



Etude numérique de la méthode du gradient conjugué déflaté pour les résolutions successives de seconds membres multiples

Inria HiePACS

01/02/2020 au 27/09/2020

Département :

ENSEIRB-MATMECA (M3)

Auteur :

Victor LEDERER

Responsable :

Marc DURUFLÉ

Rapport datant du :

20 mars 2023

Laboratoire :

INRIA HIEPACS

Tuteurs :

Emmanuel AGULLO

Gilles MARAIT

Luc GIRAUD

Résumé

Le présent rapport rend compte du stage 3A effectué à l’Inria Bordeaux Sud-Ouest au sein de l’équipe HiePACS du 01/02/2020 au 31/07/2020 et du 03/09/2020 au 27/09/2020 dans le cadre de la formation à l’ENSEIRB-MATMECA. Le sujet principal est l’étude d’une méthode de déflation appliquée au gradient conjugué pour la résolution successive de systèmes linéaires à matrice fixée mais à plusieurs seconds membres. Cette méthode utilisant l’information spectrale, on s’attachera aussi à détailler le problème aux valeurs propres sous-jacent. Pour mener à bien cette étude, un prototype de la méthode est réalisé en python.

Abstract

Conjugate gradient method is used to solve linear system $Ax = b$ when the linear operator A is symmetric positive definite. This method is based on Krylov projection method and uses a Lanczos procedure to build a Krylov subspace. When it’s come to solve a linear system with multiple right-hand sides, one way to speed up the convergence is to reuse information build in the previous solve. The idea behind the deflated conjugate gradient method is to reuse information from conjugate gradient coefficients in order to build and solve an eigenvalue problem. From this, between each solve deflated conjugate gradient method computes k approximate eigenvectors associated to the extreme eigenvalues. Those eigenvectors are used to build a deflated subspace that is added to the Krylov subspace for the next solve. During this internship, we study which eigenvectors need to be deflated to speed up the convergence between each solve. Several methods such that Lanczos Rayleigh-Ritz, Harmonic-Ritz to find which one can produce the best eigenvector approximation.

Table des matières

1 INTRODUCTION	2
1.1 Systèmes linéaires et problème aux valeurs propres	2
1.1.1 Définition d'un système linéaire	2
1.1.2 Hypothèses sur la matrice du système linéaire	3
1.1.3 Méthodes de résolution	3
1.1.4 Valeurs et vecteurs propres	4
2 Cadre	5
2.1 Considérations numériques	5
2.1.1 Méthodes de sous-espace de Krylov	5
2.1.2 Méthode de Lanczos	6
2.1.3 Méthodes de Gradient	7
2.1.4 Correspondance entre les coefficients du gradient conjugué et l'algorithme de Lanczos	9
2.1.5 Gradient conjugué déflaté	9
2.2 Comparaison de la convergence du <i>CG</i> et du <i>CG-DEF</i>	9
2.3 Objectif du PFE	10
2.4 Environnement expérimental	12
3 Injection spectrale donnée a priori et effet sur la convergence du <i>CG-DEF</i>	13
3.1 Déflation avec des vecteurs propres exacts	13
3.1.1 Matrices avec cluster central	13
3.1.2 Matrices avec clusters <i>LDC</i> et <i>MDC</i> de taille 1	15
3.1.3 Approximation de rang faible	15
3.1.4 Matrices avec clusters <i>LDC</i> et <i>MDC</i> de taille 4	19
3.2 Déflation sur un espace perturbé	19
4 Convergences des paires de Ritz/Harmonic Ritz vers les paires propres	22
4.1 Etude avec les approximations de Rayleigh-Ritz	22
4.1.1 Définition	22
4.1.2 Expérimentations	23
4.1.3 Phénomène de Ghost values et ré-orthogonalisation	25
4.2 Étude avec les approximations d'Harmonic-Ritz	25
4.2.1 Définition	25
4.2.2 Expérimentations	26
5 Déflation avec l'information spectrale d'une résolution précédente	26
5.1 Approche basée sur les approximations de Rayleigh-Ritz	28
5.2 Approche basée sur les approximations d'Harmonic Ritz	28
6 Déflation avec l'information spectrale de multiples résolutions successives, approche basée HR	33
6.1 Algorithme du <i>CG-DEF</i> pour la résolution successive de multiples seconds membres	33
6.1.1 Expérimentations	33
7 Conclusion	34
8 Annexes	37

A Annexes	37
A.1 Annexe A : construction du problème (22)	37
A.2 Annexe B : Lanczos Rayleigh-Ritz, convergences vers les paires propres	37
A.3 Annexe C : Lanczos Rayleigh-Ritz, convergences des paires propres	38
A.4 Annexe D : Lanczos Harmonic-Ritz, Quotient de Rayleigh	38
A.5 Annexe E : Lanczos Harmonic-Ritz, convergences des paires propres	38

Table des figures

1	Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater.	14
2	Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater. LDC et MDC sont de taille 1.	16
3	Evolution de l'erreur inverse pour la Matrice A_4 , dimension 300×300 , LDC et MDC de taille 1.	17
4	Zoom sur les premières itérations, effet de la déflation de MDC sur la précision de la solution à l'itération 0, cluster de taille 1.	18
5	Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater. LDC et MDC sont de taille 4.	20
6	Matrice A_2 , cluster LDC de taille 1 : évolution de l'erreur inverse en fonction du sinus de l'angle entre le vecteur propre exact u et le vecteur perturbé w du cluster.	21
7	Matrice A_3 , cluster MDC de taille 1 : évolution de l'erreur inverse en fonction du sinus de l'angle entre le vecteur propre exact u et le vecteur perturbé w du cluster.	22
8	Évolution du sinus de l'angle entre les vecteurs propres calculés via Lanczos Rayleigh-Ritz et les vecteurs propres exacts, cluster de taille 4.	24
9	Evolution du sinus de l'angle entre les vecteurs propres calculés via Lanczos Harmonic-Ritz et les vecteurs propres exacts, cluster de taille 4.	27
10	Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de RR, matrice A_2	29
11	Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de RR, matrice A_3	30
12	Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de HR, matrice A_2	31
13	Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de HR, matrice A_3	32
14	Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ pour la résolution successives de $\nu = 20$ systèmes, approche basée sur les approximations de HR. .	39
15	Convergence du quotient de Rayleigh vers les valeurs propre de A_i via Lanczos Rayleigh-Ritz, cluster de taille 4.	40
16	Convergence des paires propres calculées via Lanczos Rayleigh-Ritz, cluster de taille 4.	41
17	Convergence du quotient de Rayleigh vers les valeurs propres de A_i via Lanczos Harmonic-Ritz.	42
18	Lanczos Harmonic-Ritez, convergence de la paire propre, cluster de taille 4.	43

Liste des symboles

Convention d'écriture

ν	Nombre de second membres
W	Matrice dont les colonnes engendrent le sous-espace de déflation
k	Nombre de colonnes de W
ℓ	Nombre d'informations à conserver
\mathbb{R}	Corps des nombres réels
$M_{m \times n}(\mathbb{R})$	Ensemble des matrices réelles de m lignes à n colonnes [1][Chap 0.2, p 5]
$[c_1 \dots c_n]$	Expression d'une matrice $C \in M_{m \times n}$ définie par une collection de colonnes [2][Chap 1, p 6]
LD	Least Dominant [3]
LDC	Least Dominant Cluster
CC	Centered Cluster
MD	Most Dominant [3]
MDC	Most Dominant Cluster
$\angle(w, u)$	Angle entre deux vecteurs w et u
$\sin \angle(w, u)$	Sinus de l'angle entre les vecteurs w et u
$\cos \angle(w, u)$	Cosinus de l'angle entre les vecteurs w et u
θ	Ordre de grandeur entre CC et MDC ou LDC

Opérateurs mathématiques

$\langle x, y \rangle = x^T \cdot y$	Produit scalaire Euclidien notation issue de [4]
\cdot	matrix-matrix multiplication operator ($M_{m \times p} \times M_{p \times n} \longrightarrow M_{m \times n}$)
$*$	Scalar multiplication operator
$C = diag(c_{11}, \dots, c_{nn})$	diag() construit $C \in M_{n \times n}$ matrice diagonale telle que $C_{ij} = 0 \forall j \neq i$ [5][chap 1.3]
$C = tridiag(c_{ii-1}, c_{ii}, c_{ii+1})$	tridiag() construit $C \in M_{n \times n}$ matrice tridiagonale telle que $C_{ij} = 0 \forall j - i > 1$ [5][chap 1.3]

Liste des solveurs

CG	Gradient conjugué
PCG	Gradient conjugué préconditionné
$-DEF$	Attribut indiquant que la méthode est déflaté
$-noreortho$	Attribut indiquant qu'il n'y a pas de re-orthogonalisation

Remerciements

Je tiens à remercier mes tuteurs de stages, Emmanuel Agullo, Luc Giraud et Gilles Marait. Emmanuel et Luc m'ont aidé tout au long du stage pour les aspects théoriques et techniques du rapport. Gilles pour son aide concernant la partie programmation.

Je souhaite aussi remercier mes collègues Marek Felsoci et Mathieu Simonnin pour leur aide sur certains détails.

Je remercie aussi Chrystel Plumejeau pour son aide concernant la partie administrative du stage.

1 INTRODUCTION

Les techniques de résolution de systèmes linéaires provenant de modélisations physiques n'ont cessé de se développer et de s'enrichir. Pour preuve on compte aujourd'hui des solveurs hybrides mêlant méthodes directes et itératives associés à des préconditionneurs de haut niveau.

Les méthodes itératives basées sur les techniques de projection sur un sous-espace de Krylov [6] ont été développés avec un nombre important de variantes, notamment afin de réduire le coût mémoire/temps de calcul dans le cas de seconds membres multiples. On compte par exemple pour le GMRES [7] les variantes restarted [8], augmented et deflated [9] et block [10].

Comme la construction du sous-espace de Krylov est réalisée par la méthode d'Arnoldi ces variantes ont été étendues au cas d'un opérateur symétrique défini positif en appliquant la procédure de Lanczos. La méthode du gradient conjugué (*CG*) [11] fut largement étudiée et étendue à la résolution de plusieurs seconds membres dans une version seed [] sa version block [12] et ses versions augmentées[13] et deflatées [14].

L'objectif de ce stage est la description de la méthode du gradient conjugué déflaté (*CG-DEF*) pour la résolution successive de plusieurs seconds membres. Cette méthode sépare l'espace de recherche en deux sous-espaces complémentaires orthogonaux, l'un étant un sous espace de Krylov, l'autre étant le sous-espace de déflation. Cette technique est particulièrement utilisée lorsque le spectre de la matrice du système est constitué de clusters de valeurs propres extrêmes qui pénalisent la convergence du gradient conjugué. Un choix naturel de l'espace de déflation est alors celui engendré par les vecteurs propres correspondants. La principale difficulté des méthodes de sous-espace de Krylov déflaté étant alors le calcul à moindre coût de ces vecteurs propres. L'avantage du *CG* est qu'il est basé sur la procédure de Lanczos pour construire le sous-espace de Krylov au cours des itérations, ce qui permet en même temps de générer une approximation de ces vecteurs. Dans le cadre du *CG-DEF* nombre de variantes de Lanczos sont à prendre en compte [15], [3], [16] pour, d'une part, faire face à la perte d'orthogonalité du sous espace généré par Lanczos et, d'autre part, obtenir la meilleure approximation spectrale possible. Une étude récente [17] souligne que l'obtention d'un taux de convergence optimal de *CG-DEF* peut se faire avec une approximation spectrale de faible précision. Ce rapport reprend les bases de la déflation et en expose le principe. Pour ce faire on rappellera la définition d'un système linéaire ainsi que les méthodes de résolutions de base. Le problème aux valeurs propres sera aussi exposé. Une seconde partie s'attachera à détailler les méthodes liées au gradient conjugué déflaté (*CG-DEF*) et définira les cas tests numériques utilisés pour décrire la convergence de la méthode. Une troisième partie étudiera la convergence du *CG-DEF* lorsque l'on dispose déjà d'un espace de déflation. Puis nous détaillerons les méthodes permettant de construire des approximations spectrales. S'en suivra les résultats de convergence du *CG* suivant plusieurs résolutions successives et un comparaison avec d'autre méthodes.

1.1 Systèmes linéaires et problème aux valeurs propres

Dans cette section nous rappellerons la définition d'un système linéaire et d'un problème aux valeurs propres, et nous expliciterons les hypothèses faites sur la matrice du système linéaire. Les différents types de résolution (direct/itératif) seront expliquées, puis nous mettrons l'accent sur la méthode itérative en évoquant les variantes stationnaires et par projection.

1.1.1 Définition d'un système linéaire

Soit le n-uplet de réels $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ avec $n \in \mathbb{N}$ un entier supérieur ou égal à 1. Soient n coefficients réels $(a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ et $b \in \mathbb{R}$ le second membre. On appelle résolution d'un système linéaire le problème suivant :

$$\boxed{\text{Trouver } (x_1, x_2, \dots, x_n) \text{ tel que :} \sum_{i=1}^n a_i x_i = b} \quad (1)$$

Les $(x_i)_{i=1,\dots,n}$ sont nommées les inconnues du problème associé à l'unique équation (1), qui est donc dit de n inconnues. Voyons maintenant le cas où le problème est constitué de n équations. Pour cela on considère maintenant que le second membre est $(b_j)_{j=1,\dots,n}$ et que les coefficients sont $(a_{ij}) \in \mathbb{R} \forall (i,j) \in \mathbb{N}^{n \times n}$. Le problème à n inconnues et n équations est donc :

$$\text{Trouver } (x_1, x_2, \dots, x_n) \text{ tel que : } \begin{cases} \sum_{j=1}^n a_{1j}x_j = b_1 \\ \vdots \\ \sum_{j=1}^n a_{ij}x_j = b_i \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j = b_n \end{cases} \quad (2)$$

On peut mettre sous forme matricielle l'équation (2), de telle sorte qu'avec l'opérateur linéaire $A \in \mathbb{M}_{n \times n}(\mathbb{R})$, qui à $x \in \mathbb{R}^n$ associe l'élément $b \in \mathbb{R}^n$, et A de la forme $A_{ij} = a_{ij}, \forall (i,j) \in \mathbb{N}^{n \times n}$, l'on constitue le système linéaire suivant :

$$Ax = b \quad (3)$$

La résolution des équations des problèmes physiques tels que la mécanique des fluides (équations de Navier-Stokes) ou de la mécanique des solides sont parfois des problèmes trop complexes pour être résolus analytiquement. Cela a amené la communauté scientifique à développer des méthodes numériques permettant d'approximer (discréteriser) le problème initial et de pouvoir l'écrire sous la forme (3). Deux de ces méthodes sont connues sous le nom de méthode des volumes finis [18], [19], particulièrement adapté aux équations des fluides ainsi que la méthode des éléments finis [20], [21] pour les équations issues de la mécanique des solides.

1.1.2 Hypothèses sur la matrice du système linéaire

On se place dans le cas où la matrice A est symétrique définie positive (SDP). A symétrique signifie $A = A^T$ et A définie positive signifie que $\forall x \in \mathbb{R}^n, x^T A x \geq 0$ et $x^T A x = 0$ si et seulement si $x = 0$. Sous ces hypothèses A est non singulière, il existe donc une unique solution x^* au problème (3) telle que $x^* = A^{-1}b$.

1.1.3 Méthodes de résolution

Pour calculer une approximation de l'unique solution x^* au problème (3), voici les deux classes principales. D'une part, les méthodes directes calculent les inconnues du problème (2) une à une et sont vues comme l'expression matricielle de l'élimination de Gauss [22][chap 4,p 147 et 172]. D'autre part, les méthodes itératives calculent une suite d'approximations notée $(x_k)_{k \in \mathbb{N}}$ qui est construite de telle sorte à converger vers x^* .

Méthodes directes. Elles consistent en une étape de décomposition de la matrice A en un produit matriciel de deux matrices par exemple tel que $A = BC$ avec B et C deux matrices triangulaires, respectivement inférieure et supérieure. Une telle décomposition peut être obtenue via la décomposition LU de A introduite par Tadeusz Banachiewicz en 1938 [23]. Elle peut également s'obtenir par la factorisation de Cholesky où $A = LL^T$. A admet une décomposition de Cholesky si et seulement si A est SDP. Une fois la matrice décomposée sous l'une de ces formes, une étape de résolution de deux systèmes linéaires intervient, telle que le problème (3) se ramène à la résolution successive de deux systèmes triangulaires

$$\begin{cases} By = b \\ Cx = y \end{cases}$$

On résout d'abord le système d'inconnue y puis celui en x . Cette méthode est robuste et précise [22][chap 4,p 165,177] mais nécessite un coût mémoire important. On peut citer *MUMPS* [24] et *PasTiX* [25] comme exemple de solveurs linéaires directs creux.

Méthodes itératives. En supposant que x^* soit la solution exacte de (3), les méthodes itératives construisent une suite d'approximations $(x_k)_{k \in \mathbb{N}}$ tel que $(x_k \rightarrow x^*)$ lorsque $(k \rightarrow \infty)$. La première classe des méthodes itératives est celle dite de relaxation. Elle constitue à modifier entre chaque itération quelques composantes de l'approximation de telle sorte que les composantes correspondantes du vecteur résidu $r_k = b - A.x_k$ soient nulles. En pratique les méthodes de relaxation se basent sur une décomposition de A telle que $A = M - N$ où M et N sont deux matrices de $\mathbb{M}_{n \times n}(\mathbb{R})$, et l'approximation à l'itération suivante est donnée par la résolution du système $M.x_{k+1} = N.x_k + b$. Le choix des matrices M et N ainsi que les résultats de convergence sont donnés en [5][chap 4]. Les méthodes de projection [5][chap 5] constituent la seconde classe des méthodes itératives. Ces méthodes cherchent à construire la suite d'approximation $(x_m)_{m \in \mathbb{N}}$ dans un sous-espace K_m de \mathbb{R}^n de dimension m et en imposant au vecteur résidu r_m d'être orthogonal à un sous espace de dimension m noté L_m de \mathbb{R}^n cette condition est appelée condition de Petrov-Galerkin. On parle de projection oblique lorsque K_m et L_m diffèrent, sinon de projection orthogonale. En notant x_0 l'approximation initiale de la solution x^* du problème (3), une méthode de projection définit le problème suivant :

$$\boxed{\text{Trouver } (x_m \in x_0 + K_m) \text{ tel que } (b - Ax_m \perp L_m)} \quad (4)$$

Soit $V = [v_1 \mid \dots \mid v_m] \in \mathbb{M}_{n \times m}(\mathbb{R})$ avec $(v_i)_{i=1,\dots,m}$ une base de K_m et $W = [w_1 \mid \dots \mid w_m] \in \mathbb{M}_{n \times m}(\mathbb{R})$ avec $(w_i)_{i=1,\dots,m}$ une base de L_m , le problème (4) revient à construire $x_{m+1} = x_m + V y_m$ avec y_m le vecteur assurant la condition de Petrov-Galerkin autrement dit, $r_m \perp L_m$ si et seulement si $(W^T A V)y_m = W^T r_m$. Les méthodes itératives sont mises en oeuvres avec un coût mémoire moins important que les méthodes directes et leur implémentation en parallèle est généralement considérée comme plus aisée.

1.1.4 Valeurs et vecteurs propres

Les problèmes aux valeurs propres jouent un rôle majeur en analyse numérique et sont utilisés dans de nombreux problèmes, tant, numériques que physiques. Comme la méthode du gradient conjugué déflaté, que nous étudions, construit un tel problème pour générer un espace de déflation, cette partie détaille le principe de base du problème aux valeurs propres sans développer les méthodes de résolution.

Problème aux valeurs propres. On considère la matrice $A \in \mathbb{M}_{n \times n}(\mathbb{R})$. Soit $\lambda \in \mathbb{R}$ et $x \in \mathbb{R}^{n*}$ un vecteur non nul. Le problème aux valeurs propres se formule comme il suit :

$$\boxed{\text{Trouver un couple } (\lambda, x) \text{ tel que } (Ax = \lambda x)} \quad (5)$$

Si (5) est vérifié alors on appelle λ valeur propre, x vecteur propre de A , et le couple (λ, x) paire propre de A . Si A est symétrique alors elle possède n valeurs propres réelles et n vecteurs propres linéairement indépendants. Dans ce cas, en notant $X = [x_1 \mid x_2 \mid \dots \mid x_n]$, et $\Lambda = \text{diag}(\lambda_1, \lambda_1, \dots, \lambda_n)$ tels que (λ_i, x_i) soit la $i^{\text{ème}}$ paire propre de A , (5) se ré-écrit sous la forme suivante :

$$\boxed{\text{Trouver le couple } (\Lambda, X) \text{ tel que } (AX = \Lambda X)} \quad (6)$$

X est appelé espace propre de A et puisque l'on considère un cadre où A est symétrique, on a $X^{-1} = X^T$.

Problème aux valeurs propres généralisé. Le problème aux valeurs propres de deux matrices A et B symétriques définies positives de $\mathbb{M}_{n \times n}(\mathbb{R})$ s'écrit sous la forme :

$$Ax_i = \lambda_i Bx_i, \forall i \in [|1, 2, \dots, n|] \quad (7)$$

qui sous forme matricelle se ré-écrit : $AX = BX\Lambda$. [26][Chap 1, p 18] défini (λ_i, x_i) comme étant la $i^{\text{ème}}$ paire propre de l'ensemble (A, B) aussi appelé pencil (A, B) . Lorsque $B = I_n$ on retrouve les problèmes aux valeurs propres précédents.

Quotient de Rayleigh et méthode de la puissance itérée. En considérant un vecteur propre x du pencil (A, B) comment déterminer la valeur propre associée ? Il s'agit du problème au moindres carrés suivant $\min_{\alpha \in \mathbb{R}} \|AX - \alpha Bx\|_2$ qui se résoud via les équations normales associées : $x^T Ax = \alpha x^T Bx$. On a alors $\alpha = \frac{x^T Ax}{x^T Bx}$, que l'on redéfinit comme étant le quotient de Rayleigh noté $\rho(x, A, B)$. Le quotient de Rayleigh d'un vecteur $x \in \mathbb{R}^n$ associé au pencil (A, B) est le scalaire défini par $\rho(x, A, B) = \frac{x^T Ax}{x^T Bx}$ tel que si (λ, x) vérifie l'équation (7) alors $\rho(x, A, B) = \lambda$. Par ailleurs, si u est un vecteur propre de A et x un vecteur de \mathbb{R}^n alors $\rho(x, A) - \rho(u, A) = O(\|x - u\|^2)$ lorsque $x \rightarrow u$ [22][chap :5, p :204]. De part cette propriété de convergence, le quotient de Rayleigh est impliqué dans de nombreuses méthodes permettant d'approximer le spectre de (A, B) . Par exemple pour $B = I_n$, la méthode de la puissance itérée construit de manière récursive l'approximation u^k d'un vecteur propre x de A tel que $u^k = Au^{k-1}$ puis calcul l'approximation de la valeur propre associée à u^k via le quotient de Rayleigh $\rho(v^k, A)$ où $v^k = \frac{u^k}{\|u^k\|_2}$. [22][chap 5,p 205] montre que le couple $(\rho(v^k, A), v^k)$ ainsi obtenu converge vers le couple de vecteur propre associé à la valeur propre de A de plus grande valeur absolue.

Dans cette section nous avons explicité les briques de base nécessaires à la méthode du *CG-DEF*. La section suivante se propose de définir la méthode de sous-espace de Krylov déflaté en commençant par expliquer le lien entre le *CG* et Lanczos.

2 Cadre

Cette section s'attache à détailler les bases de la méthode du gradient conjugué déflaté. Comme la méthode déflatée utilise de l'information spectrale de la matrice A , on étudiera également la méthode de Lanczos, qui, à partir de coefficients calculés par le gradient conjugué, permet de générer des paires de Ritz, autrement dit des approximations des paires propres de A .

2.1 Considérations numériques

Nous allons présenter le cadre général des méthodes de sous-espace de Krylov, qui sont vues comme des méthodes de projections particulières. Le gradient conjugué est une de ces méthodes pour la résolution de systèmes linéaires faisant intervenir des opérateurs SDP.

