9. Projections et moindres carrés

Sections 4.2 et 4.3

MTH1007

J. Guérin, N. Lahrichi, S. Le Digabel
Polytechnique Montréal

H2023

(v2)

Plan

1. Projections

2. Approximations par moindres carrés

1. **Projections**

2. Approximations par moindres carrés

Projection sur une droite (1/2)

Soit L le sous-espace vectoriel de $\mathbb{R}^m$ correspondant à la droite engendrée par le vecteur non nul $\mathbf{a} \in \mathbb{R}^m$

- La *projection orthogonale* du vecteur $\mathbf{b} \in \mathbb{R}^m$ sur L est le vecteur $\mathbf{p} \in L$ le plus proche de $\mathbf{b}$
- La projection de $\mathbf{b}$ sur L est $\mathbf{p} = \hat{x}\mathbf{a} = \mathbf{a}\hat{x}$ où $\hat{x} = \frac{\mathbf{a}^\top \mathbf{b}}{\mathbf{a}^\top \mathbf{a}}$
- On peut le voir aussi comme $\mathbf{p} = (\mathbf{u}^\top \mathbf{b})\mathbf{u}$ avec
  - $\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ la *normalisation* de $\mathbf{a}$
  - $\mathbf{a} = \|\mathbf{a}\|\mathbf{u}$ et $\|\mathbf{u}\| = \mathbf{u}^\top \mathbf{u} = 1$ ($\mathbf{u}$ est *unitaire*)
- La droite allant de $\mathbf{p}$ à $\mathbf{b}$ est orthogonale à $\mathbf{a}$: $\mathbf{a} \perp \mathbf{e}$ avec $\mathbf{e} = \mathbf{b} - \mathbf{p}$ ($\mathbf{e}$ est appelé l'*erreur*)
- **Exemple :** Avec $\mathbf{a} = (3, 2)$ et $\mathbf{b} = (5, 5)$, on a $\mathbf{p} = \frac{1}{13}(75, 50)$

Projection sur une droite (2/2)

- Matriciellement, le résultat précédent se reformule comme suit :

  La projection de $\mathbf{b} \in \mathbb{R}^m$ sur L est $\mathbf{p} = P\mathbf{b}$, où $P \in \mathbb{R}^{m \times m}$ est la *matrice de projection* symétrique suivante :

$$P = \frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top \mathbf{a}} = \mathbf{u}\mathbf{u}^\top \text{ avec } \mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$$

$$\text{car } \mathbf{p} = \frac{\mathbf{a}^\top \mathbf{b}}{\mathbf{a}^\top \mathbf{a}}\mathbf{a} = \mathbf{a}\frac{\mathbf{a}^\top \mathbf{b}}{\mathbf{a}^\top \mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top \mathbf{a}}\mathbf{b} = P\mathbf{b}$$

- **Exemple :**

  Avec $\mathbf{a} = (3, 2)$ et $\mathbf{b} = (5, 5)$, on a $P = \frac{1}{13}\begin{bmatrix} 9 & 6 \\ 6 & 4 \end{bmatrix}$

Remarques

- P symétrique, $r(P) = 1$, $C(P) = C([\mathbf{a}]) = L$
- Si $\mathbf{b}$ est orthogonal à $\mathbf{a}$, alors $\mathbf{a}^\top \mathbf{b} = 0$. La projection est $\boldsymbol{p} = \mathbf{0}$
- Si $\mathbf{b} = \mathbf{a}$, alors $\hat{x} = 1$. La projection de $\mathbf{a}$ sur lui-même donne $\mathbf{a}$: $P\mathbf{a} = \mathbf{a}$ (mais $P \neq I$)
- Si $\mathbf{b} \in L$, alors $\mathbf{p} = \mathbf{b}$ et $\mathbf{e} = \mathbf{0}$
- Projeter une deuxième fois ne change rien :

$$P^2 = PP = (\mathbf{u}\mathbf{u}^\top)(\mathbf{u}\mathbf{u}^\top) = \mathbf{u}(\mathbf{u}^\top\mathbf{u})\mathbf{u}^\top = \mathbf{u}\mathbf{u}^\top = P$$

