Méthodes aux moindres carrés,

des données aux modèles

Part 1. **Moindres carrés linéaires**

1. Contenu

Le **plan** de cette session est le suivant:

- Des données à représenter / modéliser,
- Solution au sens des moindres carrés,
- Equations normales,
- Exemples,
- Interprétation géométrique.

Pour information, la **partie II** abordera:

- Les moindres carrés **non** linéaires et le choix des fonctions de base,
- La résolution numérique: l'algorithme de Gauss-Newton et la méthode de Householder.

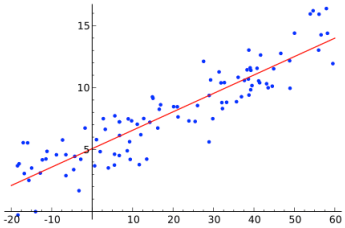
Enfin une **partie III** abordera le lien entre la solution aux moindres carrés et la SVD.

NB. Les parties II et II ne seront pour l'instant pas présentées.

2. Des données à représenter / modéliser

Etant données m observations / mesures / données $(d_k)_{1\leq k\leq m}$, nous cherchons à représenter ces données par un *modèle* mathématique.

L'exemple type est le *modèle dit de régression linéaire* qui consiste à représenter *au mieux* un ensemble de points par une droite, c'est à dire le modèle le plus simple qui soit : une équation scalaire, linéaire (une droite !).



Etant données les m observations / mesures / données $(d_k)_{1\leq k\leq m}$, nous cherchons un modèle l*inéaire à* n paramètres $(x_i)_{1\leq i\leq n}$ (et donc pas forcément à 2 paramètres comme dans l'exemple de la droite).

Ces n paramètres sont les inconnues de notre modèle. Le modèle s'écrit alors comme un système de m équations linéaires à n inconnues:

$$\boxed{\left\{\begin{array}{rcl} a_{11}x_1 + ... + a_{1n}x_n & = & d_1 \\ ... & = & ... \\ a_{m1}x_1 + ... + a_{mn}x_n & = & d_m \end{array}\right.}$$

On note A la matrice *rectangulaire* $m \times n$ correspondante, par $d \in \mathbb{R}^m$ le vecteur des observations et par $x \in \mathbb{R}^n$ le vecteur des inconnues.

Le problème s'écrit alors comme suit.

Etant donné $d \in \mathbb{R}^m$, *trouver* $x \in \mathbb{R}^n$ *tel que :* $Ax = d$.

Ce problème peut être vu comme un problème de *calibration de modèle*, à savoir on cherche à calibrer le modèle sur les données disponibles.

3. Des données à représenter / modéliser (suite)

On a vu que l'on cherche à résoudre le système suivant:

$$\boxed{Ax = d}$$

avec: A matrice *rectangulaire* $m \times n$, $d \in \mathbb{R}^m$ donné et l'inconnue $x \in \mathbb{R}^n$.

Dans le cas (très improbable...) où l'on ait autant de paramètres x_i que d'observations d_k, i.e. $n = m$, le problème est *bien posé* si et seulement si A est de rang maximal, i.e. $r = rank(A) = dim(Im(A)) = n$.

Dans ce cas, il existe un *unique* jeu de paramètres x décrivant *exactement* les données (on les interpole).

Toujours dans le cas $n = m$ mais si le modèle est *mal posé* dans le sens $r = rank(A) = dim(Im(A)) < n$, alors il existe *des* solutions, qui ne sont pas uniques. C'est à dire que le noyau de $Ker(A)$ contient des vecteurs non nuls $z \in \mathbb{R}^n$ tels que : $Az = 0$ dans $\mathbb{R}^m$.

Dans le cas $n > m$, c'est à dire qu'il y a plus de paramètres que de données à représenter, alors le système est dit *sous-déterminé*. Il existe alors une infinité de solutions.

Dans le cas $n < m$, c'est à dire qu'il y a moins de paramètres que de données, alors le système est dit *sur-déterminé*. C'est le cas qui nous intéresse tout particulièrement ici.

Généralement le système n'admet *pas* de solution.

(Pensez au fait qu'il peut y avoir une unique solution avec les m premières équations, mais cette unique solution ne va pas forcément satisfaire les $(n - m)$ équations suivantes !).

