Mathématiques Numériques 1A
MN2

Bruno Pinçon /J.-F. Scheid

Télécom Nancy

2025-2026

Modules MN

☛ 2 modules de 7 CM (1h) + 7 TD + 1 examen chacun.

- **MN1**
  1. Arithmétique flottante : représentation des nombres et calculs sur ordinateur.
  2. Factorisations de matrice (LU, Cholesky, QR, SVD).
  3. Normes vectorielles et matricielles ; conditionnement de systèmes (linéaires).
  4. Classification non-supervisée : k-means.

- **MN2**
  1. Approximation par moindres carrés ; rappels et compléments de calcul différentiel.
  2. Optimisation non-linéaire, applications (moindres carrés non-linéaires, systèmes de recommandations,....)
  3. Introduction aux réseaux de neurones (perceptron multi-couches, rétropropagation du gradient)

Mathématiques Numériques 1A
MN2

Chapitre 1 - Moindres carrés

Bruno Pinçon /J.-F. Scheid

Télécom Nancy

2025-2026

version du 10/5/2026

I. Introduction

Un problème, appelé *identification ou estimation de paramètres*, que l'on rencontre assez souvent en pratique est le suivant :

Problèmatique.

On a un modèle $y = f(t, x)$ décrivant une relation fonctionnelle entre :

- une variable d'entrée t
- une variable de sortie y

(pour le moment on considèrera t et y comme des scalaires réels mais ils peuvent être des vecteurs).

Ce modèle dépend de n **paramètres inconnus** $x_1, \ldots, x_n$ que l'on range dans le vecteur x i.e. $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$.

Le modèle est :

- soit issu (plus ou moins) de lois physiques : on sait que le phénomène décrit doit effectivement suivre la relation $y = f(t, x)$, mais on ne connaît pas certains coefficients (qui forment le vecteur x) ;
- soit n'est pas issu de lois physiques mais construit pour obtenir une relation "entrée-sortie" que l'on pourra évaluer sur d'autres entrées (une fois les paramètres inconnus x "identifiés").

  Exemple : les réseaux de neurones (l'identification des paramètres inconnus s'appelle alors "phase d'apprentissage").

Identification des paramètres inconnus.

Pour **estimer/identifier les paramètres**, on dispose de m mesures $(t_i, y_i), i = 1, \ldots, m$ avec $m \geq n$, où le plus souvent les t_i sont exacts (ou suffisamment précis pour être considérés comme exacts) mais les mesures y_i sont entachées d'erreur.

On cherche alors les paramètres x de sorte que :

$$\forall i \in [\![1, m]\!], \quad f(t_i, x) \simeq y_i, \text{ soit } \boxed{e_i(x) := f(t_i, x) - y_i \simeq 0}.$$

☛ Si on note $e(x)$ le vecteur rassemblant les m erreurs $e_i(x)$:

$$\boxed{e(x) = (e_1(x), \cdots, e_m(x))^\top}$$

alors il est naturel de chercher le vecteur x qui va rendre $\|e(x)\|$ *minimale*.

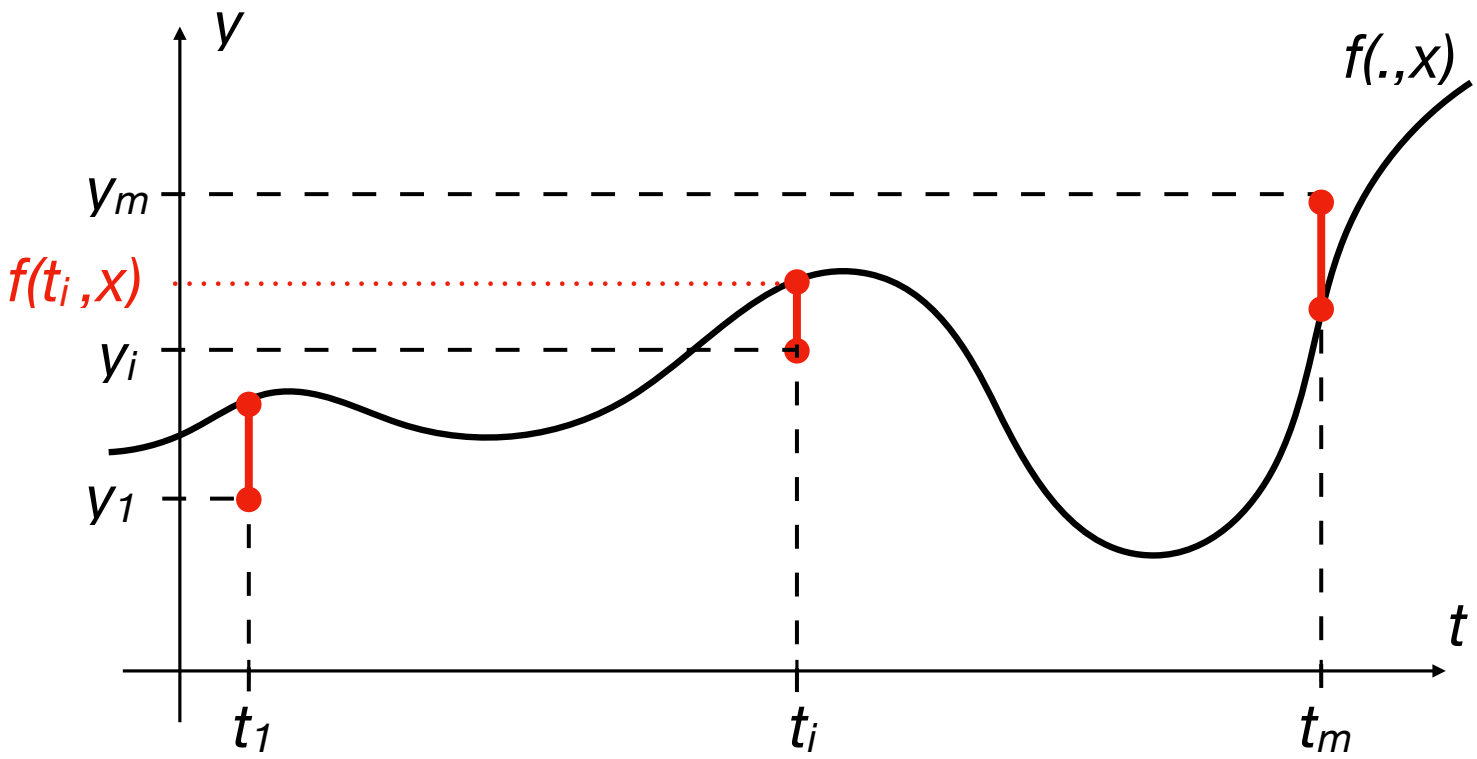
☛ Le choix de la norme euclidienne $\|.\| := \|.\|_2$ conduit précisément à la *méthode des moindres carrés* dans laquelle on minimise *le carré* de $\|e(x)\|_2$, ce qui est équivalent.

Le problème des moindres carrés s'énonce donc ainsi :

Formulation des moindres carrés.

Etant données les mesures $(t_i, y_i)_{1 \leq i \leq m}$, trouver $x \in \mathbb{R}^n$ qui minimise :

$$E(x) = \|e(x)\|_2^2 = \sum_{i=1}^{m} e_i(x)^2 = \sum_{i=1}^{m} (f(t_i, x) - y_i)^2 \qquad (1)$$



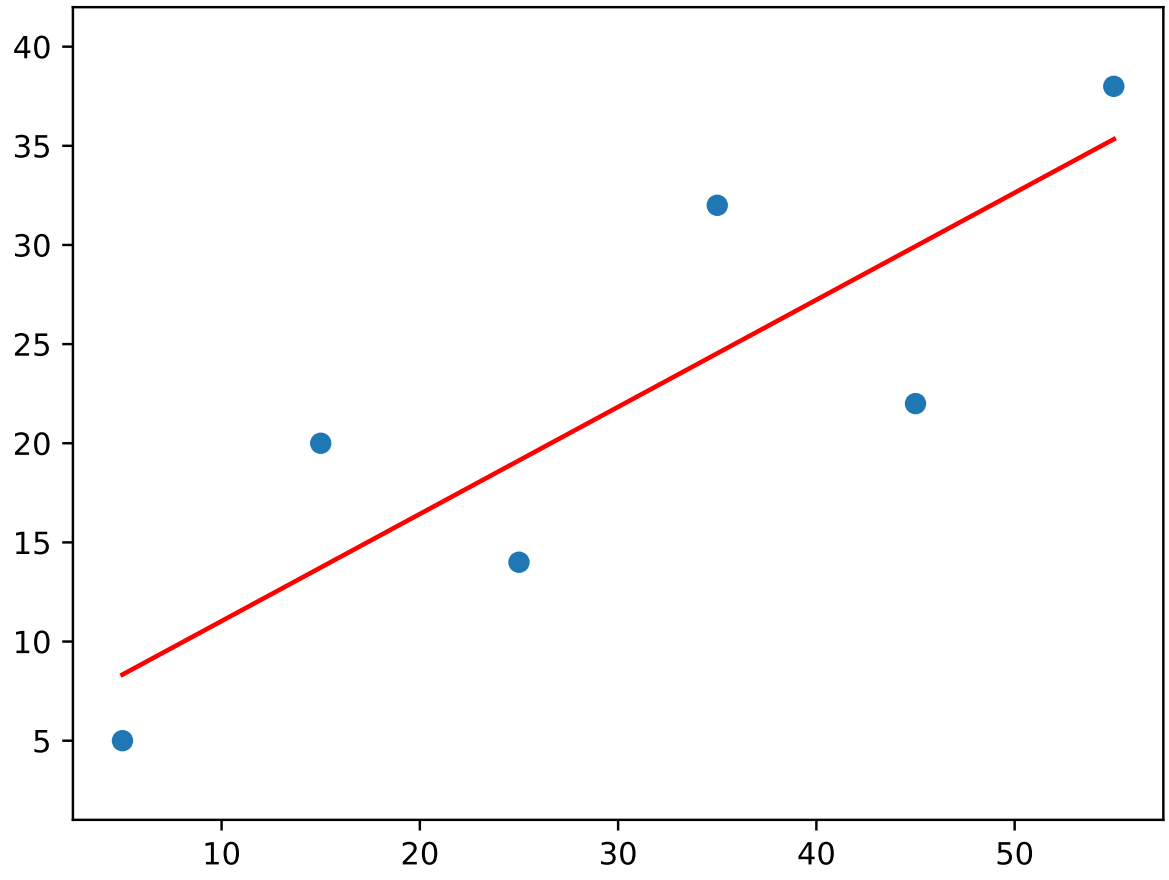
Remarque. La norme euclidienne a beaucoup d'avantages liés à l'interprétation statistique ainsi qu'à la facilité à résoudre le problème mais d'autres choix sont possibles ; par exemple si on veut minimiser le plus grand écart entre le modèle et les mesures il faut utiliser $\|.\|_\infty$.

