UNIVERSITÉ MOHAMMED V
FACULTÉ DES SCIENCES
Rabat

N° d'ordre : 2732

Thèse de Doctorat

Présentée par

Awatif ROUIJEL

Titre

Méthodes algébriques multilinéaires pour la séparation aveugle de sources : Algorithmes, simulations et application aux systèmes de communication radio mobile.

Discipline : Sciences de l'ingénieur
Spécialité : Informatique et Télécommunications

Soutenue le 25/11/2014, devant le jury composé de :

Président :

Driss ABOUTAJDINE	PES, FSR, Rabat.

Examinateurs :

Abdellah ADIB	PES, FST, Mohammedia.
M'hamed BAKRIM	PES, UCA, Marrakech.
Pierre COMON	DR CNRS, GIPSA-LAB, Grenoble.
Hamid TOUMA	PES, FSR, Rabat.
Zytoune OUADOUDI	PH, ENCG, Kenitra.

Invité :

Khalid MINAOUI	PA, FSR, Rabat.

Avant-Propos

Les travaux présentés dans ce mémoire ont été effectués au Laboratoire de Recherche en Informatique et Télécommunications (LRIT), à la Faculté des Sciences de Rabat (FSR), sous l'encadrement de Monsieur Driss ABOUTAJDINE et le co-encadrement de Monsieur Khalid MINAOUI ; au Laboratoire d'Informatique, Signaux et Système de Sophia Antipolis (I3S) et au laboratoire de Grenoble Images Parole Signal Automatique (GIPSAlab) sou l'encadrement de Monsieur Pierre COMON.

Je tiens à exprimer ma gratitude à Monsieur Driss ABOUTAJDINE, Professeur d'enseignement supérieur à la Faculté des Sciences de Rabat, Directeur du Centre National pour la Recherche Scientifique et Technique (CNRST), Responsable du Laboratoire de Recherche LRIT et Directeur de ma thèse, de m'avoir accueilli au sein de son équipe, pour ces précieux conseils, pour l'intérêt constant qu'il a porté à ce travail ainsi que pour son soutien sans faille sur les plans académiques et administratifs. Je le remercie également de m'avoir fait l'honneur de présider mon jury de thèse.

J'adresse de vifs remerciements à mon encadrant de thèse, Monsieur Pierre COMON, Directeur de Recherche CNRS, GIPSA-lab Grenoble. C'est une chance pour moi qu'un grand chercheur passionné comme lui ait accepté d'encadrer mon travail. Je le remercie infiniment pour sa rigueur scientifique, sa patience et sa vision en matière de recherche. Je le remercie également pour ses précieuses remarques, ses idées éclairées qui m'ont souvent permis de résoudre les problèmes auxquels j'étais confronté et son efficacité lors des révisions de nos articles.

Je tiens particulièrement à exprimer ma reconnaissance à Monsieur Khalid MINAOUI, professeur assistant à la Faculté des Sciences de Rabat, d'avoir co-encadré ce travail de thèse. Merci pour vos nombreux conseils, votre bonne humeur, votre disponibilité, et pour toute votre aide. Je lui témoigne toute ma reconnaissance pour la totale confiance et liberté qu'il m'a accordé depuis le début.

Je tiens également à exprimer mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail de thèse.

Merci à M. Abdellah ADIB, professeur d'enseignement supérieur à la Faculté des Sciences et Techniques de Mohemmadia, d'avoir accepté de rapporter ce travail. Ses remarques constructives m'ont beaucoup aidé à améliorer ce manuscrit.

Merci également à M. M'hamed BAKRIM, professeur d'enseignement supérieur à l'Université Cadi Ayyad à Marrakech, d'être un rapporteur de ce manuscrit. Ses suggestions lors de la lecture de ma thèse m'ont permis d'apporter des améliorations à la qualité de cette dernière.

Je tiens à remercier M. Hamid TOUMA, professeur d'enseignement supérieur à la Faculté des Sciences de Rabat, pour avoir accepté de faire partie de mon jury de thèse en tant qu'examinateur.

Merci également à M. Zytoune OUADOUDI, professeur habileté à l'École Nationale de Commerce et Gestion de Kénitra, d'avoir accepté d'examiner ce travail.

Au cours de cette thèse, j'ai bénéficié d'une bourse d'excellence octroyée par le Centre National pour la Recherche Scientifique et Technique (CNRST), dans le cadre du programme des bourses de recherche initié par le ministère de l'éducation nationale de l'enseignement supérieur, de la recherche scientifique et de la formation des cadres. Pour les stages effectués aux deux laboratoires d'accueil en France, j'ai bénéficié d'une bourse dans le cadre du programme de coopération franco-marocain STIC.

A l'issue de la rédaction de cette recherche, je suis convaincue que la thèse est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate de "l'apprenti-chercheur".

Je souligne le soutien amical et chaleureux de tous les doctorants du Laboratoire LRIT, qui ont croisé ma route durant ce parcours doctoral et, plus particulièrement, à tous ceux qui m'ont soutenu et encouragé. Bien entendu, je n'oublie pas la clique des doctorants et stagiaires que j'ai croisé tout au long de mes séjours en France.

Un merci spécial à Zakaria, mon cher binôme d'étude et de vie, qui a toujours été présent lorsque j'en ai eu besoin. Merci pour son amitié, son aide et ses encouragements.

Finalement, mais pas pour autant moins important à mes yeux, je voudrais témoigner tout mon amour et ma reconnaissance à ma chère famille, à qui je dédie cette thèse. Je vous remercie pour votre présence dans ma vie, pour vos conseils, vos encouragements, votre patience, et votre soutien plein et inconditionnel, même dans les moments les plus difficiles. Je ne pourrai jamais vous remercier assez.

Résumé

La séparation aveugle de sources consiste à estimer des signaux inconnus à partir d'un mélange observé de ces signaux, sans information sur ces signaux et/ou le mélange. Les premiers travaux sur la séparation aveugle de sources ont été initiés par C.Jutten et J.Hérault dans le cas d'un mélange instantané. Plus récemment, l'utilisation de méthodes d'algèbre multilinéaire a attiré l'attention dans plusieurs domaines tels que le data-mining, le traitement du signal et en particulier dans les systèmes de communication sans fil. En effet, dans plusieurs applications de traitement du signal pour les systèmes de communication sans fil, le signal reçu est de nature multidimensionnelle et possède une structure algébrique multilinéaire. Résoudre le problème de la séparation de sources revient alors à trouver une décomposition du tenseur des observations et de déterminer ses paramètres. Une des décompositions les plus populaires de tenseur est la décomposition canonique polyadique (CP). Cette décomposition est connue généralement sous différentes terminologies : "Canonique Polyadique" (CP en anglais), "CandDecomp", ou encore "PARAFAC". Elle peut être vu comme un analogue de la décomposition des matrices en valeurs singulières (SVD), car elle décompose le tenseur en une somme minimale de tenseurs de rang un. L'intérêt de cette décomposition réside dans son unicité sous certaines conditions. Dans la littérature, plusieurs algorithmes typiques de calcul des composantes de la décomposition CP ont été conçus, parmi lesquels on trouve l'algorithme des moindres carrés alternés et les algorithmes de descentes. Cependant, ces algorithmes ne traitent pas la matrice de facteur d'échelle.

Cette thèse traite les problèmes de modélisation des systémes multi-utilisateurs, de séparation et d'estimation de signaux de différents systèmes à l'aide d'approches tensorielles. Dans un premier temps, une décomposition tensorielle basée sur la décomposition CP a été proposée et étudiée, et qui résout le problème de l'indétermination d'échelle. De plus, il est bien connu que les matrices facteurs sont identifiées à une indétermination d'échelle près. Cette indétermination est compliquée à prendre en compte, étant donné que le produit de tous les facteurs d'échelle doit être égal à la matrice identité. Pour cette raison, seuls les indices de performances approximatifs ont été utilisés jusqu'à présent, en ignorant la dernière contrainte. Par ailleurs, il s'est avéré alors intéressant de trouver un nouveau critère de performance qui peut être calculé d'une manière exacte et qui prend en considération la contrainte d'indétermination d'échelle et de permutaion inhérente au problème de séparation de sources.

Dans un deuxième temps, divers algorithmes d'optimisation ont été étudiés et deux versions de l'algorithme de gradient ont été proposées afin de pouvoir prendre en compte

la contrainte d'égalité et la matrice de facteur d'échelle. À des fins d'évaluation, les différentes solutions proposées ont ensuite été testées sur des mélanges de trois systèmes de communication radio mobile. Nous avons montré que le modèle analytique des signaux CDMA, IDMA et la technique proposée OWDM-IDMA peut s'écrire sous forme d'un tenseur des observations dont les signaux peuvent être estimé d'une manière aveugle en s'appuyant sur les algorithmes d'optimisations proposés.

