

Splines Lissantes

François Boulier

5 octobre 2023

Ce document fournit une justification de la notion de *spline lissante* — *smoothing spline* en Anglais en n'utilisant que des arguments mathématiques accessibles à un étudiant de la spécialité IS. Il s'agit d'une notion importante en *data science*, abordée dans [2] et implantée dans le logiciel R.

Le texte de référence est celui de Reinsch [6]. Ce chapitre a beaucoup bénéficié du livre de Prenter [5] qui constitue un texte de référence très lisible sur les splines en général, de l'article de Pollock [4] pour la présentation calculatoire et de la relecture du texte de Reinsch par Cline [1].

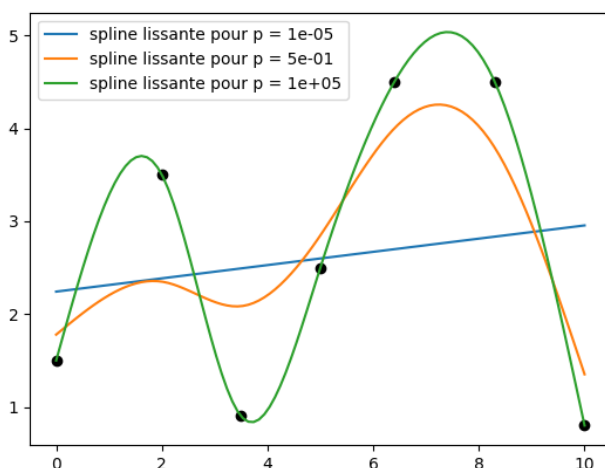


FIGURE 1 – Des données expérimentales et les splines lissantes pour trois valeurs du paramètre de lissage p . On observe que, quand p tend vers zéro, la spline tend vers la droite de régression linéaire et que, quand p tend vers $+\infty$, elle tend vers la spline interpolante.

1 Le problème posé par Reinsch

On considère $n + 1$ points (x_i, y_i) pour $0 \leq i \leq n$. À chaque point est associé un réel $\sigma_i > 0$. On suppose que les abscisses $a = x_0 < x_1 < \dots < x_n = b$ (appelées aussi les *nœuds* — *knots* en Anglais) subdivisent l'intervalle $[a, b]$ en n sous-intervalles.

On suppose fixé un paramètre de lissage $S \geq 0$. Parmi toutes les fonctions $g(x)$ qui vérifient la

contrainte (1)

$$\sum_{i=0}^n \left(\frac{g(x_i) - y_i}{\sigma_i} \right)^2 \leq S. \quad (1)$$

Reinsch cherche une fonction deux fois dérivable sur $[a, b]$ qui minimise l'intégrale

$$\int_a^b (g''(x))^2 dx. \quad (2)$$

2 La solution est une spline cubique

2.1 Rappels sur les splines cubiques

Définition 1 Une spline cubique est une fonction $s(x)$ deux fois dérivable sur l'intervalle $[a, b]$, qui se réduit à un polynôme de degré 3 sur chaque sous-intervalle $[x_i, x_{i+1}]$ de $[a, b]$.

Concrètement, toute spline cubique est donc une fonction définie par morceaux. Concrètement, toute spline cubique est donc une fonction définie par morceaux de la forme suivante. Des conditions sur les coefficients font que les raccords entre les morceaux (au niveau des nœuds intérieurs) sont C^2 .

$$s(z) = \begin{cases} s_0(z) = a_0 + b_0(z - x_0) + c_0(z - x_0)^2 + d_0(z - x_0)^3, & (z \leq x_1), \\ s_1(z) = a_1 + b_1(z - x_1) + c_1(z - x_1)^2 + d_1(z - x_1)^3, & (x_1 \leq z \leq x_2), \\ s_2(z) = a_2 + b_2(z - x_2) + c_2(z - x_2)^2 + d_2(z - x_2)^3, & (x_2 \leq z \leq x_3), \\ \vdots & \\ s_i(z) = a_i + b_i(z - x_i) + c_i(z - x_i)^2 + d_i(z - x_i)^3, & (x_i \leq z \leq x_{i+1}), \\ \vdots & \\ s_{n-1}(z) = a_{n-1} + b_{n-1}(z - x_{n-1}) + c_{n-1}(z - x_{n-1})^2 + d_{n-1}(z - x_{n-1})^3, & (x_{n-1} \leq z). \end{cases}$$

où chaque morceau $s_i(z)$ est défini par un polynôme de degré 3.

Entre deux nœuds consécutifs, le graphe d'une spline cubique est celui d'un polynôme : il est donc indéfiniment dérivable (il est de classe C^∞). Au niveau des nœuds intérieurs (on appelle *nœuds intérieurs* les $n-1$ nœuds x_1, \dots, x_{n-1}), le raccord entre deux cubiques consécutives est de classe C^2 seulement. Mathématiquement, une spline est donc *moins lisse* qu'un polynôme d'interpolation, bien qu'on puisse avoir le sentiment contraire visuellement !

Le fait que les raccords soient C^2 suffit à assurer à la spline cubique un graphe proche de celui qu'on aurait naturellement tracé à la main. La Figure 2 montre qu'un raccord C^1 n'est en général pas suffisant visuellement ; il est très difficile d'apprécier à l'œil nu la différence entre un raccord C^2 et un raccord C^3 .

