Resolution of nonregular systems by a method of decomposition in singular values

Summary :

This document is devoted to the resolution of the systems of linear equations nonregular. The matrices taken into account can be square noninvertible or rectangular.

After having recalled the theoretical framework of the solutions within the meaning of least squares, we concentrate the talk on the method by decomposition in singular values which provides, on the one hand, diagnostic tools of the degree of regularity of the system, and, on the other hand, a family of algorithms of resolution at the same time more generals and more stable than those drifting of the approach by the normal equations.

Lastly, we detail the algorithm implemented in *Code_Aster* who reabsorbs the equivalent functionality of the bookstore *Nag* (F04JDF for version 12 and F04JDE for version 15) used for the modeling of the metallurgical behavior of steels [R4.04.01].

Contents

1 Introduction

Being given a real matrix $\mathbf{A}$ of order $m \times n$ and a vector $\mathbf{b}$ element of $\mathbb{R}^m$, we consider the problem of the determination of a vector $\mathbf{x}$ element of $\mathbb{R}^n$ who checks the following linear system:

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

It is well-known ([bib3] p. 9) that this system admits one, and only one solution, for all b element of $\mathbb{R}^m$ under the requirements and sufficient that it is équi-constrained $(m=n)$ and that its matrix A that is to say regular. Also, the investigation of the under-constrained case $(m \leq n)$ and overstrained $(m \geq n)$ we will confront with the one of the three following situations:

[1] The linear system [éq 1-1] admits a solution and only one,
[2] The linear system [éq 1-1] does not admit a solution,
[3] The linear system [éq 1-1] admits an infinity of solutions.

In practice, the situation 2) meets in general in the case of an overstrained system whereas the singular and under-constrained équi-constrained systems lead in general to the situation 3).

To claim to solve a linear system of the type [éq 1-1], we should initially define what we will call **solution**. This is the object of **paragraph 2** who is based mainly on the concept of **least squares** and on **differentiable optimization** to define, some is the type of system, a solution which is always single.

paragraph 3 is devoted to **decomposition in singular values** matrices (in summary SVD: *Been worth Singular Decomposition*), which, not only constitutes a tool to diagnose which of the three preceding situations corresponds to the linear system studied, but also provides a method of determination of the solution defined in paragraph 2.

Method using **decomposition SVD** is presented to **paragraph 4** and there is compared with the method of **normal equations**.

paragraph 5 detail on the algebraic level the application of method SVD to **the resolution of a linear équi or under-constrained system** such as it is put in work in *Code_Aster*.

In the following paragraphs, we will use the notations below:

- $\|\mathbf{x}\|$ and $(\mathbf{x},\mathbf{y})$ for, respectively, the euclidian norm of the vector $\mathbf{x}$ and the associated scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$ elements of $\mathbb{R}^m$ or of $\mathbb{R}^n$,
- $\mathbf{M}^{\mathrm{T}}$ for transposed of the matrix $\mathbf{M}$,
- $\mathrm{Ker}\,\mathbf{M}$ and $\mathrm{Im}\,\mathbf{M}$ for, respectively, the core and the image of (the linear application associated with) the matrix $\mathbf{M}$,
- $\mathbf{X}^{\perp}$ for the orthogonal one of under space $\mathbf{X}$ of $\mathbb{R}^m$ or of $\mathbb{R}^n$.
-

2 Solution of a rectangular linear system

In this paragraph we will define a concept of *solution* for the linear system [éq 1-1] which enjoys the properties**existence** and of**unicity.** The approach proceeds in two times:

- Initially, by an approach of the type least squares we build problem of a differentiable and convex optimization (section 2.1) which admits always at least a solution (section 2.2). The situation 2) paragraph 1 is then eliminated,
- Then, analyzing the property of unicity (section 2.3) to note that it is not always guaranteed we will impose an additional constraint (section 2.4) on the solution characterized in section 2.1 in order to restore unicity.

2.1 Formalism of least squares

The single solution of a linear system $\mathbf{Ax}=\mathbf{b}$ of square and regular matrix the minimum of quantity realizes $\|\mathbf{Ay}-\mathbf{b}\|$ when $\mathbf{y}$ described $\mathbb{R}^n$. This property opens up the way to us which leads to a concept of solution for a linear system general of the type [éq 1-1] which confers the same properties to him as those of the typical case of the regular system. We will thus say point $\mathbf{x}$ of $\mathbb{R}^n$ that it is solution of the system [éq 1-1] if it is solution of **problem of optimization** :

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

This approach is natural because it defines a solution of which the residue $\mathbf{r}=\mathbf{Ax}-\mathbf{b}$ is null if the second member is element of $\mathrm{Im}\,\mathbf{A}$ and is of minimal standard in the contrary case, which constitutes best than one can wait.
To analyze the problem [éq 2.1-1], it is convenient to substitute to him the problem of optimization without constraints are equivalent according to:

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

where $\mathrm{J}(.)$ is the functional calculus defined by:

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

The interest of the problem [éq 2.1-2] is due to the fact that the functional calculus $\mathrm{J}(.)$ check the following properties:

- $\mathrm{J}(.)$ is twice continuously differentiable:

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

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

- $\mathrm{J}(.)$ is quadratic and convex.
-

Thus, the problem [éq 2.1-2] lies within the scope of differentiable and convex optimization so that we have the following results ([bib1] p. 156 and 146):

1) Convexity: any local optimum is in fact a total optimum, i.e. a solution of [éq 2.1-2],
2) Differentiability: any local optimum checks the equation of Euler $\mathrm{DJ}(\mathbf{x})=0$ on Ä^N who, account - held of [éq 2.1-3], led to the characterization by **equations** known as **normals** :

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

2.2 Existence of optima

In [bib1] p. 171 one finds a demonstration of the existence of at least a solution to the normal equations [éq 2.1-5]. This demonstration is based on arguments intended for the taking into account of the case of infinite dimension (theorem of projection on convex closed of a space of Hilbert).

