

Available online at www.sciencedirect.com



C. R. Acad. Sci. Paris, Ser. I 346 (2008) 107-112

COMPTES RENDUS Mathematique

http://france.elsevier.com/direct/CRASS1/

Numerical Analysis

# Error estimates in smoothing noisy data using cubic B-splines

Slim Chaabane<sup>a</sup>, Chokri Elhechmi<sup>b</sup>, Mohamed Jaoua<sup>c</sup>

<sup>a</sup> ENIT-LAMSIN, BP 37, 1002 Tunis-Bélvédère & Faculté des Sciences de Sfax, Tunisia <sup>b</sup> ENIT-LAMSIN & Faculté des Sciences de Gabès, Tunisia <sup>c</sup> Laboratoire J.-A. Dieudonné, Université de Nice, parc Valrose, 06108 Nice cedex 02, France

Received 20 May 2007; accepted after revision 29 October 2007

Available online 18 December 2007

Presented by Philippe G. Ciarlet

### Abstract

We are interested in finding out the best way to smooth noisy data using cubic B-splines. Two small parameters are involved: the noise level  $\epsilon$ , and the splines path h. The error in maximum norm between the actual data and the smoothed noisy ones is proved to be an  $O(\epsilon + h^2)$ , which leads to the choice of a splining path  $h = O(\sqrt{\epsilon})$ . Error estimates on the derivatives are also provided, and a splining protocol is derived from them, which ensures an  $O(\epsilon^{\frac{1}{2p}})$  maximum norm error on the p-th derivatives. To cite this article: S. Chaabane et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008). © 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

### Résumé

Estimations d'erreur sur des données lissées par B-splines cubiques. On s'intéresse à déterminer la meilleure facon de lisser des données bruitées en utilisant des B-splines cubiques. Nous prouvons que l'erreur en norme  $L^{\infty}$  entre les données exactes et les données lissées est en  $O(\epsilon + h^2)$ , où  $\epsilon$  désigne le niveau du bruit et h le pas du lissage, ce qui nous amène à choisir  $h = O(\sqrt{\epsilon})$ .

Des estimations d'erreur sur les dérivées d'ordre p sont aussi démontrées en choisissant un pas  $h_p = O(\epsilon^{\frac{1}{2^p}})$ . Pour citer cet article : S. Chaabane et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008).

© 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

# Version française abrégée

La plupart des applications basées sur l'exploitation de données provenant de mesures, ou de résultats de calculs antérieurs, nécessitent que ces données soient au préalable lissées. C'est le cas par exemple lorsqu'on résout des problèmes inverses d'identification de défauts ou de coefficients à partir de mesures de frontière. Les fonctions Bsplines sont souvent utilisées à cet effet, car elles conjuguent commodité d'usage - les données lissées étant obtenues grâce à une formule de type interpolation - à des propriétés de localité, les modifications apportées aux données en un point n'ayant d'impact sur les données lissées que dans un voisinange restreint de ce point.

E-mail addresses: slim.chaabane@fsm.rnu.tn (S. Chaabane), chokri.elhechmi@issatm.rnu.tn (C. Elhechmi), jaoua@math.unice.fr (M. Jaoua).

<sup>1631-073</sup>X/\$ - see front matter © 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved. doi:10.1016/j.crma.2007.10.045

Lorsqu'on s'intéresse aux questions de robustesse d'algorithmes, par exemple quand il s'agit de résoudre des problèmes inverses, il est important de connaître, pour un niveau de bruit donné, l'erreur commise en approchant les données exactes par les données bruitées lissées. Assez curieusement, cette question relativement élémentaire d'analyse numérique ne semble pas avoir reçu de réponse jusqu'ici. C'est la motivation de la présente étude.

Soit donc f une fonction de classe  $C^2$  sur l'intervalle I := [0, 1] que l'on divisera en n intervalles  $[x_i, x_{i+1}]$ , i = 0, ..., n-1 de longueur égale à  $h = \frac{1}{n}$ . On désigne par  $f^{\varepsilon} := f + \varepsilon$  les données bruitées,  $\varepsilon \in L^{\infty}(I)$  représentant le bruit qui vérifie  $\|\varepsilon\|_{0,\infty,I} = \epsilon$ . Le lissage de ces données par des B-splines cubiques nous conduit à la fonction  $\tilde{f}^{\varepsilon}$ (cf. [3,4]) définie par :

$$\tilde{f}^{\varepsilon}(x) = \sum_{i=-3}^{n-1} f^{\varepsilon}(x_{i+2}) B_{i,3}(x) \quad \forall x \in [0,1]$$

