Résolution de systèmes non réguliers par une méthode de décomposition en valeurs singulières

Résumé :

Ce document est consacré à la résolution des systèmes d'équations linéaires non réguliers. Les matrices prises en compte peuvent être carrées non inversibles ou rectangulaires.

Après avoir rappelé le cadre théorique des solutions au sens des moindres carrés, nous concentrons l'exposé sur la méthode par décomposition en valeurs singulières qui fournit, d'une part, un outil de diagnostic du degré de régularité du système, et, d'autre part, une famille d'algorithmes de résolution à la fois plus généraux et plus stables que ceux dérivant de l'approche par les équations normales.

Enfin, nous détaillons l'algorithme mis en œuvre dans le *Code_Aster* qui résorbe la fonctionnalité équivalente de la librairie *Nag* (F04JDF pour la version 12 et F04JDE pour la version 15) utilisée pour la modélisation du comportement métallurgique des aciers [R4.04.01].

Table des Matières

1 Introduction....3
2 Solution d'un système linéaire rectangulaire....3
2.1 Formalisme des moindres carrés....4
2.2 Existence d'optimums....4
2.3 Unicité de l'optimum et rang du système....5
2.4 Solution au sens des moindres carrés....5
3 Valeurs singulières....6
3.1 Décomposition en valeurs singulières....6
3.2 Rang, image et noyau....8
3.3 Pseudo-inverse et solution au sens des moindres carrés....8
4 Résolution d'un système linéaire rectangulaire....10
4.1 Méthode des équations normales....10
4.2 Méthode par décomposition en valeurs singulières....10
4.3 Comparaison de la méthode des équations normales à la méthode de décomposition en valeurs singulières....11
4.3.1 Conditionnement....11
4.3.2 Perte de précision....11
4.3.3 Structure creuse....12
4.3.4 Conclusion....12
5 Algorithme SVD pour la résolution d'un système linéaire équi ou sous-contraint....12
5.1 Réduction du problème et principe de l'algorithme....12
5.1.1 Réduction à la forme triangulaire supérieure....14
5.1.2 Réduction à la forme bi-diagonale supérieure....14
5.1.3 Décomposition SVD de la bi-diagonale supérieure....14
5.2 Réduction à la forme triangulaire supérieure....14
5.3 Réduction à la forme bi-diagonale supérieure....16
5.4 Décomposition SVD d'une bidiagonale supérieure....19
5.4.1 Principe de l'algorithme....19
5.4.2 Diagonalisation implicite de la matrice normale....20
5.4.3 Analyse de décomposition....22
5.4.4 Organisation de l'algorithme....22
6 Bibliographie....23
7 Description des versions du document....23

1 Introduction

Étant donnée une matrice réelle $\mathbf{A}$ d'ordre $m \times n$ et un vecteur $\mathbf{b}$ élément de $\mathbb{R}^m$, nous considérons le problème de la détermination d'un vecteur $\mathbf{x}$ élément de $\mathbb{R}^n$ qui vérifie le système linéaire suivant :

$$\mathbf{Ax} = \mathbf{b} \qquad \text{éq 1-1}$$

Il est bien connu ([bib3] p. 9) que ce système admet une, et une seule solution, pour tout b élément de $\mathbb{R}^m$ sous les conditions nécessaires et suffisantes qu'il soit équi-contraint $(m=n)$ et que sa matrice A soit régulière. Aussi, l'investigation des cas sous-contraint $(m \leq n)$ et sur-contraint $(m \geq n)$ nous confrontera à l'une des trois situations suivantes :

[1] Le système linéaire [éq 1-1] admet une solution et une seule,
[2] Le système linéaire [éq 1-1] n'admet pas de solution,
[3] Le système linéaire [éq 1-1] admet une infinité de solutions.

Dans la pratique, la situation 2) se rencontre en général dans le cas d'un système sur-contraint alors que les systèmes équi-contraints singuliers et sous-contraints conduisent en général à la situation 3).

Pour prétendre résoudre un système linéaire du type [éq 1-1], il nous faut d'abord définir ce que nous appellerons **solution**. Ceci est l'objet du **paragraphe 2** qui s'appuie principalement sur la notion de **moindres carrés** et sur **l'optimisation différentiable** pour définir, quelque soit le type de système, une solution qui est toujours unique.

Le **paragraphe 3** est consacré à **la décomposition en valeurs singulières** des matrices (en abrégé SVD : *Singular Value Decomposition*), qui, non seulement constitue un outil pour diagnostiquer laquelle des trois situations précédentes correspond au système linéaire étudié, mais aussi fournit une méthode de détermination de la solution définie dans le paragraphe 2.

La méthode utilisant la **décomposition SVD** est présentée au **paragraphe 4** et y est comparée à la méthode des **équations normales**.

Le **paragraphe 5** détaille sur le plan algébrique l'application de la méthode SVD à **la résolution d'un système linéaire équi ou sous-contraint** telle qu'elle est mise en œuvre dans le *Code_Aster*.

Dans les paragraphes suivants, nous utiliserons les notations ci-dessous :

- $\|\mathbf{x}\|$ et $(\mathbf{x},\mathbf{y})$ pour, respectivement, la norme euclidienne du vecteur $\mathbf{x}$ et le produit scalaire associé des vecteurs $\mathbf{x}$ et $\mathbf{y}$ éléments de $\mathbb{R}^m$ ou de $\mathbb{R}^n$,
- $\mathbf{M}^{\mathbf{T}}$ pour la transposée de la matrice $\mathbf{M}$,
- $\mathrm{Ker}\,\mathbf{M}$ et $\mathrm{Im}\,\mathbf{M}$ pour, respectivement, le noyau et l'image de (l'application linéaire associée à) la matrice $\mathbf{M}$,
- $\mathbf{X}^{\perp}$ pour l'orthogonal du sous espace $\mathbf{X}$ de $\mathbb{R}^m$ ou de $\mathbb{R}^n$.

2 Solution d'un système linéaire rectangulaire

Dans ce paragraphe nous allons définir une notion de *solution* pour le système linéaire [éq 1-1] qui jouit des propriétés d'**existence** et d'**unicité.** La démarche procède en deux temps :

- D'abord, par une approche du type moindres carrés nous construisons un problème d'optimisation différentiable et convexe (section 2.1) qui admet toujours au moins une solution (section 2.2). La situation 2) du paragraphe 1 est alors éliminée,

- Puis, analysant la propriété d'unicité (section 2.3) pour constater qu'elle n'est pas toujours garantie nous imposerons une contrainte supplémentaire (section 2.4) à la solution caractérisée dans la section 2.1 de façon à rétablir l'unicité.

2.1 Formalisme des moindres carrés

L'unique solution d'un système linéaire $\mathbf{Ax}=\mathbf{b}$ de matrice carrée et régulière réalise le minimum de la quantité $\|\mathbf{Ay}-\mathbf{b}\|$ quand $\mathbf{y}$ décrit $\mathbb{R}^n$. Cette propriété nous ouvre la voie qui conduit à une notion de solution pour un système linéaire général du type [éq 1-1] qui lui confère les mêmes propriétés que celles du cas particulier du système régulier. Nous dirons donc d'un point $\mathbf{x}$ de $\mathbb{R}^n$ qu'il est solution du système [éq 1-1] s'il est solution du **problème d'optimisation** :

$$\|\mathbf{Ax}-\mathbf{b}\|=\underset{\mathbf{y}\in\mathbb{R}^n}{\mathrm{Min}}\|\mathbf{Ay}-\mathbf{b}\| \qquad \text{éq 2.1-1}$$

Cette approche est naturelle car elle définit une solution dont le résidu $\mathbf{r}=\mathbf{Ax}-\mathbf{b}$ est nul dans le cas où le second membre est élément de $\mathrm{Im}\,\mathbf{A}$ et est de norme minimale dans le cas contraire, ce qui constitue le mieux qu'on puisse attendre.
Pour analyser le problème [éq 2.1-1], il est commode de lui substituer le problème d'optimisation sans contraintes équivalent suivant :

$$\text{trouver } \mathbf{x}\in\mathbb{R}^n \text{ tel que } \mathrm{J}(\mathbf{x})=\underset{\mathbf{y}\in\mathbb{R}^n}{\mathrm{Min}}\mathrm{J}(\mathbf{y}) \qquad \text{éq 2.1-2}$$

où $\mathrm{J}(.)$ est la fonctionnelle définie par :

$$\mathrm{J}:\mathbf{y}\in\mathbb{R}^n\rightarrow\mathrm{J}(\mathbf{y})=\frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|$$

L'intérêt du problème [éq 2.1-2] tient au fait que la fonctionnelle $\mathrm{J}(.)$ vérifie les propriétés suivantes :

- $\mathrm{J}(.)$ est deux fois continument différentiable :

$$\mathrm{DJ}(\mathbf{x}):\mathbf{h}\in\mathbb{R}^n\rightarrow\mathrm{DJ}(\mathbf{x})\,\mathbf{h}=\left(\mathbf{A}^{\mathbf{T}}\mathbf{Ax}-\mathbf{A}^{\mathbf{T}}\mathbf{b},\mathbf{h}\right)\in\mathbb{R} \qquad \text{éq 2.1-3}$$

$$D^2\mathrm{J}(\mathbf{x}):(\mathbf{h},\mathbf{k})\in\mathbb{R}^n\times\mathbb{R}^n\rightarrow\mathrm{DJ}(\mathbf{x})(\mathbf{h},\mathbf{k})=\left(\mathbf{A}^{\mathbf{T}}\mathbf{Ah},\mathbf{k}\right)\in\mathbb{R} \qquad \text{éq 2.1-4}$$