Our case being definitely simpler, we give a demonstration of the result which uses only simple algebraic arguments which, moreover, we will be useful in paragraph 3. To show that, for all $\mathbf{b}$ element of $\mathbb{R}^m$, the normal equations [éq 2.1-5] admit a solution is equivalent to the establishment of inclusion $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\subset\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. However, for any real matrix $\mathbf{M}$ of order $m\times n$ we have $\mathrm{Im}\,\mathbf{M}^{\mathbf{T}}=(\mathrm{Ker}\,\mathbf{M})^{\perp}$ ([bib3] p. 28). Also, the inclusion to be established which is equivalent to $(\mathrm{Ker}\,\mathbf{A})^{\perp}\subset(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A})^{\perp}$ who is it even equivalent to $\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\subset\mathrm{Ker}\,\mathbf{A}$. That is to say thus

$\mathbf{x}\in\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$; then $\mathbf{Ax}\in\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}$, i.e. $\mathbf{Ax}\in\left(\mathrm{Im}\,\mathbf{A}\right)^{\perp}$. Like $\mathbf{Ax}$ is also element of $\mathrm{Im}\,\mathbf{A}$, it can be only null what means that $\mathbf{x}\in\mathrm{Ker}\,\mathbf{A}$ and the demonstration completes.

This stage of the matter, we can say that any system of the type [éq 1-1] admits at least a solution within the meaning of [éq 2.1-3] and all these solutions are characterized as solution within the meaning of Being on fire normal equations of [éq 2.1-5]. The situation 2) paragraph 1 is eliminated.

Remain to eliminate the situation 3), i.e. to guarantee unicity.

2.3 Unicity of the optimum and row of the system

It is clear that the normal equations [éq 2.1-5], characterizing the optima which we seek, admit a single solution under the requirement and sufficient that $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ that is to say regular. Like $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ is always semi-definite positive, its inversibility is equivalent to its definite positivity, so that, taking into account the expression [éq 2.1-4] of the derivative second of the functional calculus $\mathrm{J}(.)$, we find the well-known theorem of unicity of the optimum of the problem [éq 2.1-2] for a convex functional calculus twice continuously differentiable ([bib1] TH 7.4-3 and 7.4-4).

In any general information, nothing prevents the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ to be singular, the solution of the system [éq 1 - 1] within the meaning of [éq 2.1-2] is thus not always single. We have nevertheless a criterion to detect this situation. With section 2.2 we established that $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\subset\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$ and as reciprocal inclusion is crudely true, we can conclude with the identity $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}=\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. The introduction of **row** $\mathrm{rg}(\mathbf{A})$ matrix $\mathbf{A}$, the dimension of its space image, then enables us to say that a requirement and sufficient so that $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ that is to say invertible is that $\mathrm{rg}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)=n$ what is equivalent to $\mathrm{rg}(\mathbf{A})=n$ because $\mathrm{rg}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)=\mathrm{rg}\left(\mathbf{A}^{\mathbf{T}}\right)=\mathrm{rg}\left(\mathbf{A}\right)$.

The interest of this criterion is due to the fact that it limits the analysis to the only matrix $\mathbf{A}$ without it being necessary to form explicitly $\mathbf{A}^{\mathbf{T}}\mathbf{A}$. This criterion also shows us that the normal equations associated with a strictly under-constrained linear system always admit an infinity of solutions. Indeed, the row of a matrix is as equal to the number of independent columns as it has; also, so that this row reaches the value n it is necessary that the columns of the matrix is of order at least n.

2.4 Solution within the meaning of least squares

We have just noted that the whole of the points which minimize the residue of the system [éq 1-1] is not necessarily tiny room to only one point. To restore unicity we refine the concept of solution of the system [éq 1-1] of section 2.1 while defining **solution within the meaning of least squares** as the element of minimal standard of the whole of the points which minimize the residue. This solution x is then characterized by:

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

This characterization is not satisfactory on the practical level because she asks for the resolution of a problem of optimization under constraints. We will substitute to him another characterization more adapted to the direction where it will lead (see section 4.2) to a procedure of calculation definitely simpler.

The unit $\mathbf{S}$ above is relocated of core of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ by any of the vectors solutions of the normal equations [éq 2.1-5]. Also, the additional condition of minimization of the standard is interpreted like a simple projection: the solution within the meaning of least squares of the system [éq 1-1] is anything else only the projection of the origin of ÄN on the whole of the solutions of the normal equations. Also, we can characterize it like the point of intersection between the unit $\mathbf{S}$ and the orthogonal one of the core of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

The definition of a solution to the system [éq 1-1] can then be summarized as follows:

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

The first condition makes $\mathbf{x}$ a vector of minimal residue while the second selects, among the vectors of minimal residue, that of minimal standard.

The definition [éq 2.4-1] is a classical generalization of the concept of solution of a regular system équi - constrained and confers on any system of the type [éq 1-1] a solution and only one.

3 Singular values

In this paragraph we have some useful results for the design of a method of operational resolution of the system [éq 1-1]. These results derive from the concept of singular values (section 3.1) and make it possible to build a base of the core and a base of the image of the matrix of the system (section 3.2) from which it is possible to give a direction, adapted to the calculation of the solution within the meaning of least squares, contrary to an unspecified matrix (section 3.3).

3.1 Decomposition in singular values

Let us start by pointing out the definition of the singular values. One calls **singular values** of a real matrix $\mathbf{A}$ of order $m \times n$ square roots of the eigenvalues of the square matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ of order n who, let us recall it, is semi-definite positive.

The concept of diagonalisation of the square matrices (when they are diagonalisables) spreads with the rectangular matrices (without restriction) by the concept of decomposition (or factorization) in singular values.

For all real matrices $\mathbf{A}$ of order $m \times n$, there exist two unit square matrices $\mathbf{Q}$ and $\mathbf{P}$ of a respective nature m and n such as:

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

where Σ is a matrix of order $m \times n$ the structure is schematized below:

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

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

μ_i are the singular values of $\mathbf{A}$ that we suppose ordered by decreasing order:

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

One can find a demonstration of this result in [bib1] p.10 for the équi-constrained case and in [bib3] p.73 for the overstrained case, the under-constrained case from of deduced then by transposition.

Factorization SVD [éq 3.1-1] of $\mathbf{A}$ give $\mathbf{A}^{\mathbf{T}}\mathbf{A}=\mathbf{P}\,\mathbf{\Sigma}^{\mathbf{T}}\mathbf{S}\mathbf{P}^{\mathbf{T}}$ and $\mathbf{A}\mathbf{A}^{T}=\mathbf{Q}\,\mathbf{\Sigma}\,\mathbf{\Sigma}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}$ so that, $\mathbf{\Sigma}^{\mathbf{T}}\mathbf{\Sigma}$ and $\mathbf{\Sigma}\,\mathbf{\Sigma}^{\mathbf{T}}$ being diagonal square matrices, the matrix $\mathbf{P}$ is made up by the orthonormalized clean vectors of the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ while the matrix $\mathbf{Q}$ is made up by the orthonormalized clean vectors of the matrix $\mathbf{A}\mathbf{A}^{\mathbf{T}}$.

