**Projet CN**

Lissage par Splines Cubiques

Maxime Sauvage

Jordan Laplace

**Table des matières**

[**Introduction**](#_hjs0pfylyp5z) **2**

[**1. Etude du problème**](#_moa8mhfqnaug) **3**

[1.1. Apprentissage supervisé](#_xnjuqwc8r8q9) 3

[1.2. Définition de la fonction de spline lissante](#_mf5x7dmplucg) 3

[1.3. Le cas de la spline interpolante](#_bo32trpg1byt) 3

[1.4. Simplification du problème en fonction du paramètre](#_9gmzcg8oda9i) 4

[1.5. Le terme général de la formule (3)](#_q7j4p858i8n5) 4

[1.6. Les matrices R et Q](#_y5w3vc8vfoek) 5

[1.7. Le Système d’équations linéaires pour obtenir b et d](#_hewfkdpp047o) 5

[**2. Résolution du problème**](#_k3ikonkuh5t0) **6**

[2.1. Le programme FORTRAN](#_5tcys2mcdvcb) 6

[2.2. Commandes de compilation et d’exécution](#_w4z0d191gdbs) 7

[2.3. Descriptions des fichiers utilisées](#_q8dov2nwmxr9) 7

[2.4. Utiliser notre programme avec un autre jeu de données](#_bkwnss1feroj) 8

[**3. Splines lissantes obtenues avec les différents jeu de données**](#_o7s98mgo0pv5) **9**

[3.1. Exemple du projet](#_x0nouyemkw90) 9

[3.2. Densité minérale osseuse relative en fonction de l’âge et du sexe.](#_r6v6aedec5sr) 12

[**Conclusion**](#_t51aeklnpdpp) **14**

# 

# Introduction

Dans le cadre module Calcul Numérique nous sommes amenés à étudier le Lissage par Splines Cubiques qui est un type de spline plus général qui dépend d’un paramètre [0,1]. On peut distinguer deux cas extrêmes le cas , la spline lissante sera la spline interpolante et le cas ou la spline lissante est la droite de régression linéaire définie par les points.

Le but de ce projet est d’effectuer des recherches sur ce type de spline mais nous devons également réaliser un programme FORTRAN permettant de calculer pour un nuage de points et un paramètre , le lissage par spline cubique. Nous avons déjà un fichier à traiter, cependant le projet nous suggère de travailler sur des exemples d’ordre n = 500.

Nous sommes fiers d’avoir pu tester notre programme sur de tel jeu de données et d’avoir pu reproduire les splines lissantes du livre *Elements of Statistical Learning* page 152 montrant la densité minérale osseuse relative en fonction de l’âge pour les hommes et les femmes.

Nous avons fait le choix de réaliser le graphique à l’aide d’un appel du logiciel Gnuplot dans notre programme, ce qui permet d’obtenir la Spline Lissante avec seulement une ligne de commande.

Dans ce compte rendu, nous allons tout d’abord présenter nos recherches sur le lissage par splines cubiques, ensuite nous détaillerons les subroutines principales présentes dans le programme FORTRAN , enfin nous discuterons autour des jeux de données que nous avons utilisées .

# 1. Etude du problème

## 1.1. Apprentissage supervisé

L’apprentissage supervisé est un apprentissage automatique à l’aide d’une fonction de prédiction sur des exemples étiquetés. La puissance de l’algorithme dépend de la précision de l’entraînement.

Il est opposé à l’apprentissage non supervisé. Dans ce cas, il n’y a plus l’intervention d’un être humain, l’algorithme va comprendre tout seul en cherchant des corrélations. Cette méthode va alors réaliser une partition et classer les données dans des groupes homogènes, on parle alors de clustering.

Ainsi les splines s’apparentent plutôt à une méthode de régression car les splines permettent de réaliser une tendance et non une classification des points.

## 1.2. Définition de la fonction de spline lissante

La fonction est une fonction définie par morceaux ou chaque morceau est définie par

(1)

Le lien entre (1) et la notion de développement limité est tel que les deux sont des approximations polynomiales, le développement limité d’une fonction en un point permet d’approximer la fonction au voisinage de ce point alors que la notation (1) permet d’approximer les points au voisinage de ce dernier à l’aide de la fonction définie par morceaux.