Définition 2 On dit qu'une spline cubique $s(x)$ interpole une fonction $f(x)$ dérivable sur $[a, b]$ si $s(x_i) = f(x_i)$ pour $0 \leq i \leq n$.

Définition 3 Une spline cubique $s(x)$ est dite naturelle si $s''(a) = s''(b) = 0$.

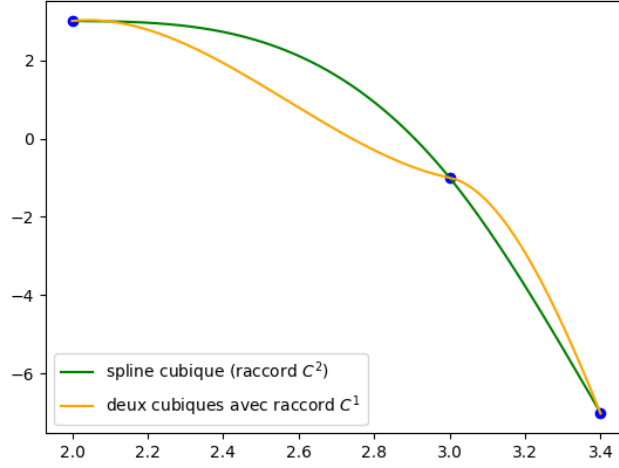


FIGURE 2 – La différence entre un raccord C^1 et un raccord C^2 peut s'apprécier visuellement. Le raccord a lieu au point $(3, -1)$.

2.2 La solution est une spline cubique

La proposition suivante correspond à la *first integral relation* de [5, page 96].

Proposition 1 Soient $f(x)$ une fonction deux fois dérivable sur l'intervalle $[a, b]$ et $s(x)$ une spline cubique naturelle interpolant f . Alors

$$\int_a^b (f'')^2 dx = \int_a^b (s'')^2 dx + \int_a^b (f'' - s'')^2 dx. \quad (3)$$

Preuve. Comme

$$(f'')^2 = ((f'' - s'') + s'')^2 = (f'' - s'')^2 + 2(f'' - s'')s'' + (s'')^2,$$

il suffit de prouver que

$$\int_a^b (f'' - s'')s'' dx = 0.$$

Le reste de la preuve est un exercice d'intégration par parties. Comme la dérivée troisième s''' est constante sur chaque sous-intervalle, on découpe l'intégrale en

$$\int_a^b (f'' - s'')s'' dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (f'' - s'')s'' dx.$$

Sur chaque sous-intervalle, on a, en intégrant par parties,

$$\int_{x_{i-1}}^{x_i} (f'' - s'')s'' dx = [(f' - s')s'']_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} (f' - s')s''' dx.$$

Mais la somme suivante est nulle : l'égalité ① est due à une simplification télescopique ; l'égalité ② est due¹ à l'hypothèse de spline naturelle : $s''(a) = s''(b) = 0$:

$$\sum_{i=1}^n [(f' - s') s'']_{x_{i-1}}^{x_i} \stackrel{\textcircled{1}}{=} [(f' - s') s'']_a^b \stackrel{\textcircled{2}}{=} 0.$$

Il ne reste plus qu'à vérifier que la somme suivante est nulle : l'égalité ③ s'établit en remarquant que s''' est une constante c_i sur chaque sous-intervalle $[x_{i-1}, x_i]$; l'égalité ④ en remarquant que $f - s$ est une primitive de $f' - s'$; l'égalité ⑤ en utilisant le fait que s interpole f .

$$\sum_{i=1}^n \int_{x_{i-1}}^{x_i} (f' - s') s''' dx \stackrel{\textcircled{3}}{=} \sum_{i=1}^n c_i \int_{x_{i-1}}^{x_i} (f' - s') dx \stackrel{\textcircled{4}}{=} \sum_{i=1}^n c_i [f - s]_{x_{i-1}}^{x_i} \stackrel{\textcircled{5}}{=} 0.$$

□

Les deux propositions suivantes sont très proches des deux théorèmes donnés dans [5, pages 100-101]. Elles découlent de la proposition 1.

Proposition 2 *La fonction identiquement nulle est l'unique spline interpolante naturelle de la fonction identiquement nulle.*

Preuve. La fonction identiquement nulle est une spline interpolante naturelle de la fonction identiquement nulle $f(x)$. Montrons l'unicité. Soit $s(x)$ une spline interpolante naturelle de $f(x)$. Alors, d'après la proposition 1 on a

$$\int_a^b (s'')^2 dx \leq \int_a^b (f'')^2 dx = 0.$$

Par conséquent, $s'' = 0$ sur $[a, b]$ et s est donc un polynôme de degré 1 sur $[a, b]$ qu'on peut écrire

$$s(x) = s(a) + s'(a)(x - a).$$

Mais comme s interpole la fonction nulle on a $s(a) = 0$ et donc $s'(a)(b - a) = 0$, qui implique $s'(a) = 0$ aussi. Par conséquent s est identiquement nulle. □

Proposition 3 *Soit f une fonction deux fois dérivable sur $[a, b]$. Il existe une unique spline naturelle s qui interpole f .*

