (least squares) fitting problem presented in the definition 2.1.i.*
ii)We define the (general) smoothing spline function as a function from an appropriate smooth function space, that is a solution to the data smoothing spline problem, presented in the definition 2.1.ii.

In the regression framework, one of the chalenge is to find an estimator of the regression function f, from the model $Y = f(X) + \epsilon$, based on data information, $(x_i, y_i), i = \overline{1,n}$. If we don't know anything about the phenomenon behind the data, but the scatter plot of the data is quite simple, the first step is to try the classical model based on linear function or a more general model as exponential, polynomial, etc. The idea is to find a function that come close to data as much as possible, so we deal with the fitting problem and the least squares method can be the appropriate criteria for closeness. The polynomial model for example is known as a flexibil model which is a linearisable model. However, if the data are too scattered, then a much more flexibile model is required. In this case, the least squares fitting spline function, from definition 2.2, can be the appropriate estimator for the regression function.

Definition 2.3. *It is called the fitting spline regression model the model $Y = f(X) + \epsilon$, where f is the spline function (definition 1.1), of m degree, $m \geq 1$, with the breaks $t_1 < t_2 < ... < t_N$. If the model is based on the least squares criteria then we use the term of least squares spline regression and the estimator is defined as the function from definition 2.2.*

According with the form (1.1) of a spline function, this spline model can be reduced to a linear one, by the substitution $X^k = Z_k, k = \overline{1,m}$ and $(X - t_k)_+^m = U_k, k = \overline{1,N}$, thus obtaining the linear model with a constant term, $Y = \alpha_0 + \alpha_1 Z_1 + ... + \alpha_m Z_m + \beta_1 U_1 + ... + \beta_N U_N$. Since such a function is a piecewise polynomial function, we gain more flexibility than in the polynomial model, so the fitting problem will have a better solution. Based on the expression (1.1), to find in the regression setting, the least squares fitting spline estimator of m degree, with the breaks $t_1 < t_2 < ... < t_N$, is equivalent with to find the estimators for $m + N + 1$ coefficients. We can observe that the polynomial spline function is completely defined if the degree and the breaks are known. In case that we don't have any information about these, the function degree, the number of breaks and the location of them become aditional parameters that need to be estimated, in order to estimate the regression function, completely. Thus, if we consider the parametric regression, one must deal with a restrictive approximation class which provides estimators with known parametric form except for the coefficients involved. Thus, the estimator comes closer to data as much as the assumed parametric forme (in this case, a spline one defined by its piecewise nature) allows. But if the data are noisy, so the data are suspected to contain errors, the smoothing framework is more appropriate to find the regression function estimator and the nonparametric regression is not so restrictive related to the class where we have to search for the estimator. In the nonparametric regression, the class of function in which one is looking for the estimator can be extended to more general functions spaces, that do not assume a certain parametric form but just some smoothness properties (continuity, derivability, integrability) of the function. Thus, one will use an estimator which even if it has great flexibility will however smooth too perturbated data, assuming some smoothness.

3.The penalized least squares smoothing spline model

The smoothing spline regression model, presented in this section, is based on an estimator, both flexible and smooth, appropriate for the cases when a parametric regression model is not sufficently motivated and the data are noisy. Also, we will meet here, the other feature of a spline function, linked with a variational problem.

Definition 3.1. *Let $H^{m,2}[a,b]$, be the functions space, defined in formula (1.2). We call the penalized least squares smoothing spline estimator of the regression function from the model $y_i = f(x_i) + \epsilon_i, i = \overline{1,n}$, with $\epsilon' = (\epsilon_1, ..., \epsilon_n) \sim N(0, \sigma^2 I)$, an element from $H^{m,2}[a,b]$, that minimizes the expression*

$$n^{-1} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_a^b \left(f^{(m)}(x)\right)^2 dx, \lambda \geq 0. \tag{3.1}$$

The related regression model will be called the smoothing spline regression model and λ is called smoothing parameter.

Remark 3.2. We can observe that this estimator is a particular case of smoothing spline function from the definition 2.2, obtained for (penalized) least squares criteria. In fact, the estimator is called the penalized least squares estimator, due to the first part (least squares criteria) and to the second part (penality functional) from the expression (3.1). Also, it can be proved (see theorem 1.3) that the unique solution for the variational problem (3.1) is the natural polynomial spline function of $2m-1$ degree (see definition 1.2), with the knots $x_i, i = \overline{1,n}$, which interpolates the fitted value of the regression function. It need to emphazise that this spline is defined not by piecewise feature but as solution to a variational problem. Anyway, in this case, since the form of the estimator derived from definition 1.2, this spline has both piecewise and variational feature.