Mots clés *:* Modélisation tensorielle, systèmes de communication sans fil, séparation aveugle, système multi-utilisateurs, décomposition CP, critère de performance, CDMA, IDMA, OWDM-IDMA.

Abstract

Blind source separation consists in estimating unknown observed signals from their mixture without any prior information about them, except some properties such as their independence. Early work on blind source separation was initiated by C.Jutten and J.Hérault in the case of an instantaneous mixture. More recently, the use of multi-linear algebra methods has attracted attention in several areas such as data mining, signal processing and particularly in wireless communication systems, among others. Indeed, in many applications of signal processing for wireless communication systems, the received signal is multidimensional in nature and has a multilinear algebraic structure. Solving the problem of source separation means to find a decomposition of the observations tensor and determin its parameters. One of the most popular tensor decompositions is the Canonical Polyadic decomposition (CP), also known as Parallel Factor Analysis (PARAFAC), which can be seen as an analogue of the matrix Singular Value Decomposition (SVD), since it decomposes the tensor into a sum of rank-one components. The interest of the CP decomposition lies in its uniqueness under certain conditions. Typical algorithms for finding the CP components include alternating least squares (ALS) and descent algorithms, which do not isolate the scaling factor matrix.

This thesis handles the problems of multi-users systems modeling as well as the separation and the estimation of signals from different systems, using the tensor approaches. First of all, a tensor decomposition based on the CP decomposition has been proposed and studied, which solves the problem of scale indeterminacy. Moreover, it is well known that factor matrices are identified up to column scaling. This indeterminacy is complicated to take into account, given that the product of all scaling matrices must be equal to the identity one. For this reason, only approximate performance indices have been used so far by ignoring the last constraint. So, it proved interesting to find a new performance criterion which can be calculated in an accurate and which can takes into account the constraint of scale and permutation indeterminacies.

Secondly, various optimization algorithms have been studied and two versions of the gradient algorithm have been proposed in order to take into account the constraint equality and the scaling factor matrix. For evaluation purposes, different proposed solutions were tested on mixtures of three mobile radio communication systems. We showed that the analytical model of CDMA, IDMA and proposed IDMA-OWDM techniques, can be written as an observations tensor whose signals can be blindly estimated using the proposed optimizations algorithms.

Keywords *:* Tensor modeling, wireless communication systems, blind separation, multi-users system, CP decomposition, performance criterion, CDMA, IDMA, OWDM-IDMA.

Table des matières

Liste des acronymes

1G	*Première Génération*
2G	*Deuxième Génération*
3G	*Troisième Génération*
3MFA	*Three-mode factor analysis*
3MPCA	*Three-mode principal components analysis*
4G	*Quatrième Génération*
ACI	*Analyse en Composantes Indépendantes*
ACP	*Analyse en Composantes Principales*
ADSL	*Asymmetric Digital Subscriber Lines*
ALS	*Alternating Least Squares*
APP	*A Posteriori Probability*
AWGN	*Additive White Gaussian Noise*
BER	*Bit Error Rate*
BFGS	*Broyden-Fletcher-Goldfarb-Shanno*
BPSK	*Binary Phase Shift Keying*
CANDECOMP	*Canonical Decomposition*
CCDF	*Complementary Cumulative Distribution Function*
CDMA	*Code Division Multiple Access*
CF	*Crest Factor*
CG	*Conjugate Gradient*
cI CDMA	*chip Interleaved Code Division Multiple Access*
CP	*Canonical Polyadic Decomposition*
CWT	*Continuous Wavelet Transform*
DEC	*Decoder*
DFP	*Davidson-Fletcher-Powell*
DFT	*Discrete Fourier Transform*
DWPT	*Discrete Wavelet Packet Transform*
DWT	*Discrete Wavelet Transform*
EEM	*Error Estimation Matrix*
ELS	*Enhanced Line Search*
ESE	*Elementary Signal Estimator*
EVD	*Eigenvalue Decomposition*
FDMA	*Frequency Division Multiple Access*
FEC	*Forward Error Correction*

GP	*Gradient Projeté*
GSM	*Global System for Mobile communication*
HOSVD	*Higher-Order Singular Value Decomposition*
IDMA	*Interleave Division Multiple Acces*
IDWPT	*Inverse Discrete Wavelet Packet Transform*
ISI	*Inter Symbol Interference*
LLR	*Logarithm Likelihood Ratio*
LM	*Levenberg-Marquardt*
LS	*Line Search*
MAI	*Multiple Access Interference*
MAP	*Maximum A Posteriori*
MIMO	*Multi-Input Multi-Output*
MRC	*Maximum Ratio Combining*
MUD	*Multiuser Detection*
NLCG	*Non Linear Conjugate Gradient*
OFDM	*Orthogonal Frequency Division Multiplexing*
OVSF	*Orthogonal Variable Spreading Factor*
OWDM	*Orthogonal Wavelet Division Multiplex*
PAPR	*Peak-to-Average Power Ratio*
PARAFAC	*Parallel Factor Analysis*
PMEPR	*Peak to Mean Enveloppe Power Ratio*
PN	*Pseudo-Noise*
QPSK	*Quadrature Phase Shift Keying*
RMS	*Root Mean Square*
SIMO	*Single-Input Multi-Output*
SNR	*Signal-to-Noise Ratio*
SVD	*Singular Values Decomposition*
TDMA	*Time Division Multiple Access*
UMTS	*Universal Mobile Telecommunications System*
UTRA	*UMTS Terrestrial Radio Access*

Liste des figures

Liste des tableaux

Chapitre

1 Introduction

le secteur des communications à distance, plus connu sous le terme de télécommunication, connaît une croissance fulgurante grâce aux progrès technologiques réalisés dans plusieurs domaines scientifiques. Cette évolution est particulièrement frappante pour les communications radio-mobiles avec l'apparition des différentes générations de téléphonie mobile. Parallèlement, les applications pouvant bénéficier de cette évolution technologique n'ont cessé de se diversifier.

Ainsi, nous assistons actuellement à l'avènement de la visiophonie et du visionnage de signaux audiovisuels sur des appareils de téléphonie mobile. En traitement du signal pour les télécommunications sans fil, ces nouvelles fonctionnalités nécessitent des transmissions de plus en plus rapides ; garantissant à la fois une nécessaire flexibilité et une impérieuse efficacité au niveau de la qualité de service. Pour ce faire, les nouveaux réseaux de télécommunications doivent permettre l'accès simultané, d'utilisateurs toujours plus nombreux, aux multiples services proposés par les différents opérateurs de téléphonie mobile. Les bandes de fréquences de transmission étant limitées, l'efficacité de l'utilisation des ressources de transmission s'avère primordiale.

Plusieurs techniques de multiplexage ont été proposées pour une utilisation à bon escient de la bande de fréquence disponible. La technique d'accès multiple à répartition en fréquence, ou plus communément FDMA (*Frequency Division Multiple Access*) est la méthode de partage de ressource spectrale la plus ancienne. Elle consiste à allouer à chaque utilisateur une bande de fréquence différente pour permettre des transmissions simultanées. Une autre technique permet à tous les utilisateurs d'émettre sur l'ensemble de la bande de fréquence mais successivement dans le temps. Il s'agit de la technique d'accès multiple à répartition en temps, appelée TDMA (*Time Division Multiple Access*). Généralement, les techniques FDMA et TDMA sont combinées pour une meilleure exploitation de la bande de fréquence. En effet, l'inconvénient majeur de ces deux techniques est la difficulté de gérer l'ensemble de la bande de fréquence le plus optimale possible. D'une part, avec la technique FDMA, si au cours du temps un utilisateur n'émet pas de signal, alors la bande de fréquence qui est allouée n'est pas utilisée. D'autre part, avec la technique TDMA, si un utilisateur n'émet pas durant l'instant qui lui attribué, alors l'intervalle de temps qui lui correspond n'est pas utilisé. C'est pourquoi, une autre technique a fait son apparition. Il s'agit de la technique d'accès multiple à répartition en code, appelée CDMA (*Code Division Multiple Access*). Cette technique permet à tous les utilisateurs de transmettre simultanément dans une même bande de fréquence et en même temps. Ainsi, toutes les ressources disponibles sont exploitées de manière optimale. Par contre, la diffi-

culté réside dans la séparation des signaux de chaque utilisateur. Pour ce faire, et dans les systèmes dits coopératifs, un code d'étalement spécifique est alloué à chaque utilisateur. Lors de la réception, ces codes servent à distinguer les signaux émis par les différents utilisateurs. Cette technique de multiplexage est désormais privilégiée par les standards de téléphonie mobile. Ainsi, des normes radio-mobiles telles que la norme américaine CDMA 2000 et la norme européenne UMTS (*Universal Mobile Telecommunications System*) l'ont retenu comme technique d'accès. Au niveau technologique, les performances optimales d'un système mono-utilisateur sont atteintes par des systèmes multi-utilisateurs de type CDMA en assignant aux différents utilisateurs des codes orthogonaux entre eux. Les codes de Hadamard qui ont cette propriété d'orthogonalité sont généralement utilisés pour des transmissions synchrones. Ce type de transmission est représentatif d'une communication sur la voie descendante d'un réseau de télécommunication mobile où la station de base émet vers les différents terminaux mobiles. A l'inverse, lors d'une transmission sur la voie montante, pour laquelle les différents terminaux mobiles émettent indépendamment vers la même station de base, la transmission est dite asynchrone. Dans ce second cas, des codes non orthogonaux ayant des inter-corrélations très faibles, comme les codes de Gold, peuvent être utilisés. Néanmoins, cette propriété d'orthogonalité nécessaire à la séparation des signaux dans les systèmes CDMA constitue l'une des contraintes majeures de la technique CDMA.