- $\mathrm{J}(.)$ est quadratique et convexe.
-

Ainsi, le problème [éq 2.1-2] s'inscrit dans le cadre de l'optimisation différentiable et convexe de sorte que nous disposons des résultats suivants ([bib1] p. 156 et 146) :

1) De la convexité : tout optimum local est en fait un optimum global, c'est à dire une solution de [éq 2.1-2],
2) De la différentiabilité : tout optimum local vérifie l'équation d'Euler $\mathrm{DJ}(\mathbf{x})=0$ sur Än qui, compte-tenu de [éq 2.1-3], conduit à la caractérisation par les **équations** dites **normales** :

$$\mathbf{A}^{\mathbf{T}}\mathbf{Ax}=\mathbf{A}^{\mathbf{T}}\mathbf{b} \qquad \text{éq 2.1-5}$$

2.2 Existence d'optimums

Dans [bib1] p. 171 on trouve une démonstration de l'existence d'au moins une solution aux équations normales [éq 2.1-5]. Cette démonstration s'appuie sur des arguments destinés à la prise en compte du cas de la dimension infinie (théorème de projection sur un convexe fermé d'un espace de Hilbert).

Notre cas étant nettement plus simple, nous donnons une démonstration du résultat qui n'utilise que des arguments algébriques simples qui, de plus, nous serons utiles dans le paragraphe 3. Montrer que, pour tout $\mathbf{b}$ élément de $\mathbb{R}^m$, les équations normales [éq 2.1-5] admettent une solution équivaut à l'établissement de l'inclusion $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\subset\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Or, pour toute matrice réelle $\mathbf{M}$ d'ordre $m\times n$ nous avons $\mathrm{Im}\,\mathbf{M}^{\mathbf{T}}=(\mathrm{Ker}\,\mathbf{M})^{\perp}$ ([bib3] p. 28). Aussi, l'inclusion à établir qui équivaut à

$(\mathrm{Ker}\,\mathbf{A})^{\perp}\subset(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A})^{\perp}$ qui est elle même équivalente à $\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\subset\mathrm{Ker}\,\mathbf{A}$. Soit donc $\mathbf{x}\in\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$; alors $\mathbf{Ax}\in\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}$, c'est-à-dire $\mathbf{Ax}\in(\mathrm{Im}\,\mathbf{A})^{\perp}$. Comme $\mathbf{Ax}$ est aussi élément de $\mathrm{Im}\,\mathbf{A}$, il ne peut être que nul ce qui signifie que $\mathbf{x}\in\mathrm{Ker}\,\mathbf{A}$ et achève la démonstration.

A ce stade du propos, nous pouvons dire que tout système du type [éq 1-1] admet au moins une solution au sens de [éq 2.1-3] et toutes ces solutions sont caractérisées comme solution au sens de Cramer des équations normales de [éq 2.1-5]. La situation 2) du paragraphe 1 est éliminée.

Reste à éliminer la situation 3), c'est-à-dire à garantir l'unicité.

2.3 Unicité de l'optimum et rang du système

Il est clair que les équations normales [éq 2.1-5], caractérisant les optimums que nous cherchons, admettent une unique solution sous la condition nécessaire et suffisante que $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ soit régulière. Comme $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ est toujours semi-définie positive, son inversibilité équivaut à sa définie positivité, de sorte que, compte tenu de l'expression [éq 2.1-4] de la dérivée seconde de la fonctionnelle $\mathrm{J}(.)$, nous retrouvons le théorème bien connu d'unicité de l'optimum du problème [éq 2.1-2] pour une fonctionnelle convexe deux fois continument différentiable ([bib1] th 7.4-3 et 7.4-4).

En toute généralité, rien n'empêche la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ d'être singulière, la solution du système [éq 1-1] au sens de [éq 2.1-2] n'est donc pas toujours unique. Nous disposons néanmoins d'un critère pour détecter cette situation. A la section 2.2 nous avons établi que $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\subset\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$ et comme l'inclusion réciproque est trivialement vraie, nous pouvons conclure à l'identité $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}=\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. L'introduction du **rang** $\mathrm{rg}(\mathbf{A})$ de la matrice $\mathbf{A}$, la dimension de son espace image, nous permet alors de dire qu'une condition nécessaire et suffisante pour que $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ soit inversible est que $\mathrm{rg}(\mathbf{A}^{\mathbf{T}}\mathbf{A})=n$ ce qui équivaut à $\mathrm{rg}(\mathbf{A})=n$ car $\mathrm{rg}(\mathbf{A}^{\mathbf{T}}\mathbf{A})=\mathrm{rg}(\mathbf{A}^{\mathbf{T}})=\mathrm{rg}(\mathbf{A})$.

L'intérêt de ce critère tient au fait qu'il limite l'analyse à la seule matrice $\mathbf{A}$ sans qu'il soit nécessaire de former explicitement $\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Ce critère nous montre aussi que les équations normales associées à un système linéaire strictement sous-contraint admettent toujours une infinité de solutions. En effet, le rang d'une matrice est aussi égal au nombre de colonnes indépendantes qu'elle possède ; aussi, pour que ce rang atteigne la valeur n il est nécessaire que les colonnes de la matrice soit d'ordre au moins n.

2.4 Solution au sens des moindres carrés

Nous venons de constater que l'ensemble des points qui minimisent le résidu du système [éq 1-1] n'est pas nécessairement réduit à un seul point. Pour rétablir l'unicité nous affinons la notion de solution du système [éq 1-1] de la section 2.1 en définissant la **solution au sens des moindres carrés** comme l'élément de norme minimale de l'ensemble des points qui minimisent le résidu. Cette solution x est alors caractérisée par :

$$\mathbf{x}\in\mathbf{S}\overset{def}{=}\left\{\mathbf{y}\in\mathbb{R}^{n}\,;\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\mathbf{x}=\mathbf{A}^{\mathbf{T}}\mathbf{b}\right\}\ \text{et}\ \|\mathbf{x}\|=\inf_{y\in\mathbf{S}}\|\mathbf{y}\|$$

Cette caractérisation n'est pas satisfaisante sur le plan pratique car elle demande la résolution d'un problème d'optimisation sous contraintes. Nous allons lui substituer une autre caractérisation plus adaptée au sens où elle conduira (voir la section 4.2) à une procédure de calcul nettement plus simple.

L'ensemble $\mathbf{S}$ ci-dessus est le translaté de noyau de $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ par l'un quelconque des vecteurs solutions des équations normales [éq 2.1-5]. Aussi, la condition supplémentaire de minimisation de la norme s'interprète comme une simple projection : la solution au sens des moindres carrés du système [éq 1-1] n'est rien d'autre que la projection de l'origine de Än sur l'ensemble des solutions des

équations normales. Aussi, nous pouvons la caractériser comme le point d'intersection entre l'ensemble $\mathbf{S}$ et l'orthogonal du noyau de $\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

La définition d'une solution au système [éq 1-1] peut alors être résumée comme suit :

$$\mathbf{x} \text{ est solution de } \mathbf{Ax}=\mathbf{b} \Leftrightarrow \begin{cases} \mathbf{A}^{\mathbf{T}}\mathbf{Ax}=\mathbf{A}^{\mathbf{T}}\mathbf{b} \\ \mathbf{x}\in\left(\operatorname{Ker}\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp} \end{cases} \qquad \text{éq 2.4-1}$$

La première condition fait de $\mathbf{x}$ un vecteur de résidu minimal tandis que la deuxième sélectionne, parmi les vecteurs de résidu minimal, celui de norme minimale.

La définition [éq 2.4-1] est une généralisation classique de la notion de solution d'un système équi-contraint régulier et confère à tout système du type [éq 1-1] une solution et une seule.

3 Valeurs singulières

Dans ce paragraphe nous présentons quelques résultats utiles pour la conception d'une méthode de résolution opérationnelle du système [éq 1-1]. Ces résultats dérivent de la notion de valeurs singulières (section 3.1) et permettent de construire une base du noyau et une base de l'image de la matrice du système (section 3.2) à partir desquelles il est possible de donner un sens, adapté au calcul de la solution au sens des moindres carrés, à l'inverse d'une matrice quelconque (section 3.3).

3.1 Décomposition en valeurs singulières

Commençons par rappeler la définition des valeurs singulières. On appelle **valeurs singulières** d'une matrice réelle $\mathbf{A}$ d'ordre $m\times n$ les racines carrées des valeurs propres de la matrice carrée $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ d'ordre n qui, rappelons-le, est semi-définie positive.

La notion de diagonalisation des matrices carrées (lorsqu'elles sont diagonalisables) se généralise aux matrices rectangulaires (sans restriction) par le concept de décomposition (ou factorisation) en valeurs singulières.