Par définition des splines cubiques, cette fonction est globalement de classe  $C^2$ , et polynomiale de degré 3 sur chacun des segments  $[x_i, x_{i+1}], i = 0, ..., n - 1$ . Le résultat essentiel du présent travail réside alors dans les estimations d'erreur suivantes :

$$\begin{cases} \|\tilde{f}^{\varepsilon} - f\|_{\infty, I} \leq c_1(\epsilon + h^2), \\ \|(\tilde{f}^{\varepsilon} - f)'\|_{\infty, I} \leq c_2(\frac{\epsilon}{h} + h) \end{cases}$$

Ces estimations conduisent naturellement à choisir un pas de lissage h d'ordre  $\sqrt{\epsilon}$ , ce qui donne une erreur en  $O(\epsilon)$  sur les données f, et en  $O(\sqrt{\epsilon})$  sur leurs dérivées f'. L'estimation sur les dérivées nous permet d'interpréter  $(\tilde{f}^{\varepsilon})'$  (dérivée premiere des données lissées) comme une perturabtion de la dérivée des données exactes, avec un niveau de bruit d'ordre  $\sqrt{\epsilon}$ . Il est alors loisible d'utiliser les résultats ci-dessus pour conclure qu'en lissant à nouveau ces données  $(\tilde{f}^{\varepsilon})'$  avec des splines cubiques, cette fois avec un pas  $\epsilon^{\frac{1}{4}}$ , puis en les dérivant, nous obtiendrons une estimation d'erreur d'ordre  $\epsilon^{\frac{1}{4}}$  sur les dérivées secondes. La procédure peut être ensuite réitérée autant de fois que la régularité des données exactes le permet, l'estimation obtenue sur la dérivée p-ème des données exactes étant alors en  $O(\epsilon^{\frac{1}{2p}})$ .

La seconde partie du travail porte sur la validation numérique de ces estimations. A cet effet, nous avons, pour plusieurs niveaux de bruit  $\epsilon$ , recherché le pas  $h_{opt}$  correspondant conduisant à la meilleure approximation des données. Nous établissons ce faisant que ce  $h_{opt}$  est bien comme le prévoit l'estimation en  $O(\sqrt{\epsilon})$ , et qu'en le choisissant ainsi, nous sommes bien conduits à une erreur en  $O(\epsilon)$  sur les données et en  $O(\sqrt{\epsilon})$  sur leurs dérivées premières. Conduite sur les dérivées d'ordre supérieur (jusqu'à l'ordre 3), cette étude permet également de valider la procédure dite de *boot strapping* décrite plus haut.

### 1. Introduction

Most of the applications based on the exploitation of noisy data provided by measurements or some computational process require the smoothing of these data by some way or other, prior to going through the application itself. This is for instance the case when solving inverse problems, aiming to recover coefficients or geometries from measured data. B-spline functions are very popular, especially in computer aided design applications, since they are quite easy to handle (the approximation is obtained by interpolation-like formulae) and have moreover local properties: modifying a value impacts only locally the approximation computed by B-splines.

When addressing robustness issues in inverse problems, one needs to know – given a noise level – how close to the actual data the smoothed noisy ones are. Quite surprisingly, up to our search, no error estimate seems to be available in the literature for this basic numerical analysis issue. This is the main motivation of the present study.

In Section 2, we shall first recall some basic properties of the B-splines, as given for example in [3] and [4]. Then, we shall establish two  $L^{\infty}$  error estimates, the first one regarding the smoothed data, and the second one their first derivatives. These estimates will provide us with a compatibility relationship between the noise level  $\epsilon$  and interval path *h* of the grid the B-splines are set up on. Provided the data are smooth enough, let's say  $f \in C^p$ ;  $p \in \mathbb{N}$ , we define in the last part of this section a splining protocol that lead to a general  $L^{\infty}$  estimate between the *p*-th order derivatives of the data. A numerical validation of the whole of these estimates is eventually presented in Section 3.

#### 2. Smoothing using B-splines

Let *I* be the interval [0, 1]. Given  $n \in \mathbb{N}^*$ , let us define the path  $h = \frac{1}{n}$ , and  $x_i = \frac{i}{n} = ih$ ,  $i \in \{0, ..., n\}$ . For  $s \in \mathbb{N}$  and a < b, we denote by  $C^s([a, b])$  the set of functions admitting *s*-th order continuous derivatives on

For  $s \in \mathbb{N}$  and a < b, we denote by  $C^{s}([a, b])$  the set of functions admitting *s*-th order continuous derivatives on [a, b], and by  $C^{-1}([a, b])$  the class of piecewise continuous functions on [a, b], admitting discontinuities only at the knots  $(x_i)_{i=0,...,n}$ .