2.1.1 Méthodes de sous-espace de Krylov

Cette partie s'attache à préciser la construction des sous-espaces K et d'approximer $A^{-1}b$ par $p(A)b$ avec p un polynôme. On définit $K_m(A, b)$ comme étant le sous-espace de Krylov vérifiant $K_m(A, b) = \text{vect}\{b, A.b, A^2.b, \dots, A^{m-1}.b\}$ [27][Chap 2.2]. En se référant à la méthode de la puissance, le choix d'une telle séquence $(A^k b)_{k=0, \dots, m-1}$ permet d'introduire de l'information spectrale dans les espaces K_m et L_m . Lorsque l'on souhaite résoudre le problème (4), le sous-espace de Krylov est défini par $K_m(A, r_0) = \text{Vect}\{r_0, A.r_0, A^2.r_0, \dots, A^{m-1}.r_0\}$ où $r_0 = b - Ax_0$. Par la suite, nous exposons certaines propriétés du sous-espace de Krylov, pour cela notons $v \in \mathbb{R}^n$ tel que $K_m(A, v) = \text{Vect}\{v, A.v, A^2.v, \dots, A^{m-1}.v\} = \text{Vect}\{v_0, v_1, \dots, v_{m-1}\}$. $K_m(A, v)$ étant un sous-espace de \mathbb{R}^n , les vecteurs $v_k = A^k v$, $\forall k \in [|0, m-1|]$, doivent être linéairement indépendants. Calculer directement $v_k = A^k v$ ne permet pas d'assurer cette propriété lorsque n ou m sont grands. On utilise la procédure de Lanczos qui permet de calculer la matrice orthogonale $V_m = [v_0, \dots, v_{m-1}] \in \mathbb{M}_{n \times m}(\mathbb{R})$, avec $(v_i)_{i=0, \dots, m-1}$ une base de $K_m(A, v)$. Ainsi la procédure de Lanczos permet à la fois de déterminer un base orthogonale pour le sous-espace de Krylov et à la fois d'obtenir une matrice $T_m = (V_m^T A V_m)$ ayant m valeurs propres dans le spectre de A tant que $m \leq n$ avec n le rang de A (cf, Section 2.1.4). Par ailleurs, $\forall x \in K_m(A, v)$, $x = (\sum_{i=0}^{m-1} \alpha_i A^i)v$. On peut donc écrire x sous la forme polynomial suivante : $x = p(A)v$. Montrons que le polynôme p est de degré au plus $(m-1)$ et est tel que $p(A) = A^{-1}$. [6][chap.5] défini le polynôme minimal $q(t)$ de A comme étant l'unique polynôme unitaire de

degré minimal tel que $q(A) = 0$. A étant symétrique réelle, elle est diagonalisable, on notera donc m' le nombre de valeurs propres distinctes parmi les n valeurs propres de A tel que $q(t) = \sum_{j=0}^{m'} \alpha'_j t^j$. Lorsque A est non singulière, [6][chap 5] montre que $\alpha'_0 \neq 0$ et par définition de q , on a :

$$0 = q(A) = \sum_{j=0}^{m'} \alpha'_j A^j$$

On peut donc écrire :

$$\begin{aligned} -\alpha'_0 * I &= \sum_{j=1}^{m'} \alpha'_j A^j \\ \Leftrightarrow -\alpha'_0 * A^{-1} &= \sum_{j=0}^{m'-1} \alpha'_{j+1} A^j \\ \Leftrightarrow A^{-1} &= \frac{-1}{\alpha'_0} \sum_{j=0}^{m'-1} \alpha'_{j+1} A^j \end{aligned}$$

Donc $p(A) = A^{-1}$ si et seulement si $m = m'$ et $\alpha_i = \frac{\alpha'_{i+1}}{\alpha'_0} \forall i \in [|0; m-1|]$. Ainsi $K_m(A, v)$ contient la solution x^* et est donc un espace adapté pour construire une approximation de la solution par la méthode de projection. D'autre part, comme la dimension de K_m est donnée par le degré du polynôme minimal de A , une méthode de projection convergera d'autant plus vite que $\dim(K_m)$ est petit.

2.1.2 Méthode de Lanczos

La méthode de Lanczos est une méthode d'Arnoldi pour le cas particulier des matrices SDP. Cette méthode permet de construire $K_m(A, v) = Vect\{v_0, v_1, \dots, v_{m-1}\}$ tel que $V_m = [v_0 | v_1 | \dots | v_{m-1}]$, soit une matrice de rang plein et, $\forall x \in K_m(A, v)$, il existe un unique polynôme p_m tel que $x = p_m(A)v = A^{-1}v$. Cette méthode permet d'autre part de construire la matrice tridiagonale $T_m = V_m^T A V_m$ tel que p_m soit le polynôme caractéristique de T_m . Autrement dit, le spectre de T_m contient m approximations des valeurs propres de A . On notera la propriété d'orthogonalité des colonnes de V_m tel que $I_m - V_m^T V_m = 0$. La procédure de Lanczos (algorithme 1) est la suivante :

Algorithm 1 Lanczos Algorithm to build T_m and V_m

- 1: compute $r_0 = b - Ax_0$, set $\eta = \|r_0\|_2$ and $v_0 = \frac{r_0}{\eta}$
 - 2: **for** $k = 1, m$ **do**
 - 3: $w_k = Av_k - \eta_{k-1}v_{k-1}$ if $k == 1$ else Av_k
 - 4: $\delta_k = v_k^T w_k$
 - 5: $w_k = w_k - \delta_k v_k$
 - 6: $\eta_{k+1} = \|w_k\|_2$ if $\eta_{k+1} == 0$ set $m = k$ break
 - 7: $v_{k+1} = \frac{w_k}{\eta_{k+1}}$
 - 8: **end for**
 - 9: set $T_m = tridiag(\eta_i, \delta_i, \eta_{i+1},)$
 - 10: set $V_m = [v_0 | v_1 | \dots | v_{m-1}]$
-

Les valeurs propres de la matrice T_m sont appelées valeurs de Ritz. [22][chap 6, p 278-279] explique que ces valeurs de Ritz tendent à converger plus rapidement vers les valeurs propres de A situées dans des clusters de petite densité. En définissant des cas tests numériques en section 2.1.4, on étudiera les propriétés de convergence de cette méthode en section 4.1.2.

2.1.3 Méthodes de Gradient

Il s'agit de méthodes itératives basées sur un problème d'optimisation. Soit la fonctionnelle suivante : $f(x) = 0,5 \langle Ax, x \rangle - \langle b, x \rangle$ et soit $h \in \mathbb{R}^n$ tel que :

$$\begin{aligned}
f(x + h) &= 0,5 \langle A(x + h), x + h \rangle - \langle b, x + h \rangle \\
&= 0,5 \langle Ax, x \rangle + 0,5 \langle Ah, x \rangle + 0,5 \langle A.x, h \rangle + 0,5 \langle Ah, h \rangle - \langle b, x \rangle - \langle b, h \rangle \\
&= f(x) + 0,5 \langle A^T x, h \rangle + 0,5 \langle Ax, h \rangle - \langle b, h \rangle + 0,5 \langle Ah, h \rangle \\
&= f(x) + \langle 0,5(A^T + A)x - b, h \rangle + 0,5 \langle Ah, h \rangle
\end{aligned} \tag{8}$$

On en déduit que $\nabla(f(x).h) = \langle 0,5(A^T + A)x - b, h \rangle$ donc en particulier, on a $\nabla f(x) = 0,5(A^T + A)x - b$. Comme nous avons fait l'hypothèse que A est symétrique, $\nabla f(x) = Ax - b$. Donc x est solution du problème (3) si et seulement si $\nabla f(x) = 0$. Par ailleurs, si $x = A^{-1}b$ et pour $y \in \mathbb{R}^n$ tel que $h = y - x$ on a en injectant h dans l'équation (8) : $f(y) = f(x) + 0,5 \langle A(y-x), (y-x) \rangle$ et comme A est aussi définie positive, on a $\langle A(y-x), (y-x) \rangle \geq 0$; ce qui implique $f(y) \geq f(x)$. Donc x est un minimum global de f . Une méthode de gradient s'attache donc à calculer une approximation de x^* en calculant une approximation du minimum de f . L'itération d'une telle méthode est de la forme : $x_{k+1} = x_k + \alpha_k p_k$, avec $\alpha_k \in \mathbb{R}$ le $k^{\text{ième}}$ pas de descente et $p_k \in \mathbb{R}^n$ la $k^{\text{ième}}$ direction de descente tels que $f(x_{k+1}) \leq f(x_k)$.

Méthode du Gradient conjugué : cette méthode est une méthode du type gradient qui s'attache à minimiser la norme A de l'erreur directe. On note $e_k = x_k - x^*$ cette erreur et $r_k = b - Ax_k$ le vecteur résidu, on a notamment l'égalité suivante : $Ae_k = -r_k$. Pour calculer le pas de descente, la méthode du gradient conjugué impose la contrainte de A -orthogonalité entre e_{k+1} et p_k . On a ainsi :

$$\begin{aligned}
0 &= \langle Ae_{k+1}, p_k \rangle \\
&= \langle A(x_{k+1} - x^*), p_k \rangle \\
&= \langle A(x_k + \alpha_k p_k - x^*), p_k \rangle \\
&= \langle Ae_k, p_k \rangle + \alpha_k \langle Ap_k, p_k \rangle \\
\implies \alpha_k &= \frac{\langle r_k, p_k \rangle}{\langle Ap_k, p_k \rangle} \\
\iff \alpha_k &= \frac{p_k^T r_k}{p_k^T A p_k}
\end{aligned} \tag{9}$$

La méthode du gradient conjugué est aussi un méthode de Krylov avec $K_m(A, r_0)$ le sous espace de Krylov construit par la procédure de Lanczos et $L_m = K_m$. La condition de Petrov-Galerkin $r_k \perp L_m$ est vérifiée si et seulement si :

$$\begin{aligned}
0 &= \langle r_{k+1}, r_k \rangle \\
&= \langle r_k, r_k \rangle - \alpha_k \langle Ap_k, r_k \rangle \\
\implies \alpha_k &= \frac{\langle r_k, r_k \rangle}{\langle Ap_k, r_k \rangle} \\
\iff \alpha_k &= \frac{r_k^T r_k}{r_k^T A p_k}
\end{aligned} \tag{10}$$

Pour monter que (10) implique (9), nous supposons comme [5][chap 6,p 190] que l'itération sur le vecteur direction de descente s'écrit $p_{k+1} = r_{k+1} + \beta_k p_k$. On a alors $\langle r_k, p_k \rangle = \langle r_k, r_k + \sum_{i=1}^k \beta_{k-i} r_{k-i} \rangle$. En supposant que (10) est vérifié on a alors $\langle r_k, p_k \rangle = \langle r_k, r_k \rangle$. Enfin, il reste à monter que $\langle Ap_k, r_k \rangle = \langle Ap_k, p_k \rangle$. En écrivant le vecteur résidu en fonction de

la direction de descente, il vient $\langle Ap_k, p_k - \beta_{k-1}p_{k-1} \rangle = \langle Ap_k, p_k \rangle$ qui est vérifiée si et seulement si $\langle Ap_k, p_{k-1} \rangle = 0$. On montre cela via l'égalité suivante :

$$\begin{aligned}
\langle Ae_{k+1}, p_{k-1} \rangle &= \langle A(x_k + \alpha_k p_k - x^*), p_{k-1} \rangle \\
&= \langle Ae_k, p_{k-1} \rangle + \alpha_k \langle Ap_k, p_{k-1} \rangle \\
&\implies \langle Ae_{k+1} - Ae_k, p_{k-1} \rangle = \alpha_k \langle Ap_k, p_{k-1} \rangle \\
&\iff \langle -r_{k+1} + r_k, p_{k-1} \rangle = \alpha_k \langle Ap_k, p_{k-1} \rangle \\
&\implies \langle Ap_k, p_{k-1} \rangle = 0
\end{aligned} \tag{11}$$

La dernière égalité de (11) s'obtient par décomposition de $p_{k-1} = r_{k-1} + \sum_{i=1}^{k-1} \beta_{k-1-i} r_{k-1-i}$ et l'orthogonalité des vecteurs résidus. De plus en appliquant le raisonnement de l'équation (11) à $\langle Ae_i, p_j \rangle \quad \forall i \leq k, \forall j \neq i$ on montre la A -orthogonalité des directions de descentes. Finalement cette dernière propriété permet de déterminer $\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$ et d'écrire l'algorithme 2 du Gradient conjugué.

Algorithm 2 CG

```

1: compute  $z = Ax_0$  and  $r_0 = b - z$ , set  $p = r_0$ 
2: for  $k = 0, maxiter$  do
3:    $\alpha_k = \frac{r_k^T r_k}{r_k^T z}$ 
4:    $x_{k+1} = x_k + \alpha_k p_k$ 
5:    $r_{k+1} = r_k - \alpha_k z$ 
6:   if  $\frac{\|b - Ax_k\|_2}{\|b\|_2} < tol$  : return  $x_k$ 
7:    $\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$ 
8:    $p_{k+1} = r_{k+1} + \beta_k p_k$ 
9:    $z = Ap_{k+1}$ 
10: end for

```

Dans le cas où le conditionnement (en norme 2) $\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}} \gg 1$ on dit que le système (3) est mal conditionné. Sur un tel système le CG mettra alors plus d'itérations pour atteindre la convergence à une précision donnée. Pour résoudre ce problème, on choisira une matrice M admettant une décomposition de Cholesky sous la forme $M = LL^T$, et telle que $\kappa(M^{-1}A) < \kappa(A)$. L'opérateur $M^{-1}A$ n'étant pas nécessairement SDP le CG ne peut pas s'appliquer sur cet opérateur. Néanmoins [5][chap 9,p 261-265] montre comment obtenir l'algorithme 3 du gradient conjugué préconditionné (PCG) correspondant où les instructions en rouge montrent la différence avec la version CG .

Algorithm 3 PCG

```

1: compute  $z = Ax_0$  and  $r_0 = b - z$ , set  $u_0 = M^{-1}r_0$  and  $p = u_0$ 
2: for  $k = 0, maxiter$  do
3:    $\alpha_k = \frac{u_k^T r_k}{r_k^T z}$ 
4:    $x_{k+1} = x_k + \alpha_k p_k$ 
5:    $r_{k+1} = r_k - \alpha_k z$ 
6:   if  $\frac{\|b - Ax_k\|_2}{\|b\|_2} < tol$  : return  $x_k$ 
7:    $u_{k+1} = M^{-1}r_{k+1}$ 
8:    $\beta_k = \frac{u_{k+1}^T r_{k+1}}{u_k^T r_k}$ 
9:    $p_{k+1} = u_{k+1} + \beta_k p_k$ 
10:   $z = Ap_{k+1}$ 
11: end for

```

2.1.4 Correspondance entre les coefficients du gradient conjugué et l'algorithme de Lanczos

La méthode du gradient conjugué utilise cette correspondance afin d'appliquer la procédure de Lanczos sur pour construire le sous espace de Krylov $K_j(A, r_0)$. A l'itération j du gradient conjugué on a $K_j(A, r_0) = \text{vect}(r_0, r_1, \dots, r_{j-1})$ tel que $\forall j, r_{j+1} \perp K_j(A, r_0)$. Cette propriété est vérifiée par construction de l'espace de recherche via la procédure de Lanczos, on a en effet $\forall j, r_j = c_j v_{j+1}$. Cette section expose les résultats de Yousef Saad sur l'équivalence entre la méthode de Lanczos et le CG. Yousef Saad montre le lien entre les itérations du gradient conjugué et les coefficients de la tridiagonale de Lanczos [5] [Chap 6, p 192–194]. On a $\delta_{j+1} = \frac{v_{j+1}^T w_{j+1}}{\|v_{j+1}\|_2^2} = \frac{v_{j+1}^T A v_{j+1}}{v_{j+1}^T v_{j+1}}$ par définition de $(v_j)_{j=0,m-1}$ et $(w_j)_{j=1,m}$. Comme $r_j = c_j v_{j+1}$ avec $c_j \in \mathbb{R}$, on a l'expression de $\delta_{j+1} = \frac{r_j^T A r_j}{r_j^T r_j}$ en fonction du résidu généré par le CG. Le numérateur n'étant pas calculé par le CG, on utilise la relation entre les vecteurs résidus et les directions de descente suivante :

$$r_j = p_j - \beta_{j-1} p_{j-1} \quad (12)$$

pour ramener l'expression de δ aux coefficients du CG, on a alors $\delta_1 = \frac{1}{\alpha_0}$ et $\forall j > 0, \delta_{j+1} = \frac{1}{\alpha_j} + \frac{\beta_{j-1}}{\alpha_{j-1}}$. Pour η_{j+1} on rappelle que la procédure de Lanczos est basé sur la relation suivante : $\eta_{j+1} v_{j+1} = A v_j - \delta_j v_j - \eta_j v_{j-1}$. Comme $\eta_{j+1} = \|w_j\|_2$, on a $\eta_{j+1} = v_{j+1}^T A v_j = \frac{|r_{j+1}^T A r_j|}{\|r_{j-1}\|_2 \|r_j\|_2}$. En utilisant (12), il vient $\eta_{j+1} = \frac{\sqrt{\beta_{j-1}}}{\alpha_{j-1}}$. Cette correspondance est un point important puisqu'elle permet d'estimer aux cours des itérations du CG le conditionnement du système en calculant le conditionnement de $T_j = \text{tridiag}(\mu_j, \delta_j, \mu_{j+1})$. La base V_j étant alors $[\frac{r_0}{\|r_0\|_2^2} | \frac{r_1}{\|r_1\|_2^2} | \dots | \frac{r_j}{\|r_j\|_2^2}]$.

2.1.5 Gradient conjugué déflaté

L'objectif de cette méthode est d'accélérer la convergence du CG en construisant un espace de recherche imposant aux vecteurs résidus générés par le CG une contrainte de Petrov-Galerkin plus forte. Concrètement cela revient à retirer des directions de descentes les composantes associées aux vecteurs propres de la matrice du système. Pour ce faire on adjoint au sous-espace de Krylov $K_m(A, r_0)$ un espace de déflation $W = \text{vect}(w_1, \dots, w_k)$ généré par k vecteurs propres de A tel que l'espace de recherche à l'itération j du CG soit $K_{kj}(A, W, r_0) = \text{vect}(W, V_j)$ avec $V_j = \text{vect}(v_1, \dots, v_j)$ le sous espace généré par la procédure de Lanczos. Néanmoins pour que $x_j \in x_0 + K_{kj}(A, W, r_0)$ et que $r_j \perp K_{kj}(A, W, r_0)$ la procédure de Lanczos ne doit plus être appliquée à A mais à une matrice auxiliaire $B = A - AW(W^T AW)^{-1}W^T A$ assurant alors l'orthogonalité de V_j par rapport à W [14]. Mais en observant que $v_1 = \frac{r_0}{\|r_0\|_2}$ rien assure que r_0 soit orthogonale à W . Pour ce faire [14] propose de choisir arbitrairement x_{-1}, r_{-1} tel que $x_0 = x_{-1} + W(W^T AW)^{-1}W^T r_{-1}$ impliquant alors $W^T r_0 = 0$. La démonstration se fait en faisant apparaître la matrice auxiliaire B dans l'expression de r_0 et en rappelant que $W^T B = 0$. L'algorithme (4) du gradient conjugué préconditionné déflaté est donné par [14].

En posant $M = I$ on a la version non préconditionné CG-DEF. En rouge on met en évidence les différences avec la version PCG (3). Les différences sont dues à l'application de Lanczos à la matrice auxiliaire B , pour $x_j \in x_0 + K_{k,j}(A, W, r_0)$, $x_j = x_0 + W\xi_j + V\eta_j$ avec $\xi_j \in \mathbb{R}_k$ et $\eta_j \in \mathbb{R}_j$. On montre que $W^T r_j = 0$ si et seulement si $\xi_j = -\Delta_j \eta_j$ avec $\Delta_j = (W^T AW)^{-1}W^T AV_j$.

2.2 Comparaison de la convergence du CG et du CG-DEF

Afin de comprendre mieux l'intérêt de la méthode déflatée, on expose les résultats de convergence des deux méthodes. En notant $e_j = x_j - x^*$ l'erreur directe à l'itération j on définit la norme A de l'erreur directe par $\|e_j\|_A = e_j^T Ae_j$. On notera κ l'opérateur de conditionnement

Algorithm 4 PCG-DEF

```

1: Choose  $k$  linearly independent vector  $w_1, \dots, w_k$ . Set  $W = [w_1 |, \dots, | w_k]$ .
2: Choose  $x_0$  such that  $W^T r_0 = 0$ . Compute  $u_0 = M^{-1}r_0$ .
3: Solve  $W^T A W \mu_0 = W^T A r_0$ . Set  $p_0 = -W\mu_0 + u_0$ .
4: for  $j = 0, maxiter$  do
5:    $\alpha_j = \frac{u_j^T r_j}{r_j^T z}$ 
6:    $x_{j+1} = x_j + \alpha_j p_j$ 
7:    $r_{j+1} = r_j - \alpha_j z$ 
8:   if  $\frac{\|b - Ax_j\|_2}{\|b\|_2} < tol$  : return  $x_j$ 
9:    $u_{j+1} = M^{-1}r_{j+1}$ 
10:   $\beta_j = \frac{u_{j+1}^T r_{j+1}}{u_j^T r_j}$ 
11:  Solve  $W^T A W \mu_j = W^T A u_j$ 
12:   $p_{j+1} = u_{j+1} + \beta_j p_j - W\mu_j$ .
13:   $z = Ap_{j+1}$ 
14: end for

```

en norme 2. Pour la méthode du *CG* on a le résultat de convergence [28] suivant :

$$\|e\|_A \leq 2 \left(\frac{(\kappa(A) - 1)^{1/2}}{(\kappa(A) + 1)^{1/2}} \right)^j \|x_o - x_j\|_A \quad (13)$$

Pour la variante déflatée, on note $H = I - W(W^T A W)^{-1} A W^T$ l'opérateur de projection A-orthogonal sur $W^{\perp A}$. Le résultat de convergence [14] [chap 5] est alors :

$$\|e\|_A \leq 2 \left(\frac{(\kappa(H^T A H) - 1)^{1/2}}{(\kappa(H^T A H) + 1)^{1/2}} \right)^j \|x_o - x_j\|_A \quad (14)$$

On suppose que les valeurs propres de A sont telles que : $\lambda_1 \leq \dots \leq \lambda_n$, et en prenant W comme la base des k vecteurs propres de A associés aux k plus petites valeurs propres de A, on a le résultat suivant :

$$\kappa(H^T A H) = \frac{\lambda_n}{\lambda_{k+1}} \leq \frac{\lambda_n}{\lambda_1} \quad (15)$$

On en déduit que le *CG-DEF* minimise plus rapidement la norme A de l'erreur directe que le *CG*. En pratique lorsque W est construit avec les approximations des k vecteurs propres de A le conditionnement du système déflaté est $\kappa(H^T A H) \approx \frac{\lambda_n}{\lambda_{k+1}}$.

L'algorithme (4) est défini lorsque l'on dispose d'une base W pouvant soit être les vecteurs propres extacts de A soit une approximation de ces derniers. La partie suivante définit les cas tests numériques permettant d'étudier la convergence du *CG-DEF* suivant ces deux cas. Une version de cette algorithme sera donnée pour le cas de la résolution successive de plusieurs seconds membres et les mêmes cas tests seront utilisés.

2.3 Objectif du PFE

L'objectif du PFE est de caractériser une version du *CG-DEF* assurant que la résolution successive de plusieurs seconds membres à opérateur linéaire fixe soit compétitive face à la résolution via le *CG*.

En notant ν le nombre de seconds membres et $s = 1, \dots, \nu$ le numéro du système, on définit un système à plusieurs seconds membres par l'équation suivante :

$$Ax^s = b^s, \quad s = 1, 2, \dots, \nu \quad (16)$$

On peut citer la résolution par différences finies de l'équation de la chaleur avec un schéma en temps implicite comme source de problèmes de la forme (16), la solution au temps t_n étant dépendante du second membre au temps t_{n-1} . Comme expliqué précédemment la résolution d'un tel système par le *CG-DEF* permet en théorie de réduire le conditionnement du système s et donc de réduire le nombre d'itérations à convergence comparativement à une résolution par le *CG*. Comme le montre l'équation (15) le conditionnement du système déflaté est lié à W . Afin de vérifier (15) plusieurs expérimentation numériques seront menées. On répondra par la même aux questions suivantes :

Q1/ Quelles partie du spectre de A doit être déflatée afin de réduire le nombre d'itérations à convergence ? Pour répondre à cette question, on fixera $\nu = 2$ tel que pour $s = 1$ le système est résolu par le *CG*, et pour $s = 2$ via le *CG-DEF*. On comparera le nombre d'itérations à convergence des deux méthodes en utilisant pour W , $k \in [|1, n|]$ vecteurs propres exacts de A . Cette étude est conduite dans les Sections 3.1.1, 3.1.2 pour des clusters de taille 1, et en 3.1.4 pour des clusters de taille 4. Pour se faire on générera des matrices cas tests avec une distribution spectrale variée. La Section 2.4 précise la procédure permettant de construire les matrices des cas tests.

Q2/ Les vecteurs propres extrêmes du spectre de A jouent-ils un rôle similaire dans la déflation ? Comme le résidu initiale doit vérifier $W^T r_0 = 0$ et que la direction de descente correspondante doit être A -orthogonale à W , l'effet de la déflation semble à priori varier suivant que l'on déflatte la partie basse ou haute du spectre de A . En se basant sur la question précédente on vérifiera l'effet de la déflation suivant la partie du spectre déflatée. Les Sections 3.1.1 à 3.1.4 font l'analyse des expérimentations répondant à cette question.

Q3/ La déflation permet-elle davantage que la réduction du conditionnement entre plusieurs résolutions successives ? Au vue des contraintes appliquées aux coefficients du *CG-DEF*, on s'attend à ce que l'effet de la déflation soit plus que celui d'un préconditionnement. Les expériences de la Section 3.1.4 utilisant des opérateurs linéaires avec des clusters de taille supérieurs à 1 permettront d'éclaircir ce point.

Q4/ Quel niveau de précision est requis sur le calcul des vecteurs propres pour observer l'effet de la déflation ? Lorsque l'on ne dispose pas des vecteurs propres exacts, la méthode du *CG-DEF* permet de construire entre chaque résolution successive une approximations de ces vecteurs. On souhaite savoir a priori à quel niveau de précision ces approximations doivent être calculées pour observer l'effet de la déflation. Pour se faire on observera les résultats de convergence du *CG-DEF* lorsque l'on utilise des vecteurs propres perturbés pour former W . Cette expérimentation est développée dans la Section 3.2.