Pour toutes matrices réelles $\mathbf{A}$ d'ordre $m\times n$, il existe deux matrices carrées unitaires $\mathbf{Q}$ et $\mathbf{P}$ d'ordre respectif m et n telles que :

$$\mathbf{A}=\mathbf{Q}\,\Sigma\,\mathbf{P}^{\mathbf{T}} \qquad \text{éq 3.1-1}$$

où Σ est une matrice d'ordre $m\times n$ dont la structure est schématisée ci-dessous :

$$\Sigma=\left|\begin{array}{ccc|c} \mu_1 & & & \\ & \mu_2 & & 0 \\ & & \ldots & \\ & & & \mu_n \end{array}\right| \quad \text{si } m\le n$$

$$\Sigma=\left|\begin{array}{ccc} \mu_1 & & \\ & \mu_2 & \\ & & \ldots \\ & & \mu_n \\ \hline & 0 & \end{array}\right| \quad \text{si } m>n$$

Les μ_i sont les valeurs singulières de $\mathbf{A}$ que nous supposons ordonnées par ordre décroissant :

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$$

On peut trouver une démonstration de ce résultat dans [bib1] p.10 pour le cas équi-contraint et dans [bib3] p.73 pour le cas sur-contraint, le cas sous-contraint s'en déduit alors par transposition.

La factorisation SVD [éq 3.1-1] de $\mathbf{A}$ donne $\mathbf{A}^{\mathbf{T}}\mathbf{A}=\mathbf{P}\,\boldsymbol{\Sigma}^{\mathbf{T}}\mathbf{S}\mathbf{P}^{\mathbf{T}}$ et $\mathbf{A}\mathbf{A}^{T}=\mathbf{Q}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}$ de sorte que, $\boldsymbol{\Sigma}^{\mathbf{T}}\boldsymbol{\Sigma}$ et $\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{\mathbf{T}}$ étant des matrices carrées diagonales, la matrice $\mathbf{P}$ est constituée des vecteurs propres orthonormalisés de la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ tandis que la matrice $\mathbf{Q}$ est constituée des vecteurs propres orthonormalisés de la matrice $\mathbf{A}\mathbf{A}^{\mathbf{T}}$.

3.2 Rang, image et noyau

Le paragraphe [§2] a montré le rôle fondamental que joue le rang de la matrice $\mathbf{A}$ et le noyau de la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ pour la résolution d'un système linéaire non régulier du type [éq 1-1]. Nous allons voir maintenant comment la factorisation [éq 3.1-1] peut être utilisée pour déterminer ce rang ainsi qu'une base de $\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

Soit r l'indice de la plus petite valeur singulière non nulle. La factorisation [éq 3.1-1] s'écrit aussi $\mathbf{Q}^{\mathbf{T}}\mathbf{A}\mathbf{P}=\boldsymbol{\Sigma}$ où la prise en compte des valeurs singulières nulles permet de préciser la décomposition en bloc de $\boldsymbol{\Sigma}$:

$$\Sigma = \left[\begin{array}{c|c|c} \Sigma_r & & 0 \\ \hline 0 & 0 & \end{array}\right] \quad \text{si } m \leq n$$

$$\Sigma = \left[\begin{array}{c|c} \Sigma_r & 0 \\ \hline 0 & 0 \\ \hline \multicolumn{2}{c}{0} \end{array}\right] \quad \text{si } m > n$$

où $\boldsymbol{\Sigma}_r=\mathrm{Diag}\left(\mu_1,\mu_2,...,\mu_r\right)$ est la matrice diagonale d'ordre r des valeurs singulières non nulles dans l'ordre croissant.

Puisque les matrices $\mathbf{Q}$ et $\mathbf{P}$ sont régulières, les matrices $\mathbf{A}$ et $\boldsymbol{\Sigma}$ sont équivalentes de sorte que leur noyau et image respectifs coïncident. Nous en déduisons donc que :

- Le rang de $\mathbf{A}$ coïncide avec le nombre de valeurs singulières non nulles :

$$\mathrm{rg}\,\mathbf{A}=r$$

- Les vecteurs colonnes de $\mathbf{P}$ d'indice $r+1$ à n forment une base de $\mathrm{Ker}\,\mathbf{A}$
- Les vecteurs colonnes de $\mathbf{Q}$ correspondant aux valeurs singulières non nulles forment une base de $\mathrm{Im}\,\mathbf{A}$

D'autre part, à la section 2.3 nous avons vu que $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}=\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. L'identité $\mathrm{Im}\,\mathbf{M}^{\mathbf{T}}=\left(\mathrm{Ker}\,\mathbf{M}\right)^{\perp}$ nous donne alors $\mathrm{Ker}\,\mathbf{A}=\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$ de sorte que la deuxième condition de la définition [éq 2.4-1] est simplement réalisée par tout vecteur qui s'exprime comme une combinaison linéaire des vecteurs colonnes de $\mathbf{P}$ correspondant aux valeurs singulières non nulles.

3.3 Pseudo-inverse et solution au sens des moindres carrés

Une autre application de la décomposition en valeurs singulières consiste en la notion de **pseudo-inverse** (ou inverse Moore-Penrose) qui généralise la notion habituelle d'inverse d'une matrice carrée régulière aux matrices rectangulaires d'une part, et aux matrices carrées singulières d'autre part.

Tout d'abord, la pseudo-inverse d'une matrice Σ de la décomposition en valeurs singulières [éq 3.1-1] est définie par :

$$\Sigma^{+}=\left[\begin{array}{c|c|c} \Sigma_r^{-1} & & 0 \\ \hline 0 & 0 & \end{array}\right] \quad \text{si } m \leq n$$

$$\Sigma^{+}=\left[\begin{array}{c|c} \Sigma_r^{-1} & 0 \\ \hline 0 & 0 \\ \hline \multicolumn{2}{c}{0} \end{array}\right] \quad \text{si } m > n$$

où $\Sigma_r^{-1}=\mathrm{Diag}\left(\frac{1}{\mu_1},\frac{1}{\mu_2},\ldots,\frac{1}{\mu_r}\right)$ est l'inverse au sens habituel de $\mathbf{S}_r$.

Ceci étant, nous utilisons la décomposition [éq 3.1-1] de la matrice $\mathbf{A}$ pour définir sa pseudo-inverse A^{+} par :

$$\mathbf{A}^{+}=\mathbf{P}\,\Sigma^{+}\mathbf{Q}^{\mathbf{T}} \qquad \text{éq 3.3-1}$$

De même, de la décomposition [éq 3.1-1] de la matrice $\mathbf{A}$ nous tirons $\mathbf{A}^{\mathbf{T}}\mathbf{A}=\mathbf{P}\,\Sigma^{\mathbf{T}}\Sigma\,\mathbf{P}^{\mathbf{T}}$, de sorte que la pseudo-inverse $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}$ de la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ est définie par :

$$\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}=\mathbf{P}\,\Sigma^{+}\left(\Sigma^{\mathbf{T}}\right)^{+}\mathbf{P}^{\mathbf{T}} \qquad \text{éq 3.3-2}$$

Nous sommes maintenant en mesure de fournir une interprétation simple de la solution au sens des moindres carrés définie par [éq 2.4-1].

La restriction à $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ de l'application linéaire associée à la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ définit un isomorphisme de $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ sur $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Comme, d'une part $\left(\mathrm{Ker}\,A^{T}A\right)^{\perp}=\left(\mathrm{Ker}\,A\right)^{\perp}=\mathrm{Im}\,A^{T}$, et, d'autre part, $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}=\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}$, cette restriction est en fait un automorphisme de $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$.

Dans la base de $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ constituée par les r premières colonnes de la matrice $\mathbf{P}$, cet automorphisme est représenté par la matrice Σ_r^{2}. Aussi, son automorphisme réciproque y est représenté par la matrice Σ_r^{-2}. L'extension à $\mathbb{R}^n$ de cet automorphisme est alors représentée, dans la base associée à la matrice $\mathbf{P}$, par la matrice $\left(\Sigma^{\mathbf{T}}\Sigma\right)^{+}=\Sigma^{\mathbf{T}}\left(\Sigma^{\mathbf{T}}\right)^{+}$, et donc, dans la base canonique, par la matrice $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}$.

Il suit que :

- Nous retrouvons le fait que, pour tout $\mathbf{b}$ élément de $\mathbb{R}^n$, il existe un unique vecteur $\mathbf{x}\in\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ solution de $\mathbf{A}^{\mathbf{T}}\mathbf{A}\mathbf{x}=\mathbf{A}^{\mathbf{T}}\mathbf{b}$, soit l'existence et l'unicité de la solution au sens des moindres carrés [éq 2.4-1] du système $\mathbf{A}\mathbf{x}=\mathbf{b}$,
- Cette unique solution est donnée par :

$$\mathbf{x}=\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}\mathbf{b} \qquad \text{éq 3.3-3}$$

La pseudo-inverse d'une matrice est définie à partir de la décomposition SVD de cette matrice. Comme la décomposition SVD n'est pas unique, la matrice pseudo-inverse n'est pas unique. Par contre, du point de vue des applications linéaires associées aux matrices, l'application pseudo-inverse est unique. Toutes les matrices pseudo-inverses associées aux différentes décompositions SVD d'une matrice donnée ne sont alors que des représentantes matricielles particulières qui expriment cette application pseudo-inverse relativement aux bases induites par les matrices orthogonales des décompositions SVD. Aussi, l'expression [éq 3.3-3] a un sens : elle définit un vecteur dont $\mathbf{x}$ représente les composantes par rapport à la base d'arrivée (matrice $\mathbf{P}$) de la décomposition SVD.