Let us first recall a few basic definitions and properties of B-splines, as presented in [3,4].

**Definition 2.1.** A *k*-th order spline function (k = 0, 1, 2, 3), w.r.t. the knots  $x_0, \ldots, x_n$ , is a  $C^{k-1}([0, 1])$  real piecewise polynomial function of degree  $\leq k$  on each segment  $[x_j, x_{j+1}], j \in \{0, 1, \ldots, n-1\}$ .

Let us denote by  $S_k$  the space of *k*-th order spline functions, the dimension of which is n + k. Additional knots are thus needed in order to define a basis constituted with functions related to the knots. Let  $k \in \mathbb{N}^*$  and  $j \in \{1, ..., k\}$ , we define the following fictitious knots thus are  $x_{-j} = -jh$ ,  $x_{n+j} = n + jh$ .

Cox and De Boor [1,2] have defined the so called B-spline functions defined by the following formulae:

$$B_{i,0}(x) = \begin{cases} 1 & \text{for } x_i \leqslant x < x_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$
(1)

$$B_{i,k}(x) = \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+1}} B_{i+1,k-1}(x)$$
(2)

for i = -k, ..., (n - 1) and k = 1, 2, 3.

The set  $\mathcal{B}_k := \{B_{i,k}; i = -k, ..., (n-1)\}$  is therefore a basis of  $\mathcal{S}_k$  the elements of which hold the following properties:

$$B_{i,k} \ge 0, \quad \sum_{i=-k}^{n-1} B_{i,k} = 1 \quad \text{and} \quad B_{i,k}(x) = 0 \quad \text{for } x \notin [x_i, x_{i+k}].$$
 (3)

Let now  $S = \sum_{i=-k}^{n-1} \gamma_i B_{i,k}$  be a spline function. Using properties (3) and referring to [3], we get:

$$S' = \sum_{i=j-k+1}^{J} \beta_i B_{i,k-1} \quad \text{on} \ [x_j, x_{j+1}] \text{ where } \beta_i = k \frac{\gamma_i - \gamma_{i-1}}{x_{i+k} - x_i}.$$
(4)

In the following, we shall only study the cubic case (k = 3). Given  $i \in \{-3, ..., (n - 1)\}$ , and referring to [3] and [4], we derive:

$$B_{i-2,3}(x_i) = \frac{2}{3}, \quad B_{i-2,3}(x_{i+1}) = B_{i-2,3}(x_{i-1}) = \frac{1}{6},$$
(5)

$$B_{i,3}(x) = B_{0,3}(x - ih)$$
 and  $B_{i-2,3}(x) = B_{i,3}(x + 2h)$  on [0, 1]. (6)

Let now  $\varepsilon \in L^{\infty}(I)$  and  $f \in C^{2}(I)$ . We denote by  $f^{\varepsilon}$  an  $L^{\infty}(I)$  data perturbation of the exact data f:

$$f^{\varepsilon} = f + \varepsilon.$$

Let us denote respectively by  $\tilde{f}$  and  $\tilde{f}^{\varepsilon}$  the following expansions of f and  $f^{\varepsilon}$  on the cubic B-spline basis:

$$\tilde{f} = \sum_{i=j-3}^{j} \alpha_i B_{i,3}$$
 on  $[x_j, x_{j+1}]$ , and  $\tilde{f}^{\varepsilon} = \sum_{i=j-3}^{j} \alpha_i^{\varepsilon} B_{i,3}$  on  $[x_j, x_{j+1}]$ , (7)

where  $\alpha_i = f_{i+2} = f(x_{i+2})$  and  $\alpha_i^{\varepsilon} = f_{i+2}^{\varepsilon} = f^{\varepsilon}(x_{i+2})$ .

For  $i \in \{-1, ..., (n+1)\}$ , we have denoted  $\varepsilon(x_i)$  by  $\varepsilon_i$  and  $f^{\varepsilon}(x_i) = f_i + \varepsilon_i$  by  $f_i^{\varepsilon}$ . One should notice here that the functions values at nodes  $x_{-1}$  and  $x_{n+1}$ , which are undefinite, are needed to write down such expansions. The usual way to overcome this difficulty is to 'double' the functions values at the interval extremities, i.e. to assign the values  $f(x_{-1}) = f(x_0)$  and  $f(x_{n+1}) = f(x_n)$  for both functions f and  $\varepsilon$ . For  $p \in \{1, 2\}$ , let us denote by  $f_i^{(p)} = f^{(p)}(x_i)$  the p-th derivative of f at the knot  $x_i$ .