Quelques exemples de modèles.

- *La droite de régression linéaire* $y = f(t, x) = at + b$.

  Ici, il y a deux paramètres inconnus a et b. Si on pose $x_1 := b$ et $x_2 := a$ alors $y = x_1 + x_2 t$. En notant $\varphi_1 : t \mapsto 1$ et $\varphi_2 : t \mapsto t$, ce modèle s'écrit

$$y = f(t, x) = x_1 \varphi_1(t) + x_2 \varphi_2(t).$$



Régression linéaire : un cas particulier des moindres carrés.

- Le modèle cherché peut être *un polynôme du second degré*

$$y = f(t, x) = at^2 + bt + c$$

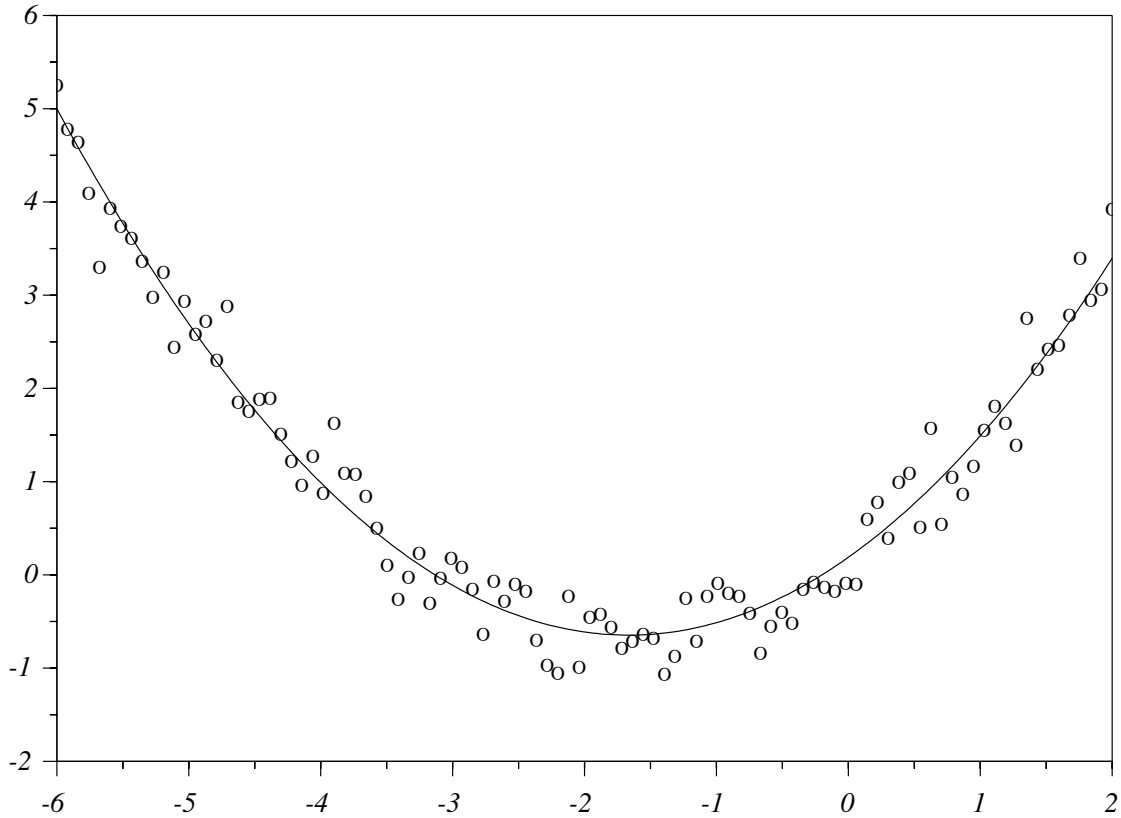
ou plutôt

$$y = f(t, x) = a(t - \bar{t})^2 + b(t - \bar{t}) + c$$

avec $\bar{t}$ l'entrée "moyenne" autour de laquelle on veut utiliser ce modèle.

En posant $x_1 := c$, $x_2 := b$ et $x_3 := a$, et $\varphi_1 : t \mapsto 1$, $\varphi_2 : t \mapsto t - \bar{t}$ et $\varphi_3 : t \mapsto (t - \bar{t})^2$, ce modèle s'écrit

$$f(t, x) = x_1\varphi_1(t) + x_2\varphi_2(t) + x_3\varphi_3(t).$$



Un polynôme du second degré.

- Si on a affaire à un phénomène périodique en t (de période T) on peut utiliser un polynôme trigonométrique (qui s'arrête à un niveau d'harmoniques adéquat) :

$$\begin{aligned} f(t,x) = x_1 + & \underbrace{x_2 \sin(\frac{2\pi t}{T}) + x_3 \cos(\frac{2\pi t}{T})}_{\textit{fondamental}} \\ + & \underbrace{x_4 \sin(\frac{4\pi t}{T}) + x_5 \cos(\frac{4\pi t}{T})}_{\textit{harmonique } 1} \\ + & \ldots \end{aligned}$$

Si la période T est connue, ce modèle est aussi, comme les deux précédents, de la forme

$$f(t,x) = \sum_j \varphi_j(t) x_j.$$

On dit alors qu'il est *linéaire*[1] en ses paramètres x_j.

1. Par contre, si la période T est inconnue, ce paramètre supplémentaire intervient de manière *non linéaire* dans le modèle.

- Un modèle souvent rencontré en physique est *le modèle exponentiel décroissant*

$$y = f(t, x) = Ce^{-t/\tau}$$

comportant les deux paramètres **inconnus** C et τ la constante de temps.

Ce modèle est linéaire en C mais non-linéaire en τ. Il est cependant possible de le linéariser (cf TD).

II. Prise en compte des erreurs de mesure

Les mesures y_i sont souvent entachées d'erreur. Pour prendre en compte ces erreurs, les y_i peuvent être considérées comme des réalisations de variables aléatoires

$$Y_i = f(t_i, x^*) + B_i$$

où $f(t_i, x^*)$ serait donc la valeur exacte (sans erreurs), correspondant aux paramètres optimaux x^*, à laquelle s'ajoute un "bruit" B_i qui suit la loi normale centrée (moyenne nulle), de variance σ_i^2 i.e.

$$B_i \sim N(0, \sigma_i^2)$$

Cela veut dire aussi qu'on suppose que les mesures expérimentales ne présentent pas de *biais* ($\mathbb{E}[B_i] = 0$) et l'écart type σ_i est alors une mesure de la précision de la i-ème mesure.

Si on dispose de cette information et que σ_i varie selon les mesures alors il est judicieux de minimiser plutôt la fonction d'erreurs avec poids (*weight*) :

$$E_w(x) = \sum_{i=1}^{m} \left(\frac{f(t_i, x) - y_i}{\sigma_i} \right)^2 \tag{2}$$

D'un point de vue intuitif cette nouvelle fonction d'erreurs permet bien de tenir compte de la précision des mesures :

- si la i-ème mesure est peu précise, c'est à dire si σ_i est grand alors elle apporte une contribution plus faible à $E_w(x)$: lorsque $\sigma_i \to \infty$ la contribution de la i-ème erreur tend vers 0 ;
- si la i-ème mesure est très précise, σ_i est petit et sa contribution dans $E_w(x)$ est amplifiée ; lorsque $\sigma_i \to 0$ cela implique qu'il faut trouver un jeu de paramètres tel que $f(t_i, x) \to y_i$: à la limite $f(t_i, x) = y_i$, le modèle doit interpoler la i-ème mesure.

Sous certaines hypothèses, on montre que la minimisation de cette fonction donne alors une solution $\bar{x}^{(m)}$ qui est, *en un certains sens*, le meilleur estimateur possible de x^* (celui ayant la plus petite variance).

Dans les deux cas (avec E et E_w), la solution tend vers la solution exacte x^* lorsque le nombre de mesures m tend vers l'infini, mais l'erreur sera en moyenne plus petite si on utilise E_w.

Cas de mesures répétées. S'il est possible d'effectuer plusieurs expériences (disons p) avec toujours les mêmes entrées $t_1, \ldots, t_m$, c'est à dire que l'on dispose pour chaque t_i de p mesures :

$$y_i^{(k)}, \quad \text{pour } k = 1, \ldots, p$$

alors on utilisera les estimations classiques de la moyenne et de l'écart type :

$$\boxed{y_i := \frac{1}{p} \sum_{k=1}^{p} y_i^{(k)}} \quad \text{et} \quad \boxed{\sigma_i := \sqrt{\frac{\sum_{k=1}^{p} (y_i^{(k)} - y_i)^2}{p-1}}} \tag{3}$$

Souvent l'appareillage de mesure apporte une indication sur la précision qui peut aussi être utilisée pour obtenir ces σ_i (par exemple si l'appareil assure une précision de 1% d'erreurs sur toute la plage des mesures observées, on prendra $\sigma_i = 0.01|y_i|$).

