Chapitre 4

Problèmes de moindres carrés

4.1 Introduction

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

> On a un modèle $y = f(t, x)$ décrivant une relation fonctionnelle entre une variable d'entrée t et une variable de sortie y [1] mais ce modèle dépend de n paramètres inconnus $x_1, \ldots, x_n$ que l'on range dans le vecteur x. Dans les bons cas ce modèle est issu de lois scientifiques, i.e. on sait que le phénomène décrit doit [2] effectivement suivre la relation $y = f(t, x)$, dans d'autres cas le modèle n'est pas vraiment guidé par des lois physiques. Pour estimer 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 [3] mais les mesures y_i sont entachées d'erreur. On cherche alors les paramètres x de sorte que :
>
> $$\forall i \in [\![1, m]\!], \quad f(t_i, x) \simeq y_i, \text{ soit } e_i(x) := f(t_i, x) - y_i \simeq 0.$$
>
> Si on note $e(x)$ le vecteur rassemblant les m erreurs $e_i(x)$ alors il est naturel [4] 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) qui s'énonce donc : "trouver $x \in \mathbb{R}^n$ qui minimise" :
>
> $$E(x) = \|e(x)\|_2^2 = \sum_{i=1}^{m} e_i(x)^2 = \sum_{i=1}^{m} (f(t_i, x) - y_i)^2 \tag{4.1}$$
>
> 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 :

— Le plus célèbre : la droite $y = at + b$. 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 $f(t, x) = x_1\varphi_1(t) + x_2\varphi_2(t)$.

1. Pour le moment on considèrera t et y comme des scalaires réels mais ils peuvent être des vecteurs.
2. Si on ne s'est pas trompé en élaborant le modèle...
3. Ou suffisamment précis pour être considérés comme exacts.
4. On rappelle que la norme d'un vecteur peut s'interpréter comme une longueur de ce vecteur.

— Le modèle cherché pourrait être un polynôme du second degré $y = at^2 + bt + c$ ou plutôt $y = a(t - \bar{t})^2 + b(t - \bar{t}) + c$ avec $\bar{t}$ l'entrée "moyenne" autour de laquelle on veut utiliser ce modèle (ou mieux une base de type Lagrange). Idem en posant $c := x_1$, $b := x_2$ et $a := x_3$, 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 $f(t, x) = x_1\varphi_1(t) + x_2\varphi_2(t) + x_3\varphi_3(t)$.

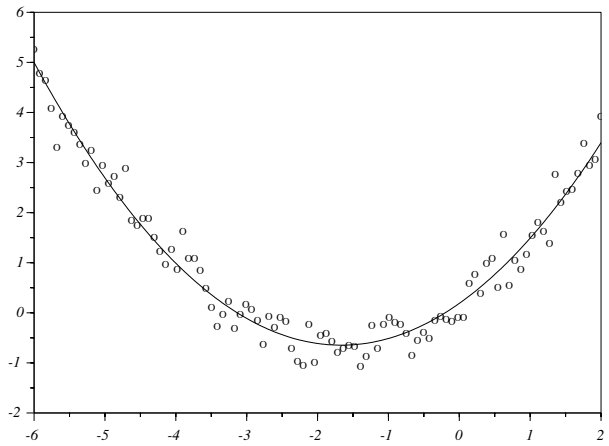

FIGURE 4.1 – Approximation par un polynôme du second degré de données bruitées

— 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) :

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

Si la période T est connue, ce modèle est aussi, comme les deux précédents, de la forme $\sum_j \varphi_j(t)x_j$. On dit alors qu'il est linéaire en ses paramètres x_i. Par contre si la période T est inconnue, ce paramètre supplémentaire (appelé x_6 si on se limite à l'harmonique 1) intervient de manière non linéaire dans le modèle.

— Un modèle souvent rencontré en physique est le modèle exponentiel décroissant $y = Ce^{-\frac{t}{\tau}}$ comportant les deux paramètres 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).