Preuve. On commence par montrer que les $4n$ coefficients des n cubiques qui constituent s sont contraints par un système de $4n$ équations linéaires. Les conditions de spline naturelle $s''(a) = s''(b) = 0$ donnent deux équations. Les conditions $s(x_i) = f(x_i)$ donnent $n + 1$ équations. Les conditions qui expriment le fait que les cubiques doivent se raccorder ainsi que leurs dérivées première et deuxième, sur les nœuds intérieurs donnent $3(n - 1)$ équations. Au total, on obtient $4n$ équations et donc un système d'équations linéaires $Ac = v$ où A est une matrice $(4n) \times (4n)$, c est le vecteur des coefficients des cubiques et v est un vecteur ne comportant que des zéros, les valeurs de f aux nœuds et les valeurs de f' en a et en b . La proposition 2 montre que dans le cas où f est la fonction identiquement nulle, ce système admet une unique solution. Donc la matrice A est inversible. Donc la spline s existe et est unique quelle que soit la fonction f . □

1. Prenter suppose $f'(a) = s'(a)$ et $f'(b) = s'(b)$, ce qui conduit à la même conclusion.

La proposition suivante aussi est une conséquence de la proposition 1. Elle est plus ou moins écrite dans [5, page 101].

Proposition 4 *Soit $f(x)$ est une fonction quelconque, deux fois dérivable sur $[a, b]$, interpolant $n+1$ points d'abscisses $a = x_0 < x_1 < \dots < x_n = b$, d'ordonnées y_0, y_1, \dots, y_n . Si $s(x)$ est la spline cubique naturelle interpolant ces mêmes points alors*

$$\int_a^b (f'')^2 dx \geq \int_a^b (s'')^2 dx. \quad (4)$$

En particulier, si l'inégalité (4) est une égalité alors f est une spline cubique (mais pas forcément la spline cubique naturelle).

Preuve. La proposition 3 montre l'existence de s . L'inégalité (4) est une conséquence de la proposition 1 et du fait que l'intégrale (5) est positive ou nulle.

$$\int_a^b (f'' - s'')^2 dx \quad (5)$$

Supposons que l'inégalité (4) soit une égalité et donc que (5) soit nulle. Comme f et s sont deux fois dérivables, leur dérivée seconde est continue et la différence $f'' - s''$ est la fonction nulle sur $[a, b]$. Par conséquent, f'' est un polynôme de degré 1 sur chaque sous-intervalle; donc f est un polynôme de degré 3 sur chaque sous-intervalle; donc f est une spline cubique d'après la définition 1. \square

Corollaire 1 *Le problème posé par Reinsch a une solution quel que soit S . Cette solution est une spline cubique.*

Preuve. L'ensemble des fonctions g satisfaisant la contrainte (1) n'est jamais vide (même pour $S = 0$), puisqu'il contient au minimum le polynôme interpolant les $n + 1$ points. La fonction f qui minimise (2) existe donc toujours. D'après la proposition 4, cette fonction est une spline cubique. \square

3 Une propriété de la dérivée troisième de la solution

Si $s(x)$ est une spline cubique, la dérivée troisième s''' n'est pas continue en général au niveau des nœuds. Pour tout nœud x_i , on distingue donc la valeur de s''' à gauche de x_i , notée $s'''(x_i)_-$, de sa valeur à droite, notée $s'''(x_i)_+$. Pour homogénéiser les notations, on pose $s'''(x_0)_- = s'''(x_n)_+ = 0$.

Soit $f(x)$ la spline cubique qui minimise (2) parmi les fonctions satisfaisant la contrainte (1). Reinsch montre qu'il existe un paramètre p tel que

$$f'''(x_i)_- - f'''(x_i)_+ = 2p \frac{f(x_i) - y_i}{\sigma_i^2} \quad (0 \leq i \leq n). \quad (6)$$

Cette section est destinée à montrer cette propriété. Dans le texte de Reinsch, p est un *multiplicateur de Lagrange*. Dans l'analyse ci-dessous, qui est due à Cline [1], on évite toute référence à cette notion.

Cline commence par démontrer la proposition suivante [1, Lemma 1], qui est une variation sur l'argument central de la proposition 1.

Proposition 5 Soient $s(x)$ une spline cubique naturelle de nœuds $a = x_0 < x_1 < \dots < x_n = b$ et $\ell(x)$ une fonction deux fois dérivable sur $[a, b]$. Alors

$$\int_a^b \ell'' s'' dx = \sum_{i=0}^n \ell(x_i) (s'''(x_i)_+ - s'''(x_i)_-). \quad (7)$$

Preuve. Comme la dérivée troisième s''' est constante sur chaque sous-intervalle, on découpe l'intégrale en

$$\int_a^b \ell'' s'' dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \ell'' s'' dx.$$

Sur chaque sous-intervalle, on a, en intégrant par parties,

$$\int_{x_{i-1}}^{x_i} \ell'' s'' dx = [\ell' s'']_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} \ell' s''' dx$$

Mais la somme suivante est nulle : l'égalité (1) est due à une simplification télescopique ; l'égalité (2) est due à l'hypothèse de spline naturelle :

$$\sum_{i=1}^n [\ell' s'']_{x_{i-1}}^{x_i} \stackrel{(1)}{=} [\ell' s'']_a^b \stackrel{(2)}{=} 0.$$