III. Modèles linéaires en leurs paramètres

Dans la suite de ce chapitre, on va considérer uniquement des modèles *linéaires* en leurs paramètres x_i, c'est à dire de la forme :

$$f(t,x) = x_1\varphi_1(t) + x_2\varphi_2(t) + \cdots + x_n\varphi_n(t) = \sum_{j=1}^{n} x_j\varphi_j(t) = \sum_{j=1}^{n} \varphi_j(t)x_j$$

avec $x = (x_1, \ldots, x_n)^\top$ et où les n fonctions φ_j sont données (suite à une phase de modélisation).

☛ *Nous allons montrer que le problème de moindres carrés revient alors à résoudre un problème d'algèbre linéaire.*

Pour cela, nous partons de la fonction d'erreurs E que nous développons en utilisant la linéarité du modèle :

$$E(x) = \sum_{i=1}^{m} (f(t_i,x) - y_i)^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \varphi_j(t_i)x_j - y_i \right)^2 \qquad (4)$$

Le terme $\sum_{j=1}^{n} \varphi_j(t_i)x_j$ correspond au produit matriciel entre le vecteur ligne $(\varphi_1(t_i), \ldots, \varphi_n(t_i))$ et le vecteur colonne x :

$$\begin{pmatrix} \varphi_1(t_i), \varphi_2(t_i), \ldots, \varphi_n(t_i) \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \varphi_1(t_i)x_1 + \cdots + \varphi_n(t_i)x_n$$

Ainsi, si on introduit la matrice A de dimension (m, n) et telle que

$$\boxed{a_{ij} = \varphi_j(t_i)} \tag{5}$$

on remarque que la i-ème ligne de A est égale à

$$A_i = \begin{pmatrix} \varphi_1(t_i) & \varphi_2(t_i) & \ldots & \varphi_n(t_i) \end{pmatrix}$$

d'où :

$$f(t_i, x) = \sum_{j=1}^{n} x_j \varphi_j(t_i) = A_i x = (Ax)_i$$

On obtient alors l'expression suivante pour la fonction erreurs :

Expression de la fonctions d'erreurs des *moindres carrés*

Soit E la fonction d'erreurs E définie par (1) et A la matrice de dimension (m, n) définie par (5). On a

$$E(x) = \sum_{i=1}^{m}((Ax)_i - y_i)^2 = \sum_{i=1}^{m}((Ax - b)_i)^2 = \|Ax - b\|_2^2. \qquad (6)$$

où on a posé $b := (y_1, y_2, \ldots, y_m)^\top$.

Remarque. Si les erreurs sur les mesures sont connues, il est facile de transformer (2) en (1). On obtient alors pour $E_w(x)$ la même expression (6) mais avec une matrice A et un vecteur b donnés par :

$$a_{ij} = \frac{\varphi_j(t_i)}{\sigma_i}, \quad b_i = \frac{y_i}{\sigma_i}$$

(au lieu de $a_{ij} = \varphi_j(t_i)$ et $b_i = y_i$), cf TD.

IV. Rappels et compléments de Calcul Différentiel

Dans ce qui suit, on considère une fonction $f : x \in \mathbb{R}^n \longmapsto f(x) \in \mathbb{R}$.

1. Dérivées partielles, gradient

Définition 1.

1. Soit $x = (x_1, \cdots, x_n)$ fixé. Si la fonction $g : \mathbb{R} \mapsto \mathbb{R}$ définie par
$$g(t) = f(x_1, \cdots, x_{j-1}, t, x_{j+1}, \cdots, x_n), \ \ t \in \mathbb{R},$$
est dérivable en $t = x_j$, on dit que la j-ème *dérivée partielle* de f existe et est égale à
$$\frac{\partial f}{\partial x_j}(x) = g'(x_j)$$

2. Si les n dérivées partielles existent, le *gradient* de f en x est alors le *vecteur ligne* défini par
$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_n}(x) \right).$$

Exemple. Soit $f(x_1, x_2) = -x_1^2 - x_2^2$ pour $(x_1, x_2) \in \mathbb{R}^2$. On a

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = -2x_1; \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = -2x_2$$

Le gradient de f est un vecteur ligne de $\mathbb{R}^2$ qui vaut donc

$$\nabla f(x) = (-2x_1, -2x_2)$$

2. Dérivée directionnelle.

Définition 2. *Dérivée directionnelle.*

On dit que la fonction f admet une *dérivée directionnelle* $Df(x,d)$ en $x \in \mathbb{R}^n$ dans la direction $d \in \mathbb{R}^n$ si et seulement si la limite suivante existe :

$$\lim_{t \to 0} \frac{f(x+td) - f(x)}{t} = Df(x,d).$$

Remarques :

1. Soit la fonction réelle g d'une variable, définie sur $\mathbb{R}$ par

   $$g(t) = f(x+td), \ \ t \in \mathbb{R}.$$

   Dire que f admet une dérivée directionnelle en x selon d est équivalent à dire que g est dérivable en $t = 0$.
2. La notion de dérivée directionnelle est simple à appréhender : on choisit un point x et une direction d dans l'espace : on regarde alors comment la fonction varie sur la droite $D_{x,d}$ définie par

   $$y \in D_{x,d} \Leftrightarrow \exists t \in \mathbb{R} \text{ tel que } y = x + td.$$

La fonction f restreinte sur cet ensemble devient alors une fonction d'une seule variable t et on cherche à calculer son "taux de variations" en $t = 0$ (c'est à dire en x).

Voir animation : `https://www.geogebra.org/m/R7nnr379`

Lien entre dérivée directionnelle et dérivée partielle.

Les dérivées partielles sont les dérivées directionnelles dans la direction des vecteurs de la base canonique e^i, c-a-d :

$$\frac{\partial f}{\partial x_i}(x) = Df(x, e^i). \tag{7}$$

☛ L'intérêt de cette remarque est qu'il est souvent plus facile de calculer les dérivées partielles via les dérivées directionnelles.

Exemples importants (exercices)

- *Forme linéaire.* Soit la fonction $f_1(x) = (y|x) = \sum_{i=1}^{n} y_i x_i$

  où y est un vecteur donné de $\mathbb{R}^n$. En calculant directement les dérivées partielles (ou bien aussi avec les dérivées directionnelles), on montre facilement que

$$\nabla f_1(x) = y^\top. \tag{8}$$

- *Forme quadratique.* Soit la fonction $f_2(x) = (Bx|x)$

  où B est une matrice réelle (n, n). En calculant la dérivée directionnelle (c'est bcp plus difficile avec les dérivées partielles), on obtient (exercice) :

$$\nabla f_2(x) = x^\top (B + B^\top). \tag{9}$$

- *Autre exemple* (où on est obligé d'utiliser la dérivée directionnelle).

  Soit $f_3 : \mathbb{R}^2 \mapsto \mathbb{R}$ définie par

$$f_3(x_1, x_2) = \begin{cases} 0 & \text{si } x = 0 \\ \dfrac{x_1 x_2}{x_1^2 + x_2^2} & \text{si } x \in \mathbb{R}^2 \setminus \{0\} \end{cases}$$

  On a $\dfrac{\partial f_3}{\partial x_1}(0,0) = \lim_{t \to 0} \dfrac{f_3(t,0) - f_3(0,0)}{t} = 0.$

3. Dérivabilité

Définition 3.

On dit que la fonction f est *dérivable* (ou *différentiable*) en $x \in \mathbb{R}^n$ si et seulement si, il existe une *application linéaire* $L_x \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ telle que :

$$\forall h \in \mathbb{R}^n, \quad f(x+h) = f(x) + L_x(h) + r(h)||h||$$

avec $r(h) : \mathbb{R}^n \to \mathbb{R}$ telle que :

$$\lim_{||h|| \to 0} r(h) = 0$$

Remarques :

1. $\|.\|$ désigne une norme quelconque de $\mathbb{R}^n$ (toutes les normes sont équivalentes dans $\mathbb{R}^n$).
2. Matriciellement une application linéaire de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ est un vecteur ligne.

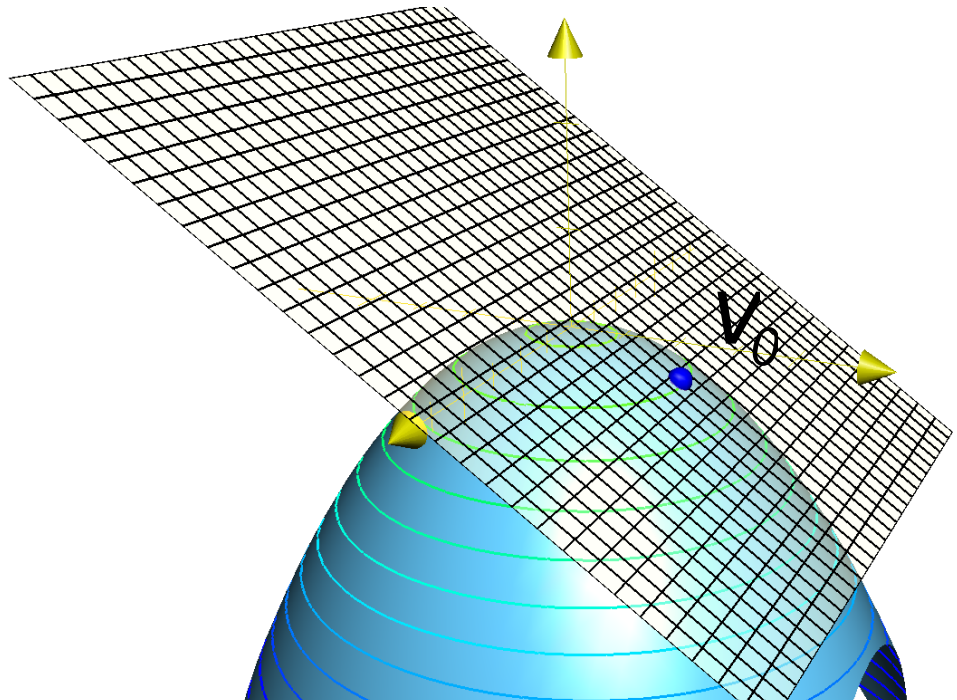
3. L_x est appelée *dérivée* ou aussi *différentielle* de f en x et notée

$$\boxed{f'(x) := L_x}$$

4. Intuitivement, une fonction est dérivable en x_0 si l'hypersurface définie par $z = f(x)$ admet un *plan tangent* au point $v_0 = (x_0, f(x_0))$, ou de façon équivalent, f peut être approchée en x_0 (au premier ordre) par une application affine (définissant un hyperplan) :

$$g(x) = f(x_0) + f'(x_0)(x - x_0)$$

Exemple : $n = 2$, une surface et son plan tangent en v_0.