Exercise. *Illustrer succintement les cas évoqués, en faisant des figures dans* $\mathbb{R}^3$.

1) Etant donné un ensemble de $m = 10$ *points de* $\mathbb{R}^3$*, et un modèle linéaire à* $n = 3$ *paramètres: cas sous-déterminé.*

2) Etant donné un ensemble de $m = 2$ *points de* $\mathbb{R}^3$*, et un modèle linéaire à* $n = 3$ *paramètres: cas sur-déterminé.*

3) Etant donné un ensemble de 3 *points de* $\mathbb{R}^3$*, et un modèle linéaire à* $n = 3$ *paramètres.*

4. Des données à représenter / modéliser: un exemple

Le défenseur d'une région a pour mission d'intercepter d'éventuels missiles à venir. Il dispose de radars relevant les positions des missiles agresseurs, et son objectif est de prévoir leurs trajectoires pour pouvoir les intercepter. Ses radars ont relevé les 5 positions suivantes d'un des missiles passé:

x_i	0	250	500	750	1000
y_i	0	8	15	19	20

Les x_i désignent la distance au sol sur une trajectoire directe, les y_i l'altitude du missile (en km).

Bien renseigné, il sait que la trajectoire des missiles est parabolique.

Elle satisfait donc une équation de la forme: $y(x) = a_0 + a_1\, x + a_2\, x^2$.

Le défenseur obtient alors un système d'équations dont les inconnues sont $a_0,\ a_1$ et a_2.

Après changement d'unité, les équations du modèle s'écrivent comme suit:

$$\begin{cases} a_0 & & & & & = & 0, \\ a_0 & + & 2.5a_1 & + & 6.25\, a_2 & = & 0.08, \\ a_0 & + & 5\, a_1 & + & 25\, a_2 & = & 0.15, \\ a_0 & + & 7.5\, a_1 & + & 56.25\, a_2 & = & 0.19, \\ a_0 & + & 10\, a_1 & + & 100\, a_2 & = & 0.20. \end{cases}$$

Soit un système linéaire sur-déterminé de $m = 5$ équations à $n = 3$ inconnues.

5. Solution au sens des moindres carrés

Au lieu de chercher une solution qui satisfasse chacune des équations du modèle (i.e. une ou des solutions hypothétique(s) interpolant exactement chacune des observations), il est plus judicieux de chercher une solution qui représente *au mieux* les observations.

Plus précisément, une solution x *minimisant la norme* $\|Ax - d\|$.

Un bon choix de norme est alors la norme Euclidienne $\|.\|_2$ car *associée à un produit scalaire* (contrairement aux normes $\|.\|_1$ ou $\|.\|_\infty$ par exemple).

Le problème devient alors de

Trouver $x^* \in \mathbb{R}^n$ tel que:

$$j(x^*) = \min_{x \in \mathbb{R}^n} j(x) \text{ avec } j(x) = \|Ax - d\|^2_{2,m}$$

Il s'agit d'un *problème d'optimisation*, dans l'espace entier $\mathbb{R}^n$.

La solution de ce problème est appelée *solution au sens des Moindres Carrés (MC).*

Un calcul simple montre immédiatement que:

$$j(x) = < A^T Ax, x > -2 < Ax, d > + < d, d >$$

où $< ., . >$ désigne le produit scalaire Euclidien. La fonctionnelle j est définie de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$.
La matrice $A^T A$ est carrée; et comme cette matrice est positive, j est quadratique donc convexe.
j admet donc un minimum (au moins) dans $\mathbb{R}^n$.
Si de plus la matrice $A^T A$ est définie alors la solution est unique.

Remark. Rappelons qu'une matrice carrée M positive est également définie si : $< Mx, x >= 0$ équivaut à $x = 0$.

Par ailleurs, on peut aisément montrer que $A^T A$ est définie si et seulement A est de rang maximal.

6. Solution au sens des moindres carrés: équations normales

Rappelons qu'une condition nécessaire d'optimum local de j est:

$$\nabla j(x) = 0$$

Ensuite, la condition suffisante de minimum (local) est que la matrice Hessienne de j, notée $H_j(x)$ ou encore $D^2 j(x)$, est *positive définie.*
(Rappelons qu'une matrice Hessienne est toujours symétrique).