- Lorsque P projette sur un sous-espace ($C(P)$), $I - P$ projette sur le sous-espace orthogonal ($N(P)$)
- **Exemple :**

  Illustrer avec $\mathbf{a} = (3, 2)$, $\mathbf{b} = (5, 5)$, et $P = \frac{1}{13}\begin{bmatrix} 9 & 6 \\ 6 & 4 \end{bmatrix}$

Projection dans un sous-espace

Soit $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ n vecteurs linéairement indépendants avec $n < m$ et W le sous-espace de $\mathbb{R}^m$ de dimension n engendré par ces vecteurs

- La *projection orthogonale* du vecteur $\mathbf{b} \in \mathbb{R}^m$ sur W est le vecteur $\mathbf{p} \in W$ le plus proche de $\mathbf{b}$
- La projection de $\mathbf{b}$ sur W est $\mathbf{p} = A\hat{\mathbf{x}} \in \mathbb{R}^m$ où
$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$
et
$$\hat{\mathbf{x}} = \left(A^\top A\right)^{-1} A^\top \mathbf{b} \in \mathbb{R}^n$$
- $\mathbf{p} = P\mathbf{b}$ où P est la *matrice de projection*
$$P = A\left(A^\top A\right)^{-1} A^\top \in \mathbb{R}^{m \times m}$$

Remarques (1/2)

- Le SÉL $A\mathbf{x} = \mathbf{b}$ n'a pas toujours de solution $(m > n)$
- $C(A^\top) = \mathbb{R}^n$ et $N(A) = \{\mathbf{0}\}$
- La matrice $A^\top A$ est symétrique et de taille $n \times n$. Elle est inversible car $r(A) = r(A^\top A)$ (preuve en exercice) et $r(A) = n$
- Intuition pour la formule de $\hat{\mathbf{x}}$: la droite allant de $\mathbf{p}$ à $\mathbf{b}$ est orthogonale au sous-espace $W = C(A)$:

$$A^\top(\mathbf{b} - \mathbf{p}) = A^\top \mathbf{e} = \mathbf{0} = A^\top(\mathbf{b} - A\hat{\mathbf{x}}) \Rightarrow A^\top A\hat{\mathbf{x}} = A^\top \mathbf{b}$$

$$\Rightarrow \hat{\mathbf{x}} = (A^\top A)^{-1} A^\top \mathbf{b}$$

- Pour trouver la projection $\mathbf{p}$, il faut résoudre le système $A^\top A\hat{\mathbf{x}} = A^\top \mathbf{b}$, qui possède une solution unique

Remarques (2/2)

- Les matrices de projection $P \in \mathbb{R}^{m \times m}$ vérifient : $P^2 = P$, $P^\top = P$, $r(P) = n < m$ (P est singulière)
- $W = C(A) = C(P) = C(P^\top)$: La projection sur W peut être vue comme la multiplication des vecteurs de $\mathbb{R}^n$ par A ($A\mathbf{x} \in W$ pour tout $\mathbf{x} \in \mathbb{R}^n$) ou bien comme la multiplication des vecteurs de $\mathbb{R}^m$ par P ($P\mathbf{b} \in W$ pour tout $\mathbf{b} \in \mathbb{R}^m$)
- $N(P) = N(P^\top) = N(A^\top)$
- $\mathbf{e} = \mathbf{b} - \mathbf{p}$ est orthogonal aux vecteurs de $W = C(A)$. Il est donc dans les noyaux à gauche de A et P : $A^\top \mathbf{e} = \mathbf{0}$ et $P^\top \mathbf{e} = \mathbf{0}$
- Décomposition de $\mathbb{R}^m$: $\mathbf{b} = \underbrace{\mathbf{p}}_{\in C(A)} + \underbrace{\mathbf{e}}_{\in C(A)^\perp = N(A^\top)} \in \mathbb{R}^m$