## 1.3. Le cas de la spline interpolante

Dans le cas de la spline interpolante, la courbe passe par les points ainsi pour nous avons :

(1)

Or

Nous obtenons ainsi (1)

## 1.4. Simplification du problème en fonction du paramètre

La formule dépend d’un paramètre or lorsque ce dernier vaut 1 ou 0, des simplifications sont possibles.

(3)

|  |  |
| --- | --- |
| Valeur de | Formulation de |
| 0 |  |
| 1 |  |

## 1.5. Le terme général de la formule (3)

Le terme intégral de la formule (3) peut s’apparenter à une “énergie”, en effet au lieu de faire passer la spline par les points (le cas de la spline interpolante) la spline passe alors par des ressorts. La spline décrit ainsi une fonction continue qui minimise l’énergie de déformations des ressorts et de la spline.

Ainsi l'énergie de courbure entre deux point et vaut :

où est la raideur de courbure élastique de la spline.

L'énergie de déformation pour toute la spline vaut :

si l'on suppose qu'elle présente une raideur constante sur toute sa longueur ce qui s’apparente à l’intégrale. Donc l’intégrale de la formule (3) constitue bien une énergie où est le rapport de raideur entre la spline et le ressort.

## 1.6. Les matrices R et Q

Lors de la détermination des coefficients R et Q nous étions amenés à résoudre le système d’équations linéaires suivants : .

Lors de la création du programme FORTRAN nous avons étudié les matrices et pour pouvoir les remplir. Ainsi nous avons commencé par déterminer les dimensions des matrices pour pouvoir les parcourir sans sortir ces dernières.

|  |  |
| --- | --- |
| Matrice | Dimension |
|  | (Nbpoints - 2) (Nbpoints - 2 ) |
|  | Nbpoints (Nbpoints - 2) |

Nbpointsétant la variable qui corresponds au nombre de couples (x,y) qui constituent le nuage de points.

a pour propriété d’être symétrique et tridiagonale.

## 1.7. Le Système d’équations linéaires pour obtenir *b* et *d*

Nous cherchons à résoudre le système linéaire suivant :

(7)

On s’intéresse aux propriétés de la matrice , est un réel, est la matrice identité, elle n’impacte donc pas le calcul de la matrice calculée à partir du membre gauche, ainsi le produit des matrices etest une matrice symétrique car le produit d’une matrice par sa transposée est une matrice symétrique de même taille que la matrice qui est symétrique également, par conséquent est une matrice symétrique.

Cette matrice étant définie positive et symétrique, nous pouvons appliquée la factorisation de Cholesky qui consiste à calculer la factorisation suivante :

Avec une matrice triangulaire supérieure.

Le système (7) est de la forme

En connaissant , obtenue grâce à la décomposition de Cholesky, on résout le système suivant :

(\*)

(\*\*)

On résout le système (\*) pour trouver le vecteur , puis le système (\*\*) pour trouver le vecteur . La résolution est facilitée par la forme triangulaire des matrices.

# 2. Résolution du problème

## 2.1. Le programme FORTRAN

Afin d’aérer le code du programme principal, nous avons réalisé des subroutines qui correspondent à chaque étape de calcul permettant de rendre plus facile la lecture des lignes de codes de ce programme.

Définition des subroutines utilisées :

* smouthing : programme principal utilisant les subroutines ci-dessous
* *make\_H* : Formation du vecteur pour à partir du vecteur
* *make\_R* : Formation de la matrice
* *make\_Q* : Formation de la matrice
* *make\_Msigma* : Formation de la matrice , une matrice diagonal où
* *make\_muqt* : Formation de ( matrice de taille (NBPNTS,NBPNTS -2))
* *CHOLESKY\_SOLVE* : Utilisation de la méthode de CHOLESKY pour résoudre le systeme linéraire
* *make\_D* : Calcul du vecteur
* *make\_A* : Calcul du vecteur
* *make\_C* : Calcul du vecteur
* *horner* : Permet de calculer un morceau de la spline lissante avec la méthode de horner
* *make\_spline* : Formation de la spline via la méthode de Horner et affichage via gnuplot

L’utilisation de ces subroutines nous a permis de réaliser la totalité des étapes de calculs permettant d’obtenir la spline lissante pour un nuage de points x et y.

Nous avons également fait appel à des fonctions de la bibliothèques BLAS notamment pour les produits matrice-vecteur et les produits matrice-matrice.