En 2002, l'équipe de Li Ping a proposé une nouvelle technique d'accès multiple baptisée IDMA (*Interleave Division Multiple Acces*) (Ping *et al.*, 2007, 2002). Cette technique s'avère être un cas particulier de la technique CDMA. En effet, les utilisateurs sont distingués à l'aide d'entrelaceurs et non plus de codes orthogonaux comme ce qui est le cas de la technique CDMA. La technique IDMA bénéficie alors de plusieurs avantages de la technique CDMA, notamment, des qualités de l'étalement de spectre. En effet, l'un des avantages de la technique d'étalement de spectre est le fait qu'elle soit robuste face aux différents types de brouillage. De plus, ses propriétés d'auto-corrélation permettent de tirer partie au mieux de la diversité des canaux multi-trajets à évanouissements. La principale caractéristique de la technique IDMA est la possibilité d'utiliser pour tous les utilisateurs un même code d'étalement. La distinction des signaux des différents utilisateurs se fait alors par des entrelaceurs spécifiques. Les performances de ce système, constitué d'entrelaceurs générés de façon aléatoire présente, sont proches de la limite théorique d'un système multi-utilisateurs (Ping *et al.*, 2006). C'est la raison pour laquelle la technique IDMA a tout de suite attiré l'attention de la communauté scientifique. Les derniers travaux publiés sur le sujet laissent à penser qu'il s'agit d'une technique d'accès prometteuse pour les futurs systèmes de télécommunication.

À la réception, la séparation et l'égalisation se fait en utilisant les codes d'étalement et les séquences d'apprentissage dans le cas d'un système CDMA, ou des entrelaceurs et le code d'étalement dans le cas d'un système IDMA. De ce fait, 25% du débit total disponible est consacré à l'apprentissage en GSM et jusqu'à 50% en UMTS. Dans ce contexte, les méthodes de séparation aveugle suscitent un vif intérêt en traitement de signal. Elle consiste à estimer des signaux inconnus à partir d'un mélange de ces signaux, sans informations ni sur les signaux ni sur le mélange, c-à-d sans utiliser la connaissance a priori des codes d'étalement ni celle des séquences d'apprentissage à la réception.

Il existe différentes méthodes pour retrouver les sources à partir du mélange. Ces méthodes sont classifiées en deux approches : l'approche purement statistique et l'approche purement déterministe. On peut par exemple s'appuyer sur l'hypothèse que les sources appartiennent à un alphabet fini (Godard, 1980; der Veen et Paulraj, 1996). Il est également possible de s'appuyer sur l'hypothèse que les sources sont mutuellement indépendantes, on parle dans ce cas d'analyse en composantes indépendantes (ACI) (Belouchrani *et al.*, 1997; Comon, 1994), qui représente le concept fondamental de l'approche statistique. En ce qui concerne l'approche déterministe, elle repose quant à elle sur la structure algébrique des signaux eux-mêmes et non pas la structure algébrique de leurs statistiques. Plus récemment, les méthodes d'algèbre multilinéaire ont retenu une attention particulière (De Lathauwer, 2006; Sidiropoulos et Bro, 2000; Comon, 2000). Les données du problème, appelées aussi les observations, peuvent en effet dans certains cas être regardées comme les éléments d'un tenseur d'ordre supérieur à trois. Il existe une décomposition des tenseurs qui propose de décomposer un tenseur sous la forme d'une somme minimale de tenseurs de rang un. Cette décomposition a été introduite de manière indépendante sous le nom de CANDECOMP (*Canonical Decomposition*) en psychométrie (Carroll et Chang, 1970) et de PARAFAC (*Parallel Factor Analysis*) en phonétique (R.Harshman, 1970), mais la términologie la plus répandue est le CP (*Canonical Polyadic Decomposition*) (Comon *et al.*, 2009). Elle a ensuite été utilisée dans des domaines variés, comme la chimiométrie (BRO, 1997), ou les télécommunications (Sidiropoulos *et al.*, 2000b,a). Résoudre le problème de séparation de sources revient alors à déterminer les paramètres de la décomposition. La caractèristique la plus intéressante de la décomposition CP est son originalité intrinsèque. Car, et contrairement à la décomposition matricielle, où le problème de liberté de rotation se manifeste, la décomposition CP des tenseurs est essentiellement unique, à l'indétermination de l'échelle et de permutation (KRUSKAL, 1977; Stegeman et Sidiropoulos, 2007). La première preuve d'unicité a été fournie par Kruskal dans (KRUSKAL, 1977). Cette preuve a été reformulée en utilisant l'algèbre linéaire de base (Stegeman et Sidiropoulos, 2007). Une preuve concise valable pour les tenseurs complexes a été donnée dans (Sidiropoulos *et al.*, 2000b). L'algorithme traditionellement utilisé pour déterminer les paramètres de la décomposition est un algorithme des moindres carrés alternés. On peut se demander alors, s'il n'existe pas d'autres techniques, plus fiables ou plus rapides pour identifier les paramètres de la décomposition. D'autre part, Il est bien connu que les matrices facteurs sont identifiées à une indétermination d'échelle près. Cette indétermination est compliquée à prendre en compte, étant donné que le produit de tous les facteurs d'échelle doit être égal à la matrice identité. Pour cette raison, seuls les indices de performance approximatifs ont été utilisés jusqu'à présent en ignorant la dernière contrainte. Cependant, on peut se demander s'il est possible de calculer l'indice de performance exacte ?

Objectifs

Le premier objectif de notre étude est de se familiariser avec les différentes notions nécessaires à la compréhension des aspects théoriques autour des techniques d'accès multiples CDMA et IDMA. Pour ce faire, une étude bibliographique sur la technique IDMA et plus globalement un état de l'art sur la technique d'accès multiples CDMA ont été

effectués. Une étude de la technique IDMA dans le contexte d'un canal mono-trajet théorique a été effectuée pour la compréhension du fonctionnement du système. Une analyse de la capacité du récepteur à traiter à la fois les interférences d'accès multiples et le bruit de transmission a alors été réalisée.

Récemment, il a été montré que la technique hybride OFDM-IDMA (Mahafeno *et al.*, 2006; Zhou *et al.*, 2005) est une approche prometteuse pour résoudre les deux majeurs obstacles à la communication sans fil, à savoir, l'interférence inter-symboles ISI et l'interférence d'accès multiple MAI. Cette technique peut atténuer l'ISI en utilisant la technique OFDM et supprimer la MAI par la détection multi-utilisateurs (MUD) de la technique IDMA (Wang *et al.*, 2006; Ping *et al.*, 2006). Cette technique utilise la forme d'onde rectangulaire comme une mise en forme de filtre. Ce choix a l'avantage de permettre une implémentation efficace de la modulation et de la démodulation en utilisant la DFT (*Discrete Fourier Transform algorithms*). Cependant, la forme d'onde rectangulaire n'est pas optimale dans une transmission radio-électrique. En effet, cette fonction est mal localisée en fréquence et sensible à la dispersion temporelle du canal de propagation. Récemment, l'utilisation du préfixe cyclique a été proposée comme solution de ce problème, mais il peut entraîner une perte d'efficacité spectrale. L'autre inconvénient majeur du signal OFDM est sa grande fluctuation d'enveloppe, ce qui peut dégrader le rendement des amplificateurs de puissance des émetteurs en les obligeant à fonctionner à une faible puissance moyenne (Armstrong, 2002). Ce phénomène se quantifie par le facteur de crête, couramment appelé PAPR (*Peak To Average Power Ratio*), qui représente le rapport entre l'amplitude du pic du signal et sa valeur efficace. La consommation de puissance de l'amplificateur dépend en grande partie de la puissance de crête. En effet, l'amplificateur devra être dimensionné par rapport au pic du signal. Cependant, la valeur efficace du signal est la valeur "utile", celle qui va vraiment caractériser la puissance transmise, et donc la valeur à maximiser. Donc, le PAPR va être le gain perdu entre la puissance maximale de l'amplificateur et la puissance réellement transmise. Pour cette raison, il est important de minimiser le PAPR, ce qui permet d'avoir des amplificateurs dimensionnés au plus juste par rapport à la puissance à transmettre. Notre deuxième objectif a été alors de proposer une nouvelle technique d'accès multiple, que nous avons baptisé OWDM-IDMA comme une combinaison de la techniques OWDM et IDMA. Cette technique est suggérée pour résoudre le problème de la forme d'onde rectangulaire dans la modulation OFDM (Rouijel *et al.*, 2011a, 2012). En effet, par le formalisme des ondelettes et de leur application aux paquets d'ondelettes et bancs de filtres, il est possible de construire une base orthogonale en temps et en fréquence dont les propriétés peut développer un système de communication multiporteuse en utilisant avec plus de précision le spectre radio-électrique. De plus, l'utilisation de la technique IDMA peut rendre le système plus robuste contre les interférences MAI, et conduire à une amélioration significative des performances.