3.2 Row, image and core

The paragraph [§2] showed the fundamental role which the row of the matrix plays $\mathbf{A}$ and the core of the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ for the resolution of a nonregular linear system of the type [éq 1-1]. We will see now how factorization [éq 3.1-1] can be used to determine this row as well as a base of $\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

That is to say r the index of the smallest nonworthless singular value. Factorization [éq 3.1-1] is written too $\mathbf{Q}^{\mathbf{T}}\mathbf{A}\mathbf{P}=\mathbf{\Sigma}$ where the taking into account of the worthless singular values makes it possible to specify the decomposition in block of $\mathbf{\Sigma}$:

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

$$\Sigma = \begin{array}{|c|c|} \hline \Sigma_r & 0 \\ \hline 0 & 0 \\ \hline 0 & \\ \hline \end{array} \quad \text{si } m > n$$

where $\mathbf{\Sigma}_r=\mathrm{Diag}\left(\mu_1,\mu_2,...,\mu_r\right)$ is the diagonal matrix of order r nonworthless singular values in the ascending order.

Since matrices $\mathbf{Q}$ and $\mathbf{P}$ are regular, the matrices $\mathbf{A}$ and $\mathbf{\Sigma}$ are equivalent so that their respective core and image coincide. We thus deduce from it that:

- The row of $\mathbf{A}$ coincide with the number of nonworthless singular values:

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

- The vectors columns of $\mathbf{P}$ of index $r+1$ with n form a base of $\mathrm{Ker}\,\mathbf{A}$
- The vectors columns of $\mathbf{Q}$ corresponding to the nonworthless singular values form a base of $\mathrm{Im}\,\mathbf{A}$

In addition, with section 2.3 we saw that $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}=\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Identity $\mathrm{Im}\,\mathbf{M}^{\mathbf{T}}=\left(\mathrm{Ker}\,\mathbf{M}\right)^{\perp}$ we gives then $\mathrm{Ker}\,\mathbf{A}=\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$ so that the second condition of the definition [éq 2.4-1] is simply carried out by any vector which is expressed like a linear combination of the vectors columns of $\mathbf{P}$ corresponding to the nonworthless singular values.

3.3 Pseudo-opposite and solution within the meaning of least squares

Another application of the decomposition in singular values consists of the concept of **pseudo - opposite** (or opposite Moore-Penrose) which generalizes the usual concept of reverse of a regular square matrix to the rectangular matrices on the one hand, and the singular square matrices on the other hand.

First of all, theopposite one of a matrix Σ decomposition in singular values [éq 3.1-1] is defined by:

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

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

where $\Sigma_r^{-1} = \mathrm{Diag}\left(\frac{1}{\mu_1}, \frac{1}{\mu_2}, \ldots, \frac{1}{\mu_r}\right)$ is the reverse with the usual direction of $\mathbf{S}_r$.

This being, we use the decomposition [éq 3.1-1] matrix $\mathbf{A}$ to define its pseudo - opposite A^{+} by:

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

In the same way, of the decomposition [éq 3.1-1] of the matrix $\mathbf{A}$ we draw $\mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{P}\,\Sigma^{\mathbf{T}}\,\Sigma\,\mathbf{P}^{\mathbf{T}}$, so that theopposite one $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}$ matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ is defined by:

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

We are now able to provide a simple interpretation of the solution within the meaning of least squares defined by [éq 2.4-1].

The restriction on $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ linear application associated with the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ an isomorphism defines of $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ on $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Like, on the one hand $\left(\mathrm{Ker}\,A^{T}A\right)^{\perp} = \left(\mathrm{Ker}\,A\right)^{\perp} = \mathrm{Im}\,A^{T}$, and, in addition, $\mathrm{Im}\,\mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathrm{Im}\,\mathbf{A}^{\mathbf{T}}$, this restriction is in fact an automorphism of $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$.

In the base of $\left(\mathrm{Ker}\,\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{\perp}$ constituted by r first columns of the matrix $\mathbf{P}$, this automorphism is represented by the matrix Σ_r^{2}. Also, its reciprocal automorphism y is represented by the matrix Σ_r^{-2}. Extension to $\mathbb{R}^n$ of this automorphism is then represented, in the base associated with the matrix $\mathbf{P}$, by the matrix $\left(\Sigma^{\mathbf{T}}\Sigma\right)^{+} = \Sigma^{\mathbf{T}}\left(\Sigma^{\mathbf{T}}\right)^{+}$, and thus, in the canonical base, by the matrix $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}$.

It follows that:

- We find the fact that, for all $\mathbf{b}$ element of $\mathbb{R}^n$, there exists a single vector $\mathbf{x}\in\left(\mathrm{Ker}\,\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{\perp}$ solution of $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}=\mathbf{A}^{\mathrm{T}}\mathbf{b}$, that is to say the existence and the unicity of the solution within the meaning of least squares [éq 2.4-1] system $\mathbf{A}\mathbf{x}=\mathbf{b}$,

- This single solution is given by:

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

Theopposite one of a matrix is defined starting from decomposition SVD of this matrix. As decomposition SVD is not single, the pseudo-opposite matrix is not single. On the other hand, from the point of view of the linear applications associated with the matrices, the application pseudo - opposite is single. All the matrices pseudo-opposite associated with the various decompositions SVD with a given matrix are not whereas representing matric particular which express this pseudo-opposite application compared to bases induced by the orthogonal matrices of decompositions SVD. Also, the expression [éq 3.3-3] has a direction: it defines a vector of which $\mathbf{x}$ represent the components compared to the base of arrival (matrix $\mathbf{P}$) decomposition SVD.

4 Resolution of a rectangular linear system

The two methods of resolution of the system [éq 1-1] which we present to sections 4.1 (method of the normal equations) and 4.2 (decomposition in singular values) aim to the resolution of the normal equations [éq 2.1-5]. These two methods are characterized not only by the choice of the algorithms which they implement (inversion against pseudo-inversion), but also by their degree of general information and their digital properties which are compared with section 4.3

4.1 Method of the normal equations

The resolution of the system $\mathbf{A}\mathbf{x}=b$ by the method of the normal equations consists in calculating the solution within the meaning of least squares [éq 2.4-1] in a “direct” way, i.e. by using the relation directly $x=\left(A^{T}A\right)^{-1}A^{T}b$. For that, it is initially a question of calculating $A^{T}A$ and $A^{T}b$, then to solve the system obtained either by iterative method or by a factorization of $A^{T}A$.