Notre subroutine *CHOLESKY\_SOLVE* provient directement du projet réalisé lors du précédent semestre.

## 2.2. Commandes de compilation et d’exécution

Une fois le programme compilé avec la ligne de commande suivante :

../gfor smoothing.f make\_H.f make\_R.f make\_Q.f make\_Msigma.f make\_muqt.f CHOLESKY\_SOLVE.f make\_spline.f make\_A.f make\_C.f make\_D.f horner.f.

Nous obtenons un exécutable ./a.out qui va nous permettre d’appliquer notre programme sur un nuage de points représenté par une colonne de x puis une colonne de y toute deux séparées par un espace dans un fichier.

La ligne de commande à utiliser depuis un terminal ouvert est la suivante sont : ./a.out < data.dat

Nous avons d’abord utilisé notre programme sur les données à traiter dans le sujet du projet. Nous avons obtenu la même trace de calcul ainsi qu’un lissage par spline cubique qui s’adapte bien aux nuages de points, de plus lorsque , nous avons bien une spline interpolante qui passe par tous les points.

## 2.3. Descriptions des fichiers utilisées

Notre dossier projet en plus de contenir les fichiers fortran indiquée dans la partie 2.1, contient aussi des fichiers permettant l’affichage et le calcul des splines. Voici une description de ces fichiers :

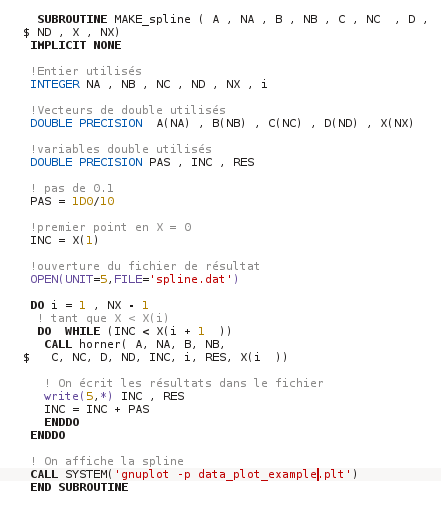
* *example.dat* : fichier contenant les données pour tester notre programme
* *BMDfemale.dat* : fichier contenant les données des mesures de la densité osseuse relative en fonction de l’âge pour les femmes
* *BMDmale.dat* : fichier contenant les données des mesures de la densité osseuse relative en fonction de l’âge pour les hommes
* *data\_plot\_example.plt* : fichier permettant d’afficher la spline du jeu de donné example.dat
* *data\_plot\_female.plt* : fichier permettant d’afficher la spline du jeu de donné BMDfemale.dat
* *data\_plot\_male.plt* : fichier permettant d’afficher la spline du jeu de donné BMDmale.dat
* *xy\_example.dat* : fichier contenant l'abscisse et l’ordonnée contenue dans example.dat
* *xy\_female.txt* : fichier contenant l'abscisse et l’ordonnée contenue dans BMDfemale.dat
* *xy\_male.txt* : fichier contenant l'abscisse et l’ordonnée contenue dans BMDmale.dat

## 2.4. Utiliser notre programme avec un autre jeu de données

Si vous souhaitez utiliser notre programme avec un autre jeu de données voici la marche à suivre.

Dans un premier temps il vous faut formater vos données pour obtenir un fichier du même format que *example.dat* , *data\_plot\_example.plt* et *xy\_example.dat*.

Dans un second temps il vout faut modifier la ligne CALL SYSTEM dans *make\_spline.f.* Dans cette ligne il vous suffit de changer le nom du fichier en .plt par le vôtre.