If we look at the (3.1), the flexibility of this spline estimator is provided by the smoothing parameter λ that controlls the tradeoff between the fidelity to data of the estimated curve and the smoothness of this curve. One can observe that, for $\lambda = 0$, the second term from (3.1) will be 0, consequently, the minimization of whole expression is reduced to the minimization "in force", of the sum of squares. This case leads to an estimator that interpolates the data. On the other side, the case $\lambda \to \infty$ makes the second term from (3.1) to grow, therefore, in compensation, the accent in the minimization must be lay on the penalty functional. This case gives as estimator, the $m-1$ degree least squares fitting polynom, obtained from data (that is, the most possible smooth curve). Between these two extremes, a large value for λ indicates the smoothness of the solution in spite of the closeness to the data, while a small value leads to a curve very close to data but which loses smoothness.

Estimating the smoothing parameter

In the related literature, several automatically data driven selection methods for λ are developed. One of these methods is based on the cross validation technique-CV (see [3]), that is constructed on the leaving out one principle (omission of one data from sample and use as a measure of regression quality of this data).

Definition 3.3. *The cross validation function related to the spline smoothing problem, that by minimizing in respect with λ, gives an estimator of the smoothing parameter, is*

$$CV(\lambda) = n^{-1} \sum_{k=1}^{n} \left(y_k - f_\lambda^{(-k)}(x_k) \right)^2 \tag{3.2}$$

where $f_\lambda^{(-k)}$ is the unique solution of the spline smoothing problem, stated for the data sample from which it was leaving out the k-th data.

An application of the smoothing spline regression to real data

In order to smooth some data by spline functions, we have implemented in Matlab environment, the following algorithm based on cross-validation method for selecting the smoothing parameter:

Algorithm 1

Step 1

Read the data, $(x_i, y_i), i = \overline{1, n}$.

Step 2

Order the data increasingly with respect to $x_i, i = \overline{1, n}$.

Step 3

If the data are not strictly increasing, for each group, $(x_i, y_i), i = \overline{k_1, k_2}$, for which $x_i = x_j, \forall i, j \in [k_1, k_2]$ and $y_i \neq y_j, \forall i, j \in [k_1, k_2], i \neq j$, weight the data as follows: instead of $k_2 - k_1 + 1$ data, consider just a single data, (x_{k_1}, y'_{k_1}), with the weight $w_{k_1} = k_2 - k_1 + 1$, where,

$$y'_{k_1} \equiv \frac{\sum\limits_{i=k_1}^{k_2} y_i}{k_2 - k_1 + 1} \overset{not}{=} y_{k_1}$$

Step 4

Read the new data (strictly increasing with respect to x), $(x_i, y_i), i = \overline{1, n'}$.

Step 5

For various values of λ, determine $f_\lambda^{(-i)}$, the smoothing spline function related to data less the i-th data.

Step 6

For the same values of λ, calculate $CV(\lambda) = \frac{1}{n} \sum\limits_{i=1}^{n} \left(y_i - f_\lambda^{(-i)}(x_i) \right)^2$.

Step 7
Determine λ_{CV}, for which $CV(\lambda_{CV}) = \underset{\lambda}{min} CV(\lambda)$.
Stop.

In what follows, we consider some observed data, taken from [6], and we smooth these data by CV based smoothing spline function. The data, $(x_i, y_i), i = \overline{1, 15}$, used in this application, represent the observed values for gas productivity and feedstock flow, in a cracking process, over 15 days. By applying the algorithm 1, based on the CV method, we obtain a value for the smoothing patrameter, equal to $\lambda_{CV} = 0.0102$, that indicates a good closeness to data . The following figure is the plot of the data together with the spline estimator.