À partir d'ici, la preuve diffère de celle de la proposition 1. Il ne reste plus qu'à simplifier la somme suivante : l'égalité (3) s'établit en remarquant que s''' est constante et égale à $s'''(x_{i-1})_+$ sur chaque sous-intervalle $[x_{i-1}, x_i]$ et que ℓ est une primitive de ℓ' ; l'égalité (4) s'obtient en découpant la somme en deux sous-sommes ; l'égalité (5) s'obtient en décalant de 1 les indices de la première sous-somme de façon à faire apparaître $\ell(x_i)$ en facteur dans les deux expressions ; l'égalité (6) vient du fait que $s'''(x_i)_- = s'''(x_{i-1})_+$; l'égalité (7) n'est due qu'à un réarrangement des sommes ; l'égalité (8) s'obtient en utilisant le fait que $s'''(x_0)_- = s'''(x_n)_+ = 0$.

$$\begin{aligned} - \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \ell' s''' dx &\stackrel{(3)}{=} - \sum_{i=0}^{n-1} s'''(x_i)_+ (\ell(x_{i+1}) - \ell(x_i)) \\ &\stackrel{(4)}{=} - \sum_{i=0}^{n-1} s'''(x_i)_+ \ell(x_{i+1}) + \sum_{i=0}^{n-1} s'''(x_i)_+ \ell(x_i) \\ &\stackrel{(5)}{=} - \sum_{i=1}^n s'''(x_{i-1})_+ \ell(x_i) + \sum_{i=0}^{n-1} s'''(x_i)_+ \ell(x_i) \\ &\stackrel{(6)}{=} - \sum_{i=1}^n s'''(x_i)_- \ell(x_i) + \sum_{i=0}^{n-1} s'''(x_i)_+ \ell(x_i) \\ &\stackrel{(7)}{=} s'''(x_0)_+ \ell(x_0) - s'''(x_n)_- \ell(x_n) + \sum_{i=1}^{n-1} \ell(x_i) (s'''(x_i)_+ - s'''(x_i)_-) \\ &\stackrel{(8)}{=} \sum_{i=0}^n \ell(x_i) (s'''(x_i)_+ - s'''(x_i)_-). \end{aligned}$$

□

La proposition suivante est [1, Lemma 2].

Proposition 6 Soient u et v deux vecteurs non nuls de \mathbb{R}^{n+1} . Si, pour tout $p \in \mathbb{R}$ on a $u \neq pv$ alors il existe un vecteur w tel que $w^T u > 0$ et $w^T v < 0$.

Preuve. On note $\|\cdot\|$ la norme euclidienne. Supposons que pour tout $p \in \mathbb{R}$ on ait $u \neq pv$. Les deux vecteurs ne sont donc pas colinéaires et, d'après l'inégalité de Cauchy-Schwartz on a $u^T v < \|u\| \|v\|$. Par conséquent,

$$\|u\| - \frac{u^T v}{\|v\|} > 0 \quad \text{et} \quad \|v\| - \frac{u^T v}{\|u\|} > 0.$$

Le vecteur w ci-dessous vérifie $w^T u > 0$ et $w^T v < 0$:

$$w = \frac{u}{\|u\|} - \frac{v}{\|v\|}.$$

□

Proposition 7 Supposons $S > 0$. Soit $s(x)$ une spline cubique naturelle de nœuds $a = x_0 < x_1 < \dots < x_n = b$ et vérifiant la contrainte (1). Si, pour tout $p \in \mathbb{R}$ on a

$$s'''(x_i)_- - s'''(x_i)_+ \neq 2p \frac{s(x_i) - y_i}{\sigma_i^2} \quad (0 \leq i \leq n). \quad (8)$$

alors il existe une autre fonction $t(x)$ deux fois dérivable sur $[a, b]$, satisfaisant la contrainte (1) et telle que

$$\int_a^b (s'')^2 dx > \int_a^b (t'')^2 dx. \quad (9)$$

En particulier, s n'est pas une solution du problème de Reinsch.

Preuve. La non existence de p implique que l'une des quantités $s'''(x_i)_- - s'''(x_i)_+$ soit non nulle. On distingue deux cas. Premier cas : on suppose que l'une au moins des quantités $s(x_i) - y_i$ est non nulle. Alors, d'après la proposition 6, il existe un vecteur w tel que

$$\sum_{i=0}^n w_i (s'''(x_i)_- - s'''(x_i)_+) > 0 \quad \text{et} \quad \sum_{i=0}^n w_i \frac{s(x_i) - y_i}{\sigma_i^2} < 0.$$

Soit $\ell(x)$ une fonction deux fois dérivable sur $[a, b]$ telle que $\ell(x_i) = w_i$ pour $0 \leq i \leq n$ (le polynôme d'interpolation convient). Alors, en remplaçant w_i par $\ell(x_i)$ dans les inégalités précédentes, on obtient :

$$\sum_{i=0}^n \ell(x_i) (s'''(x_i)_- - s'''(x_i)_+) > 0, \quad (10)$$

$$\sum_{i=0}^n \ell(x_i) \frac{s(x_i) - y_i}{\sigma_i^2} < 0. \quad (11)$$

Pour tout $\mu \in \mathbb{R}$ on a les égalités suivantes : l'égalité ① s'obtient en développant le carré ; l'égalité ② s'obtient en appliquant la proposition 5 sur l'intégrale du milieu :

$$\begin{aligned} \int_a^b (s'' + \mu \ell'')^2 dx &\stackrel{\textcircled{1}}{=} \int_a^b (s'')^2 dx + 2\mu \int_a^b \ell'' s'' dx + \mu^2 \int_a^b (\ell'')^2 dx \\ &\stackrel{\textcircled{2}}{=} \int_a^b (s'')^2 dx + 2\mu \underbrace{\sum_{i=0}^n \ell(x_i) (s'''(x_i)_+ - s'''(x_i)_-)}_{\text{négatif pour } \mu > 0 \text{ très petit}} + \mu^2 \int_a^b (\ell'')^2 dx. \end{aligned}$$