4. Lien entre dérivée directionnelle et dérivabilité.

On a les propriétés importantes suivantes.

Théorème 1.

Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ dérivable en x, de dérivée $f'(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$. Alors

1. f est continue en x.
2. Les dérivées partielles $\dfrac{\partial f}{\partial x_i}(x)$, $i = 1, \cdots, n$ de f existent.
3. Pour tout $d \in \mathbb{R}^n \setminus \{0\}$, f admet une dérivée directionnelle selon d et
$$Df(x, d) = f'(x)d \tag{10}$$
En particulier, $\dfrac{\partial f}{\partial x_i}(x) = Df(x, e^i) = f'(x)e^i, \quad \forall i = 1, \cdots, n.$
Autrement dit,
$$\nabla f(x)v = f'(x)v, \quad \forall v \in \mathbb{R}^n \tag{11}$$

Remarque. D'après (11), on peut identifier l'espace $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ avec les vecteurs lignes de $\mathbb{R}^n$ et faire l'identification "$f'(x) = \nabla f(x)$".

Ainsi, si $f : \mathbb{R}^n \mapsto \mathbb{R}$ est dérivable en x alors toutes les dérivées partielles $\dfrac{\partial f}{\partial x_i}(x)$, $i = 1, \cdots, n$ existent bien.

La réciproque est fausse !

Ce n'est pas parce que toutes les dérivées partielles existent que la fonction est dérivable.

Exemple. Soit $f : \mathbb{R}^2 \mapsto \mathbb{R}$ définie par

$$f(x_1, x_2) = \begin{cases} 0 & \text{si } x = 0 \\ \dfrac{x_1 x_2}{x_1^2 + x_2^2} & \text{si } x \in \mathbb{R}^2 \setminus \{0\} \end{cases}$$

On montre que F n'est pas continue en $x = 0$ (calculer $f(t, t)$ et faire tendre $t \to 0$), donc pas dérivable en 0, mais les dérivées partielles existent bien en 0.

Cependant, pour la réciproque, on a quand même le résultat suivant.

Théorème 2.

Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ et $x \in \mathbb{R}^n$. On suppose qu'il existe $\delta > 0$ tel que toutes les dérivées partielles $\dfrac{\partial f}{\partial x_i}(x)$ existent dans $B(x, \delta)$ et sont continues en x. Alors f est dérivable en x.

5. Minimum d'une fonction de plusieurs variables.

(a) Minimum local et minimum global.

Définition 4.

1. On dit que $\bar{x} \in \mathbb{R}^n$ est un *minimum local* de f si et seulement si il existe une boule ouverte $B(\bar{x}, r)$ telle que :

$$\forall x \in B(\bar{x}, r),\ f(x) \geq f(\bar{x}).$$

Le point $\bar{x}$ est un minimum local *strict* si $\forall x \in B(\bar{x}, r)$, $x \neq \bar{x}$, $f(x) > f(\bar{x})$.

2. On dit que $\bar{x}$ est un *minimum global* de f si et seulement si

$$\forall x \in \mathbb{R}^n,\ f(x) \geq f(\bar{x}).$$

Le point $\bar{x}$ est le minimum global *strict* si $\forall x \in \mathbb{R}^n$, $x \neq \bar{x}$, $f(x) > f(\bar{x})$.

(b) Condition nécessaire d'optimalité.

Théorème 3. Condition nécessaire d'optimalité

On suppose que f est dérivable sur $\mathbb{R}^n$. Si $\bar{x}$ est un minimum local de f alors

$$\nabla f(\bar{x}) = 0. \tag{12}$$

On dit alors que $\bar{x}$ est un *point critique* de f.

Remarques.

- Si $\bar{x}$ est un *maximum* local de f alors c'est aussi un point critique i.e. $\nabla f(\bar{x}) = 0$
- La réciproque n'est pas vraie en général : un point critique n'est pas forcément un minimum ou un maximum local. Il peut s'agir d'un **point selle** (pour $n \geq 2$) i.e. un point critique $\bar{x}$ qui n'est justement ni un minimum local, ni un maximum local, autrement dit pour tout voisinage ouvert V de $\bar{x}$, il existe au moins deux points $x_1, x_2 \in V$ tels que

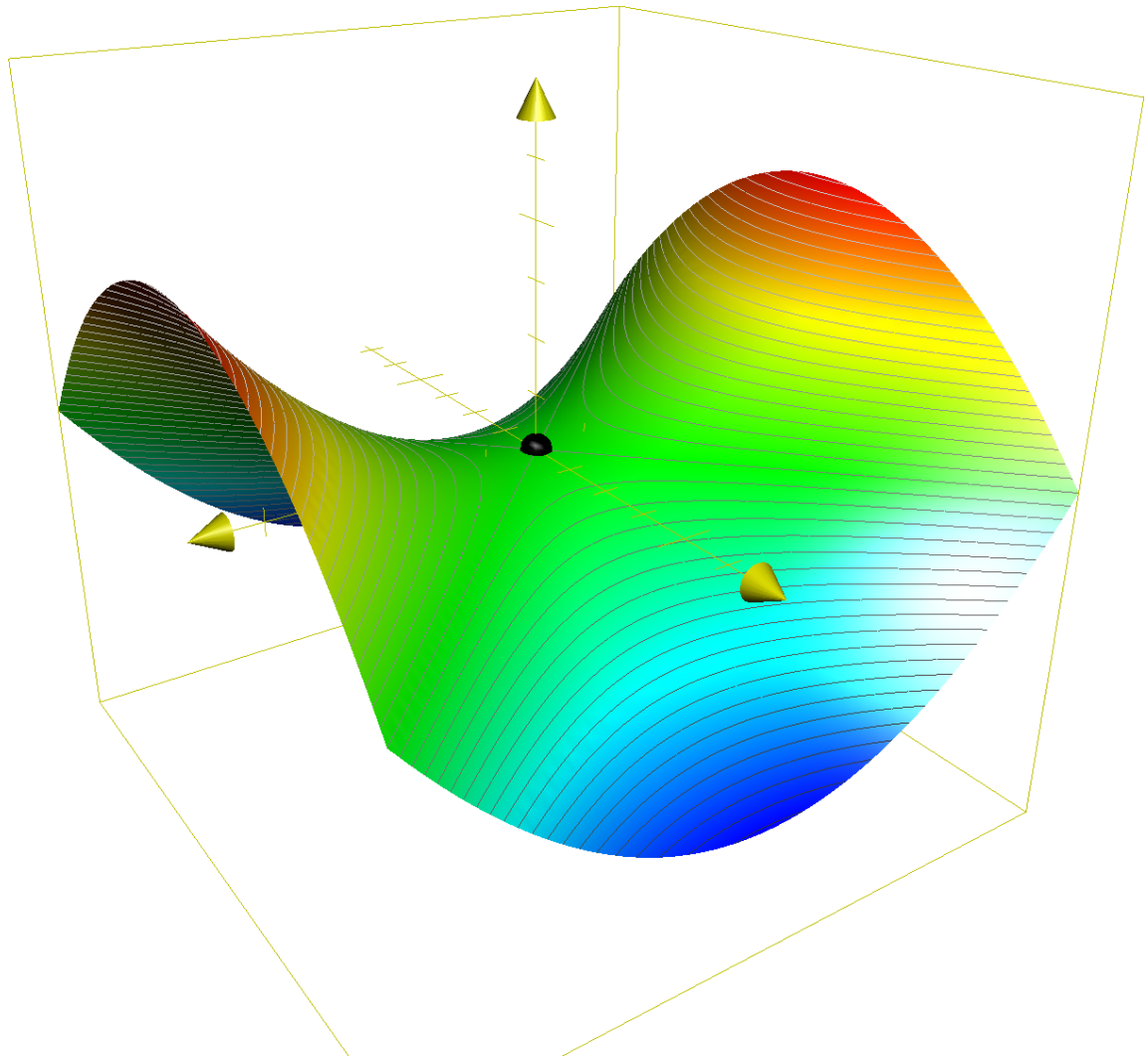
$$f(x_1) < f(\bar{x}) < f(x_2)$$

Ainsi, partant du point $\bar{x}$, on pourra toujours trouver des directions où la fonction f augmente et d'autres où elle diminue.

Exemples.

- Pour $\mathbb{R} \ni x \overset{f}{\mapsto} x^3$, en $x = 0$, on a $f'(0) = 0$ mais $\bar{x} = 0$ n'est ni un minimum local, ni un maximum local ($x = 0$ est un point d'inflexion avec $f''(0)$).

- La fonction $f(x, y) = x^2 - y^2$ admet $(\bar{x}, \bar{y}) = (0, 0)$ comme point selle.