4 Résolution d'un système linéaire rectangulaire

Les deux méthodes de résolution du système [éq 1-1] que nous présentons aux sections 4.1 (méthode des équations normales) et 4.2 (décomposition en valeurs singulières) visent à la résolution des équations normales [éq 2.1-5]. Ces deux méthodes se distinguent non seulement par le choix des algorithmes qu'elles mettent en œuvre (inversion contre pseudo-inversion), mais aussi par leur degré de généralité et par leurs propriétés numériques qui sont comparées à la section 4.3

4.1 Méthode des équations normales

La résolution du système $\mathbf{A}\mathbf{x}=b$ par la méthode des équations normales consiste à calculer la solution au sens des moindres carrés [éq 2.4-1] de façon "directe", c'est-à-dire en utilisant directement la relation $x=\left(A^T A\right)^{-1}A^T b$. Pour cela, il s'agit d'abord de calculer $A^T A$ et $A^T b$, puis de résoudre le système obtenu soit par méthode itérative soit par une factorisation de $A^T A$.

Nous pouvons d'ores et déjà remarquer que cette méthode est limitée aux matrices $A^T A$ régulières, ce qui limite son domaine d'application au système [éq 1-1] dont la matrice est de plein rang (voir la section 2.3). En particulier, la méthode des équations normales ne peut traiter ni les systèmes strictement sous-contraints, ni les systèmes équi-contraints singuliers (voir la section 2.4).

4.2 Méthode par décomposition en valeurs singulières

Nous avons vu à la section 3.3 que la solution du système $\mathbf{A}\mathbf{x}=\mathbf{b}$ au sens des moindres carrés définie par [éq 2.4-1] peut être caractérisée par la relation $\mathbf{x}=\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}\mathbf{b}$ [éq 3.3-3]. La méthode de résolution du système basée sur cette propriété **est dite méthode par décomposition en valeurs singulières** car elle construit la pseudo-inverse [éq 3.3-2] de $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ via la décomposition SVD [éq 3.1-1] de la matrice $\mathbf{A}$.

Comme toute matrice peut être décomposée en valeurs singulières, il suit que tout système du type [éq 1-1] peut être résolu au sens de [éq 2.4-1] par cette méthode qui présente donc, au moins, l'avantage de la généralité par rapport à la méthode des équations normales.

Ce n'est pas tout. La méthode par décomposition en valeurs singulières, contrairement à la méthode des équations normales, ne demande pas la construction explicite de la matrice $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ et du vecteur

$\mathbf{A}^{\mathbf{T}}\mathbf{b}$ (nous verrons à la section 4.3 l'intérêt sur le plan numérique de cette propriété). En effet, il est aisé de vérifier que la matrice $\mathbf{\Sigma}$ des valeurs singulières de la factorisation [éq 3.1-1] satisfait à l'identité $\mathbf{\Sigma}^{+}\left(\mathbf{\Sigma}^{\mathbf{T}}\right)^{+}\mathbf{\Sigma}^{\mathbf{T}}=\mathbf{\Sigma}^{+}$, de sorte que, prémultipliant $\mathbf{A}^{\mathbf{T}}$ par la pseudo-inverse de $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ et tenant compte de la factorisation [éq 3.1-1], nous obtenons $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}=\mathbf{P}\,\mathbf{\Sigma}^{+}\left(\mathbf{\Sigma}^{\mathbf{T}}\right)^{+}\mathbf{\Sigma}^{\mathbf{T}}\mathbf{P}^{\mathbf{T}}\,\mathbf{P}\,\mathbf{\Sigma}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}$, qui, par orthogonalité de $\mathbf{P}$ nous donne $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}=\mathbf{A}^{+}$. Dès lors, les caractérisations [éq 3.3-3] et [éq 2.4-1] de la solution cherchée sont équivalentes à la caractérisation :

$$\mathbf{x}\ \text{est solution de}\ \mathbf{A}\mathbf{x}=\mathbf{b}\Leftrightarrow\mathbf{x}=\mathbf{A}^{+}\mathbf{b} \qquad \text{éq 4.2-1}$$

4.3 Comparaison de la méthode des équations normales à la méthode de décomposition en valeurs singulières

Dans les deux sections précédentes, nous venons de constater, qu'algébriquement, la méthode de décomposition en valeurs singulières est plus générale et plus simple que la méthode des équations normales. Nous allons maintenant constater, en suivant [bib3] p. 336, qu'elle lui est aussi supérieure sur le plan numérique. Cette supériorité s'exprime d'une part, en terme de stabilité non seulement de la résolution, mais aussi de la construction du problème et, d'autre part, à un niveau moins critique, en terme d'adaptation au traitement des matrices creuses.

4.3.1 Conditionnement

Le conditionnement d'une matrice $\mathbf{A}$ d'ordre $m\times n$ est défini comme le rapport de ses valeurs singulières extrêmes et non nulles :

$$\mathrm{cond}\left(\mathbf{A}\right)=\frac{\mu_1}{\mu_r}$$

où r est le rang de la matrice $\mathbf{A}$.

Les résultats présentés dans [bib3] p.184, utilisant les équations normales comme un outil d'analyse et non comme un outil de calcul, montrent que la perturbation de la solution du problème d'optimisation [éq 2.1-1] due aux erreurs d'arrondis peut être proportionnelle à $\mathrm{cond}\left(\mathbf{A}\right)^2$. Mais les résultats classiques de l'analyse de stabilité de la solution d'un système linéaire par rapport à ces mêmes erreurs montrent une proportionnalité au nombre de conditionnement de la matrice. Si bien que dans le cas d'une résolution directe des équations normales, nous obtenons une erreur toujours proportionnelle à $\mathrm{cond}\left(\mathbf{A}^{\mathbf{T}}A\right)=\mathrm{cond}\left(\mathbf{A}\right)^2$, ce qui est moins bon que $\mathrm{cond}\left(\mathbf{A}\right)$.

La méthode de résolution par décomposition en valeurs singulières n'utilise que des transformations orthogonales (voir le paragraphe 4), si bien qu'elle ne modifie pas le conditionnement initial du problème [éq 1-1] et est donc, de ce point de vue, plus attrayante que la méthode des équations normales.

4.3.2 Perte de précision

Nous venons de voir que les erreurs d'arrondis conduisent à une dégradation de la solution plus sensible lorsqu'elle est calculée via les équations normales plutôt que par une décomposition en valeurs singulières. L'exemple suivant, tiré de [bib 2], montre que la construction même du système [éq 2.1-5] des équations normales est perturbée par les erreurs d'arrondi.

Soit donc la matrice suivante :

$$\mathbf{A}=\begin{bmatrix}1 & 1\\ \varepsilon & 0\\ 0 & \varepsilon\end{bmatrix}\ \text{conduisant à}\ \mathbf{A}^{\mathbf{T}}\mathbf{A}=\begin{bmatrix}1+\varepsilon^2 & 1\\ 1 & 1+\varepsilon^2\end{bmatrix}$$

dont les valeurs singulières sont $\mu_1=\sqrt{2+\varepsilon^2}$ et $\mu_2=|\varepsilon|$, de sorte que le rang de $\mathbf{A}$ est 2 dès que $\varepsilon\neq 0$. Si ε vérifie $\varepsilon^2<\varepsilon_{mach}<\varepsilon$ où ε_{mach} est la précision machine, alors tous les coefficients de $\mathbf{A}^T\mathbf{A}$ seront calculés à la valeur 1 et les valeurs singulières calculées seront, au mieux, $\mu_1=\sqrt{2}$ et $\mu_2=0$. Il suit que le rang numérique, calculé par les équations normales, sera 1, alors que celui calculé par une décomposition SVD de la matrice $\mathbf{A}$ serait égal à 2

4.3.3 Structure creuse

A un niveau moindre, la construction de la matrice des équations normales induit un remplissage du système associé que la méthode utilisant la décomposition en valeurs singulières évite.

4.3.4 Conclusion

Le tableau suivant résume la discussion des sous-sections précédentes :

	Généralité	Conditionnement	Perte de précision à la construction du problème	remplissage
Équations normales	systèmes de plein rang	$\mathrm{cond}(\mathbf{A})^2$	possible	oui
SVD	tout système	$\mathrm{cond}(\mathbf{A})$	impossible	non

5 Algorithme SVD pour la résolution d'un système linéaire équi ou sous-contraint

Dans ce paragraphe, nous détaillons la méthode de résolution des systèmes non-réguliers mise en œuvre dans le *Code_Aster*. Cette méthode s'applique au système sous-contraint ou équi-contraint singulier et fournit la solution au sens des moindres carrés [éq 2.4-1].

Le calcul d'une décomposition SVD de $\mathbf{A}$ est équivalent au calcul du spectre de la matrice normale associée $\mathbf{A}^T\mathbf{A}$. Aussi, il ne peut être obtenu qu'à la convergence d'un procédé itératif.

La **section 5.1** expose le **principe de l'algorithme** et montre en particulier comment l'application de deux transformations orthogonales permet de **réduire le problème** à la simple recherche de la décomposition SVD d'une matrice bi-diagonale supérieure. Les **sections 5.2** et **5.3** sont consacrées à **l'algorithmique** de ces **réductions**. La **section 5.4** présente l'algorithme de la **décomposition SVD** de la matrice bi-diagonale.