D'après l'inégalité (10), le terme du milieu est négatif pour tout $\mu > 0$ et, pour peu que ce même μ soit suffisamment petit, la somme des deux derniers termes est négative. La fonction $t = s + \mu \ell$ vérifie l'inégalité (9). Pour montrer que s n'est pas la solution optimale (pour le premier cas), il ne reste plus qu'à montrer que t vérifie la contrainte (1). On a l'égalité suivante :

$$\begin{aligned} \sum_{i=0}^n \left(\frac{s(x_i) + \mu \ell(x_i) - y_i}{\sigma_i^2} \right)^2 &= \\ \sum_{i=0}^n \left(\frac{s(x_i) - y_i}{\sigma_i^2} \right)^2 &+ \underbrace{2\mu \sum_{i=0}^n \ell(x_i) \frac{s(x_i) - y_i}{\sigma_i^2} + \mu^2 \sum_{i=0}^n \left(\frac{\ell(x_i)}{\sigma_i^2} \right)^2}_{\text{négatif pour } \mu > 0 \text{ très petit}}. \end{aligned} \quad (12)$$

D'après l'inégalité (11), le terme du milieu est négatif pour tout $\mu > 0$ et, pour peu que ce même μ soit suffisamment petit, la somme des deux derniers termes est négative. Par conséquent, si s vérifie la contrainte (1) alors la fonction $t = s + \mu \ell$ la vérifie également, ce qui achève la preuve du premier cas.

Second cas : on suppose que toutes les quantités $s(x_i) - y_i$ sont nulles. Il est encore possible de trouver un vecteur w tel que l'inégalité (10) soit satisfaite (il suffit de prendre des w_i de même signe que les $s'''(x_i)_- - s'''(x_i)_+$). Par conséquent, en construisant $\ell(x)$ comme dans le premier cas et en prenant $\mu > 0$ suffisamment petit, on obtient une fonction $t = s + \mu \ell$ qui vérifie (9). Pour montrer que s n'est pas la solution optimale (pour le second cas), il ne reste plus qu'à montrer que t vérifie la contrainte (1). On a l'égalité suivante qui est une forme simplifiée de (12) :

$$\sum_{i=0}^n \left(\frac{s(x_i) + \mu \ell(x_i) - y_i}{\sigma_i^2} \right)^2 = \mu^2 \sum_{i=0}^n \left(\frac{\ell(x_i)}{\sigma_i^2} \right)^2$$

Comme on a supposé $S > 0$, en prenant μ suffisamment petit, cette expression est inférieure ou égale à S . La fonction $t = s + \mu \ell$ vérifie donc la contrainte (1) et la proposition est prouvée dans tous les cas. \square

4 Calcul des coefficients de la spline

Dans cette section-ci, on revient aux explications données par Reinsch. On obtient des formules pour les coefficients de la spline qui utilisent la relation (6) et dépendent donc du paramètre $p > 0$ (le *multiplieur de Lagrange*) introduit dans cette relation. On verra en section 5 comment calculer p à partir de S et réciproquement.

On note $s(x)$ la spline cubique naturelle solution du problème de Reinsch. Cette spline est composée de n polynômes de degré 3 :

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad (x_i \leq x \leq x_{i+1}). \quad (13)$$

Écrire chaque polynôme s_i ($0 \leq i \leq n-1$) sous la forme d'un développement limité en $x = x_i$ simplifie considérablement les formules qui donnent leurs coefficients. Notons

$$h_i = x_{i+1} - x_i, \quad (0 \leq i \leq n-1). \quad (14)$$

La continuité de la dérivée seconde de s aux nœuds intérieurs se traduit par le fait que $s''_{i+1}(x_{i+1}) = 2c_{i+1}$ doit être égal à $s''_i(x_{i+1}) = 6d_i h_i + 2c_i$. En tirant d_i on obtient :

$$d_i = \frac{c_{i+1} - c_i}{3h_i}, \quad (0 \leq i \leq n-1). \quad (15)$$

La continuité de s aux nœuds intérieurs se traduit par le fait que $s_{i+1}(x_{i+1}) = a_{i+1}$ doit être égal à $s_i(x_{i+1}) = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3$. En tirant b_i on obtient :

$$b_i = \frac{a_{i+1} - a_i}{h_i} - c_i h_i - d_i h_i^2, \quad (0 \leq i \leq n-1). \quad (16)$$

La continuité de la dérivée première de s aux nœuds intérieurs se traduit par le fait que $s'_i(x_i) = b_i$ doit être égal à $s'_{i-1}(x_i) = b_{i-1} + 2c_{i-1} h_{i-1} + 3d_{i-1} h_{i-1}^2$. En remplaçant b_{i-1} , d_{i-1} , b_i et d_i par leur valeur (15), (16), on obtient

$$\begin{aligned} \frac{h_{i-1}}{3} c_{i-1} + 2 \left(\frac{h_{i-1}}{3} + \frac{h_i}{3} \right) c_i + \frac{h_i}{3} c_{i+1} = \\ \frac{1}{h_{i-1}} a_{i-1} - \left(\frac{1}{h_{i-1}} + \frac{1}{h_i} \right) a_i + \frac{1}{h_i} a_{i+1}, \quad (1 \leq i \leq n-1). \end{aligned} \quad (17)$$