We can notice right now that this method is limited to the matrices $A^{T}A$ regular, which limits its scope of application to the system [éq 1-1] whose matrix is of full row (see section 2.3). In particular, the method of the normal equations can treat neither the strictly under-constrained systems, nor the singular équi-constrained systems (see section 2.4).

4.2 Method by decomposition in singular values

We saw with section 3.3 that the solution of the system $\mathbf{A}\mathbf{x}=\mathbf{b}$ within the meaning of least squares defined by [éq 2.4-1] can be characterized by the relation $\mathbf{x}=\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathrm{T}}\mathbf{b}$ [éq 3.3-3]. Method of resolution of the system based on this property **is known as method by decomposition in singular values** because it builds theopposite one [éq 3.3-2] of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ via decomposition SVD [éq 3.1 - 1] of the matrix $\mathbf{A}$.

As any matrix can be broken up into singular values, it follows that any system of the type [éq 1-1] can be solved within the meaning of [éq 2.4-1] by this method which thus presents, at least, the advantage of the general information compared to the method of the normal equations.

It is not all. The method by decomposition in singular values, contrary to the method of the normal equations, does not require the explicit construction of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and of the vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$ (we will see with section 4.3 the interest on the digital level of this property). Indeed, it is easy to check that the matrix $\mathbf{\Sigma}$ singular values of factorization [éq 3.1-1] satisfied with identity $\mathbf{\Sigma}^{+}\left(\mathbf{\Sigma}^{\mathrm{T}}\right)^{+}\mathbf{\Sigma}^{\mathrm{T}}=\mathbf{\Sigma}^{+}$, so

that, prémultipliant $\mathbf{A}^{\mathbf{T}}$ by theopposite one of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ and taking account of factorization [éq 3.1-1], we obtain $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}=\mathbf{P}\boldsymbol{\Sigma}^{+}\left(\boldsymbol{\Sigma}^{\mathbf{T}}\right)^{+}\boldsymbol{\Sigma}^{\mathbf{T}}\mathbf{P}^{\mathbf{T}}\mathbf{P}\boldsymbol{\Sigma}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}$, which, by orthogonality of $\mathbf{P}$ we gives $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{+}\mathbf{A}^{\mathbf{T}}=\mathbf{A}^{+}$. Consequently, the characterizations [éq 3.3-3] and [éq 2.4-1] of the sought solution are equivalent to the characterization:

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

4.3 Comparison of the method of the normal equations to the method of decomposition in singular values

In the two preceding sections, we come to note, that algebraically, the method of decomposition in singular values is more general and simpler than the method of the normal equations. We now will note, while following [bib3] p. 336, qu ' it is also higher to him on the digital level. This superiority is expressed on the one hand, in term of stability not only of the resolution, but also of construction of the problem and, on the other hand, on a less critical level, in term of adaptation to the treatment of the hollow matrices.

4.3.1 Conditioning

The conditioning of a matrix $\mathbf{A}$ of order $m\times n$ is defined like the report of its extreme and nonworthless singular values:

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

where r is the row of the matrix $\mathbf{A}$.

Results presented in [bib3] p.184, using the normal equations as a tool of analysis and not like a computational tool, show that the disturbance of the solution of the problem of optimization [éq 2.1-1] due to the rounding errors can be proportional to $\operatorname{cond}\left(\mathbf{A}\right)^2$. But the classical results of the analysis of stability of the solution of a linear system compared to these same errors show a proportionality with the number of conditioning of the matrix. So that in the case of a direct resolution of the normal equations, we obtain an error always proportional to $\operatorname{cond}\left(\mathbf{A}^{\mathbf{T}}A\right)=\operatorname{cond}\left(\mathbf{A}\right)^2$, which is less good than $\operatorname{cond}\left(\mathbf{A}\right)$.

The method of resolution by decomposition in singular values uses only orthogonal transformations (see paragraph 4), so that it does not modify the initial conditioning of the problem [éq 1-1] and is thus, from this point of view, more attractive than the method of the normal equations.

4.3.2 Loss of precision

We have just seen that the rounding errors lead to a degradation of the more significant solution when it is calculated via the normal equations rather than by a decomposition in singular values. The following example, drawn from [feeding-bottle 2], watch that the construction even of the system [éq 2.1-5] of the normal equations is disturbed by the rounding errors.

That is to say thus the following matrix:

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

whose singular values are $\mu_1=\sqrt{2+\varepsilon^2}$ and $\mu_2=|\varepsilon|$, so that the row of $\mathbf{A}$ is 2 as soon as $\varepsilon\neq 0$. If ε check $\varepsilon^2<\varepsilon_{mach}<\varepsilon$ where ε_{mach} is the precision machine, then all the coefficients of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ will be calculated with the value 1 and the calculated singular values will be, at best, $\mu_1=\sqrt{2}$ and

$\mu_2=0$. It follows that the digital row, calculated by the normal equations, will be 1 , whereas that calculated by a decomposition SVD of the matrix $\mathbf{A}$ would be equal to 2

4.3.3 Hollow structure

With a less level, the construction of the matrix of the normal equations induced a filling of the associated system that the method using the decomposition in singular values avoids.

4.3.4 Conclusion

The following table summarizes the discussion of the preceding sub-sections:

	General information	Conditioning	Loss of precision to the construction of the problem	filling
Normal equations	systems of full row	$\text{cond}(\mathbf{A})^2$	possible	yes
SVD	any system	$\text{cond}(\mathbf{A})$	impossible	not

5 Algorithm SVD for the resolution of a linear équi or under-constrained system

In this paragraph, we detail the method of resolution of the not-regular systems put in work in *Code_Aster*. This method applies to the under-constrained or équi-constrained system singular and provides the solution within the meaning of least squares [éq 2.4-1].

The calculation of a decomposition SVD of $\mathbf{A}$ is equivalent to the calculation of the spectrum of the associated normal matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$. Also, it can be obtained only with the convergence of an iterative process.

section 5.1 expose it **principle of the algorithm** and shows in particular how the application of two orthogonal transformations allows **to reduce the problem** with the simple research of decomposition SVD of a matrix higher bi--diagonal. **sections 5.2** and **5.3** are devoted to **the algorithmic one** of these **reductions**. **section 5.4** present the algorithm of **decomposition SVD** matrix bi--diagonal.