Dans un deuxième temps, au vu de la perte de débit causée par l'utilisation des séquences d'apprentissage et les codes d'étalement, et ces derniers et les entrelaceurs dans le cas des systèmes CDMA et IDMA respectivement, il s'est avéré intéressant de se placer dans un contexte de séparation aveugle des signaux de chacune de ces deux systèmes. pour ce faire, une étude d'une approche purement déterministe, qui est la décomposition tensorielle, a été faite. La séparation et l'égalisation des signaux reçus en utilisant la

décomposition tensorielle CP nous a amené à nous intéresser d'une part aux algorithmes utilisés pour effectuer cette décomposition et à proposer de nouveaux algorithmes de la décomposition tensorielle, et d'autre part au calcul d'un critère de performance exacte qui prend en compte la contrainte d'indétermination d'échelle et de permutation.

Organisation du document

Les études et contributions que nous avons effectué durant les années de thèse sont présentées dans ce manuscrit en quatres chapitres. Dans ce qui suit, nous décrivons brièvement le contenu de chaque chapitre.

Dans le chapitre 2, nous introduisons les notations et les définitions d'algèbre multilinéaire qui seront utilisées dans ce document. La définition mathématique d'un tenseur d'ordre supérieur, les propriétés propres à leur utilisation ainsi que les opérateurs tensoriels manipulés dans ce manuscrit, sont alors exposés.

Dans le chapitre 3, nous présenterons certaines méthodes de décomposition tensorielle. Une plus grande attention sera accordée aux décompositions d'un tenseur de troisième ordre, puisque ça sera le cas dans la plupart des applications étudiées dans cette thèse. Puis, nous introduirons notre première contribution, dans laquelle nous proposerons une décomposition tensorielle basée sur la décomposition CP, et qui résout le problème de l'indétermination de l'échelle. Une étude de l'existence et de l'unicité de cette décomposition sera alors effectuée. Ensuite, nous proposerons un nouveau critère de performance qui est calculé d'une manière exacte et qui prend en considération la contrainte d'indétermination d'échelle et de permutation. C'est ce qui représentera notre deuxième contribution.

Dans le chapitre 4, nous proposerons plusieurs algorithmes itératifs pour le calcul des décompositions tensorielles. Tout d'abord, nous détaillerons les algorithmes généralement employés pour déterminer les paramètres de la décomposition. L'algorithme ALS standard, fréquemment utilisé dans la littérature, sera d'abord développé, puis l'insertion d'une recherche linéaire dans cet algorithme sera montrée. Ensuite, nous proposerons deux algorithmes itératifs basés sur la descente de gradient et qui prennent en considération la contrainte d'égalité présentée au chapitre précedent. Le problème du choix du pas de recherche sera également traité.

Le chapitre 5 sera consacré à la séparation et à l'égalisation des signaux CDMA, IDMA et OWDM-IDMA. Tout d'abord, une chaîne de transmission numérique de base sera décrite et les différents modèles de canaux, choisis pour cette étude, seront définis. Puis, nous donnerons un bref rappel des techniques d'accès multiple, et nous rappelerons le modèle de transmission des signaux CDMA dans le cas coopératif. Ensuite, nous montrerons que le modèle analytique de ces signaux peut s'écrire sous forme d'un tenseur des observations et nous estimerons les signaux de manière aveugle en s'appuyant sur les algorithmes d'optimisation proposés dans le chapitre précédent.

Dans un second temps, nous présenterons la technique IDMA qui est un cas particulier de la technique CDMA, et nous étudierons son fonctionnement. Dans cette étude, des canaux mono-trajets théoriques seront alors considérés. En premier lieu, le principe de l'émetteur et celui du récepteur IDMA seront successivement décrits. L'algorithme de détection multi-utilisateurs ainsi que le processus itératif à la réception seront alors dé-

taillés. Puis, nous proposerons une séparation aveugle des signaux IDMA. Nous verrons que la structure particulière de ces signaux, obtenue grâce à l'étalement de spectre et de la nouvelle forme des matrices d'entrelacement que nous proposerons dans ce chapitre, nous permet d'utiliser la technique développée au chapitre précédent. En effet, la repésentation tensorielle des signaux IDMA, que nous proposerons dans ce chapitre, nous permettera d'estimer les signaux reçus en décomposant le tenseur des observations. Pour ce faire, les algorithmes de décomposition proposés et présentés auparavant, et qui joueront ici le rôle d'un récepteur aveugle, seront alors utilisés.

La dernière partie de ce chapitre sera dédiée au système proposé, combinant la technique multi-porteuses OWDM avec la technique IDMA, que nous baptiserons OWDM-IDMA. Les principes de l'émetteur et du récepteur du système OWDM-IDMA seront tout d'abord décrits et l'étude des performances sera effectuée. Ensuite, une modélisation analytique et tensorielle des signaux OWDM-IDMA sera alors présentée.

Enfin, nous terminerons ce mémoire par une conclusion générale synthétisant les apports méthodologiques et applicatifs, et nous proposerons quelques perspectives pour la poursuite de ce travail.

Ce manuscrit est complété par des annexes où le lecteur pourra trouver des informations complémentaires.

Contributions

Les contributions de cette thèse porteront sur les trois principaux axes de recherche suivants :

- Dans l'axe mathématique, deux contributions originales ont été proposées dans ce document :
  - Une nouvelle décomposition tensorielle basée sur la décomposition CP, que nous avons baptisé "la décomposition CP avec isolement du facteur d'échelle". Le principe de cette nouvelle décomposition est de calculer la valeur optimale du facteur d'échelle que nous avons isolé et mis en dehors des matrices facteur dans la décomposition CP. L'isolement de la matrice de facteur d'échelle permet de réduire les indéterminations d'échelle en module unitaire et non pas les fixer complètement, d'où la difficulté d'estimer l'erreur d'identiffication des matrices facteurs.
  - Le calcul d'un indice de performance exacte plus réaliste. Vu que l'indétermination a été caractérisée dans la décomposition CP par $3R$ nombres complexes de module unitaire, notre travail consistait alors à trouver la distance minimale exacte sous une contrainte angulaire, en calculant les $3R$ phases optimales.
- Dans l'axe algorithmique, deux nouveaux algorithmes d'optimisation plus performants ont été développés : "Algorithme 1" et "Algorithme 2". Ces deux algorithmes sont basés sur la méthode du gradient projeté et prennent en compte l'isolement du facteur d'échelle.
- Dans l'axe applicatif, nous avons proposé une nouvelle technique d'accès multiple OWDM-IDMA qui combat conjointement les interférences ISI et MAI. Ensuite, une modélisation des données reçues sous forme d'un tenseur d'ordre trois a été établie dans le cas des trois systèmes suivants : CDMA, IDMA et OWDM-IDMA.

Publications de l'auteur

Revues

1. A. Rouijel, K. Minaoui, P. Comon, D. Aboutajdine, "CP Decomposition Approach to Blind Separation for DS-CDMA System using a New Performance Index", *EURASIP Journal on Advances in Signal Processing*, vol. 2014, n. 1, pp. 128, August, 2014.
2. A. Rouijel, K. Minaoui, P. Comon, D. Aboutajdine, "Short : Blind Separation of the Multiarray Multisensor Systems Using the CP Decomposition", *Lecture Notes in Computer Science Springer*, vol. 8593, pp 313-318 , May, 2014.
3. A. Rouijel, B. Nsiri, D. Aboutajdine, "Performance Analysis of a Novel OWDM-IDMA Approach for Wireless Communication System", *Journal of Communications Software and Systems (JCOMSS)*, vol. 8, n. 3, p. 61, 2012.