Les algorithmes seront décrits avec la convention de notation dans laquelle :

- $\mathrm{R}(i,j,\theta)$ désigne la rotation de Givens du plan (i,j) et d'angle θ,
- $\mathbf{A}^{(k)}$ désigne l'itéré d'indice k d'une itération matricielle et $\mathbf{A}^{(k,l)}$ l'itéré l d'une itération interne à l'itéré $\mathbf{A}^{(k)}$.

5.1 Réduction du problème et principe de l'algorithme

Dans cette section, nous présentons l'algorithme de résolution d'un sytème linéaire équi ou sous-contraint par la méthode SVD.

Nous réduisons le problème à la recherche de la décomposition SVD d'une matrice bidiagonale comme dans [bib2] mais nous effectuons la réduction d'une autre façon que celle proposée dans [bib2] : nous commençons par réduire la matrice à une forme triangulaire supérieure, puis, nous réduisons cette triangulaire à une forme bidiagonale supérieure. Ces deux réductions sont effectuées par transformations orthogonales.

Les opérations de **calcul de la décomposition SVD** s'enchaînent comme suit :

5.1.1 Réduction à la forme triangulaire supérieure

$$\begin{aligned} \mathbf{A}=[\mathbf{U} \quad \mathbf{0}]\mathbf{P}_1^{\mathrm{T}} \quad & \text{si} \quad m<n \\ \mathbf{A}=\mathbf{U}\mathbf{P}_1^{\mathrm{T}} \quad & \text{si} \quad m=n \end{aligned} \qquad \textbf{éq 5.1-1}$$

où $\mathbf{P_1}$ est une matrice orthogonale d'ordre n et $\mathbf{U}$ une matrice triangulaire supérieure d'ordre m.

5.1.2 Réduction à la forme bi-diagonale supérieure

$$\mathbf{U}=\mathbf{Q}_2\mathbf{B}\mathbf{P}_2^{\mathrm{T}} \qquad \text{éq 5.1-2}$$

où $\mathbf{Q_2}$ et $\mathbf{P_2}$ sont deux *matrices orthogonales* d'ordre m et $\mathbf{B}$ une *matrice bidiagonale* supérieure d'ordre m.

5.1.3 Décomposition SVD de la bi-diagonale supérieure

$$\mathbf{B}=\mathbf{Q}_3\mathbf{\Sigma}\,\mathbf{P}_3^{\mathrm{T}} \qquad \text{éq 5.1-3}$$

où $\mathbf{Q_3}$ et $\mathbf{P_3}$ sont deux *matrices orthogonales* d'ordre m et $\mathbf{\Sigma}$ une *matrice diagonale* d'ordre m de la forme :

$$\Sigma = \begin{array}{|c|c|}\hline \begin{matrix}\mu_1 & & \\ & \ddots & \\ & & \mu_2\end{matrix} & \begin{matrix}0\end{matrix} \\ \hline \begin{matrix}0\end{matrix} & \begin{matrix}0 & & \\ & \ddots & \\ & & 0\end{matrix} \\ \hline \end{array} = \begin{array}{|c|c|}\hline \Sigma_r & 0 \\ \hline 0 & 0 \\ \hline \end{array}$$

Combinant les relations [éq 5.1-1], [éq 5.1-2] et [éq 5.1-3], nous obtenons une **décomposition SVD de la matrice** $\mathbf{A}$:

$$\mathbf{A}=\mathbf{Q}_2\mathbf{Q}_3 \begin{array}{|cc|c|}\hline \Sigma_r & 0 & \\ 0 & 0 & 0 \\ \hline \end{array} \begin{array}{|cc|}\hline \mathbf{P}_3^T\mathbf{P}_2^T & 0 \\ \hline 0 & \mathbf{I} \\ \hline \end{array} \mathbf{P}_1^T \qquad \text{éq 5.1-4}$$

La **solution au sens des moindres carrés** [éq 2.4-1] du système [éq 1-1] est alors obtenue par l'application de la pseudo-inverse [éq 3.3-1] de A déduite de la décomposition en valeurs singulières [éq 5.1-4]. Nous obtenons donc :

$$\mathbf{x}=\mathbf{P}_1\begin{bmatrix}\mathbf{P}_2\mathbf{P}_3\mathbf{S}^{+}\mathbf{Q}_3^{\mathrm{T}}\mathbf{Q}_2^{\mathrm{T}}\mathbf{b} \\ 0\end{bmatrix} \qquad \text{éq 5.1-5}$$

L'algorithme proposé consiste donc en l'enchaînement des factorisations [éq 5.1-1], [éq 5.1-2] et [éq 5.1-3] préalablement à l'application de la relation [éq 5.1-5].

5.2 Réduction à la forme triangulaire supérieure

A partir d'une matrice $\mathbf{A}$ d'ordre $m\times n$ pour $m\le n$ on détermine une matrice triangulaire supérieure $\mathbf{U}$ d'ordre m et une matrice orthogonale $\mathbf{P}$ d'ordre n telles que :

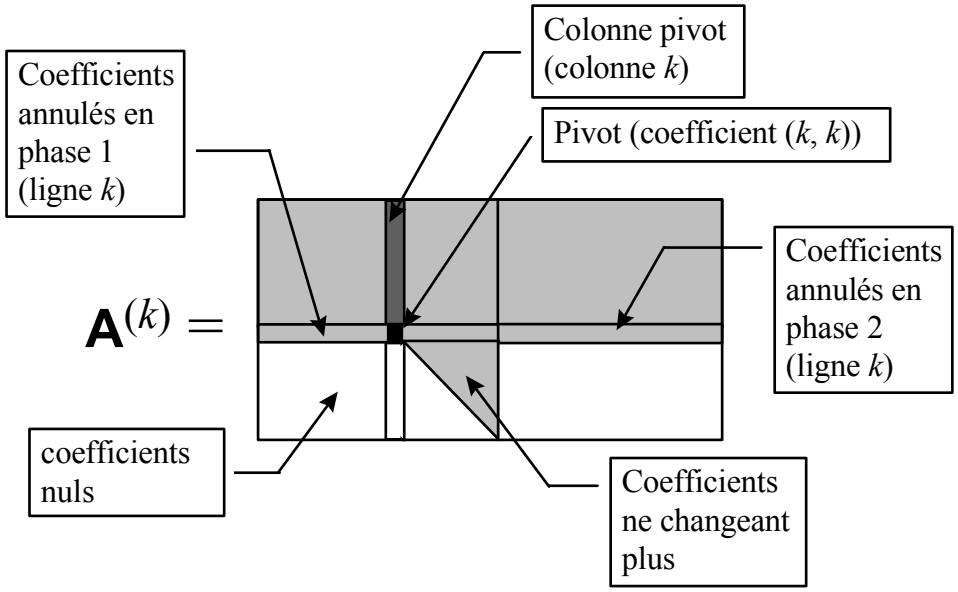
$$\mathbf{A}=\begin{bmatrix}\mathbf{U} & 0\end{bmatrix}\mathbf{P}^{\mathrm{T}} \quad \text{si} \quad m<n$$

$$\mathbf{A}=\mathbf{U}\mathbf{P}^{\mathrm{T}} \quad \text{si} \quad m=n$$

Algorithmiquement, la factorisation utilise une méthode d'élimination qui s'interprète, comme la construction d'une suite de matrices $\mathbf{A}^{(k)}$ par :

$$\begin{cases} \mathbf{A}^{(\mathbf{m})} = \mathbf{A} \\ \mathbf{A}^{(k-1)} = \mathbf{A}^{(k)} \mathbf{P}^{(k)} \quad \text{pour} \quad k = m, m-1, \ldots, 1 \end{cases}$$

où chaque matrice courante $\mathbf{A}^{(k)}$ présente la structure schématisée ci-dessous :



Les coefficients $a_{i,j}^{(k)}$ des matrices $\mathbf{A}^{(k)}$ de l'itération vérifient donc :

$$a_{i,j}^{(k)} = 0 \text{ si} \begin{cases} k+1 \le i \le m \quad \text{et} \quad m+1 \le j \le n \\ k+1 \le i \le m \quad \text{et} \quad 1 \le j \le k \\ k+1 \le j \le i \le m \end{cases} \qquad \text{éq 5.2-1}$$

de sorte qu'à l'issue de la récurrence, nous aurons :

$$\mathbf{U} = \mathbf{A}^{(1)} \text{ et } \mathbf{P} = \prod_{k=m, m-1, \ldots, 1} \mathbf{P}^{(k)}$$

Le problème se réduit donc à la préservation de la structure [éq 5.2-1] lors du passage de $\mathbf{A}^{(k)}$ à $\mathbf{A}^{(k-1)}$ par une transformation $\mathbf{P}^{(k)}$ qui doit être orthogonale. Le problème de l'orthogonalité est réglé en choisissant la transformation comme un produit de rotations de Givens et le problème de la préservation de la structure est résolu en effectuant ce produit dans un ordre qui ne détruit pas les zéros créés.