Enfin, la relation (6) se traduit par le fait que $s'''_i(x_i) - s'''_{i-1}(x_i) = 6d_i - 6d_{i-1}$ doit être égal à $2p(y_i - s_i(x_i))/\sigma_i^2$. En remplaçant d_i et d_{i-1} par leur valeur (15) on obtient

$$\frac{1}{h_{i-1}} c_{i-1} - \left(\frac{1}{h_{i-1}} + \frac{1}{h_i} \right) c_i + \frac{1}{h_i} c_{i+1} = p \frac{y_i - a_i}{\sigma_i^2}, \quad (1 \leq i \leq n-1). \quad (18)$$

4.1 Formulation matricielle

Introduisons la matrice T de dimension $(n-1) \times (n-1)$ suivante (attention au coefficient 1/3)

$$T = \frac{1}{3} \begin{pmatrix} 2(h_0 + h_1) & h_1 & 0 & \dots & 0 & 0 \\ h_1 & 2(h_1 + h_2) & h_2 & \dots & 0 & 0 \\ 0 & h_2 & 2(h_2 + h_3) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2(h_{n-3} + h_{n-2}) & h_{n-2} \\ 0 & 0 & 0 & \dots & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix},$$

la matrice Q de dimension $(n+1) \times (n-1)$ suivante où $g_i = 1/h_i$:

$$Q = \begin{pmatrix} g_0 & 0 & \dots & 0 & 0 \\ -g_0 - g_1 & g_1 & \dots & 0 & 0 \\ g_1 & -g_1 - g_2 & \dots & 0 & 0 \\ 0 & g_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -g_{n-3} - g_{n-2} & g_{n-2} \\ 0 & 0 & \dots & g_{n-2} & -g_{n-2} - g_{n-1} \\ 0 & 0 & \dots & 0 & g_{n-1} \end{pmatrix},$$

la matrice diagonale Σ formée de $\sigma_0, \sigma_1, \dots, \sigma_n$, ainsi que les vecteurs $a = (a_0, a_1, \dots, a_n)^T$, $c = (c_1, c_2, \dots, c_{n-1})^T$ et $y = (y_0, y_1, \dots, y_n)^T$. Les relations (15) et (18) auxquelles il faut ajouter les conditions de spline naturelle $c_0 = c_n = 0$ s'écrivent alors

$$T c = Q^T a, \quad (19)$$

$$Q c = p \Sigma^{-2} (y - a). \quad (20)$$

En multipliant (20) à gauche par $Q^T \Sigma^2$ et en éliminant $Q^T a$ grâce à (19) on trouve une équation matricielle qui ne dépend que du vecteur c :

$$(Q^T \Sigma^2 Q + p T) c = p Q^T y \quad (21)$$

ainsi qu'une équation pour le vecteur a :

$$a = y - p^{-1} \Sigma^2 Q c. \quad (22)$$

La matrice des coefficients de l'équation (21) est symétrique à cinq bandes. Elle est définie positive si $p > 0$. Il ne reste plus qu'à résoudre (21) (Cholesky) pour obtenir c , reporter dans (22) pour obtenir a puis utiliser (15) et (16) pour obtenir d et b .

5 Calcul du paramètre de lissage

Pour comparer le texte fondateur de Reinsch avec des travaux plus modernes [2], il est utile de présenter la formulation du problème de Reinsch par la formule (24). On en profite pour montrer que la borne S est atteinte à l'optimum (ce qui peut aussi se montrer sans la formulation Lagrangienne, comme le fait remarquer Cline). Dans les sections suivantes, on montre comment calculer p à partir de S , en utilisant d'intéressantes techniques de dérivation d'expressions matricielles. Puis on introduit la notion de degré de liberté, qui fournit un paramétrage plus élégant, utilisable dans des contextes plus généraux et qui est implanté dans le logiciel R. Enfin, on montre comment définir le paramètre optimal (en un certain sens), ce qui permet de déterminer automatiquement le lissage à appliquer lorsque l'utilisateur ne souhaite pas préciser le degré de liberté qu'il souhaite.

5.1 Formulation avec multiplicateur de Lagrange

La méthode du *multiplicateur de Lagrange* est une méthode générale permettant de transformer un problème d'optimisation sous contrainte en un problème d'optimisation équivalent, sans contrainte, mais au prix de l'introduction d'une variable supplémentaire appelée *multiplicateur de Lagrange*.