Calculons le gradient et la Hessienne de j.

Lemma. *La fonctionnelle j définie précédemment est deux fois continuement différentiable (i.e. de classe C^2).*
Le gradient de j, vecteur de $\mathbb{R}^n$, s'écrit: $\nabla j(x) = A^T Ax - A^T d$
La matrice Hessienne de j, matrice carrée $n \times n$, s'écrit: $H_j(x) = A^T A$

Proof. Une manière de montrer ce lemme (bien que pas la plus directe) est de calculer le développement limité de Taylor à l'ordre 2 de j au voisinage d'un point x.
Soient x et h deux vecteurs de$\mathbb{R}^n$, on écrit formellement:

$$\begin{aligned} j(x+h) &= \langle A\,(x+h) - d,\ A\,(x+h) - d \rangle\,, \\ &= \langle (A\,x - d) + A\,h,\ (A\,x - d) + A\,h \rangle\,, \\ &= j(x) + \langle A\,h,\ A\,x - d \rangle + \langle A\,x - d,\ A\,h \rangle + \langle A\,h,\ A\,h \rangle, \\ &= j(x) + 2\,\langle A^T\,(A\,x - d),\, h \rangle + \langle A^T\,Ah,\, h \rangle. \end{aligned}$$

On reconnait alors le développement à l'ordre 2 d'une fonction f de classe C^2,

$$f(x+h) = f(x) + \langle \nabla f(x),\, h \rangle + \frac{1}{2}\,\langle H_f(x)\,h,\, h \rangle + \|h\|_2^2\,\varepsilon(h)$$

On en déduit alors les expressions de $\nabla j(x)$ et $H_j(x)$. □

7. Solution au sens des M.C.: équations normales (suite)

On obtient finalement le résultat central suivant.

Theorem. *Soit le problème aux moindres carrés dans le cas sur-déterminé, i.e. $n < m$.*
Toute solution du problème est également solution du système linéaire $n \times n$ suivant:

$$\boxed{A^T Ax = A^T d}$$

Ces équations sont appelées les *équations normales.*
De plus, la solution est unique si et seulement si A est de rang maximal i.e. $r = rang(A) = n$.

Proof. La condition nécéssaire d'optimum local $\nabla j(x) = 0$ donne directement les équations normales.
La matrice $A^T\,A$ est symétrique positive, cet extremum est alors un minimum global.
De plus ce minimum global est unique si $A^T\,A$ est définie, soit A de rang maximal.

□

8. **Retour à l'exemple**

Revenons à notre exemple de trajectoire parabolique de missiles. Les équations normales s'écrivent ainsi:

$$\begin{bmatrix} 5 & 25 & 187.5 \\ 25.00 & 187.5 & 1562.5 \\ 187.50 & 1562.5 & 13828.125 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.62 \\ 4.375 \\ 34.9375 \end{bmatrix}$$

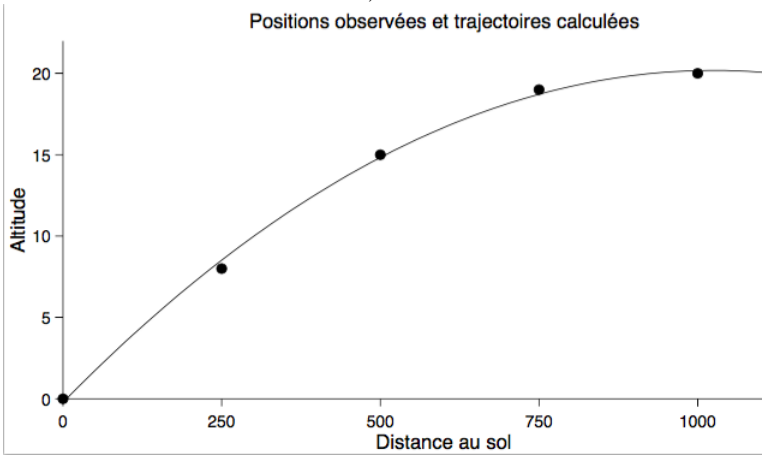
Il s'agit d'un système de 3 équations à 3 inconnues que l'on peut résoudre très facilement par la méthode d'élimination de Gauss.