Comme il a été dit avant, les mesures y_i sont souvent entachées d'erreur. Si le modèle décrit bien le phénomène observé, les y_i peuvent être considérées comme la réalisation de variables aléatoires $Y_i = f(t_i, x^*) + B_i$ où $f(t_i, x^*)$ serait donc la valeur exacte (correspondant aux paramètres optimaux x^*) plus un bruit B_i qui suit (approximativement) la loi normale centrée et de variance σ_i^2 (c'est à dire que $B_i \sim N(0, \sigma_i^2)$). Cela veut dire aussi qu'on suppose que les mesures expérimentales ne présentent pas de biais (elles sont justes en moyenne, $E[B_i] = 0$) 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 :

$$E_w(x) = \sum_{i=1}^{m} \left(\frac{f(t_i, x) - y_i}{\sigma_i} \right)^2 \tag{4.2}$$

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

— 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.

Mathématiquement, 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 (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).

Notons que, 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 $y_i^{(k)}, k = 1, \ldots, p$ alors on utilisera les estimations classiques de la moyenne et de l'écart type :

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

Souvent l'appareillage de mesure apporte une indication sur la précision qui peut aussi être utilisée pour obtenir ces σ_i [5].

4.2 Modèles linéaires en leurs paramètres

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

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

où les n fonctions φ_j sont données (suite à la phase de modélisation). Dans ce cas 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'erreur E que nous développons en utilisant la linéarité du modèle :

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

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 :

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

5. Par exemple si l'appareil assure une précision de 1% sur toute la plage des mesures observées, on prendra $\sigma_i = 0.01|y_i|$

Ainsi, si on introduit la matrice A de dimension (m, n) et telle que $a_{ij} = \varphi_j(t_i)$, on remarque que sa i ème ligne est égale à $A_i = \left[\begin{array}{cccc} \varphi_1(t_i) & \varphi_2(t_i) & \dots & \varphi_n(t_i) \end{array} \right]$ d'où :

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

et on obtient l'expression suivante pour la fonction erreur :

$$E(x) = \sum_{i=1}^{m} ((Ax)_i - y_i)^2 = \sum_{i=1}^{m} ((Ax - b)_i)^2 = \|Ax - b\|_2^2. \tag{4.3}$$

où on a posé $b := [y_1, y_2, \dots, y_m]^\top$.

Remarque : si les erreurs sur les mesures sont connues, il est facile de transformer (4.2) en (4.1). On obtient alors pour $E_w(x)$ la même expression que ci-dessus mais avec une matrice A de coefficients $a_{i,j} = \frac{\varphi_j(t_i)}{\sigma_i}$ (au lieu de $a_{i,j} = \varphi_j(t_i)$) et un vecteur b de composantes $\frac{y_i}{\sigma_i}$ (au lieu de y_i), cf TD.

4.3 Aspects mathématiques

Idéalement, c'est à dire avec le bon modèle et aucune erreur de mesure, il faudrait que $Ax^* = b$. On retombe donc sur un système linéaire mais celui-ci comporte en général plus d'équations que d'inconnues (on a fait l'hypothèse que $m \geq n$). De plus du fait des erreurs dans les mesures il faut disposer de beaucoup plus de mesures que le modèle ne contient de paramètres (soit $m >> n$) pour espérer obtenir une bonne précision (aspect statistique de la méthode).

Evoquons rapidement le cas $m = n$. Celui-ci consiste alors à résoudre un système carré. Si $\det(A) \neq 0$, on a une solution unique $x^* = A^{-1}b$ que l'on peut obtenir via une factorisation de la matrice. De plus on aura $f(t_i, x^*) = y_i, \forall i$: on est revenu à un problème d'interpolation ! Cela peut sembler intéressant mais sauf si les mesures sont très précises la solution obtenue ne sera pas satisfaisante. Pour s'en rendre compte retournez voir la figure (4.1) et imaginez que l'on fasse passer une parabole par 3 points de l'échantillon choisi au hasard : il y a peu de chance qu'elle soit adéquate pour l'ensemble des points de mesure apparaissant dans la figure !

Pour un système quelconque $Ax = b$ avec $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, l'algèbre linéaire (cf chap 2) nous dit qu'on a existence d'une solution si $b \in ImA$. Dans le cas $m > n$, comme ImA [6] est de dimension inférieure ou égale à n il a peu de chance que le second membre b appartienne à l'image de A d'où la recherche d'une solution au sens des moindres carrés, c'est à dire celle qui va minimiser :

$$E(x) = \|Ax - b\|_2^2$$

Avant d'aller plus loin, observons que si l'hypothèse $KerA = \{0\}$ sera le plus souvent vraie [7] dans le cas contraire on aura $dim(KerA) \geq 1$ et si x^* est une solution au sens des moindres carrés alors $x^* + v$ où $v \in KerA$ sera aussi solution :

$$E(x^* + v) = \|A(x^* + v) - b\|_2^2 = \|Ax^* + \underbrace{Av}_{=0} - b\|_2^2 = \|Ax^* - b\|_2^2 = E(x^*)$$

et il y aura donc une infinité de solutions.

6. On a chap 2 on a vu que $ImA := \{y \in \mathbb{R}^m : \exists x \in \mathbb{R}^n \text{ tel que } y = Ax\}$ est un s.e.v. vectoriel de $\mathbb{R}^m$ dont une famille génératrice est constituée des colonnes de A : $\mathcal{A} := (A^1, \dots, A^n)$. Si $\mathcal{A}$ est libre, elle constitue alors une base de ImA d'où $dim(ImA) = n$. Cette hypothèse sera vraie la plupart du temps : n vecteurs choisis au hasard dans un espace de dimension $m \geq n$ seront généralement linéairement indépendants ; ici A n'est pas choisie au hasard mais à partir de fonctions φ_i qui permettent d'assurer généralement cette hypothèse.

7. Si $dim(ImA) = n$ alors $dim(KerA) = 0$ par le théorème du rang ($dim(KerA) + dim(ImA) = dim(\mathbb{R}^n) = n$).

4.3.1 Solution du problème par résolution des équations normales

Plusieurs méthodes pour résoudre le problème :

$$\min_{x\in\mathbb{R}^n} E(x) = \|Ax - b\|_2^2 \tag{4.4}$$

sont basées sur le résultat suivant.

Théorème 1 : Soit E un espace vectoriel sur $\mathbb{R}$ de dimension m et S un sous espace vectoriel de E de dimension n (donc $n \leq m$).E est muni d'un produit scalaire (noté $(.|.)$), on note $\|.\|$ la norme associée ($\|x\| = \sqrt{(x|x)}$). Le problème :

$$\min_{y\in S} \|y - b\|^2$$

où $b \in E$ est donné, admet une solution unique y^*, appelée projection orthogonale de b sur S et le vecteur résidu $r = y^* - b$ appartient au sous espace vectoriel orthogonal à S (noté $S^\perp$), c-a-d $(y^* - b|v) = 0,\ \forall v \in S$.

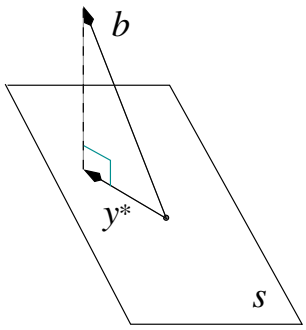

FIGURE 4.2 – Projection orthogonale sur un sous-espace

Démonstration : soit $\mathcal{S} = (s^1, s^2, \ldots, s^n)$ une b.o.n. (base orthonormée) de S. On la complète par $m-n$ vecteurs pour former une b.o.n. de E notée $\mathcal{E} = (s^1, \ldots, s^n, s^{n+1}, \ldots, s^m)$ (il est clair que les $m-n$ derniers vecteurs forment une b.o.n. de $S^\perp$). Tout vecteur $y \in S$ s'écrit de manière unique $y = \sum_{i=1}^n y_i s^i$ et le vecteur b a aussi une décomposition unique sur la base $\mathcal{E}$: $b = \sum_{i=1}^m b_i s^i$. Développons le terme $\|y - b\|^2$:

$$\begin{aligned}
\|y - b\|^2 &= (y - b|y - b) \\
&= (\sum_{i=1}^{n}(y_i - b_i)s^i - \sum_{i=n+1}^{m} b_i s^i | \sum_{i=1}^{n}(y_i - b_i)s^i - \sum_{i=n+1}^{m} b_i s^i) \\
&= \sum_{i=1}^{n}(y_i - b_i)^2 + \sum_{i=n+1}^{m} b_i^2
\end{aligned}$$

car comme $\mathcal{E}$ est une b.o.n. on a $(s^i|s^j) = \delta_{ij}$. Cette quantité est trivialement minimisée pour $y^* = \sum_{i=1}^n b_i s^i$ et tout autre choix pour y conduit à :

$$\|y - b\|^2 > \|y^* - b\|^2.$$

D'autre part le vecteur résidu $r = -\sum_{i=n+1}^m b_i s^i$ est bien dans $S^\perp$.

Ce résultat s'applique à notre problème de la façon suivante : le rôle de y est joué par Ax, S par ImA et $E = \mathbb{R}^m$. Le théorème nous dit qu'à l'optimum, le résidu $Ax^* - b$ doit être orthogonal à ImA. Comme $ImA = Vect(A^1, \ldots, A^n)$ cela est équivalent à dire que :

$$(Ax^* - b|A^j) = 0 \iff (A^j)^\top (Ax^* - b) = 0, \forall j \in [\![1, n]\!]$$

Les n équations ainsi obtenues peuvent s'écrire de manière plus compacte par :

$$A^\top (Ax^* - b) = 0$$

en effet la j ème composante du vecteur $A^\top (Ax^* - b)$ est égale à $(A^j)^\top (Ax^* - b)$ (le vérifier). Cette équation vectorielle se récrit aussi :

$$A^\top Ax^* = A^\top b \tag{4.5}$$

et cette forme est appelée équations normales du problème de moindres carrés (4.4).

Avant de rentrer dans le détail des équations normales notons que, si on calcule une b.o.n. $\mathcal{S}$ de ImA, alors la solution s'obtient très facilement dans cette base en calculant des produits scalaires. Cette remarque est à la base de la très robuste méthode QR.

4.3.2 Détail sur les équations normales

Si on pose $G = A^\top A$ et $h = A^\top b$, le système linéaire (4.5) s'écrit donc $Gx^* = h$ où :

- G est une matrice carrée (n, n) ($G \in \mathbb{R}^{n \times n}$) car produit d'une matrice (n, m) (taille de $A^\top$) par une matrice (m, n) ;
- $h = A^\top b$ est un vecteur de $\mathbb{R}^n$.

Les équations normales sont donc un simple système linéaire carré : on a une unique solution x^* si et seulement si G est inversible et cette solution peut s'obtenir par la méthode de Gauss vue au chapitre 2. En fait on peut montrer qu'il existe toujours une solution car $h = A^\top b \in ImG$ (cf TD).

Définition : Une matrice carrée $M \in \mathbb{R}^{n \times n}$ est dite semi-définie positive lorsque $(Mx|x) = x^\top Mx \geq 0, \forall x \in \mathbb{R}^n$. Si $(Mx|x) = x^\top Mx > 0, \forall x \neq 0$ on dit que M est définie positive. *Remarque :* 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 $M^{\mathcal{S}} := (M + M^\top)/2$ (appelée symétrisée de M) est symétrique et telle que $(M^{\mathcal{S}}x|x) = (Mx|x)$.

Proposition 1 : Si $M \in \mathbb{R}^{n \times n}$ est définie positive alors elle est inversible.

Preuve : Comme M est carrée il suffit de montrer que $KerM = \{0\}$. Supposons le contraire, il existe donc $x \neq 0$ tel que $Mx = 0$. Pour ce vecteur non nul on a donc $(Mx|x) = (0|x) = 0$ ce qui contredit la définie positivité de M. □

Venons en maintenant à énoncer les propriétés de $G = A^\top A$:

Proposition 2 : G est symétrique et semi-définie positive. Si de plus l'hypothèse naturelle $KerA = \{0\}$ est vérifiée, G est définie positive (et donc inversible).

Preuve : G est symétrique, en effet :

$$G^\top = (A^\top A)^\top = A^\top (A^\top)^\top = A^\top A = G$$

(petit rappel $(AB)^\top = B^\top A^\top$), mais aussi semi définie positive, puisque :

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

on remarque aussi que :

$$(Gx|x) = 0 \Leftrightarrow \|Ax\|_2 = 0 \Leftrightarrow Ax = 0 \Leftrightarrow x \in KerA$$

(rappel : la norme d'un vecteur est nulle si et seulement si le vecteur est nul) et donc G est bien définie positive ssi $KerA = \{0\}$. $\square$

Remarque : on peut montrer que sur une matrice symétrique et définie positive 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 (cf feuille 2) qui est une adaptation de la factorisation LU pour une telle matrice.

En résumé, résoudre un problème de moindres carrés (dont le modèle est linéaire en ses paramètres) et résolvant les équations normales consiste en 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 on peut calculer la matrice A et le vecteur b ;
3. on peut alors calculer la matrice $G = A^\top A$ et le vecteur $h = A^\top b$;
4. puis résoudre le 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.

4.4 Écriture de E comme une forme quadratique

Cette partie a juste été évoquée en cours (j'ai tracé les 2 dessins de la fin et expliqué qu'une forme quadratique était la généralisation d'un polynôme du second degré).

La fonction $E : x \in \mathbb{R}^n \longmapsto \|Ax - b\|^2 \in \mathbb{R}$ se développe en écrivant la norme au carré comme un produit scalaire :

$$\begin{aligned} E(x) &= (Ax - b|Ax - b) \\ &= (Ax|Ax) - 2(Ax|b) + (b|b) \\ &= x^\top A^\top Ax - 2(A^\top b)^\top x + \|b\|^2 \\ &= x^\top Gx - 2h^\top x + c \end{aligned}$$

avec $G = A^\top A$, $h = A^\top b$ et $c = \|b\|^2$ une constante positive. Nous venons d'écrire E comme une forme quadratique. Nous avons vu précédemment les propriétés de G.

A quoi ressemble la fonction E : une forme quadratique est la généralisation d'un polynôme du second degré dans le cas multidimensionnel. Lorsque la matrice G est définie positive (cad lorsque $KerA = \{0\}$), on obtient en fait un paraboloïde dont "le sommet est dirigé vers le bas", ce sommet constituant le point où E est minimale. Ceci peut s'expliquer rigoureusement avec un peu d'algèbre linéaire. Toute matrice symétrique est diagonalisable dans une base orthonormée, c-a-d qu'il existe une matrice orthogonale P (ses vecteurs colonnes p^j sont les vecteurs propres de G et forment une b.o.n. de $\mathbb{R}^n$) telle que :

$$\Lambda = P^\top GP \Leftrightarrow G = P\Lambda P^\top$$

la matrice diagonale Λ étant formée des n valeurs propres λ_i (les valeurs propres multiples apparaissant autant de fois que leur multiplicité) de G. L'inverse d'une matrice orthogonale carrée s'obtient simplement en prenant la transposée, en effet :

$$(P^\top P)_{ij} = P_i^\top P^j = (p^i)^\top p^j = (p_j|p_i) = \delta_{ij}$$

(l'élément en position (i,j) du produit de 2 matrices AB est égal à $(AB)_{ij} = A_i B^j$). D'où le résultat attendu $P^\top P = I$. On passe maintenant dans la b.o.n. des vecteurs propres, c-a-d que l'on va travailler avec $y = P^\top x$, il vient :

$$\begin{aligned} E(x) &= x^\top P\Lambda P^\top x - 2h^\top x + c \\ &= (P^\top x)^\top \Lambda P^\top x - 2h^\top PP^\top x + c \\ &= y^\top \Lambda y - 2(P^\top h)^\top y + c \\ &= y^\top \Lambda y - 2w^\top y + c \end{aligned}$$

où w est le vecteur h exprimé dans la base des vecteurs propres ($w = P^\top h$). Comme Λ est une matrice diagonale, l'expression obtenue pour E (en tant que fonction de y) est beaucoup plus simple :

$$E(x) = E(Py) = \hat{E}(y) = \sum_{i=1}^{n} (\lambda_i y_i^2 - 2w_i y_i) + c$$

Comme la matrice G est définie positive, ses valeurs propres sont strictement positives :

$$0 < (Gp^i|p^i) = \lambda_i (p^i|p^i) = \lambda_i$$

Continuons le développement de E :

$$\begin{aligned} \hat{E}(y) &= \sum_{i=1}^{n} (\lambda_i (y_i^2 - 2\frac{w_i}{\lambda_i} y_i)) + c \\ &= \sum_{i=1}^{n} \left(\lambda_i \left(y_i - \frac{w_i}{\lambda_i} \right)^2 - \frac{w_i^2}{\lambda_i} \right) + c \\ &= \sum_{i=1}^{n} \lambda_i \left(y_i - \frac{w_i}{\lambda_i} \right)^2 + \tilde{c} \end{aligned}$$

où $\tilde{c} = c - \sum_{i=1}^{n} w_i^2/\lambda_i$. Cette expression montre que le minimum de E (vue comme fonction de y) est atteind au point y^* tel que :

