Université Grenoble Alpes
MAP350 – L3 MI

TP8 : Approximation aux moindres carrés – Erreur résiduelle

Ce TP est consacré au problème de l'approximation. Précisément, étant donnée une suite de points (x_i, y_i) du plan, nous nous intéressons désormais à la construction d'une courbe polynomiale qui passe à *proximité* de ces points. La courbe étant supposée ne pas pouvoir *passer* par l'ensemble des points (typiquement, une droite devant passer à proximité de n ($n \geq 3$) points non alignés).
L'approximation privilégie plutôt la *forme globale* de la courbe recherchée (définie par un modèle mathématique) approchant *au mieux* (selon un critère à définir) les données dont le nombre est en général beaucoup plus grand que le nombre de paramètres libres dans le modèle mathématique.

Récupérer le script Python `TP8LeastSquaresStudents.py` permettant d'acquérir à la souris une suite de points (x, y) dont les abscisses sont deux à deux distinctes.

Droite de régression linéaire

Considérons une suite de n points (x_i, y_i) du plan à approcher par une droite $Y = a\,X + b$. L'approche est ici matricielle, de type *moindres carrés*, et permet de considérer des situations plus générales que celle d'une droite (ou même d'un polynôme) passant à proximité de points.

Idéalement, la droite devrait passer par l'ensemble des points (x_i, y_i) qui devraient donc satisfaire le système linéaire suivant :

$$\left\{\begin{array}{rcl} a\,x_1 + b & = & y_1 \\ a\,x_2 + b & = & y_2 \\ & \vdots & \\ a\,x_n + b & = & y_n \end{array}\right. \quad \Leftrightarrow \quad \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{noté} \quad A\,X = B.$$

Ce système linéaire $A\,X = B$ n'admettant généralement pas de solution, on cherche un vecteur $\hat{X}$ qui minimise la *distance* du vecteur $A\,X$ au vecteur B, autrement dit qui minimise la quantité $||A\,X - B||$, ce qui revient à minimiser la quantité $||A\,X - B||^2 = \sum_{i=1}^{n} \big((a\,x_i + b) - y_i\big)^2$.
On rappelle le résultat suivant du cours. *Etant donné un système linéaire $A\,X = B$, où le nombre de lignes de la matrice A est strictement supérieur au nombre de colonnes, le vecteur $\hat{X}$ qui minimise la quantité $||A\,X - B||^2$ est solution du système linéaire carré suivant*

$$\big(A^T A\big)\,X = A^T\,B, \tag{1}$$

où A^T est la matrice transposée de A. Si les colonnes de la matrice A sont linéairement indépendantes, la solution $\hat{X}$ est unique.

Exercise 1

> *Compléter le script Python* `TP8LeastSquaresStudents.py` *afin de déterminer (par la méthode des moindres carrés) la droite de régression linéaire associée à une suite de points acquise à la souris. Déterminer (à l'aide de la formule donnée dans le cours) le coefficient de corrélation de Pearson associé à ces données. Expérimenter ce programme.*

Approximation polynomiale aux moindres carrés

On souhaite maintenant approcher une suite de n points (x_i, y_i) du plan par une fonction polynomiale $Y = F(X) = a_0 + a_1 X + a_2 X^2 + \cdots + a_p X^p$ avec $p+1 < n$. L'approche est identique. Idéalement, l'ensemble des n données (x_i, y_i) devrait satisfaire le système linéaire

$$\left\{\begin{array}{rcl} a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_p x_1^p & = & y_1 \\ a_0 + a_1 x_2 + a_2 x_2^2 + \cdots + a_p x_2^p & = & y_2 \\ & \vdots & \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_p x_n^p & = & y_n \end{array}\right. \quad \Leftrightarrow \quad \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Ce système linéaire $A\,X = B$ n'admettant généralement pas de solution, on cherchera encore à déterminer une solution aux moindres carrrés.

Exercise 2

Modifier le programme de l'exercice précédent afin de déterminer le polynôme de degré p approchant au mieux, au sens des moindres carrés, une suite de n points (x_i, y_i) acquis à la souris $(n > p+1)$. Expérimenter ce programme.

Erreur résiduelle

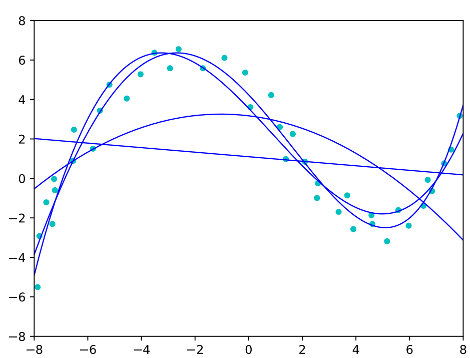
La méthode des moindres carrés consiste à minimiser la somme des erreurs au carrés définie par $||A\,X - B||^2 = \sum_{i=1}^n \big(F(x_i) - y_i\big)^2 = \sum_{i=1}^n \epsilon_i^2$, où ϵ_i est l'erreur entre la donnée y_i et le modèle $F(x_i)$. Lorsque les paramètres optimaux ont été déterminés, il reste en général une erreur appelée *erreur résiduelle* $E_{res} = \sum_{i=1}^n \epsilon_i^2$ qui permet de mesurer l'écart entre les données et le modèle. Afin d'analyser la pertinence d'un modèle, on considère ici *l'erreur résiduelle moyenne* $\left(\sum_{i=1}^n \epsilon_i^2\right)/n$ dont la racine carrée est l'*écart type*.

Exercise 3

Le script Python de l'exercice précédent permet de déterminer le meilleur polynôme de degré p approximant une suite de points (par exemple acquise à la souris) selon la méthode des moindres carrés.
Faire varier p entre 1 et une valeur p_{max} raisonable. Vérifier que l'écart type diminue lorsque le degré du modèle polynomial augmente. Cela vous semble-t'il naturel ? Que se passe-t'il si le nombre n de points est égal à $p_{max}+1$? Pouvez-vous proposer une méthode permettant de choisir un modèle pertinent (c-à-d, un degré) associé à une suite de points donnée.

```
degree  1 Residual standard deviation = 3.040348677
degree  2 Residual standard deviation = 2.57235040843
degree  3 Residual standard deviation = 0.934328391479
degree  4 Residual standard deviation = 0.76968149107
```

Dans cet exemple, les points sont entrés à la souris, puis on détermine selon le critère des moindres carrés, le meilleur polynôme d'approximation de degré 1,2,3,4,... (figure à droite), et on affiche l'écart type (ci-dessus).



Le compte rendu de ce TP consistera en un fichier pdf dont le nom sera `TP8_NOM1_NOM2.pdf`
Ce fichier pdf contiendra en entête les noms NOM1 et NOM2, puis les réponses aux questions de l'exercice 3 ci-dessus avec un exemple graphique pertinent montrant l'intérêt de choisir un polynôme de degré 4.