Le problème posé par Reinsch consiste à minimiser (2) sous la contrainte d'inégalité (1). Pour appliquer la méthode du multiplicateur de Lagrange, il est nécessaire de transformer la contrainte d'inégalité en une contrainte d'égalité. On introduit une première variable z (une variable d'écart) pour représenter la différence entre les deux membres de l'inégalité (1) et on code le sens de l'inégalité en posant que cette différence est égale à z^2 (nécessairement positif ou nul) plutôt qu'à z :

$$\sum_{i=0}^n \left(\frac{g(x_i) - y_i}{\sigma_i} \right)^2 = S - z^2. \quad (23)$$

D'après la théorie du multiplicateur de Lagrange, minimiser (2) sous la contrainte (23) équivaut à minimiser le critère suivant, sans contrainte, où on a introduit une seconde variable p (le multiplicateur de Lagrange) :

$$\mathcal{L} = \int_a^b (g'')^2 dx + p \left(\sum_{i=0}^n \left(\frac{g(x_i) - y_i}{\sigma_i} \right)^2 - S + z^2 \right). \quad (24)$$

Comme dans le cas classique des fonctions de plusieurs variables réelles, toute solution optimale de (24) annule le système des dérivées partielles du critère, par rapport à chacune des variables dont il dépend. En particulier, toute solution optimale de (24) annule

$$\frac{\partial \mathcal{L}}{\partial z} = 2pz.$$

À l'optimum, on a donc soit $p = 0$ et la solution de (24) est une droite, soit $z = 0$ (le cas qui nous intéresse). Supposons $z = 0$. D'après (23), on conclut que S est égale à

$$\sum_{i=0}^n \left(\frac{g(x_i) - y_i}{\sigma_i} \right)^2. \quad (25)$$

Avec les notations de la section 4.1 et les formules (21) et (22), la somme (25) est le carré de la norme deux d'un certain vecteur et peut s'écrire :

$$(\|\Sigma^{-1}(a - y)\|_2)^2 = (\|\Sigma Q(Q^T \Sigma^2 Q + pT)^{-1} Q^T y\|_2)^2 = F(p)^2.$$

Le paramètre p de la formule ci-dessus est celui introduit dans la relation (6). Il se trouve que c'est le même paramètre que le multiplicateur de Lagrange p (voir l'article de Reinsch) mais c'est sans importance pour ce qui nous occupe. On a donc établi que

$$S = F(p)^2. \quad (26)$$

5.2 Calcul de p en fonction de S

Pour calculer S en fonction de p , il suffit d'évaluer $F(p)$. Pour calculer p en fonction de S , Reinsch propose d'utiliser une méthode de Newton, ce qui suppose de mettre au point une formule pour la dérivée $F'(p)$. C'est ce qu'on explique dans cette section.

On définit la dérivée u' d'un vecteur dont les coordonnées sont des fonctions u_i comme le vecteur des dérivées u'_i .

Notons $u = (Q^T \Sigma^2 Q + pT)^{-1} Q^T y$. Le vecteur u dépend de p . Ses coordonnées sont donc des fonctions de p . Notons $v = \Sigma Q u$. D'une part, d'après la règle sur la transposée d'un produit et le fait que Σ est égale à sa transposée, on a $v^T = u^T Q^T \Sigma$; d'autre part, comme les matrices Σ et Q ne dépendent pas de p et que la dérivée d'une somme est la somme des dérivées, on voit que $v' = \Sigma Q u'$. Enfin, en développant les expressions, on voit que

$$F(p)^2 = v^T v. \quad (27)$$

Par conséquent,

$$(F(p)^2)' = 2v^T v' = 2u^T Q^T \Sigma^2 Q u'.$$

On cherche maintenant une formule pour u' . La définition de u implique que

$$(Q^T \Sigma^2 Q + pT) u = Q^T y \text{ et donc} \quad (28)$$

$$Q^T \Sigma^2 Q u + T(pu) = Q^T y. \quad (29)$$

Comme les matrices et vecteurs Q , T , Σ et y ne dépendent pas de p on a, en dérivant par rapport à p ,

$$Q^T \Sigma^2 Q u' = -T(u + pu') \text{ et donc} \quad (30)$$

$$u' = -(Q^T \Sigma^2 Q + pT)^{-1} T u \quad (31)$$

Par conséquent,

$$\frac{1}{2} (F(p)^2)' = pu^T T (Q^T \Sigma^2 Q + pT)^{-1} T u - u^T T u. \quad (32)$$

On a tout ce qui est nécessaire pour appliquer la méthode de Newton. Dans [7, page 453], Reinsch conseille de partir de l'équation suivante, pour une convergence plus rapide :

$$\frac{1}{F(p)} = \frac{1}{\sqrt{S}} \quad (33)$$

5.3 Paramétrage par le degré de liberté

Paramétrer l'importance du lissage par p ou par S n'est pas très satisfaisant. Qui plus est, ces deux paramètres sont étroitement liés à la définition des splines lissantes telle que Reinsch l'a donnée. Dans la littérature, on présente parfois les splines comme des fonctions g qui minimisent d'autres critères comme [2, Eq (5.9), page 151] :

$$\text{RSS}(g, \lambda) = \sum_{i=0}^n (y_i - g(x_i))^2 + \lambda \int_{x_0}^{x_n} (g''(x))^2 dx. \quad (34)$$

Le paramètre λ ressemble fort à un multiplicateur de Lagrange mais ce n'en est pas un : il n'est pas multiplié par la contrainte ! Pour autant, il joue approximativement le rôle de $1/p$ et permet de contrôler l'importance du lissage. Et il existe plein de variantes de ces définitions. Le critère (34) est probablement adapté de [3, Eq (1.2), page 378] où la somme est divisée par le nombre de points.