The algorithms will be described with the convention of notation in which:

- $\mathrm{R}(i,j,\theta)$ indicate the rotation of Givens of the plan (i,j) and of angle θ ,
- $\mathbf{A}^{(\mathbf{k})}$ indicate reiterated index k of a matric iteration and $\mathbf{A}^{(\mathbf{k},\mathbf{l})}$ reiterated l of an internal iteration with reiterated $\mathbf{A}^{(\mathbf{k})}$.

5.1 Reduction of the problem and principle of the algorithm

In this section, we present the algorithm of resolution of a linear sytème équi or under - constrained by method SVD.

We reduce the problem in search of decomposition SVD of a matrix bidiagonale as in [bib2] but we carry out the reduction in another way that proposed in [bib2]: we start by reducing the matrix to a higher triangular form, then, we reduce this triangular to a higher form bidiagonale. These two reductions are carried out by orthogonal transformations.

Operations of **calculation of decomposition SVD** are connected as follows:

5.1.1 Reduction with the higher triangular form

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

where $\mathbf{P}_1$ is an orthogonal matrix of order n and $\mathbf{U}$ a higher triangular matrix of order m.

5.1.2 Reduction with the form higher bi--diagonal

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

where $\mathbf{Q}_2$ and $\mathbf{P}_2$ are two *orthogonal matrices* of order m and $\mathbf{B}$ one *matrix bidiagonale* higher of order m.

5.1.3 Decomposition SVD of the higher bi--diagonal

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

where $\mathbf{Q}_3$ and $\mathbf{P}_3$ are two *orthogonal matrices* of order m and Σ one *diagonal matrix* of order m form:

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

Combining the relations [éq 5.1-1], [éq 5.1-2] and [éq 5.1-3], we obtain one **decomposition SVD of the matrix** $\mathbf{A}$:

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

solution within the meaning of least squares [éq 2.4-1] from the system [éq 1-1] is then obtained by the application of pseudo-opposite [éq 3.3-1] of A deduced from the decomposition in singular values [éq 5.1-4]. We thus obtain:

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

The algorithm proposed thus consists of the sequence of factorizations [éq 5.1-1], [éq 5.1-2] and [éq 5.1-3] prior to the application of the relation [éq 5.1-5].

5.2 Reduction with the higher triangular form

Starting from a matrix $\mathbf{A}$ of order $m\times n$ for $m\le n$ a higher triangular matrix is determined $\mathbf{U}$ of order m and an orthogonal matrix $\mathbf{P}$ of order n such as:

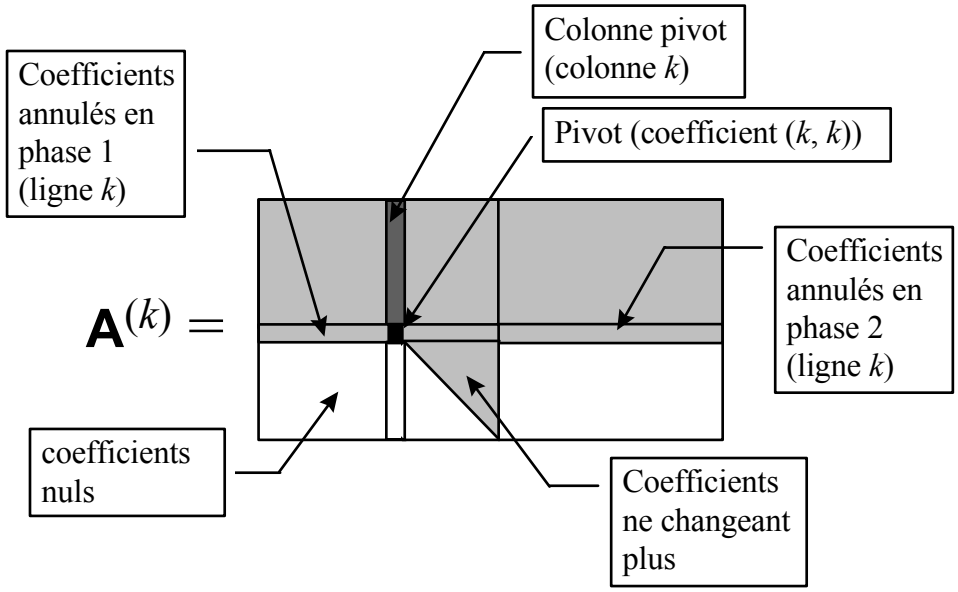
$$\mathbf{A}=\begin{bmatrix}\mathbf{U} & 0\end{bmatrix}\mathbf{P}^{\mathrm{T}} \qquad \text{si} \quad m<n$$

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

Algorithmiquement, factorization uses a method of elimination which is interpreted, like the construction of a succession of matrices $\mathbf{A}^{(k)}$ by:

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

where each current matrix $\mathbf{A}^{(k)}$ have the structure schematized below:



Coefficients $a_{i,j}^{(k)}$ matrices $\mathbf{A}^{(k)}$ iteration thus check:

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

so that at the conclusion of the recurrence, we will have:

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

The problem is thus reduced to the safeguarding of the structure [éq 5.2-1] at the time of the passage of $\mathbf{A}^{(k)}$ with $\mathbf{A}^{(k-1)}$ by a transformation $\mathbf{P}^{(k)}$ who must be orthogonal. The problem of orthogonality is solved by choosing the transformation like a product of rotations of Givens and the problem of the safeguarding of the structure is solved by carrying out this product in an order which does not destroy the zeros not created.

Taking account of the rectangular structure of the matrix $\mathbf{A}^{(k)}$, we build reiterated in $\mathbf{A}^{(k-1)}$ two phases:

- The phase 1 cancel successively the coefficients $a_{k,j}^{(k-1)}$ corresponding to the columns $j=k-1, k-2,...,1$, , which results in:
$$\mathbf{A}^{(k-1,k-1)}=\mathbf{A}^{(k)}$$
$$\mathbf{A}^{(k-1, j-1)}=\mathbf{A}^{(k-1, j)}\mathrm{R}\left(k, j, \theta_j^{(k)}\right)^T \quad \text{pour} \quad j=k-1, k-2,...,1$$
- The phase 2 cancel successively the coefficients $a_{k,j}^{(k-1)}$ corresponding to the columns $j=n, n-1,..., m+1,$, which results in the recurrence:

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

The angle $\theta_j^{(k)}$ rotation of Givens of the plan (k,j) is selected to cancel the coefficient in position (k,j) of $a^{(k-1,j)}$. The application of each rotation thus modifies only the columns k and j what does not destroy the worthless coefficients produced by the preceding stages. We note that the column k play a particular part (that of pivot) because it only is systematically modified by each rotation whereas the other columns are modified only by the rotation which cancels their coefficient with the line k.