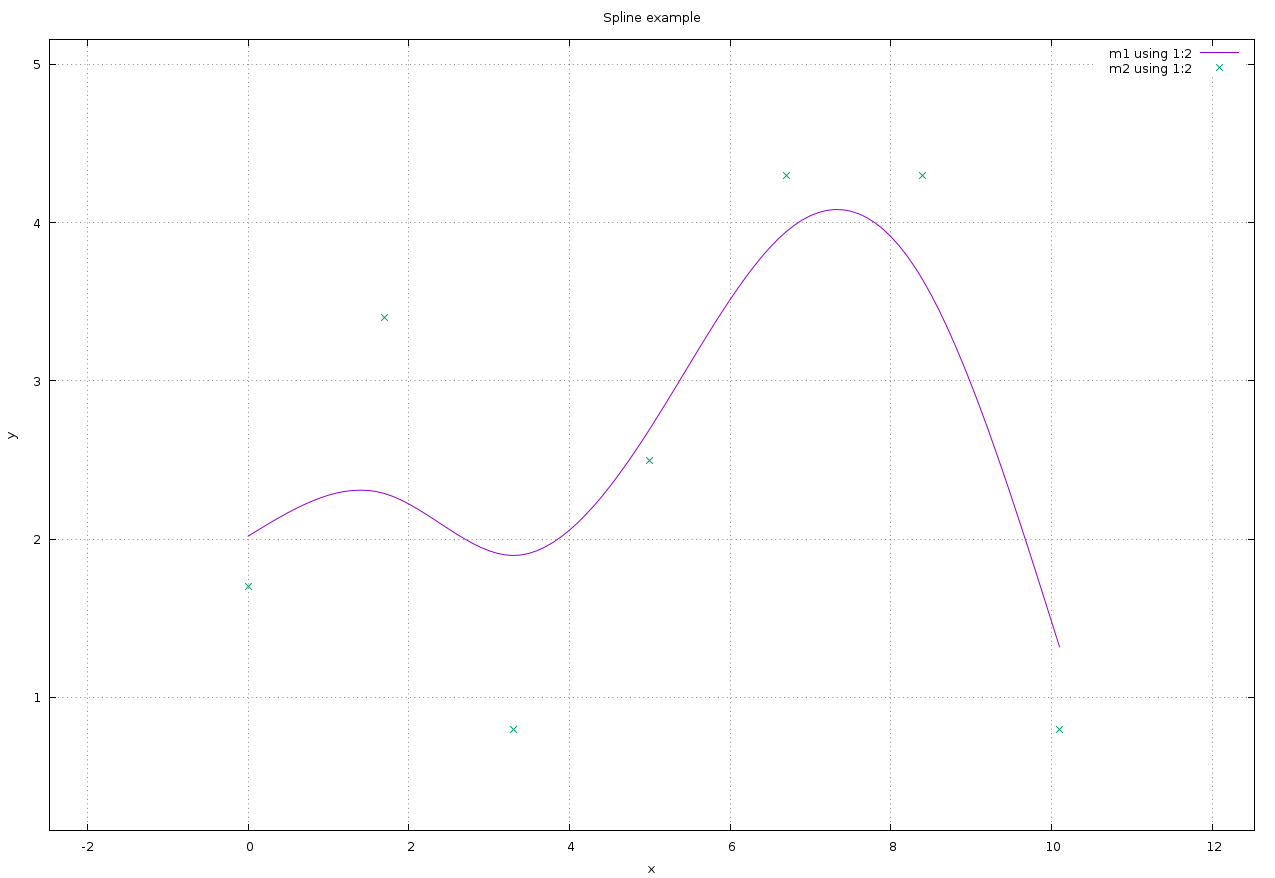
*Figure 1. Ligne à changer dans le code de la subroutine MAKE\_spline*