Les valeurs de la solution sont approximativement les suivantes:

$$x^* \simeq \begin{bmatrix} -0.0023 \\ 0.0398 \\ -0.0019 \end{bmatrix}$$

En traçant la trajectoire obtenue à partir de cette valeur optimale de x^*, on constate que la parabole approche bien les mesures faites par le radar

(ce qui tout à fait normal puisque construite ainsi, à savoir il s'agit de la trajectoire minimisant la distance aux données en norme euclidienne).



9. **Exemple: changeons de modèle pour voir...**

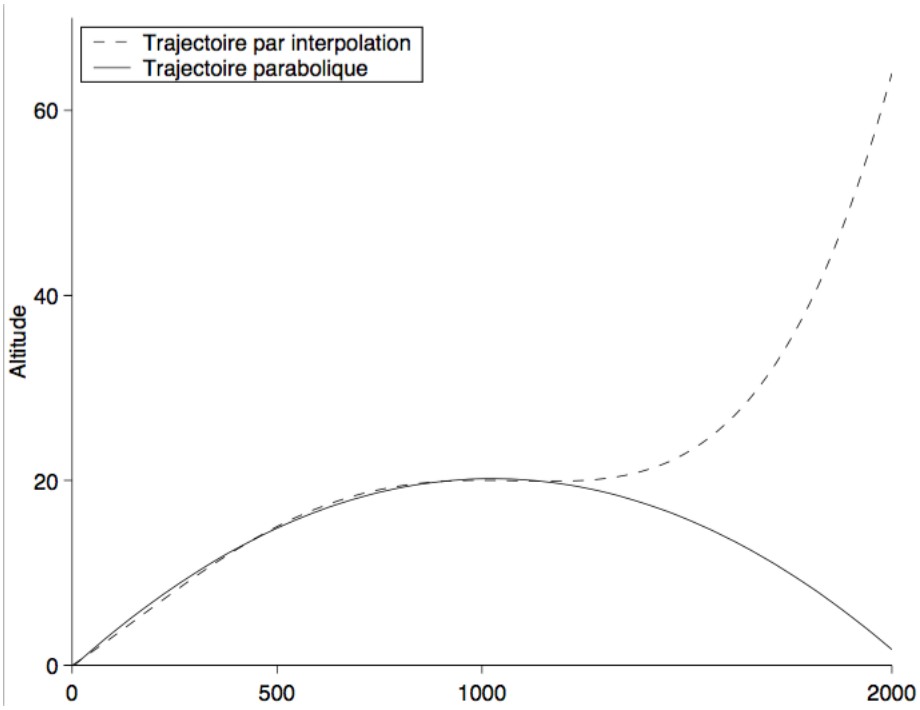
A présent changeons de *modèle a-priori*; à savoir supposons que le défenseur *ne savait pas* que la trajectoire était parabolique.

Au vu des 5 données dont il disposait, il aurait alors pu "naturellement" chercher à représenter *exactement* chacune d'entre elles en calculant l'unique polynôme de degré 4, $p_4(x)$, passant par ces 5 points (i.e. effectuer une interpolation de Lagrange).

Ensuite, pour prévoir la trajectoire du missile, il aurait pu extrapoler la position à l'aide du polynôme d'interpolation, soit hors de son intervalle de définition (ce qu'il ne faut a-priori pas faire...).

Regardons ce que cela aurait donné en traçant la trajectoire décrite par ce polynôme p_4 sur une distance totale de 2000 km,

i.e. sur l'intervalle de définition [0,1000] mais aussi sur [1000,2000] intervalle d'extrapolation.



Prévision trajectoire [1000-2000] Moindres carrés vs interpolation : le missile s'échappe vers le ciel... Fort heureusement que le défenseur connaissait a-priori le bon modèle à calibrer...

Retenons donc que ce n'est pas parce qu'un modèle est bien calibré (au sens des moindres carrés) que le modèle sera nécessairement bon pour effectuer une prédiction... Par contre si le modèle mathématique choisi a-priori est le bon, alors on peut a-priori extrapoler pour prédire...