Tenant compte de la structure rectangulaire de la matrice $\mathbf{A}^{(k)}$, nous construisons l'itéré en $\mathbf{A}^{(k-1)}$ deux phases :

- La phase 1 annule successivement les coefficients $a_{k,j}^{(k-1)}$ correspondant aux colonnes $j = k-1, k-2, \ldots, 1$, , ce qui se traduit par :

$$\mathbf{A}^{(k-1,k-1)} = \mathbf{A}^{(k)}$$

$$\mathbf{A}^{(k-1,j-1)} = \mathbf{A}^{(k-1,j)} \mathrm{R}\left(k, j, \theta_j^{(k)}\right)^T \quad \text{pour} \quad j = k-1, k-2, \ldots, 1$$

- La phase 2 annule successivement les coefficients $a_{k,j}^{(k-1)}$ correspondant aux colonnes $j = n, n-1, \ldots, m+1$, , ce qui se traduit par la récurrence :

$$\mathbf{A}^{(k-1,n)}=\mathbf{A}^{(k-1,0)}$$
$$\mathbf{A}^{(k-1,j-1)}=\mathbf{A}^{(k-1,j)}\mathrm{R}\left(k,j,\theta_j^{(k)}\right)^T \quad \text{pour} \quad j=n,n-1,\ldots,k+1$$

L'angle $\theta_j^{(k)}$ de la rotation de Givens du plan (k,j) est choisi pour annuler le coefficient en position (k,j) de $a^{(k-1,j)}$. L'application de chaque rotation ne modifie donc que les colonnes k et j ce qui ne détruit pas les coefficients nuls produits par les étapes précédentes. Nous constatons que la colonne k joue un rôle particulier (celui de pivot) car elle seule est systématiquement modifiée par chaque rotation alors que les autres colonnes ne sont modifiées que par la rotation qui annule leur coefficient à la ligne k.

A l'issue de ces récurrences, nous avons $\mathbf{A}^{(k-1)}=\mathbf{A}^{(k-1,k)}$. La matrice $\mathbf{P}^{(k)}$ est alors donnée par :

$$\mathbf{P}^{(k)}=\prod_{j=m+1}^{j=n}\mathrm{R}\left(k,j,\theta_j^{(k)}\right)\prod_{\mathrm{j}=1}^{\mathrm{j=k-1}}\mathrm{R}\left(k,j,\theta_j^{(k)}\right)$$

de sorte que la matrice $\mathbf{P}$ vaut :

$$\mathbf{P}=\prod_{k=1}^{k=m}\left(\prod_{j=m+1}^{j=n}\mathrm{R}\left(k,j,\theta_j^{(k)}\right)\prod_{j=1}^{j=k-1}\mathrm{R}\left(k,j,\theta_j^{(k)}\right)\right)$$

5.3 Réduction à la forme bi-diagonale supérieure

Réduire une matrice carrée triangulaire supérieure $\mathbf{A}$ d'ordre m à la forme bidiagonale supérieure consiste à trouver deux matrices orthogonales $\mathbf{P}$ et $\mathbf{Q}$ et une matrice bidiagonale supérieure $\mathbf{B}$, toutes trois d'ordre m, telles que :

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

Algorithmiquement, la factorisation procède comme celle de la section précédente en utilisant une méthode d'élimination qui s'interprète algébriquement comme la construction d'une suite de matrices $\mathbf{A}^{(k)}$ par :

$$\begin{cases}\mathbf{A}^{(1)}=\mathbf{A}\\ \mathbf{A}^{(k+1)}=\mathbf{Q}^{(k)^T}\mathbf{A}^{(k)}\mathbf{P}^{(k)} \quad \text{pour} \quad k=1,2,\ldots,m-2\end{cases}$$

où chaque matrice courante $\mathbf{A}^{(k)}$ présente la structure diagonale par bloc suivante :

- Le bloc diagonal supérieur (indices de ligne et de colonne variant de 1 à $k-1$) est une matrice bi-diagonale supérieure d'ordre $k-1$,
- Le bloc diagonal inférieur (indices de ligne et de colonne variant de k à m) est une matrice triangulaire supérieure d'ordre $m-k$.

Les coefficients $a_{i,j}^{(k)}$ des matrices $\mathbf{A}^{(k)}$ de l'itération vérifient donc :

$$a_{i,j}^{(k)}=0 \text{ si} \begin{cases}1\le i\le k-1 & \text{et } i+2\le j\le m\\ 1\le i\le k-1 & \text{et } i<j\\ k\le i\le m & \text{et } 1\le j<i\end{cases} \qquad \text{éq 5.3-1}$$

de sorte qu'à l'issue de la récurrence, nous aurons :

$$\mathbf{B}=\mathbf{A}^{(m+1)} \quad , \quad \mathbf{Q}=\prod_{k=1}^{k=m}\mathbf{Q}^{(k)} \quad \text{et} \quad \mathbf{P}=\prod_{k=1}^{k=m}\mathbf{P}^{(k)}$$

Comme pour la factorisation de la section précédente, le problème se réduit à la préservation de la structure [éq 5.3-1] lors du passage de $\mathbf{A}^{(k)}$ à $\mathbf{A}^{(k+1)}$. L'orthogonalité des transformations $\mathbf{Q}^{(k)}$ et $\mathbf{P}^{(k)}$ est obtenue en les construisant comme produit de rotations de Givens et le problème de la préservation de la structure est résolu en effectuant ces produits dans un ordre qui ne détruit pas les zéros créés par les étapes précédentes.

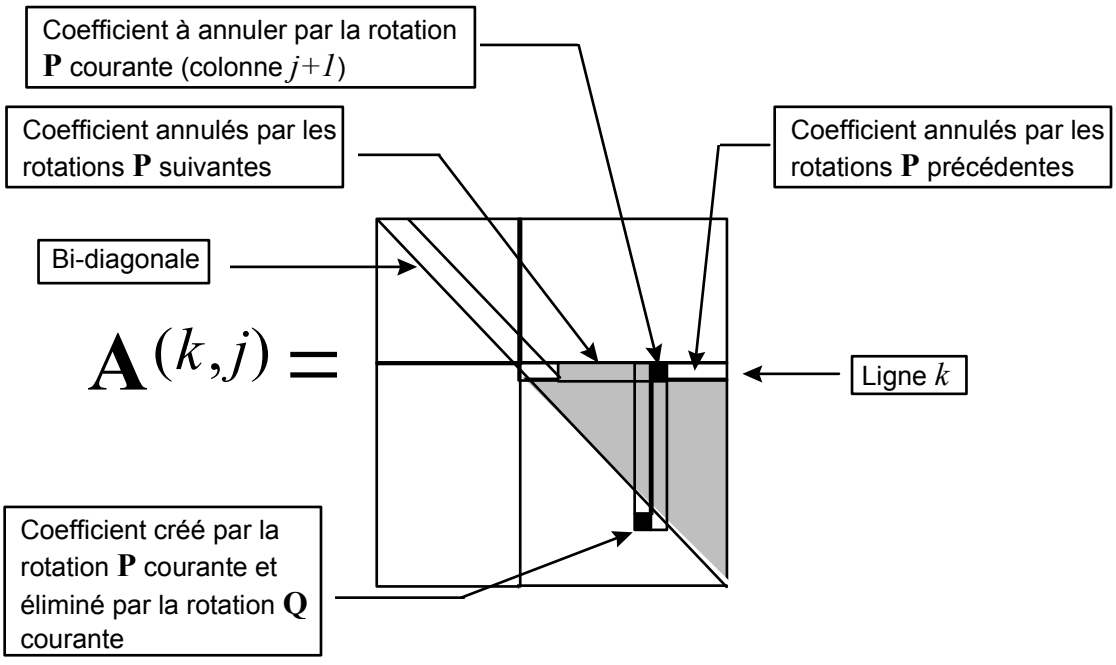
L'algorithme annule donc successivement les coefficient $(k, j+1)$ pour $j=m-1, m-2, \ldots, k+2$ par l'application à droite d'une rotation de Givens du plan $(j, j+1)$. Cette rotation ne modifie que les colonnes j et $j+1$, ce qui crée un coefficient parasite en position $(j+1, j)$. Ce coefficient parasite est alors éliminé par l'application à gauche de la transposée d'une rotation de Givens dans le plan $(j, j+1)$.

Le procédé de passage de $\mathbf{A}^{(k)}$ à $\mathbf{A}^{(k+1)}$ peut être alors formalisé par :

$$\mathbf{A}^{(k+1,\,m-1)}=\mathbf{A}^{(k)}$$
$$\begin{cases}\mathbf{A}^{(k+1,\,j=1/2)}=\mathbf{A}^{(k+1,\,j)}\mathrm{R}\left(J, J+1, \theta_j^{(k)}\right)\\ \mathbf{A}^{(k+1,\,j=1)}=\mathrm{R}\left(J, J+1, \theta_{j=1/2}^{(k)}\right)^T\mathbf{A}^{(k+1,\,j=1/2)} \qquad \text{pour} \quad j=m-1, m-2, \ldots, k+1\end{cases}$$
$$\mathbf{A}^{(k+1)}=\mathbf{A}^{(k+1,\,k)}$$

où les angles $\theta_j^{(k)}$ et $\theta_{j=1/2}^{(k)}$ sont choisis pour annuler respectivement le coefficient en position $(k, j+1)$ de $\mathbf{A}^{(k+1,\,j)}$ et le coefficient en position $(j+1, j)$ de $\mathbf{A}^{(k+1,\,j=1/2)}$.