# 

# 3. Splines lissantes obtenues avec les différents jeu de données

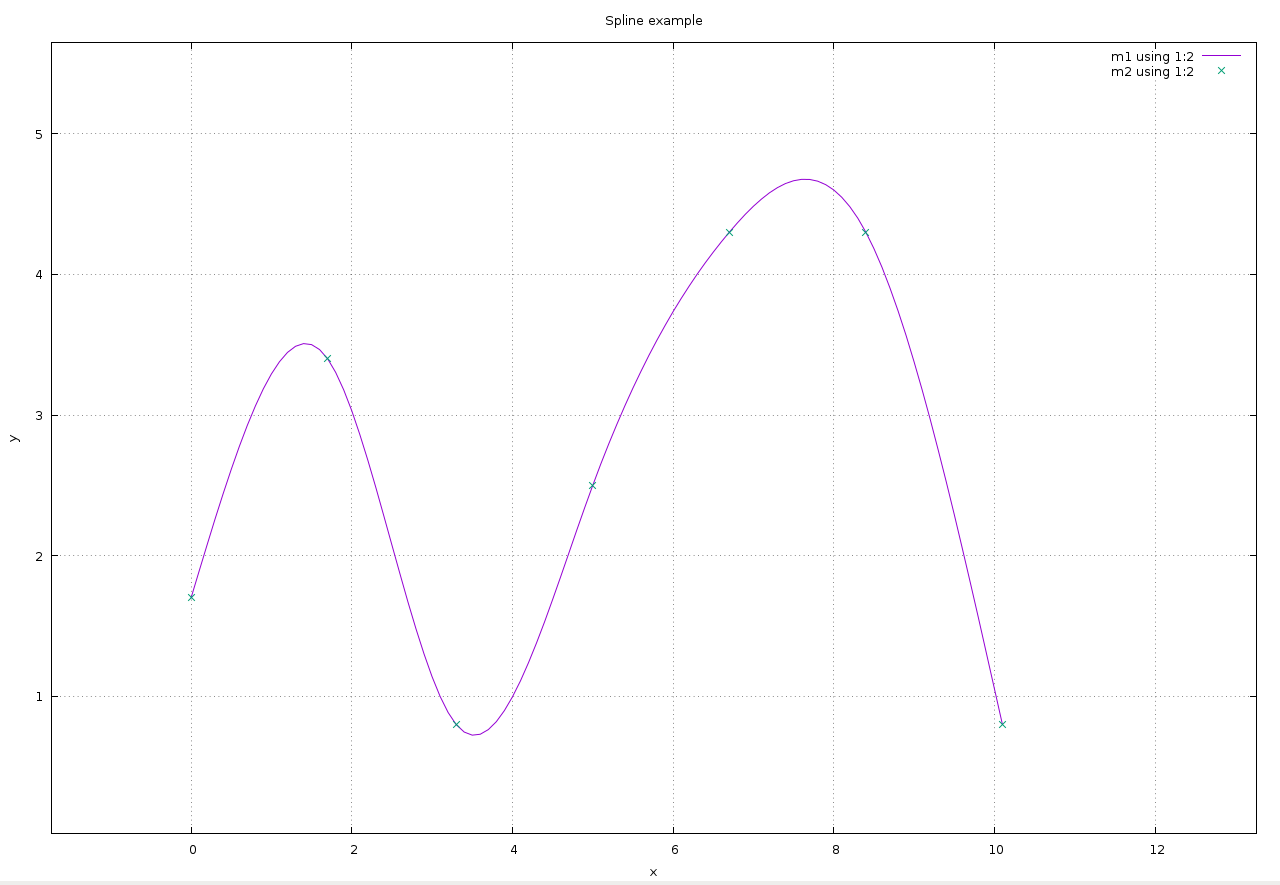
## 3.1. Exemple du projet

Voici la spline que nous obtenons avec le fichier example.dat ().