Conférences

1. A. Rouijel, K. Minaoui, P. Comon, D. Aboutajdine, "Short : Blind Separation of the Multiarray Multisensor Systems Using the CP Decomposition", In *Second International Conference, NETYS 2014*, May 15-17, Marrakech, Morocco, 2014.
2. P. Comon, K. Minaoui, A. Rouijel, D. Aboutajdine, "Performance Index for Tensor Polyadic Decompositions", In *21th EUSIPCO conference (EUSIPCO'13)*, September 9-13, Marrakech, Morocco, 2013.
3. A. Rouijel, B. Nsiri, D. Aboutajdine, "Peak to Average Power Ratio Analysis for OWDM-IDMA System", In *IEEE International Conference on Multimedia Computing and Systems (ICMCS'2011)*, April 7-9, Ouarzazate, Morocco, 2011.
4. A. Rouijel, B. Nsiri, M. ET.TOLBA, "Performance of Iterative Multiuser Detection in OWDM-IDMA system", In *IEEE International Symposium on signal, Image, Video and Communications (ISIVC'2010)*, Rabat, Morocco, 2010.
5. A. Rouijel, B. Nsiri, A. Faqihi, D. Aboutajdine, "A New Approach for Wireless Communication Systems Based on IDMA Technique", In *IEEE International Symposium on signal, Image, Video and Communications (ISIVC'2010)*, Rabat, Morocco, 2010.
6. A. Rouijel, Z. Mohammadi, B. Nsiri, D. Aboutajdine, "Etude d'une technique hybride basée sur la technique d'accès multiple IDMA", In *Workshop on Information Technologies and Communication (WOTIC'09)*, Décembre 24-25, Agadir, Morocco, 2009.
7. A. ROUIJEL, B. Nsiri, A. FAQIHI, D. Aboutajdine, "Etude des Performances de la technique IDMA", In *Journées Doctorales en Technologies de l'Information et de la Communication (JDTIC'09)*, July 16-18, Rabat, Morocco, 2009.

Chapitre 2

Généralités sur l'algèbre tensorielle

2.1 Introduction

Dans un nombre croissant d'applications, on est obligé de manipuler des quantités à plusieurs indices. Ces quantités sont appelées tenseurs. La définition du terme tenseur dépendra généralement du domaine scientifique dans lequel il est utilisé. Dans le cas général, un tenseur est traité comme une entité mathématique qui jouit de la propriété de multilinéarité après changement du système de coordonées (Comon, 2000). Nous considérons qu'un tenseur d'ordre N représente un tableau multidimensionnel dont chaque élément est accessible via N indices. Un tenseur d'ordre un est un vecteur, un tenseur d'ordre deux est une matrice alors qu'un tenseur d'ordre zéro est un scalaire. L'analyse des tenseurs d'ordre supérieur à trois est dite algèbre multilinéaire. Nous s'intéressons aux tenseurs d'ordre supérieur ou égal à trois car ils possèdent des propriétés qui ne sont pas valables pour les matrices ou les vecteurs.

Dans ce chapitre, nous introduirons les notions et les notations des outils d'algèbre multilinéaire. Plusieurs définitions générales ainsi que des présentations détaillées des méthodes d'algèbres multilinéaires sont données dans les références bibliographiques suivantes : (Kroonenberg, 1983a; de Silva et Lim, 2008; Bader et Kolda, 2006). La suite de ce chapitre est organisée de la manière suivante : Dans la section 2.2, nous donnerons les notations qui seront adoptées dans ce manuscrit. Après avoir définit mathématiquement la notion d'un tenseur, une énumeration des différents outils tensoriels ainsi que leurs propriétés exploitées par la suite sera faite dans la section 2.3.

2.2 Notations

Commençons d'abord par l'introduction des notations essentielles qui seront utilisées dans ce document. $\mathbb{R}$ désigne l'ensemble des nombres réels, alors que $\mathbb{C}$ est l'ensemble des nombres complexes. Les tenseurs seront notés par des lettres calligraphiques $\mathcal{T}$, les matrices par des lettres majuscules grasses $\mathbf{A}$, les vecteurs par des miniscules grasses $\mathbf{a}$ et les scalaires par des miniscules en italiques a. De plus, la $p^{\text{ème}}$ colonne de la matrice $\mathbf{A}$ est notée $\mathbf{a}_p$ et le $p^{\text{ème}}$ élément d'un vecteur $\mathbf{a}$ est a_p. L'élément d'indice $\{i, j\}$ de la matrice $\mathbf{A}$ sera noté a_{ij}, alors que l'élément $\{i, j, k\}$ d'un tenseur d'ordre trois $\mathcal{T}$ se notera $t_{i,j,k}$. $\mathbf{1}$ représentera un vecteur contenant des uns, et $\mathbf{I}$ la matrice identité.

Le conjugé d'un élément a sera noté a^*, alors que son module sera $|a|$. Le conjugé de la matrice $\mathbf{A}$ sera notée $\mathbf{A}^*$, sa transposée $\mathbf{A}^T$, sa transposée hermitienne $\mathbf{A}^H$ et son

pseudo inverse $\mathbf{A}^{\dagger}$.

Dans ce document, nous utiliserons aussi les deux opérateurs *diag* et *vec*. Le premier opérateur range les éléments d'un vecteur sur la diagonale d'une matrice :

$$diag\{\mathbf{a}\} = \begin{pmatrix} a_1 & & & 0 \\ & a_2 & & \\ & & \ddots & \\ 0 & & & a_R \end{pmatrix}. \tag{2.1}$$

alors que l'opérateur *vec* permet d'écrire une matrice sous forme de vecteur par concaténation de toutes ses colonnes. Si on dispose d'une matrice $\mathbf{A}$ de taille $I \times J$, les éléments du vecteur $vec(\mathbf{A})$ sont ainsi définis par $(vec(\mathbf{A}))_{i+(j-1)I} = a_{ij}$.

$$vec(\mathbf{A}) = \begin{pmatrix} a_{11} \\ \vdots \\ a_{I,1} \\ a_{12} \\ \vdots \\ a_{I2} \\ \vdots \\ a_{1J} \\ \vdots \\ a_{IJ} \end{pmatrix}. \tag{2.2}$$

L'opérateur inverse de *vec* sera noté *unvec*.

2.3 Définitions et propriétés mathématiques

2.3.1 Définition d'un tenseur

La définition d'un tenseur est donnée en détails par DE SILVA et al (de Silva et Lim, 2008), qui rappelle que ce dernier ne doit pas être confondu avec le tenseur utilisé en physique et en ingénierie. Ces derniers sont souvent désignés par le tenseur champ (*field tensor*) en mathématique. Nous considérons qu'un tenseur d'ordre N représente un tableau multidimensionnel dont chaque élément est accessible via N indices $\{I_1, ..., I_N\}$. Ce tenseur est noté par :

$$\mathcal{T} \in \mathbb{C}^{I_1 \times \cdots \times I_N}. \tag{2.3}$$

et ses éléments sont notés $t_{I_1,\cdots,I_N}$. Chaque dimension du tenseur est appelée «n-mode» en référence au $n^{\text{ème}}$ indice du tenseur (Smilde *et al.*, 2004; Lathauwer *et al.*, 2000; Kroonenberg, 1983a). Un tenseur d'ordre zéro est un scalaire, un tenseur d'ordre un est un vecteur, alors qu'un tenseur d'ordre deux est une matrice. Un tenseur d'ordre trois peut être resprésenté par une somme de produits tensoriels entre trois vecteurs, donc il peut être généré au moyen de trois vecteurs et représenté par un tableau à trois dimensions. L'ordre d'un tenseur désigne alors le nombre de ces dimensions. La Figure 2.1 illustre un

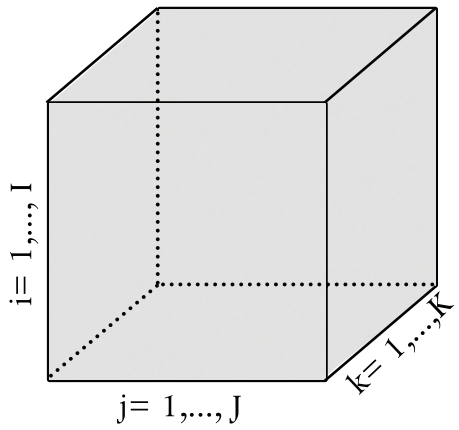

FIGURE 2.1 – Exemple d'un tenseur d'ordre 3.

tenseur d'ordre 3 dont les dimensions successives sont I, J et K.

Considérons un tenseur $\mathcal{T}$ d'ordre 3. En spécifiant un indice, on obtient une tranche (une matrice) dans une orientation donnée et, en spécifiant deux indices, on obtient une fibre (un vecteur). La Figure 2.2 illustre les tranches horizontales $\mathcal{T}_{I,:,:}$, les tranches latérales $\mathcal{T}_{:,J,:}$ et les tranches frontales $\mathcal{T}_{:,:,K}$. De la même façon, la fibre d'un tenseur est similaire à une tranche mais en fixant l'ensemble des indices sauf un seul. La Figure 2.3 présente les fibres en colonnes $\mathcal{T}_{:,J,K}$, les fibres en lignes $\mathcal{T}_{I,:,K}$ et en tubes $\mathcal{T}_{I,J,:}$, tel que la ponctuation (:) représente le "full rang" d'un indice donné.