La structure des matrices $\mathbf{A}^{(k+1,\,j)}$ est illustrée dans la figure suivante :



A l'issue de cette récurrence, les matrices et $\mathbf{P}^{(k)}$ et $\mathbf{Q}^{(k)}$ sont données par :

$$\mathbf{P}^{(k)}=\prod_{j=m-1}^{j=k+1}\mathrm{R}\left(j, j+1, \theta_j^{(k)}\right) \quad \text{et} \quad \mathbf{Q}^{(k)}=\prod_{j=m-1}^{j=k+1}\mathrm{R}\left(k, j, \theta_{j=1/2}^{(k)}\right)$$

de sorte que les matrices $\mathbf{P}$ et $\mathbf{Q}$ valent :

$$\mathbf{P}=\prod_{k=1}^{k=m-2}\prod_{j=m-1}^{j=k+1}\mathrm{R}\left(j,j+1,\theta_{j}^{(k)}\right) \quad \text{et} \quad \mathbf{Q}=\prod_{k=1}^{k=m-2}\prod_{j=m-1}^{j=k+1}\mathrm{R}\left(j,j+1,\theta_{j=1/2}^{(k)}\right)$$

5.4 Décomposition SVD d'une bidiagonale supérieure

Nous présentons un algorithme de construction de la décomposition SVD d'une matrice bidiagonale supérieure $\mathbf{A}$ d'ordre m. L'algorithme construit donc deux matrices orthogonales $\mathbf{Q}$ et $\mathbf{P}$ et une matrice diagonale $\mathbf{D}$ telles que :

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

L'algorithme est tiré de [bib2].

5.4.1 Principe de l'algorithme

L'algorithme de calcul de la décomposition SVD diagonalise itérativement la matrice A au moyen de la récurrence :

$$\begin{cases} \mathbf{A}^{(1)}=\mathbf{A} \\ \mathbf{A}^{(k+1)}=\mathbf{Q}^{(k)^{T}}\mathbf{A}^{(k)}\mathbf{P}^{(k)} \quad \text{pour} \quad k=1,2,... \end{cases}$$ éq 5.4.1-1

où les matrices $\mathbf{Q}^{(k)}$ et $\mathbf{P}^{(k)}$ sont orthogonales et les matrices $\mathbf{A}^{(k)}$ sont bi-diagonales supérieures.

A la convergence, nous aurons :

$$\mathbf{D}=\mathbf{A}^{(\infty)},\ \mathbf{P}=\prod_{k=1}^{k=\infty}\mathbf{P}^{(k)} \text{ et } \mathbf{Q}=\prod_{k=1}^{k=\infty}\mathbf{Q}^{(k)}$$

L'idée de l'itération consiste à :

- Choisir $\mathbf{P}^{(k)}$ pour faire converger l'algorithme QR appliqué à la diagonalisation de la matrice (dite normale) $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ sans la former explicitement. En effet, la matrice $\mathbf{P}$ de la décomposition SVD de $\mathbf{A}$ n'est rien d'autre que la matrice des vecteurs propres de $\mathbf{A}^{\mathbf{T}}\mathbf{A}$,
- Choisir $\mathbf{Q}^{(k)}$ pour préserver la structure bidiagonale supérieure des itérés successifs.

Comme dans le cas des factorisations présentées aux sections 5.2 et 5.3, les matrices $\mathbf{Q}^{(k)}$ et $\mathbf{P}^{(k)}$ sont construites comme produit de rotations de Givens. Le passage de $\mathbf{A}^{(k)}$ à $\mathbf{A}^{(k+1)}$ est alors réalisé par :

$$\mathbf{A}^{(k+1)}=\left[\mathbf{Q}^{(k,2)}\mathbf{Q}^{(k,3)}...\mathbf{Q}^{(k,m)}\right]^{T}\mathbf{A}^{(k)}=\left[\mathbf{P}^{(k,2)}\mathbf{P}^{(k,3)}...\mathbf{P}^{(k,m)}\right]$$ éq 5.4.1-2

où les $\mathbf{Q}^{(k,i)}$ et $\mathbf{P}^{(k,i)}$ sont deux rotations du plan $(i-1,i)$ d'angle respectif θ_j et φ_i :

$$\mathbf{Q}^{(k,i)}=R\left(i-1,i,\theta_i^{(k)}\right) \text{ et } \mathbf{P}^{(\mathbf{k,i})}=R\left(i-1,i,\varphi_i^{(k)}\right)$$

Les rotations sont alternativement appliquées à droite puis à gauche de façon à ce que $\mathbf{A}^{(k+1)}$ conserve la structure bidiagonale supérieure de $\mathbf{A}^{(k)}$. Pour ce faire :

L'angle $\varphi_2^{(k)}$ est, pour le moment, choisi arbitrairement ; l'application de la rotation $\mathbf{P}^{(k,2)}$ crée alors un coefficient en position $(2,1)$,

L'angle $\theta_2^{(k)}$ est choisi pour que l'application de la rotation $\mathbf{Q}^{(k,2)}$ annule le coefficient en position $(2,1)$, ce qui crée un coefficient non nul en position $(1,3)$,

L'angle $\varphi_3^{(k)}$ est choisi pour que l'application de la rotation $\mathbf{P}^{(k,3)}$ annule le coefficient en position $(1,3)$, ce qui crée un coefficient non nul en position $(3,2)$,

.
.
.

L'angle $\theta_{m-1}^{(k)}$ est choisi pour que l'application de la rotation $\mathbf{Q}^{(k,m-1)}$ annule le coefficient en position $(m-1,m-2)$, ce qui crée un coefficient non nul en position $(m-2,m)$,

L'angle $\varphi_m^{(k)}$ est choisi pour que l'application de la rotation $\mathbf{P}^{(k,m)}$ annule le coefficient en position $(m-2,m)$, ce qui crée un coefficient non nul en position $(m,m-1)$,

L'angle $\theta_m^{(k)}$ est choisi pour que l'application de la rotation $\mathbf{Q}^{(k,m)}$ annule le coefficient en position $(m,m-1)$, et la matrice $\mathbf{A}^{(k+1)}$ soit bidiagonale supérieure.

Pour toute valeur de l'angle, ce procédé assure le maintient de la structure bidiagonale supérieure aux itérés [éq 5.4.1-2]. Nous allons voir maintenant comment il est possible de choisir cet angle pour faire converger l'itération [éq 5.4.1-1].

5.4.2 Diagonalisation implicite de la matrice normale

L'algorithme de la sous-section précédente laisse indéterminé l'angle $\varphi_2^{(k)}$ de la première rotation de $\mathbf{P}^{(k)}$. Nous allons lever cette indétermination de façon à faire de la matrice $\mathbf{P}^{(k)}$ la matrice orthogonale d'un pas QR, avec décalage spectral, appliqué à la diagonalisation de la matrice normale $\mathbf{M}=\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

A l'itération SVD [éq 5.4.1-1] de la matrice $\mathbf{A}$, nous associons une itération sur la matrice normale $\mathbf{M}=\mathbf{A}^{\mathbf{T}}\mathbf{A}$:

$$\mathbf{M}^{(k+1)}=\mathbf{A}^{(k+1)^{\mathrm{T}}}\mathbf{A}^{(k+1)}=\mathbf{P}^{(k)^{T}}\mathbf{M}^{(k)}\mathbf{P}^{(k)}$$

Itération QR pour la diagonalisation de la matrice normale

La transformation QR, avec décalage spectral σ_k, appliquée à $\mathbf{M}^{(k)}$ s'écrit :

Factoriser $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}$ sous la forme $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}=\mathbf{P}_\sigma\mathbf{R}_\sigma$
Construire $\mathbf{M}_\sigma^{(k+1)}$ par $\mathbf{M}_\sigma^{(k+1)}=\mathbf{P}_\sigma\mathbf{R}_\sigma+\sigma_k\mathbf{I}$

où $\mathbf{P}_\sigma$ et $\mathbf{R}_\sigma$ sont deux matrices respectivement orthogonale et triangulaire supérieure. Les matrices $\mathbf{M}^{(k)}$ et $\mathbf{M}_\sigma^{(k+1)}$ sont donc tridiagonales et semblables :

$$\mathbf{M}_\sigma^{(k+1)}=\mathbf{P}_\sigma^{T}\mathbf{M}^{(k)}\mathbf{P}_\sigma$$

Du point de vue pratique, la matrice $\mathbf{P}_\sigma$ se présente comme un produit de rotations de Givens :

$$\mathbf{P}_\sigma^{T}=\mathrm{R}(n-1,n,\psi_n)\,\mathrm{R}(n-2,n-1,\psi_{n-1})\ldots\mathrm{R}(1,2,\psi_2)$$

Les angles ψ_k sont choisis pour que l'application à gauche de $\mathrm{R}\left(k-1,k,\psi_k\right)$ à la matrice $\left[\prod_{I=k-1,k-2,\ldots,2}\mathrm{R}\left(I-1,I,\psi_I\right)\right]\left(\mathbf{M}^{(k)}-\sigma_k\mathbf{I}\right)$ annule le coefficient de position $\left(k,k-1\right)$ dans la matrice résultat.