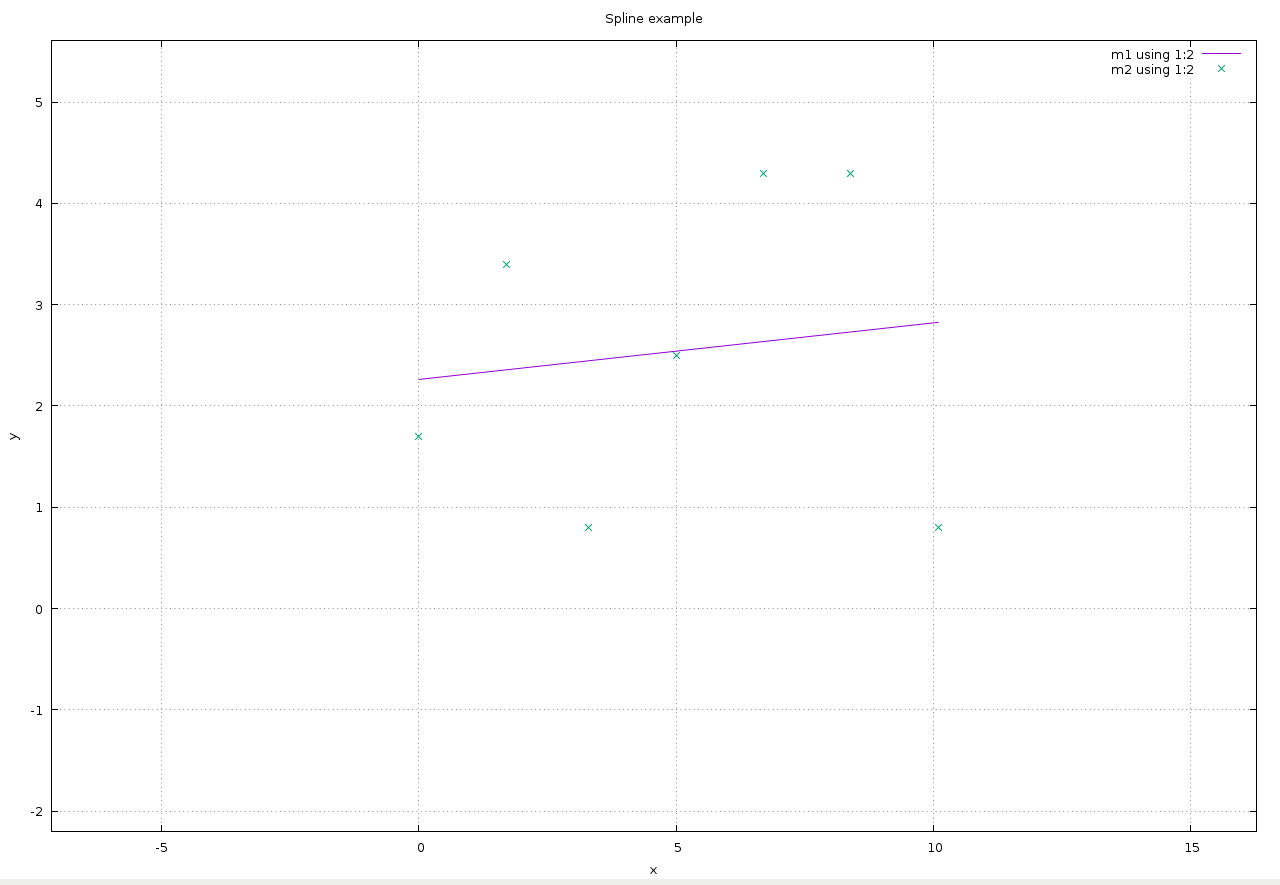
*Figure 2. Spline lissante réalisée à partir des données de example.dat avec*

Voici la spline que nous obtenons avec le fichier example.dat avec . Comme attendu cette spline passe tous les points.



*Figure 3. Spline lissante réalisée à partir des données de example.dat avec*

Voici la spline que nous obtenons avec le fichier example.dat avec proche de 0. En effet, il est difficile de mettre lambda à zéro car nous faisons une division par dans nos calcul. Comme attendu cette spline est une droite.