At the conclusion of these recurrences, we have $\mathbf{A}^{(k-1)}=\mathbf{A}^{(k-1,k)}$. The matrix $\mathbf{P}^{(k)}$ is then given by:

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

so that the matrix $\mathbf{P}$ is worth:

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

5.3 Reduction with the form higher bi--diagonal

To reduce a higher triangular square matrix $\mathbf{A}$ of order m with the higher form bidiagonale consists in finding two matrices orthogonal $\mathbf{P}$ and $\mathbf{Q}$ and a higher matrix bidiagonale $\mathbf{B}$, all three of order m, such as:

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

Algorithmiquement, factorization proceeds as that of the preceding section by using a method of elimination which is interpreted algebraically like the construction of a succession of matrices $\mathbf{A}^{(k)}$ by:

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

where each current matrix $\mathbf{A}^{(k)}$ have the diagonal structure per block following:

- The higher diagonal block (indices of line and column varying from 1 with $k-1$) is a matrix higher bi--diagonal of order $k-1$,
- The lower diagonal block (indices of line and column varying from k with m) is a higher triangular matrix of order $m-k$.

Coefficients $a_{i,j}^{(k)}$ matrices $\mathbf{A}^{(k)}$ iteration thus check:

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

so that at the conclusion of the recurrence, we will have:

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

As for the factorization of the preceding section, the problem is reduced to the safeguarding of the structure [éq 5.3-1] at the time of the passage of $\mathbf{A}^{(k)}$ with $\mathbf{A}^{(k+1)}$. The orthogonality of the transformations $\mathbf{Q}^{(k)}$ and $\mathbf{P}^{(k)}$ is obtained by building them like product of rotations of Givens and the problem of safeguarding of the structure is solved by carrying out these products in an order which does not destroy the zeros not created by the preceding stages.

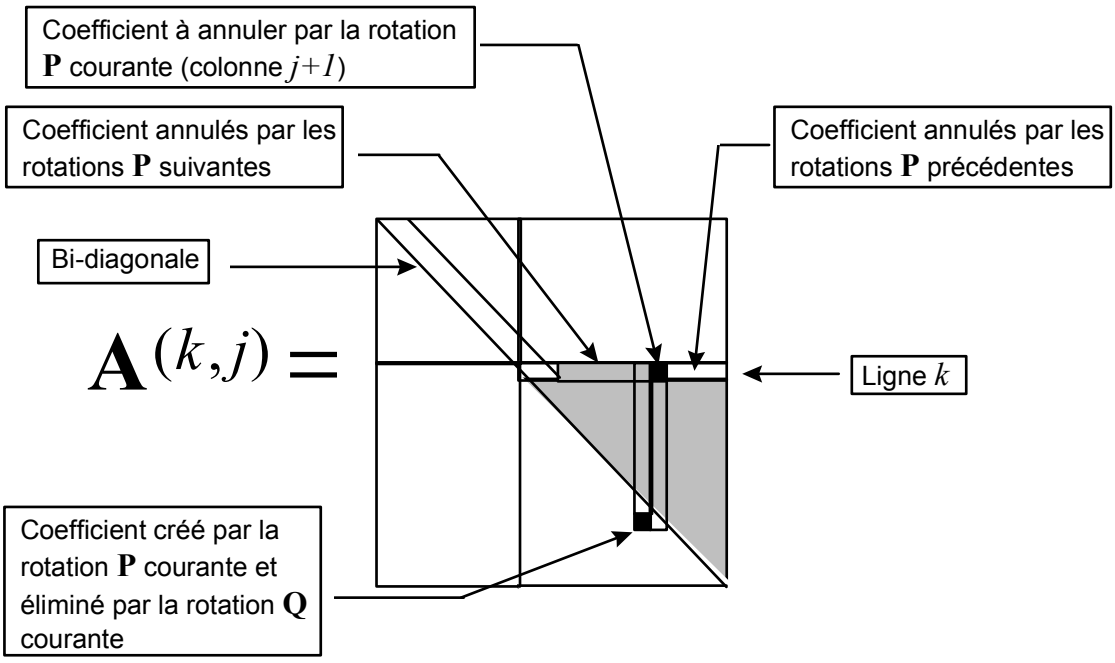
The algorithm thus cancels successively the coefficient $(k, j+1)$ for $j=m-1, m-2, ..., k+2$ by the application on the right of a rotation of Givens of the plan $(j, j+1)$. This rotation modifies only the columns j and $j+1$, which creates a parasitic coefficient in position $(j+1, j)$. This parasitic coefficient is then eliminated by the application on the left from transposed of a rotation of Givens in the plan $(j, j+1)$.

The process of passage of $\mathbf{A}^{(k)}$ with $\mathbf{A}^{(k+1)}$ can be then formalized by:

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

where angles $\theta_j^{(k)}$ and $\theta_{j=1/2}^{(k)}$ are selected to cancel the coefficient in position respectively $(k, j+1)$ of $\mathbf{A}^{(k+1,\,j)}$ and the coefficient in position $(j+1, j)$ of $\mathbf{A}^{(k+1,\,j=1/2)}$.

The structure of the matrices $\mathbf{A}^{(k+1,\,j)}$ is illustrated in the following figure:



At the conclusion of this recurrence, matrices and $\mathbf{P}^{(k)}$ and $\mathbf{Q}^{(k)}$ are given by:

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

so that matrices $\mathbf{P}$ and $\mathbf{Q}$ are worth:

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

5.4 Decomposition SVD of a higher bidiagonale

We present an algorithm of construction of decomposition SVD of a higher matrix bidiagonale $\mathbf{A}$ of order m. The algorithm thus builds two orthogonal matrices $\mathbf{Q}$ and $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such as:

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

The algorithm is drawn from [bib2].

5.4.1 Principle of the algorithm

The calculation algorithm of decomposition SVD diagonalise repeatedly the matrix A by means of the recurrence:

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

where matrices $\mathbf{Q}^{(k)}$ and $\mathbf{P}^{(k)}$ are orthogonal and the matrices $\mathbf{A}^{(k)}$ are higher bi--diagonals.

