Rapport Tâche 4 : détection de points aberrants

Interpolaspline

Avril 2020

1 Introduction

Il y a plusieurs méthodes différentes pour interpoler des données malgré des points aberrants. La méthode la plus intuitive est de détecter ces points et de les supprimer avant de tracer la spline. C'est cette détection de points aberrants qui a été étudiée durant la tâche 4. Trois membres du projet ont travaillé sur cette tâche : Béryl, Mohamed et Zakaria. Les méthodes ont été réparties entre ces membres du projet. Zakaria a également étudié le moyen de rendre l'étude locale (ce problème sera détaillé plus tard).

2 Objectif

L'objectif de la tâche était de rechercher et d'implémenter quelques méthodes de détection de points aberrants. Peu de temps était prévu pour cette tâche car on pensait que les méthodes seraient plutôt simples. C'est en effet le cas, néanmoins un problème s'est posé alors qu'on n'y avait pas pensé durant la planification des tâches : l'étude doit être faite localement, pour un groupe de points proches les uns des autres. En effet, si par exemple on prend la fonction identité discrétisée sur [0,10] avec un point aberrant (1,10), ce point ne sera pas detecté si l'on considère tous les points en même temps, car si celui-ci est décrété aberrant (car il se situe trop loin de la moyenne par exemple), alors (10,10) sera aussi considéré comme aberrant bien que ce ne soit pas le cas. Il faut donc trouver un moyen de séparer les points en groupes de points pas trop éloignés (exceptés éventuellement les aberrants) afin de les étudier groupe par groupe.

3 Implémentation

Nous allons détailler dans cette section les algorithmes des méthodes, puis ceux étudiés afin de créer des intervalles d'étude. Ensuite, nous allons effectuer des tests.

3.1 Algorithmes

3.1.1 Algorithmes des méthodes

Le premier algorithme créé prend en argument une liste d'ordonnées et une méthode de détection des points aberrants, et traite toute la liste.

Dans un second temps, un deuxième algorithme a été créé, permettant de ne traiter que le point d'indice i. Cela permettra à l'avenir de reprendre cet algorithme pour supprimer les points aberrants pendant l'interpolation.

Méthode des k plus proches voisins Calculer pour chaque observation la distance au K plus proche voisin k-distance ; Ordonner les observations selon ces distances k-distance ; Les données

FIGURE 1 – Exemple de traitements des données

aberrantes ont les plus grandes distances k-distance ; Les observations qui ont les n pourcent plus grandes distances k-distance sont des données aberrante, n étant un paramètre à fixer. Dans un premier temps, nous avons crée un algorithme qui prend en entrée une liste et un indice, et retourne une liste contenant les distances de l'élément à la position i à aux éléments de la liste. Ensuite, un deuxième qui comme la précédente prend en entrée une liste, un indice et un entier k et retourne la k-distance de l'élément à la position indice qui représente la moyenne ses k petites distances. Enfin,undernierquielleprenduneliste,unentierketunentierncommeindiquéaudessus, retourne la liste contenant les valeurs de la liste qui ont les n pourcent plus grande k-distance.

3.1.2 Algorithmes de création d'intervalles

3.2 tests

Les tests ont été réalisés sur la définition des points aberrant suivante :
Un point est dit aberrant lorsqu'il n'appartient pas à l'intervalle

$$[Q_1 - 1.5 * (Q_3 - Q_1), Q_3 + 1.5 * (Q_3 - Q_1)]$$

Le résultat est visible dans la figure 1.