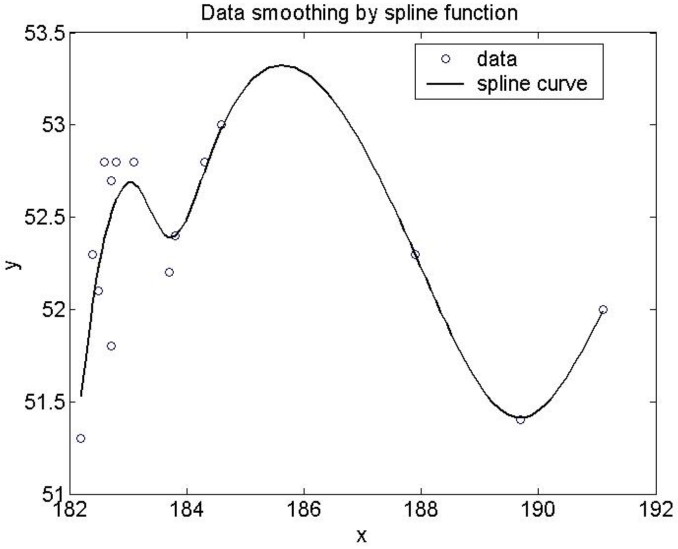

Figure 4.1

We can observe that spline estimator comes close to the data as much as its smoothness allows (cross-validation method for selection of λ, realises the tradeoff between smoothness and closeness to data), in fact ignoreing those points that seem to be outliers. But if we have some knowledge about the phenomenon behind the data regarding that the data are exactly measured, then we can choose directly a smaller value for λ, in order to come closer with spline curve to the data. By contrary, if we suppose that the data are very noisy, we can choose a greater value, to obtain more smoothness for the estimator.

4. Some extensions of smoothing spline model

Certainly, at the same time with the notion of spline functions and although with the large variety of related practical problems, the notion of smoothing spline estimator bears a lot of extensions among which we remind in this paper the case of the models with bounded linear observational functionals and with the estimator searched in reproducing kernel Hilbert spaces (see the papers: [10], [11]). Also, we consider the case of models with data proceeded from multiple sources ([1], [2], [5]).

Spline smoothing in reproducing kernel Hilbert spaces. Smoothing Lg-spline estimators

This problem starting with the assumption of searching for smoothing spline estimator in reproducing kernel Hilbert spaces (definition (1.8)). Then in general, the estimator doesn't keep the piecewise polynomial structure anymore, but it maintains the titulature of spline, being a solution to a certain variational problem, another facet of a spline function.

We consider the observational model, based on the data $(x_i, y_i), i = \overline{1, n}$,

$$y_i = L_i f + \epsilon_i, i = \overline{1, n} \tag{4.1}$$

with $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)' \sim N(0, \sigma^2 I$ and L_i, bounded linear functionals, on $H_{\mathcal{R}}$, the reproducing kernel Hilbert space of real-valued functions on $\mathcal{I}$, with reproducing kernel $\mathcal{R}$.

Definition 4.1. *We call the smoothing Lg-spline (nonparametric) regression model, the model (4.1), for which we search for the estimator of f , by searching for $f \in H_{\mathcal{R}}$, based on the criteria*

$$\min_{j} \left\{ n^{-1} \sum_{i=1}^{n} (y_i - L_i f)^2 + \lambda \left\| P_1 f \right\|_{H_{\mathcal{R}}}^2 \right\}, \lambda \geq 0, \tag{4.2}$$

where $P_1 f$ is the orthogonal projection of f onto H_1, in $H_{\mathcal{R}} = H_0 \oplus H_1$, $dim(H_0) = M \leq n$. Such an estimator is called the smoothing Lg-spline function, λ being the related smoothing parameter.

Remark 4.2. If $\mathcal{I} = [a, b]$, $H_{\mathcal{R}} = H^{m,2}[a, b]$, the observational functionals, $L_i, i = \overline{1, n}$, are the evaluation functionals defined by $L_i f = f(x_i)$ and $J_m(f) = \int\limits_a^b \left(f^{(m)}(x)\right)^2 dx$, then the Lg-spline smoothing problem is reduced to the (polynomial) spline smoothing problem presented in the previous section.

The next theorem gives, under certain conditions, the existence and uniqueness of the Lg-spline estimator and also gives a characterization of this estimator. The related demonstration can be found in [10].