With convergence, we will have:

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

The idea of the iteration consists with:

- To choose $\mathbf{P}^{(k)}$ to make converge the algorithm QR applied to the diagonalisation of the matrix (known as normal) $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ without forming it explicitly. Indeed, the matrix $\mathbf{P}$ decomposition SVD of $\mathbf{A}$ is anything else only the matrix of the clean vectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$,
- To choose $\mathbf{Q}^{(k)}$ to preserve the higher structure bidiagonale the reiterated successive ones.

As in the case of the factorizations presented to sections 5.2 and 5.3, the matrices $\mathbf{Q}^{(k)}$ and $\mathbf{P}^{(k)}$ are built like product of rotations of Givens. The passage of $\mathbf{A}^{(k)}$ with $\mathbf{A}^{(k+1)}$ is then realized by:

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

where them $\mathbf{Q}^{(k,i)}$ and $\mathbf{P}^{(k,i)}$ are two rotations of the plan $(i-1,i)$ of respective angle θ_j and φ_i :

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

Rotations are alternatively applied on the right then on the left so that $\mathbf{A}^{(k+1)}$ preserve the higher structure bidiagonale of $\mathbf{A}^{(k)}$. With this intention:

The angle $\varphi_2^{(k)}$ is, for the moment, arbitrarily selected; the application of rotation $\mathbf{P}^{(k,2)}$ create a coefficient in position then $(2,1)$,

The angle $\theta_2^{(k)}$ is selected so that the application of rotation $\mathbf{Q}^{(k,2)}$ cancel the coefficient in position $(2,1)$, which creates a coefficient not no one in position $(1,3)$,

The angle $\varphi_3^{(k)}$ is selected so that the application of rotation $\mathbf{P}^{(k,3)}$ cancel the coefficient in position $(1,3)$, which creates a coefficient not no one in position $(3,2)$,

.
.
.

The angle $\theta_{m-1}^{(k)}$ is selected so that the application of rotation $\mathbf{Q}^{(k,m-1)}$ cancel the coefficient in position $(m-1,m-2)$, which creates a coefficient not no one in position $(m-2,m)$,

The angle $\varphi_m^{(k)}$ is selected so that the application of rotation $\mathbf{P}^{(k,m)}$ cancel the coefficient in position $(m-2,m)$, which creates a coefficient not no one in position $(m,m-1)$,

The angle $\theta_m^{(k)}$ is selected so that the application of rotation $\mathbf{Q}^{(k,m)}$ cancel the coefficient in position $(m,m-1)$, and stamps it $\mathbf{A}^{(k+1)}$ that is to say bidiagonale higher.

For any value of the angle, this process ensures maintains it structure bidiagonale higher than reiterated [éq 5.4.1-2]. We will see now how it is possible to choose this angle to make converge the iteration [éq 5.4.1-1].

5.4.2 Implicit Diagonalisation of the normal matrix

The algorithm of the preceding sub-section leaves unspecified the angle $\varphi_2^{(k)}$ the first rotation of $\mathbf{P}^{(k)}$. We will raise this indetermination in order to make matrix $\mathbf{P}^{(k)}$ the orthogonal matrix of a step QR, with spectral shift, applied to the diagonalisation of the normal matrix $\mathbf{M}=\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

With iteration SVD [éq 5.4.1-1] of the matrix $\mathbf{A}$, we associate an iteration on the normal matrix $\mathbf{M}=\mathbf{A}^{\mathrm{T}}\mathbf{A}$:

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

Iteration QR for the diagonalisation of the normal matrix

The transformation QR, with spectral shift σ_k, applied to $\mathbf{M}^{(k)}$ is written:

To factorize $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}$ in the form $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}=\mathbf{P}_\sigma\mathbf{R}_\sigma$
To build $\mathbf{M}_\sigma^{(k+1)}$ by $\mathbf{M}_\sigma^{(k+1)}=\mathbf{P}_\sigma\mathbf{R}_\sigma+\sigma_k\mathbf{I}$

where $\mathbf{P}_\sigma$ and $\mathbf{R}_\sigma$ are two matrices respectively orthogonal and triangular higher. Matrices $\mathbf{M}^{(k)}$ and $\mathbf{M}_\sigma^{(k+1)}$ are thus tridiagonales and similar:

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

From the practical point of view, the matrix $\mathbf{P}_\sigma$ is presented in the form of a product of rotations of Givens:

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

Angles ψ_k are selected so that the application on the left of $\mathrm{R}\left(k-1,k,\psi_k\right)$ with the matrix $\left[\prod_{I=k-1,k-2,\ldots,2}\mathrm{R}\left(1-1,1,\psi_1\right)\right]\left(\mathbf{M}^{(k)}-\sigma_k\mathbf{I}\right)$ cancel the coefficient of position $\left(k,k-1\right)$ in the matrix result.

Francis showed that the passage of $\mathbf{M}^{(k)}$ with $\mathbf{M}_{\sigma}^{(k+1)}$ do not require the explicit formation of the matrix $\mathbf{M}^{(k)}-\sigma_k\mathbf{I}$: the shift can be carried out implicitly. The theorem is stated as follows:

Theorem (Francis) : That is to say $\mathbf{X}$ an orthogonal matrix whose first column coincides with that of $\mathbf{P}_{\sigma}$. Under the assumptions:

1) $\mathbf{M}^{(k+1)}=\mathbf{X}^T\mathbf{M}^{(k)}\mathbf{X}$
2) $\mathbf{M}^{(k+1)}$ is tridiagonale,
3) Under-diagonal elements of $\mathbf{M}^{(k)}$ are all nonworthless (irreducibility of $\mathbf{M}^{(k)}$),

one has

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

where $\mathbf{D}$ is a diagonal matrix of diagonal coefficients all equal to ± 1.

Application to algorithm SVD

Consequently, the choice of the angle $\varphi_2^{(k)}$ the first rotation of iteration SVD [éq 5.4.1-1] consists of $\varphi_2^{(k)}=-\psi_2$. Thus $R\left(1,2,\varphi_2^{(k)}\right)=R\left(1,2,\psi_2\right)^T$ so that the first column of $\mathbf{P}^{(k)}$ coincide with the first column of $\mathbf{P}_{\sigma}$. Therefore, if all under-diagonal elements of $\mathbf{M}^{(k)}$ are nonworthless, then, the matrices $\mathbf{P}^{(k)}$ and $\mathbf{P}_{\sigma}$ are identified (with a multiplicative factor ± 1 close to the columns) and iteration SVD [éq 5.4.1-1] is equivalent to the application of the transformation QR, with shift, with the matrix $\mathbf{M}^{(k)}$.