Démonstration du Théorème 3. Comme $\bar{x}$ est un minimum local de f, il existe une boule $B(\bar{x}, r)$ telle que $f(x) \geq f(\bar{x}),\ \forall x \in B(\bar{x}, r)$. La fonction f étant dérivable on a :

$$f(x) = f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + r(h)||h||, \quad \lim_{h \to 0} r(h) = 0 \quad (13)$$

Prenons alors $h = \tau d$, où τ est un scalaire et $d \in \mathbb{R}^n$ une direction que l'on se fixe arbitrairement (avec $||d|| = 1$). On a pour tout τ tel que $|\tau| < r,\ x = \bar{x} + h \in B(\bar{x}, r)$, donc :

$$f(x) - f(\bar{x}) \geq 0 \quad (14)$$

soit, en utilisant (13) :

$$f'(\bar{x})\tau d + r(\tau d)|\tau| \geq 0 \quad (15)$$

- Pour $\tau > 0$, on déduit de (15) que

$$f'(\bar{x})d + r(\tau d) \geq 0$$

et donc en passant à la limite quand $\tau \to 0+$, on obtient :

$$f'(\bar{x})d \geq 0 \quad (16)$$

- Pour $\tau < 0$, on déduit de (15) que

$$f'(\bar{x})d - r(\tau d) \leq 0$$

et donc en passant à la limite quand $\tau \to 0-$, on obtient :

$$f'(\bar{x})d \leq 0 \tag{17}$$

D'où $f'(\bar{x})d = 0$ mais comme la direction d est arbitraire (avec cependant $||d|| = 1$) on en déduit que $f'(\bar{x}) = 0$ (par exemple en prenant successivement $d = e^1, d = e^2$, etc...). □

(c) Convexité.

Définition 5.

Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dite

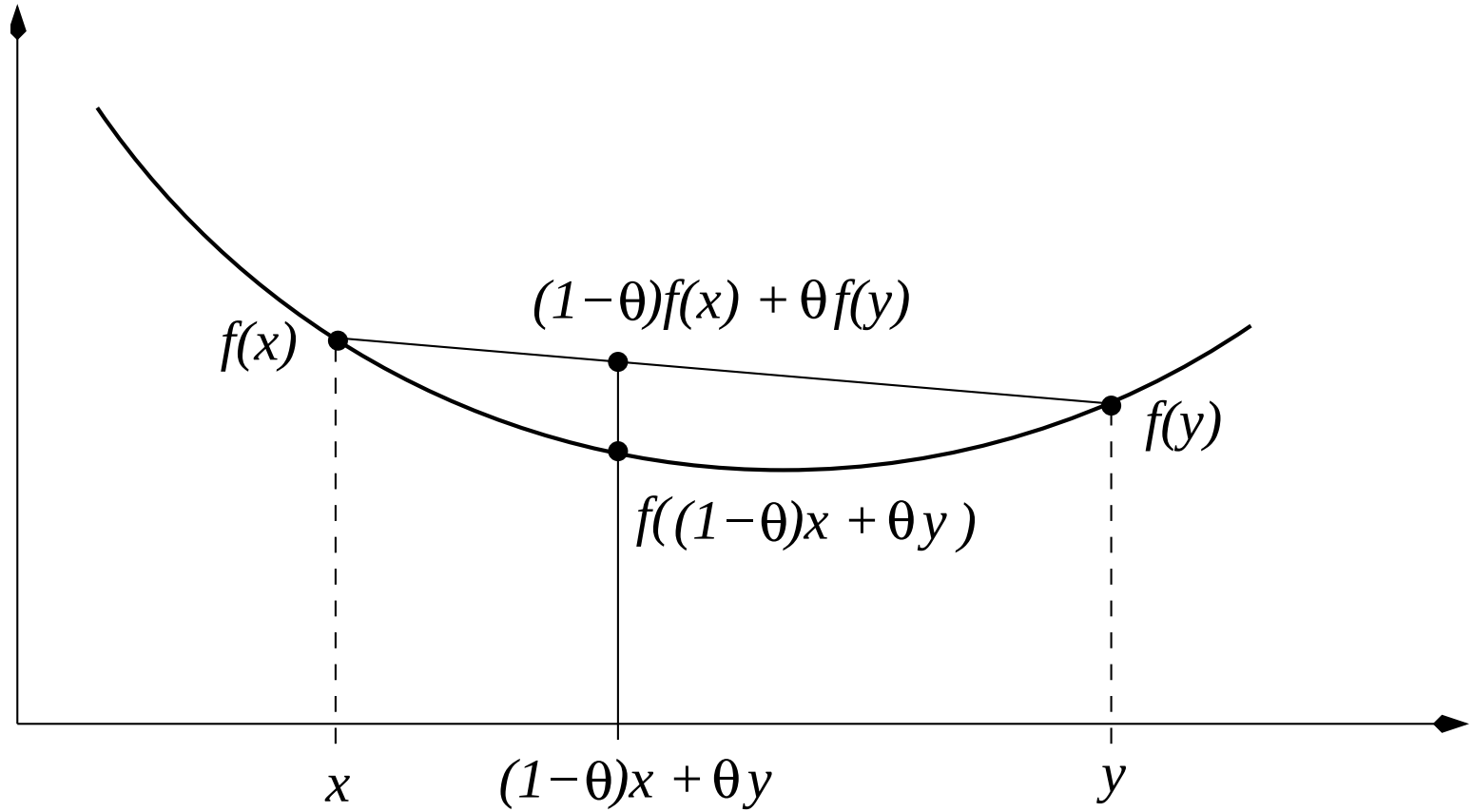
- *convexe* si $\forall x, y \in \mathbb{R}^n,\ \forall \theta \in [0,1]$, on a :

$$f((1-\theta)x + \theta y) \leq (1-\theta)f(x) + \theta f(y)$$

- *strictement convexe* si $\forall x, y \in \mathbb{R}^n, x \neq y,\ \forall \theta \in]0,1[$, on a :

$$f((1-\theta)x + \theta y) < (1-\theta)f(x) + \theta f(y)$$

Illustration de la convexité (avec $\theta \simeq 1/3$) :



La convexité permet d'obtenir une condition suffisante d'optimalité globale.

Théorème 4.

Si $f : \mathbb{R}^n \to \mathbb{R}$ est dérivable et *convexe* et s'il existe un point $\bar{x}$ vérifiant $\nabla f(\bar{x}) = 0$ alors $\bar{x}$ est un *minimum global* de f.

De plus, si f est *strictement* convexe alors ce minimum global est unique.

Démonstration.

- *Existence.* On écrit la propriété de convexité de f avec le point $\bar{x}$ et un point quelconque $x \neq \bar{x}$: $\forall \theta \in]0, 1[$, on a

$$f((1-\theta)\bar{x} + \theta x) < (1-\theta)f(\bar{x}) + \theta f(x)$$

On obtient avec $\theta \neq 0$,

$$F(\theta) := \frac{f(\bar{x} + \theta(x - \bar{x})) - f(\bar{x})}{\theta} < f(x) - f(\bar{x})$$

et en passant à la limite lorsque $\theta \to 0+$, il vient :

$$\lim_{\theta \to 0+} F(\theta) = Df(\bar{x}, x - \bar{x}) = f'(\bar{x})(x - \bar{x}) \leq f(x) - f(\bar{x})$$

Comme $f'(\bar{x}) = 0$, on obtient :

$$f(\bar{x}) \leq f(x), \ \forall x \in \mathbb{R}^n.$$

Ce qui montre que $\bar{x}$ est un minimum global de f.

- *Unicité.* On suppose de plus que f est strictement convexe. Soit $x^* \neq \bar{x}$ un autre minimum global, c'est-à-dire que $f(\bar{x}) = f(x^*) = m$ alors en utilisant la stricte convexité de f il vient :

$$f(\frac{1}{2}\bar{x} + \frac{1}{2}x^*) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(x^*) = m$$

on aurait donc trouvé un point $\hat{x} = \frac{1}{2}\bar{x} + \frac{1}{2}x^*$ pour lequel la valeur de f est strictement inférieure à son minimum global, ce qui est une contradiction. D'où $\bar{x} = x^*$.

□

V. Solution par les équations normales

1. Introduction.

Le problème des moindres carrés consiste à résoudre

$$\min_{x \in \mathbb{R}^n} E(x) = \|Ax - b\|_2^2 \tag{18}$$

où A est une matrice de taille (m, n) avec $m \geq n$.

- Si $m = n$ et A inversible alors on a une solution unique donnée par $x^* = A^{-1}b$. On aura de plus $f(t_i, x) = y_i$, $\forall i$: il s'agit d'un problème d'*interpolation*.
  Cette situation est rare (dans le cadre des mdc) et sans intérêt puisqu'il n'y a rien à minimiser...
- Si $m > n$ (en pratique $m \gg n$), il y a plus d'équations que d'inconnues et en général le système $Ax = b$ n'admet pas de solution. Par exemple, la droite de régression linéaire (ou une parabole...) ne peut généralement pas passer par $n > 2$ points (sauf si ceux-ci sont alignés...).

Hypothèse fondamentale sur A.

On suppose que

$$\ker A = \{0\}$$

c'est-à-dire que les colonnes de A sont linéairement indépendantes

On rappelle que le noyau de A est un s.e.v de $\mathbb{R}^n$, défini par

$$\ker A = \{x \in \mathbb{R}^n \text{ tel que } Ax = 0\}.$$

Si $\ker A = \{0\}$ alors la famille $\mathcal{A} = (A^1, \ldots, A^n)$ constituée des colonnes de A, est libre i.e. les A^j sont linéairement indépendants [2].

Cette hypothèse est réaliste : n vecteurs choisis au hasard dans un espace de dimension $m \geq n$ seront généralement linéairement indépendants.

---

2. *Vérification.* Soit $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ tels que $\sum_{j=1}^n \alpha_j A^j = 0$. On a $A^j = Ae^j$ et donc $0 = \sum_{j=1}^n \alpha_i Ae^j = A\left(\sum_{j=1}^n \alpha_j e^j\right)$ ce qui implique que $\sum_{j=1}^n \alpha_j e^j \in \ker A$. Sous l'hypothèse fondamentale, on a $\sum_{j=1}^n \alpha_j e^j = 0$ d'où l'on déduit $\alpha_j = 0$, $\forall j$.