Using Eqs. (5) and (6), we obtain:

$$\tilde{f}(x) = \sum_{i=j-1}^{j+2} f_i B_{i,3}(x+2h), \quad \text{and} \quad \tilde{f}^\varepsilon(x) = \sum_{i=j-1}^{j+2} f_i^\varepsilon B_{i,3}(x+2h) \quad \text{on} \ [x_j, x_{j+1}].$$
(8)

We have the following lemma:

**Lemma 2.2.** Let  $f \in C^2(I)$ , then, there exist a positive constant *c* such that:

$$\|\tilde{f}' - f'\|_{L^{\infty}(I)} \leqslant ch.$$

$$\tag{9}$$

The main result of this Note is the following:

**Theorem 2.3.** Let  $f \in C^2(I)$ , then there exist two constants  $c_1, c_2 > 0$  such that:

$$\|\tilde{f}^{\varepsilon} - f\|_{\infty,I} \leq c_1(\epsilon + h^2) \quad and \quad \left\| (\tilde{f}^{\varepsilon} - f)' \right\|_{\infty,I} \leq c_2 \left(\frac{\epsilon}{h} + h\right), \quad where \ \epsilon = \|\varepsilon\|_{0,\infty,I}.$$

**Remark 1.** Theorem 2.3 suggests that, given a noise level  $\epsilon$ , the best smoothing using cubic B-splines would be obtained by choosing a splining path  $h = \sqrt{\epsilon}$ . Doing so, we get:

$$\|\tilde{f}^{\varepsilon} - f\|_{0,\infty,I} = 2c_1\epsilon = \mathcal{O}(\epsilon); \left\| (\tilde{f}^{\varepsilon} - f)' \right\|_{0,\infty,I} = 2c_2\sqrt{\epsilon} = \mathcal{O}(\sqrt{\epsilon}).$$
<sup>(10)</sup>

Thanks to the above estimate,  $(\tilde{f}^{\varepsilon})'$  may be seen as noisy data related to f', with a noise level  $\eta = O(\sqrt{\varepsilon})$ . This gives raise to iterating the process: these data may be smoothed again using cubic B-splines, with the appropriate path which is actually of order  $\epsilon^{\frac{1}{4}}$ . By boot-strapping, we can generalize this result to any order p, provided the function f is smooth enough.

Suppose  $f \in C^{n+1}$ ;  $n \ge 1$ , and let us denote by  $f_0^{\varepsilon} = f^{\varepsilon}$  and by  $f_p^{\varepsilon} = (\widetilde{f_{p-1}^{\varepsilon}})'$ ;  $p \in \{1, ..., n\}$ , where  $\widetilde{f_{p-1}^{\varepsilon}}$  denotes the function obtained by smoothing  $f_{p-1}^{\varepsilon}$  with  $h = \varepsilon^{\frac{1}{2p}}$ . The following result then holds true:

**Corollary 2.4.** Let  $f \in C^{n+1}n \ge 1$  and  $p \in \{1, ..., n\}$ , then,  $\|f_p^{\epsilon} - f^{(p)}\|_{0,\infty,I} = O(\epsilon^{\frac{1}{2^p}})$ .

## 3. Numerical validations

# 3.1. Validating the optimal path $h_{\text{opt}} \simeq \sqrt{\epsilon}$

Given a noise level  $\epsilon$ , we have run several smoothings with various paths, and plotted the errors w.r.t. the splining path *h*. From this plot, we can derive the optimal path  $h_{opt}(\epsilon)$  for each noise level  $\epsilon$ . Reiterating this operation for various noise levels, we are able to draw a log–log plot  $\ln(h_{opt}) v \sin(\epsilon)$ , the slope of which gives us the power of  $\epsilon$  in the  $h_{opt}$  relationship. The numerical slope obtained, as shown in Fig. 1, is 0.51.

### 3.2. Validating the error estimates

We aim here to validate the estimates 10, provided the optimal path is used in the splining process. For various values of  $\epsilon$ , we have thus computed the maximum norm errors  $||f - \tilde{f}^{\varepsilon}||_{0,\infty,I}$  and  $||f' - \tilde{f}^{\varepsilon'}||_{0,\infty,I}$  w.r.t.  $\epsilon$ , when using the optimal path  $h_{\text{opt}}(\epsilon)$  in the splining process. Fig. 2 shows that the numerical slope obtained in the log–log plots are 0.97 instead of 1 and 0.51 instead of 0.5

### 3.3. Validating the boot-strapping estimates

As described above, the first iteration in the boot-strapping procedure consists to reiterate the splining on the derivatives  $f_1^{\varepsilon} := (\tilde{f}^{\varepsilon})'$ , which may be seen – thanks to the estimates (10) – as the actual data derivative f' plus and



Fig. 1. (a) Smoothing using B-splines; (b) seeking the optimal path  $h_{\text{opt}}$  minimizing  $\|\tilde{f}^{\epsilon} - f\|_{0,\infty,I}$ ; (c) validation of  $h_{\text{opt}} = O(\sqrt{\epsilon})$  – the numerical slope obtained is 0.51.

Fig. 1. (a) Lissage des données utilisant les B-splines ; (b) pas optimal  $h_{\text{opt}}$  minimisant  $\|\tilde{f}^{\epsilon} - f\|_{0,\infty,I}$ ; (c) validation de  $h_{\text{opt}} = O(\sqrt{\epsilon})$  – la pente calculée est 0,51.



Fig. 2. (a)  $\|\tilde{f}^{\varepsilon} - f\| \approx cte\epsilon$  (numerical slope = 0.97); (b)  $\|(\tilde{f}^{\varepsilon} - f)'\| \approx cte\sqrt{\epsilon}$  (numerical slope = 0.51). Fig. 2. (a)  $\|\tilde{f}^{\varepsilon} - f\| \approx cte\epsilon$  (pente calculée = 0,97); (b)  $\|(\tilde{f}^{\varepsilon} - f)'\| \approx cte\sqrt{\epsilon}$  (pente calculée = 0,51).

Table 1 Validation of $h \simeq \epsilon^{\frac{1}{2^p}}$ , $p = 1$ , $p = 2$ , $p = 3$ Tableau 1 Validation de $h \simeq \epsilon^{\frac{1}{2^p}}$ , $p = 1$ , $p = 2$ , $p = 3$								
					Optimal path	$h_0$	$h_1$	$h_2$
					Theoretical power	0.5	0.25	0.125
					Computed power	0.51	0.24	0.11

additional noise of order  $\sqrt{\epsilon}$ . The best way to smooth these data is thus, thanks to Theorem 2.3, to choose a splining path of order  $\epsilon^{\frac{1}{4}}$ , thus yielding the following estimate on the second order derivative:

$$\|f_2^{\epsilon} - f''\|_{L^{\infty}(I)} = \mathcal{O}(\epsilon^{\frac{1}{4}})$$

and by reiterating the process as many times as allowed by the smoothness properties of f:

$$\|f_p^{\epsilon} - f^p\|_{0,\infty,I} = \mathcal{O}(\epsilon^{\frac{1}{2^p}}).$$

In Tables 1 and 2, we have displayed the outputs of our run numerical trials for p = 2, 3, trying to validate as described in the above subsection the optimal path needed, and the obtained errors once this path has been chosen.

Table 2	1				
Validation of $  f_p^{\epsilon} - f^{(p)}  _{\infty, I} = O(\epsilon^{\frac{1}{2^{p}}}), p = 1, p = 2, p = 3$					
Tableau 2					
Validation de $  f_p^{\epsilon} - f^{(p)}  _{\infty,I} = O(\epsilon^{\frac{1}{2^p}}), p = 1, p = 2, p = 3$					
Error estimates	$\ f_1^{\epsilon} - f^{(1)}\ _{\infty,I}$	$\ f_2^{\epsilon} - f^{(2)}\ _{\infty,I}$	$\ f_3^{\epsilon} - f^{(3)}\ _{\infty,I}$		
Slope	0.51	0.28	0.17		

## 4. Conclusions

We have proved, and numerically validated, error estimates on noisy data when smoothed using cubic B splines, as well their first derivatives. These estimates lead to a 'best choice' of the splines path h with respect to the noise level  $\epsilon$ . Estimates on the higher order derivatives have also been obtained by reiterating the smoothing process. Although cubic splines are the most popular, it might be interesting to extend these results to any order splines.

### Acknowledgements

The LAMSIN's researchers work is supported by the Tunisian Ministry for Higher Education, Scientific research and Technology, through the LR99ES20 program.

### References

[1] M.G. Cox, The numerical evaluation of B-splines, J. Inst. Math. Appl. 10 (1972) 134-149.

[2] C. De Boor, On calculating with B-splines, J. Approx. Theory 6 (1972) 50-62.

[3] C. De Boor, A Practical Guide to Splines, Springer Series in Applied Mathematics, vol. 27, 2001.

[4] P. Lancaster, K. Salkauskas, Curve and Surface Fitting: An Introduction, Academic Press, London, 1986.