Q5/ Quelles méthodes utiliser pour calculer W ? Deux variantes de la méthode de Lanczos permettent de calculer une approximation des vecteurs propres de A au cours des itérations du *CG-DEF*. D'une part la méthode Lanczos Rayleigh-Ritz qui utilise directement la tridiagonale de Lanczos T_j tel que $W = V_j Y_j$ avec Y_j les vecteurs de Ritz. D'autre part la méthode Lanczos Harmonic-Ritz présentée dans la Section 4.2 qui nécessite de résoudre un problème de la forme (7). Afin de mesurer la qualité des approximations calculées, on comparera le sinus de l'angle entre les vecteurs propres exacts de A , notés u , et les vecteurs w calculés par les deux variantes. De même pour le quotient de Rayleigh $\rho(w) = \rho(w, A, B)$ et le résidu de la paire propre $\|Aw - \rho(w)w\|_2$. L'objectif est d'obtenir une méthode qui minimise le sinus entre ces vecteurs. [3], [16] ont développé de nouvelles variantes permettant d'enrichir au cours d'une même résolution la qualité des vecteurs propres calculés. On abordera cette question dans un premier temps via les sections 4.1.2 et 4.2.2, puis par les Sections 5.1 et 5.2, traitant le calcul

de W dans la méthode déflatée respectivement en se basant sur l'approche Rayleigh-Ritz, et l'approche Harmonic-Ritz.

Q6/ Etude du CG-DEF avec calcul de W via Harmonic-Ritz pour deux résolutions successives. Pour calculer une approximation des vecteurs propres via Lanczos Harmonic-Ritz ou Rayleigh-Ritz, certains coefficients du *CG-DEF* doivent être sauvegardés au cours des itérations afin de pouvoir former le problème aux valeurs propres correspondant. On appelle *maxiter* le nombre d'itérations maximal pour converger à une précision donnée. La méthode consiste à résoudre le système ($s=1$) avec le *CG* en considérant que $W^{(s=1)} = \emptyset$. Lors de cette première résolution, on conserve les $\ell \in [|0, \text{maxiter}|]$ premières données que sont les pas de descente $\alpha_\ell^{(s)} = \{\alpha_0^{(s)}, \dots, \alpha_{\ell-1}^{(s)}\}$ et $\beta_\ell^{(s)} = \{\beta_0^{(s)}, \dots, \beta_{\ell-1}^{(s)}\}$, la matrice $P_\ell^{(s)} = [p_0^{(s)}, \dots, p_{\ell-1}^{(s)}]$ des ℓ premières directions de descente ainsi que le vecteur $d_\ell^{(s)} = \{d_0^{(s)}, \dots, d_{\ell-1}^{(s)}\}$ avec $(d_i^{(s)})_{i=0,\ell-1} = p_i^T A p_i$. La méthode Lanczos Rayleigh-Ritz ne nécessite que $\alpha_\ell^{(s)}$ et $\beta_\ell^{(s)}$ pour construire T_ℓ comme expliqué en 2.1.4. L'objectif de cette étude est de faire varier ℓ , le niveau d'informations à conserver lors de la première résolution, afin d'étudier la convergence du *CG-DEF* lors de la seconde résolution. Cette étude permettra de déterminer le volume d'information à conserver suivant la partie du spectre que l'on souhaite approcher, elle est menée dans la Section 5.1 pour Rayleigh-Ritz et en 5.2 pour Harmonic-Ritz.

Q7/ Application à ν résolutions successives Lorsque $s > 2$ la méthode du *CG-DEF* permet de raffiner W entre chaque résolution de telle sorte que $W^{s+1} = [W^s | P_\ell^s] Y^s$ avec Y^s la matrice dont les colonnes sont les vecteurs propres du problème construit par Lanczos Rayleigh-Ritz ou Lanczos Harmonic-Ritz. On étudiera la convergence de *CG-DEF* dans le cas où $\nu > 2$ afin d'observer une réduction du nombre d'itérations à convergence après chaque résolution. Cette étude est à retrouver au sein de la Section 6.

2.4 Environnement expérimental

On souhaite ici étudier la convergence du *CG-DEF* appliquée à différents systèmes linéaires impliquant des matrices ayant une distribution spectrale variée. On entend par cela que les matrices ont une répartition des paires propres par clusters. On réfère ainsi la plus petite valeur propre au terme **Least Dominant** (*LD*) et la plus grande au terme **Most Dominant** (*MD*), tels qu'un cluster dont les valeurs propres sont les plus petites ou les plus grandes en module est appelé respectivement **Least Dominant Cluster** (*LDC*) et **Most Dominant Cluster** (*MDC*) [3]. On réfère un cluster central à la notation (*CC*). Cette étude doit permettre de déterminer les vecteurs propres devant être déflatés afin d'accélérer la convergence de la méthode suivant la répartition spectrale de la matrice. On définit au préalable les cas tests et notations utilisés lors des expérimentations :

Construction des matrices cas tests. On définit les matrices D_0 , D_1 ayant un cluster (*CC*) de dimension $N \times N$ tel que

- D_0 : matrice diagonale avec N éléments pris aléatoirement dans l'intervalle $[|1,N|]$ sans remise
- D_1 : matrice diagonale avec N éléments pris aléatoirement dans l'intervalle $[0.5,1.5]$
- et les matrices D_2 , D_3 , D_4 de dimension $N \times N$ ayant des clusters *LDC* et *MDC* de taille k distants de θ du cluster central de I_1 tels que :
 - D_2 : k éléments pris dans $(1.0/\theta) * [1/2, 3/2]$
 - D_3 : k éléments pris dans $\theta * [1/2, 3/2]$
 - D_4 : combine les clusters *LDC* de D_2 et *MDC* de D_3

On construit les matrices A_i des cas tests telles que $A_i = Q^T D_i Q$ avec $i \in [|0, 4|]$ et Q matrice unitaire issue de la décomposition *QR* d'une matrice carré aléatoire. Ainsi le spectre de A_i est

donné par les éléments diagonaux de D_i et les vecteurs propres sont les vecteurs donnés par les colonnes de Q^t .

Paramètres du *CG-DEF*. Le *CG-DEF* prenant comme paramètres le nombre de vecteurs à déflater k , ainsi que le volume d'information à conserver ℓ entre deux résolutions, on définit les notations suivantes :

- S0 : résolution via le CG, $\ell = 0$ et $k = 0$
- S1 : $\ell = \text{maxiter}$ et $k = \text{maxiter}$
- S2 : $\ell = \text{maxiter}$ et $k = 1, 2, \dots, \text{maxiter}$
- S3 : $\ell \subset [|0, \text{maxiter}|]$ et $k = 1, 2, \dots, \text{card}(\ell)$

où maxiter est le nombre d'itérations à convergence du solveur lors de la résolution précédente. On utilise les notations *LD*, *MD* et *LMD* pour préciser la partie du spectre à laquelle les k vecteurs déflatés appartiennent.

3 Injection spectrale donnée a priori et effet sur la convergence du *CG-DEF*

Cette partie s'attache à décrire numériquement le phénomène de déflation sur la convergence du gradient conjugué lorsque l'espace de déflation contient les vecteurs propres exacts. Comme ces vecteurs sont généralement des approximations, on réalisera une seconde étude en ajoutant une perturbation sur la base de déflation. On étudiera la convergence du *CG-DEF* en fonction de la qualité spectrale de la base.

3.1 Déflation avec des vecteurs propres exacts

On étudie ici l'effet de la déflation sur la convergence suivant l'erreur inverse. Le critère d'arrêt du solveur est basé sur le résidu itéré tel que $\frac{\|r\|_2}{\|b\|_2} < 1e - 16$. Les figures suivantes présentent l'erreur inverse en fonction des itérations du solveur. Les labels en haut des graphiques indiquent la procédure utilisée pour la déflation comme expliqué dans la Section 2.2. Les labels à droite font référence à la distance θ entre un cluster central et un cluster *LDC* ou *MDC*. Les sections 3.1.1 et 3.1.2 permettent de savoir quelle partie du spectre de l'opérateur linéaire doit être déflatée, question Q1, afin de réduire le nombre d'itérations à convergence entre deux résolutions successives. Ces sections répondent en plus à la question Q2 de la section 2.3 en exposant des résultats de convergence différents suivant la partie du spectre déflatée et la distribution spectrale de la matrice.

3.1.1 Matrices avec cluster central

Lorsque les matrices n'ont pas de *LDC* ou de *MDC*, il faut déflater avec un grand nombre de vecteurs propres afin de réduire le nombre d'itérations à convergence (figure 1a et 1b S2-LD). Lorsque l'espace de déflation utilise toute l'information spectrale la méthode converge à l'itération 0 (figures 1a, 1b S2-LD). On constate aussi que déflater par rapport aux 4 premiers vecteurs propres de *LD* permet de réduire le nombre d'itérations (figures 1a S3-LD) contrairement à une déflation par rapport aux 4 derniers vecteurs propres de *MD* (figures 1a S3-MD).

Lorsque les matrices n'ont pas de cluster *LDC* ou *MDC*, déflater les vecteurs propres proches de *LD* permet d'accélérer la convergence et l'on vérifie ainsi (7). Voyons maintenant en section 3.1.2 le cas des clusters de taille 1.

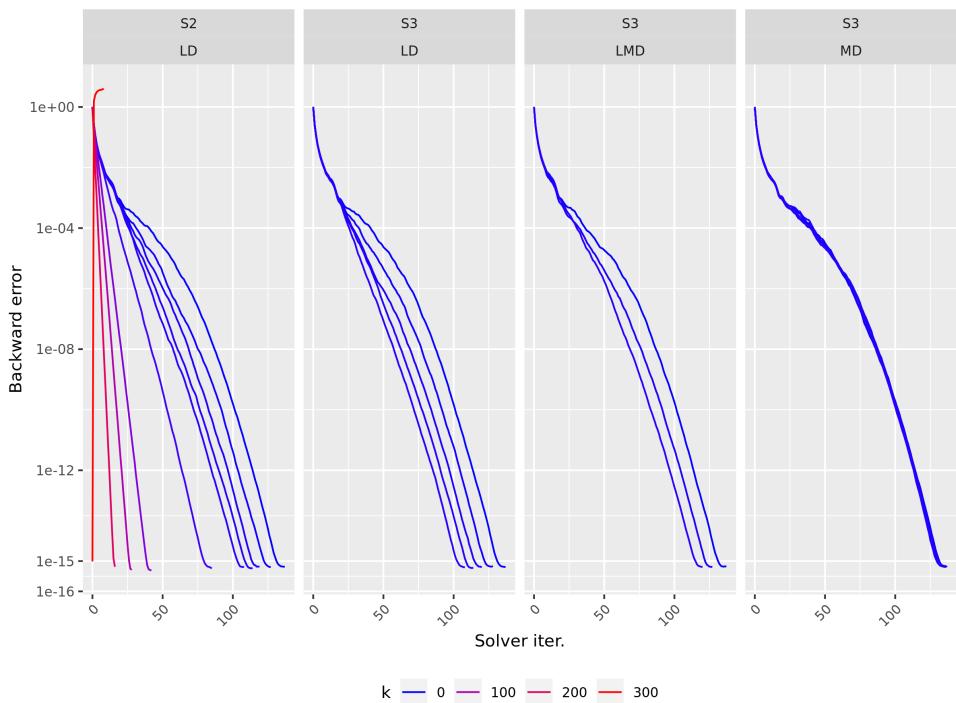
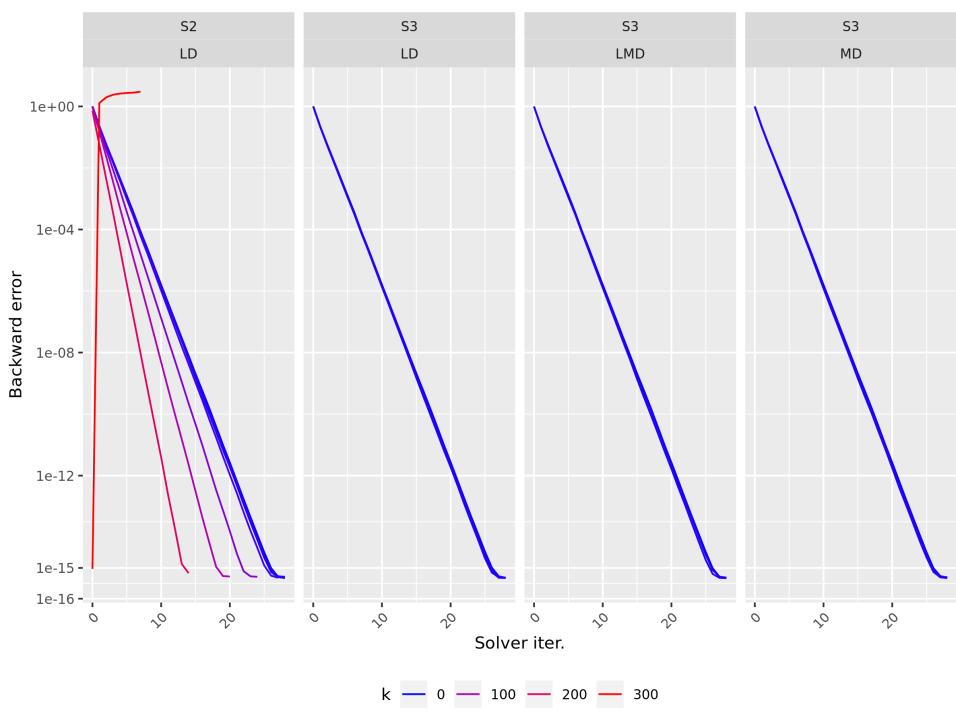
(a) Matrice A_0 , dimension 300×300 (b) Matrice A_1 , dimension 300×300

FIGURE 1 – Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater.

3.1.2 Matrices avec clusters *LDC* et *MDC* de taille 1

Matrices avec un unique cluster. En présence d'un cluster, déflater avec le vecteur propre correspondant au cluster permet de réduire significativement le nombre d'itérations à convergence (figures 2a S3-LD, 2b S3-MD). Se faisant on retrouve une convergence linéaire similaire à celle de la matrice A_1 pour $k = 0$ (figure 1b). Cela permet aussi de valider l'inégalité (15), puisque en déflatant l'unique vecteur propre pénalisant la convergence, l'on réduit significativement le nombre d'itérations à convergence par rapport au premier système.

Matrices avec les deux clusters *LDC* et *MDC*. Lorsque l'on déflate par rapport à *LD* ou à *MD* on a une nette diminution du nombre d'itérations à convergence tant que l'espace de déflation comprend le vecteur propre correspondant au cluster *LDC* ou *MDC* (figures 3 S3-LD et S3-MD) mais la convergence n'est pas linéaire. Par contre lorsque l'on déflate par rapport aux deux parties du spectre, la convergence est bien linéaire (figure 3 S3-LMD).

Ainsi la matrice de déflation doit contenir l'ensemble des vecteurs propres extrêmes qu'ils soient de *LDC* ou de *MDC*. L'effet de la déflation dépend donc de la répartition du spectre de la matrice. On a ainsi donné des éléments de réponse aux questions Q1 et Q2.

3.1.3 Approximation de rang faible

Cette section montre de nouveau que l'effet de la déflation varie suivant le spectre de A et la partie déflaté (Q1 et Q2), et d'autre part, que la déflation apporte plus qu'une réduction du préconditionnement comme le suggérait la question Q3. En effet, pour les matrices possédant un cluster *MDC* (A_3, A_4) la déflation du vecteur propre de *MDC* permet d'obtenir à l'itération 0 une approximation de la solution à une précision plus élevée que celle du *CG* (figures 4a et 4b S3-MD). On constate que plus θ est grand plus la solution calculée à l'itération 0 est précise. Comme à l'itération 0 du *CG-DEF* on a $x_0 = x_{-1} + W(W^T AW)^{-1}W^T r_{-1}$, l'erreur inverse correspondante est :

$$\frac{\|b - Ax_0\|_2}{\|b\|_2} = \frac{\|r_{-1} - AW(W^T AW)^{-1}W^T r_{-1}\|_2}{\|b\|_2} \quad (17)$$

On retrouve $I - AW(W^T AW)^{-1}W^T = H^T$ qui est la matrice du la projection $A^{-1} - orthogonal$ sur W^\perp . Comme A est SDP sa décomposition en valeurs propres est équivalent à sa décomposition en valeur spectrale et l'on a $(W^T AW) = \Lambda$ lorsque $rank(W) = n$. Si l'on déflate par rapport à $k \leq n$ vecteurs propres, il vient [22][chap 5, p35] l'approximation A_k de A par des matrices de rang 1 : $A_k = \sum_{j=\nu_1}^{\nu_2} \lambda_j w_j w_j^T$ avec $1 \leq \nu_1 \leq \nu_2 \leq n$, $\nu_2 - \nu_1 + 1 = k$. tel que $\|A - A_k\|_F = \inf(B \in \mathbb{M}_{n \times n}, rank(B) \leq k) \|A - B\|_F = (\lambda_1 + \dots + \lambda_{\nu_1-1} + \lambda_{\nu_2+1} + \dots + \lambda_n)^{\frac{1}{2}}$. Comme $(W^T AW)^{-1} = \Lambda^{-1}$, on a $W(W^T AW)^{-1}W^T = A_k^{-1} = \sum_{j=\nu_1}^{\nu_2} \frac{1}{\lambda_j} w_j w_j^T$ l'approximation de rang faible de l'inverse de A . Ainsi pour $rank(W) = k$, $\|r_{-1} - AW(W^T AW)^{-1}W^T r_{-1}\|_2 = \|r_{-1} - AA_k^{-1}r_{-1}\|_2$ qui tend vers 0 lorsque k tend vers n . Cette approximation de rang faible de l'inverse de A permet d'expliquer que l'erreur inverse du *CG-DEF* soit inférieure à celle du *CG* à l'itération 0 notamment lorsque l'on déflat de *MD* vers *LD*, les vecteurs propres de *MDC* génèrent des matrices de rang 1 approchant au mieux A^{-1} et lorsque k est proche de n .

On observe ainsi que le *CG-DEF* est une méthode de choix lorsque la matrice posséde un cluster *MDC*, et que dans ce cas les vecteurs propres de *LD* et de *MD* n'ont pas le même effet sur la déflation.

Toujours afin de répondre aux question Q1 à Q3 de la Section 2.3, on se place pour la suite dans le cas où les clusters sont de tailles 4.

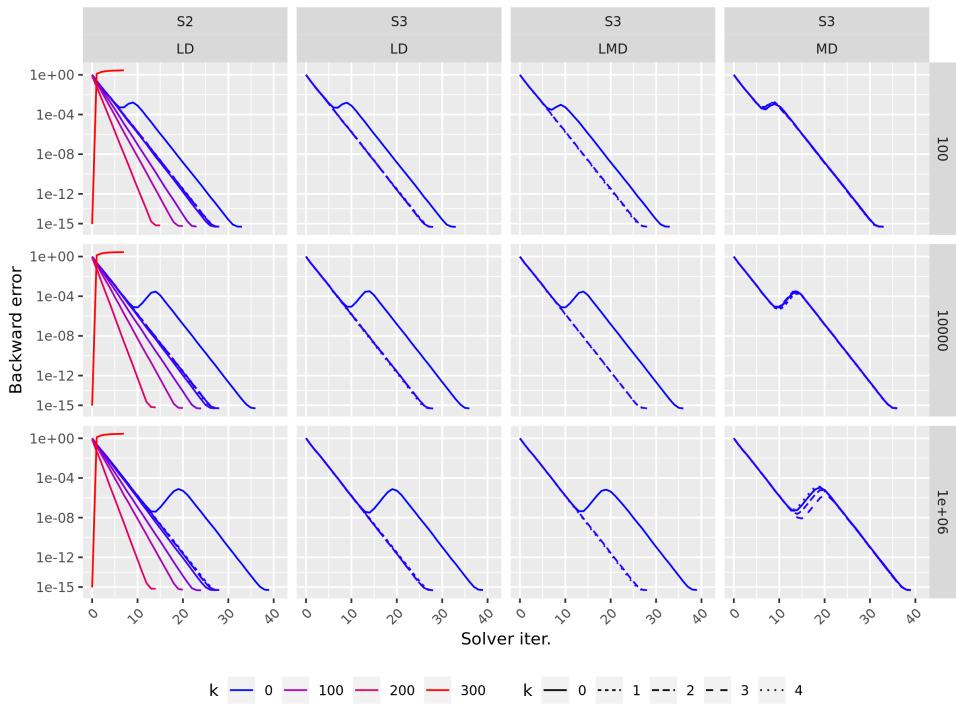
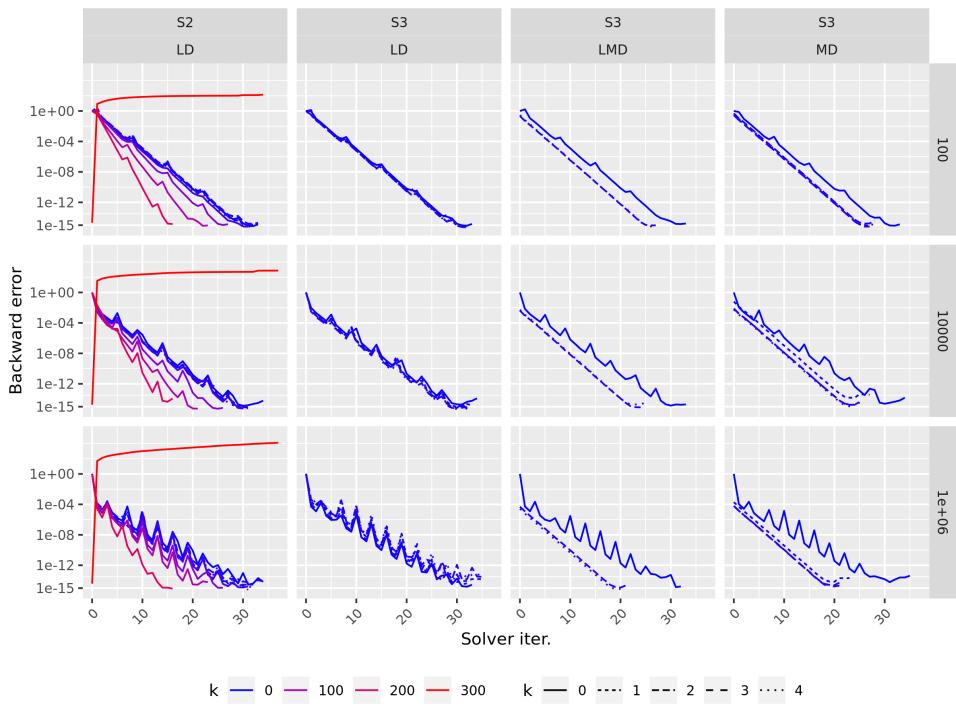
(a) Matrice A_2 , dimension 300×300 , cluster de taille 1.(b) Matrice A_3 , dimension 300×300 , cluster de taille 1.

FIGURE 2 – Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater. LDC et MDC sont de taille 1.

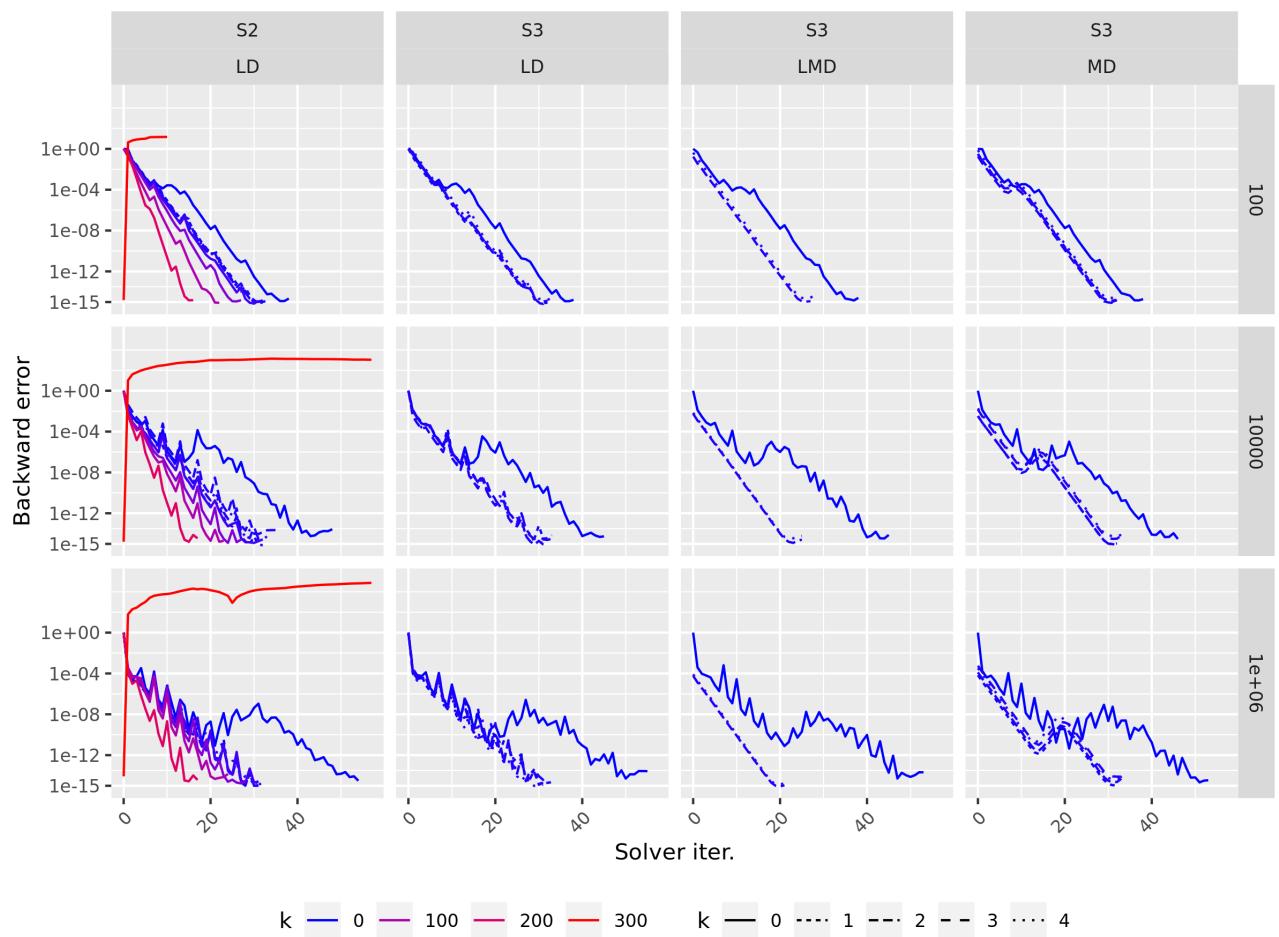


FIGURE 3 – Evolution de l'erreur inverse pour la Matrice A_4 , dimension 300×300 , *LDC* et *MDC* de taille 1.