2. E comme une forme quadratique.

Une forme quadratique est la généralisation d'un polynôme du second degré dans le cas multidimensionnel. La fonction d'erreur $E : x \in \mathbb{R}^n \longmapsto \|Ax - b\|^2 \in \mathbb{R}$ se développe en écrivant le carré de la norme comme un produit scalaire [3] :

$$\begin{aligned} E(x) &= \|Ax - b\|_2^2 = (Ax - b \,|\, Ax - b) \\ &= (Ax \,|\, Ax) - 2(Ax \,|\, b) + (b \,|\, b) \\ &= (A^\top Ax \,|\, x) - 2(x \,|\, A^\top b) + \|b\|^2 \\ &= x^\top A^\top Ax - 2x^\top (A^\top b) + \|b\|^2. \end{aligned}$$

On a donc

$$\boxed{E(x) = (Gx \,|\, x) - 2(x \,|\, h) + c} \tag{19}$$

avec

$$G = A^\top A, \;\; h = A^\top b$$
$$c = \|b\|^2 \text{ une constante positive.}$$

3. On rappelle la propriété du produit scalaire : $(Ax \,|\, y) = (x \,|\, A^\top y)$, $\forall x, y \in \mathbb{R}^n$

L'expression (19) de E peut aussi s'écrire

$$\boxed{E(x) = x^\top Gx - 2h^\top x + c} \tag{20}$$

3. Propriétés de E.

(a) Rappels et compléments sur les matrices symétriques définies positives (voir MN1, Chap.3, *factorisation de Cholesky*).

Définition 6. Matrice semi-définie/définie positive

- Une matrice carrée $M \in \mathbb{R}^{n\times n}$ est dite *semi-définie positive* si :

$$(Mx \,|\, x) = x^\top Mx \geq 0, \ \forall x \in \mathbb{R}^n.$$

- Si $(Mx \,|\, x) > 0, \ \forall x \neq 0$, on dit que M est *définie positive*.

Remarque importante.

Le plus souvent ces deux notions s'appliquent sur des matrices *symétriques*.

En effet, si M n'est pas symétrique, on montre facilement que la matrice *symétrisée* de M

$$\boxed{M^{\mathcal{S}} := (M + M^{\top})/2}$$

est symétrique et

$$(M^{\mathcal{S}} x \,|\, x) = (Mx \,|\, x)$$

☛ Ainsi, dans le terme $(Mx \,|\, x)$, **seule la partie symétrisée $M^{\mathcal{S}}$ de M intervient** et on peut donc toujours considérer la matrice M symétrique.

On a les propriétés suivantes sur les matrices semi-définie/définie positive (rappels et compléments, voir MN1, Chap.3).

Proposition 1.

Soit $M \in \mathbb{R}^{n\times n}$.

1. Si M est symétrique toutes ses valeurs propres sont réelles.
2. Si M est définie positive alors elle est inversible. De plus, les valeurs propres de sa partie symétrisée $M^{\mathcal{S}}$ sont toutes >0 [a].
3. La matrice $M^{\top}M$ est symétrique et semi-définie positive.

---

a. Ainsi, si M est symétrique définie positive, toutes ses valeurs propres sont réelles et >0.

Preuve.

1. On introduit le produit scalaire $(\cdot|\cdot)$ pour les complexes dans $\mathbb{C}^n$. Pour $x, y \in \mathbb{C}^n$, on définit

$$(x|y) = \sum_{i=1}^{n} x_i \bar{y}_i \tag{21}$$

où $\bar{y}_i$ est le complexe conjugué de la composante $y_i \in \mathbb{C}$ de y.
On rappelle les propriétés suivantes de ce produit scalaire. Pour $x, y \in \mathbb{C}^n$,

- $(x|y) = \overline{(y|x)}$
- $(x|x) \in \mathbb{R}^+$. On peut alors définir la norme $\|x\| = \sqrt{(x|x)}$.
- $(\alpha x|y) = \alpha(x|y)$ et $(x|\alpha y) = \overline{\alpha}(x|y)$ pour $\alpha \in \mathbb{C}$.

Soit (λ, u) un élément propre de M i.e $\lambda \in \mathbb{C}$ et $u \in \mathbb{C}^n$, $u \neq 0$ t.q.

$$Mu = \lambda u$$

En prenant le produit scalaire avec u, on obtient

$$(Mu|u) = (\lambda u|u) = \lambda(u|u) = \lambda\|u\|^2 \tag{22}$$

Par ailleurs, on a $(Mu|u) = (u|M^\top u) = (u|Mu)$ car M est symétrique. On obtient

$$\lambda\|u\|^2 = (Mu|u) = (u|Mu) = (u|\lambda u) = \overline{\lambda}(u|u) = \overline{\lambda}\|u\|^2$$

d'où $(u \neq 0)$

$$\lambda = \overline{\lambda} \quad \text{i.e. } \lambda \in \mathbb{R}.$$

2. Comme M est carrée, pour montrer que M est inversible, on montre que $\ker M = \{0\}$ (exercice). Pour montrer que toute v.p. λ de $M^{\mathcal{S}}$ est > 0, on utilise (22). On a $(u \neq 0)$:
$0 < (Mu|u) = (M^{\mathcal{S}}u|u) = \lambda\|u\|^2$.
3. La matrice $M^\top M$ est symétrique, en effet [4] :

$$(M^\top M)^\top = M^\top (M^\top)^\top = M^\top M$$

Elle est aussi semi-définie positive, puisque

$$(M^\top Mx \,|\, x) = (Mx \,|\, Mx) = \|Mx\|_2^2 \geq 0, \ \forall x \in \mathbb{R}^n$$

□

4. $(AB)^\top = B^\top A^\top$

(b) Propriétés de la matrice $G = A^\top A$.

Proposition 2.

Sous l'hypothèse $\ker A = \{0\}$, la matrice $G = A^\top A$ est symétrique, définie positive et donc inversible.

Preuve. D'après la Proposition 1., la matrice G est symétrique et semi-définie positive. De plus, on a

$$(Gx \,|\, x) = 0 \Leftrightarrow \|Ax\|_2 = 0 \Leftrightarrow Ax = 0 \Leftrightarrow x \in \ker A$$

et donc G est bien définie positive ssi $\ker A = \{0\}$. □

(c) Convexité de E.

Proposition 3.

La fonction d'erreur E est convexe et sous l'hypothèse $\ker A = \{0\}$, elle est *strictement* convexe.

Preuve. Soit $q = (1-\theta)E(x) + \theta E(y) - E((1-\theta)x + \theta y)$. On voit assez facilement que le terme linéaire et le terme constant de E n'interviennent pas et on a, pour $0 < \theta < 1$,

$$q = (1-\theta)(Gx \,|\, x) + \theta(Gy \,|\, y) - (G((1-\theta)x + \theta y) \,|\, (1-\theta)x + \theta y).$$

Après calculs (exercice) et en utilisant le fait que G est symétrique, on obtient

$$q = \theta(1-\theta)(G(x-y) \,|\, x-y).$$

La matrice G étant semi-définie positive, on a $q \geq 0$ i.e. E est convexe. De plus, sous l'hypothèse $\ker A = \{0\}$, la matrice G est définie positive et donc avec $x \neq y$, on a $q > 0$ i.e. E est *strictement* convexe. □

4. Résolution par le zéro de la dérivée de E.

(a) Calcul du gradient de E.

On veut appliquer les résultats de Calcul Différentiel et en particulier le Théorème 4. Comme E est une forme quadratique, c-a-d un polynôme du second degré en les variables $x_1, x_2, \ldots, x_n$, E est (indéfiniment) dérivable sur $\mathbb{R}^n$. Par conséquent une condition nécessaire pour que $\bar{x}$ soit un minimum est que (cf. Théorème 3.) :

$$\nabla E(\bar{x}) = 0. \tag{23}$$

On rappelle que $E = x^\top G x - 2h^\top x + c$ avec $G = A^\top A$ et $h = A^\top b$.

Le gradient de E vaut (exercice, cf. exercice TD) :

$$\boxed{\nabla E(x) = x^\top (G + G^\top) - 2h^\top = 2(x^\top G^\top - h^\top)} \tag{24}$$

car G est symétrique.

(b) Equations normales.

L'équation (23) avec (24) devient :

$$\bar{x}^\top G^\top - h^\top = 0.$$

En transposant (pour se ramener à un système linéaire classique), on obtient finalement :

$$\boxed{A^\top A\,\bar{x} = A^\top b} \tag{25}$$

C'est un système d'équations linéaires que l'on appelle *équations normales* du problème de moindres carrés.

Sous l'hypothèse $\ker A = \{0\}$, la matrice $G = A^\top A$ est inversible (cf. Proposition 2) et le système (25) admet une unique solution. De plus, E est strictement convexe (cf. Proposition 3). D'après le Théorème 4, on en déduit le résultat suivant.

Proposition 4. Equations normales.

Soit A une matrice de taille (m, n) telle que $\ker A = \{0\}$ et b un vecteur de $\mathbb{R}^m$. Le problème de moindres carrés

$$min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \tag{26}$$

admet une unique solution $\bar{x} \in \mathbb{R}^n$ qui vérifie le système

$$A^\top A \bar{x} = A^\top b \tag{27}$$

appelé *équations normales* du problème de moindres carrés.

Remarques.

- On peut montrer que les équations normales admettent **toujours au moins** une solution i.e.

  $$h = A^\top b \in Im\,(A^\top A) = \{z \in \mathbb{R}^m,\ \exists y \in \mathbb{R}^n,\ A^\top A\, y = z\}$$

  et ceci **sans l'hypothèse** que $\ker A = \{0\}$. On peut même établir que si $\ker A \neq \{0\}$, il y a une **infinité** de solution aux équations normales.