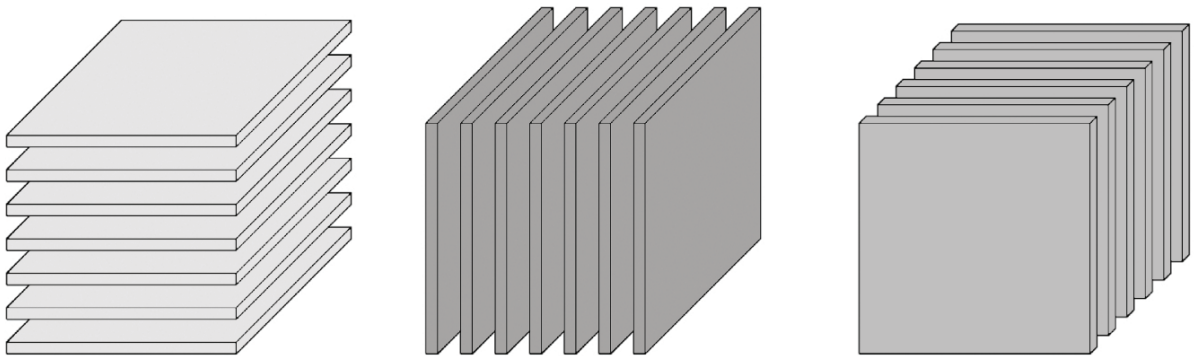

FIGURE 2.2 – Les tranches pour un tenseur d'ordre 3. De gauche à droite, $\mathcal{T}_{I,:,:}$, $\mathcal{T}_{:,J,:}$ et $\mathcal{T}_{:,:,K}$.

Un tenseur peut être symétrique, antisymétrique ou ni l'un ni l'autre. On dira alors qu'un tenseur $\mathcal{T}$ d'ordre N est symétrique s'il est inchangé par permutations des indices (Comon *et al.*, 2008; Comon, 2006). Cela veut dire que les éléments de ce tenseur vérifient $t_{I_1 I_2 \cdots I_N} = t_{\Pi(I_1 I_2 \cdots I_N)}$ pour toute permutation Π des indices. Un tenseur est dit antisymétrique si, par une permutation quelconque des indices, il subit un changement de signe qui n'est d'autre que celui de la permutation. Pour un tenseur antisymétrique, les composantes dans lesquelles un indice se répète au moins deux fois sont toutes nulles. Par exemple, les k composantes t_{iik} du tenseur $\mathcal{T}$ sont nulles.

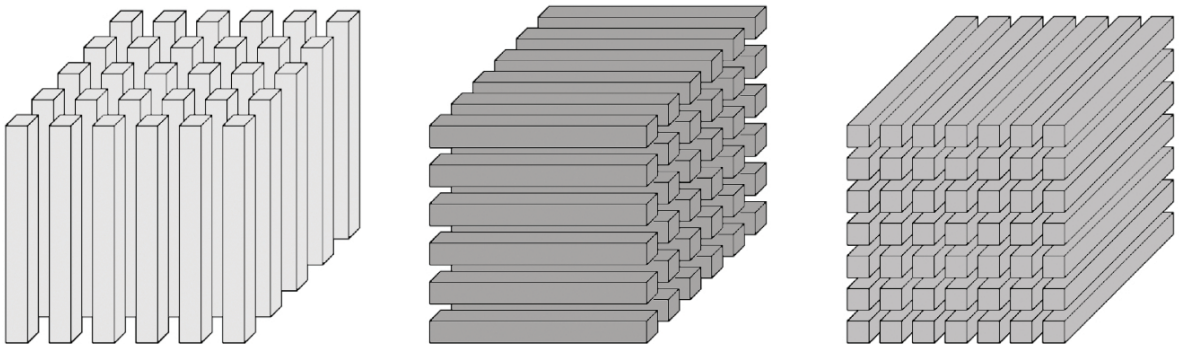

FIGURE 2.3 – Les fibres pour un tenseur d'ordre 3. De gauche à droite, $\mathcal{T}_{:,J,K}$, $\mathcal{T}_{I,:,K}$ et $\mathcal{T}_{I,J,:}$.

2.3.2 Dépliement d'un tenseur

Afin d'étudier les propriétés des données multidimensionnelles dans une direction n-modale particulière, définissons la matrice dépliante (Bader et Kolda, 2006; Kiers, 2000) dans le n-mode du tenseur $\mathcal{T}$, notée $\mathbf{T}_n$. Typiquement, le dépliement selon le n-mode d'un tenseur revient à juxtaposer toutes les tranches associées à ce mode de façon à obtenir une matrice. Il existe trois façons de déplier un tenseur en matrice selon trois modes. Pour mieux comprendre, considérons un tenseur $\mathcal{T}$ de taille $I \times J \times K$. Le dépliement de ce tenseur dans le premier mode «1-mode» nous donne une matrice $\mathbf{T}_1^{I,KJ}$. De même, les dépliements du tenseur dans le deuxième et le troisième mode mènent respectivement au matrices $\mathbf{T}_2^{J,KI}$ et $\mathbf{T}_3^{K,JI}$. Les tranches d'une dépliante peuvent être rangées dans n'importe quel ordre. Les matrices résultantes ont la même taille que celles présentées avant, sauf que l'ordre dans lequel apparaissent les valeurs est différent. La Figure 2.4 illustre l'opération de dépliement d'un tenseur dans les trois mode.

Dans les sections suivantes, nous présenterons les principaux opérateurs tensoriels que nous sommes amenés à utiliser tout le long de ce manuscrit.

2.3.3 Produit de Kronecker

Le produit de Kronecker, noté $\otimes$, de deux matrices $\mathbf{A}$ et $\mathbf{B}$, dont les dimensions sont respectivement $I \times K$ et $J \times L$, est une matrice de taille $IJ \times KL$ définie comme suit :

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1K}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & \cdots & a_{IK}\mathbf{B} \end{pmatrix}. \tag{2.4}$$

Propriétés

Considérons les quatres matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ et $\mathbf{D}$. Le produit de Kronecker vérifie les propriétés suivantes sous réserve de compatibilité des tailles pour les quatres matrices :

- Associativité : $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$
- Distrubitivité par rapport à l'addition des matrices :
  - $\mathbf{A} \otimes 0 = 0 = 0 \otimes \mathbf{A}$
  - $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$

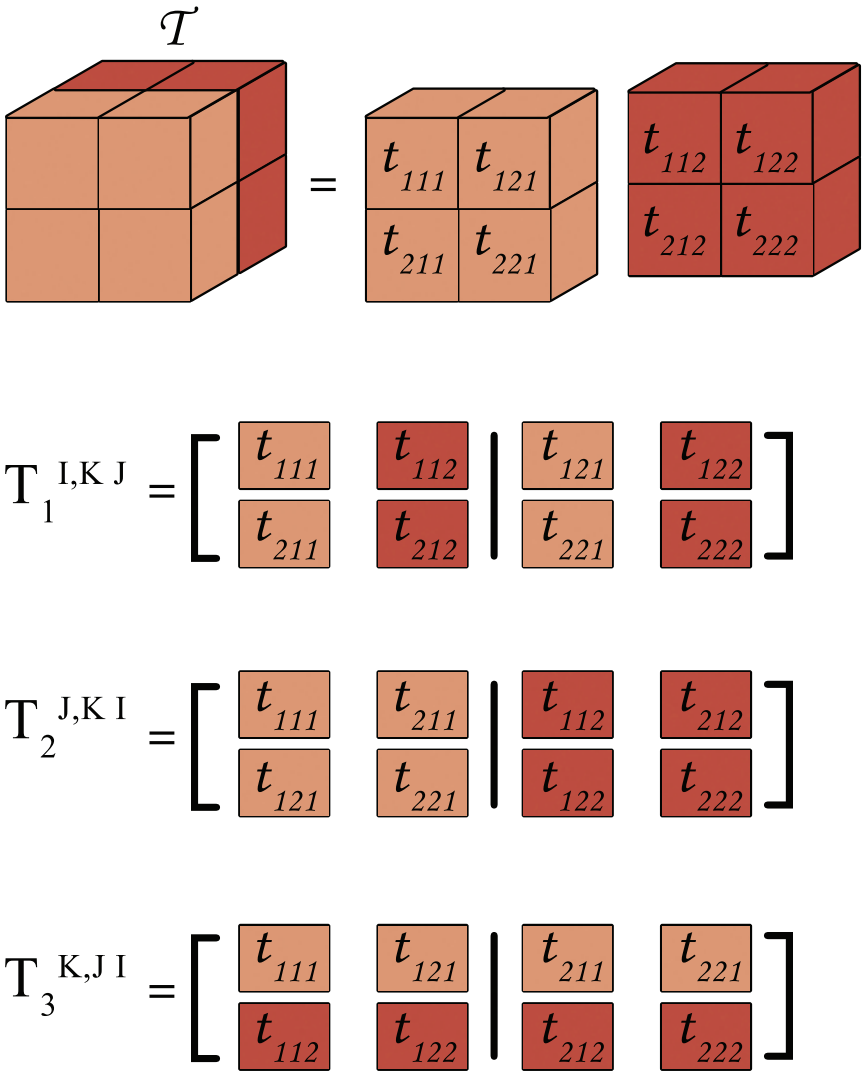

FIGURE 2.4 – Exemple de dépliement d'un tenseur d'ordre 3.