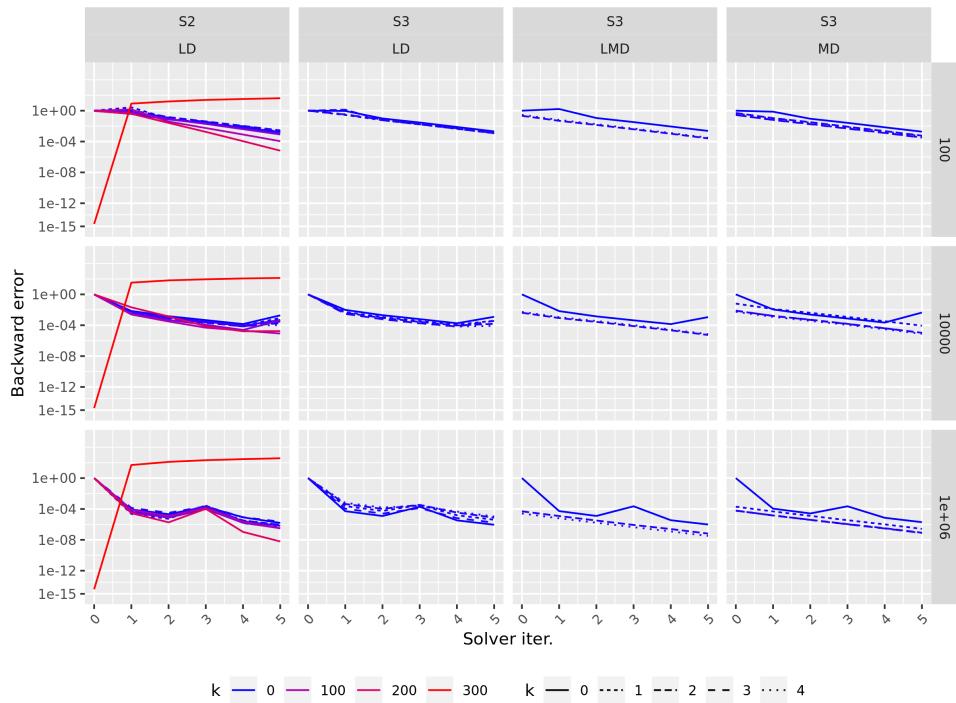
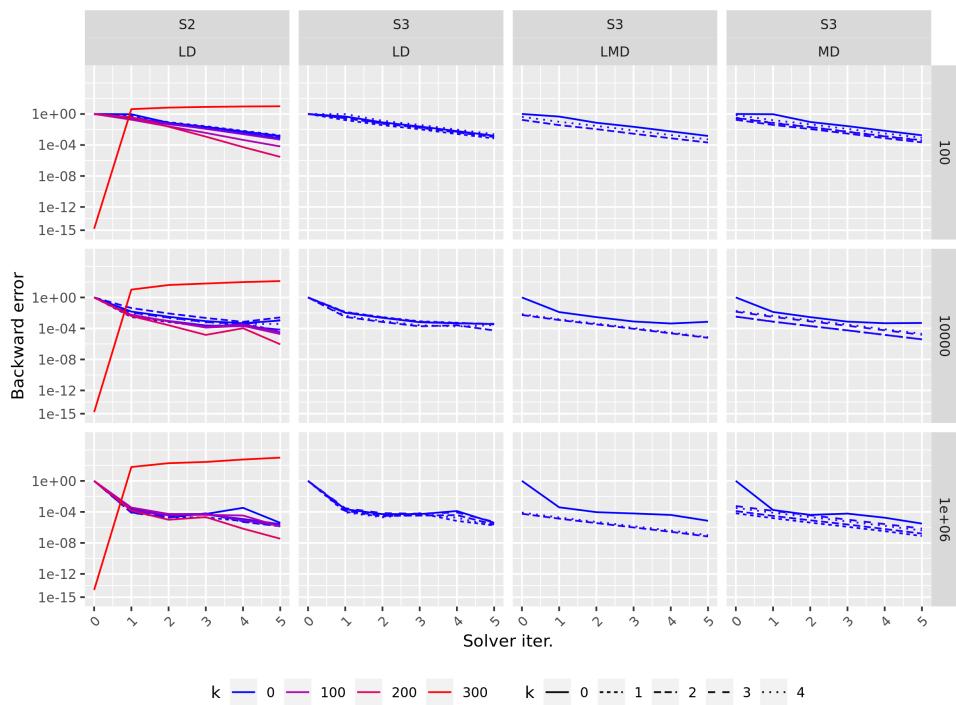
(a) Matrice A_3 , dimension 300×300 , cluster de taille 1.(b) Matrice A_4 , dimension 300×300 , cluster de taille 1.

FIGURE 4 – Zoom sur les premières itérations, effet de la déflation de MDC sur la précision de la solution à l'itération 0, cluster de taille 1.

3.1.4 Matrices avec clusters LDC et MDC de taille 4

Lorsque les clusters LDC et MDC ont plusieurs paires propres, déflater une partie des vecteurs propres correspondant permet de réduire le nombre d'itérations comme on le constate sur la figure 5a S3-LD avec $k = 1, 2, 3$. Comme par construction les valeurs propres d'un cluster ont le même ordre de grandeur, l'accélération de la convergence observée ici n'est pas due à la priorité (15). Il y a là un effet sur l'espace de recherche de la solution, en rappelant que les directions de descente du $CG\text{-}DEF$ sont A-orthogonales à W , déflater une partie du cluster permet d'éliminer de l'espace de recherche des directions de descentes non optimales. On a ici une nouvelle réponse à la question Q4 de la section 2.3.

Mais pour que la convergence soit linéaire, il faut déflater l'ensemble des vecteurs propres du cluster (figures 5a S3-LD et 5b S3-MD , $k = 4$). L'effet de l'approximation de faible rang sur la précision du résidu à l'itération 0 n'est visible que lorsque W contient tous les vecteurs du cluster MDC .

En résumé, lorsque l'on dispose de l'espace généré par les vecteurs propres de la matrice du système linéaire à résoudre, l'espace de déflation doit être construit avec les vecteurs propres correspondants aux clusters LDC et/ou MDC (Q1, Q3) afin de réduire significativement le nombre d'itérations nécessaires pour converger. Le $CG\text{-}DEF$ est plus performant lorsque la matrice a un mauvais conditionnement (Q2) (comparaison des figures pour les matrices A_0 et A_1 et des courbes de convergence en fonction de θ). Cette méthode permet lorsque l'opérateur linéaire possède un cluster MDC d'obtenir une approximation de la solution à une meilleure précision lors des premières itérations comparativement au CG (Q3). Maintenant que l'on sait répondre aux questions Q1, Q2 et Q3 portant sur le choix et l'effet de l'espace de déflation, on peut s'interroger sur la précision à laquelle la matrice de déflation doit être calculée afin de pouvoir réduire au mieux le nombre d'itérations à convergence. On étudiera dans la Section suivante la question Q4.

3.2 Déflation sur un espace perturbé

En général on ne dispose pas des paires propres exactes d'une matrice mais d'une approximation de ces dernières. On s'attache ainsi dans cette partie à étudier le $CG\text{-}DEF$ lorsque l'on perturbe la base de déflation exacte. On étudie ici la convergence de l'erreur inverse en fonction des itérations du $CG\text{-}DEF$ (Q4). Les matrices utilisées sont A_2 et A_3 avec les clusters respectivement LDC et MDC de taille 1 . Les labels en haut des courbes sont les valeurs de θ et la légende représente la valeur du sinus de l'angle (noté $\sin<(w,u)$) entre le vecteur propre du cluster u et le vecteur perturbé w tel que :

$$w = \frac{u + \alpha \frac{v}{\|v\|_2}}{\|u + \alpha \frac{v}{\|v\|_2}\|_2}$$

avec $\alpha \in [10^{-6}, 1]$ et $v \in \mathbb{R}^n$ dont les éléments sont pris aléatoirement entre 0 et 1. On a alors :

$$\begin{aligned} \text{Cos } < (u, w) &= \frac{w^T u}{\|u\|_2 \|w\|_2} \\ \implies \text{Sin } < (u, w) &= (1 - \text{Cos } < (u, w) * \text{Cos } < (u, w))^{\frac{1}{2}} \end{aligned}$$

Ainsi pour $\alpha \rightarrow 0$ on a ($w \rightarrow u$ et $\text{Sin } < (w, u) \rightarrow 0$).

Matrice A_2 . On constate (figure 6) que plus θ est grand plus le sinus de l'angle entre les deux vecteurs doit être faible afin de déflater le vecteur propre du cluster. Pour $\theta = 10^2$ on peut se satisfaire d'un sinus de 10^{-3} , pour $\theta = 10^6$, on doit générer une approximation du vecteur propre tel que $\sin<(w,u) \leq 10^{-6}$.

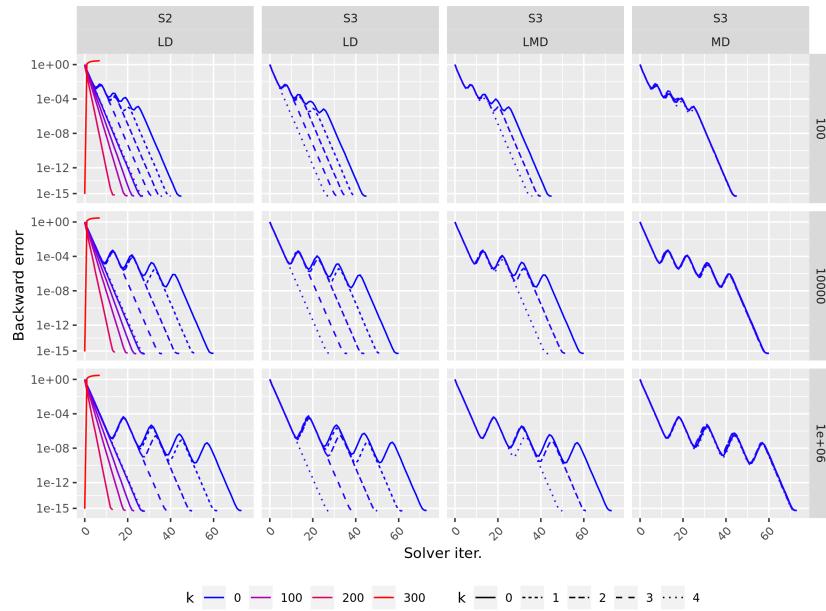
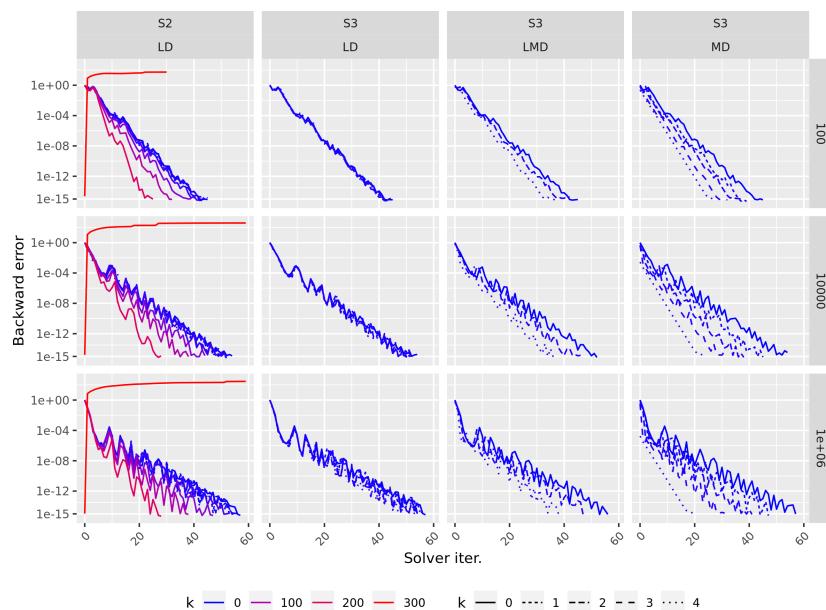
(a) Matrice A_2 , dimension 300×300 , cluster de taille 4.(b) Matrice A_3 , dimension 300×300 , cluster de taille 4.

FIGURE 5 – Influence de la déflation sur la convergence en erreur inverse. k étant le nombre de vecteurs propres utilisés pour déflater. LDC et MDC sont de taille 4.

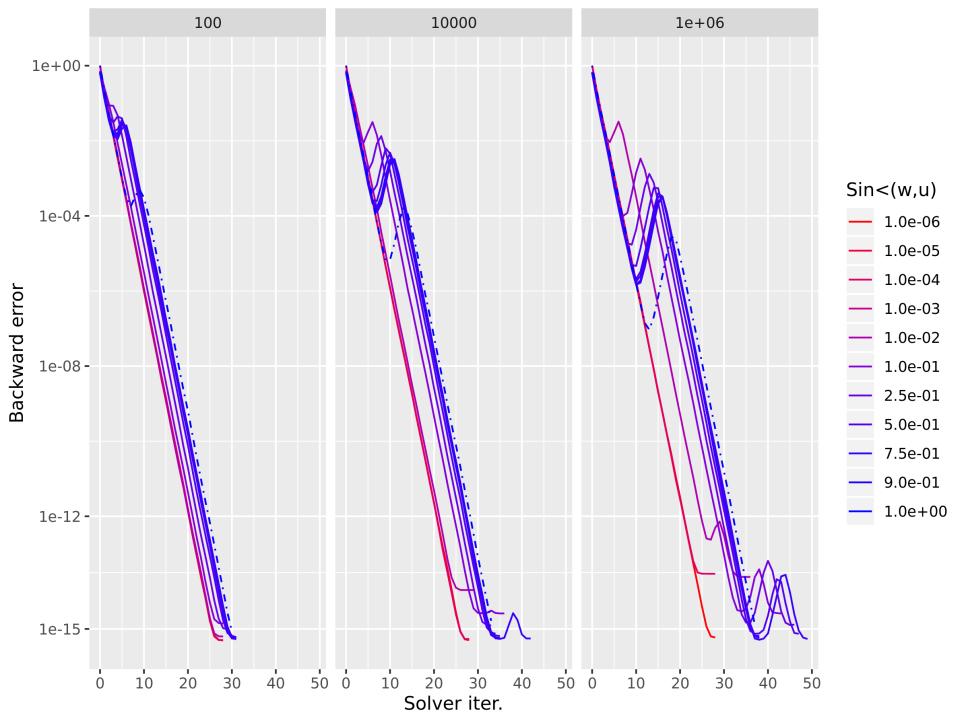


FIGURE 6 – Matrice A_2 , cluster LDC de taille 1 : évolution de l’erreur inverse en fonction du sinus de l’angle entre le vecteur propre exact u et le vecteur perturbé w du cluster.

Matrice A_3 . Même lorsque le $\sin\angle(w,u)$ est proche de 1, le *CG-DEF* déflate correctement le vecteur propre du cluster (figures 7). On voit aussi que l’effet de l’approximation de rang faible est conservée.

On a montré pour des matrices avec LDC que plus θ est grand plus l’approximation du vecteur propre doit être précise. Pour un MDC il semble que cette approximation peut être calculée à une faible précision. Connaissant maintenant une précision minimale sur le critère du sinus de l’angle entre w et u , on peut dans la section 4 comparer deux procédures permettant d’approcher les paires propres d’une matrice. On répondra ainsi à la question Q5.

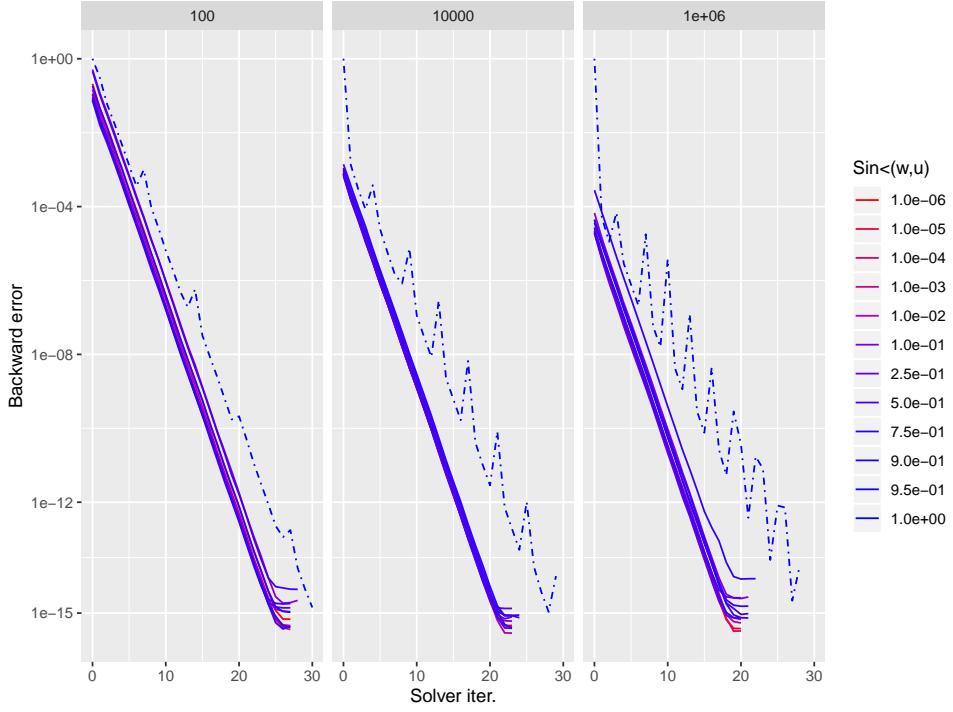


FIGURE 7 – Matrice A_3 , cluster MDC de taille 1 : évolution de l’erreur inverse en fonction du sinus de l’angle entre le vecteur propre exact u et le vecteur perturbé w du cluster.

4 Convergences des paires de Ritz/Harmonic Ritz vers les paires propres

On compare ici différentes procédures pour le calcul des paires propres afin de répondre à la question Q5 de la Section 2.3, l’objectif étant de déterminer les méthodes permettant de générer des approximations des vecteurs propres de bonne qualité en un nombre d’itérations faible devant la dimension du problème aux valeurs propres correspondant. Se faisant, on implantera les méthodes dans le *CG-DEF* afin de pouvoir calculer la matrice de déflation W comme expliqué en section 2.3.

4.1 Etude avec les approximations de Rayleigh-Ritz

La méthode de Lanczos Rayleigh-Ritz (RR) permet de calculer de manière itérative les valeurs et vecteurs propres de A . Cette méthode permettant de générer une bonne approximation des paires propres de A correspondantes aux parties extrêmes de son spectre [26][chap 13, p 289 & chap 12, p 274], nous étudions numériquement ses propriétés de convergence.

4.1.1 Définition

On définit ici la méthode de Lanczos Rayleigh-Ritz (RR) comme étant la projection du problème (6) sur l’espace $R(V)$ engendré par les colonnes de V_m construit par la méthode de Lanczos. On cherche alors (Λ, X) tel que :

$$AX = X\Lambda, \quad (18)$$

$$(AX - X\Lambda) \perp V_m \quad (19)$$

La condition d’orthogonalité permet de réécrire le problème (18) tel que :

$$V_m^T(AX - X\Lambda) = 0 \iff V_m^TAX - V_m^TX\Lambda = 0,$$

$$\text{si } X \in R(V) \text{ alors } V_m^T A V_m \hat{X} - V_m^T V_m \hat{X} \Lambda = 0$$

Comme la procédure de Lanczos génère $T_m = V_m^T A V_m$ tel que $V_m^T V_m = I_m$, on en déduit que :

$$T_m \hat{X} = \hat{X} \Lambda$$

On montre donc que si (Λ, X) est paire propre de A alors (Λ, \hat{X}) est paire propre de (T_m) et $X = V_m \hat{X}$. La méthode (RR) consiste à calculer itérativement les paires paires propres (Λ, \hat{X}) de T_m puis à calculer les paires de Ritz $(\Lambda, V_m \hat{X})$ [22][chap 6, p 278] approximant les pairs propres de A .

4.1.2 Expérimentations

La dimension des matrices est de 100×100 . On applique l'algorithme de Lanczos 1 pour calculer la tridiagonale de Lanczos T_j et V_j à chaque itération. On se limitera aux matrices A_2 et A_3 avec des clusters de taille 4. Un code couleur est mis en légende afin de pouvoir suivre la convergence d'une paire de Ritz entre les différents graphiques. Les quotients de Rayleigh compris entre 0,5 et 1,5 en rouge correspondent au cluster central. En bleu en représentera soit les quotients de Rayleigh inférieurs à 0,5 soit supérieurs à 1,5.

Convergence vers les valeurs propres On représente l'évolution du quotient de Rayleigh en fonction des itérations de Lanczos. Les valeurs en haut représentent les valeurs de θ . Les croix sont les quotients calculés et les points sont les valeurs propres de la matrice. La méthode de Lanczos RR permet au bout de $n = 100$ itérations d'obtenir une approximation de toutes les valeurs propres de la matrice indépendamment de la position du cluster. On remarque cependant que pour une cluster LDC (figure 15a, annexe A.2) les valeurs propres du cluster LDC sont obtenues plus tard que celles de CC . Plus θ est grand plus le nombre d'itérations pour approcher toutes les valeurs propres de LDC est grand. Pour la matrice A_3 ayant un cluster MDC le phénomène inverse se produit (figure 15b, annexe A.2).

La méthode de Lanczos RR est plus efficace pour calculer les approximations des valeurs propres d'une matrice lorsque le spectre de celle-ci est composé d'un cluster MDC .

Convergence vers les vecteurs propres, réponse à la question Q5 Pour la convergence vers les vecteurs propres, les vecteurs de RR approchant les vecteurs propre LDC convergent plus rapidement que ceux approchant CC , mais le nombre d'itération nécessaires pour atteindre un sinus de 10^{-7} augmente avec θ (figure 8a). Pour la matrice A_3 (figure 8b) la convergence des vecteurs de RR vers les vecteurs de MDC se fait au cours des premières itérations.

La méthode RR permet d'obtenir une très bonne approximation des vecteurs propres des clusters. On note que pour un MDC ces approximations sont obtenues en très peu d'itérations indépendamment de θ . En faisant le lien avec l'étude de la déflation via une base perturbée (section 3.2), cette méthode construit des approximations suffisamment précises pour être utilisées dans le *CG-DEF*.

Convergence vers les paires propres On constate que le résidu de la paire propre calculée par RR pour la matrice A_2 est stable suivant θ (figure 16a, annexe A.3), là où pour la matrice A_3 (figure 16b, annexe A.3), le résidu augmente avec θ .

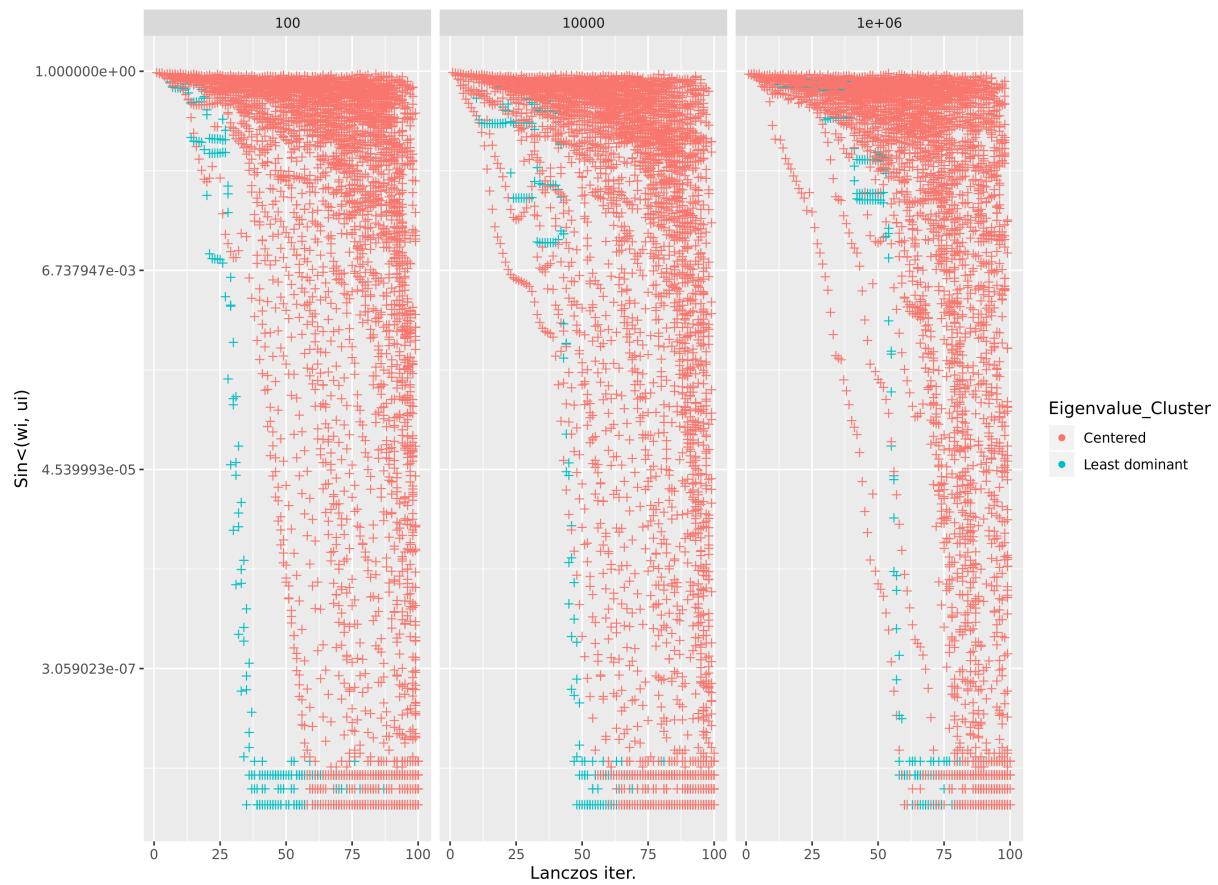
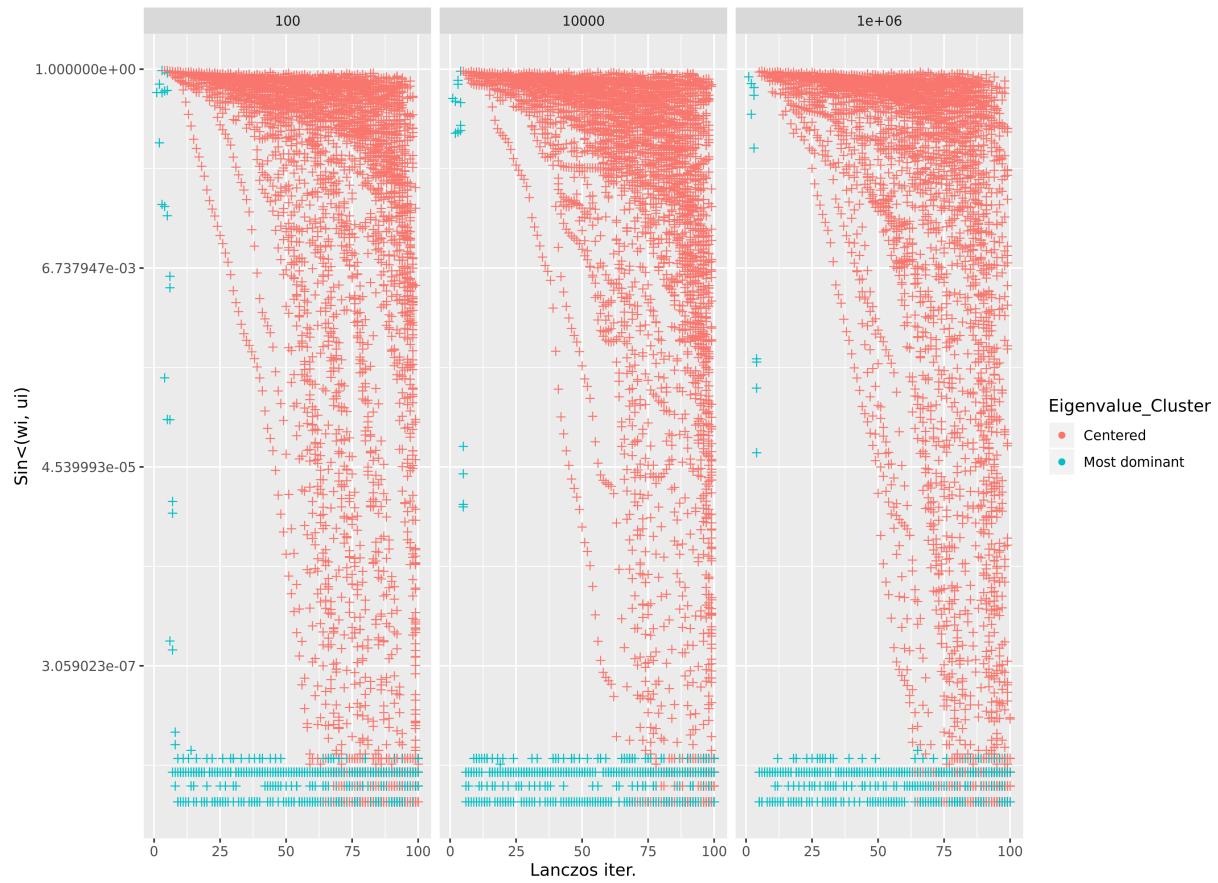
(a) Matrice A_2 , dimension 100×100 (b) Matrice A_3 , dimension 100×100

FIGURE 8 – Évolution du sinus de l'angle entre les vecteurs propres calculés via Lanczos Rayleigh-Ritz et les vecteurs propres exacts, cluster de taille 4.

4.1.3 Phénomène de Ghost values et ré-orthogonalisation

La similarité entre les matrices A et T_m repose sur $V_m^T V_m = I_m$. En précision finie, les erreurs d'arrondis accumulées au cours des itérations de Lanczos font perdre cette propriété d'orthonormalité de V_m . Sans ré-orthogonalisation, plusieurs auteurs [26][chap 13, p 293] et [22][chap 6, p 282] expliquent que la procédure de Lanczos (RR) génère alors plusieurs valeurs de ritz "Ghost Eigenvalues" pour la même valeur propre de A , les vecteurs de Ritz correspondant seront alors multiples les uns des autres. Dans les expériences réalisées, la base V_m est ré-orthogonalisée.

4.2 Étude avec les approximations d'Harmonic-Ritz

La méthode Lanczos Harmonic-Ritz (HR), aussi appelée Modified Rayleigh-Ritz ou Interior Rayleigh-Ritz, est une variante de (RR) développée par Morgan et Zeng [29] en 1991 afin de déterminer les paires propres intérieures au spectre des matrices symétriques. La méthode, se basant sur un shift-invert, permet de cibler la partie du spectre que l'on souhaite approcher et est moins sensible au phénomène de Ghost Eigenvalues.

4.2.1 Définition

Soit $\sigma \in \mathbb{R}$ le shift désignant la valeur propre λ de A que l'on veut approcher. Comme expliqué par [30] la meilleure méthode est celle de RR appliquée à $(A - \sigma I)^{-1}$ car les valeurs propres intérieures de A proches de σ sont alors les valeurs propres extérieures de $(A - \sigma I)^{-1}$. Cela nous amène à considérer le problème aux valeurs propres suivant :

Trouver un couple $(\mu, z) \in (\mathbb{R} \times \mathbb{R}^n)$ avec $\mu = \frac{1}{\lambda - \sigma}$ tel que $(A - \sigma I)^{-1}z = \mu z$

(20)

L'opérateur inverse du problème (20) étant défini explicitement, il n'est pas possible de résoudre ce problème tel quel. Pour contourner cela, on projette (20) sur l'espace engendré par $(A - \sigma I)R(V)$, on a alors :

$$\begin{aligned} & ((A - \sigma I)V)^T ((A - \sigma I)^{-1}z - \mu z) = 0 \\ \iff & V^T(A - \sigma I)((A - \sigma I)^{-1}z - \mu V^T(A - \sigma I)z) = 0 \\ \iff & V^Tz = \mu V^T(A - \sigma I)z \end{aligned} \quad (21)$$

Si $z \in (A - \sigma I)R(V)$, alors $z = (A - \sigma I)V\hat{z}$ et (21) correspond au problème aux valeurs propres généralisé suivant :

$$F\hat{z} = \mu G\hat{z} \quad \text{avec } F = V^T(A - \sigma I).V \text{ et } G = V^T(A - \sigma I)^2V \quad (22)$$

Par construction si (μ, \hat{z}) est une paire propre du pencil (F, G) , alors (μ, z) est paire propre de $(A - \sigma I)^{-1}$. Par ailleurs pour revenir sur le problème aux valeurs propres de A , il faut considérer l'équation (22) sous la forme suivante : $G\hat{z} = \frac{1}{\mu}F\hat{z}$ qui peut être vue comme la projection du problème aux valeurs propres de $(A - \sigma I)$ soit sur $R(V)$, soit sur $(A - \sigma I)R(V)$:

$$V^T \left(\frac{1}{\mu}(A - \sigma I)V\hat{z} - (A - \sigma I)^2V\hat{z} \right) = 0 \quad (23)$$

$$V^T(A - \sigma I) \left(\frac{1}{\mu}V\hat{z} - (A - \sigma I)V\hat{z} \right) = 0 \quad (24)$$

Les deux cas impliquent que $(\frac{1}{\mu}, z)$ soit une paire propre de $(A - \sigma I)$ et donc que (λ, z) soit la paire propre de A mais pour (23) on a $z = (A - \sigma I)V\hat{z}$ et pour (24), $z = V\hat{z}$. Mais comme [31][p 292] le détaille, pour obtenir la paire propre de A , il est préférable d'utiliser (24)

et de considérer [29][p 36] $\rho(z, A) = \sigma + \rho(z, A - \sigma I)$ à la place de λ . Afin d'éviter de calculer explicitement les matrices de l'équation (22), on utilisera les relations définies en annexe A (A.1). Ces relations permettent de construire F et G en utilisant les coefficients du *CG-DEF* calculés lors de la résolution précédente.

4.2.2 Expérimentations

Les expérimentations sont menées sur les mêmes matrices que pour Lanczos RR. Pour obtenir les approximations des vecteur propres des matrices, on applique la procédure de Lanczos RR pour calculer V_j puis l'on résout le problème aux valeurs propres généralisé (22) avec $\sigma = 0$ pour obtenir \hat{z} le vecteur propre du pencil (G, F) . On en déduit $w = V_j \hat{z}$ l'approximation du vecteur propre de A à l'itération j de Lanczos HR.

Convergence vers les valeurs propres En observant les figures 17a et 17b (en annexe 17) les mêmes remarques que pour les figures 15a et 15b s'appliquent, à la différence que pour la matrice A_2 , Lanczos HR génère au cours d'une même itération un plus grand nombre d'approximations des valeurs propres de *LDC*.

Convergence vers les vecteurs propres, répond à la question Q5 Pour la matrice A_2 (figure 9a), le sinus entre les vecteurs propre exacts de *LDC* et les approximations restent à un niveau inférieur à 10^{-7} tant que $\theta \leq 10^4$. Pour des valeurs de θ plus grandes, les approximations sont de moins bonne qualité. Pour la matrice A_3 (figure 9b) la qualité de l'approximation ne varie pas en fonction de θ et atteint une précision de 10^{-8} .

La procédure de Lanczos RR est donc plus stable et précise que la variante HR pour le calcul des approximations des vecteurs propres et ce indépendamment du fait que le spectre de la matrice ait une cluster *LDC* ou *MDC*.

Convergence vers les pairs propres Les résidus des paires propres calculées par Lanczos HR augmentent lorsque θ augmente, jusqu'à atteindre une précision de 10^{-7} (figures 18a, annexe A.5).

La procédure de Lanczos RR est plus précise et stable que la variante HR, indépendamment du type de cluster, et notamment pour le calcul des vecteurs appartenant aux clusters qui doivent être déflatés. De cette étude on peut favoriser l'implantation de Lanczos RR dans le *CG-DEF* par rapport à HR, ce qui répond à la question Q5. Néanmoins il est nécessaire de rappeler que les procédures de Lanczos RR ou HR en tant que telles sont différentes de leur implantations au sein du *CG-DEF* puisque, dans ce cas, les coefficients utilisés pour construire les problèmes aux valeurs propres sont ceux du *CG-DEF*. Cela nous amène à considérer la question Q6, tel que l'on comparera les approches RR et RH dans le *CG-DEF* dans les sections 5.1 et 5.2.

5 Déflation avec l'information spectrale d'une résolution précédente

Pour déterminer l'approche basée sur les coefficients du *CG-DEF* générant les approximations des vecteurs propres de A entre Lanczos Rayleigh-Ritz et Lanczos Harmonic-Ritz à un niveau de précision suffisamment élevé pour retrouver les résultats de convergence de la Section 3.1 où les colonnes de W étaient générées en utilisant les vecteurs propres exacts de A , on réalise dans les sections suivantes les mêmes études qu'en Section 3.1 mais en appliquant la procédure détaillée à la question Q6 de la Section 2.3.

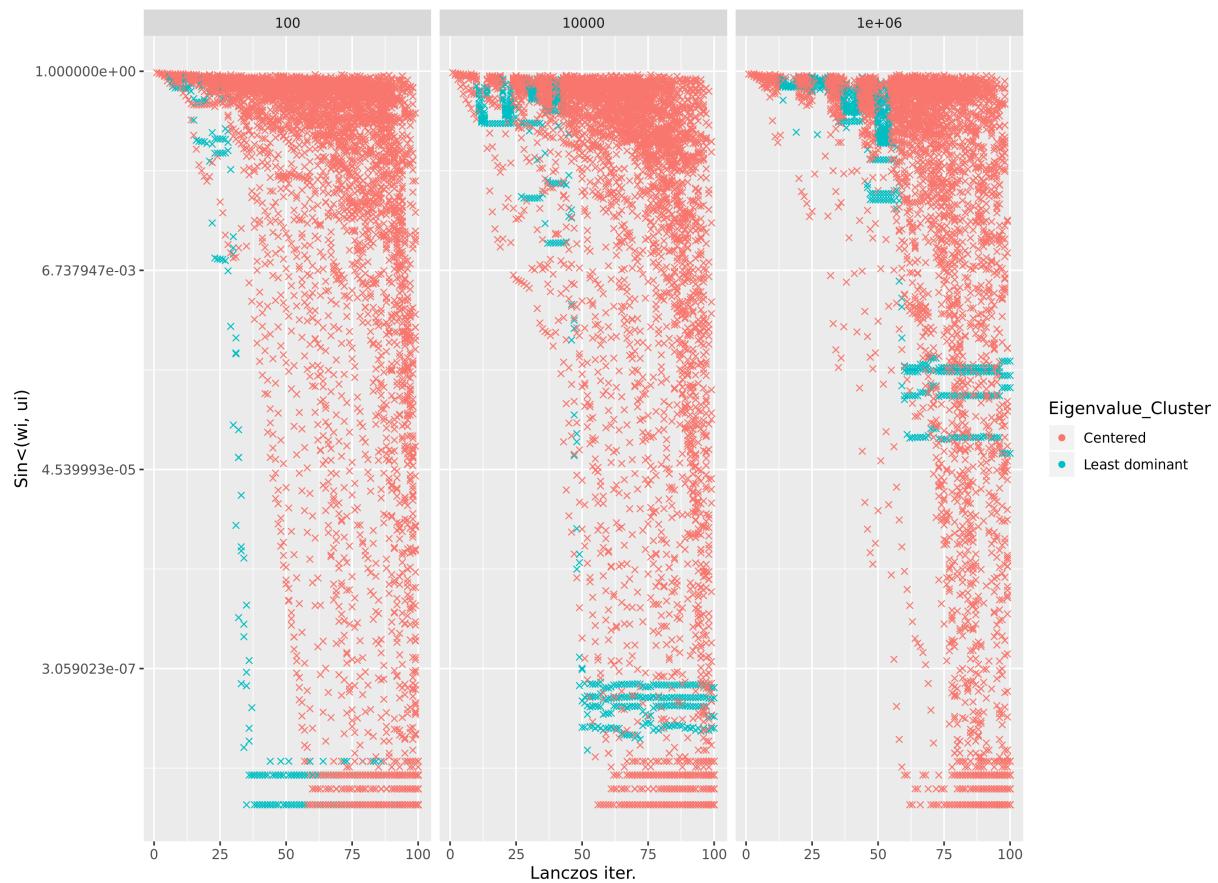
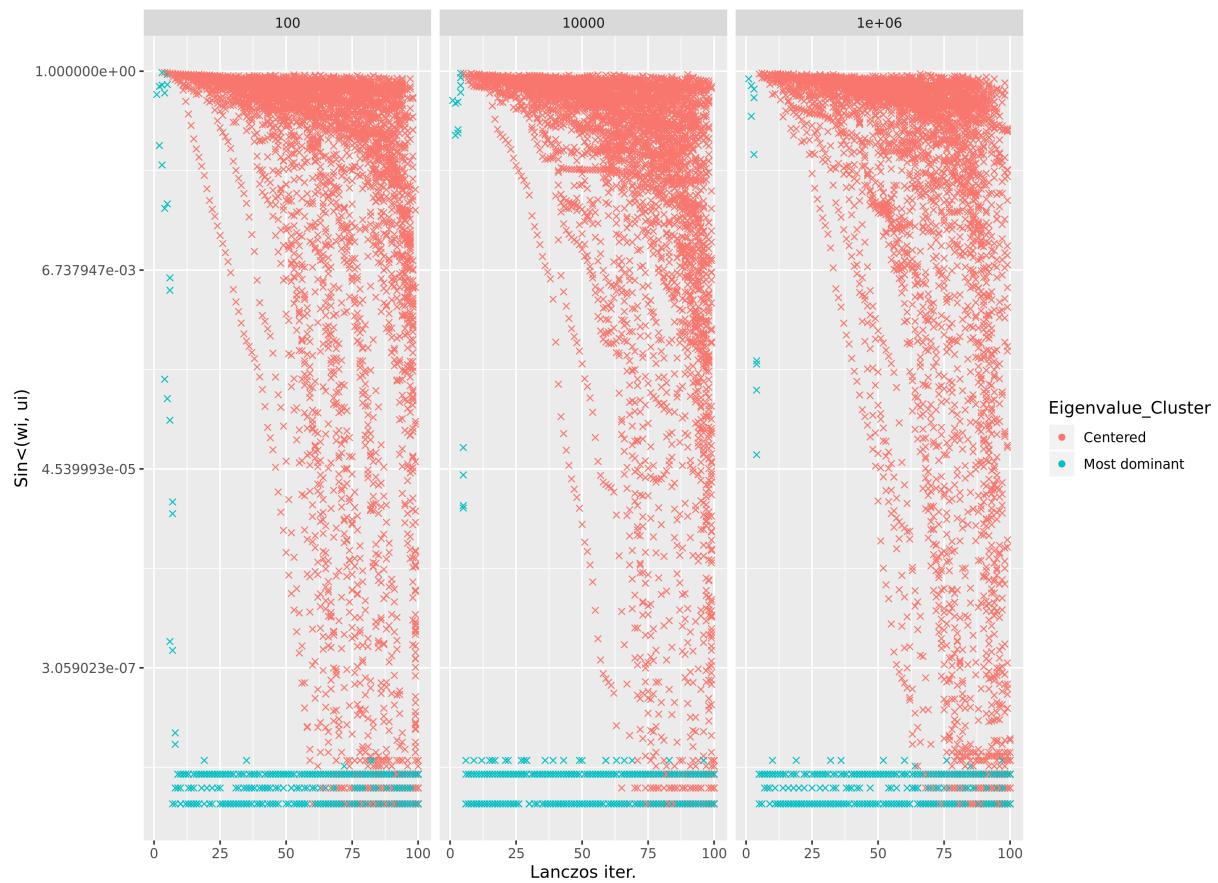
(a) Matrice A_2 , dimension 100×100 (b) Matrice A_3 , dimension 100×100

FIGURE 9 – Evolution du sinus de l'angle entre les vecteurs propres calculés via Lanczos Harmonic-Ritz et les vecteurs propres exacts, cluster de taille 4.

Pour les cas tests suivants, on se limite aux matrices A_2 , A_3 , avec un cluster de taille 4, dans l'objectif de répondre à la question Q6. Les figures représentant la convergence de l'erreur inverse en fonction des itérations de la méthode déflatée suivant le paramètre θ à droite, et les labels en hauts des graphiques font références aux paramètres définis dans la Section 2.4.

5.1 Approche basée sur les approximations de Rayleigh-Ritz

Nous avons vérifié dans la Section 4.1 que la méthode Lanczos RR permettait de calculer une bonne approximation des vecteurs propres des clusters, et ce en très peu d'itérations pour un cluster MDC et suivant un nombre d'itérations croissant pour un cluster LDC . Voyons maintenant à travers la convergence de la méthode déflatée si la procédure de Lanczos RR basée sur les coefficients du $CG\text{-}DEF$ permet d'assurer une accélération de la convergence. Comme on l'observe sur la figure 10, le $CG\text{-}DEF$ n'accélère pas la convergence du système $s = 2$ par rapport au système $s = 1$, indépendamment de tous les paramètres de déflation. En se référant à la figure 6, on peut conclure que l'approche basée sur RR du $CG\text{-}DEF$ ne permet pas de déterminer des approximations des vecteurs de LDC à la précision nécessaire suivant θ pour une matrice ayant un cluster LDC .

Lorsque l'opérateur linéaire possède un cluster MDC figure 11, on voit que déflater k vecteurs de LD a le même effet que déflater par rapport à MD . On comprend que l'approche basée sur RR calcule uniquement des approximations des vecteurs de MD . On remarque cependant qu'en déflatant les k vecteurs de MDC avec $\ell = [|0, 5|]$, on retrouve les profils de convergences de la figure 5b. Autrement dit, l'approche basée RR permet d'obtenir très rapidement des approximations des vecteurs de MDC de bonnes qualités, qui utilisées pour la déflation permettent de réduire significativement le nombre d'itérations à convergence du second système comparativement à la résolution du premier système. Pour répondre à la question Q6, l'approche basée RR appliquée au $CG\text{-}DEF$ peut ainsi être appliquée à la résolution successive de plusieurs seconds membres lorsque l'opérateur linéaire du système à un cluster MDC . De plus le volume d'information ℓ à sauvegarder entre deux résolutions successive doit être au moins égale au nombre de vecteurs propres k devant être déflatés.

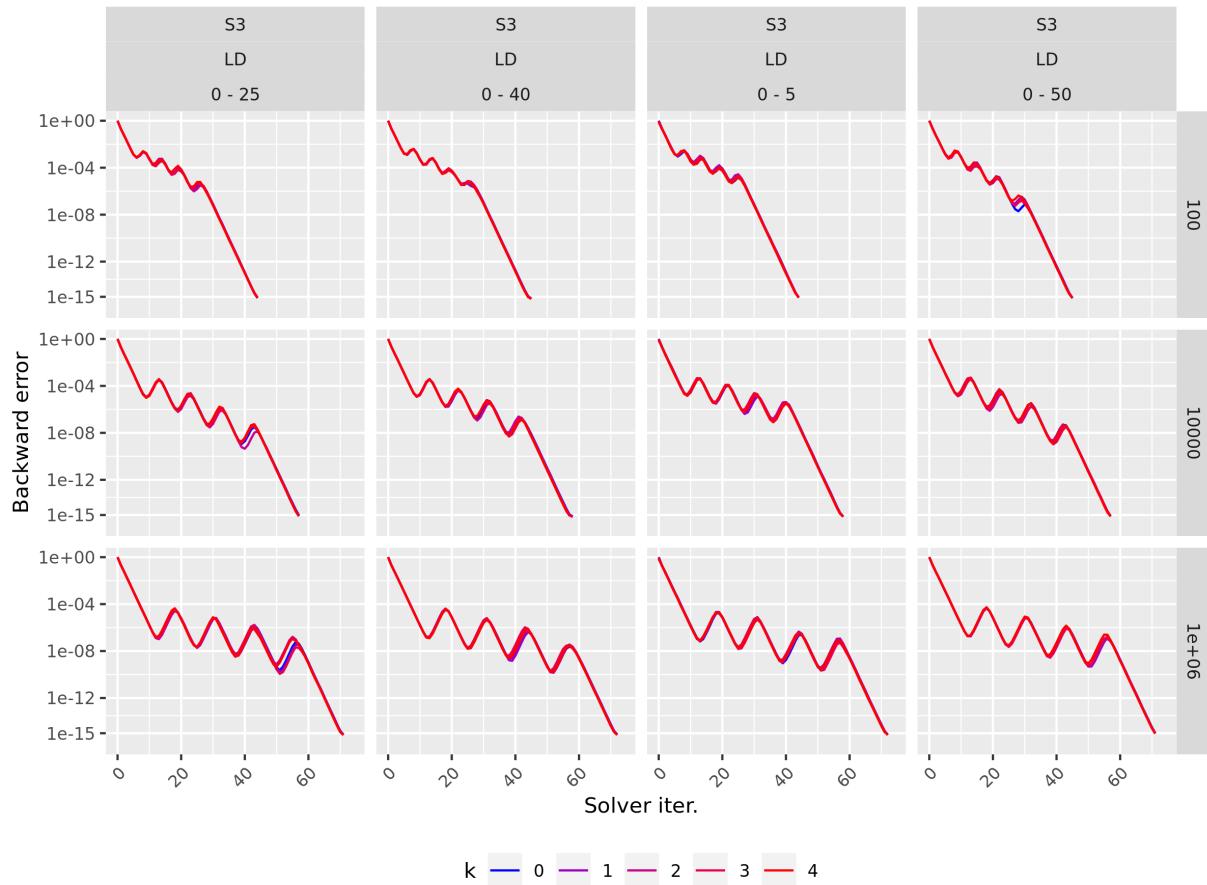
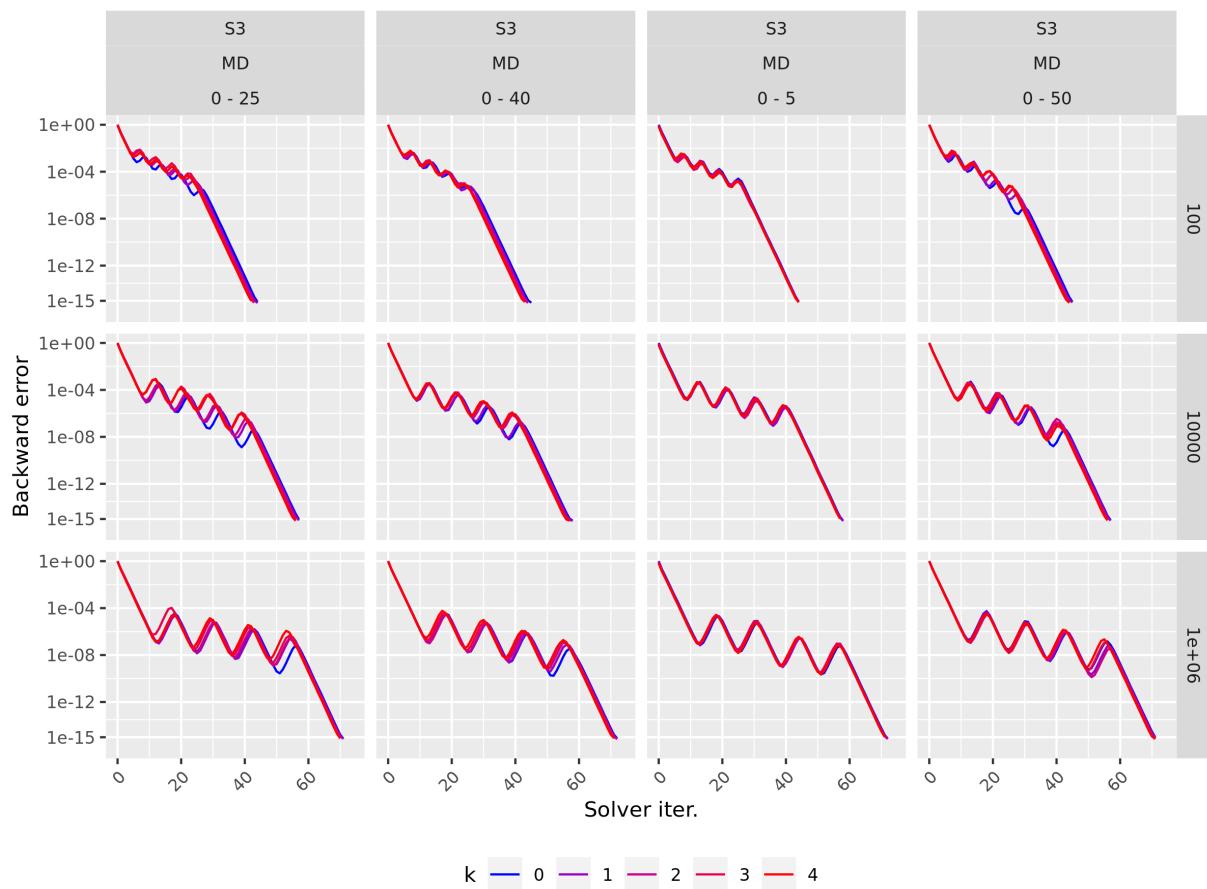
5.2 Approche basée sur les approximations d'Harmonic Ritz

Voyons maintenant l'approche basée HR appliquée au $CG\text{-}DEF$. Cette méthode ne permet pas d'accélérer la convergence entre deux résolutions successives pour une opérateur linéaire dont le spectre est composé d'une cluster LDC comme le montre les résultats de convergences de la figure 12b lorsque l'on déflat par rapport aux k vecteurs de MD . Cependant, la méthode permet de calculer les approximations des vecteurs du cluster LDC à une qualité suffisante pour accélérer la converge, figure 12a. On remarque que cette accélération est obtenue lorsque le volume d'information ℓ sauvegardé entre deux résolutions est grand. On remarque aussi que plus θ est grand plus ℓ doit être grand pour pouvoir déflater l'ensemble du cluster.

Dans les cas où le spectre de la matrice dispose d'un cluster MDC figure 13 , on retrouve des résultats similaires à l'approche basée sur RR, à la différence que la méthode calcule les approximations des vecteurs de MDC uniquement pour ℓ proche de k .

Finalement pour répondre à la question Q6, les deux approches sont équivalentes pour des matrices dont le spectre a un cluster MDC et ces méthodes sont d'autant plus efficaces que le conditionnement de la matrice est élevé. Pour des matrices dont le spectre posséde un cluster LDC seule la méthode basée sur HR permet d'obtenir une accélération de la convergence. Néanmoins dans ce cas, plus la matrice à un mauvais conditionnement, plus le volume d'information à sauvegarder entre deux résolutions doit être grand.

Pour une matrice ayant les deux clusters, comme le volume d'information à utiliser pour construire le problème aux valeurs propres varie en fonction du cluster à déflater, il faut néces-

(a) Déflation suivant les k vecteurs de LD cluster de taille 4.(b) Déflation suivant les k vecteurs de MD cluster de taille 4.**FIGURE 10** – Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de RR, matrice A_2 .

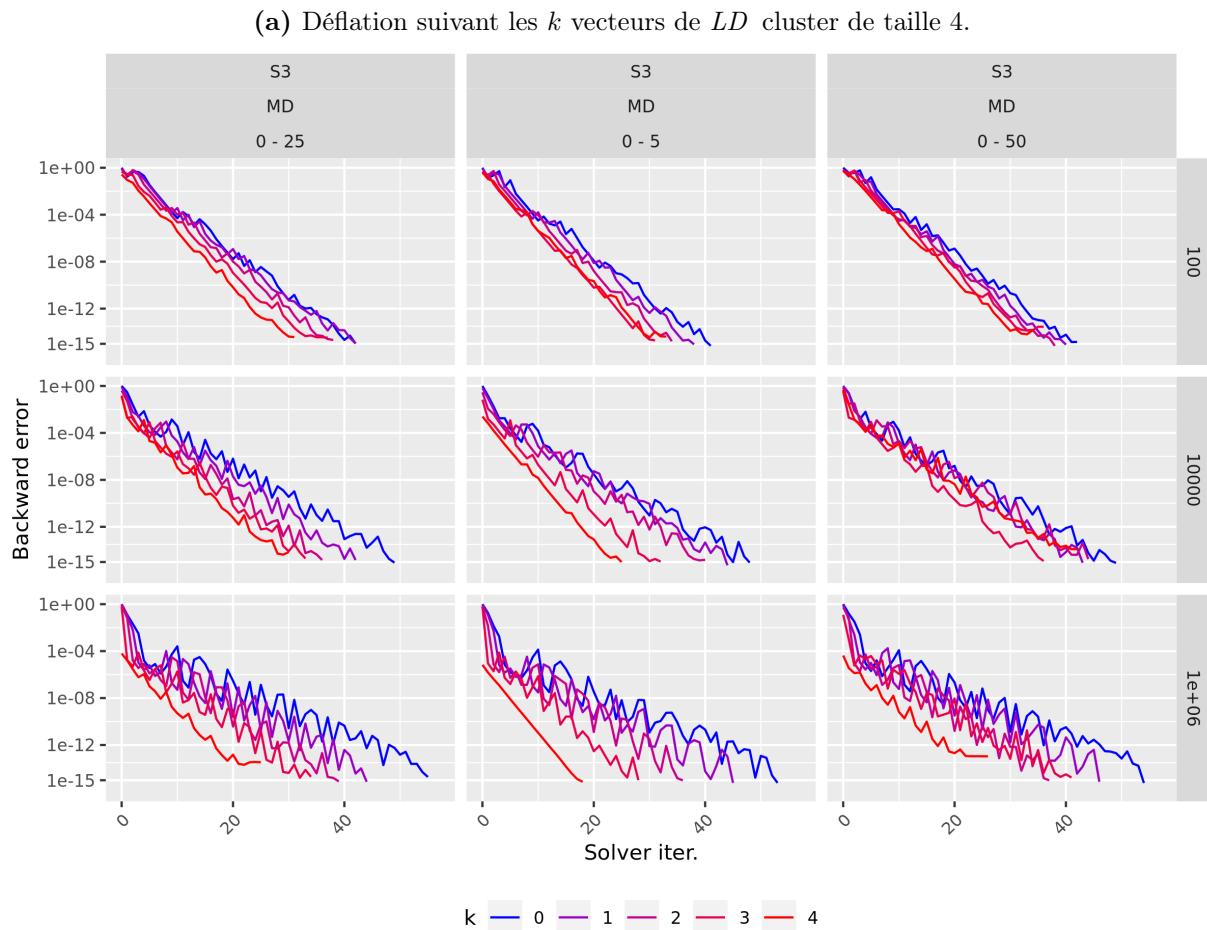
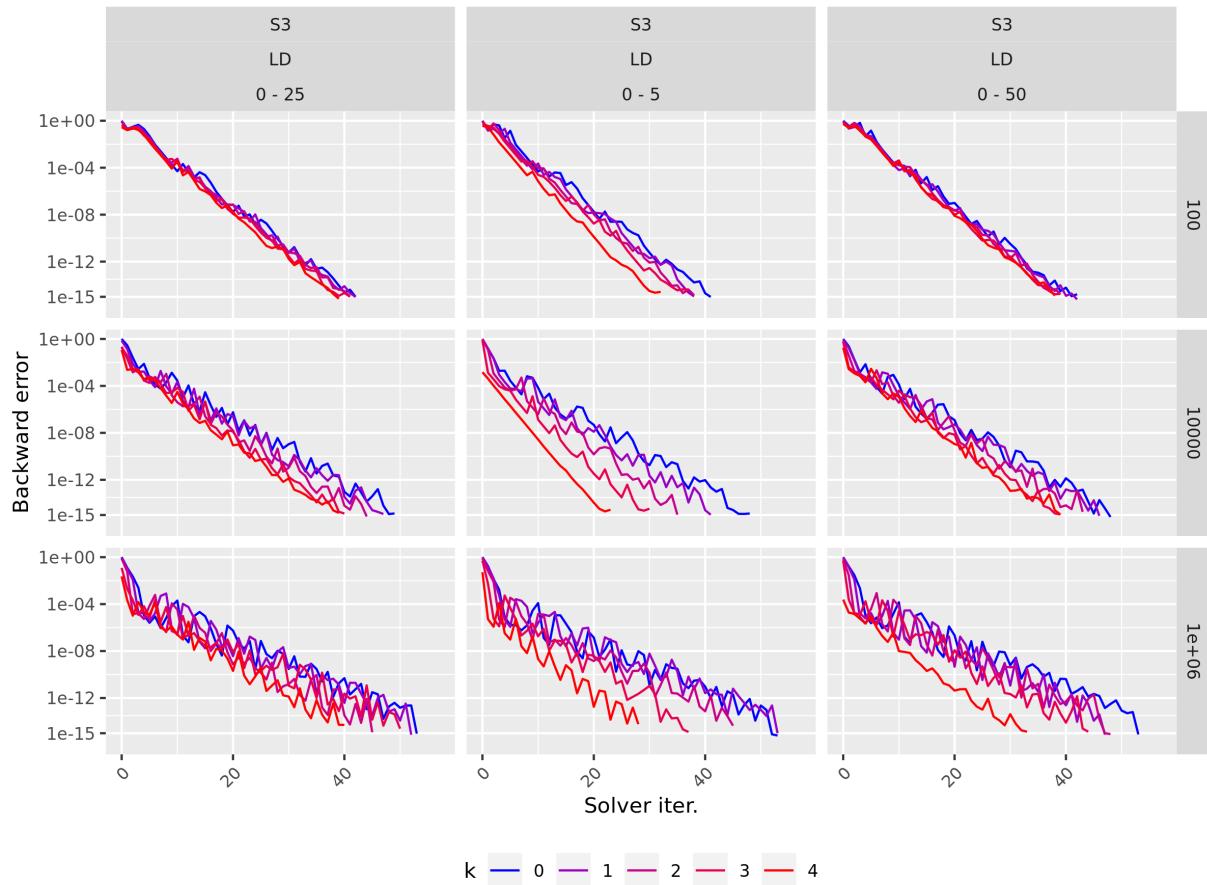
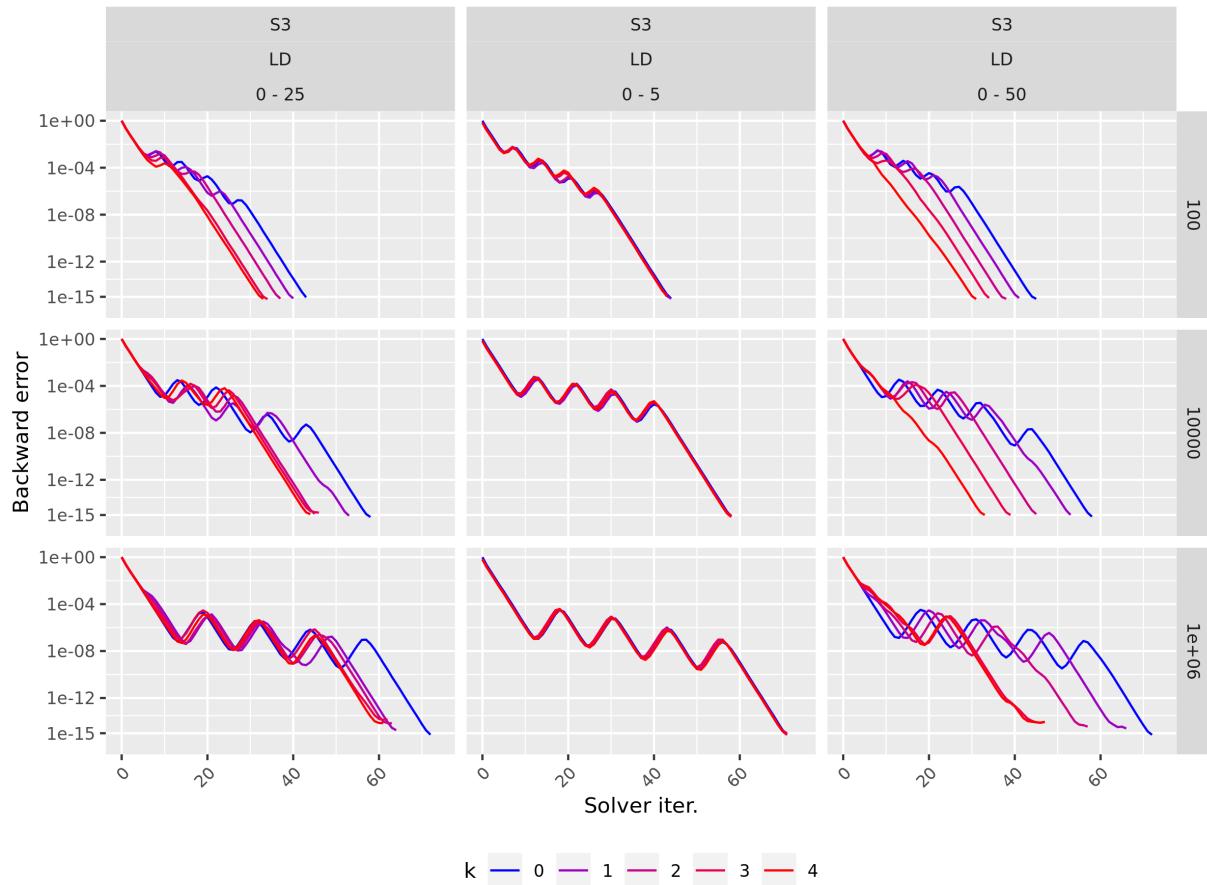
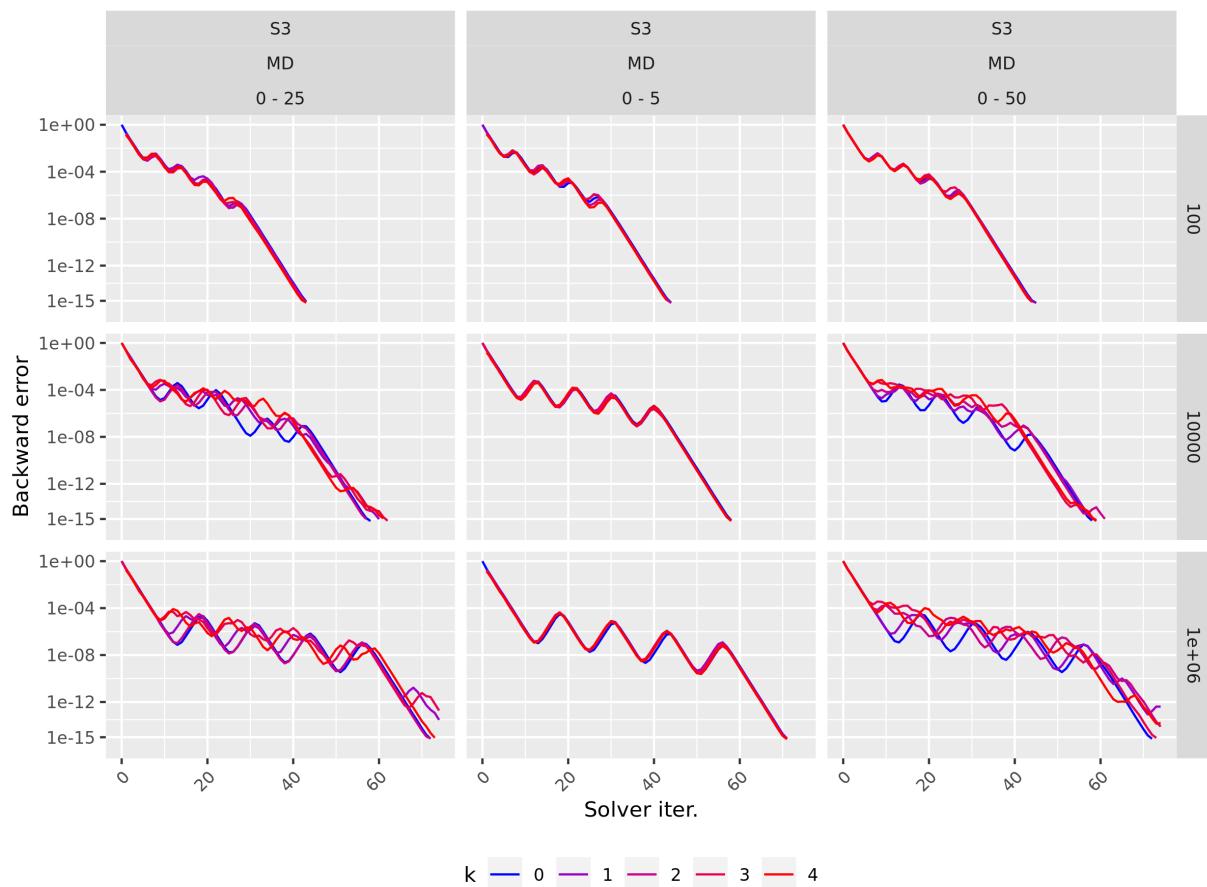
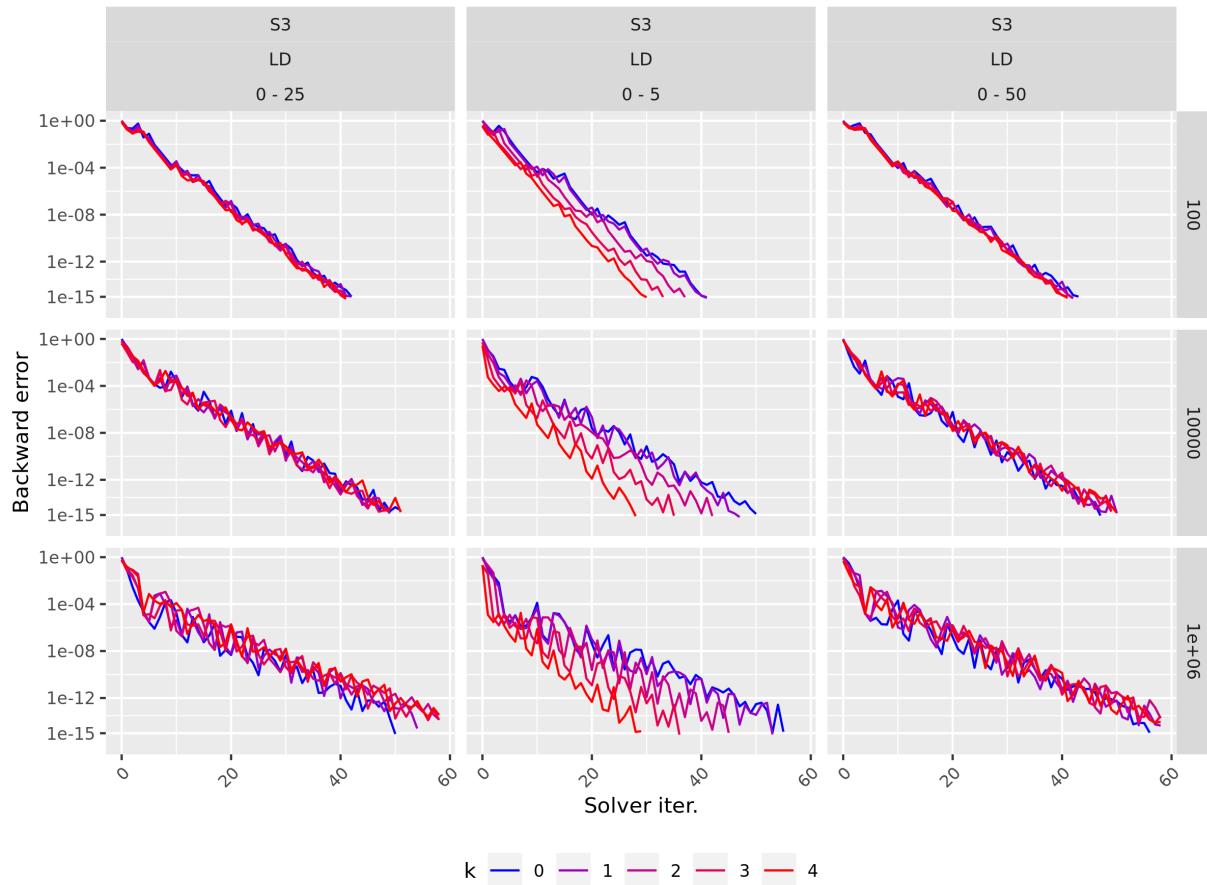
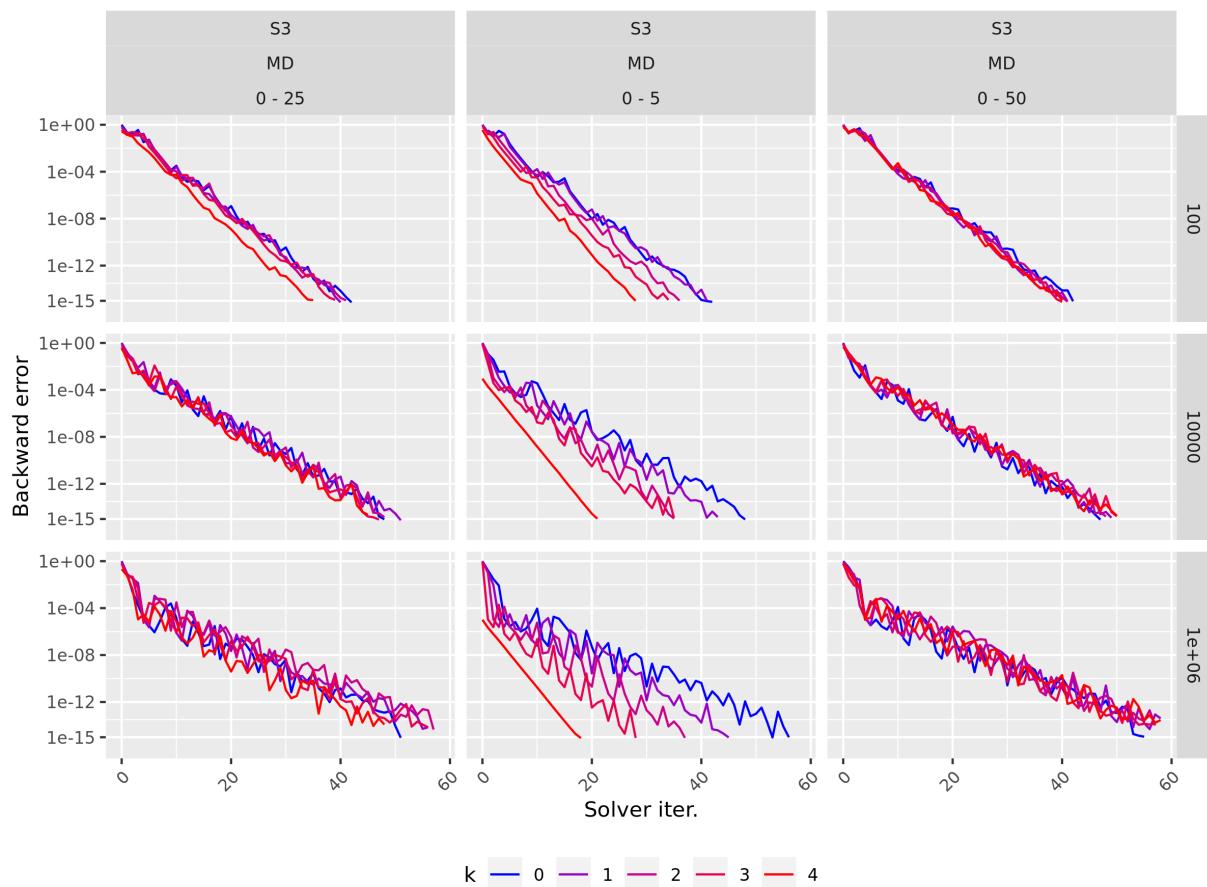


FIGURE 11 – Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de RR, matrice A_3 .

(a) Déflation suivant les k vecteurs de LD cluster de taille 4.(b) Déflation suivant les k vecteurs de MD cluster de taille 4.**FIGURE 12** – Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de HR, matrice A_2 .