- Le système $Ax = b$ n'admet généralement pas de solution alors que les équations normales $A^\top Ax = A^\top b$ admettent toujours (au moins) une solution.
- Les équations normales sont donc un simple système linéaire carré : on a une unique solution $\bar{x}$ si et seulement si G est *inversible* et alors cette solution peut s'obtenir par la méthode de Gauss (ou Cholesky) (cf. MN1, Chap.2).
- On peut montrer que sur une matrice symétrique et définie positive (sdp) la factorisation LU **sans échange d'équations** est très stable. Par ailleurs, on peut diminuer le coût calcul par deux en utilisant la méthode de Cholesky puisque G est sdp.

Exercice. La droite des moindres carrés

On dispose de données $(t_i, y_i)_{1 \leq i \leq m}$ sensées suivre le modèle affine $y = f(t) := at + b$. On cherche à identifier les paramètres a et b.

1. On pose $x_1 := b$ et $x_2 := a$, écrire le modèle sous la forme $f(t) = x_1 \varphi_1(t) + x_2 \varphi_2(t)$
2. Pour un jeu de paramètres $x := (x_1, x_2)^\top$ donné, quelle est l'erreur $e_i(x)$ entre le modèle et la i-ème mesure ?
3. Sans connaissance des erreurs de mesure, quelle fonction (notée E) de x faut-il minimiser pour identifier les paramètres au sens des moindres carrés ? On écrira d'abord E à l'aide des m erreurs $e_i(x)$ puis sous la forme $E(x) = \|Ax - b\|_2^2$ en explicitant la matrice A et le vecteur b.
4. Écrire les équations normales permettant de trouver la valeur optimale des paramètres.

Solution.

- Le problème s'écrit

$$\min_{x=(x_1,x_2)\in\mathbb{R}^2} E(x)$$

où la fonction erreur E vaut :

$$\begin{aligned} E(x) &= \sum_{i=1}^{m} e_i^2(x) = \sum_{i=1}^{m} (f(t_i,x) - y_i)^2 = \sum_{i=1}^{m} (x_1 + x_2 t_i - y_i)^2 \\ &= \sum_{i=1}^{m} ((Ax-b)_i)^2 \end{aligned}$$

avec A matrice de taille $(m,2)$ et $b \in \mathbb{R}^m$ donnés par

$$A = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

- Equations normales $A^\top Ax = A^\top b$ avec

$$A^\top A = \begin{pmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix}, \quad A^\top b = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m t_i y_i \end{pmatrix}$$

5. Résolution par minimisation directe d'une forme quadratique.

On a la propriété suivante pour les formes quadratiques.

Lemme 1. Minimum d'une forme quadratique.

Soit $M \in \mathbb{R}^{n\times n}$ une matrice symétrique, définie positive (donc inversible), v un vecteur de $\mathbb{R}^n$ et c un scalaire. La forme quadratique

$$Q(x) = x^\top Mx - 2v^\top x + c, \quad x \in \mathbb{R}^n$$

admet un minimum unique x^* solution du système linéaire $Mx^* = v$.

Preuve en exercice.

Indication : calculer $(M(x - x^*) \,|\, x - x^*)$ en fonction de $Q(x)$ et $Q(x^*)$ et montrer que $Q(x) > Q(x^*), \forall x \neq x^*$.

Solution. On a $$\begin{aligned} Q(x^*) &= x^{*\top} \underbrace{Mx^*}_{=v} - 2v^\top x^* + c \\ &= x^{*\top} v - 2v^\top x^* + c \\ &= -v^\top x^* + c = -(v \,|\, x^*) + c \end{aligned}$$

On développe le produit scalaire

$$\begin{aligned} (M(x - x^*) \,|\, x - x^*) &= (Mx \,|\, x) - (Mx \,|\, x^*) \\ &\quad -(Mx^* \,|\, x) + (Mx^* \,|\, x^*) \\ &= (Mx \,|\, x) - 2(Mx^* \,|\, x) + (Mx^* \,|\, x^*) \end{aligned}$$

car on a $(Mx \,|\, y) = (x \,|\, M^\top y) = (x \,|\, My)$ puisque M est symétrique.

On obtient ainsi

$$\begin{aligned} (M(x - x^*) \,|\, x - x^*) &= (Mx \,|\, x) - 2(v \,|\, x) + (v \,|\, x^*) \\ &= (Mx \,|\, x) - 2(v \,|\, x) + c - (-(v \,|\, x^*) + c) \\ &= Q(x) - Q(x^*) \end{aligned}$$

Puisque M est définie positive, on a $(M(x - x^*) \,|\, x - x^*) > 0$ pour tout $x \neq x^*$. Par conséquent,

$$Q(x^*) < Q(x) \text{ pour tout } x \neq x^*$$

ce qui prouve que x^* est l'unique minimum de Q. □

En appliquant le Lemme 1, avec $M = G = A^\top A$ et $v = A^\top b$, on retrouve le résultat de la Proposition 4 pour l'existence d'un (unique) vecteur minimisant le problème de moindres carrés, solution des équations normales.

6. Résumé de la méthode des moindres carrés.

En résumé, résoudre un problème de moindres carrés (dont le modèle est linéaire en ses paramètres) en résolvant les équations normales, se fait selon les étapes suivantes :

1. choix du modèle, c'est à dire des fonctions φ_j.
2. avec les (t_i, y_i) et éventuellement les σ_i, calcul de la matrice A et du vecteur b.
3. calcul de la matrice $G = A^\top A$ et du vecteur $h = A^\top b$.
4. résolution du système linéaire $Gx = h$ via donc une factorisation de Cholesky ; c'est à ce niveau qu'on peut détecter le problème de non inversibilité de G via une estimation du conditionnement de cette matrice.

VI. Autres résolutions : QR, SVD, régularisation.

1. Introduction.

Il peut arriver que la matrice $G = A^\top A$ des équations normales :

- soit singulière lorsque $\ker A \neq \{0\}$. On rappelle que dans ce cas, il y a une infinité de solutions.
- ou bien soit mal conditionnée avec $\kappa(G) = \|G^{-1}\|\|G\|$ "grand" ($\kappa(G)$ se rapprochant de $\frac{1}{\mathbf{u}}$)

Remèdes possibles.

1. Résoudre le problème de moindres carrés avec une méthode plus stable comme utilisant une factorisation QR (algorithme standard comme par exemple dans python/scipy ou numpy).
2. Rajouter une contrainte sur les paramètres : si on considère le cas où G est singulière, il y a une infinité de solutions. L'ajout d'une contrainte va permettre de choisir une solution particulière, comme par exemple *la solution de norme minimale* : c'est cette solution qui est obtenue avec la SVD de A.
3. Minimiser une fonction d'erreur "régularisée".

2. Résolution par la méthode QR.

La matrice A de taille (m, n) admet une décomposition $A = QR$ avec Q une matrice de taille (m, m), orthogonale ($Q^{-1} = Q^\top$) et R est une matrice de taille (m, n) de la forme (cf. MN1, Chap.2, cas d'une matrice rectangulaire) :

$$R = \begin{pmatrix} \boxed{S} \\ \boxed{0} \end{pmatrix} \begin{matrix} \} n \\ \} m - n \end{matrix}$$

où S est une matrice *triangulaire supérieure* de taille (n, n).
On a alors

$$\|Ax - b\|_2^2 = \|QRx - b\|_2^2 = \|Q(Rx - Q^\top b)\|_2^2$$

car $QQ^\top = I_m$. Or, pour tout $x \in \mathbb{R}^m$, on a

$$\|Qx\|_2^2 = (Qx \mid Qx) = (x \mid Q^\top Qx) = \|x\|_2^2$$

(on a $\|Q\|_2 = 1$). Donc

$$\|Ax - b\|_2^2 = \|Rx - Q^\top b\|_2^2$$

On a

$$Rx = \begin{pmatrix} Sx \\ 0 \end{pmatrix}, \quad Q^\top b = c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \in \mathbb{R}^m$$

avec $c_1 \in \mathbb{R}^n$ et $c_2 \in \mathbb{R}^{m-n}$. Par conséquent, on obtient

$$\|Ax - b\|_2^2 = \|Rx - Q^\top b\|_2^2 = \|Sx - c_1\|_2^2 + \|c_2\|_2^2$$

On voit ainsi que le problème des moindres carrés, se réduit à

$$\min_{x \in \mathbb{R}^n} \|Sx - c_1\|_2^2 \tag{28}$$

Sous l'hypothèse que $\ker A = \{0\}$, la matrice S est inversible [5] et dans ce cas, la solution $x^* \in \mathbb{R}^n$ du problème des moindres carrés est donnée par

$$\boxed{Sx^* = c_1} \tag{29}$$

Ce système est facile à résoudre car S est triangulaire supérieure.

5. Soit $x \in \mathbb{R}^n$ tel que $Sx = 0$. On a alors $Rx = 0$ et $Ax = QRx = 0$ d'où $x = 0$ avec l'hypothèse $\ker A = \{0\}$.

3. Résolution par SVD.

La matrice A de taille (m, n) avec $m \geq n$ et de rang[6] $r = \mathit{rang}\, A \leq n$ admet la décomposition en SVD suivante (cf. MN1, Chap.2) :

$$A = U\Sigma V^\top$$

où $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ sont orthogonales et $\Sigma \in \mathbb{R}^{m\times n}$ est donnée par

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \text{ avec } \Sigma_1 = \mathit{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$$

(Σ_1 est une matrice diagonale) et

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

Si $\ker A = \{0\}$ alors $r = \mathit{rang}\, A = n$ (*rang plein*) i.e. toutes les colonnes de A sont linéairement indépendantes et on a vu que la matrice $G = A^\top A$ des équations normales est inversible.

6. Le rang de A noté $\mathit{rang}\, A$ est le nombre maximal de colonnes de A linéairement indépendantes. Si $\ker A = \{0\}$ alors $\mathit{rang}\, A = n$ et on dit que A est de *rang plein* dans ce cas-là.

Mais ici, on ne fait pas cette hypothèse de rang plein et on considère donc de façon générale que $r = \mathit{rang}\, A \leq n$. Un des points forts de la SVD est justement de permettre de traiter le cas $r = \mathit{rang}\, A < n$.