La projection sur une droite est un cas particulier

Si $n = 1$, on retrouve les formules de la projection sur une droite dans $\mathbb{R}^m$:

$$
\begin{aligned}
A \in \mathbb{R}^{m \times n} \quad &\rightarrow \quad \mathbf{a} \in \mathbb{R}^m \\
\mathbf{p} = P\mathbf{b} = A\hat{\mathbf{x}} \in \mathbb{R}^m \quad &\rightarrow \quad \mathbf{p} = P\mathbf{b} = \hat{x}\mathbf{a} \in \mathbb{R}^m \\
\hat{\mathbf{x}} = \left(A^\top A\right)^{-1} A^\top \mathbf{b} \in \mathbb{R}^n \quad &\rightarrow \quad \hat{x} = \frac{\mathbf{a}^\top \mathbf{b}}{\mathbf{a}^\top \mathbf{a}} \in \mathbb{R} \\
P = A\left(A^\top A\right)^{-1} A^\top \in \mathbb{R}^{m \times m} \quad &\rightarrow \quad P = \frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top \mathbf{a}} \in \mathbb{R}^{m \times m}
\end{aligned}
$$

1. Projections

2. Approximations par moindres carrés

Introduction

- $A \in \mathbb{R}^{m \times n}$ avec $m > n$ et $r(A) = n$
- Le SÉL $A\mathbf{x} = \mathbf{b}$ est de plein rang colonne et ne possède pas toujours de solution
- Au lieu on résout le système $A^\top A\hat{\mathbf{x}} = A^\top \mathbf{b}$ qui possède toujours une solution (ce sont les *équations normales*)
- $\mathbf{p} = A\hat{\mathbf{x}}$ est la projection de $\mathbf{b}$ dans $C(A)$: C'est donc le point de $C(A)$ le plus proche de $\mathbf{b}$
- $\hat{\mathbf{x}}$ minimise l'erreur

$$\|\mathbf{e}\|^2 = \|\mathbf{b} - \mathbf{p}\|^2 = \|\mathbf{b} - P\mathbf{b}\|^2 = \|\mathbf{b} - A\hat{\mathbf{x}}\|^2$$

- On peut donc voir $\hat{\mathbf{x}}$ comme la *meilleure "solution" possible* à $A\mathbf{x} = \mathbf{b}$
- $\hat{\mathbf{x}}$ est appelée la solution de $A\mathbf{x} = \mathbf{b}$ au sens des *moindres carrés*

Application : Droite d'ajustement (1/3)

- Étant donnés $m > 2$ *points de données* $(t_1, b_1), (t_2, b_2), \ldots, (t_m, b_m)$ de $\mathbb{R}^2$, on essaie de trouver l'équation d'une droite qui passe par les m points
- Cette équation est $b(t) = c + dt$ avec $c, d \in \mathbb{R}$
- c et d devraient être les solutions du système $A\mathbf{x} = \mathbf{b}$ avec

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} c \\ d \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Droite d'ajustement (2/3)

- Or le système $A\mathbf{x} = \mathbf{b}$ n'a a priori pas de solution car il est peu probable que les m points soient alignés
- Au lieu, on résout
$$A^\top A\,\hat{\mathbf{x}} = A^\top \mathbf{b}$$
- La solution $\hat{\mathbf{x}} = (c, d)$ donnera la *droite d'ajustement* ou de *régression* qui minimise la somme des erreurs verticales avec les points de données : $\hat{\mathbf{x}}$ est tel que $E(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2$ est le plus petit possible :
$$\|\mathbf{e}\|^2 = E(\hat{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^2} E(\mathbf{x})$$

Droite d'ajustement (3/3)