$$y_i^* = \frac{w_i}{\lambda_i},\ i = 1, 2, \ldots, n \Leftrightarrow y = \Lambda^{-1} w = \Lambda^{-1} P^\top h$$

et donc au point :

$$x^* = P\Lambda^{-1} P^\top h$$

en tant que fonction de x. Comme le calcul des valeurs propres et vecteurs propres est généralement beaucoup plus difficile (et cher) que la solution d'un problème de moindres carrés, on n'utilise pas cette méthode. On vient quand même de montrer que la solution existe et est unique. En effet tout autre choix pour y et donc pour x conduit à une valeur de la fonction telle que :

$$\hat{E}(y) = \underbrace{\sum_{i=1}^{n} \lambda_i \left(y_i - \frac{w_i}{\lambda_i} \right)^2}_{>0} + \tilde{c} > E(x^*) = \tilde{c}.$$

Mais revenons à notre paraboloïde, est-ce bien ce que nous venons d'obtenir ? Oui car la matrice orthogonale P peut être choisie telle que $\det(P) = 1$ (le déterminant d'une matrice orthogonale est toujours égal à 1 ou -1 et si le déterminant est égal à -1 il suffit de permuter deux colonnes pour obtenir 1) et le changement de base $y = P^\top x$ est donc une rotation généralisée dans l'espace $\mathbb{R}^n$, c'est à dire que la forme de l'objet constitué par E n'a pas été modifiée. Maintenant en procédant à des coupes $\hat{E}(y) = Cte$, on obtient :

1. l'ensemble vide si $Cte < \tilde{c}$ la valeur minimale de E ;
2. le point y^* pour $Cte = \tilde{c}$,
3. un ellipsoïde (une ellipse si $n = 2$) pour $Cte > \tilde{c}$ (à faire en exercice),

ce qui correspond bien (du moins pour $n = 2$) à un paraboloïde dont le sommet est dirigé vers le bas ! Et que se passe-t-il si $dim(KerA) \geq 1$? Et bien 0 est alors valeur propre de G et en dimension 2 (en considérant une valeur propre nulle et l'autre strictement positive) on obtient un cylindre parabolique (cf dessin 4.3) (il y a alors une infinité de solutions correspondant à la droite D sur le dessin).

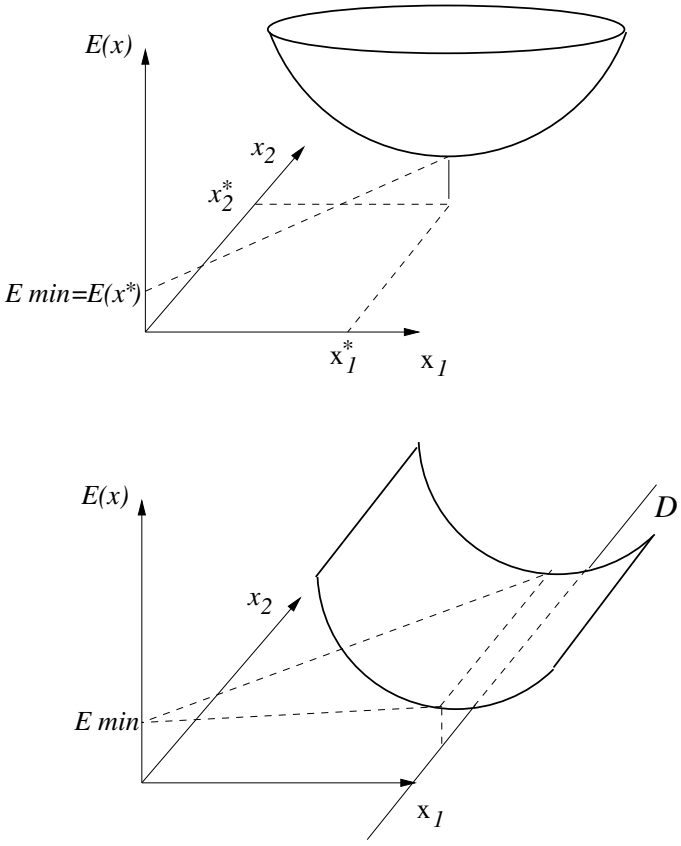

FIGURE 4.3 – Formes quadratiques : le cas λ_1, $\lambda_2 > 0$ et le cas $\lambda_1 > 0$, $\lambda_2 = 0$