Francis a montré que le passage de $\mathbf{M}^{(k)}$ à $\mathbf{M}_{\sigma}^{(k+I)}$ ne nécessite pas la formation explicite de la matrice $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}$: le décalage peut être effectué implicitement. Le théorème s'énonce comme suit :

Théorème (Francis) : Soit $\mathbf{X}$ une matrice orthogonale dont la première colonne coïncide avec celle de $\mathbf{P}_{\sigma}$. Sous les hypothèses :

1) $\mathbf{M}^{(k+I)}=\mathbf{X}^T\mathbf{M}^{(k)}\mathbf{X}$
2) $\mathbf{M}^{(k+I)}$ est tridiagonale,
3) Les éléments sous-diagonaux de $\mathbf{M}^{(k)}$ sont tous non nuls (irréductibilité de $\mathbf{M}^{(k)}$),

on a

$$\mathbf{M}^{(k+I)}=\mathbf{D}\mathbf{M}_{\sigma}^{(k+I)}\mathbf{D}$$

où $\mathbf{D}$ est une matrice diagonale de coefficients diagonaux tous égaux à ± 1.

Application à l'algorithme SVD

Dès lors, le choix de l'angle $\varphi_2^{(k)}$ de la première rotation de l'itération SVD [éq 5.4.1-1] consiste en $\varphi_2^{(k)}=-\psi_2$. Ainsi $R\left(1,2,\varphi_2^{(k)}\right)=R\left(1,2,\psi_2\right)^T$ de sorte que la première colonne de $\mathbf{P}^{(k)}$ coïncide avec la première colonne de $\mathbf{P}_{\sigma}$. Donc, si tous les éléments sous-diagonaux de $\mathbf{M}^{(k)}$ sont non nuls, alors, les matrices $\mathbf{P}^{(k)}$ et $\mathbf{P}_{\sigma}$ s'identifient (à un facteur multiplicatif ± 1 près des colonnes) et l'itération SVD [éq 5.4.1-1] est équivalente à l'application de la transformation QR, avec décalage, à la matrice $\mathbf{M}^{(k)}$.

Choix du décalage spectral

Le décalage est habituellement choisi comme la valeur propre du mineur inférieur d'ordre deux de $\mathbf{A}^{(k)}$ la plus proche de $a_{m,m}^{(k)}$. Ce choix assure une convergence globale qui, généralement, est cubique.

Applicabilité du théorème de Francis et phénomène de décomposition

L'utilisation du théorème de Francis suppose la non nullité de tous les coefficients sous-diagonaux des matrices $\mathbf{M}^{(k)}$, qui n'est en rien garantie. De plus, dans le cadre de la méthode QR de diagonalisation d'une matrice tridiagonale symétrique, l'apparition de coefficients sous-diagonaux nuls est :

- Souhaitable : les coefficients nuls découplent les blocs diagonaux qu'ils encadrent, ce qui ramène la diagonalisation de la matrice complète à la diagonalisation de ses blocs diagonaux (ce phénomène est souvent appelé "décomposition"),
- Inévitable : la convergence de l'algorithme vers une valeur propre s'interprète algébriquement comme l'apparition d'un bloc diagonal précédent d'ordre 1.

Dans la sous-section suivante, nous allons voir quel traitement il convient d'adopter en présence d'une décomposition.

5.4.3 Analyse de décomposition

L'analyse de décomposition porte sur chaque itéré pris indépendamment des autres, aussi, nous n'utiliserons pas l'indice supérieur (k).

Soit $d_1, d_2, \ldots, d_m$ et $e_2, e_3, \ldots, e_m$ les éléments respectivement diagonaux et sur-diagonaux de A. Les éléments sous-diagonaux de la matrice normale $\mathbf{M}=\mathbf{A}^T\mathbf{A}$ sont alors donnés par :

$$m_{i+1,i}=d_i e_{l+1} \quad \text{pour} \quad i=1,2,\ldots,m-1$$

Supposons, pour simplifier, que seul le coefficient $m_{l-1,l}$ est nul. La matrice $\mathbf{M}$ présente alors une structure de deux blocs diagonaux dont la réunion des spectres respectifs donne le spectre de $\mathbf{M}$. Cette décomposition a lieu soit pour $e_1=0$ soit pour $e_1\neq 0 \text{ et } d_{l-1}=0$.

Le cas $e_1=0$ ne pose aucune difficulté. La matrice $\mathbf{A}$ possède alors une structure diagonale de deux blocs qui fournissent chacun une partie complémentaire de la décomposition SVD de $\mathbf{A}$. Chaque bloc étant bi-diagonal supérieur, sans coefficients sur-diagonaux nuls, sa décomposition SVD est calculable par l'itération [éq 5.4.1-1].

Le cas $e_1\neq 0 \text{ et } d_{l-1}=0$ est plus délicat. En effet, l'itération [éq 5.4.1-1] ne peut être appliquée ni à la matrice $\mathbf{A}$, pour ne pas violer les hypothèses du théorème de Francis, ni à aucune sous matrices de $\mathbf{A}$, pour assurer la structure bi-diagonale aux itérés. Ce problème est contourné par la post-multiplication de $\mathbf{A}$ par une série de rotations de Givens dans les plans successifs $(l-1,l),(l-1,l+1),\ldots,(l-1,m)$:

- La rotation du plan $(l-1,l)$ annule le coefficient $(l-1,l)$ et crée un coefficient en position $(l-1,l+1)$,
- La rotation du plan $(l-1,l+1)$ annule le coefficient $(l-1,l+1)$ et crée un coefficient en position $(l-1,l+2)$,
- La rotation du plan $(l-1,l+1)$ annule le coefficient $(l-1,l+1)$ et crée un coefficient en position $(l-1,l+2)$,

  .
  .
  .
- La rotation du plan $(l-1,m)$ annule le coefficient $(l-1,m)$ et ne crée pas de coefficient.

De sorte que la matrice produite par ce procédé présente la même structure que celle correspondant au cas $e_1=0$.

5.4.4 Organisation de l'algorithme

L'algorithme isole successivement chaque valeur singulière, aussi, il existe un indice k_p tel que l'itéré $\mathbf{A}^{(k_p)}$ se décompose en deux blocs diagonaux :

$$\mathbf{A}^{(k_p)}=\begin{array}{|c|c|}\hline \mathbf{B}^{(k_p)} & 0 \\ \hline 0 & \mathbf{D}^{(k_p)} \\ \hline \end{array}$$

où $\mathbf{B}^{(k_p)}$ est une matrice bidiagonale supérieure d'ordre p et $\mathbf{D}^{(k_p)}$ est une matrice diagonale d'ordre $m-p+1$ rassemblant sur sa diagonale les valeurs singulières trouvées.

A partir de cet itéré, l'algorithme applique l'itération de la sous-section 5.4.1 à la sous-matrice $\mathbf{B}^{(k_p)}$ jusqu'à l'annulation du coefficient en position $(p-1,p)$, signal de la convergence de la $p^{ième}$ valeur singulière. Chaque pas de l'itération interne, ainsi définie, s'organise comme suit :

- Analyse de décomposition,
- Dans le cas où la décomposition trouvée correspond à un élément diagonal nul, une série de rotations supplémentaires est appliquée pour retrouver la structure de décomposition générée par un élément sur-diagonal nul. Ces rotations sont construites suivant la méthode présentée à la sous-section 5.4.3,
- Si le coefficient en position $(p-1,p)$ ne produit pas de décomposition alors la sous-matrice $\mathbf{B}^{(k_p)}$ est l'objet d'un pas de l'itération [éq 5.4.1-1] où, conformément à l'analyse de la sous-section 5.4.2, l'angle de la première rotation est choisi pour que ce pas soit équivalent à l'application d'une transformation QR, avec décalage spectral implicite, sur la matrice normale associée.

La convergence complète de l'itération est alors obtenue à l'indice k_m pour lequel la sous-matrice $\mathbf{D}^{(k_p)}$ est d'ordre m.

Bien entendu, dans la pratique, un coefficient est considéré comme nul dès qu'il est inférieur, en valeur absolue, à une certaine tolérance. La tolérance généralement utilisée pour les problèmes de valeurs singulières est choisie comme le produit de la précision machine par $\|\mathbf{A}\|_1$. Remarquons que dans le cas d'une décomposition produite par un élément diagonal nul à la tolérance choisie, l'application de la série de rotations supplémentaires décrites à la sous-section 5.4.3 crée une sous colonne d'éléments non nuls sous ce coefficient. Ces éléments ne sont pas gênants car ils sont tous nuls à la tolérance choisie.

6 Bibliographie

1) P. G. CIARLET : "Introduction à l'analyse numérique matricielle et à l'optimisation" _ MASSON (1985).

2) G. H. GOLUB, C. REINSCH "Singular value decomposition and least squares solutions" in "Handbook for Automatic Computation - Linear Algebra, Vol. 2" J.H. WILKHINSON, C. REINSH Editors _ SPINGER VERLAG (1971).

3) P. LASCAUX, R. THEODOR : " Analyse numérique matricielle appliquée à l'art de l'ingénieur", tomes 1 et 2 _ MASSON (1986).

7 Description des versions du document

Version Aster	Auteur(s) Organisme(s)	Description des modifications
3	B.QUINNEZ, R.MICHEL EDF-R&D/MMN	Texte initial