Explicitement, le système d'équations 2×2 qui définit la droite d'ajustement est

$$\underbrace{\begin{bmatrix} m & \displaystyle\sum_{i=1}^{m} t_i \\ \displaystyle\sum_{i=1}^{m} t_i & \displaystyle\sum_{i=1}^{m} t_i^2 \end{bmatrix}}_{A^\top A} \begin{bmatrix} c \\ d \end{bmatrix} = \underbrace{\begin{bmatrix} \displaystyle\sum_{i=1}^{m} b_i \\ \displaystyle\sum_{i=1}^{m} t_i b_i \end{bmatrix}}_{A^\top b}$$

Point de vue géométrique

- On cherche le vecteur de $C(A)$ qui est le plus proche de $\mathbf{b}$
- Il s'agit du vecteur $\mathbf{p}$ qui minimise (le carré de) la distance

$$\|\mathbf{e}\|^2 = \|\mathbf{p} - \mathbf{b}\|^2$$

- Cette distance représente la somme des carrés des erreurs verticales entre les points et la droite

Point de vue algébrique

- Chaque vecteur $\mathbf{b} \in \mathbb{R}^m$ se décompose en $\mathbf{b} = \mathbf{p} + \mathbf{e}$ où $\mathbf{p} \in C(A)$ et $\mathbf{e} \in C(A)^\perp = N(A^\top)$
- Le système $A\,\mathbf{x} = \mathbf{p} + \mathbf{e}$ n'a pas de solution mais $A\,\hat{\mathbf{x}} = \mathbf{p}$ possède une solution
- $A\mathbf{x} \in \mathbb{R}^m$ et $\mathbf{p} \in \mathbb{R}^m$ sont dans $C(A)$, donc $A\mathbf{x} - \mathbf{p} \in C(A)$. Ainsi $(A\mathbf{x} - \mathbf{p}) \perp \mathbf{e}$
- Cette solution minimise l'erreur

$$E(\mathbf{x}) = \|A\,\mathbf{x} - \mathbf{b}\|^2 = \|A\,\mathbf{x} - \mathbf{p}\|^2 + \|\mathbf{e}\|^2$$

et

$$E(\hat{\mathbf{x}}) = \|A\,\hat{\mathbf{x}} - \mathbf{b}\|^2 = \|\mathbf{e}\|^2$$

Point de vue de l'optimisation

- On minimise la fonction d'erreur

$$E(\mathbf{x}) = \|A\,\mathbf{x} - \mathbf{b}\|^2$$

- Pour cela, on résout $\nabla E(\mathbf{x}) = \mathbf{0}$ (donne $\hat{\mathbf{x}}$) et on montre que $\nabla^2 E(\hat{\mathbf{x}})$ est *définie-positive*

- Ce qui revient à

$$A^\top A\,\hat{\mathbf{x}} = A^\top \mathbf{b}$$

Solution analytique

- Il n'est pas nécessaire de résoudre explicitement $A^\top A\,\hat{\mathbf{x}} = A^\top \mathbf{b}$, car on peut trouver une solution analytique

- Par exemple, par l'optimisation, on obtient

$$\begin{cases} d = \dfrac{\sum_{i=1}^m t_i b_i - m\ \hat{t}\ \hat{b}}{\sum_{i=1}^m t_i^2 - m\hat{t}^2} = \frac{\sum_{i=1}^m (t_i - \hat{t})(b_i - \hat{b})}{\sum_{i=1}^m (t_i - \hat{t})^2} \\ c = \hat{b} - d\hat{t} \end{cases}$$

avec $\hat{b} = \frac{1}{m}\sum_{i=1}^m b_i$ et $\hat{t} = \frac{1}{m}\sum_{i=1}^m t_i$ (*moyennes* des b_i et des t_i)

Généralisation

Le même principe s'applique à l'ajustement d'une courbe de forme connue à un ensemble de points donnés

Par exemple : Ajuster une parabole sur les points $(0,6),(1,0),(2,0)$ et $(1/2,1)$: