Algorithmes numériques
pour
l'algèbre linéaire

Version 1.0

8 janvier 2009

Michel Llibre

Ref. DCSD-2009_008-NOT-001-1.0

isae
Institut Supérieur de l'Aéronautique et de l'Espace

ONERA
THE FRENCH AEROSPACE LAB

Table des matières

Chapitre 1

Introduction

Cette note présente les principaux algorithmes numériques utilisés en algèbre linéaire.

Le deuxième chapitre présente des *rappels sur quelques transformations linéaires* :

— projections orthogonales sur les sous-espaces image et noyau,
— symétrie par rapport à un hyperplan (transformations de Householder),
— rotation dans un plan de la base (rotation de Givens),
— opérations élémentaires sur les matrices.

Le troisième chapitre présente des *algorithmes numériques de résolution des systèmes linéaires* :

— factorisations de Cholesky historique et $\mathbf{U}^T\mathbf{DU}$,
— factorisations **LU** (Doolittle et Crout),
— factorisations **QR** et **QRE**,
— décomposition en valeur singulière (SVD) $\mathbf{UDV}^T$,

et l'utilisation de ces algorithmes pour résoudre des systèmes ou pour calculer le noyau d'une matrice.

Le quatrième chapitre présente des algorithmes utilisés pour calculer *les valeurs et vecteurs propres d'une matrice* :

— les formes de Hessenberg d'une matrice,
— les algorithmes LR, QR et HQR.

Le dernier chapitre présente diverses méthodes pour calculer une *exponentielle de matrice*.

Cette note est à rapprocher de la note "*Résolution de systèmes linéaires, Moindres carrés récursifs et Filtre de Kalman discret*" (Ref. DCSD-2008_069-NOT-001-1.0 qui présente des méthodes de résolution qui s'appuient sur ces algorithmes.

Chapitre 2

Rappels sur quelques transformations linéaires

Les problèmes linéaires mettent en jeu des transformations linéaires que nous étudions à l'aide du calcul matriciel. Sauf spécification contraire on suppose, dans tout ce qui suit, que tous les scalaires sont des réels.

2.1 Définitions

2.1.1 Espace image

Considérons une matrice $\mathbf{A}$ de dimension $m \times n$ qui transforme un vecteur $\mathbf{x}$ de $\mathbb{R}^n$ en un vecteur $\mathbf{y}$ de $\mathbb{R}^m$: $\mathbf{y} = \mathbf{A}\mathbf{x}$ et notons $\mathbf{a}_j$ pour $j = 1$ à n les n vecteurs colonnes de $\mathbf{A}$ (ce sont des vecteurs de $\mathbb{R}^m$). $\mathbf{y}$ s'écrit :

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \sum_{j=1}^{n} x_j \mathbf{a}_j$$

Lorsque $\mathbf{x}$ engendre toutes les directions possibles de l'espace $\mathbb{R}^n$, le vecteur $\mathbf{y}$ est limité au sous-espace de $\mathbb{R}^m$ engendré par les n vecteurs $\mathbf{a}_i$, appelé *espace image* de $\mathbf{A}$ et noté $\mathrm{Im}(\mathbf{A})$.

2.1.2 Rang

Le *rang* de la matrice $\mathbf{A}$ est la dimension de son espace image :

$$\mathrm{rang}(\mathbf{A}) = \dim(\mathrm{Im}(\mathbf{A}))$$

On a forcément $\mathrm{rang}(\mathbf{A}) \leq \min(m,n)$ puisque l'espace $\mathrm{Im}(\mathbf{A})$ n'est engendré que par n vecteurs de $\mathbb{R}^m$.

Comme le rang est également l'ordre du plus grand des mineurs non nul que l'on peut extraire de $\mathbf{A}$, on a :

$$\mathrm{rang}(\mathbf{A}) = \mathrm{rang}\left(\mathbf{A}^T\right) = \dim\left(\mathrm{Im}\left(\mathbf{A}^T\right)\right)$$

2.1.3 Noyau

Un vecteur $\mathbf{k}$ de $\mathbb{R}^n$ appartient au sous-espace noyau de $\mathbf{A}$, noté $\ker(\mathbf{A})$ si :

$$\mathbf{A}\mathbf{k} = \mathbf{0}$$

Les m vecteurs lignes $\mathbf{l}_i$ (appartenant à $\mathbb{R}^n$) de $\mathbf{A}$, engendrent un sous-espace de dimension $\text{rang}(\mathbf{A})$. Les vecteurs $\mathbf{k}$ appartiennent au sous-espace orthogonal à $\text{Im}\left(\mathbf{A}^T\right)$ dans $\mathbb{R}^n$, d'où :

$$n = \dim(\ker(\mathbf{A})) + \text{rang}(\mathbf{A})$$

2.1.4 Rang complet

On dit que $\mathbf{A}$ est de *rang complet* si $\text{rang}(\mathbf{A}) = n$, ce qui suppose donc que $n \leq m$ et implique $\text{rang}(\mathbf{A}) = \min(m,n)$.

2.1.5 Noyau orthonormé

La matrice $\mathbf{N}$ est un noyau orthonormé de A, si :

1. $\mathbf{AN} = \mathbf{0}$ (vecteurs du noyau)
2. $\mathbf{N}^T\mathbf{N} = \mathbf{I}$ (orthonormés)
3. $\text{rang}(\mathbf{N}) = \dim(\ker(\mathbf{A}))$ (noyau complet)

2.2 Projections

On rappelle que les matrices de projections $\mathbf{P}$ sont idempotentes et symétriques :

$$\mathbf{P}^2 = \mathbf{P} \text{ et } \mathbf{P} = \mathbf{P}^T$$

2.2.1 Projection orthogonale sur le sous-espace image

Quel que soit $\mathbf{x}$, les vecteurs $\mathbf{y}$ générés par $\mathbf{y} = \mathbf{Ax}$ appartiennent à $\text{Im}(\mathbf{A})$ et les vecteurs $\mathbf{z}$ générés par $\mathbf{z} = \mathbf{A}^T\mathbf{y} = \mathbf{A}^T\mathbf{Ax}$ appartiennent à $\text{Im}\left(\mathbf{A}^T\right)$. Si $\mathbf{A}$ est de rang complet, on a $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$, soit $\mathbf{y} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$. Posons :

$$\mathbf{P} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$$

On a $\mathbf{P} = \mathbf{P}^T$ et $\mathbf{P}^2 = \mathbf{P}$. C'est donc une matrice de projection telle que $\mathbf{Py} = \mathbf{y}$ quand $\mathbf{y} \in \text{Im}(\mathbf{A})$. Si $\mathbf{y} \notin \text{Im}(\mathbf{A})$, notons $\mathbf{z} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$ son transformé et calculons son produit scalaire avec les vecteurs du noyau orthonormé $\mathbf{N}$ de $\mathbf{A}$. Il vient :

$$\mathbf{z}^T\mathbf{N} = \mathbf{y}^T\mathbf{A}^T\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{AN} = \mathbf{0}$$

La matrice $\mathbf{P}$ est donc la matrice de projection orthogonale sur le sous-espace image de $\mathbf{A}$ (sous l'hypothèse que $\mathbf{A}$ soit de rang complet).

Projection orthogonale sur un vecteur : Si $\mathbf{A} = \mathbf{v}$ est constituée par un seul vecteur, on retrouve en $\mathbf{P} = \mathbf{v}\left(\mathbf{v}^T\mathbf{v}\right)^{-1}\mathbf{v}^T = \mathbf{uu}^T$ avec $\mathbf{u} = \dfrac{1}{\|\mathbf{v}\|}\mathbf{v}$ la projection orthogonale sur le vecteur unitaire $\mathbf{u}$ porté par $\mathbf{v}$.

$$\mathbf{P}_u = \mathbf{uu}^T$$

2.2.2 Projection orthogonale sur le sous-espace noyau

La matrice :

$$\Pi = \mathbf{I} - \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$$

réalise la projection orthogonale sur le sous-espace noyau de $\mathbf{A}$. Si $\mathbf{N}$ est un noyau orthonormé de $\mathbf{A}$, on a également :

$$\Pi = \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T = \mathbf{N}\mathbf{N}^T$$

d'où la relation :

$$\mathbf{N}\mathbf{N}^T + \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T = \mathbf{I}$$

Projection orthogonale sur un hyperplan : La projection orthogonale sur l'hyperplan (le noyau) orthogonal à $\mathbf{u}$ s'écrit comme dans $\mathbb{R}^3$:

$$\Pi_{\perp\mathbf{u}} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$$

2.3 Symétrie par rapport à un hyperplan

On peut étendre à $\mathbb{R}^n$ le raisonnement fait dans $\mathbb{R}^3$. Si $\mathbf{S}_u$ réalise une symétrie par rapport à l'hyperplan orthogonal au vecteur unitaire $\mathbf{u}$, telle que $\mathbf{w} = \mathbf{S}_u\mathbf{v}$ alors $\mathbf{w}+\mathbf{v}$ est égal au double de la projection sur l'hyperplan orthogonal à $\mathbf{u}$, d'où $\mathbf{S}_u\mathbf{v}+\mathbf{v} = 2\left(\mathbf{I} - \mathbf{u}\mathbf{u}^T\right)\mathbf{v}$ ce qui implique :

$$\mathbf{S}_u = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$$

qui est symétrique ($\mathbf{S}_u = \mathbf{S}_u^T$), involutive ($\mathbf{S}_u^2 = \mathbf{I}$) et par conséquent unitaire ($\mathbf{S}_u^{-1} = \mathbf{S}_u^T$).

La symétrie $\mathbf{S}_u$ est utilisée dans les factorisations QR (cf. 3.5) comme transformation orthogonale d'une matrice $\mathbf{A}$ d'éléments $a_{i,j}$ pour annuler tous les termes de la colonne j_p en dessous de l'élément de la ligne i_p par une pré-multiplication , ou pour annuler tous les termes de la ligne i_p, à droite de l'élément de la colonne j_p par une post-multiplication, sans toucher, dans les deux cas, les éléments des lignes $i < i_p$ et des colonnes $j < j_p$. Dans ce cas on la nomme transformation de Householder .

2.3.1 Transformation de Householder à gauche

On cherche à annuler les termes a_{i,j_p} de la colonne j_p, pour $i > i_p$, par pré-multiplication par $\mathbf{S}_u$.

Posons :

$$\mathbf{S}_u = \mathbf{I} - \frac{1}{h}\mathbf{v}\mathbf{v}^T \text{ avec } \|\mathbf{v}\|^2 = 2h$$

Notons v_i les composantes du vecteur $\mathbf{v}$, $\mathbf{a}_j$ les vecteurs colonnes de la matrice $\mathbf{A}$ et $\mathbf{a}'_j$ ceux de la matrice

$$\mathbf{A}' = \mathbf{S}_u\mathbf{A}$$

Il vient :

$$\mathbf{a}'_j = \mathbf{a}_j - s\mathbf{v} \text{ avec } s = \frac{1}{h}\mathbf{v}^T\mathbf{a}_j$$

Pour que les éléments des lignes $i < i_p$ de $\mathbf{A}'$ soient inchangés, il suffit de choisir :

$$v_i = 0 \text{ pour } i < i_p$$

Pour que les éléments des lignes $i > i_p$ de la colonne j_p de $\mathbf{A}'$ soient nuls, il faut que :

$$a_{i,j_p} - v_i \left(\frac{1}{h} \mathbf{v}^T \mathbf{a}_{j_p} \right) = 0 \text{ pour } i > i_p$$

ce qui est obtenu en choisissant :

$$v_i = a_{i,j_p} \text{ pour } i > i_p \text{ et } \left(\frac{1}{h} \mathbf{v}^T \mathbf{a}_{j_p} \right) = 1$$

Les deux dernières inconnues v_{i_p} et h sont déterminées par les deux équations $\|\mathbf{v}\|^2 = 2h$ et $\mathbf{v}^T \mathbf{a}_{j_p} = h$. Posons :

$$R^2 = \sum_{i=i_p}^{m} a_{i,j_p}^2 \text{ avec } R > 0$$
$$r^2 = \sum_{i=i_p+1}^{m} a_{i,j_p}^2 = R^2 - a_{i_p,j_p}^2$$

Il vient :

$$v_{i_p} a_{i_p,j_p} + r^2 = h$$
$$v_{i_p}^2 + r^2 = 2h$$

En éliminant h, on trouve que v_{i_p} est solution de :

$$v_{i_p}^2 - 2 v_{i_p} a_{i_p,j_p} - r^2 = 0$$

D'où :

$$v_{i_p} = a_{i_p,j_p} + \varepsilon R$$
$$h = R \left(R + \varepsilon a_{i_p,j_p} \right) = \varepsilon R v_{i_p}$$

avec $\varepsilon = \pm 1$. Traditionnellement on choisit $\varepsilon = \text{signe}\left(a_{i_p,j_p}\right)$:

$$v_{i_p} = a_{i_p,j_p} + R \text{ signe}\left(a_{i_p,j_p}\right)$$

D'où :

$$h = \left(a_{i_p,j_p} + R \text{ signe}\left(a_{i_p,j_p}\right) \right) a_{i_p,j_p} + r^2 = R^2 + R \left| a_{i_p,j_p} \right|$$

soit :

$$h = R^2 + R \left| a_{i_p,j_p} \right|$$

Par ailleurs :

$$a'_{i_p,j_p} = a_{i_p,j_p} - v_{i_p} = -R \text{ signe}\left(a_{i_p,j_p}\right)$$

Remarques :

1) Si on désire que a'_{i_p,j_p} soit positif, c'est-à-dire $a'_{i_p,j_p} = R$, il faut choisir $\varepsilon = -1$:

$$v_{i_p} = a_{i_p,j_p} - R$$

Dans ce cas $\frac{1}{h} = 1/\left(R \left(R - a_{i_p,j_p} \right) \right)$ n'est pas calculable lorsque $R = a_{i_p,j_p}$. Mais comme les éléments que l'on désire annuler sont déjà nuls, il n'y a rien à faire, ou plus précisément : $\mathbf{S}_u = \mathbf{I}$, $\mathbf{A}' = \mathbf{A}$.

2) L'écriture $\mathbf{S}_u = \mathbf{I} - \frac{1}{h}\mathbf{v}\mathbf{v}^T$ à la place de $\mathbf{S}_u = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$ permet d'éviter la racine carrée du calcul de la norme de $\mathbf{v}$. Une autre économie est réalisée dans les algorithmes de triangularisation (décomposition QR ou SVD). L'opérateur $\mathbf{S}_u$ y est souvent écrit :

$$\mathbf{S}_u = \mathbf{I} - \frac{1}{w_{i_p}}\mathbf{w}\mathbf{w}^T$$

où w_{i_p} est la première composante non nulle de $\mathbf{w}$. En effet, en posant :

$$\mathbf{w} = \frac{1}{R \text{ signe}\left(a_{i_p,j_p}\right)}\mathbf{v}$$

la première composante non nulle de $\mathbf{w}$ s'écrit :

$$w_{i_p} = \frac{a_{i_p,j_p} + R \text{ signe}\left(a_{i_p,j_p}\right)}{R \text{ signe}\left(a_{i_p,j_p}\right)} = 1 + \frac{1}{R}\left|a_{i_p,j_p}\right|$$

et h devient $h' = \frac{1}{R^2}h$, soit :

$$h' = \frac{R^2 + R\left|a_{i_p,j_p}\right|}{R^2} = 1 + \frac{1}{R}\left|a_{i_p,j_p}\right| = w_{i_p}$$

A l'époque où les algorithmes QR (cf. 3.5) et SVD (cf. 3.8.1) ont été conçus, cette écriture a été adoptée car elle permettait d'économiser la mémorisation des h.

2.3.2 Transformation de Householder à droite

On cherche à annuler les termes a_{i,j_p} de la ligne i_p, pour $j > j_p$, par post-multiplication par $\mathbf{S}_u$ avec :

$$\mathbf{S}_u = \mathbf{I} - \frac{1}{h}\mathbf{v}\mathbf{v}^T \quad ; \quad \|\mathbf{v}\|^2 = 2h$$

Notons $\mathbf{a}_i^T$ les vecteurs lignes de la matrice $\mathbf{A}$ et $\mathbf{a}_i'^T$ ceux de la matrice

$$\mathbf{A}' = \mathbf{A}\mathbf{S}_u$$

Il vient :

$$\mathbf{a}_i'^T = \mathbf{a}_i^T - s\mathbf{v}^T \text{ avec } s = \frac{1}{h}\mathbf{a}_i^T\mathbf{v}$$

Pour que les éléments des colonnes $j < j_p$ de $\mathbf{A}'$ soient inchangés, il suffit de choisir :

$$v_j = 0 \text{ pour } j < j_p$$

Pour que les éléments des colonnes $j > j_p$ de la ligne i_p de $\mathbf{A}'$ soient nuls, il faut que :

$$a_{i_p,j} - sv_j = 0 \text{ pour } j > j_p \text{ avec } s = \frac{1}{h}\mathbf{a}_{i_p}^T\mathbf{v}$$

ce qui est obtenu en choisissant :

$$v_j = a_{i_p,j} \text{ pour } j > j_p \text{ et } s = 1$$

Les deux dernières inconnues v_{j_p} et h sont déterminées par les deux équations $\|\mathbf{v}\|^2 = 2h$ et $\mathbf{a}_{i_p}^T \mathbf{v} = h$. Comme précédemment, on pose :

$$R^2 = \sum_{j=j_p}^{n} a_{i_p,j}^2 \text{ avec } R > 0$$

et on trouve :

$$v_{j_p} = a_{i_p,j_p} \pm R$$

Traditionnellement on choisit :

$$v_{j_p} = a_{i_p,j_p} + R \text{ signe}\left(a_{i_p,j_p}\right)$$

D'où :

$$h = R^2 + R\left|a_{i_p,j_p}\right|$$

et :

$$a'_{i_p,j_p} = -R \text{ signe}\left(a_{i_p,j_p}\right)$$

Remarque : Si on désire que a'_{i_p,j_p} soit positif, c'est-à-dire $a'_{i_p,j_p} = R$, on choisit

$$v_{j_p} = a_{i_p,j_p} - R \rightarrow h = R\left(R - a_{i_p,j_p}\right)$$

Dans ce cas, si $R = a_{i_p,j_p}$ on fait $\mathbf{S}_u = \mathbf{I}$.

Remarque : Ici aussi, on peut utiliser l'écriture $\mathbf{S}_u = \mathbf{I} - \frac{1}{w_{ip}}\mathbf{w}\mathbf{w}^T$ pour éconiser la mémorisation des h.

2.4 Rotation dans un des plans de la base

Notons $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ la base orthonormée canonique de $\mathbb{R}^n$ dans laquelle sont exprimées toutes les composantes des vecteurs colonnes des vecteurs $\mathbf{x}$, $\mathbf{y}$ et des vecteurs colonnes des matrices $\mathbf{A}$ carrées considérées. La matrice réalisant une rotation d'angle θ des vecteurs du plan $\{\mathbf{e}_{k_1}, \mathbf{e}_{k_2}\}$ et laissant inchangés les vecteurs du sous-espace orthogonal s'écrit, par exemple dans le cas $n = 6$, $k_1 = 2$ et $k_2 = 4$:

$$\mathbf{G}_{k_1k_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & -s & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & s & 0 & 0 & c & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

avec $c = \cos(\theta)$, $s = \sin(\theta)$ et $\varepsilon = \pm 1$.

La matrice de rotation $\mathbf{G}$ est utilisée dans les factorisations QR ou SVD (cf. 3.8.1) comme transformation orthogonale d'une matrice $\mathbf{A}$ pour annuler un seul de ses termes, en ne modifiant que deux lignes, par une pré-multiplication, ou deux colonnes, par une post-multiplication. Dans ce cas on la nomme rotation de Givens .

Remarque : $\mathbf{G}_{k_1k_2}^{-1} = \mathbf{G}_{k_1k_2}^T = \mathbf{G}_{k_2k_1}$ et réciproquement $\mathbf{G}_{k_2k_1}^{-1} = \mathbf{G}_{k_2k_1}^T = \mathbf{G}_{k_1k_2}$

2.4.1 Rotation de Givens à gauche

Notons $\mathbf{a}_j$ les vecteurs colonnes de la matrice $\mathbf{A}$ et $\mathbf{a}'_j$ ceux de la matrice :

$$\mathbf{A}' = \mathbf{A}\mathbf{G}_{k_1k_2}$$

Seules les colonnes k_1 et k_2 sont modifiées :

$$\begin{aligned}\mathbf{a}'_j &= \mathbf{a}_j \text{ pour } j \neq k_1 \text{ et } i \neq k_2\\ \mathbf{a}'_{k_1} &= c\mathbf{a}_{k_1} + s\mathbf{a}_{k_2}\\ \mathbf{a}'_{k_2} &= -s\mathbf{a}_{k_1} + c\mathbf{a}_{k_2}\end{aligned}$$

Pour avoir :

$$a'_{i,k_2} = 0$$

on fait :

$$\begin{aligned}\rho &= \mu\sqrt{a_{i,k_1}^2 + a_{i,k_2}^2} \quad ; \quad \mu = \pm 1\\ c &= \frac{1}{\rho}a_{i,k_1} \quad ; \quad s = \frac{1}{\rho}a_{i,k_2}\end{aligned}$$

Il en résulte que :

$$a'_{i,k_1} = \rho$$

Traditionnellement, si $|a_{i,k_1}| > |a_{i,k_2}|$, on donne à ρ le signe de a_{i,k_1} et inversement si $|a_{i,k_2}| > |a_{i,k_1}|$, on donne à ρ le signe de a_{i,k_2}.

Lorsque $\rho = 0$, l'élément a_{i,k_2} étant déjà nul, il n'y a rien à faire. On prend alors $\mathbf{G}_{k_1k_2} = \mathbf{I}$, c'est-à-dire $c = 1$ et $s = 0$.

2.4.2 Rotation de Givens à droite

Notons $\mathbf{a}_i^T$ les vecteurs lignes de la matrice $\mathbf{A}$ et $\mathbf{a}_i'^T$ ceux de la matrice :

$$\mathbf{A}' = \mathbf{G}_{k_1k_2}^T\mathbf{A}$$

Seules les lignes k_1 et k_2 sont modifiées :

$$\begin{aligned}\mathbf{a}_i'^T &= \mathbf{a}_i^T \text{ pour } i \neq k_1 \text{ et } i \neq k_2\\ \mathbf{a}_{k_1}'^T &= c\mathbf{a}_{k_1}^T + s\mathbf{a}_{k_2}^T\\ \mathbf{a}_{k_2}'^T &= -s\mathbf{a}_{k_1}^T + c\mathbf{a}_{k_2}^T\end{aligned}$$

Pour avoir :

$$a'_{k_2,j} = 0$$

on fait :

$$\begin{aligned}\rho &= \mu\sqrt{a_{k_1,j}^2 + a_{k_2,j}^2} \quad ; \quad \mu = \pm 1\\ c &= \frac{1}{\rho}a_{k_1,j} \quad ; \quad s = \frac{1}{\rho}a_{k_2,j}\end{aligned}$$

Il en résulte que :

$$a'_{k_1,j} = \rho$$

Traditionnellement, si $\left|a_{k_1,j}\right| > \left|a_{k_2,j}\right|$, on donne à ρ le signe de $a_{k_1,j}$ et inversement si $\left|a_{k_2,j}\right| > \left|a_{k_1,j}\right|$, on donne à ρ le signe de $a_{k_2,j}$.

Lorsque $\rho = 0$, l'élément $a_{k_2,j}$ étant déjà nul, il n'y a rien à faire. On prend alors $\mathbf{G}_{k_1k_2} = \mathbf{I}$, c'est-à-dire $c = 1$ et $s = 0$.

2.5 Opérations élémentaires sur les matrices

Aux opérations élémentaires utilisées pour trianguliser, ou diagonaliser une matrice on peut associer des matrices de pré- ou post-multiplication de la matrice à transformer. D'une manière générale une opération portant sur des lignes se fait par une pré-multiplication et une opération portant sur des colonnes se fait par une post-multiplication. Etant donné que $\mathbf{A} = \mathbf{IA} = \mathbf{AI}$, si on fait subir à la matrice $\mathbf{I}$ (qui pré- ou post-multiplie $\mathbf{A}$) l'opération considérée, on obtient l'expression de la matrice qui réalise l'opération élémentaire désirée.

2.5.1 Permutations de lignes et colonnes

Notons $\mathbf{E}_{k_1,k_2}$ la matrice qui permute soit les lignes, soit les colonnes d'indices k_1 et k_2 de la matrice $\mathbf{A}$. Par la pré-multiplication, $\mathbf{E}_{k_1,k_2}\mathbf{A}$ permute les lignes k_1 et k_2 de la matrice. Par la post-multiplication, $\mathbf{A}\mathbf{E}_{k_1,k_2}$ permute les colonnes. Faisons subir cette opération à la matrice identité $\mathbf{I}$, pour obtenir l'expression de $\mathbf{E}_{k_1,k_2}$. On a par exemple dans le cas où $m = 6$, $k_1 = 2$ et $k_2 = 4$:

$$\mathbf{E}_{k_1,k_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Remarque : $\mathbf{E}_{k_1,k_2}$ est symétrique ($\mathbf{E}_{k_1,k_2} = \mathbf{E}_{k_1,k_2}^T$), involutive ($\mathbf{E}_{k_1,k_2}^2 = \mathbf{I}$) et par conséquent unitaire ($\mathbf{E}_{k_1,k_2}^{-1} = \mathbf{E}_{k_1,k_2}^T$).

2.5.2 Multiplication d'une ligne ou colonne

Notons $\mathbf{K}_{k,\mu}$ la matrice qui multiplie la k-ième ligne ou colonne de $\mathbf{A}$ par μ. Par la pré-multiplication, $\mathbf{K}_{k,\mu}\mathbf{A}$ multiplie la ligne k de la matrice $\mathbf{A}$ par μ. Par la post-multiplication, $\mathbf{A}\mathbf{K}_{k,\mu}$ multiplie la colonne k de la matrice $\mathbf{A}$ par μ. On a par exemple dans le cas où $m = 5$, $k = 2$:

$$\mathbf{K}_{k,\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Remarque : $\mathbf{K}_{k,\mu}$ est symétrique ($\mathbf{K}_{k,\mu} = \mathbf{K}_{k,\mu}^T$) et $\mathbf{K}_{k,\mu}^{-1} = \mathbf{K}_{k,1/\mu}$

2.5.3 Cumulation de lignes ou colonnes

Notons $\mathbf{C}^l_{k_1,k_2,\mu}$ la matrice qui à la ligne k_1, ajoute la ligne k_2 multipliée par μ par la pré-multiplication $\mathbf{C}^l_{k_1,k_2,\mu}\mathbf{A}$. On a par exemple dans le cas où $m = 6$, $k_1 = 2$ et $k_2 = 4$:

$$\mathbf{C}^l_{k_1,k_2,\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \mu & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Notons $\mathbf{C}^c_{k_1,k_2,\mu}$ la matrice qui à la colonne k_1, ajoute la colonne k_2 multipliée par μ par la post-multiplication $\mathbf{AC}^c_{k_1,k_2,\mu}$. On peut vérifier que $\mathbf{C}^c_{k_1,k_2,\mu} = \mathbf{C}^{lT}_{k_1,k_2,\mu}$. On a par exemple dans le cas où $m = 6$, $k_1 = 2$ et $k_2 = 4$:

$$\mathbf{C}^c_{k_1,k_2,\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \mu & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2.5.4 Réduction en vecteur de base

Une opération classique dans la résolution des systèmes linéaires consiste à sélectionner un élément appelé pivot, ligne i_p, colonne j_p, soit a_{i_p,j_p}, le réduire à 1 par l'opération $\mathbf{K}_{j_p,1/a_{i_p,j_p}}$ (respectivement $\mathbf{K}_{i_p,1/a_{i_p,j_p}}$), puis à annuler les autres éléments de la ligne i_p par des opérations de cumulation $\mathbf{C}^c_{j,j_p,-a_{i_p},j}$ pour $j = 1,n$ et $j \neq j_p$ (resp. annuler les autres éléments de la colonne j_p par des opérations de cumulation $\mathbf{C}^l_{i,i_p,-a_i,j_p}$ pour $i = 1,m$ et $i \neq i_p$). Notons $\mathbf{T}^c_{i_p,j_p}$ l'opération qui réduit la ligne i_p au vecteur $(0,\ldots,1,\ldots,0)$ et $\mathbf{T}^l_{i_p,j_p}$ l'opération qui réduit la colonne j_p au vecteur $(0,\ldots,1,\ldots,0)^T$. Ces deux opérations s'écrivent :

$$\begin{aligned} \mathbf{T}^c_{i_p,j_p} &= \mathbf{K}_{j_p,1/a_{i_p,j_p}} \mathbf{C}^c_{1,j_p,-a_{i_p},1} \mathbf{C}^c_{2,j_p,-a_{i_p},2} \ldots \mathbf{C}^c_{n,j_p,-a_{i_p},n} \quad , \quad j \neq j_p \\ \mathbf{T}^l_{i_p,j_p} &= \mathbf{C}^l_{m,i_p,-a_m,j_p} \ldots \mathbf{C}^l_{2,i_p,-a_2,j_p} \mathbf{C}^l_{1,i_p,-a_1,j_p} \mathbf{K}_{i_p,1/a_{i_p,j_p}} \quad , \quad i \neq i_p \end{aligned}$$

A titre d'exemple, dans le cas $m = n = 5$, avec $i_p = 3$ et $j_p = 4$, on a :

$$\mathbf{T}^c_{3,4} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -\frac{a_{31}}{a_{34}} & -\frac{a_{32}}{a_{34}} & -\frac{a_{33}}{a_{34}} & \frac{1}{a_{34}} & -\frac{a_{35}}{a_{34}} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

et :

$$\mathbf{T}^l_{3,4} = \begin{pmatrix} 1 & 0 & -\frac{a_{14}}{a_{34}} & 0 & 0 \\ 0 & 1 & -\frac{a_{24}}{a_{34}} & 0 & 0 \\ 0 & 0 & \frac{1}{a_{34}} & 0 & 0 \\ 0 & 0 & -\frac{a_{44}}{a_{34}} & 1 & 0 \\ 0 & 0 & -\frac{a_{54}}{a_{34}} & 0 & 1 \end{pmatrix}$$

Attention, ces deux matrices ne sont pas transposées l'une de l'autre.

La matrice $\mathbf{T}^c_{3,4}$ est une matrice de manipulations de colonnes qui transforme la ligne d'indice $i_p = 3$ de la matrice $\mathbf{A}_{5\times5} = [a_{ij}]$ en $(0,0,0,1,0)$ par la post-multiplication $\mathbf{AT}^c_{3,4}$.

La matrice $\mathbf{T}^l_{3,4}$ est une matrice de manipulations de lignes qui transforme la colonne d'indice $j_p = 4$ de la matrice $\mathbf{A}_{5\times5} = [a_{ij}]$ en $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ par la pré-multiplication $\mathbf{T}^l_{3,4}\mathbf{A}$.

Chapitre 3

Algorithmes de résolution des systèmes linéaires

Considérons la résolution du système linéaire :

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

où $\mathbf{x}$ et $\mathbf{y}$ sont deux vecteurs respectivement de dimension n et m et $\mathbf{A}$ une matrice $m \times n$ de rang r.

$$n = \dim(\mathbf{x})$$
$$m = \dim(\mathbf{y})$$
$$r = \text{rang}(\mathbf{A})$$

Nous dirons que ce système est :

— déterminé si $m = n$,
— déterminé régulier si $m = n = r$
— sur-déterminé si $m > n$
— sur-déterminé régulier si $m > n = r$
— redondant (sous-déterminé) si $m < n$
— redondant régulier (sous-déterminé régulier) si $r = m < r$

Pour résoudre les systèmes déterminés réguliers nous présentons l'algorithme de factorisation LU. Si la matrice $\mathbf{A}$ est symétrique on utilisera avantageusement la factorisation de Cholesky sans racine carrée.

Les systèmes sur-déterminés réguliers n'ont pas de solution exacte (a priori). Parmi les solutions approchées, nous proposons des algorithmes pour calculer la solution qui minimise l'erreur quadratique $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$ à partir des factorisations de Cholesky et QR.

Les systèmes redondants réguliers ont une infinité de solution. Dans cette infinité nous proposons la solution de norme minimale (plus généralement celle qui minimise $\|\mathbf{x} - \mathbf{x}_0\|^2$) que l'on peut calculer à partir de la factorisations de Cholesky.

Les systèmes singuliers, ou plus précisément tels que $r < \min(m,n)$ mettent en échec toutes ces méthodes. Nous proposons deux algorithmes pour calculer la solution de norme minimale qui minimise l'erreur quadratique. L'un basé sur la factorisation QRE et l'autre basé sur la décomposition en valeurs singulières.

Généralement, on part de l'hypothèse que le système est régulier, ou plus précisément que $r = \min(m,n)$. Dans ce cas la résolution peut se faire par une des trois factorisations : Cholesky, LU ou QR.

Mais si l'hypothèse perd de sa crédibilité, nous suggérons d'utiliser la factorisation QRE qui présente l'avantage d'être configurable pour tous les cas.

Si la matrice $\mathbf{A}$ est carrée de dimension $n \times n$ avec $\det \mathbf{A} \neq 0$ alors $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. Le calcul de $\mathbf{A}^{-1}$ revient à résoudre n fois ce système avec pour $\mathbf{y}$ les n vecteurs de base.

3.1 Cas particuliers des matrices triangulaires

Considérons le cas des systèmes déterminés réguliers à matrice triangulaire.

3.1.1 Matrice triangulaire inférieure (Lower)

Si $\mathbf{A} = \mathbf{L}$ avec $l_{ij} = 0$ pour $i > j$, alors $\mathbf{y} = \mathbf{A}\mathbf{x}$ s'écrit :

$$y_i = \sum_{j=1}^{i} l_{ij} x_j \quad , \quad i = 1, n$$

La régularité implique que $l_{ii} \neq 0$ pour $i = 1, n$. Ces équations se résolvent en une boucle directe :

$$x_i = \frac{1}{l_{ii}} \left(y_i - \sum_{j=1}^{i-1} l_{ij} x_j \right) \quad , \quad i = 1, n \tag{3.1}$$

3.1.2 Matrice triangulaire supérieure (Upper)

Si $\mathbf{A} = \mathbf{U}$ avec $u_{ij} = 0$ pour $i < j$, alors $\mathbf{y} = \mathbf{A}\mathbf{x}$ s'écrit :

$$y_i = \sum_{j=i}^{n} u_{ij} x_j \quad , \quad i = 1, n$$

La régularité implique que $u_{ii} \neq 0$ pour $i = 1, n$. Ces équations se résolvent en une boucle rétrograde :

$$x_i = \frac{1}{l_{ii}} \left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \quad , \quad i = n, 1 \tag{3.2}$$

3.2 Factorisations de Cholesky

Considérons le cas des systèmes déterminés réguliers à matrice symétrique. La résolution classique utilise la factorisation de Cholesky.

3.2.1 Factorisation historique

Si $\mathbf{A}$ est *symétrique définie positive*, la résolution numérique peut être effectuée par la factorisation de la matrice $\mathbf{A}$ en un produit de deux matrices triangulaires sous la forme :

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

avec $l_{ij} = 0$ si $i > j$. Il en résulte que :

$$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk} \quad \text{pour } i \geq j$$

D'où :

$$\left.\begin{array}{c} l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2} \\ l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) \quad i = j+1, n \end{array}\right\} \quad j = 1, n$$

Après avoir calculé $\mathbf{L}$, on résout $\mathbf{Lz} = \mathbf{y}$ en une boucle directe , puis $\mathbf{L}^T\mathbf{x} = \mathbf{z}$ en une boucle rétrograde.

3.2.2 Factorisation $\mathbf{U}^T\mathbf{DU}$

Si $\mathbf{A}$ est symétrique définie positive ou négative (valeurs propres non nulles toutes de même signe), on peut utiliser une décomposition en un triple produit :

$$\mathbf{A} = \mathbf{U}^T\mathbf{DU}$$

où $\mathbf{U}$ est triangulaire supérieure à diagonale unité ($u_{ij} = 0$ pour $i < j$ et $u_{ii} = 1$) et $\mathbf{D}$ est diagonale. Il vient :

$$a_{ij} = d_i u_{ij} + \sum_{k=1}^{i-1} u_{ki} d_k u_{kj} \quad \text{pour } j \geq i$$

Il en résulte que :

$$\left.\begin{array}{c} d_i = a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2 d_k \\ u_{ij} = \frac{1}{d_i} \left(a_{ij} - \sum_{k=1}^{i-1} u_{ki} d_k u_{kj} \right) \quad j = i+1, n \end{array}\right\} \quad i = 1, n$$

On résout ensuite $\mathbf{U}^T\mathbf{z} = \mathbf{y}$ en une boucle directe (cf. 3.1) et $\mathbf{Ux} = \mathbf{D}^{-1}\mathbf{z}$ en une boucle rétrograde (cf. 3.2).

Remarque 1 : On peut vérifier que le mineur principal de $\mathbf{A}$ de niveau k est égal au produit des k premiers éléments d_i. Ainsi :

$$D_{1,2\ldots k-1,k}^{1,2\ldots k-1,k} = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1,k-1} & a_{1,k} \\ a_{12} & a_{22} & \ldots & a_{2,k-1} & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{1,k-1} & a_{2,k-1} & \ldots & a_{k-1,k-1} & a_{k-1,k} \\ a_{1,k} & a_{2,k} & \ldots & a_{k-1,k} & a_{k,k} \end{vmatrix} = d_1 d_2 \ldots d_{k-1} d_k$$

Il en résulte que si un des mineurs principaux est nul, c'est algorithme n'est pas applicable, car le d_i correspondant sera nul. Ce cas ne peut pas se produire lorsque les valeurs propres sont toutes de même signe, car dans ce cas les signes des mineurs principaux sont tous strictement positifs (cas définie positive) , soit de signes strictement alternés, avec le premier négatif (cas définie négative).

A titre d'exemple, la matrice symétrique :

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$$

a pour mineurs principaux :

$$D_1^1 = 1 = d_1$$
$$D_{12}^{12} = \begin{vmatrix} 1 & 2 \\ 2 & 4 \end{vmatrix} = 0 = d_1 d_2 \rightarrow d_2 = 0$$
$$D_{123}^{123} = \begin{vmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{vmatrix} = -1 = d_1 d_2 d_3$$

mais comme d_2 est nul, d_3 n'est pas défini et ne peut être calculé par la dernière identité. Ainsi, bien que la matrice **A** soit définie (son déterminant vaut -1) la factorisation de Cholesky ne peut lui être appliquée. On pourra vérifier qu'elle n'est ni positive, ni négative (valeurs propres : -0.52, 0.17 et 11.3).

A toutes fins utiles, les éléments u_{ik} sont donnés par :

$$u_{i,k} = \frac{D_{1,2\ldots i-1,k}^{1,2\ldots i-1,k}}{D_{1,2\ldots i-1,i}^{1,2\ldots i-1,i}}$$

où le mineur du numérateur n'est pas principal, c'est le déterminant de la matrice constituée des i premières lignes, avec les $i-1$ premières colonnes et la k-ième.

Remarque 2 : Une matrice symétrique provient souvent d'un produit du type :

$$\mathbf{A} = \mathbf{B}^T\mathbf{B}$$

Comme nous supposons que $|\mathbf{A}| \neq 0$, nous supposons que $\text{rang}(\mathbf{B}) = n$, ce qui implique le nombre m de lignes de **B** est au moins égal à n :

$$\mathbf{B}_{m\times n} \text{ avec } m \geq n$$

La forme quadratique $\mathbf{x}^T\mathbf{A}\mathbf{x}$ s'écrit :

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = \|\mathbf{B}\mathbf{x}\|^2$$

Elle est donc positive ou nulle (mais dans ce cas **B** ne serait pas de rang complet et son noyau contiendrait **x**).

Il en résulte que si **B** *est de rang complet, la matrice* $\mathbf{B}^T\mathbf{B}$ *est définie positive. On peut lui appliquer la décomposition de Cholesky.*

3.3 Les factorisations LU

Dans le cas d'un système déterminé régulier, on peut utiliser une décomposition en un produit de deux matrices triangulaires inférieure et supérieure :

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

soit :

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj}$$

Parmi les diverses factorisations possibles, les plus fréquemment utilisés sont les factorisations de Doolittle et de Crout.

3.3.1 Décomposition par la méthode de Doolittle

Elle impose $l_{ii} = 1$. Il en résulte que :

$u_{1k} = a_{1k}$ pour $k = 1,n$

$l_{k1} = a_{k1}/u_{11}$ pour $k = 2,n$

Puis pour $j = 2$ à n ($n-1$ pour les termes l)

$$u_{jk} = a_{jk} - \sum_{s=1}^{j-1} l_{js}u_{sk} \text{ pour } k = j,n$$
$$l_{kj} = \left(a_{kj} - \sum_{s=1}^{j-1} l_{ks}u_{sj}\right)/u_{jj} \text{ pour } k = j+1,n$$

3.3.2 Décomposition par la méthode de Crout

Elle impose $u_{ii} = 1$. Il en résulte que :

$l_{k1} = a_{k1}$ pour $k = 1,n$

$u_{1k} = a_{1k}/l_{11}$ pour $k = 2,n$

Puis pour $k = 2$ à n ($n-1$ pour les termes u)

$$l_{jk} = a_{jk} - \sum_{s=1}^{k-1} l_{js}u_{sk} \text{ pour } j = k,n$$
$$u_{kj} = \left(a_{kj} - \sum_{s=1}^{k-1} l_{ks}u_{sj}\right)/l_{kk} \text{ pour } j = k+1,n$$

Comme précédemment on résout en deux boucles, d'abord directe $\mathbf{Lz} = \mathbf{y}$ puis rétrograde $\mathbf{Ux} = \mathbf{z}$.

3.4 La solution des moindres carrés

Considérons le cas des systèmes réguliers ayant plus de conditions indépendantes que d'inconnues. Ils sont tels que :

$$m = \dim(\mathbf{y}) > \dim(\mathbf{x}) = n = \text{rang}(\mathbf{A})$$

On suppose donc que la matrice $\mathbf{A}$ est de rang complet. Dans ce cas, il n'y a, a priori, pas de solution. On cherche malgré tout le vecteur $\mathbf{x}$ qui produit une erreur $\varepsilon = \mathbf{y} - \mathbf{Ax}$ minimale. D'où le critère :

$$\mathbf{r} = \tfrac{1}{2}(\mathbf{y} - \mathbf{Ax})^T \mathbf{P}^{-1}(\mathbf{y} - \mathbf{Ax})$$

où $\mathbf{P}$ est une matrice régulière de pondérations.

$$\frac{\partial r}{\partial \mathbf{x}} = -\mathbf{A}^T\mathbf{P}^{-1}(\mathbf{y} - \mathbf{A}x) = 0 \rightarrow \mathbf{x} = \left(\mathbf{A}^T\mathbf{P}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{P}^{-1}\mathbf{y}$$

Cette solution est surtout employée avec $\mathbf{P}^{-1} = \mathbf{1}$, d'où la pseudo-inverse des moindre-carrés :

$$\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$$

On remarquera que cette solution inverse une matrice $n \times n$ (la plus petite des dimensions de $\mathbf{A}$). Elle n'existe que si $\mathbf{A}$ est de rang complet.

On peut calculer $\mathbf{x}$ à l'aide de la factorisation de Cholesky. On évaluera d'abord le vecteur $\mathbf{z} = \mathbf{A}^T\mathbf{P}^{-1}\mathbf{y}$, puis on résoudra $\left(\mathbf{A}^T\mathbf{P}^{-1}\mathbf{A}\right)\mathbf{x} = \mathbf{z}$ en utilisant la factorisation de Cholesky.

On peut calculer directement $\mathbf{x}$ par la factorisation QR présentée ci-après.

3.5 La factorisation QR

La factorisation QR est due aux travaux de Householder [4], Francis [2] et Wilkinson [7]. Cette première version, *également très utilisée pour les systèmes déterminés réguliers* (tels que $r = n = m$), fournit la solution des moindres carrés lorsque $r = n < m$. La matrice $\mathbf{A}$ de dimension $m \times n$, est mise sous la forme :

$$\mathbf{A} = \mathbf{QR} \Leftrightarrow \mathbf{R} = \mathbf{Q}^T\mathbf{A}$$

où $\mathbf{Q}$ est une matrice unitaire de dimension $m \times m$ ($\mathbf{Q}^{-1} = \mathbf{Q}^T$) et $\mathbf{R}$ une matrice triangulaire supérieure de dimension $m \times n$, avec forcément ses $m-n$ dernières lignes nulles.

$$\mathbf{R} = \begin{bmatrix} \mathbf{U} \\ 0 \end{bmatrix}$$

Comme $\mathbf{Q}$ ne modifie pas la norme des vecteurs, $\|\mathbf{y} - \mathbf{Ax}\| = \left\|\mathbf{Q}^T(\mathbf{y} - \mathbf{Ax})\right\| = \left\|\mathbf{Q}^T\mathbf{y} - \mathbf{Rx}\right\|$. Le vecteur $\mathbf{z} = \mathbf{Q}^T\mathbf{y}$ ne peut être réalisé par $\mathbf{Rx}$ qui a ses $m-n$ dernières composantes nulles. Seul le vecteur constitué par les n premières composantes de $\mathbf{z}$, noté $\mathbf{z}_n$ est réalisable par $\mathbf{Ux}$. La solution obtenue en résolvant $\mathbf{z}_n = \mathbf{Ux}$ par une boucle rétrograde est donc optimale au sens des moindres carrés.

Pour calculer $\mathbf{R}$ et $\mathbf{Q}^T$, la matrice $\mathbf{A}$ est pré-multipliée par m matrices de Householder $\mathbf{S}_k$ qui annulent à chaque fois les $k-1$ dernières composantes de la k-ième colonne de la matrice en cours de calcul.

En posant $\mathbf{A}_0 = \mathbf{A}$, on définit :

$$\mathbf{A}_1 = \mathbf{S}_1\mathbf{A}_0,\ \mathbf{A}_2 = \mathbf{S}_2\mathbf{A}_1,\ \ldots,\ \mathbf{A}_k = \mathbf{S}_k\mathbf{A}_{k-1},\ \ldots,\ \mathbf{U} = \mathbf{A}_n = \mathbf{S}_n\mathbf{A}_{n-1}$$

D'où :

$$\mathbf{Q}^T = \mathbf{S}_n\mathbf{S}_{n-1}\ldots\mathbf{S}_2\mathbf{S}_1$$

avec :

$$\mathbf{S}_k = \mathbf{I} - \frac{1}{h_k}\mathbf{v}_k\mathbf{v}_k^T$$

Il vient d'après le paragraphe 2.3.1 :

$$\begin{aligned} d_k^2 &= a_{k,k}^2 + a_{k+1,k}^2 + \ldots + a_{m,k}^2 \\ \mathbf{v}_k^T &= \begin{pmatrix} 0 & \ldots & 0 & v_{k,k} & a_{k+1,k} & \ldots & a_{m,k} \end{pmatrix} \\ v_{k,k} &= a_{k,k} + d_k \text{signe}\,(a_{k,k}) \\ h_k &= d_k\,(d_k + |a_{k,k}|) \end{aligned}$$

où les $a_{i,j}$ sont les éléments de la matrice $\mathbf{A}_{k-1}$ (ils devraient être notés $a_{i,j}^{(k-1)}$).

La matrice $\mathbf{Q}^T$ peut être calculée au fur et à mesure en faisant subir à une matrice identité, les mêmes transformations que celles qui sont appliquées à la matrice $\mathbf{A}$.

Dans le cas de la résolution de $\mathbf{y} = \mathbf{Ax}$, le calcul de $\mathbf{Q}^T$ n'est pas nécessaire, seul celui de $\mathbf{z} = \mathbf{Q}^T\mathbf{y} = \mathbf{S}_n\mathbf{S}_{n-1}\ldots\mathbf{S}_2\mathbf{S}_1\mathbf{y}$ l'est. Ce vecteur peut être calculé au fur et à mesure, sans réservation mémoire, en remplacement de $\mathbf{y}$. De même, la matrice $\mathbf{R}$ est calculée au fur et à mesure, sans réservation mémoire, en remplacement de la matrice $\mathbf{A}$. La résolution de $\mathbf{z} = \mathbf{Rx}$ peut se faire directement en remplaçant les composantes de $\mathbf{z}$ par les composantes calculées de $\mathbf{x}$.

Remarque : Les éléments diagonaux $r_{k,k}$ de $\mathbf{R}$ sont données à $r_{k,k} = -d_k$ signe $\left(a_{k,k}^{(k-1)}\right)$.

Si on désire réaliser une *factorisation* QR *à diagonale positive*, à savoir $r_{k,k} = +d_k$, il faut choisir $v_{k,k} = a_{k,k} - d_k$ et $h_k = d_k\,(d_k - a_{k,k})$.

3.6 La pseudo inverse de Moore-Penrose

Considérons le cas des systèmes réguliers ayant plus d'inconnues que d'équations. Ils sont tels que :

$$\text{rang}(\mathbf{A}) = m = \dim(\mathbf{y}) \leq \dim(\mathbf{x}) = n$$

On suppose donc que la matrice $\mathbf{A}$ est de rang complet. Si $m < n$ (strictement), il y a une infinité de solutions $\mathbf{x}$. Dans cette infinité on choisit celle qui minimise la norme $\|\mathbf{x} - \mathbf{x}_0\|$ avec généralement $\mathbf{x}_0 = 0$. D'où le critère à minimiser :

$$r = \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{P}^{-1}(\mathbf{x} - \mathbf{x}_0) + (\mathbf{y} - \mathbf{A}\mathbf{x})^T \lambda$$

λ est le vecteur des multiplicateurs de Lagrange associés aux contraintes $\mathbf{y} - \mathbf{A}\mathbf{x} = \mathbf{0}$ et $\mathbf{P}^{-1}$ est une matrice de pondérations choisie pour favoriser la proximité de certaines composantes par rapport à celle des autres. Les conditions de stationnarité du premier ordre donnent :

$$\begin{aligned} \frac{\partial r}{\partial \mathbf{x}} &= \mathbf{P}^{-1}(\mathbf{x} - \mathbf{x}_0) - \mathbf{A}^T \lambda = 0 \rightarrow \mathbf{x} = \mathbf{x}_0 + \mathbf{P}\mathbf{A}^T \lambda \\ \mathbf{y} &= \mathbf{A}\mathbf{x}_0 + \mathbf{A}\mathbf{P}\mathbf{A}^T \lambda \rightarrow \lambda = \left(\mathbf{A}\mathbf{P}\mathbf{A}^T\right)^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}_0) \end{aligned}$$

d'où :

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{P}\mathbf{A}^T \left(\mathbf{A}\mathbf{P}\mathbf{A}^T\right)^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}_0)$$

On remarquera que cette solution inverse une matrice $\dim(\mathbf{y}) \times \dim(\mathbf{y})$ (la plus petite des dimensions de $\mathbf{A}$).

Cette solution est surtout employée avec $\mathbf{P} = \mathbf{1}$ et $\mathbf{x}_0 = 0$, d'où la pseudo-inverse classique dite de Moore-Penrose :

$$\mathbf{x} = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1} \mathbf{y}$$

Elle n'existe que si $\mathbf{A}$ est de rang complet (cf. hypothèse).

Un raisonnement géométrique permet d'obtenir plus rapidement ce résultat. Décomposons $\mathbf{x}$ en deux parties $\mathbf{x}_e$ et $\mathbf{x}_n$:

$$\mathbf{x} = \mathbf{x}_e + \mathbf{x}_n \quad \text{avec} \quad \begin{cases} \mathbf{x}_e = \mathbf{A}^T \lambda \\ \mathbf{A}\mathbf{x}_n = 0 \end{cases}$$

$\mathbf{x}_e$ appartient aux sous-espace engendré par les lignes de $\mathbf{A}$, et $\mathbf{x}_n$ est orthogonal à ce sous-espace. Il en résulte que :

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}_e + \mathbf{x}_n) = \mathbf{A}\mathbf{A}^T \lambda$$

Si $\mathbf{A}$ est de rang complet, alors $\lambda = \left(\mathbf{A}\mathbf{A}^T\right)^{-1} \mathbf{y}$ d'où :

$$\mathbf{x}_e = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1} \mathbf{y}$$

La solution générale consiste à ajouter à la partie $\mathbf{x}_e$ produisant effectivement $\mathbf{y}$, un vecteur $\mathbf{x}_n$ dans le noyau de $\mathbf{A}$ dite solution homogène ou solution nulle car elle ne produit rien au niveau de $\mathbf{y}$. Si on note :

$$\mathbf{A}^\dagger = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1}$$

la pseudo-inverse de Moore-Penrose, la matrice $\mathbf{A}^\dagger\mathbf{A}$ est un projecteur orthogonal dans le sous-espace des lignes de $\mathbf{A}$ et la matrice :

$$\mathbf{P}_N = \mathbf{1} - \mathbf{A}^\dagger\mathbf{A}$$

est un projecteur dans le noyau de $\mathbf{A}$, d'où quelque soit le vecteur $\mathbf{x}'$ de l'espace des $\mathbf{x}$:

$$\mathbf{x}_n = \left(\mathbf{1} - \mathbf{A}^{\dagger}\mathbf{A}\right)\mathbf{x}'$$

Remarque : Pour résoudre ces systèmes, qui font toujours intervenir des matrices symétriques, on utilise souvent la factorisation de Cholesky.

3.7 La factorisation QRE

Les méthodes que nous examinons maintenant s'appliquent dans tous les cas, même lorsque $r = \text{rang}(\mathbf{A}) < \min(n,m)$.

Dans le pire des cas, le vecteur $\mathbf{y}$ n'appartient pas à l'espace engendré par les colonnes de $\mathbf{A}$. Le système $\mathbf{y} = \mathbf{Ax}$ n'est pas compatible. Seule la projection $\mathbf{y}_A$ de $\mathbf{y}$ dans l'espace image de $\mathbf{A}$ est réalisable et le système $\mathbf{y}_A = \mathbf{Ax}$ a une infinité de solutions. La factorisation QRE, variante de la factorisation QR, nous semble la méthode la mieux adaptée à la résolution de ce problème. Elle peut même fournir la solution de norme minimale (au prix de deux factorisations successives) pour laquelle on fait généralement appel à la décomposition en valeurs singulières. La factorisation QRE permet également de déterminer le rang de la matrice $\mathbf{A}$.

La matrice $\mathbf{A}$ est mise sous la forme :

$$\mathbf{A} = \mathbf{QRE}^T \Leftrightarrow \mathbf{R} = \mathbf{Q}^T\mathbf{AE}$$

où $\mathbf{E}$ est une matrice unitaire produit de matrices de permutations.

Pour le calcul de $\mathbf{Q}$ et $\mathbf{R}$, la procédure est quasiment identique à celle qui vient d'être exposée, à ceci près qu'avant chaque étape, on effectue une permutation de colonne de la matrice à traiter. Par exemple, au moment de traiter la k-ième colonne $\mathbf{A}_{k-1}$ on effectue une permutation des colonnes k et $j \geq k$ de $\mathbf{A}_{k-1}$ de manière à amener en k-ième colonne, la colonne de rang j qui a la norme maximale pour les lignes de rang $i \geq k$:

$$d_j^2 = \max_{l=k \text{ à } n}\left(d_l^2 = \sum_{i=k}^{m} a_{i,l}^2\right) \tag{3.3}$$

Si $d_j^2 = 0$, *la triangularisation est terminée. Le rang r de la matrice est égal à* $k-1$.

Dans le cas contraire, on poursuit l'algorithme : Notons $\mathbf{E}_k$ la matrice qui effectue la permutation par une post-multiplication de $\mathbf{A}_{k-1}$. C'est une matrice unitaire symétrique, telle que $\mathbf{E}_k = \mathbf{E}_k^T = \mathbf{E}_k^{-1}$ ($\mathbf{E}_k$ s'obtient en permutant les colonnes j et k d'une matrice identité).

Après la permutation, on effectue l'annulation de la partie basse de $\mathbf{A}_{k-1}\mathbf{E}_k$, avec une matrice $\mathbf{S}_k$ calculée comme pour la factorisation QR.

Il en résulte une matrice que :

$$\mathbf{A}_k = \mathbf{S}_k\mathbf{A}_{k-1}\mathbf{E}_k$$

Finalement, on a :

$$\mathbf{A} = \mathbf{QRE}^T \text{ avec } \begin{cases} \mathbf{R} = \mathbf{A}_r = \begin{pmatrix} \mathbf{U}_r & \mathbf{B} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ \mathbf{Q}^T = \mathbf{S}_r \ldots \mathbf{S}_2\mathbf{S}_1 \\ \mathbf{E} = \mathbf{E}_1\mathbf{E}_2 \ldots \mathbf{E}_r \end{cases}$$

avec :

— $r = \text{rang}(\mathbf{A}) \leq \min(m,n)$

— $\mathbf{U}_r$ matrice $r \times r$ triangulaire supérieure avec ses éléments diagonaux non nuls, décroissants en valeur absolue,
— $\mathbf{B}$ matrice quelconque de dimension $(m-r) \times (n-r)$,
— $\mathbf{Q}$ matrice unitaire $m \times m$,
— $\mathbf{E}$ matrice unitaire $n \times n$.

La résolution de $\mathbf{y} = \mathbf{Ax} = \mathbf{QRE}^T\mathbf{x}$ consiste à :

1. calculer $\mathbf{z} = \mathbf{Q}^T\mathbf{y} = \mathbf{S}_r \ldots \mathbf{S}_2\mathbf{S}_1\mathbf{y}$,
2. calculer $\mathbf{w}_r$ solution de $\mathbf{U}_r\mathbf{w}_r = \mathbf{z}_r$ où $\mathbf{z}_r$ ne comporte que les r premières composantes de $\mathbf{z}$,
3. former $\mathbf{w}$ à partir de $\mathbf{w}_r$ en ajoutant $n-r$ dernières composantes nulles,
4. calculer $\mathbf{x} = \mathbf{E}_1\mathbf{E}_2 \ldots \mathbf{E}_r\mathbf{w}$.

Dans le cas d'une simple résolution, $\mathbf{z}$ est calculé au fur et à mesure en lieu et place de $\mathbf{y}$ par les produits par $\mathbf{S}_k$. $\mathbf{R}$ est calculée en lieu et place de $\mathbf{A}$. $\mathbf{w}$ est calculé en lieu et place de $\mathbf{z}$ (c'est-à-dire de $\mathbf{y}$). Seuls les indices des permutations effectuées sont mémorisés et $\mathbf{x}$ est directement obtenu en appliquant la permutation aux composantes de $\mathbf{w}$.

Si $r < \min(n,m)$, la solution trouvée n'est pas de norme minimale, mais elle est exacte au sens que $\mathbf{Ax}$ est égal à la partie réalisable de $\mathbf{y}$ (la projection de $\mathbf{y}$ dans l'espace image de $\mathbf{A}^T$). Pour obtenir la solution de norme minimale, il faut retrancher à $\mathbf{x}$ sa projection sur le noyau de $\mathbf{A}$. Si $\mathbf{N}$ représente le noyau orthonormé de $\mathbf{A}$ (qui peut être calculé par une factorisation QRE de $\mathbf{A}^T$, cf. 3.9.1), la solution de norme minimale s'écrit :

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbf{NN}^T\mathbf{x} \tag{3.4}$$

3.8 La décomposition en valeurs singulières

3.8.1 Utilisation de la décomposition en valeurs singulières

Souvent notée SVD (Singular Value Decomposition), la décomposition en valeurs singulières de $\mathbf{A}$ s'écrit :

$$\mathbf{A} = \mathbf{UDV}^T$$

Si $\mathbf{A}$ est de dimension $m \times n$, alors $\mathbf{U}$ est une matrice unitaire $m \times m$, $\mathbf{V}$ une matrice unitaire $n \times n$ et $\mathbf{D}$ une matrice $m \times n$ de la forme :

$$\mathbf{D} = \begin{pmatrix} \Lambda & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{pmatrix}$$

où Λ est une matrice diagonale de dimension $r \times r$ dont les éléments diagonaux sont les valeurs singulières de la matrice $\mathbf{A}$, généralement classées de la plus grande en haut à gauche, à la plus petite en bas à droite. On peut vérifier que :

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{V} = \mathbf{V}\begin{pmatrix} \Lambda^2 & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{pmatrix}$$
$$\left(\mathbf{AA}^T\right)\mathbf{U} = \mathbf{U}\begin{pmatrix} \Lambda^2 & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{pmatrix}$$

ce qui montre que les valeurs singulières sont les racines carrées des valeurs propres de $\mathbf{A}^T\mathbf{A}$ et de $\mathbf{AA}^T$, que $\mathbf{U}$ est la matrice des vecteurs propres de $\mathbf{AA}^T$ et que $\mathbf{V}$ est la matrice des vecteurs propres de $\mathbf{A}^T\mathbf{A}$.

En partitionnant la matrice $\mathbf{U}$ en $\mathbf{U} = \begin{pmatrix} \mathbf{U}_r & \mathbf{N}_g \end{pmatrix}$ avec $\mathbf{U}_r$ de dimension $m \times r$, et $\mathbf{N}_g$ de dimension $m \times (m-r)$, on peut vérifier que $\mathbf{N}_g^T \mathbf{A} = 0$, autrement dit $\mathbf{N}_g$ est le noyau orthonormé à gauche de $\mathbf{A}$:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_r & \mathbf{N}_g \end{pmatrix} \text{ tel que } \mathbf{N}_g^T \mathbf{A} = 0$$

Les colonnes de $\mathbf{U}_r$ constituent une base orthonormée de l'espace engendré par les colonnes de $\mathbf{A}$. Les colonnes de $\mathbf{N}_g$ constituent une base de l'espace orthogonal.

En partitionnant la matrice $\mathbf{V}$ en $\mathbf{V} = \begin{pmatrix} \mathbf{V}_r & \mathbf{N}_d \end{pmatrix}$ avec $\mathbf{V}_r$ de dimension $n \times r$, et $\mathbf{N}_d$ de dimension $n \times (n-r)$, on peut vérifier que $\mathbf{A}\mathbf{N}_d = 0$, autrement dit $\mathbf{N}_d$ est le noyau orthonormé à droite de $\mathbf{A}$ **:**

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_r & \mathbf{N}_d \end{pmatrix} \text{ tel que } \mathbf{A}\mathbf{N}_d = 0$$

Les colonnes de $\mathbf{V}_r$ constituent une base orthonormée de l'espace engendré par les lignes de $\mathbf{A}$. Les colonnes de $\mathbf{N}_d$ constituent une base de l'espace orthogonal.

$\mathbf{U}$ et $\mathbf{V}$ étant unitaires, on a $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{1}$ et $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{1}$, d'où :

$$\begin{cases} \mathbf{U}_r\mathbf{U}_r^T + \mathbf{N}_g\mathbf{N}_g^T = \mathbf{1} \\ \mathbf{U}_r^T\mathbf{U}_r = \mathbf{1} \\ \mathbf{N}_g^T\mathbf{N}_g = \mathbf{1} \\ \mathbf{U}_r^T\mathbf{N}_g = 0 \\ \mathbf{N}_g^T\mathbf{U}_r = 0 \end{cases} \text{ et } \begin{cases} \mathbf{V}_r\mathbf{V}_r^T + \mathbf{N}_d\mathbf{N}_d^T = \mathbf{1} \\ \mathbf{V}_r^T\mathbf{V}_r = \mathbf{1} \\ \mathbf{N}_d^T\mathbf{N}_d = \mathbf{1} \\ \mathbf{V}_r^T\mathbf{N}_d = 0 \\ \mathbf{N}_d^T\mathbf{V}_r = 0 \end{cases}$$

Par ailleurs :

$$\mathbf{A} = \begin{pmatrix} \mathbf{U}_r & \mathbf{N}_g \end{pmatrix} \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}_r^T \\ \mathbf{N}_d^T \end{pmatrix}$$

$$\mathbf{A} = \mathbf{U}_r \Lambda \mathbf{V}_r^T$$

La pseudo-inverse qui fournit la solution de norme minimale au système $\mathbf{y} = \mathbf{A}\mathbf{x}$ s'écrit :

$$\mathbf{A}^\dagger = \mathbf{V} \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{U}^T = \mathbf{V}_r \Lambda^{-1} \mathbf{U}_r^T$$

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y}$$

La solution générale s'écrit alors :

$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{x}_N \text{ avec } \begin{cases} \hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y} \text{ de norme minimale} \\ \mathbf{x}_N = \mathbf{N}_d \mathbf{N}_d^T \mathbf{x}', \forall \mathbf{x}' \in \mathbb{R}^n \end{cases}$$

Remarque : $\mathbf{x}_N$ s'écrit également $\mathbf{x}_N = \mathbf{N}_d \mathbf{z}$ où $\mathbf{z}$ est un vecteur quelconque de dimension $\in \mathbb{R}^{n-r}$.

Le calcul des matrices $\mathbf{U}$, $\mathbf{D}$ et $\mathbf{V}$ est relativement difficile. Il met en oeuvre des procédures itératives qui convergent vers les matrices cherchées.

Si $\mathbf{A}$ est de rang complet, c'est-à-dire si $r = \min(m,n)$, on dispose d'expressions plus simples ne nécessitant pas cette décomposition :

- Si $r = m = n$, on est dans le cas déterminé où $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.
- Si $r = m < n$, on est dans le cas redondant où $\mathbf{A}^\dagger$ est la pseudo-inverse de Moore-Penrose : $\mathbf{A}^\dagger = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1}$ présentée en section 3.6.
- Si $r = n < m$, on est dans le cas sur-déterminé où $\mathbf{A}^\dagger$ est la pseudo-inverse des moindres-carrés : $\mathbf{A}^\dagger = \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T$ présentée en section 3.4.

Dans le cas contraire, c'est-à-dire si $r < \min(m,n)$, seule la SVD fournit la pseudo-inverse $\mathbf{A}^\dagger$.

3.8.2 Procédure itérative de décomposition en valeurs singulières

Dans un premier temps, la matrice **A** est transformée en une matrice bi-diagonale **B** par une série de k produits à droite et l produits à gauche par des matrices qui réalisent des symétries par rapport à des hyperplans :

$$\mathbf{B} = \mathbf{S}_K \cdots \mathbf{S}_2 \mathbf{S}_1 \mathbf{A} \mathbf{P}_1 \mathbf{P}_2 \ldots \mathbf{P}_L$$

avec $K = \min(n, m-1)$ et $L = \min(m, n-2)$.

de telle manière que :

cas $m \geq n$

$$\mathbf{B} = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_n \\ 0 & 0 & 0 & \ldots & \ldots & p_n \\ 0 & 0 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & \ldots & 0 \end{pmatrix}$$

cas $n > m$

$$\mathbf{B} = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_m & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \ldots & p_m & q_{m+1} & \ldots & 0 \end{pmatrix}$$

La matrice de symétrie $\mathbf{S}_k = \mathbf{1} - \frac{1}{h_k}\mathbf{v}_k\mathbf{v}_k^T$ est calculée par les formules données pour la décomposition QR, afin de produire la colonne de zéros sous le terme p_k. La matrice $\mathbf{P}_k = \mathbf{1} - \frac{1}{h'_k}\mathbf{v}'_k\mathbf{v}'^T_k$ est calculée de manière similaire pour produire les zéros en ligne à droite de q_k.

Multiplication à gauche par $\mathbf{S}_k$:

Calcul des éléments h_k et $\mathbf{v}_k$ de $\mathbf{S}_k$: En notant $a_{i,j}$ les éléments de la matrice qui a subi les transformations précédentes, il vient :

$$\begin{aligned} d_k^2 &= a_{k,k}^2 + a_{k+1,k}^2 + \ldots + a_{m,k}^2 \\ \mathbf{v}_k^T &= \begin{pmatrix} 0 & \ldots & 0 & v_{k,k} & a_{k+1,k} & \ldots & a_{m,k} \end{pmatrix} \\ v_{k,k} &= a_{k,k} + d_k \text{signe}(a_{k,k}) \\ h_k &= d_k (d_k + |a_{k,k}|) \end{aligned}$$

Dans la multiplication par $\mathbf{S}_k$ les colonnes et les lignes de rang inférieur à k ne sont pas touchées.

Pour la colonne k, on a $a'_{k,k} = -d_k \text{signe}(a_{k,k})$ et $a'_{l,k} = 0$ pour $l > k$.

Pour les colonnes $j = k+1$ à n, on a $a'_{i,j} = a'_{i,j} - \frac{s}{h_k} v_{i,j}$, avec $s = \sum_{l=k}^{m} v_{l,k} a_{l,j}$.

Remarque : Quand $d_k^2 = 0$ la multiplication par $\mathbf{S}_k$ est inutile.

Multiplication à droite par $\mathbf{P}_k$:

Calcul des éléments h'_k et $\mathbf{v}'_k$ de $\mathbf{P}_k$: En notant $a_{i,j}$ les éléments de la matrice qui résulte du produit par $\mathbf{S}_k$, il vient :

$$\begin{aligned}
\rho_k^2 &= a_{k,k+1}^2 + a_{k,k+2}^2 + \ldots + a_{k,n}^2 \\
\mathbf{v}_k'^T &= \begin{pmatrix} 0 & \ldots & 0 & v_{k,k+1} & a_{k,k+2} & \ldots & a_{k,n} \end{pmatrix} \\
v_{k,k+1} &= a_{k,k+1} + \rho_k \text{signe}\,(a_{k,k+1}) \\
h_k' &= \rho_k\,(\rho_k + |a_{k,k+1}|)
\end{aligned}$$

Dans la multiplication par $\mathbf{P}_k$ les colonnes et les lignes de rang inférieur à $k+1$ ne sont pas touchées.
Pour la ligne k, on a $a'_{k,k+1} = -\rho_k \text{signe}\,(a_{k,k+1})$ et $a'_{k,l} = 0$ pour $l > k+1$.

Pour les lignes $i = k+1$ à m, on a $a'_{i,l} = a'_{i,l} - \dfrac{s}{h'_k} v'_{k,l}$, avec $s = \sum_{l=k+1}^{n} v'_{k,l} a_{i,l}$.

Remarque : Quand $\rho_k^2 = 0$, la multiplication par $\mathbf{P}_k$ est inutile.

Procédure de diagonalisation

A l'issue du processus précédent, la matrice $\mathbf{A}$ a été transformé en la matrice bi-diagonale $\mathbf{B}$.

On considère les deux cas suivants :

Cas $m \geq n$

$$\mathbf{B} = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_n \\ 0 & 0 & 0 & \ldots & \ldots & p_n \\ 0 & 0 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & \ldots & 0 \end{pmatrix} \rightarrow \mathbf{B}' = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_n \\ 0 & 0 & 0 & \ldots & \ldots & p_n \end{pmatrix}$$

et :

Cas $n > m$

$$\mathbf{B} = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_m & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \ldots & p_m & q_{m+1} & \ldots & 0 \end{pmatrix} \rightarrow \mathbf{B}' = \begin{pmatrix} p_1 & q_2 & 0 & \ldots & \ldots & 0 & 0 \\ 0 & p_2 & q_3 & \ldots & \ldots & 0 & 0 \\ 0 & 0 & p_3 & \ddots & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ddots & q_m & 0 \\ 0 & 0 & 0 & \ldots & \ldots & p_m & q_{m+1} \\ 0 & 0 & 0 & \ldots & \ldots & 0 & 0 \end{pmatrix}$$

où $\mathbf{B}'$ est une matrice carrée d'ordre $b = n$ quand $n \leq m$ et d'ordre $b = m+1$ quand $n > m$. Dans ce dernier cas, une ligne de zéros est ajoutée à la matrice $\mathbf{B}$ (à laquelle on supprime les colonnes de rang $j > m+1$, s'il y en a).

Ensuite un processus itératif de transformation est entrepris pour faire converger la matrice :

$$\mathbf{B}_k = \mathbf{S}_k \mathbf{B}_{k-1} \mathbf{T}_k$$

vers la matrice diagonale $\mathbf{D}$ recherchée, où la matrice $\mathbf{B}_0$ est un bloc diagonal extrait de la matrice $\mathbf{B}'$. Initialement, si aucun élément q_k n'est nul, on a $\mathbf{B}_0 = \mathbf{B}'$. Mais si $q_k = 0$, alors $\mathbf{B}'$ est sous forme de deux

blocs diagonaux respectivement d'ordres $k-1$ et $b-k+1$. La procédure itérative, appliquée à ces blocs de dimension réduite, est beaucoup plus rapide. Dans ce qui suit, l'indice 1 correspond à la première ligne ou colonne de ces blocs et l'indice b à la dernière.

$\mathbf{S}_k$ et $\mathbf{T}_k$ sont des produits de matrice de rotation de Givens respectivement définis par :

$$\mathbf{T}_k = \mathbf{G}_{12}\mathbf{G}_{23}\dots\mathbf{G}_{b-1,b}$$
$$\mathbf{S}_k = \mathbf{G}'^T_{b-1,b}\dots\mathbf{G}'^T_{1,2}$$

Les produits sont effectuées dans l'ordre suivant : $\mathbf{B}_{k-1}\mathbf{G}_{12}$, $\mathbf{G}'^T_{1,2}\mathbf{H}$, $\mathbf{H}\mathbf{G}_{23}$, $\mathbf{G}'^T_{2,3}\mathbf{H}$, ... ,$\mathbf{H}\mathbf{G}_{b-1,b}$, $\mathbf{G}'^T_{b-1,b}\mathbf{H}$ ou $\mathbf{H}$ est le résultat du produit précédent.

— $\mathbf{B}_{k-1}\mathbf{G}_{12} \rightarrow \mathbf{H}$ fait apparaître un nouvel élément en h_{21}
— $\mathbf{G}'^T_{1,2}\mathbf{H} \rightarrow \mathbf{H}$ annule $h_{2,1}$ et fait apparaître un nouvel élément en $h_{1,3}$
— $\mathbf{H}\mathbf{G}_{23} \rightarrow \mathbf{H}$ annule $h_{1,3}$ et fait apparaître un nouvel élément en $h_{3,2}$
— $\mathbf{G}'^T_{2,3}\mathbf{H} \rightarrow \mathbf{H}$ annule $h_{3,2}$ et fait apparaître un nouvel élément en $h_{2,4}$
— ...
— $\mathbf{H}\mathbf{G}_{b-1,b} \rightarrow \mathbf{H}$ annule $h_{b-2,b}$ et fait apparaître un nouvel élément en $h_{b,b-1}$
— $\mathbf{G}'^T_{b-1,b}\mathbf{H} \rightarrow \mathbf{B}_k$ annule simplement $h_{b,b-1}$.

A l'issue de ces multiplications, la matrice $\mathbf{B}_k$ est bi-diagonale comme $\mathbf{B}_{k-1}$.

Calcul de $\mathbf{G}_{12}$

A l'exception de $\mathbf{G}_{12}$, chaque matrice de Givens est parfaitement définie de manière à annuler le terme créé par la multiplication précédente. Le calcul de ces matrices est détaillé au paragraphe 2.4. Pour calculer les éléments c et s de $\mathbf{G}_{12}$ on considère une ligne fictive ayant respectivement en colonne 1 et 2 des valeurs proportionnelles à :

$$\rho c = p_1^2 - p_b^2$$
$$\rho s = p_1 q_2$$

Wilkinson (dans [7] et [9]) garantit une convergence au moins cubique si on remplace $\rho c = p_1^2 - p_b^2$, par $\rho c = p_1^2 - \lambda$ avec λ valeur propre du dernier bloc 2×2 diagonal de la matrice $\mathbf{C} = \mathbf{B}^T_{k-1}\mathbf{B}_{k-1}$, la plus proche du dernier élément de la diagonale de $\mathbf{C}$. Considérons ce bloc 2×2, et posons :

$$\begin{pmatrix} c_{b-1,b-1} & c_{b-1,b} \\ c_{b,b-1} & c_{b,b} \end{pmatrix} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

La valeur proche λ recherchée (la plus proche de $c = c_{b,b}$) vaut :

$$\lambda = \begin{cases} \dfrac{c+a}{2} + \frac{1}{2}\sqrt{(a-c)^2+4b^2} \text{ si } c \geq a \\ \dfrac{c+a}{2} - \frac{1}{2}\sqrt{(a-c)^2+4b^2} \text{ si } c < a \end{cases} \tag{3.5}$$

Les sources fournies par LINPACK ([6]) et par Wilkinson dans [7] sont très confus à ce niveau.Ils mettent λ sous la forme :

$$\lambda = p_b^2 + \varepsilon \tag{3.6}$$

avec une expression très complexe pour ε et de plus ils sont en désaccord sur le signe de ε. Les quelques tests que nous avons effectués ne donnent pas de différence significative en rapidité de convergence avec λ calculé par (3.5) ou par (3.6) quel que soit le signe affecté à ε ou quand on fait $\varepsilon = 0$. Il semble qu'on puisse simplifier le codage de cette étape en conservant tout simplement $\lambda = p_b^2$, c'est-à-dire $\rho c = p_1^2 - p_b^2$.

But de l'itération

Au bout d'un certain nombre d'itérations (inférieur à 10 en général) l'élément q_b devient négligeable et p_b est une valeur singulière recherchée. On élimine alors la dernière ligne et la dernière colonne de $\mathbf{B}_k$ pour constituer une nouvelle matrice $\mathbf{B}_0$ d'ordre $b = b-1$. On itère le processus jusqu'à ce que b soit égal à 1. Toutes les valeurs singulières sont trouvées.

Au cours des calculs, tous les produits à gauche et à droite sont cumulés pour former les matrices $\mathbf{U}^T$ et $\mathbf{V}$. Au cours du processus, les valeurs singulières sont ordonnées de la plus grande à la plus petite et les permutations correspondantes sont reportés sur $\mathbf{U}^T$ et $\mathbf{V}$.

Conclusion

On retiendra que la puissance de cette méthode est en rapport avec la complexité des calculs qu'elle met en oeuvre. Contrairement aux méthodes précédentes, le nombre des opérations dépend de la précision recherchée.

3.9 Calcul du noyau d'une matrice

Le calcul du noyau d'une matrice peut être effectué à partir des factorisations QRE ou SVD.

3.9.1 Noyau par la factorisation QR

La factorisation QRE d'une matrice $\mathbf{A}$ contient le noyau à gauche de $\mathbf{A}$ ou le noyau à droite de la matrice $\mathbf{A}^T$. Notons $\mathbf{Q}_r$ les r premières colonnes de $\mathbf{Q}$ et $\mathbf{N}_g$ les $m-r$ dernières :

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_r & \mathbf{N}_g \end{pmatrix}$$

Il vient :

$$\begin{pmatrix} \mathbf{Q}_r^T \\ \mathbf{N}_g^T \end{pmatrix} \mathbf{A} = \begin{pmatrix} \mathbf{U}_r & \mathbf{B} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{E}^T$$

qui montre que :

$$\mathbf{N}_g^T \mathbf{A} = \mathbf{0} \quad ; \quad \mathbf{A}^T \mathbf{N}_g = \mathbf{0}$$

Les $m-r$ dernières colonnes de $\mathbf{Q}$ constituent une base orthonormée du noyau de $\mathbf{A}^T$.

Remarque : Dans la résolution précédente, avec $r < \min(n,m)$, le noyau nécessaire pour calculer la solution optimale est celui de $\mathbf{A}$, alors qu'on dispose dans $\mathbf{Q}$ de celui de $\mathbf{A}^T$. Si on veut le noyau de $\mathbf{A}$, pour calculer la solution de norme minimale (3.4), il faut effectuer une décomposition QRE de $\mathbf{A}^T$. Bien que peu usitée cette méthode peut s'avérer plus rapide que la méthode généralement utilisée qui passe par la décomposition en valeurs singulières de $\mathbf{A}$.

3.9.2 Noyau par la factorisation SVD

Rappelons les résultats présentés en section 3.8.1. La factorisation SVD d'une matrice $\mathbf{A}$ contient les noyaux à droite et à gauche de la matrice $\mathbf{A}$. Notons $\mathbf{U}_r$ les r premières colonnes de $\mathbf{U}$, $\mathbf{N}_g$ les $m-r$ dernières, $\mathbf{V}_r$ les r premières colonnes de $\mathbf{V}$ et $\mathbf{N}_d$ les $n-r$ dernières. Il vient :

$$\mathbf{A} = \begin{pmatrix} \mathbf{U}_r & \mathbf{N}_g \end{pmatrix} \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_r^T \\ \mathbf{N}_d^T \end{pmatrix}$$

Ce qui montre que :

$$\mathbf{A}\mathbf{N}_d = \mathbf{0} \text{ et } \mathbf{N}_g^T \mathbf{A} = \mathbf{0}$$

Les $n-r$ dernières colonnes de $\mathbf{V}$ *constituent une base orthonormée du noyau de* $\mathbf{A}$ *et les $m-r$ dernières colonnes de* $\mathbf{U}$ *constituent une base orthonormée du noyau de* $\mathbf{A}^T$.

Chapitre 4

Valeurs et vecteurs propres

Nous ne rappelons ici que les principaux résultats, en particulier ceux qui sont utilisés dans les algorithmes de résolution de système ou de calcul des valeurs et vecteurs propres.

4.1 Généralités

Etant donné une matrice carrée réelle $\mathbf{A}$, les vecteurs $\mathbf{v}$ tels que :

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

où λ est un scalaire, sont appelés vecteurs propres à droite de $\mathbf{A}$. λ est appelée valeur propre de $\mathbf{A}$. Si on cherche les solutions $\mathbf{v}$ non triviales (avec $\mathbf{v} \neq \mathbf{0}$) à l'équation $(\mathbf{A}-\lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}$, cela implique que $\mathbf{A}-\lambda\mathbf{I}$ possède un noyau (est singulière). Ainsi, les valeurs propres sont solutions de :

$$\det(\mathbf{A}-\lambda\mathbf{I}) = 0$$

En remarquant que $\det\left(\mathbf{A}^T-\lambda\mathbf{I}\right)$ est le même polynôme en λ que $\det(\mathbf{A}-\lambda\mathbf{I})$, il en résulte que $\mathbf{A}$ et $\mathbf{A}^T$ ont les mêmes valeurs propres. Notons $\mathbf{w}$ le vecteur propre $\mathbf{w}$ de $\mathbf{A}^T$ associé à λ (tel que $\mathbf{A}^T\mathbf{w} = \lambda\mathbf{w}$ ou $\mathbf{w}^T\mathbf{A} = \lambda\mathbf{w}^T$). On l'appelle également vecteur propre à gauche de $\mathbf{A}$. Si $\mathbf{A}$ n'est pas symétrique $\mathbf{w}$ et $\mathbf{v}$ sont a priori différents.

Notons $\mathbf{v}_i$ le vecteur propre à droite associé à λ_i et $\mathbf{w}_j$ le vecteur propre à gauche associé à λ_j. Calculons l'expression :

$$\mathbf{w}_j^T\mathbf{A}\mathbf{v}_i = \lambda_j\mathbf{w}_j^T\mathbf{v}_i = \mathbf{w}_j^T\lambda_i\mathbf{v}_i$$

ce qui implique :

$$0 = \mathbf{w}_j^T\mathbf{A}\mathbf{v}_i - \mathbf{w}_j^T\mathbf{A}\mathbf{v}_i = (\lambda_j-\lambda_i)\,\mathbf{w}_j^T\mathbf{v}_i$$

On a ainsi le théorème :

Les vecteurs propres à droite et à gauche associés à des valeurs propres distinctes sont orthogonaux.

Une matrice $\mathbf{A}$ réelle d'ordre n à n valeurs propres solution d'un polynôme caractéristique d'ordre n à coefficients réels. Elles sont donc soit réelles, soit complexes conjuguées. Certaines peuvent être multiples.

Par ailleurs on démontre par l'absurde que *si les n valeurs propres λ_i sont distinctes, les n vecteurs propres (à droite et à gauche) sont indépendants*.

Vecteur propre et cofacteurs

Considérons une ligne non nulle de la matrice des cofacteurs de $(\mathbf{A}-\lambda_i\mathbf{I})$, la ligne numéro k par exemple. Quel que soit le scalaire λ, le déterminant $\det(\mathbf{A}-\lambda\mathbf{I})$ est égal au produit scalaire de la ligne k des cofacteurs de $(\mathbf{A}-\lambda\mathbf{I})$ par la ligne k de $(\mathbf{A}-\lambda\mathbf{I})$. Si $\lambda=\lambda_i$ ce produit scalaire est nul. De plus, si on fait le produit scalaire de la même ligne k des cofacteurs de $(\mathbf{A}-\lambda\mathbf{I})$ par une autre ligne de $(\mathbf{A}-\lambda\mathbf{I})$, la ligne j par exemple, on trouvera un produit scalaire nul, car il correspond au déterminant d'une matrice qui a les lignes k et j identiques. Il en résulte que la ligne k des cofacteurs de $(\mathbf{A}-\lambda_i\mathbf{I})$ est orthogonale à toutes les lignes de $(\mathbf{A}-\lambda_i\mathbf{I})$. Ces coefficients sont donc proportionnels aux composantes du vecteur propre à droite associé à λ_i. Il en résulte que toute colonne non nulle de l'adjointe de $(\mathbf{A}-\lambda_i\mathbf{I})$ est vecteur propre de $\mathbf{A}$ associé à λ_i.

4.2 Valeurs et vecteurs propres d'une matrice réelle symétrique

Les valeurs propres et les vecteurs propres d'une matrice réelle symétrique sont réelles.

En effet, soit λ la valeur propre de la matrice symétrique $\mathbf{A}$ associée au vecteur propre $\mathbf{v}$. Notons $\bar{\mathbf{v}}$ le vecteur conjugué de $\mathbf{v}$. On a alors :

$$\mathbf{A}\mathbf{v}=\lambda\mathbf{v} \text{ implique } \bar{\mathbf{v}}^T\mathbf{A}\mathbf{v}=\lambda\bar{\mathbf{v}}^T\mathbf{v}=\lambda\|\mathbf{v}\|^2$$

Par ailleurs, $\mathbf{A}\mathbf{v}=\lambda\mathbf{v}$ implique $\mathbf{A}\bar{\mathbf{v}}=\bar{\lambda}\bar{\mathbf{v}}$ car $\mathbf{A}$ est réelle et $\bar{\mathbf{v}}^T\mathbf{A}=\bar{\lambda}\bar{\mathbf{v}}^T$ car $\mathbf{A}$ est symétrique, d'où :

$$\bar{\mathbf{v}}^T\mathbf{A}\mathbf{v}=\bar{\lambda}\bar{\mathbf{v}}^T\mathbf{v}=\bar{\lambda}\|\mathbf{v}\|^2=\lambda\|\mathbf{v}\|^2\rightarrow\bar{\lambda}=\lambda$$

la valeur propre est réelle. Il en résulte que le vecteur propre l'est aussi.

Les valeurs propres d'une matrice réelle symétrique du type $\mathbf{A}=\mathbf{B}^T\mathbf{B}$ sont réelles positives ou nulles.

En effet, si $\mathbf{A}=\mathbf{B}^T\mathbf{B}$ alors $\lambda\|\mathbf{v}\|^2=\mathbf{v}^T\mathbf{A}\mathbf{v}=\mathbf{v}^T\mathbf{B}^T\mathbf{B}\mathbf{v}=\|\mathbf{B}\mathbf{v}\|^2\geq 0$. λ est réel positif ou nul.

Les vecteurs propres d'une matrice réelle symétrique associés à des valeurs propres distinctes non nulles sont orthogonaux.

En effet, soient λ_1 et λ_2 deux valeurs propres non nulles distinctes de la matrice $\mathbf{A}$ associée aux vecteurs propres $\mathbf{v}_1$ et $\mathbf{v}_2$. On a :

$$\mathbf{A}\mathbf{v}_1=\lambda_1\mathbf{v}_1 \text{ implique } \mathbf{v}_2^T\mathbf{A}\mathbf{v}_1=\lambda_1\mathbf{v}_2^T\mathbf{v}_1 \text{ ou bien } \mathbf{v}_2^T\mathbf{A}=\lambda_2\mathbf{v}_2^T \text{ implique } \mathbf{v}_2^T\mathbf{A}\mathbf{v}_1=\lambda_2\mathbf{v}_2^T\mathbf{v}_1$$

D'où par soustraction :

$$0=(\lambda_2-\lambda_1)\,\mathbf{v}_2^T\mathbf{v}_1.$$

Or $\lambda_2\neq\lambda_1$ implique $\mathbf{v}_2^T\mathbf{v}_1=0$.

4.3 Valeurs et vecteurs propres complexes

Si λ est valeur propre complexe de $\mathbf{A}$ réelle carrée quelconque, associée au vecteur propre complexe $\mathbf{v}$, alors $\mathbf{A}\mathbf{v}=\lambda\mathbf{v}$ devient en conjuguant $\mathbf{A}\bar{\mathbf{v}}=\bar{\lambda}\bar{\mathbf{v}}$ qui montre que $\bar{\lambda}$ est valeur propre complexe de $\mathbf{A}$ associée à $\bar{\mathbf{v}}$.

La manipulation de quantités complexes n'étant pas toujours aisée (dans les logiciels qui ne traitent pas les complexes par exemple), on associe souvent les paires de valeurs propres et vecteurs propres complexes conjuguées pour ne manipuler que des quantités réelles.

Posons :

$$\lambda = \alpha + i\beta$$
$$\mathbf{v} = \mathbf{v}_x + i\mathbf{v}_y$$

Il vient :

$$\begin{aligned} \mathbf{Av} &= \lambda \mathbf{v} \\ \mathbf{A}(\mathbf{v_x} + \mathbf{iv_y}) &= (\alpha + i\beta)(\mathbf{v}_x + i\mathbf{v}_y) \\ \mathbf{Av_x} + \mathbf{iAv_y} &= (\alpha \mathbf{v}_x - \beta \mathbf{v}_y) + i(\alpha \mathbf{v}_y + \beta \mathbf{v}_x) \end{aligned}$$

D'où :

$$\mathbf{A}\begin{bmatrix} \mathbf{v}_x & \mathbf{v}_y \end{bmatrix} = \begin{bmatrix} \mathbf{v}_x & \mathbf{v}_y \end{bmatrix}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \tag{4.1}$$

Ainsi, à la paire de valeurs propres complexes $\alpha \pm i\beta$ on peut associer un bloc diagonal $\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$ couplé aux parties réelles et imaginaires des vecteurs propres par la relation (4.1).

Remarque : Comme tout vecteur propre, un vecteur propre complexe est défini à un coefficient *complexe* multiplicatif près :

$$\xi \mathbf{Av} = \mathbf{A}(\xi \mathbf{v}) = \lambda(\xi \mathbf{v})$$

Même s'ils sont normés, il est donc très difficile de comparer les directions des vecteurs propres complexes. Pour les comparer, il faut en plus de les normer, imposer à la première composante non nulle (ou la plus grande en valeur absolue) d'être réelle (ou imaginaire) pure.

4.4 Matrices semblables

Soit une transformation de matrice régulière **P** qui transforme les vecteurs de $\mathbb{R}^n$:

$$\begin{cases} \mathbf{v} = \mathbf{Pv}' \\ \mathbf{v}' = \mathbf{P}^{-1}\mathbf{v} \end{cases} \text{ et } \begin{cases} \mathbf{w} = \mathbf{Pw}' \\ \mathbf{w}' = \mathbf{P}^{-1}\mathbf{w} \end{cases}$$

Dans cette transformation, la matrice **A** qui transforme **w** en **v** :

$$\mathbf{w} = \mathbf{Av}$$

devient une *matrice semblable* **B** qui transforme $\mathbf{w}'$ en $\mathbf{v}'$:

$$\begin{aligned} \mathbf{w} &= \mathbf{Av} \\ \mathbf{Pw}' = \mathbf{APv}' &\rightarrow \begin{cases} \mathbf{w}' = \mathbf{Bv}' \\ \mathbf{B} = \mathbf{P}^{-1}\mathbf{AP} \end{cases} \end{aligned}$$

Le couple de valeur et vecteur propre λ et **v** devient :

$$\mathbf{Av} = \lambda \mathbf{v} \rightarrow \mathbf{PBP}^{-1}\mathbf{v} = \lambda \mathbf{v}$$

d'où :

$$\mathbf{Bw} = \lambda \mathbf{w} \text{ avec } \mathbf{w} = \mathbf{P}^{-1}\mathbf{v}$$

Les matrices semblables ont les mêmes valeurs propres. Les vecteurs propres sont multipliés par $\mathbf{P}^{-1}$.

4.5 Diagonalisation

4.5.1 Valeurs propres distinctes

Si une matrice carrée $n \times n$ possède n valeurs propres distinctes, en regroupant ses valeurs propres (réelles et/ou complexes) sous la forme d'une matrice diagonale Λ et ses vecteurs propres (réels et/ou complexes) à droite sous forme d'une matrice $\mathbf{V}$, et ses vecteurs propres à gauche sous forme d'une matrice $\mathbf{W}$ on peut écrire :

$$\mathbf{AV} = \mathbf{V}\Lambda$$
$$\mathbf{W}^T\mathbf{A} = \Lambda\mathbf{W}^T$$

avec :

$$\mathbf{W}^T\mathbf{V} = \mathbf{D}$$

où $\mathbf{D}$ est une matrice diagonale d'éléments $d_{ii} = \mathbf{w}_i^T\mathbf{v}_i \neq 0$ car si un des produit était nul, cela voudrait dire (par exemple) que $\mathbf{w}_i$ étant orthogonal aux n $\mathbf{v}_j$, ceux-ci sont liés, ce qui est contradictoire avec le théorème précédemment énoncé. Les vecteurs $\mathbf{v}_i$ et $\mathbf{w}_i^T$ étant arbitraires à un coefficient multiplicatif complexe (non nul) près, on peut les choisir tels que $d_{ii} = 1$. Il en résulte qu'on peut choisir :

$$\mathbf{W}^T = \mathbf{V}^{-1}$$

D'où :

$$\mathbf{V}^{-1}\mathbf{AV} = \Lambda \text{ et } \mathbf{A} = \mathbf{V}\Lambda V^{-1}$$

où on retrouve que les matrices semblables $\mathbf{A}$ et Λ ont mêmes valeurs propres.

4.5.2 Valeurs propres multiples

Si une valeur propre λ est racine d'ordre p du polynôme caractéristique, p est appelé la *multiplicité algébrique* de la valeur propre λ. Le nombre q de vecteurs propres indépendants que l'on peut associer à cette valeur propre λ est appelé la *multiplicité géométrique* de la valeur propre λ. Ces multiplicités vérifient :

$$q \leq p$$

Lorsque $q < p$, on ne peut pas constituer une base de $\mathbb{R}^n$ avec les vecteurs propres, ce qui est génant pour procéder à la diagonalisation de la matrice $\mathbf{A}$. Heureusement, pour la plupart des matrices usuelles (symétriques, hermitiques, anti-symétriques, unitaires) possédant des valeurs propres multiples, les vecteurs propres permettent de constituer une base de $\mathbb{R}^n$, et la diagonalisation s'opère comme dans le cas des valeurs propres distinctes.

L'exemple classique d'inégalité stricte entre multiplicité algébrique et géométrique, est le bloc de Jordan :

$$\mathbf{J} = \begin{pmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{pmatrix}$$

avec des termes non nuls, tous égaux (pris arbitrairement égaux à 1), au-dessus (ou en dessous) de la diagonale. Dans ce cas, $\lambda = a$ est valeur propre de multiplicité algébrique $p = 3$. Seul $\mathbf{v}^T = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}$ est vecteur propre. Il en résulte que $\lambda = a$ est valeur propre de multiplicité géométrique $q = 1$. Cette matrice ne peut être rendue semblable à une matrice diagonale (elle ne peut pas être diagonalisée).

Ainsi, il existe des matrices présentant une ou plusieurs valeurs propres avec une multiplicité géométrique strictement inférieure à la multiplicité algébrique. Dans ce cas, à toutes fins utiles, on peut considérer leur matrice semblable sous *forme canonique de Jordan*. La forme canonique de Jordan est une matrice constituée de blocs diagonaux, dont les seuls éléments non nuls sont, soit des valeurs (propres) isolées sur la diagonale, soit des blocs diagonaux de Jordan (associés à des valeurs propres multiples).

Toute matrice est semblable à une forme canonique de Jordan.

A une valeur propre λ de multiplicité géométrique q inférieure à sa multiplicité algébrique p, seront associés q blocs diagonaux de Jordan. Les dimensions de ces q blocs peuvent être réparties de diverses manières. La dimension minimale est 1. La somme des dimensions vaut p. Il y aura en tout $p-q$ termes à 1 au-dessus de la diagonale.

Malheureusement, on ne dispose pas d'algorithme efficace calculant la forme canonique de Jordan des matrices mal conditionnées. (cf. G.H. Golub et J.H. Wilkinson. "Ill-conditioned eigensystems and the computation of the Jordan canonical form", SIAM Rev. 18(1976), pp 578-619). Ainsi pour le calcul des valeurs propres, on utilise un algorithme de triangularisation spectrale (section suivante) et pour le calcul de l'exponentielle de matrice, on passe par une approximation de Padé.

4.6 Algorithmes de triangularisation spectrale

Quelle que soit la matrice carrée $\mathbf{A}$ d'ordre n, il existe une matrice régulière $\mathbf{P}$ telle que $\mathbf{PAP}^{-1} = \mathbf{U}$ matrice triangulaire supérieure, *avec les valeurs propres de* $\mathbf{A}$ *sur la diagonale de* $\mathbf{U}$. Comme Evariste Gallois a montré qu'il n'y a pas d'expression analytique pour les racines des polynômes de degré supérieur à 4, il en résulte qu'en théorie, pour $n > 4$, les matrices $\mathbf{P}$ ne peuvent pas être obtenues par un nombre fini d'opérations élémentaires (sinon, on exprimerait les racines du polynôme caractéristiques en fonctions des coefficients par un nombre fini d'opérations). D'excellentes approximations sont obtenues en un nombre limité d'itérations par les algorithmes LR, QR et HQR présentés ci-après. Ces algorithmes sont généralement appliqués après que la matrice $\mathbf{A}$ ait été mise sous forme de Hessenberg supérieure.

4.6.1 Forme de Hessenberg

Une matrice $\mathbf{B}$ est sous forme de Hessenberg supérieure lorsque tous ses élements sous la sous-diagonale sont nuls, soit $b_{i,j} = 0$ pour $1 \leq j \leq i-2$.

Une matrice carrée $\mathbf{A}$ d'ordre n peut être mise sous cette forme par des produits de symétries (transformations de Householder) ou par des rotations planes (transformations de Givens). A titre d'exemple, l'algorithme utilisant les transformations de Householder est le suivant.

L'algorithme transforme $\mathbf{A}_0 = \mathbf{A}$ en matrice de Hessenberg supérieure $\mathbf{B} = \mathbf{A}_{n-2}$ au moyen des $n-2$ symétries suivantes :

$$\left\{ \begin{array}{l} \mathbf{P}_k = \mathbf{I} - \dfrac{1}{h}\mathbf{v}\mathbf{v}^T \\ \mathbf{A}_k = \mathbf{P}_k \mathbf{A}_{k-1} \mathbf{P}_k \end{array} \right. , \quad k = 1, \ldots n-2$$

En notant $a_{i,j}$ les éléments de $\mathbf{A}_{k-1}$, h et $\mathbf{v}$ sont obtenus par :

$$
\begin{aligned}
&R = \sqrt{\sum_{i=k+1}^{n} a_{i,k}^2} \\
&\begin{cases} v_i = 0\,, & 1 \le i \le k \\ v_{k+1} = a_{k+1,k} + \varepsilon R \\ v_i = a_{i,k}\,, & k+2 \le i \le n \end{cases} \\
&h = R\,(R + \varepsilon a_{k+1,k}) = \varepsilon R v_{k+1}
\end{aligned}
$$

où $\varepsilon = \pm 1$ est traditionnellement choisi égal à signe $(a_{k+1,k})$.

Le calcul de $\mathbf{A}_k$ à partir de $\mathbf{A}_{k-1}$ est effectué de la manière suivante :

$$
\begin{aligned}
&\mathbf{x}^T = \mathbf{v}^T \mathbf{A}_{k-1}\ ,\ \mathbf{y} = \mathbf{A}_{k-1}\mathbf{v}\ ,\ \mathbf{z} = \frac{1}{h}\mathbf{v}\ ,\ \alpha = \mathbf{x}^T\mathbf{z} \\
&\mathbf{A}_k = \mathbf{A}_{k-1} - \mathbf{z}\mathbf{x}^T - (\mathbf{y} - \alpha\mathbf{u})\,\mathbf{z}^T
\end{aligned}
$$

La matrice $\mathbf{B} = \mathbf{A}_{n-2}$ est sous forme de Hessenberg supérieure.

C'est algorithme peut être simplifié lorsque $\mathbf{A}$ est symétrique.

4.6.2 Algorithme LR et QR

La triangularisation de $\mathbf{A}$ est réalisée, à partir $\mathbf{A}_0 = \mathbf{A}$, par les itérations suivantes. On effectue, soit une factorisation QR *à diagonale positive*, soit une factorisation LU (de Doolittle à diagonale unité), de $\mathbf{A}_k$:

$$
\mathbf{A}_k = \begin{cases} \mathbf{Q}_k\mathbf{R}_k \\ \mathbf{L}_k\mathbf{U}_k \end{cases} = \mathbf{P}_k\mathbf{R}_k
$$

où $\mathbf{P}_k$ représente $\mathbf{Q}_k$ dans le cas de la factorisation QR et $\mathbf{L}_k$ dans le cas factorisation LU. On définit $\mathbf{A}_{k+1}$ par :

$$
\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{P}_k
$$

Comme $\mathbf{A}_{k+1} = \mathbf{P}_k^{-1}\mathbf{P}_k\mathbf{R}_k\mathbf{P}_k$, on a :

$$
\mathbf{A}_{k+1} = \mathbf{P}_k^{-1}\mathbf{A}_k\mathbf{P}_k
$$

Ainsi, $\mathbf{A}_{k+1}$ est semblable à $\mathbf{A}_0 = \mathbf{A}$ et possède les mêmes valeurs propres que $\mathbf{A}$.

Lorsque toutes les valeurs propres sont distinctes et réelles, Rutishauser à montré en 1958 que les éléments diagonaux $a_{i,j}^{(k)}$ de la matrice $\mathbf{A}_k$ tendent vers les valeurs propres λ_i de $\mathbf{A}$ lorsque $k \to \infty$ alors que $\mathbf{L}_k \to \mathbf{I}$ ([8] pages 487 à 492). Dans le cas avec la factorisation QR de Francis [2] les éléments diagonaux $a_{i,j}^{(k)}$ de la matrice $\mathbf{A}_k$ tendent également vers les valeurs propres λ_i de $\mathbf{A}$ lorsque $k \to \infty$ alors que $\mathbf{Q}_k$ tend vers une matrice à diagonale $q_{ii}^{(k)} = 1\text{signe}(\lambda_i)$ ([8] pages 517 à 521 et [1] pages 253 à 255). Dans les deux cas, la convergence est d'autant plus rapide que les rapports $|\lambda_i/\lambda_{i+1}|$ sont élevés).

Pour chaque couple de valeurs propres complexes conjuguées λ_p et $\lambda_{p+1} = \bar{\lambda}_p$ (égales en module), il subsiste dans $\mathbf{A}_k$, lorsque $k \to \infty$, des paires d'éléments non nuls, immédiatement de part et d'autre de la diagonale, qui constituent des blocs diagonaux 2×2 :

$$
\begin{pmatrix} a_{p,p}^{(k)} & a_{p,p+1}^{(k)} \\ a_{p+1,p}^{(k)} & a_{p+1,p+1}^{(k)} \end{pmatrix}
$$

ayant pour valeurs propres λ_p et $\bar{\lambda}_p$.

Cette procédure est appliquée après que $\mathbf{A}_0$ ait été mise sous *forme de Hessenberg supérieure* car cette forme simplifie les factorisations (toutes les matrices $\mathbf{A}_k$ restent sous cette forme).

Partition des $\mathbf{A}_k$

Lorsqu'un élément de la sous-diagonale de Hessenberg devient négligeable, la matrice $\mathbf{A}_k$ peut être partitionnée comme ci-dessous :

$$\mathbf{A}_k = \begin{pmatrix} \mathbf{B}_k & \mathbf{C}_k \\ \mathbf{0} & \mathbf{D}_k \end{pmatrix}$$

pour être traitée comme somme directe des deux matrices $\mathbf{B}_k$ et $\mathbf{D}_k$, puisque la matrice $\mathbf{C}_k$ n'intervient pas dans le déterminant de $\mathbf{A}_k - \lambda\mathbf{I}$. Si $\mathbf{B}_k$ et/ou $\mathbf{D}_k$ sont d'ordre 2, le calcul des valeurs propres est immédiat.

Décalage spectral

Présentons son principe dans le cas où les valeurs propres sont réelles. En remarquant que la matrice $\mathbf{A}_k - p\mathbf{I}$ à comme valeurs propres $\lambda_i - p$, et que la convergence est d'autant plus rapide que les rapports $\left|\lambda_i/\lambda_j\right|$ sont élevés, on voit que pour $\mathbf{A}_k - p\mathbf{I}$ la convergence qui dépend $\left|(\lambda_i - p)/(\lambda_j - p)\right|$ sera très rapide si p est proche de λ_j. Il y a donc intérêt à procéder à un décalage spectral avec p égal à la plus petite des valeurs propres, c'est-à-dire λ_n, car la factorisation les range de la plus grande à la plus petite. Comme approximation, on peut utiliser $a_{n,n}^{(k)}$, ou mieux, la valeur propre du dernier bloc diagonal 2×2 qui est la plus proche de $a_{n,n}^{(k)}$.

L'algorithme de *décalage spectral avec restauration* factorise $\mathbf{A}_k - p_k\mathbf{I}$ et restaure la valeur classique de $\mathbf{A}_{k+1}$ de la manière suivante :

$$\begin{aligned} \mathbf{A}_k - p_k\mathbf{I} &= \mathbf{P}_k\mathbf{R}_k \\ \mathbf{A}_{k+1} &= \mathbf{R}_k\mathbf{P}_k + p_k\mathbf{I} \end{aligned} \tag{4.2}$$

Comme

$$\begin{aligned} \mathbf{A}_{k+1} &= \mathbf{P}_k^{-1}\mathbf{P}_k\mathbf{R}_k\mathbf{P}_k + p_k\mathbf{I} \\ &= \mathbf{P}_k^{-1}\left(\mathbf{A}_k - p_k\mathbf{I}\right)\mathbf{P}_k + p_k\mathbf{I} \end{aligned}$$

on a toujours :

$$\mathbf{A}_{k+1} = \mathbf{P}_k^{-1}\mathbf{A}_k\mathbf{P}_k \tag{4.3}$$

semblable à $\mathbf{A}$.

Sans restauration l'algorithme est le suivant :

$$\begin{aligned} \mathbf{A}_k - p_k\mathbf{I} &= \mathbf{P}_k\mathbf{R}_k \\ \mathbf{A}_{k+1} &= \mathbf{R}_k\mathbf{P}_k \end{aligned}$$

Comme :

$$\begin{aligned} \mathbf{A}_{k+1} &= \mathbf{P}_k^{-1}\mathbf{P}_k\mathbf{R}_k\mathbf{P}_k \\ &= \mathbf{P}_k^{-1}\left(\mathbf{A}_k - p_k\mathbf{I}\right)\mathbf{P}_k \\ &= \mathbf{P}_k^{-1}\mathbf{A}_k\mathbf{P}_k - p_k\mathbf{I} \end{aligned}$$

il vient :

$$\mathbf{A}_{k+1} = \mathbf{P}\mathbf{A}_0\mathbf{P}^{-1} - \sum_{i=0}^{k}(p_i)\,\mathbf{I}$$

Les valeurs propres de $\mathbf{A}_{k+1}$ sont celles de $\mathbf{A}_0$ décalées de $\sum_{i=0}^{k}(p_i)$.

Déflation

Lorsque $a^{(k)}_{n-1,n}$ est devenu négligeable à la précision de calcul, $a^{(k)}_{n,n}$ est la valeur propre cherchée (au décalage spectral près). Les autres valeurs propres sont celles de la sous-matrice principale d'ordre $n-1$ qui, comme la matrice initiale, est sous forme de Hessenberg. Le processus est alors appliqué à cette sous-matrice, en prenant tout de suite $a^{(k)}_{n-1,n-1}$ comme valeur du décalage spectral p_k. Ce processus de déflation est itéré jusqu'à ce qu'il ne reste qu'un bloc 2×2 pour lequel le calcul des valeurs propres est immédiat.

4.6.3 L'algorithme HQR

Cet algorithme attribué à Francis et Kublanovskaya ([2] et [5]) est mis en oeuvre dans la librairie Eispack : [7], pages 359-371. Il est basé sur la factorisation QR d'une matrice de Hessenberg, avec décalage spectral et restauration, effectués par paires, qui utilisent les deux valeurs propres du dernier bloc diagonal 2×2 dont la somme et le produit (réels) sont données par :

$$\begin{aligned} p_k + p_{k+1} &= a^{(k)}_{n-1,n-1} + a^{(k)}_{n,n} \\ p_k p_{k+1} &= a^{(k)}_{n-1,n-1} a^{(k)}_{n,n} - a^{(k)}_{n-1,n} a^{(k)}_{n,n-1} \end{aligned}$$

Si les deux itérations étaient effectuées l'une après l'autre, la matrice $\mathbf{A}_{k+2}$ serait calculée par :

$$(\mathbf{A}_{k+1} - p_{k+1}\mathbf{I}) = \mathbf{P}_{k+1}\mathbf{R}_{k+1} \quad \rightarrow \quad \mathbf{A}_{k+2} = \mathbf{R}_{k+1}\mathbf{P}_{k+1} + p_{k+1}\mathbf{I}$$

où $\mathbf{A}_{k+2}$ vérifie de plus l'égalité :

$$\mathbf{A}_{k+2} = \mathbf{P}^{-1}_{k+1}\mathbf{A}_{k+1}\mathbf{P}_{k+1} = \mathbf{P}^{-1}_{k+1}\mathbf{P}^{-1}_{k}\mathbf{A}_k\mathbf{P}_k\mathbf{P}_{k+1} = \mathbf{Q}^T\mathbf{A}_k\mathbf{Q}$$

où on a posé :

$$\mathbf{Q} = \mathbf{P}_k\mathbf{P}_{k+1}$$

Posons par ailleurs :

$$\mathbf{R} = \mathbf{R}_{k+1}\mathbf{R}_k$$

et calculons le produit $\mathbf{QR}$. Il vient en utilisant les égalités 4.2 et 4.3 :

$$\begin{aligned} \mathbf{QR} &= \mathbf{P}_k\mathbf{P}_{k+1}\mathbf{R}_{k+1}\mathbf{R}_k = \mathbf{P}_k\left(\mathbf{A}_{k+1} - p_{k+1}\mathbf{I}\right)\mathbf{R}_k \\ &= \mathbf{P}_k\mathbf{A}_{k+1}\mathbf{R}_k - p_{k+1}\mathbf{P}_k\mathbf{R}_k = \mathbf{A}_k\mathbf{P}_k\mathbf{R}_k - p_{k+1}\mathbf{P}_k\mathbf{R}_k \\ &= \left(\mathbf{A}_k - p_{k+1}\mathbf{I}\right)\mathbf{P}_k\mathbf{R}_k = \left(\mathbf{A}_k - p_{k+1}\mathbf{I}\right)\left(\mathbf{A}_k - p_k\mathbf{I}\right) \end{aligned}$$

La matrice réelle :

$$\mathbf{M} = \left(\mathbf{A}_k - p_{k+1}\mathbf{I}\right)\left(\mathbf{A}_k - p_k\mathbf{I}\right) = \mathbf{A}^2_k - \left(p_k + p_{k+1}\right)\mathbf{A}_k + p_k p_{k+1}\mathbf{I}$$

permet d'obtenir directement la matrice réelle $\mathbf{Q}$ par sa factorisation QR.

$$\mathbf{M} = \mathbf{QR} \tag{4.4}$$

puis $\mathbf{A}_{k+2}$ peut être calculée par $\mathbf{A}_{k+2} = \mathbf{Q}^T\mathbf{A}_k\mathbf{Q}$.

Toutefois, $\mathbf{A}_{k+2}$ n'est pas obtenu par cette relation dans l'algorithme de LINPACK. Il détermine directement $\mathbf{A}_{k+2}$ à partir de $\mathbf{A}_k$ et de la première colonne de $\mathbf{M}$. Il est basé sur l'*implicit Q theorem* ([3], pages 346-347). Ce théorème indique que :

- si $\mathbf{H}$ est une forme de Hessenberg supérieure semblable à $\mathbf{A}_k$:

$$\mathbf{H} = \mathbf{S}^T\mathbf{A}_k\mathbf{S}$$

- et si la matrice orthogonale $\mathbf{S}$ a sa première colonne $\mathbf{s}_1$ égale à celle de $\mathbf{Q}$ alors on a :

$$\mathbf{H} = \mathbf{A}_{k+2}$$
$$\mathbf{S} = \mathbf{Q}$$

Cette propriété est utilisée pour calculer $\mathbf{S}^T$ sous forme d'un produit de $n-1$ matrices de symétrie $\mathbf{P}_i = \mathbf{I} - 2\mathbf{u}_i\mathbf{u}_i^T$ (matrices de Householder) :

$$\mathbf{S}^T = \mathbf{P}_{n-1}\ldots\mathbf{P}_2\mathbf{P}_1$$

La relation (4.4) montre que $\mathbf{R} = \mathbf{S^T M}$, c'est-à-dire que les matrices $\mathbf{P}_i$ triangularisent $\mathbf{M}$. Les matrices $\mathbf{P}_2$, $\mathbf{P}_2$, $\ldots,\mathbf{P}_{n-1}$ ne modifiant pas la première colonne, il en résulte que $\mathbf{s}_1$ est la première colonne de $\mathbf{P}_1$, matrice de symétrie déterminée par la première colonne de $\mathbf{M}$ dont elle annule les $n-1$ derniers termes. Ensuite, les $n-2$ matrices $\mathbf{P}_2$, $\mathbf{P}_2$, $\ldots,\mathbf{P}_{n-1}$ sont déterminées de manière à amener la matrice $\mathbf{P}_1\mathbf{A}_k\mathbf{P}_1^T$ à la forme de Hessenberg supérieure :

$$\mathbf{H} = \mathbf{P}_{n-1}\ldots\mathbf{P}_2\mathbf{P}_1\mathbf{A}_k\mathbf{P}_1\mathbf{P}_2\ldots\mathbf{P}_{n-1}$$

par la procédure décrite en 4.6.1. Etant donné que $\mathbf{A}_k$ était déjà sous forme de Hessenberg, il en résulte que les vecteurs $\mathbf{u}_i$ des matrices $\mathbf{P}_i$ n'ont que 3 composantes non nulles (i, $i+1$ et $i+2$).

Chapitre 5

L'exponentielle de matrice

5.1 L'intégration des systèmes différentiels linéaires

L'exponentielle de matrice intervient dans la résolution des systèmes différentiels linéaires à coefficients constant, du type :

$$\dot{\mathbf{x}} + \mathbf{A}\mathbf{x} = \mathbf{e}(t) \tag{5.1}$$

où $\mathbf{x} = \mathbf{x}(t)$ est un vecteur inconnu de dimension n, fonction du paramètre t, $\mathbf{A}$ une matrice $n \times n$ et $\mathbf{e}(t)$ un vecteur de dimension n donné en fonction de t. La dérivée relative au paramètre t est notée $\dot{\mathbf{x}} = \frac{d}{dt}\mathbf{x}(t)$. Supposons connu :

$$\mathbf{x}(0) = \mathbf{x}_0$$

Définissons l'exponentielle de matrice par la formule :

$$e^{\mathbf{A}} = 1 + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \cdots \tag{5.2}$$

Il en résulte que :

$$\begin{aligned} e^{\mathbf{A}t} &= \mathbf{1} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^2 + \cdots \\ \frac{d}{dt}\left(e^{\mathbf{A}t}\right) &= 0 + \mathbf{A}\left(1 + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^2 + \cdots\right) \\ \frac{d}{dt}\left(e^{\mathbf{A}t}\right) &= \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A} \end{aligned}$$

Ainsi, on peut vérifier que $\mathbf{x}_s(t) = e^{-\mathbf{A}t}\mathbf{x}_0$ est solution du système différentiel (5.1) sans deuxième membre. En effet :

$$\dot{\mathbf{x}}_s(t) = \frac{d}{dt}\left(e^{-\mathbf{A}t}\right)\mathbf{x}_0 = -\mathbf{A}e^{-\mathbf{A}t}\mathbf{x}_0$$

d'où :

$$\dot{\mathbf{x}}_s + \mathbf{A}\mathbf{x}_s = -\mathbf{A}e^{-\mathbf{A}t}\mathbf{x}_0 + \mathbf{A}e^{-\mathbf{A}t}\mathbf{x}_0 = 0$$

La solution du système complet (5.1) est obtenue par la méthode de variation de la constante. On pose :

$$\mathbf{x}(t) = e^{-\mathbf{A}t}\mathbf{y}(t)$$

avec par conséquent :

$$\mathbf{y}(0) = \mathbf{x}(0) = \mathbf{x}_0 \tag{5.3}$$

Il en résulte que :

$$\dot{\mathbf{x}} = -\mathbf{A}e^{-\mathbf{A}t}\mathbf{y} + e^{-\mathbf{A}t}\dot{\mathbf{y}} = \mathbf{e}(t) - \mathbf{A}\mathbf{x}$$

D'où :

$$e^{-\mathbf{A}t}\dot{\mathbf{y}} = \mathbf{e}(t) \rightarrow \dot{\mathbf{y}} = e^{\mathbf{A}t}\mathbf{e}(t)$$

et :

$$\mathbf{y}(t) = cte + \int_0^t e^{\mathbf{A}\tau}\mathbf{e}(\tau)\,d\tau$$

La constante est donnée par la relation (5.3). Il en résulte :

$$\mathbf{x}(t) = e^{-\mathbf{A}t}\left(\mathbf{x}_0 + \int_0^t e^{\mathbf{A}\tau}\mathbf{e}(\tau)\,d\tau\right)$$

que l'on écrit parfois :

$$\mathbf{x}(t) = e^{-\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(\tau-t)}\mathbf{e}(\tau)\,d\tau$$

Remarque :

Si $\mathbf{A}$ est diagonale d'éléments a_{ii}, alors le système sans second membre est composé de n équations différentielles indépendantes dont les solutions sont :

$$x_i(t) = x_i(0)e^{-a_{ii}t}$$

Il en résulte que $e^{\mathbf{A}}$ est la matrice diagonale ayant pour éléments diagonaux les $e^{a_{ii}}$.

5.2 Calcul de l'exponentielle de matrice

5.2.1 Utilisation du théorème de Cayley-Hamilton

Le calcul de $e^{\mathbf{A}}$ est rarement fait en utilisant la série (5.2). Dans certains cas, avec des matrices d'ordre réduit, on peut mettre à profit le théorème de Cayley-Hamilton pour effectuer cette sommation. Considérons par exemple la matrice suivante :

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3\omega^2 & 0 & 0 & -2\omega \\ 0 & 0 & 2\omega & 0 \end{pmatrix}$$

Ella a pour polynôme caractéristique :

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \lambda^4 + \omega^2\lambda^2 = \lambda^2\left(\lambda^2 + \omega^2\right) = 0$$

Elle n'est pas diagonalisable, car la valeur propre double $\lambda = 0$ n'a qu'un seul vecteur propre $\vec{e}_2$. Le théorème de Cayley-Hamilton indique que $\mathbf{A}$ est racine de son équation caractéristique. Ainsi on a :

$$\mathbf{A}^4 + \omega^2\mathbf{A}^2 = 0 \rightarrow \begin{cases} \mathbf{A}^4 = -\omega^2\mathbf{A}^2 \\ \mathbf{A}^5 = -\omega^2\mathbf{A}^3 \\ \mathbf{A}^6 = -\omega^2\mathbf{A}^4 = \omega^4\mathbf{A}^2 \\ \mathbf{A}^7 = \omega^4\mathbf{A}^3 \\ \cdots \end{cases}$$

D'où :

$$\begin{aligned}
e^{\mathbf{A}t} &= 1+\mathbf{A}t+\frac{1}{2!}\mathbf{A}^2t^2+\frac{1}{3!}\mathbf{A}^3t^3+\frac{1}{4!}\mathbf{A}^4t^4+\frac{1}{5!}\mathbf{A}^5t^5+\frac{1}{6!}\mathbf{A}^6t^6+\frac{1}{7!}\mathbf{A}^7t^7+\dots \\
&= 1+\mathbf{A}t+\frac{1}{2!}\mathbf{A}^2t^2+\frac{1}{3!}\mathbf{A}^3t^3-\frac{1}{4!}\omega^2\mathbf{A}^2t^4-\frac{1}{5!}\omega^2\mathbf{A}^3t^5+\frac{1}{6!}\omega^4\mathbf{A}^2t^6+\frac{1}{7!}\omega^4\mathbf{A}^3t^7-\dots \\
&= 1+\mathbf{A}t+\mathbf{A}^2\left(\frac{1}{2!}t^2-\frac{1}{4!}\omega^2t^4+\frac{1}{6!}\omega^4t^6-\dots\right)+\mathbf{A}^3\left(\frac{1}{3!}t^3-\frac{1}{5!}\omega^2t^5+\frac{1}{7!}\omega^4t^7-\dots\right) \\
&= 1+\mathbf{A}t+\frac{1}{\omega^2}\mathbf{A}^2\left(1-\cos\omega t\right)+\frac{1}{\omega^3}\mathbf{A}^3\left(\omega t-\sin\omega t\right)
\end{aligned}$$

Or :

$$\mathbf{A}^2=\begin{pmatrix} 3\omega^2 & 0 & 0 & -2\omega \\ 0 & 0 & 2\omega & 0 \\ 0 & 0 & -\omega^2 & 0 \\ 6\omega^3 & 0 & 0 & -4\omega^2 \end{pmatrix} \quad ; \quad \mathbf{A}^3=\begin{pmatrix} 0 & 0 & -\omega^2 & 0 \\ 6\omega^3 & 0 & 0 & -4\omega^2 \\ -3\omega^4 & 0 & 0 & 2\omega^3 \\ 0 & 0 & -2\omega^3 & 0 \end{pmatrix}$$

D'où :

$$e^{\mathbf{A}t}=\begin{pmatrix} 4-3\cos\omega t & 0 & \frac{1}{\omega}\sin\omega t & \frac{2}{\omega}\left(\cos\omega t-1\right) \\ 6\left(\omega t-\sin\omega t\right) & 1 & \frac{2}{\omega}\left(1-\cos\omega t\right) & \frac{1}{\omega}\left(4\sin\omega t-3\omega t\right) \\ 3\omega\sin\omega t & 0 & \cos\omega t & -2\sin\omega t \\ 6\omega\left(1-\cos\omega t\right) & 0 & 2\sin\omega t & 4\cos\omega t-3 \end{pmatrix}$$

5.2.2 Matrice semblables

Supposons $\mathbf{A}$ soit semblable à une matrice $\mathbf{B}$ dont on connaît l'exponentielle de matrice $e^{\mathbf{B}}$. Par hypothèse il existe $\mathbf{V}$ régulière telle que :

$$\mathbf{A}=\mathbf{V}\mathbf{B}V^{-1}$$

On a alors :

$$\begin{aligned}
\mathbf{A}^2 &= \mathbf{V}\mathbf{B}V^{-1}\mathbf{V}\mathbf{B}V^{-1}=\mathbf{V}\mathbf{B}^2V^{-1} \\
\mathbf{A}^3 &= \mathbf{V}\mathbf{B}^2V^{-1}\mathbf{V}\mathbf{B}V^{-1}=\mathbf{V}\mathbf{B}^3V^{-1} \\
&\dots
\end{aligned}$$

d'où :

$$\begin{aligned}
e^{\mathbf{A}} &= \mathbf{1}+\mathbf{V}\mathbf{B}V^{-1}+\frac{1}{2!}\mathbf{V}\mathbf{B}^2V^{-1}+\frac{1}{3!}\mathbf{V}\mathbf{B}^3V^{-1}+\cdots \\
&= \mathbf{V}\left(\mathbf{1}+\mathbf{B}+\frac{1}{2!}\mathbf{B}^2+\frac{1}{3!}\mathbf{B}^3+\cdots\right)V^{-1} \\
&= \mathbf{V}e^{\mathbf{B}}V^{-1}
\end{aligned}$$

On en déduit que :

$$\mathbf{A}=\mathbf{V}\mathbf{B}V^{-1}\rightarrow e^{\mathbf{A}}=\mathbf{V}e^{\mathbf{B}}V^{-1}$$

5.2.3 Matrice diagonalisable

Supposons $\mathbf{A}=\mathbf{V}\Lambda V^{-1}$ avec Λ matrice diagonale. D'après la remarque faite en section précédente, on sait calculer e^{Λ} qui est la matrice diagonale ayant pour éléments diagonaux e^{λ_i}. Le calcul de $e^{\mathbf{A}}$ est alors immédiat.

5.2.4 Forme canonique de Jordan

Supposons $\mathbf{A} = \mathbf{VJV}^{-1}$ avec $\mathbf{J}$ forme canonique de Jordan. L'exponentielle d'une forme canonique de Jordan est une matrice bloc-diagonale, qui à chaque bloc de Jordan, associe un bloc triangulaire supérieur avec la correspondance suivante :

$$\mathbf{B} = \begin{pmatrix} b & 1 & 0 & 0 & 0 \\ 0 & b & 1 & 0 & 0 \\ 0 & 0 & b & 1 & 0 \\ 0 & 0 & 0 & b & 1 \\ 0 & 0 & 0 & 0 & b \end{pmatrix} \rightarrow e^{\mathbf{B}} = e^b \begin{pmatrix} 1 & 1 & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} \\ 0 & 1 & 1 & \frac{1}{2!} & \frac{1}{3!} \\ 0 & 0 & 1 & 1 & \frac{1}{2!} \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

5.2.5 L'approximation de Padé

L'approximation de Padé de e^x (avec x scalaire) sous forme d'une fraction de polynômes limités à l'ordre N s'écrit :

$$e^x \simeq \frac{P(x)}{Q(x)} = \frac{\sum_{k=0}^{N} a_k x^k}{\sum_{k=0}^{N} b_k x^k}$$

Il en résulte les approximations suivantes :

$$\begin{aligned} P(x) &\simeq e^x Q(x) \\ P'(x) &\simeq e^x \left[Q'(x) + Q(x)\right] \\ P''(x) &\simeq e^x \left[Q''(x) + 2Q'(x) + Q(x)\right] \\ &\cdots \end{aligned}$$

d'où le terme générique relatif à la dérivée d'ordre n :

$$e^x \sum_{k=0}^{n} C_n^k Q^{(k)}(x) = P^{(n)}(x)$$

où C_n^k désigne le coefficient du binôme :

$$C_n^k = \frac{n!}{k!\,(n-k)!}$$

et où $P^{(n)}(x)$ représente la n-ième dérivée de $P(x)$ qui s'écrit :

$$\begin{aligned} P^{(n)}(x) &= \sum_{k=n}^{N} k\,(k-1)\,(\cdots)\,(k-n+1)\,a_k x^{k-n} \\ &= \sum_{k=n}^{N} \frac{k!}{(k-n)!} a_k x^{k-n} \end{aligned}$$

Effectuons l'approximation au voisinage au voisinage de $x = 0$. On a :

$$\begin{aligned} P^{(n)}(0) &= n!a_n \triangleq A_n \\ Q^{(k)}(0) &= k!b_k \triangleq B_k \end{aligned}$$

Ecrivons les relations obtenues entre ces coefficients en faisant $x = 0$ dans les relations précédentes. Il en résulte un système d'équations dont l'équation générique s'écrit :

$$A_n = \sum_{k=0}^{n} C_n^k B_k \tag{5.4}$$

quel que soit n, avec $A_k = B_k = 0$ pour tout $k > N$.

Comme les a_k et b_k sont définis à un facteur constant multiplicatif près on peut choisir l'un d'entre eux. On n'a donc que $2N+1$ inconnues. Il suffit donc de prendre les $2N+1$ équations ci-dessus pour $n = 0$ à $2N$.

Les $N+1$ dernières équations s'écrivent :

$$\sum_{k=0}^{n} C_n^k B_k = A_n \text{ pour } n = N \text{ à } 2N$$

où le deuxième membre est connu si on se donne A_N. La matrice de ce système de $N+1$ équations à $N-1$ inconnues s'écrit :

$$M = \begin{pmatrix} C_N^1 & C_N^2 & C_N^3 & \cdots & C_N^N \\ C_{N+1}^1 & C_{N+1}^2 & C_{N+1}^3 & \cdots & C_{N+1}^N \\ C_{N+2}^1 & C_{N+2}^2 & C_{N+2}^3 & \cdots & C_{N+2}^N \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{2N}^1 & C_{2N}^2 & C_{2N}^3 & \cdots & C_{2N}^N \end{pmatrix}$$

On pourra vérifier que son déterminant est remarquable et vaut :

$$|M| = 1$$

Pour simplifier on choisit :

$$A_N = 1$$

Le second membre du système étant alors égal à $\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}^T$, il en résulte que pour résoudre, il suffit de calculer les mineurs de la première ligne, qui ont également une valeur remarquable. Le mineur du $(k+1)$-ième terme de la première ligne vaut :

$$\begin{vmatrix} C_{N+1}^1 & \cdots & C_{N+1}^k & C_{N+1}^{k+2} & \cdots & C_{N+1}^N \\ C_{N+2}^1 & \cdots & C_{N+2}^k & C_{N+2}^{k+2} & \cdots & C_{N+2}^N \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{2N}^1 & \cdots & C_{2N}^k & C_{2N}^{k+2} & \cdots & C_{2N}^N \end{vmatrix} = C_{2N-k}^N$$

Il en résulte que :

$$B_k = (-1)^k C_{2N-k}^N \text{ pour } k = 0 \text{ à } N$$

Ensuite on calcule les A_n à partir des relations (5.4). Ils ont également des valeurs remarquables :

$$A_n = \sum_{k=0}^{n} (-1)^k C_n^k C_{2N-k}^N = C_{2N-n}^N$$

Comme les a_n et b_n sont définis à un facteur constant multiplicatif près, on peut tous les diviser par C_{2N}^N pour avoir $a_0 = b_0 = 1$. Les valeurs traditionnellement utilisées sont :

$$a_n = (-1)^n b_n = \frac{1}{n!} \frac{(N!)}{(N-n)!} \frac{(2N-n)!}{(2N)!}$$

ou encore :

$$a_n = (-1)^n b_n = \frac{1}{n!} \frac{N(N-1)(\cdots)(N-n+1)}{2N(2N-1)(\cdots)(2N-n+1)}$$

Mise à l'échelle : L'approximation obtenue est faite autour de $x = 0$. Elle est d'autant plus précise que $|x|$ est faible. Si x est grand (en valeur absolue) devant 1, au lieu de calculer $y = e^x$ par l'approximation de Padé, on peut calculer $z = e^{x/2}$ par l'approximation de Padé et faire ensuite $y = z \times z$, ou encore mieux calculer $z = e^{x/2^n}$ et faire ensuite $y = z \times z \times \cdots \times z$, n fois. Mais ces multiplications introduisent des erreurs. Il y a donc un compromis à trouver. En fait, on se contente de ramener $x/2^n$ à une valeur inférieure (en valeur absolue) à $\frac{1}{2}$.

Application à l'exponentielle de matrice : On choisit généralement N de l'ordre de 6 (en double précision) à 10 (en quadruple précision). Les a_k et b_k étant pré-calculés, on calcule les puissances $\mathbf{A}^k$ de la matrice, puis les deux matrices :

$$\mathbf{P} = \left(\sum_{k=0}^{N} b_k \mathbf{A}^k \right) \text{ et } \mathbf{Q} = \left(\sum_{k=0}^{N} a_k \mathbf{A}^k \right)$$

et on approxime $e^{\mathbf{A}}$ par :

$$e^{\mathbf{A}} \simeq \mathbf{P}^{-1}\mathbf{Q}$$

Bibliographie

[1] Blum E. K. "*Numerical analysis and computation theory and practice*" Addison-Wesley. 1972.

[2] J.G.F. Francis, "*The QR Transformation - a unitary analogue to the LR transformation*". Computer journal. Volume 4, 1961. Part 1 pages 265-271, part II pages 332-345.

[3] Gene H Golub, Charles F. Van Loan, "*Matrix Computations*", Third Edition, Third Edition, Johns Hopkins University Press - April 1996.

[4] Householder A.S., "*Unitary Triangularization of a Nonsymmetric Matrix*", J. Assoc. Comput. March. 5 (1958), pages 339- 342.

[5] Kublanovskaya V.N. "*On some algorithms for the solution of the complete eigenvalue problem*". Z. Vycisl. Mat. i Mat. Fiz. Vol 1, pages 555-570, 1961.

[6] Dongara J.J., Moler C.B., Bunch J.R., Steward G.W. "*LINPACK users' Guide*". Siam, Philadelphia. 1979.

[7] Wilkinson J. H., Reinsch C. "*Handbook for Automatic Computation, Volume II : Linear Algebra*",Springer-Verlag, New-York 1971.

[8] Wilkinson J. H. "*The Algebraic Eigenvalue Problem*", Clarendo Press, Oxford 1968.

[9] Wilkinson J. "*Global convergence of tridiagonal QR algorith with origin shifts*", Linear Algebra and its Applications. 1, pages 409-420. 1968