Theorem 4.3. *Let $\phi_1, \phi_2, ..., \phi_M$, span the null space H_0 of P_1 and let T be the $n \times M$ matrix, defined by $T = \{L_i \phi_k\}_{i=\overline{1,n},\, k=\overline{1,M}}$. If T is of full column rank (equal to M), then the smoothing Lg-spline estimator, f_λ, is uniquely given by*

$$f_\lambda = \sum_{k=1}^{M} d_k \phi_k + \sum_{k=1}^{n} c_i \xi_i \tag{4.3}$$

where $\xi_i = P_1 \eta_i$, $d = (d_1, d_2, ..., d_M)' = (T'S^{-1}T)^{-1} T'S^{-1}y$, $c = (c_1, c_2, ..., c_n)' = S^{-1}\left(I - T(T'S^{-1}T)^{-1} T'S^{-1}\right) y$, $S = \Sigma + n\lambda I, \Sigma = \{(\xi_i, \xi_j)\}_{i,j=\overline{1,n}}$ and $\eta_i \in H_{\mathcal{R}}$, is the representer of the bounded linear functional, L_i.

Remark 4.4. i)The form (4.3) of the estimator reduces the Lg-spline smoothing problem to searching for $c \in \mathbb{R}^n$ and $d \in \mathbb{R}^M$, that minimizes the expression

$$\frac{1}{n} \|y - (\Sigma c + Td)\|^2 + \lambda c' \Sigma c. \tag{4.4}$$

ii)The cross validation function related to the Lg-spline smoothing problem, that by minimizing in respect with λ, gives an estimator of the smoothing parameter, is

$$CV(\lambda) = n^{-1} \sum_{k=1}^{n} \left(y_k - L_k f_\lambda^{(-k)}\right)^2 \tag{4.5}$$

where $f_\lambda^{(-k)}$ is the unique solution of the Lg-spline smoothing problem, stated for the data sample from which it was leaving out the k-th data.

Spline smoothing in the case of data from multiple sources with different weights

In this section, we have treated an extension of the spline data smoothing problem, namely, the model with data provided from more sources, for which the variances are different. The case of two sources with the same volume of data, related to functions defined on circle, was study in [5].

After the statement of the problem, we are concerned with the construction of a selection criteria for both the smoothing parameter and the relative weights corresponding to the sources. We begin with the regression model

$$Y = f(X) + \epsilon \tag{4.6}$$

for which, those $n = N_1 + N_2 + ... + N_l$ observations come from l sources, with different unknown acccuracies. We search for the estimator of f in the space $H^{m,2}[0,1]$. In this case, we deal with l observational models,

$$y_{1i} = f(x_{1i}) + \epsilon_{1i}, i = \overline{1, N_1},$$

$$y_{2i} = f(x_{2i}) + \epsilon_{2i}, i = \overline{1, N_2},$$

$$..................................$$

$$y_{li} = f(x_{li}) + \epsilon_{li}, i = \overline{1, N_l}, \tag{4.7}$$

with the errors, $\epsilon_j = (\epsilon_{j1}, \epsilon_{j2}, ..., \epsilon_{jN_j})' \sim N(0, \sigma_j^2 I)$. From those l sources, we have extracted $N_j, j = \overline{1, l}$, observations, materialized in the data $y_j = (y_{j1}, ..., y_{jN_j})'$ and $x_j = \left(x_{j1}, ..., x_{jN_j}\right)'$, each of such observational models being homoscedastic, that is, the data from the same sources have the same variance. If we knew the variances, $\sigma_j^2, j = \overline{1, l}$ and the parameter λ, then we could estimate the function f, by searching for $f \in H^{m,2}[0,1]$, that minimizes the expression

$$\frac{1}{n}\left[\frac{1}{\sigma_1^2}\sum_{i=1}^{N_1}(y_{1i}-f(x_{1i}))^2+\frac{1}{\sigma_2^2}\sum_{i=1}^{N_2}(y_{2i}-f(x_{2i}))^2+...+\frac{1}{\sigma_l^2}\sum_{i=1}^{N_l}(y_{li}-f(x_{li}))^2\right]$$
$$+\lambda J(f) \tag{4.8}$$

where $J(f)=\int\limits_0^1\left[f^{(m)}(x)\right]^2dx$.

It can be observed that the expression (4.8) is likewise the penalized least squares expression (3.1), with the mention that for each group of data we gave an importance directly proportional with the accuracy of the related source. This accuracy is obviously, invers proportional with the source variance (under the hypothesis of l homoscedastic models).