  - $(\mathbf{A}+\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes\mathbf{C}+\mathbf{B}\otimes\mathbf{C}$

- Le transposé : $(\mathbf{A}\otimes\mathbf{B})^T=(\mathbf{A}^T\otimes\mathbf{B}^T)$
- Si $\mathbf{A}$ et $\mathbf{B}$ sont inversibles, $\mathbf{A}\otimes\mathbf{B}$ l'est aussi : $(\mathbf{A}\otimes\mathbf{B})^{-1}=(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1})$
- Soit α et β deux scalaires :
  - $\alpha(\mathbf{A}\otimes\mathbf{B})=(\alpha\mathbf{A})\otimes\mathbf{B}=\mathbf{A}\otimes(\alpha\mathbf{B})$
  - $(\alpha\mathbf{A})\otimes(\beta\mathbf{B})=\alpha\beta(\mathbf{A}\otimes\mathbf{B})$
- La propriété suivante mélange les aspects liés au produit matriciel usuel et au produit de Kronecker : $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=(\mathbf{AC})\otimes(\mathbf{BD})$.

Le produit de Kronecker n'est pas commutatif ; cependant pour toutes $\mathbf{A}$ et $\mathbf{B}$, il existe deux matrices de permutation $\mathbf{P}$ et $\mathbf{Q}$ telles que $\mathbf{A}\otimes\mathbf{B}=\mathbf{P}(\mathbf{B}\otimes\mathbf{A})\mathbf{Q}$.

2.3.4 Produit de Khatri Rao

Le produit de Khatri-Rao (Rao et Mitra, 1972), noté $\odot$, entre deux matrices A et B ayant le même nombre de colonnes J est défini comme étant le produit de Kronecker selon

les colonnes. Il s'exprime par :

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \cdots, \mathbf{a}_J \otimes \mathbf{b}_J]. \tag{2.5}$$

Propriétés

Le produit de Khatri Rao présente les propriétés suivantes. Si l'on considère trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, dont les deux premières possédent le même nombre de colonnes alors que la troisième possède la même taille que $\mathbf{B}$

- Associativité : $\mathbf{A} \odot (\mathbf{B} \odot \mathbf{C}) = (\mathbf{A} \odot \mathbf{B}) \odot \mathbf{C}$
- Distrubitivité par rapport à l'addition des matrices : $\mathbf{A} \odot (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \odot \mathbf{B}) + (\mathbf{A} \odot \mathbf{C})$
- Soit α un scalaire : $\alpha(\mathbf{A} \odot \mathbf{B}) = (\alpha\mathbf{A}) \odot \mathbf{B} = \mathbf{A} \odot (\alpha\mathbf{B})$
- Non-commutativité : $\mathbf{A} \odot \mathbf{B} \neq \mathbf{B} \odot \mathbf{A}$

2.3.5 Produit de Hadamard

Le produit de Hadamard, noté $\boxdot$, entre deux matrices $\mathbf{A}$ et $\mathbf{B}$ de même taille $I \times J$ est une matrice de même dimension que les opérandes. Ce produit n'est rien d'autre que le produit terme à terme entre chaque élément des matrices, et qui s'exprime comme suit :

$$(\mathbf{A} \boxdot \mathbf{B})_{ij} = a_{ij} \cdot b_{ij}. \tag{2.6}$$

Propriétés

Considèrons trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ de même taille et un scalaire α. Le produit de Hadamard vérifie les propriétés suivantes :

- Associativité : $\mathbf{A} \boxdot (\mathbf{B} \boxdot \mathbf{C}) = (\mathbf{A} \boxdot \mathbf{B}) \boxdot \mathbf{C}$
- Distrubitivité par rapport à l'addition des matrices : $\mathbf{A} \boxdot (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \boxdot \mathbf{B}) + (\mathbf{A} \boxdot \mathbf{C})$
- Soit α un scalaire : $\alpha(\mathbf{A} \boxdot \mathbf{B}) = (\alpha\mathbf{A}) \boxdot \mathbf{B} = \mathbf{A} \boxdot (\alpha\mathbf{B})$
- Commutativité : $\mathbf{A} \boxdot \mathbf{B} = \mathbf{B} \boxdot \mathbf{A}$
- Le transposé : $(\mathbf{A} \boxdot \mathbf{B})^T = \mathbf{A}^T \boxdot \mathbf{B}^T$
- $(\mathbf{A} \boxdot \mathbf{B})^T(\mathbf{A} \boxdot \mathbf{B}) = \mathbf{A}^T\mathbf{A} \boxdot \mathbf{B}^T\mathbf{B}$

2.3.6 La trace d'une matrice

Considérons une matrice carrée $\mathbf{A}$ de dimension $N \times N$. La trace de la matrice $\mathbf{A}$ est la somme de ses éléments diagonaux donnée par :

$$trace\{\mathbf{A}\} = \sum_{i=1}^{N} a_{ii}. \tag{2.7}$$

2.4 Opérateurs tensoriels

2.4.1 Le produit n-modal

Le produit n-mode, noté $\times_n$, généralise le produit matriciel au produit entre un tenseur et les vecteurs n-modaux dans un n-mode particulier (Lathauwer *et al.*, 2000). Considèrons un tenseur d'ordre trois $\mathcal{T} \in \mathbb{C}^{I \times J \times K}$ et une matrice $\mathbf{A} \in \mathbb{C}^{L \times I}$. Le produit en 1-mode entre le tenseur $\mathcal{T}$ et la matrice $\mathbf{A}$ est un tenseur de taille $L \times J \times K$ dont les éléments selon le premier mode sont définis par :

$$(\mathcal{T} \times_1 \mathbf{A})_{ljk} = \sum_{i=1}^{I} t_{ijk} a_{li}. \tag{2.8}$$

De même, le produit en 2-mode de $\mathcal{T}$ par une matrice $\mathbf{B} \in \mathbb{C}^{L \times J}$ est un tenseur de taille $I \times L \times K$ tel que ses éléments sont définis par :

$$(\mathcal{T} \times_1 \mathbf{A})_{ilk} = \sum_{j=1}^{J} t_{ijk} b_{lj}. \tag{2.9}$$

Ainsi que le produit en 3-mode de $\mathcal{T}$ par une matrice $\mathbf{C} \in \mathbb{C}^{L \times K}$ est un tenseur de taille $I \times L \times K$:

$$(\mathcal{T} \times_1 \mathbf{A})_{ijl} = \sum_{k=1}^{K} t_{ijk} c_{lk}. \tag{2.10}$$

Nous pouvons généraliser l'écriture de ces trois dernières équations aux produits n-modaux entre le tenseur $\mathcal{A} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ et les matrices $\mathbf{X^{(n)}} \in \mathbb{C}^{J_n \times I_n}$, avec $n = 1, \cdots, N$. Le résultat de ce produit est un tenseur $\mathcal{T} \in \mathbb{C}^{J_1 \times \cdots \times J_N}$:

$$\mathcal{T} = \mathcal{A} \times_1 \mathbf{X^{(1)}} \times_2 \cdots \times_N \mathbf{X^{(N)}}. \tag{2.11}$$

Propriétés

Quel que soit le tenseur $\mathcal{A} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$, le produit n-mode vérifie les propriétés suivantes (De Lathauwer *et al.*, 2000) :

– Quelles que soient les matrices $\mathbf{U} \in \mathbb{C}^{J_n \times I_n}$ et $\mathbf{V} \in \mathbb{C}^{K_n \times J_n}$:

$$(\mathcal{A} \times_n \mathbf{U}) \times_n \mathbf{V} = \mathcal{A} \times_n (\mathbf{U})\mathbf{V}). \tag{2.12}$$

– Quelles que soient les matrices $\mathbf{U} \in \mathbb{C}^{J_n \times I_n}$ et $\mathbf{V} \in \mathbb{C}^{J_m \times I_m}$:

$$(\mathcal{A} \times_n \mathbf{U}) \times_m \mathbf{V} = (\mathcal{A} \times_m \mathbf{U}) \times_n \mathbf{V}). \tag{2.13}$$