(a) Déflation suivant les k vecteurs de LD cluster de taille 4.(b) Déflation suivant les k vecteurs de MD cluster de taille 4.**FIGURE 13** – Courbes de convergence suivant l'erreur inverse du $CG\text{-}DEF$ approche basée sur les approximations de HR, matrice A_3 .

sairement résoudre en deux temps le problème aux valeurs propres afin d'obtenir des approximations des vecteurs propres des deux clusters. Sinon un seul des deux clusters sera déflaté.

6 Déflation avec l'information spectrale de multiples résolutions successives, approche basée HR

Cette section a pour objectif de répondre à la question Q7 de la Section 2.3. La déflation dépend de la qualité de l'information spectrale calculée, l'étape 8 de l'algorithme 5 permet de raffiner l'approximation des vecteurs propres. On veut vérifier que le *CG-DEF* permet de réduire de plus en plus le nombre d'itérations à convergence au fur et à mesure que le nombre de systèmes résolus augmente, comparativement au *CG*.

6.1 Algorithme du *CG-DEF* pour la résolution successive de multiples seconds membres

Lorsque $\nu > 2$, pour l'approche basée sur HR, on doit conserver en plus des éléments exposés pour aborder la question Q6, les ℓ vecteurs μ calculés à l'étape 11 de l'algorithme 4 tel que $\hat{\Delta}_{\ell+1}^{(s)} = [\mu_0^{(s)}, \dots, \mu_\ell^{(s)}]$ pour former le problème (22). On reporte en annexe A.1 le détail de la construction du problème aux valeurs propres généralisé (22). L'algorithme 5 du *CG-DEF* dans ce cas est donné par [14] tel que :

Algorithm 5 Deflated *CG* for multiples right-hand sides

- 1: Select ℓ and k with $k \leq \ell$.
 - 2: Solve the first system of (16) using algorithm 4, with $W = W^{(s=1)} = \emptyset$.
 - 3: Solve the Harmonic-Ritz problem (28) for k eigenvectors using (26) and (27).
 - 4: Set $W^{(s=2)} = [P_l^{(s=1)}]Y^{(s=1)}$
 - 5: **for** $s = 2, \nu$ **do**
 - 6: Solve (16) using algorithm 4, with $W = W^{(s)}$.
 - 7: Solve the Harmonic-Ritz problem (29) for k eigenvectors using (30) and (31)
 - 8: Set $W^{(s+1)} = [W^{(s)} | P_l^{(s)}]Y^{(s)}$
 - 9: **end for**
-

6.1.1 Expérimentations

On étudie ici la convergence du *CG-DEF* lors de la résolution successive de $\nu = 20$ systèmes. On comparera les résultats suivant que la matrice du système est A_2 ou A_3 pour répondre à la question Q7 de la section 2.3. Les matrices sont de dimension 500×500 avec des clusters de taille 4. Pour les expérimentations suivantes, le nombre de vecteurs à déflater k est fixé à 4. Les courbes en bleu correspondent à la résolution via le *CG* (étape 2 de l'algorithme 5). Le code couleur des historique de convergence permet de représenter le numéro du système linéaire dans la séquence de 1 (bleu) à 20 (rouge).

Matrice A_2 . Pour $\ell = [0, 5]$, les courbes de convergences de la résolution des systèmes $s = 2, \dots, \nu$ sont confondues avec celle de la résolution du système $s = 1$ (figure 14a S3-LD $\ell = [0, 5]$). L'étape 8 de l'algorithme 5 ne permet donc pas de raffiner la qualité des vecteurs calculés. Pour $\ell = [0, 30]$ (figure 14a S3-LD $\ell = [0, 30]$), le nombre d'itérations à convergence diminue au fur et à mesure que le nombre de systèmes résolus augmente. On voit que passé la résolution des 10 premiers systèmes, l'ensemble des vecteurs du cluster *LDC* sont déflatés et les courbes de convergences sont linéaires. Pour répondre à la question Q7, lorsque le volume d'information

à conserver entre deux résolutions est suffisamment grand, la méthode du gradient conjugué déflaté permet d'accélérer la convergence après chaque résolution.

Matrice A_3 . Pour $\ell = [0, 5]$ la résolution successive des systèmes permet pour les systèmes déflatés ($s > 1$) d'atteindre une précision de $1O^{-15}$ sur l'erreur inverse en deux fois moins d'itérations qu'une résolution avec le CG , ($s = 1$) (figure 14b S3-LD $\ell = [0, 5]$). Le nombre de résolutions successives nécessaire pour déflater la totalité des vecteurs du cluster MDC varie suivant θ . Pour $\ell = [0, 30]$ la résolution successive des systèmes ne permet pas d'accélérer la convergence.

Pour répondre à la question Q7, la méthode $CG\text{-}DEF$ avec une approche basée sur HR pour le calcul de W appliquée à la résolution successive d'un système de la forme (16) est compétitive par rapport au CG . L'étape 8 de l'algorithme 5 permettant d'améliorer la qualité des approximations des vecteurs propres après chaque résolution.

7 Conclusion

Les expérimentations des Sections 3.1.1, 3.1.2, 3.1.3 et 3.1.4 ont montré que le $CG\text{-}DEF$ accélère la convergence comparativement au CG lorsque l'on déflat les vecteurs propres des clusters du spectre de la matrice du système linéaire. Nous avons aussi observé que le $CG\text{-}DEF$ appliqué à des matrices ayant un cluster MDC de valeurs propres, permet d'obtenir à l'itération initiale une solution plus précise suivant l'erreur inverse comparativement à la solution calculée par le CG . Lorsque l'on ne dispose pas des vecteurs propres exacts pour construire W , nous avons étudié deux méthodes permettant de générer des approximations de ces vecteurs (Section 5). L'approche basée sur la procédure de Lanczos Harmonic-Ritz permet de construire des approximations des vecteurs des clusters LDC et MDC à une précision suffisante pour que la méthode déflatée soit compétitive par rapport au CG . L'approche basée sur la procédure de Rayleigh-Ritz génère uniquement des approximations des vecteurs de MDC . Les expériences réalisées dans ce cadre montrent que pour calculer des approximations des vecteurs du cluster LDC , le volume d'information à sauvegarder entre deux résolutions ℓ doit être grand devant le nombre d'approximations à calculer k . Lorsque l'on cherche à approcher les vecteurs de MDC , ℓ doit être proche de k . Enfin, dans le cadre de multiples résolutions successives (Section 6), en utilisant le $CG\text{-}DEF$ avec l'approche basée sur HR comme solveur aux valeurs propres, le nombre d'itérations à convergence diminue après chaque résolution. Cette méthode est bien adaptée dans le cas d'un très grand nombre de seconds membres.

Pour pousser cette étude plus loin, il serait nécessaire d'étudier les approches développées par [3] et [16] pour le calcul des vecteurs propres. Il serait aussi intéressant de proposer une version restart du $CG\text{-}DEF$, l'idée étant d'actualiser W^s au cours d'une même résolution et non plus entre chaque résolution. Lorsque l'itération j de l'algorithme (4) atteint ℓ , l'algorithme 5 serait relancé avec $x_0^s = x_j^s$ et $W^s = [W^s | P_\ell^s]Y^s$.

Lors de ce stage, en étudiant la méthode du gradient conjugué déflaté, j'ai pu travailler à la fois sur les méthodes de résolution des systèmes linéaires et sur les problèmes aux valeurs propres. Cela m'a permis de découvrir une partie de la richesse de l'analyse numérique. Travailler au sein du HiePACS fut une bonne expérience, me donnant l'opportunité de découvrir les outils d'intégration continue tels que gitlab, et les outils de reproductibilité des expérimentations numériques via l'utilisation la programmation lettrée (emacs org-mode) pour développer en python ces expériences.

Références

- [1] Charles R.Johnson ROGER A.HORN. "Matrix Analysis".
- [2] G.H. Golub & AL. "Matrix Computations".
- [3] Venkovic et AL. "Comparative study of harmonic and Rayleigh-Ritz procedures with application to deflated conjugate gradients".
- [4] D. Steven Mackey & AL. "G-Reflectors : analogues of Householder transformations in scalar ..."
- [5] Yousef SAAD. "Iterative Methods for Sparse Linear Systems". ISBN : 0-89871-534-2.
- [6] Ilse C. F. Ipsen & Carl D. MEYER. "The Idea behind Krylov Methods".
- [7] Y. Saad et M.H. SCHULTZ. "GMRES : A generalized minimal residual algorithm for solving nonsymmetric linear systems".
- [8] R. B. MORGAN. "A restarted GMRES method augmented with eigenvectors".
- [9] A. CHAPMAN et Y. SAAD. "Deflated and augmented Krylov subspace techniques".
- [10] R. B. MORGAN. "Restarted block-GMRES with deflation of eigenvalues".
- [11] STEVEN F. ASHBYt & AL. "A Taxonomy for Conjugate Gradient Methods".
- [12] D. P. O'LEARY. "The Block Conjugate Gradient Algorithm and Related Methods".
- [13] Frédéric Guyomarc'h JOCELYNE ERHEL. "An augmented conjugate gradient method for solving consecutive symmetric positive definite linear systems".
- [14] Y. Saad & AL. "A Deflated Version of The Conjugate Gradient Algorithm".
- [15] R. B. MORGAN. "Deflated and Restarted Symmetric Lanczos Methods for Eigenvalues and Linear Equations with Multiple Right-hand Sides".
- [16] A.Stathopoulos & K.ORGINOS. "Computing and Deflating Eigenvalues While Solving Multiple Right-Hand Side Linear Systems with an application to Quantum Chromodynamics".
- [17] K. Kahl & H. RITTICH. "The Deflated Conjugate Gradient Method : Convergence Perturbation and Accuracy".
- [18] Randall J. LEVEQUE. "Finite Volume Methods for Hyperbolic Problems". 2002.
- [19] C. Berthon ALL. "An efficient scheme on wet/dry transitions for shallow water equations with friction".
- [20] O.C. Zienkiewicz R.L. TAYLOR. "The Finite Element Method". 2000.
- [21] Stéphane LEJEUNES. "Modélisation de structures lamifiées élastomère-métal à l'aide d'une méthode de réduction de modèles". Université de la méditerranée - Aix-Marseille II, 2006. ISBN : HAL Id : tel-00090600.
- [22] Lloyd N. Trefethen & David BAU. "Numerical Linear Algebra". ISBN : 0-89871-361-7.
- [23] A. SCHWARZENBERG-CZERNY. "On matrix factorization and efficient least squares solution", 405-410, Series 110 (1995).
- [24] Multifrontal Massively Parallel sparse direct Solver. URL : <http://mumps.enseeiht.fr/index.php?page=home>.
- [25] Parallel Sparse matriX package. URL : <https://solverstack.gitlabpages.inria.fr/pastix/>.
- [26] I.Beresford N. PARLETT. "The Symmetric Eigenvalue Problem".
- [27] O.Coulaud & AL. Deflation and augmentation techniques in Krylov subspace... Rapp. tech. Inria, 0249-6399.

- [28] Luc. GIRAUD. “Introduction to Krylov Subspace Methods for The Solution of Linear Systems”, p. 168.
- [29] Ronald B. Morgan & Min ZENG. “Harmonic Projection Methods for Large Non-Symmetric Eigenvalue Problems”.
- [30] D. S. SCOTT. “The advantages of inverted operators in Rayleigh-Ritz approximations”.
- [31] Ronald B. MORGAN. “Computing Interior Eigenvalues of Large Matrices”.

8 Annexes

A Annexes

A.1 Annexe A : construction du problème (22)

Lorsque l'on résout successivement ν systèmes linéaires :

$$A.x^s = b^s, \forall s \in [|1, \nu|] \quad (25)$$

la méthode *CG-DEF* consiste à résoudre le système ($s=1$) avec le *CG* en considérant que $W^{(1)} = \emptyset$. Lors de cette première résolution, on conserve les ℓ premières données que sont les pas de descente $\alpha_\ell^s = \{\alpha_0^{(s)}, \dots, \alpha_{\ell-1}^{(s)}\}$ et $\beta_\ell^s = \{\beta_0^{(s)}, \dots, \beta_{\ell-1}^{(s)}\}$, la matrice $P_\ell^{(s)} = [p_0^{(s)}, \dots, p_{\ell-1}^{(s)}]$ des ℓ premières directions de descente ainsi que le vecteur $d_\ell^{(s)} = \{d_0^{(s)}, \dots, d_{\ell-1}^{(s)}\}$ avec $(d_i^{(s)})_{i=0,\ell-1} = p_i^T A.p_i$. Avec ces informations, on construit les matrices $\tilde{D}_\ell^{(s)}$ et $\tilde{G}_\ell^{(s)}$ telles que :

$$\tilde{D}_\ell^{(s)} = \text{diag}(d_\ell^{(s)}) \quad (26)$$

$$\tilde{G}_\ell^{(s)} = \begin{bmatrix} \frac{d_0^{(s)}}{\alpha_0^{(s)}} * (1 + \beta_0^{(s)}) & -\frac{d_1^{(s)}}{\alpha_0^{(s)}} \\ -\frac{d_1^{(s)}}{\alpha_0^{(s)}} & \frac{d_1^{(s)}}{\alpha_1^{(s)}} * (1 + \beta_1^{(s)}) & -\frac{d_2^{(s)}}{\alpha_1^{(s)}} \\ & -\frac{d_2^{(s)}}{\alpha_1^{(s)}} & \ddots & \ddots \\ & & \ddots & \ddots & -\frac{d_{\ell-1}^{(s)}}{\alpha_{\ell-2}^{(s)}} \\ & & & -\frac{d_{\ell-1}^{(s)}}{\alpha_{\ell-2}^{(s)}} & \frac{d_{\ell-1}^{(s)}}{\alpha_{\ell-1}^{(s)}} * (1 + \beta_{\ell-1}^{(s)}) \end{bmatrix} \quad (27)$$

On résout alors le problème aux valeurs propres généralisés suivant :

$$\tilde{G}_\ell^{(s)}.Y^s - \Theta.\tilde{D}_\ell^{(s)}.Y^{(s)} = 0 \quad (28)$$

avec $Y^{(s)} = [y_1^{(s)}, \dots, y_k^{(s)}]$ les k vecteurs propres de Ritz associés aux k plus petites valeurs propres de Ritz $\Theta^{(s)} = [\theta_1^{(s)}, \dots, \theta_k^{(s)}]$. On peut maintenant actualiser la base W en vue de résoudre le ($s+1$)-ième système pour $s = 1, \dots, \nu - 1$, on pose $W^{(s+1)} = [W^{(s)}, P_\ell^{(s)}].Y^{(s)}$ qui pour ($s=1$) se restreint à $W^{(s+1)} = [P_\ell^{(s)}].Y^{(s)}$. On résout le ($s+1$)-ième système avec le *CG-DEF* (algorithme 4), lors de cette résolution, en plus de sauvegarder $\alpha_\ell^{(s)}$, $\beta_\ell^{(s)}$, $P_\ell^{(s)}$ et $d_\ell^{(s)}$, l'on conservera $\tilde{\Delta}_{\ell+1}^{(s)} = [\mu_0^{(s)}, \dots, \mu_\ell^{(s)}]$ et au lieu de résoudre (28), on aura :

$$G_\ell^{(s)}.Y^s - \Theta.F_\ell^{(s)}.Y^s = 0 \quad (29)$$

avec :

$$G_\ell^{(s)} = \begin{bmatrix} (A.W^{(s)})^T.M^{-1}(A.W^{(s)}) & (W^{(s)})^T.(A.W^{(s)})\tilde{\Delta}_{\ell+1}^{(s)}.\tilde{L}_\ell^{(s)} \\ (\tilde{\Delta}_{\ell+1}^{(s)}.\tilde{L}_\ell^{(s)})^T.(W^{(s)})^T.(A.W^{(s)}) & \tilde{G}_\ell^{(s)} \end{bmatrix} \quad (30)$$

$$F_\ell^{(s)} = \begin{bmatrix} (W^{(s)})^T.(A.W^{(s)}) & 0 \\ 0 & \tilde{D}_\ell^{(s)} \end{bmatrix} \quad (31)$$

avec M la matrice de préconditionnement qui est restreinte à la matrice identité si l'on ne préconsigne pas.

A.2 Annexe B : Lanczos Rayleigh-Ritz, convergences vers les paires propres

Figure 15.

A.3 Annexe C : Lanczos Rayleigh-Ritz, convergences des paires propres

Figure 16.

A.4 Annexe D : Lanczos Harmonic-Ritz, Quotient de Rayleigh

Figure 17.

A.5 Annexe E : Lanczos Harmonic-Ritz, convergences des paires propres

Figure 18

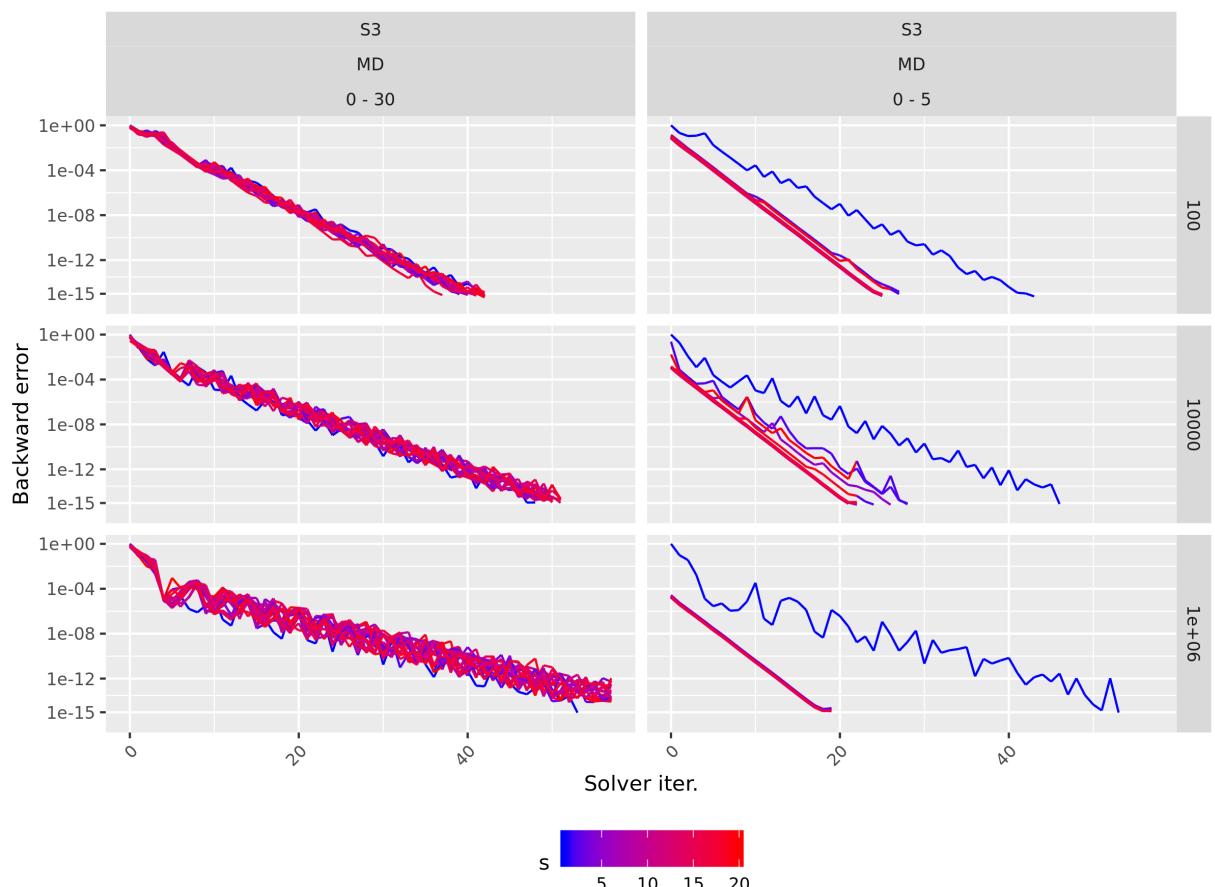
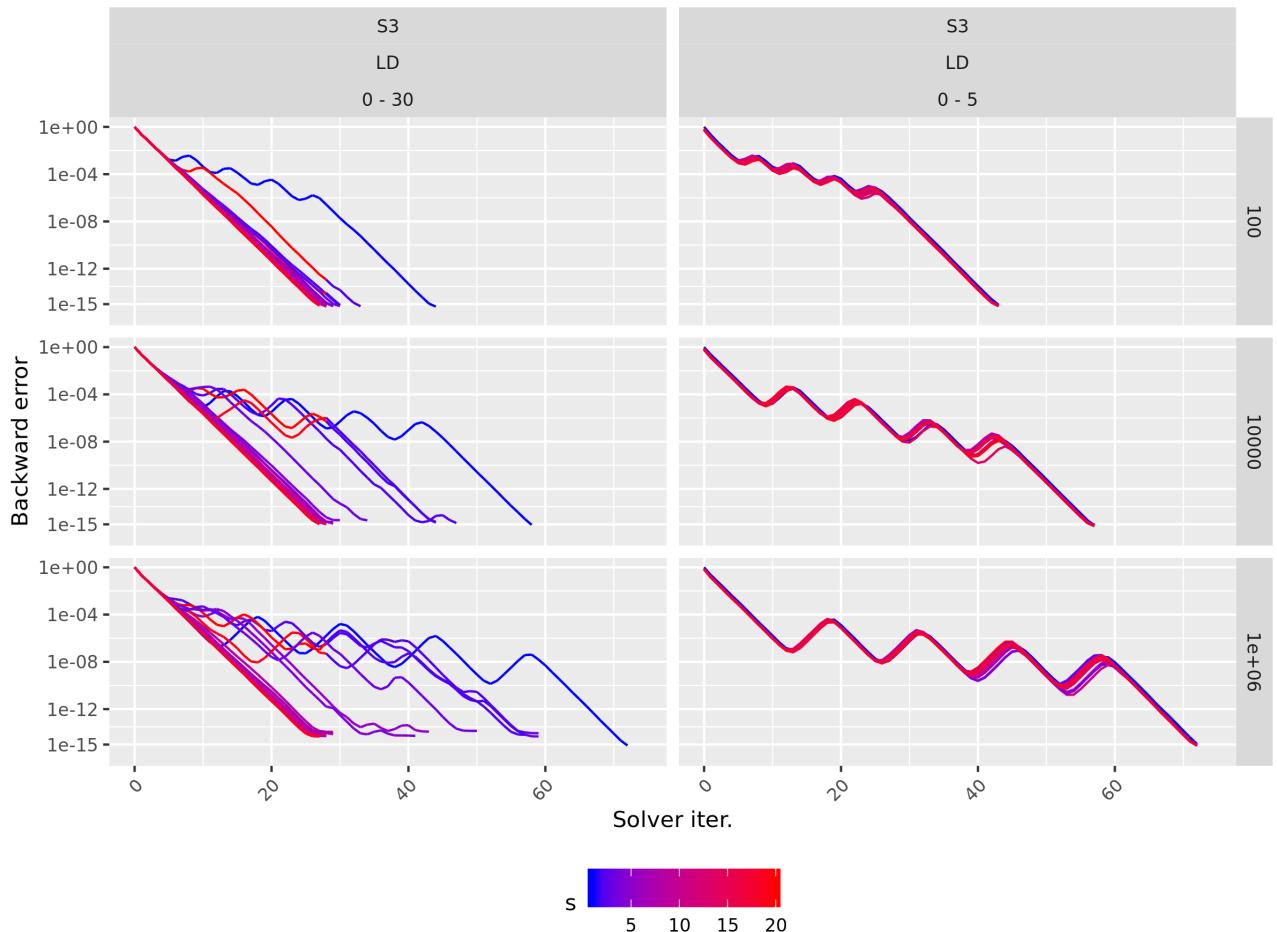
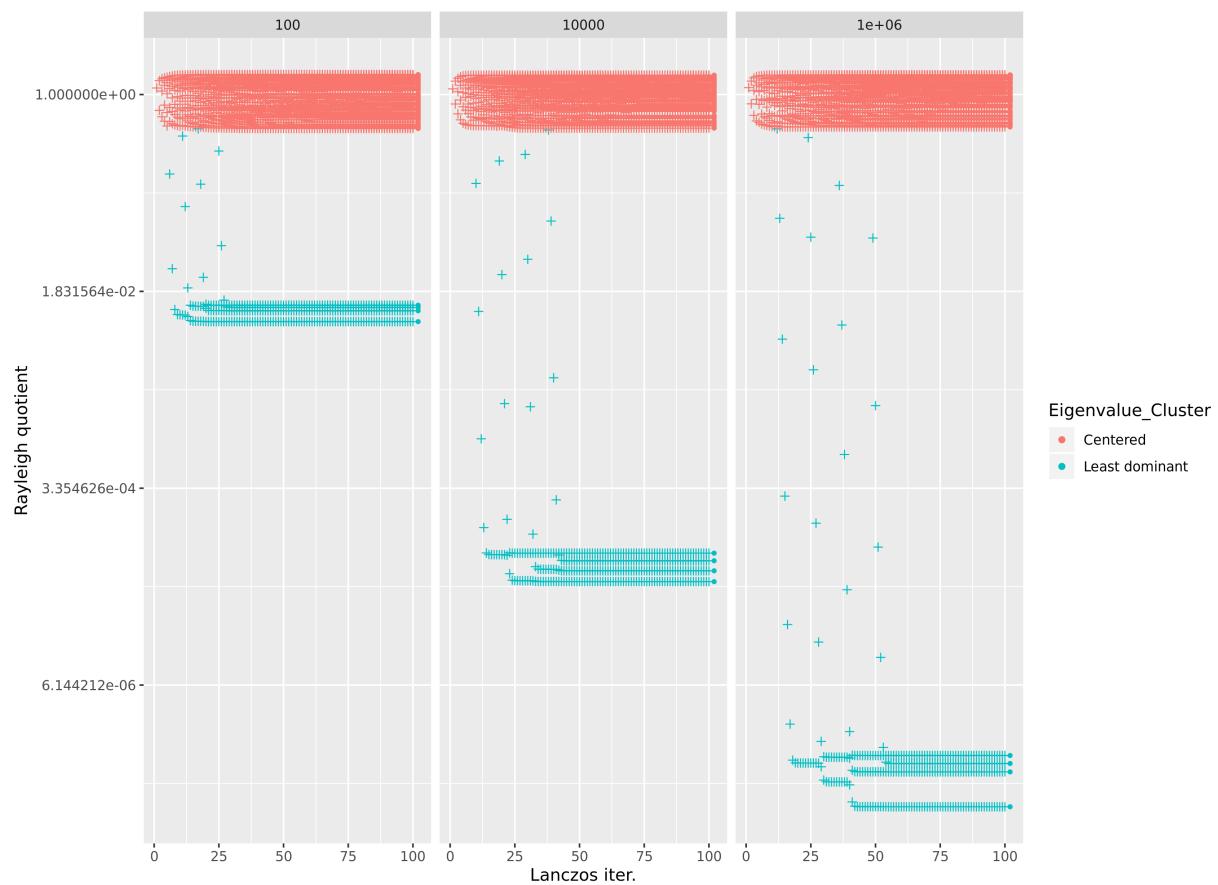
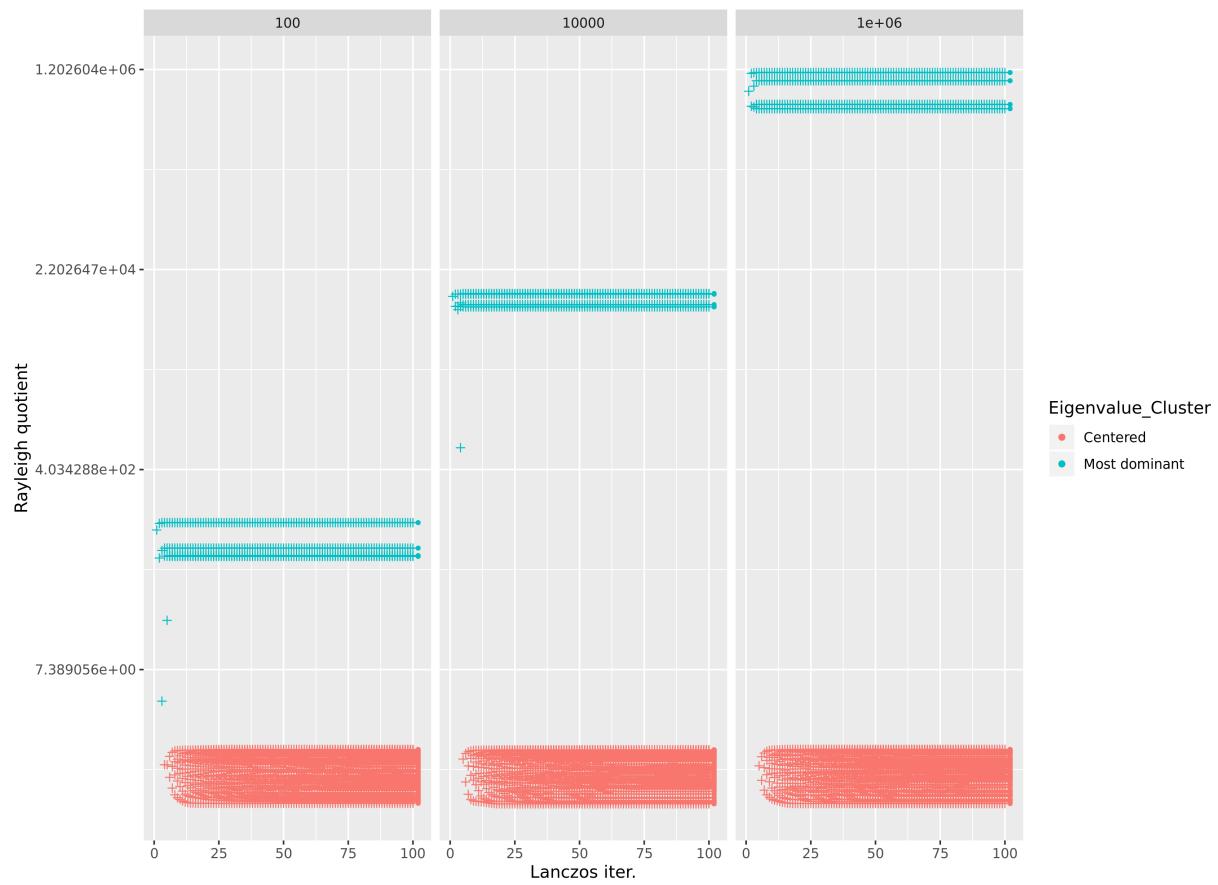