From the computational point of view, an appropriate solution to the minimization problem corresponding to (4.8), consists in searching for f in a subspace of $H^{m,2}[0,1]$, such that we can write

$$f\cong\sum_{k=1}^{N}c_kB_k \tag{4.9}$$

where B_k, $k=\overline{1,N}$ are the basis functions that span the disscused subspace. Such a subspace is $S_{2m-1}(\Delta)$, the space of $2m-1$ degree, natural spline functions, associated with a partition Δ, with N knots, and an appropriate basis is that which contains among those N basis functions, $N-2m$, B-spline functions, about which we know that they have a compact support (see theorem 1.6). In this case, we consider the partition Δ of the interval $[0,1]$, formed with all those n values $x_{ji}, i=\overline{1,N_j}, j=\overline{1,l}$, arranged in increasing order. Consequently, the problem of estimation for $f\in H^{m,2}[0,1]$ is reduced to the problem of estimation for $f\in S_{2m-1}(\Delta)$, that is the problem of estimation for $c\in\mathbb{R}^n$, $c=(c_1,c_2,...,c_n)'$.

If we replace (4.9) in the observational models (4.7) and use the matriceal writing, then we have l matriceal models,

$$y_1=B_1c+\epsilon_1,$$

$$........................ \tag{4.10}$$

$$y_l = B_l c + \epsilon_l,$$

with the errors, $\epsilon_j \sim N(0, \sigma_j^2 I), j = \overline{1,l}$ the matrices $B_j, j = \overline{1,l}$, being given by

$$B_1 = (B_{ik})_{\substack{1 \le i \le N_1 \\ 1 \le k \le n}}; B_{ik} = B_k(x_{1i}),$$

$$B_2 = (B_{ik})_{\substack{1 \le i \le N_2 \\ 1 \le k \le n}}; B_{ik} = B_k(x_{2i}),$$

....

$$B_l = (B_{ik})_{\substack{1 \le i \le N_l \\ 1 \le k \le n}}; B_{ik} = B_k(x_{li}).$$

With these notations, we can prove the next propozition (see [1]):

Propozition 4.5. *i) The solution to the variational problem based on the minimization of the expression (4.8), is given by the function* $f = \sum_{k=1}^{n} c_k B_k$, *where c satisfies the variational problem,*

$$\min_{c \in \mathbb{R}^n} \left\{ r_1 \|y_1 - B_1 c\|^2 + r_2 \|y_2 - B_2 c\|^2 + ... + r_l \|y_l - B_l c\|^2 + \alpha c' \Omega c \right\} \quad (4.11)$$

with α*, the resulted smoothing parameter,* θ*, a nuisance parameter and* $r_j, j = \overline{1,l}$*, the weighting parameters, given by the relations*

$$\alpha = \sigma_1 \sigma_2 ... \sigma_l \cdot \lambda \cdot n,$$

$$\theta = \sigma_1 \sigma_2 ... \sigma_l,$$

$$r_j = \frac{\sigma_1 \sigma_2 ... \sigma_{j-1} \sigma_{j+1} ... \sigma_l}{\sigma_j}, j = \overline{1,l} \quad (4.12)$$

and with the matrix $\Omega = \left\{ \int_0^1 B_k^{(m)}(x) B_s^{(m)}(x) dx \right\}_{k,s=\overline{1,n}}$.

ii) For fixed $\alpha > 0$ and $r = (r_1, r_2, ..., r_l)' \in \mathbb{R}^l$, the solution to the variational problem (4.11) is

$$c_{r,\alpha} = \left(r_1 B_1' B_1 + r_2 B_2' B_2 + ... + r_l B_l' B_l + \alpha\Omega\right)^{-1} \left(r_1 B_1' y_1 + ... + r_l B_l' y_l\right) \tag{4.13}$$

We can observe that the estimator depends on the smoothing parameter α, and the estimation of the ratios $r_j, j = \overline{1,l}$, defined by (4.12). Beyond the implication of r in the full estimation of c, respectively of f, the estimation of r is important in itself, since it gives information regarding the relative accuracy of various sources (instruments), from the data are taken. In order to constructe a CV function for estimation of these parameters, we make the following notations:

Notations 4.6. In order to unify the parts corresponding to those l sources, we will use the following notations:

$$y = \left(y_1', y_2', ..., y_l'\right)', B = \left(B_1', B_2', ..., B_l'\right)',$$

$$I(r) = \begin{pmatrix} \frac{1}{\sqrt{r_1}} I_{N_1} & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & \frac{1}{\sqrt{r_2}} I_{N_2} & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & \frac{1}{\sqrt{r_{l-1}}} I_{N_{l-1}} & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & \sqrt{r_1 r_2 ... r_{l-1}} I_{N_l} \end{pmatrix}$$

$$M = \left(r_1 B_1' B_1 + r_2 B_2' B_2 + ... + r_l B_l' B_l + \alpha\Omega\right),$$

$$y^r = I^{-1}(r) \cdot y, B^r = I^{-1}(r) \cdot B.$$

Also, we denote by $c_{r,\alpha}^{(-k)}$, the spline estimator obtained as a solution to the variational problem (4.11) applied on the sample with $n = N_1 + N_2 + ... + N_l$ data, from which the k-th data was left out. Also, by y_k, we denote the k-th element of the vector y, and by B^k and $B^{k,r}$, the k-th row of the matrix B, respectively B^r.

Now, we can define a CV-like function for the model (4.10), based on those l sources:

Definition 4.7. *The cross validation function related to the multiple sources spline smoothing problem, that by minimizing in respect with α and r, gives an estimator of the smoothing parameter and of the relative accuracies parameter, is*

$$CV(r,\alpha) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - B^k c_{r,\alpha}^{(-k)}\right)^2 \tag{4.14}$$

where $c_{r,\alpha}^{(-k)}$ is the unique solution of the spline smoothing problem (4.11), stated for the data sample from which it was leaving out the k-th data.

The minimization of this function, made in the purpose of the data driven estimation of r and α, represents an extension of the single source related criteria from (3.2) based on cross-validation method.

Obviously this problem can be written also, in the more general context of Lg-spline functions. Moreover, besides of these two extensions of a smoothing spline problem, presented here, a lot of extensions can be done in respect with problems that arise in practical context. We remind here only, the case of thin plate spline, the case of exponential data, partial spline models, the case of nonlinear observational functional and so on (see [10]).

References

[1].Breaz N., *On the smoothing parameter in case of data from multiple sources*, Acta Universitatis Apulensis, Mathematics-Informatics, Proc. of Int. Conf. on Theory and Appl. of Math. and Inf., Alba Iulia, no. 6, Part A, 75-84, 2003.

[2] Breaz N., *Modele de regresie bazate pe funcţii spline*, Presa Universitară Clujeană, 2007.

[3].Craven P., Wahba G., *Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math., 31, 377-403, 1979.

[4].Eubank R.L., *Nonparametric Regression and Spline Smoothing*-Second Edition, Marcel Dekker, Inc., New York, Basel, 1999.

[5].Gao F., *On Combining Data from Multiple Sources with Unknown Relative Weights*, Tech. Report 894, Department of Statistics, University of Wisconsin, Madison, WI, 1993.

[6].Marinoiu C., *Choosing a smoothing parameter for a curve fitting by minimizing the expected prediction error*, Acta Universitatis Apulensis, Mathematics-Informatics, no 5, 91-96, 2003.

[7].Micula Gh., Micula S., *Handbook of Splines*, Kluwer Academic Publishers, Dordrecht-Boston-London, 1999.

[8].Schoenberg I.J., *Contribution to the problem of approximation of equidistant data by analytic functions*, Parts A and B, Quart. Appl. Math. no. 4, 45-88, 112-441, 1946.

[9].Wahba G., *Smoothing noisy data by spline functions*, Numer. Math., 24, 383-393, 1975.

[10].Wahba G., *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[11].Wahba G., *An Introduction to Model Building With Reproducing Kernel Hilbert Spaces*, Tech. Report 1020, Univ of Wisconsin, Madison, 2000.

Authors:

Nicoleta Breaz
Department of Mathematics
"1 Decembrie 1918" University of Alba Iulia,
str. N. Iorga, No. 11-13, 510009,
Alba, Romania,
e-mail: nbreaz@uab.ro

Mihaela Aldea
Department of Mathematics
"1 Decembrie 1918" University of Alba Iulia,
str. N. Iorga, No. 11-13, 510009,
Alba, Romania,
e-mail: maldea@uab.ro