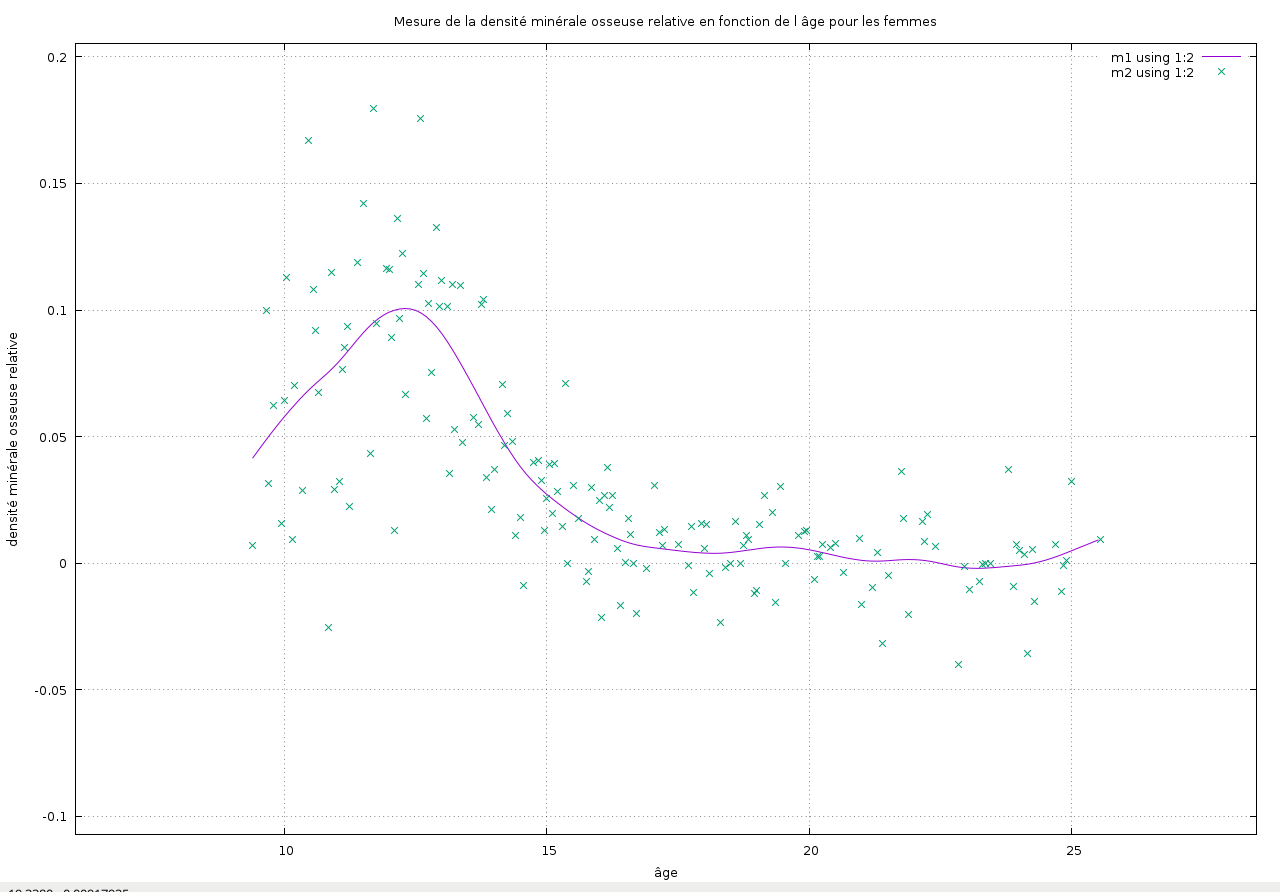
*Figure 4. Spline lissante réalisée à partir des données de example.dat avec proche* 0

## 

## 3.2. Densité minérale osseuse relative en fonction de l’âge et du sexe.

Le jeu de données liées aux densités minérales osseuses relative en fonction de l’âge et du sexe était disponible depuis le site <https://web.standford.edu/~hastie/ElemStatLearn> dans la rubrique data. Cependant ces données réclamaient un traitement, en effet il y avait une colonne *ind* qui ne nous concernait pas pour réaliser la spline et valeurs concernant les hommes et les femmes étaient mélangés De plus, il y avait plusieurs valeurs de densité pour un même âge.

C’est la raison pour laquelle nous avons utilisé le logiciel R, afin de séparé les données des hommes des données des femmes et nous avons décidé de faire la moyenne des densités qui correspondent à un même âge. Par exemple, pour un même X, nous faisons la moyenne des Y pour définir le couple (X,Y). Si nous avons trois X = 10 avec comme Y 31, 29 et 30, alors nous obtenons le couple (10,30). Enfin nous avons supprimé les colonnes avec les données inutile tel que le genre ou la colonne d’indice puis nous avons généré le fichier avec le même format que celui de l’exemple ainsi que le fichier avec les abscisses et les ordonnées correspondantes en colonne.

*Figure 5. Spline lissante réalisée à partir des données concernant les femmes avec* 

# 

*Figure 6. Spline lissante réalisée à partir des données concernant les hommes avec*

# 

# Conclusion

En guise de conclusion, nous pouvons dire que le projet nous a permis d’étudier en profondeur le lissage par spline cubique. Nous avons vu notamment leur utilisation sur des jeux de données conséquents tel que l’analyse de la mesure de la densité minérale osseuse en fonction de l’âge. Nous avons constaté que la construction d’une telle courbe demande beaucoup d’étape de calcul.

Nous avons réussi à traiter l’exemple donné par le projet et les données liées aux mesures de la densité minérale osseuse à l’aide de notre programme en langage FORTRAN. Nous pouvons appliquer notre programme sur tous les nuages de points tant que le jeu de données ne soit pas trop grand.

Pour aller plus loin, on aurait pu essayer de travailler sur la subroutine générant les graphiques afin de pouvoir superposer les splines. Par exemple, on aurait pu superposer la spline des données des hommes avec celle des données de femmes. De plus, on aurait pu également travaillé avec des données approximatives. Ainsi, on aurait pu travaillé avec la matrice avec l’erreur associé aux données.