Dans ce cas ($r < n$), le problème de moindres carrés admet une infinité de solutions et on va chercher une solution x de *norme minimale* avec la SVD.

Utilisation de la SVD de A.

$$\begin{aligned} \text{On a } \|Ax - b\|_2^2 &= \|U\Sigma V^\top x - b\|_2^2 \\ &= \|U(\Sigma V^\top x - U^\top b)\|_2^2 \\ &= \|\Sigma \underbrace{V^\top x}_{=y} - \underbrace{U^\top b}_{=c}\|_2^2 \quad (\text{car } U \text{ orthogonale}) \end{aligned}$$

On pose

$$\boxed{y = V^\top x} \text{ et } \boxed{c = U^\top b} \tag{30}$$

On a de plus (U et V étant orthogonales),

$$\|y\|_2 = \|x\|_2 \text{ et } \|c\|_2 = \|b\|_2 \tag{31}$$

Ainsi, $\|Ax - b\|_2^2 = \|\Sigma y - c\|_2^2$

$$= \sum_{i=1}^{r} |\sigma_i y_i - c_i|^2 + \sum_{i=r+1}^{n} c_i^2$$

Toutes les solutions du problème de moindres carrés sont alors données par

$$y_i = \begin{cases} c_i/\sigma_i & \text{pour } i = 1, \ldots, r \\ \alpha_i \text{ quelconque} & \text{pour } i = r+1, \ldots, n \end{cases} \tag{32}$$

- Dans le cas où $r = \mathit{rang}\, A = n$ (rang plein), on retrouve la solution unique.
- Dans le cas où $r < n$, on choisit la solution y de *norme minimale* et donc la solution x correspondante aussi de *norme minimale* puisque $\|y\|_2 = \|x\|_2$ (cf. (31)).

On a

$$\|y\|_2^2 = \sum_{i=1}^{r} (c_i/\sigma_i)^2 + \sum_{i=r+1}^{n} \alpha_i^2.$$

On choisit donc $\alpha_i = 0$ pour tout $i = r+1, \ldots, n$, de sorte que la solution de *norme minimale* est donnée par

$$\boxed{y_i^* = \begin{cases} c_i/\sigma_i & \text{pour } i = 1, \ldots, r \\ 0 & \text{pour } i = r+1, \ldots, n \end{cases}} \qquad (33)$$

On récupère alors la solution x^* de norme minimale grâce à (30) :

$$x^* = Vy^*$$

Remarque. La factorisation SVD d'une matrice est un outil puissant pour résoudre les problèmes de moindres carrés ; cependant le coût de calculs est plus important (que les équations normales ou QR) et la méthode est limitée à des problèmes de moindres carrés de taille raisonnable.

4. Régularisation.

Une alternative à la SVD (en particulier dans le cas où $\mathit{rang}\, A < n$) est de minimiser la fonction d'erreur "régularisée" :

$$E_\epsilon(x) = \|Ax - b\|_2^2 + \epsilon \|x\|_2^2 \tag{34}$$

avec une valeur $\epsilon > 0$:

- pour ϵ "petit", on favorise la minimisation de $E(x) = \|Ax - b\|_2^2$;
- pour ϵ "grand", on favorise la minimisation de $\|x\|_2^2$.

Il faut trouver un équilibre (beaucoup de papiers sur ce sujet...).

Trouver le minimum de E_ϵ est équivalent à résoudre le système linéaire

$$(A^\top A + \epsilon I)x^* = A^\top b.$$

De plus, la matrice $A^\top A + \epsilon I$ est définie positive et donc toujours inversible (avec $\epsilon > 0$), sans avoir besoin de l'hypothèse $\ker A = \{0\}$ (exercice).

VII. Application des moindres carrés aux systèmes de recommandation

1. Rang d'une matrice.

On rappelle que le rang d'une matrice $A \in \mathbb{R}^{m\times n}$ noté rang(A) est la dimension du sous-espace engendré par ses vecteurs colonnes et c'est aussi la dimension de son espace image :

$$\text{rang}(A) = \dim\left(\text{vect}(A^1, \ldots, A^n)\right) = \dim(\text{Im}(A)) \qquad (35)$$

où A^i désigne le i-ème vecteur colonne de A et

$$\text{Im}(A) = \{y \in \mathbb{R}^m,\ \exists x \in \mathbb{R}^n \text{ tel que } Ax = y\}$$

Le rang de A est aussi le nombre maximal de vecteurs colonnes linéairement indépendants.

Proposition 5. Caractérisation du rang d'une matrice

Soit une matrice $A \in \mathbb{R}^{m\times n}$ et un entier $k \leq p = \min(m, n)$. On a

$$\text{rang}(A) \leq k \ \Leftrightarrow\ \exists X \in \mathbb{R}^{m\times k} \text{ et } Y \in \mathbb{R}^{n\times k} \text{ telles que } A = XY^\top$$

2. Systèmes de recommandation.

Dans un contexte de systèmes de recommandation tels qu'ils peuvent être proposés par des plateformes de streaming (Youtube, Spotify, Netflix...), les utilisateurs expriment leurs préférences de certains items (avec une note) et le système de recommandation cherche à calculer une note aux items qui n'en ont pas reçus des utilisateurs.

Dans l'exemple ci-dessous, 4 utilisateurs expriment leurs préférences parmi 6 films.

	Star Wars épisode VIII	Pulp Fiction	Le fabuleux destin d'Amélie Poulain	Matrix	Il était une fois dans l'Ouest	Les Bronzés à TELECOM Nancy
Cathy	5	?	3	2	3	?
Stéphane	4	5	4	1	2	1
Sophie	1	?	?	?	?	?
Anna	?	3	5	4	2	?

Les points d'interrogations indiquent l'absence de notes. On dispose ainsi d'une matrice de notes/préférences *incomplète*.

Les systèmes de recommandation reposent souvent sur l'idée que les préférences peuvent être expliquées par un petit nombre de facteurs dits *latents*. Un système de recommandation va évaluer les notes manquantes en déterminant une approximation de rang faible $k \ll 1$ de la matrice incomplète des notes.

Modélisation. On considère m utilisateurs exprimant des préférences (notes) de certains objets d'un ensemble de n objets (ou items).

☛ On dispose ainsi d'une matrice des notes/préférences $A \in \mathbb{R}^{m \times n}$, où A_{ij} représente la note donnée par l'utilisateur i à l'objet j.

Certaines entrées peuvent être manquantes. L'objectif est de construire un modèle de recommandation basé sur une factorisation de rang faible $k \ll 1$ de A. En suivant la propriété de la Proposition 5, on cherche une approximation

$$A \simeq XY^\top$$

avec $X \in \mathbb{R}^{m \times k}$, $Y \in \mathbb{R}^{n \times k}$ et $k \ll p = \min(m, n)$.

Ainsi, on voudrait déterminer les matrices X et Y qui minimisent la norme

$$\|A - XY^\top\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij} - X_i Y_j{}^\top\right)^2$$

où X_i désigne la i-ème ligne de X et $\|\cdot\|_F$ désigne la norme de Frobenius[7] définie par

$$\|M\|_F = \left(\sum_{i,j} M_{ij}^2\right)^{1/2}$$

Puisque A est incomplètement connue, on va minimiser uniquement sur les quantités connues dans la matrice A. On note

$$\Omega \subset [\![1, m]\!] \times [\![1, n]\!]$$

l'ensemble des indices des coefficients connus i.e. $(i, j) \in \Omega$ si et seulement si le coefficient A_{ij} de A est connu.

7. La norme de Frobenius n'est pas une norme subordonnée à une norme vectorielle mais elle vérifie quand même la propriété de sous-multiplicativité $\|AB\|_F \leq \|A\|_F \|B\|_F$

On définit alors

$$\|A - XY^\top\|_{F,\Omega} := \left(\sum_{(i,j)\in\Omega} \left(A_{ij} - X_i Y_j^\top\right)^2 \right)^{1/2} \tag{36}$$

On cherche alors les matrices X et Y solution de

$$\boxed{\min_{X,Y} \|A - XY^\top\|_{F,\Omega}^2} \tag{37}$$

Le but de l'approximation $\hat{A} := XY^\top$ de A est de prédire les valeurs $\hat{A}_{ij}$ pour $(i,j) \notin \Omega$ c'est-à-dire prédire les valeurs de A_{ij} inconnues.

3. Méthode ALS (Alternating Least Squares).

Pour résoudre (37), on utilise la méthode ALS qui consiste à résoudre itérativement et alternativement des problèmes de moindres carrés *linéaires* pour Y fixé puis pour X fixé (via des équations normales). On définit ainsi une suite de matrices $(X^{(l)}, Y^{(l)})_{l\geq 0}$ qui va converger vers une solution de (37).

La fonction objectif dans (37) n'est pas convexe par rapport au couple (X, Y), il peut donc y avoir existence et convergence vers un minimum local. On peut ajouter à la fonction objectif des termes de régularisation du type $\lambda\|X\|_2^2$ et $\lambda\|Y\|_2^2$ (avec $\lambda > 0$) pour sélectionner les solutions de normes petites (cf. section *VI*.§5 "Régularisation").

☛ Voir exercice de TD (Feuille 2) pour une description détaillée de la méthode ALS.

Avec l'exemple précédent, on obtient l'approximation $\hat{A} := XY^\top$ suivante (avec $k = 2$) :

	Star Wars épisode VIII	Pulp Fiction	Le fabuleux destin d'Amélie Poulain	Matrix	Il était une fois dans l'Ouest	Les Bronzés à TELECOM Nancy
Cathy	5	4.62	3.59	1.62	2.39	0.91
Stéphane	4	5	3.5	1.32	2.51	1
Sophie	1	0.17	0.35	0.31	0.13	0.03
Anna	12.93	3	4.93	4.05	2.07	0.57

On peut proposer "Star Wars" à Anna et "Pulp Fiction" à Cathy...