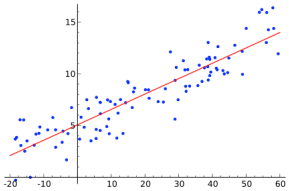
10. **Exemples historique & classique de solution aux moindres carrés**

La méthode des moindres carrés, c'est génial et... pas nouveau ! Historiquement, J.C. Gauss (mathématicien allemand, 1777-1855) et A.-M. Legendre (mathématicien français, 1752-1833) ont été les premiers à définir et calculer des solutions au sens des moindres carrés.

J.C. Gauss, âgé de 24 ans, a réussi l'exploit de prévoir de manière précise la trajectoire de l'asteroïde Céres sur la base de quelques observations (et sur l'hypothèse à l'époque originale d'une trajectoire elliptique -et non circulaire-).

L'exemple aussi basique qu'utile: modèle de régression lineaire. En statistiques, un modèle de régression linéaire consiste à représenter un ensemble de variables (les données) par une loi linéaire.

Autrement dit, dans le cas le plus simple, faire passer une droite "au mieux" dans un "nuage" de points.



Sur la base de m points (x_i, y_i) de $\mathbb{R}^2$ à représenter au mieux, la matrice A s'écrit:

$$\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}$$

sachant que la base de fonctions dans laquelle nous recherchons une solution aux MC est l'ensemble: $\{1, x\}$.

Un simple calcul donne les équations normales $A^T Aa = A^T d$, qui s'écrivent comme suit:

$$\begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}$$

C'est un système 2×2 immédiat à résoudre et qui donne les coefficients de la droite approchant au mieux (au sens de la norme euclidienne) le "nuage" de points.

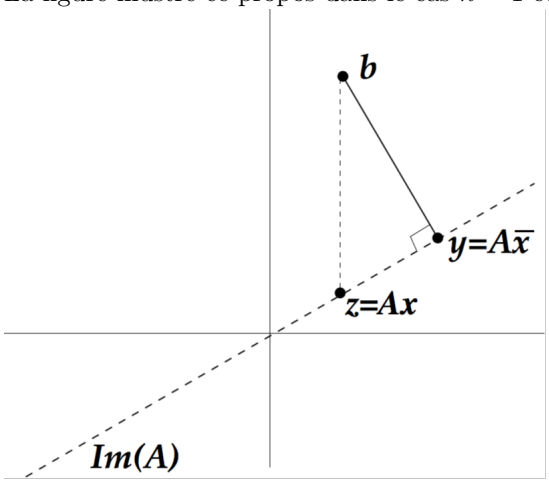
11. **Interprétation géométrique**

Etablissons d'une manière différente les équations normales tout en interprétant géométriquement la solution.

La solution aux Moindres Carrés consiste à minimiser la distance euclidienne entre les données, vecteur d de $\mathbb{R}^m$, et $Im(A)$ sous-espace vectoriel de $\mathbb{R}^m$.

La solution optimale $x^* \in \mathbb{R}^n$ est alors tel que Ax^* est le projeté orthogonal de d, $d \in \mathbb{R}^m$, sur $Im(A)$.

La figure illustre ce propos dans le cas $n = 1$ et $m = 2$.



Montrons donc ce résultat à savoir: "$\|Ax - d\|_{2,m}$ minimal si et seulement si $(Ax - d) \perp Im(A)$ ".

Cela est équivalent d'écrire que: "le résidu $r = (Ax - d) \in Im(A)^T$".

Notons x^* la solution optimale. On a:

$$\forall x \in \mathbb{R}^n,\ \|Ax - d\|_{2,m}^2 = \|Ax^* - d\|_{2,m}^2 + \|Ax - Ax^*\|_{2,m}^2 \geq \|Ax^* - d\|_{2,m}^2$$

Par ailleurs, on a:

$$\begin{array}{rcl} r \in \mathrm{Im}(A)^\perp & \Leftrightarrow & \forall y \in \mathrm{Im}(A),\ r^T y = 0, \\ & \Leftrightarrow & \forall z \in \mathbb{R}^n,\ r^T A z = 0, \\ & \Leftrightarrow & \forall z \in \mathbb{R}^n,\ z^T A^T r = 0, \\ & \Leftrightarrow & A^T r = 0, \\ & \Leftrightarrow & r \in \mathrm{Ker}(A^T) \end{array}$$

Ce qui nous conduit donc aux équations normales: $A^T(Ax^* - b) = 0$.