La notion de *degré de liberté* [2, section 5.4.1, page 153] a le gros avantage d'être définie à partir de la spline solution et pas du critère. L'idée consiste à exprimer le vecteur a , qui contient les ordonnées de la spline aux abscisses x_i ($0 \leq i \leq n$), comme une fonction linéaire du vecteur y des ordonnées des points initiaux. Notons donc S_p la matrice telle que

$$a = S_p y. \quad (35)$$

Cette matrice se définit facilement à partir des équations de la section 4.1 par

$$S_p = I - \Sigma^2 Q (Q^T \Sigma^2 Q + pT)^{-1} Q^T. \quad (36)$$

Elle est symétrique définie positive. Ses valeurs propres sont donc réelles et positives. On peut montrer que les deux plus grandes sont exactement égales à 1 [2, page 156]. L'idée consiste à définir le *degré de liberté* de la spline comme la trace de S_p :

$$\text{df}_p = \text{Tr}(S_p). \quad (37)$$

La trace d'une matrice est la somme de ses valeurs propres. C'est aussi la somme de ses éléments diagonaux. On a donc $2 < \text{df}_p \leq n + 1$.

L'idée, c'est que le rang d'une matrice est égal au nombre de ses valeurs propres non nulles et que la trace de S_p est une mesure de la dimension de l'image de S_p et donc une mesure du nombre de paramètres effectifs dont dépend la spline.

Les splines lissantes implantées dans le logiciel R sont paramétrées par ce degré de liberté.

5.4 Calcul du paramètre optimal

On peut enfin vouloir déterminer le paramètre de lissage (quel qu'il soit) à partir d'un autre critère à optimiser. Pour fixer les idées, supposons que ce paramètre soit le paramètre p de la section 4.1 et notons g_p la spline lissante définie par p . Une idée populaire consiste à chercher le paramètre p_{opt} qui minimise le critère [2, Eq. (5.26) ; page 161]

$$\text{CV}(g_p) = \sum_{i=0}^n (y_i - g_p^{[i]}(x_i))^2 \quad (38)$$

où $g_p^{[i]}$ est la spline lissante obtenue, avec le paramètre p , en prenant en compte tous les points sauf le i -ème. Il s'agit d'une technique de *validation croisée* où on ne laisse de côté qu'un seul point (*leave one out cross validation*). Il est remarquable que (38) peut être calculée à partir de la spline lissante g_p obtenue en tenant compte de tous les points, grâce à la formule suivante [2, Eq. (5.27) ; page 161], où on a supposé que les lignes et colonnes de la matrice S_p sont indicées à partir de 0 :

$$\text{CV}(g_p) = \sum_{i=0}^n \left(\frac{y_i - g_p(x_i)}{1 - S_p(i, i)} \right)^2. \quad (39)$$

C'est ce type de méthode qui permet de déterminer automatiquement le paramètre de lissage dans un logiciel comme R, lorsque l'utilisateur ne spécifie pas le degré de liberté voulu.

Pour déterminer la valeur optimale p_{opt} du paramètre de lissage, il est utile de savoir que la fonction qui à p associe $\text{CV}(g_p)$ a la forme convexe dessinée Figure 3.

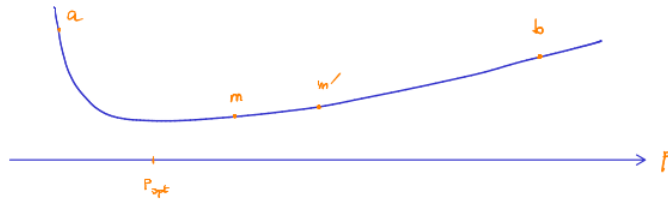


FIGURE 3 – Représentation schématique de la fonction qui à un paramètre de lissage p associe $\text{CV}(g_p)$. Pour déterminer le paramètre optimal — l'abscisse du minimum de la courbe — un algorithme simple consiste à partir de trois points choisis de telle sorte que le point intermédiaire m soit plus bas que les deux points extérieurs a et b . Il suffit ensuite de prendre un point $m' \neq m$ entre a et b . Supposons que m' soit entre m et b . S'il est plus bas que m alors remplacer (a, m, b) par (m, m', b) sinon remplacer (a, m, b) par (a, m, m') . L'intervalle $[a, b]$ s'est rétréci. En répétant ce procédé, on isole le paramètre optimal dans un intervalle arbitrairement petit.

Références

- [1] Alan Kaylor Cline. An Expansion of the Derivation of the Spline Smoothing Theory. <https://pdfs.semanticscholar.org/ae2c/118aaa18beeded133276c0cf0c9ee4c62d1d.pdf>, 2017.
- [2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 2009. Available at <https://hastie.su.domains/ElemStatLearn>.
- [3] Peter Craven and Grace Wahba. Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31 :377–403, 1979.
- [4] D. S. G. Pollock. Smoothing with cubic splines. *Handbook of Time Series Analysis, Signal Processing, and Dynamics*, 12 1999. Available at www.physics.muni.cz/~jancely/NM/Texty/Numerika/CubicSmoothingSpline.pdf.
- [5] Paddy M. Prenter. *Splines and Variational Methods*. Pure and Applied Mathematics. Wiley, New York, 1975.
- [6] Christian H. Reinsch. Smoothing by Spline Functions. *Numerische Mathematik*, 10 :177–183, 1967.
- [7] Christian H. Reinsch. Smoothing by Spline Functions II. *Numerische Mathematik*, 16 :451–454, 1971.