Choice of the spectral shift

The shift is usually selected like the eigenvalue of the lower minor of order two of $\mathbf{A}^{(k)}$ nearest to $a_{m,m}^{(k)}$. This choice ensures a total convergence which, generally, is cubic.

Applicability of the theorem of Francis and phenomenon of decomposition

The use of the theorem of Francis supposes to it not nullity of all the under-diagonal coefficients of the matrices $\mathbf{M}^{(k)}$, which of anything is not guaranteed. Moreover, within the framework of the method QR of diagonalisation of a symmetrical matrix tridiagonale, the appearance of worthless under-diagonal coefficients is:

- Desirable: the worthless coefficients uncouple the diagonal blocks which they frame, which brings back the diagonalisation of the complete matrix to the diagonalisation of its diagonal blocks (this phenomenon is often called “decomposition”),
- Inevitable: the convergence of the algorithm towards an eigenvalue is interpreted algebraically like the appearance of a preceding diagonal block of order 1.

In the following sub-section, we will see which treatment it is appropriate to adopt in the presence of a decomposition.

5.4.3 Analysis of decomposition

The analysis of decomposition relates to each reiterated taken independently of the others, also, we will not use the superscript (k).

That is to say $d_1, d_2, \ldots, d_m$ and $e_2, e_3, \ldots, e_m$ respectively diagonal and on-diagonal elements of A. Under-diagonal elements of the normal matrix $\mathbf{M}=\mathbf{A}^T\mathbf{A}$ are then given by:

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

Let us suppose, to simplify, that only the coefficient $m_{l-1,l}$ is null. The matrix $\mathbf{M}$ then have a structure of two diagonal blocks whose meeting of the respective spectra gives the spectrum of $\mathbf{M}$. This decomposition takes place is for $e_1=0$ maybe for $e_1\neq 0 \text{ et } d_{l-1}=0$.

The case $e_1=0$ do not raise any difficulty. The matrix $\mathbf{A}$ then have a diagonal structure of two blocks which provide each one a part complementary to decomposition SVD of $\mathbf{A}$. Each block being bi--diagonal superior, without worthless on-diagonal coefficients, its decomposition SVD is calculable by the iteration [éq 5.4.1-1].

The case $e_1\neq 0 \text{ et } d_{l-1}=0$ is more delicate. Indeed, the iteration [éq 5.4.1-1] cannot be applied nor to the matrix $\mathbf{A}$, to violate the assumptions of the theorem of Francis, nor with none under matrices of $\mathbf{A}$, to ensure the structure bi--diagonal reiterated. This problem is circumvented by the post-multiplication of $\mathbf{A}$ by a series of rotations of Givens in the successive plans $(l-1,l),(l-1,l+1),\ldots,(l-1,m)$:

- The rotation of the plan $(l-1,l)$ cancel the coefficient $(l-1,l)$ and creates a coefficient in position $(l-1,l+1)$,
- The rotation of the plan $(l-1,l+1)$ cancel the coefficient $(l-1,l+1)$ and creates a coefficient in position $(l-1,l+2)$,
- The rotation of the plan $(l-1,l+1)$ cancel the coefficient $(l-1,l+1)$ and creates a coefficient in position $(l-1,l+2)$,

  .
  .
  .
- The rotation of the plan $(l-1,m)$ cancel the coefficient $(l-1,m)$ and does not create a coefficient.

So that the matrix produced by this process has the same structure as that corresponding to the case $e_1=0$.

5.4.4 Organization of the algorithm

The algorithm isolates successively each singular value, also, it exists an index k_p such as reiterated $\mathbf{A}^{(k_p)}$ breaks up into two diagonal blocks:

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

where $\mathbf{B}^{(k_p)}$ is a higher matrix bidiagonale of order p and $\mathbf{D}^{(k_p)}$ is a diagonal matrix of order $m-p+1$ gathering on its diagonal the found singular values.

From this reiterated, the algorithm applies the iteration of sub-section 5.4.1 to the submatrix $\mathbf{B}^{(k_p)}$ until the cancellation of the coefficient in position $(p-1,p)$, signal of the convergence of $p^{ième}$ singular value. Each step of the internal iteration, thus defined, is organized as follows:

- Analysis of decomposition,
- If the found decomposition corresponds to a diagonal element no one, a series of additional rotations is applied to find the structure of decomposition generated by a on-diagonal element no one. These rotations are built according to the method presented to sub-section 5.4.3,
- If the coefficient in position $(p-1,p)$ then does not produce decomposition under - matrix $\mathbf{B}^{(k_p)}$ is the object of a step of the iteration [éq 5.4.1-1] where, in accordance with the analysis of sub-section 5.4.2, the angle of the first rotation is selected so that this step is equivalent to the application of a transformation QR, with implicit spectral shift, on the associated normal matrix.

The complete convergence of the iteration is then obtained with the index k_m for which the submatrix $\mathbf{D}^{(k_p)}$ is of order m.

Of course, in practice, a coefficient is regarded as null as soon as it is lower, in absolute value, with a certain tolerance. The tolerance generally used for the problems of singular values is selected like the product of the precision machine by $\|\mathbf{A}\|_1$. Let us notice that in the case of a decomposition produced by a diagonal element no one with the tolerance chosen, the application of the series of additional rotations described to sub-section 5.4.3 creates under column of nonworthless elements under this coefficient. These elements are not awkward because they all are worthless with the selected tolerance.

6 Bibliography

1) P.G. CIARLET: “Introduction to the matric digital analysis and optimization” _MASSON (1985).

2) G.H. GOLUB, C. REINSCH “Singular been worth decomposition and least public gardens solutions” in “Handbook for Automatic Computation - Linear Algebra, vol. 2” J.H. WILKHINSON, C. REINSH Editors _SPINGER VERLAG (1971).

3) P. LASCAUX, R. THEODOR: “Matric digital analysis applied to the art of the engineer”, volumes 1 and 2 _MASSON (1986).

7 Description of the versions of the document

Version Aster	Author (S) Organization (S)	Description of the modifications
3	B.QUINNEZ, R.MICHEL EDF-R&D/MMN	Initial text