FIGURE 14 – Courbes de convergence suivant l'erreur inverse du *CG-DEF* pour la résolution successives de $\nu = 20$ systèmes, approche basée sur les approximations de HR.

(a) Matrice A_2 , dimension 100×100 (b) Matrice A_3 , dimension 100×100 **FIGURE 15** – Convergence du quotient de Rayleigh vers les valeurs propre de A_i via Lanczos Rayleigh-Ritz, cluster de taille 4.

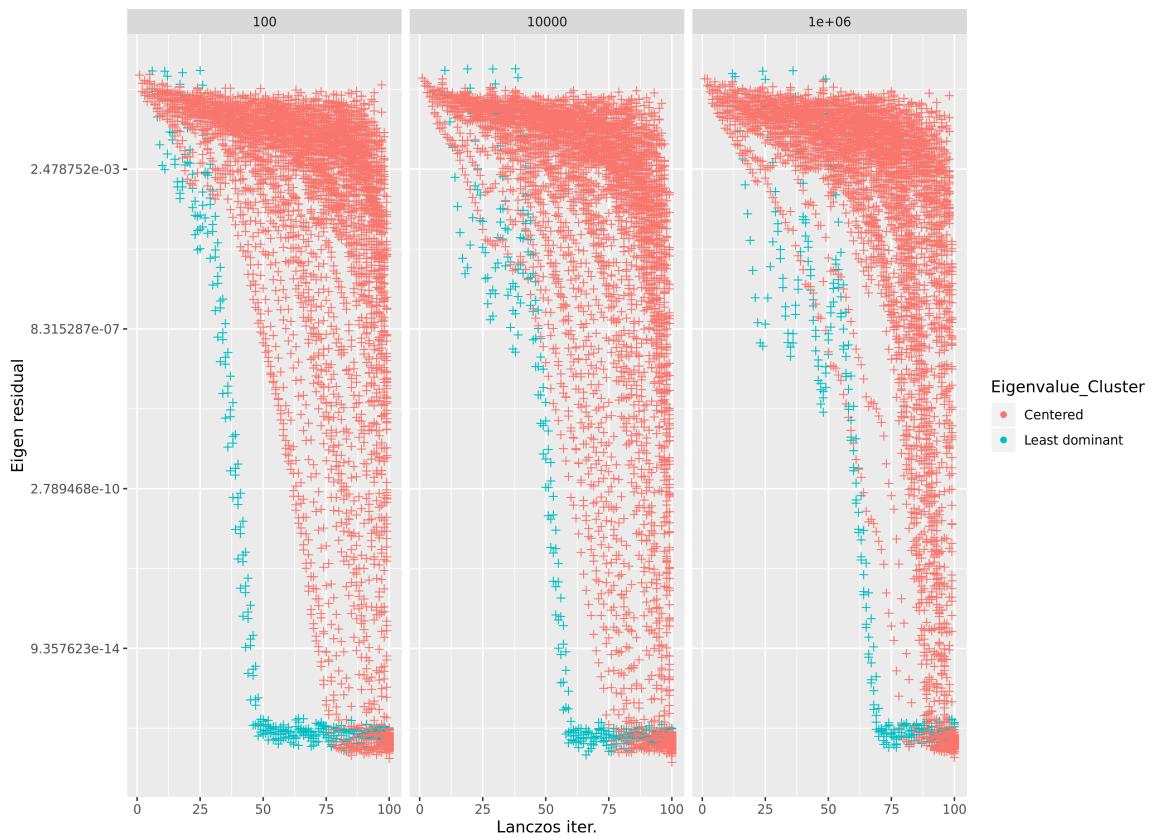
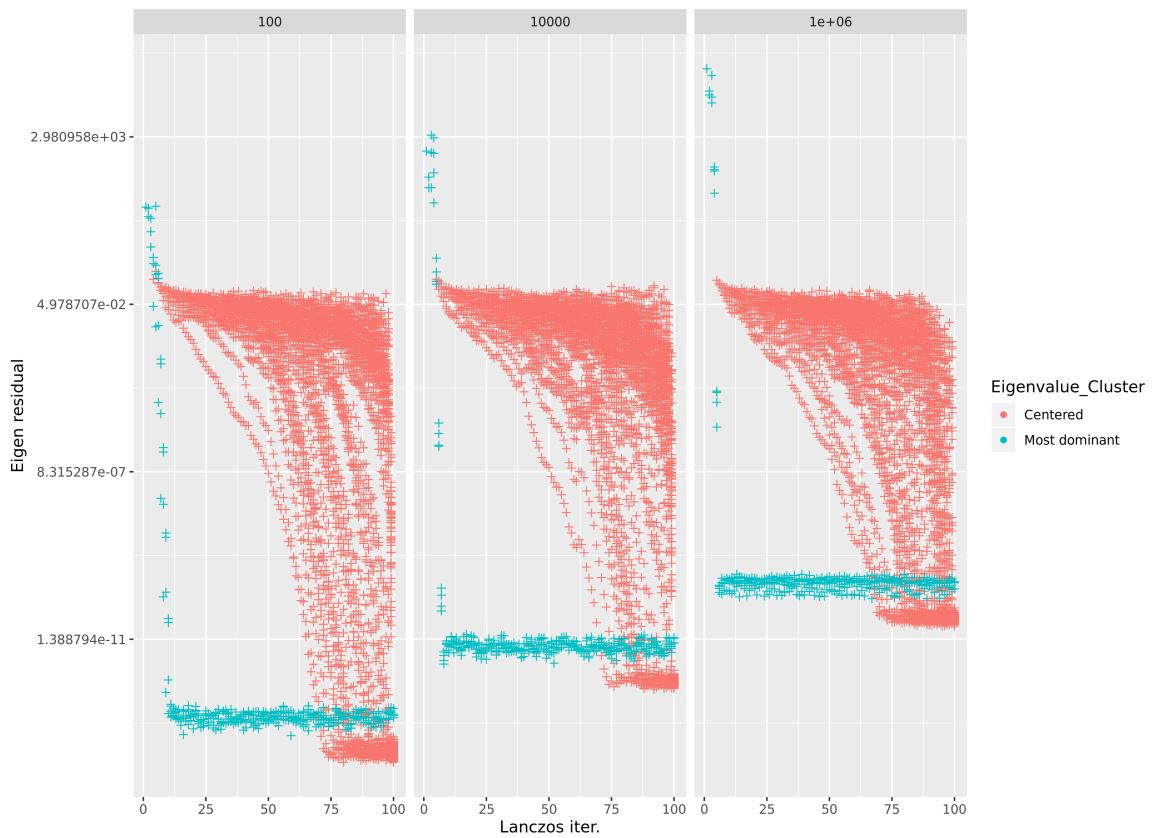
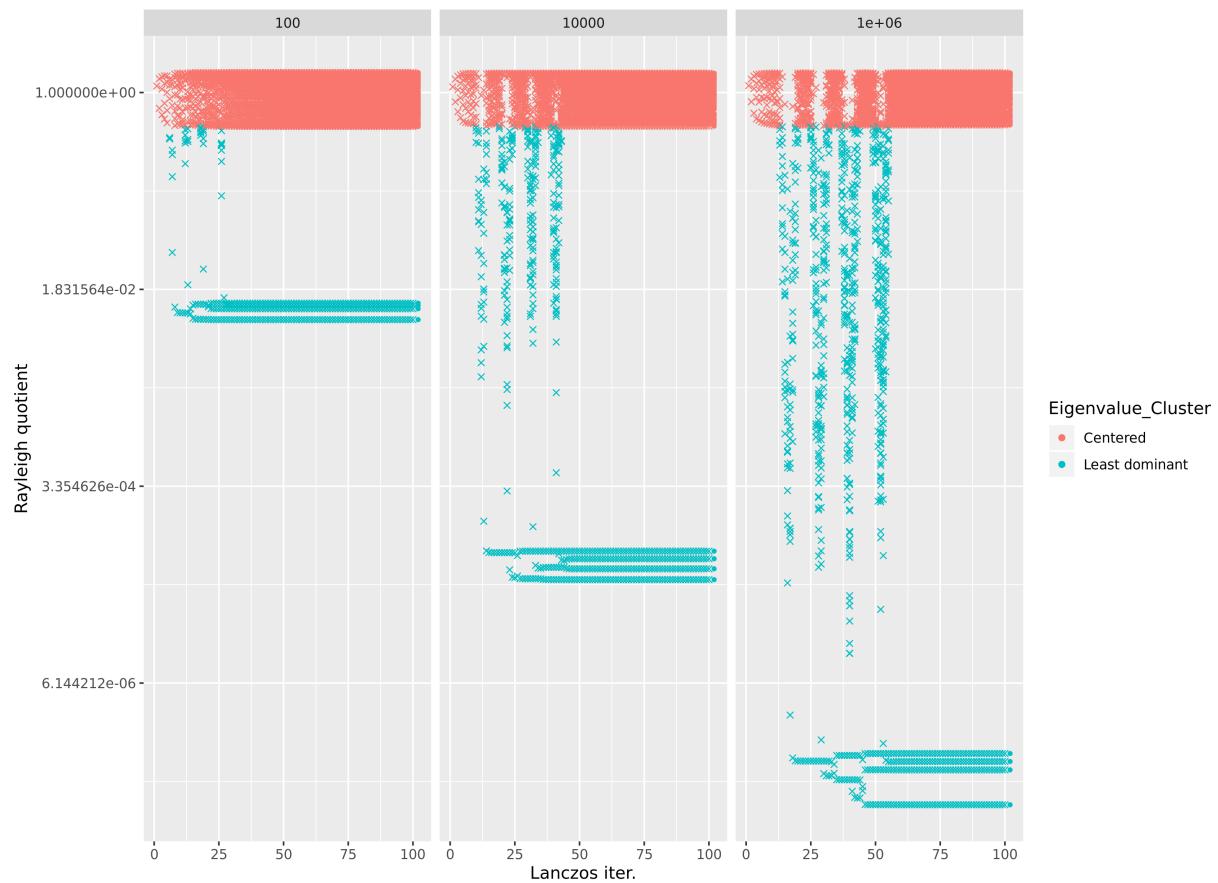
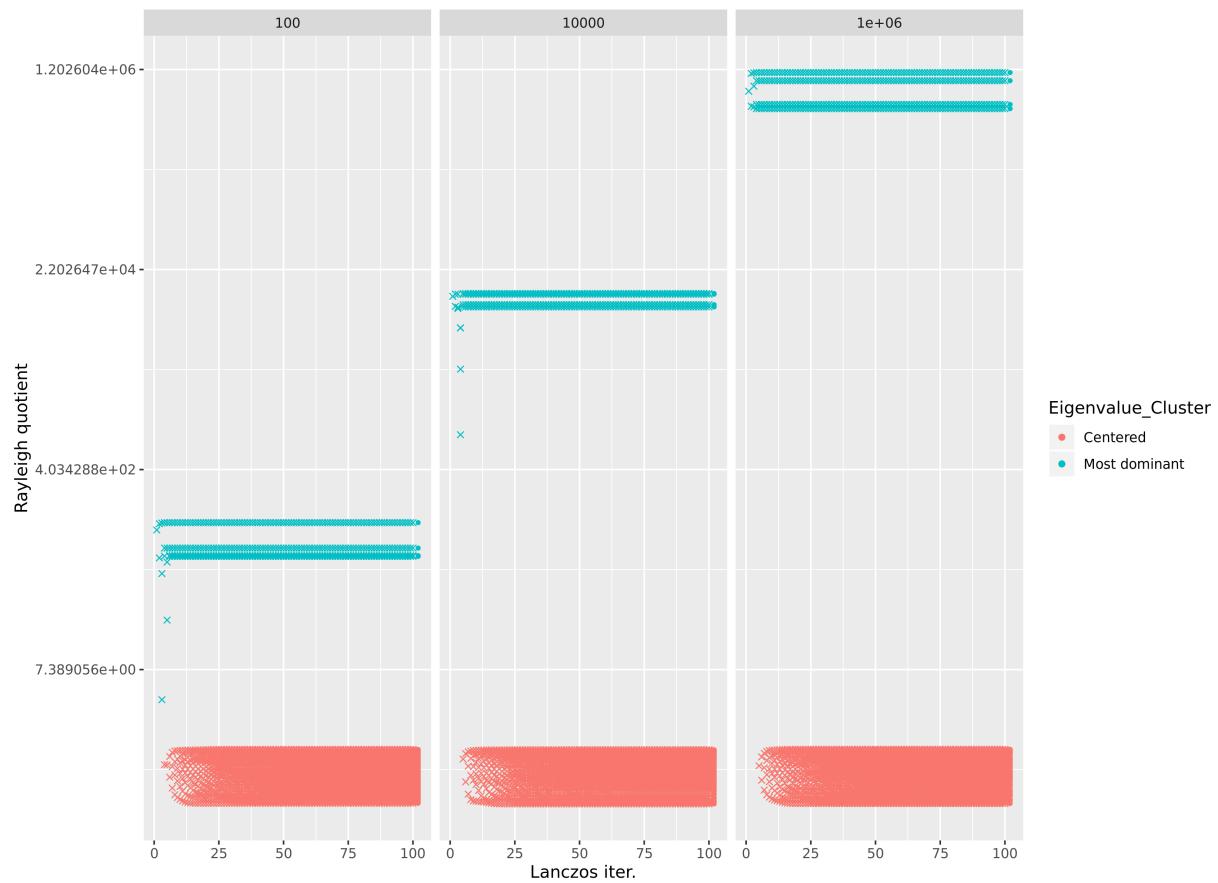
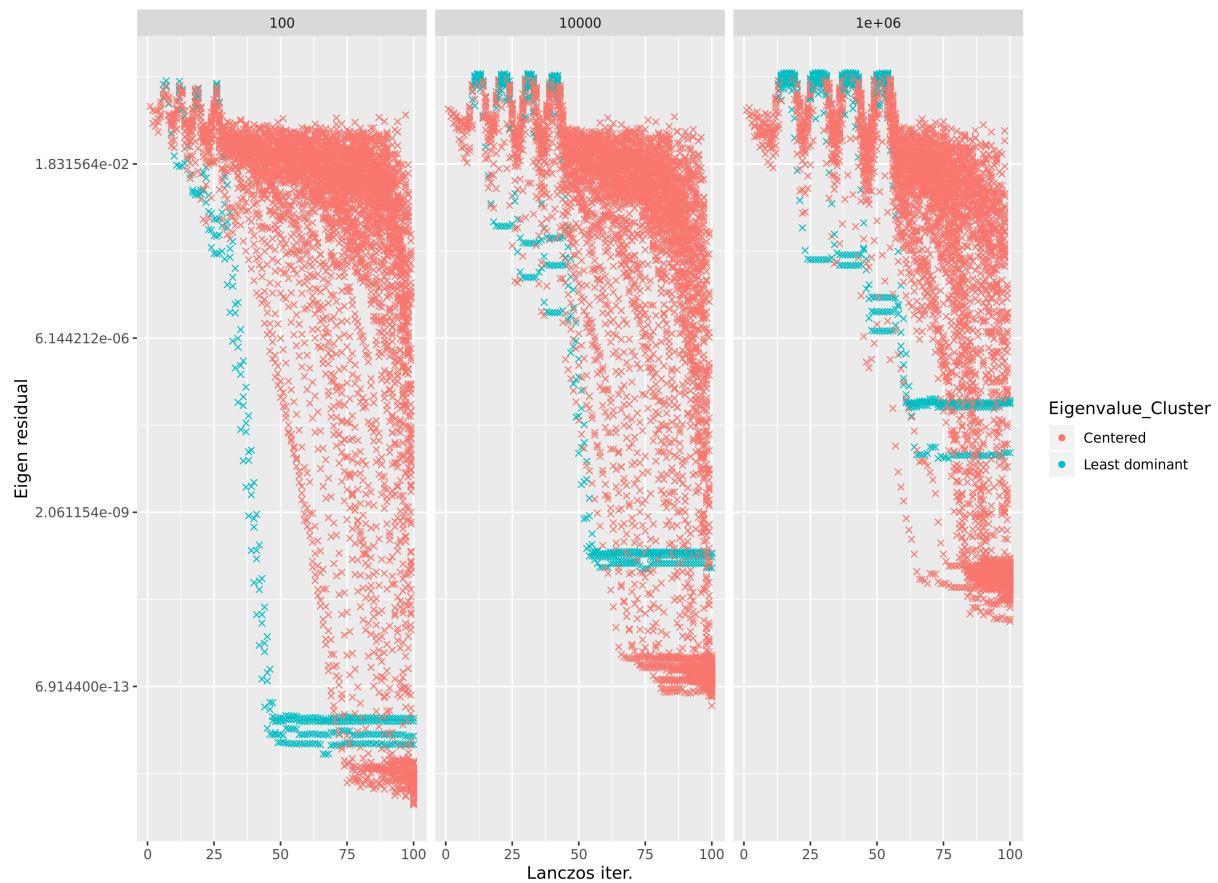
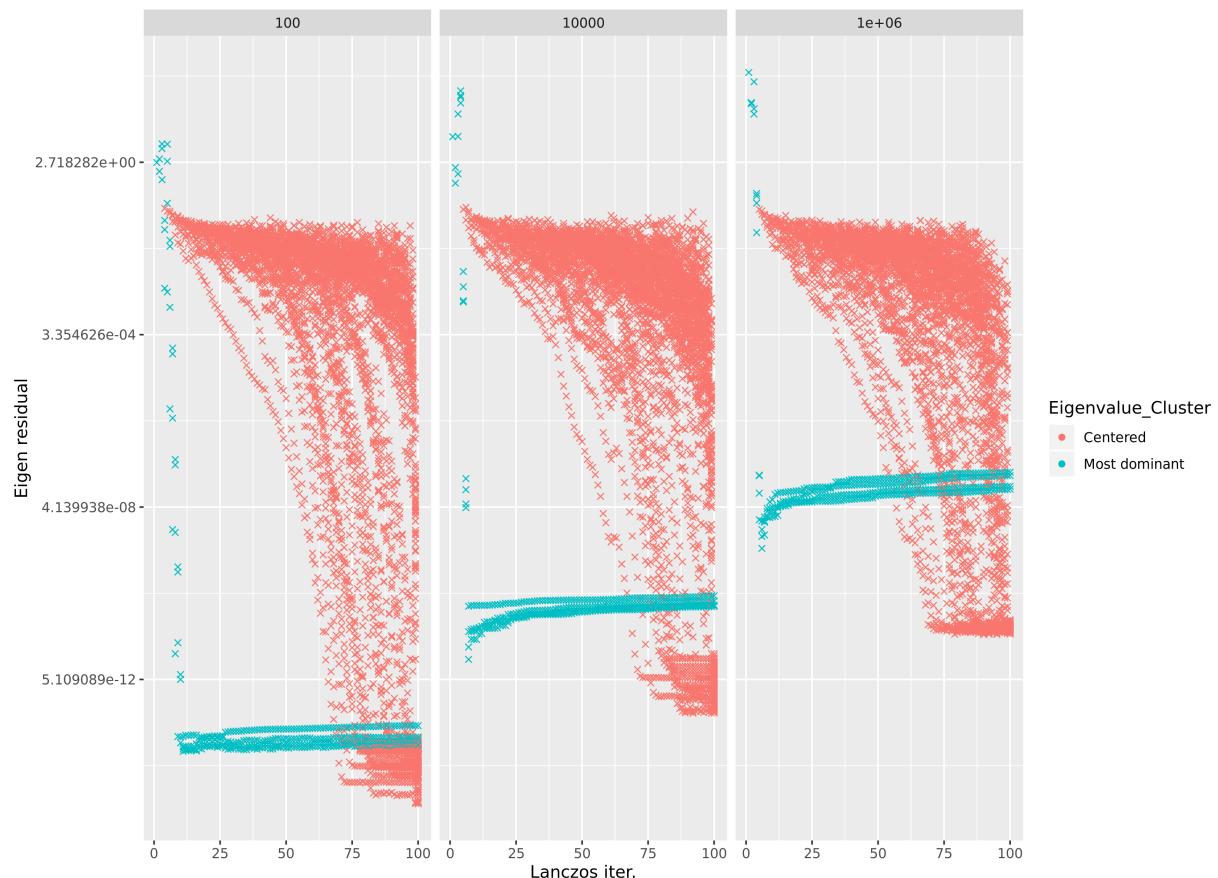
(a) Matrice A_2 , dimension 100×100 (b) Matrice A_3 , dimension 100×100

FIGURE 16 – Convergence des paires propres calculées via Lanczos Rayleigh-Ritz, cluster de taille 4.

(a) Matrice A_2 , dimension 100×100 (b) Matrice A_3 , dimension 100×100 **FIGURE 17** – Convergence du quotient de Rayleigh vers les valeurs propres de A_i via Lanczos Harmonic-Ritz.

(a) Matrice A_2 , dimension 100×100 (figures 18b)(b) Matrice A_3 , dimension 100×100 **FIGURE 18** – Lanczos Harmonic-Ritez, convergence de la paire propre, cluster de taille 4.