– Si les matrices $\mathbf{U}^{(n)} \in \mathbb{C}^{I_n \times R_n}$ sont orthogonales $\forall n = 1, \cdots, N$ alors (De Lathauwer *et al.*, 2000; Kroonenberg, 1983b; Kroonenberg et Leeuw, 1980) :

$$\mathcal{T} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \cdots \times_N \mathbf{U}^{(N)} \Rightarrow \mathcal{T} \times_1 \mathbf{U}^{(1)^T} \times_2 \cdots \times_N \mathbf{U}^{(N)^T} = \mathcal{A}. \tag{2.14}$$

2.4.2 Le produit extérieur

Le produit extérieur entre deux vecteurs $\mathbf{a}$ et $\mathbf{b}$ définit classiquement une matrice $\mathbf{A} \in \mathbb{C}^{I_1 \times I_2}$ de rang 1 :

$$\mathbf{A} = \mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T. \tag{2.15}$$

où I_1 est la dimension du vecteur $\mathbf{a}$ et I_2 est la dimension du vecteur $\mathbf{b}$. En général, le produit extérieur de plusieurs vecteurs $\mathbf{a}^{(1)}, \cdots, \mathbf{a}^{(N)}$ de dimension respectivement $I_1, \cdots, I_N$, définit le tenseur $\mathcal{T} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ de rang 1 :

$$\mathcal{T} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}. \tag{2.16}$$

dont les éléments sont définis par le produit :

$$t_{i_1,\ldots,i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}. \tag{2.17}$$

dans lequel $a^{(n)}_{i_n}$ est la $i_n{}^{\text{ème}}$ composante du vecteur $a^{(n)}$.
Le produit extérieur entre les trois vecteurs $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ engendre un tenseur $\mathcal{T} \in \mathbb{C}^{I \times J \times K}$ d'ordre 3 et de rang 1 :

$$\mathcal{T} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}, \tag{2.18}$$

La Figure 2.5 schématise le produit extérieur de trois vecteurs $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ de dimension respectivement I, J, K. Le produit extérieur de deux tenseurs se fait en multipliant leurs

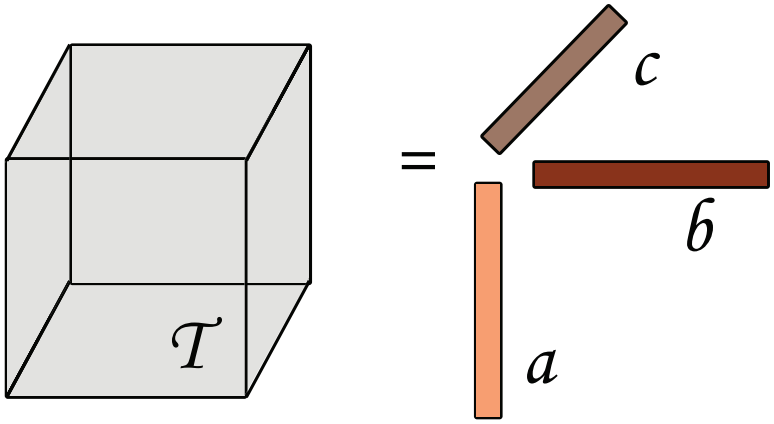

FIGURE 2.5 – Représentation d'un tenseur de rang 1 généré à partir de produits tensoriels entre trois vecteurs.

composantes. Par contre, le tenseur obtenu dans ce cas possède un ordre égal à la somme des ordres des tenseurs multipliés. Pour mieux comprendre, soit $\mathcal{X} \in \mathbb{C}^{I_1 \times \cdots \times I_P}$ et $\mathcal{Y} \in \mathbb{C}^{J_1 \times \cdots \times J_Q}$ deux tenseurs d'ordre P et Q respectivement. Le produit extérieur $(\mathcal{X} \circ \mathcal{Y})$ est un tenseur d'ordre $P + Q$ dont chaque élément est défini par :

$$(\mathcal{X} \circ \mathcal{Y})_{i_1 i_2 \ldots i_P j_1 j_2 \ldots j_Q} = x_{i_1 i_2 \ldots i_P} y_{j_1 j_2 \ldots j_Q}. \tag{2.19}$$

2.4.3 Produit scalaire

Le produit scalaire tensoriel, noté $\langle , \rangle$, est défini de la façon suivante. Considérons deux tenseurs $\mathcal{X}$ et $\mathcal{Y} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ de même ordre N. Le produit scalaire entre ces deux tenseurs est donné par :

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1,i_2,\ldots,i_N} y_{i_1,i_2,\ldots,i_N}. \tag{2.20}$$

Une propriété intéressante du produit scalaire tensoriel est qu'il peut s'écrire en fonction de la trace :

$$\langle \mathcal{X}, \mathcal{Y} \rangle = trace\{\mathbf{X}_n \mathbf{Y}_n^T\}, \quad \forall n = 1, \cdots, N. \tag{2.21}$$

tel que $\mathbf{X}_n$ et $\mathbf{Y}_n^T$ sont les matrices dépliantes de $\mathcal{X}$ et $\mathcal{Y}$ selon le n-mode.

2.4.4 Norme Frobenius

La norme Frobenius $\|.\|_F$ d'un tenseur $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \cdots \times I_N}$, héritée du produit scalaire tensoriel défini précédemment, est définie comme suit :

$$\|\mathcal{T}\|_F = \sqrt{(\langle \mathcal{T}, \mathcal{T} \rangle)} = \sqrt{(\sum_{i_1,\ldots,i_N} |t|^2_{i_1,i_2,\ldots,i_N})}. \tag{2.22}$$

La distance quadratique entre les tenseurs $\mathcal{X}$ et $\mathcal{Y} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ peut être définie par la quantité $\|\mathcal{X} - \mathcal{Y}\|_F^2$.

2.4.5 Rang d'un tenseur

Le rang d'un tenseur d'ordre supérieur à 2 n'est pas trivial à définir (Bro, 1998; De Lathauwer *et al.*, 2000; Lathauwer *et al.*, 2000; de Silva et Lim, 2008). Il n'est pas évident de généraliser la définition du rang d'une matrice pour un tenseur d'ordre N (avec $N > 2$). En effet, le rang d'une matrice quelconque $\mathbf{A}$ vérifie deux définitions :

Définition 1 : Le rang de la matrice correspond à la dimension du sous-espace vectoriel engendrée par les vecteurs (lignes et colonnes) de $\mathbf{A}$,

Définition 2 : Le rang de $\mathbf{A}$ est lié au produit extérieur, étant le nombre minimum de matrices de rang 1 qui génèrent la matrice $\mathbf{A}$.

Pour un tenseur d'ordre N (avec $N > 2$), l'égalité entre ces deux définitions n'est plus vérifiée (de Silva et Lim, 2008), et conduit à deux définitions de rang distinctes : Un rang multilinéaire et un rang tensoriel. Au début du 19$^{\text{ème}}$ siècle, Hitchcock a présenté pour la première fois le concept d'un rang multiple (HITCHCOCK, 1927a,b), qui généralise directement le rang des vecteurs colonnes et des vecteurs lignes d'une matrice pour un tenseur d'ordre supérieur. Plus tard, Kruskal introduit l'idée du rang n-modal (Kruskal, 1989), popularisé par De Lathauwer (De Lathauwer *et al.*, 2000). La différence entre le rang n-modal et le rang multiple est que le premier se restreint aux matrices dépliantes du tenseur $\mathcal{T}$, alors que le deuxième est défini pour un déploiement arbitraire.

Dans les années 70, le concept de rang a été indépendamment proposé dans des applications par Carroll et Chang (Carroll et Chang, 1970), Harshman (R.Harshman,

1970) et Kruskal (KRUSKAL, 1977; Kruskal, 1989). Le nouveau concept de rang lié au produit extérieur est connu sous le nom de rang tensoriel.

2.4.5.1 Tenseur de rang 1

Un tenseur $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \times \cdots I_N}$ est de rang 1 s'il peut s'écrire comme le produit extérieur de N vecteurs : $\mathcal{T} = \mathbf{a}_1 \circ \mathbf{a}_2 \circ \cdots \circ \mathbf{a}_N$. Ainsi, un tenseur d'ordre trois $\mathcal{T} \in \mathbb{C}^{I \times J \times K}$ est de rang 1 si ses éléments peuvent s'écrire comme : $t_{ijk} = a_i b_j c_k$.

Comme cas particulier, une matrice $\mathbf{C} \in \mathbb{C}^{I \times J}$ est de rang 1 si $\mathbf{C} = \mathbf{a}\mathbf{b}^T$.

2.4.5.2 Rang d'un tenseur

Le rang d'un tenseur arbitraire $\mathcal{T}$, noté $R = rank(\mathcal{T})$, est défini comme le nombre minimum des tenseurs de rang 1 qui génèrent $\mathcal{T}$ par combinaison linéaire.

Le rang d'une matrice $\mathbf{C} \in \mathbb{C}^{I \times J}$ est toujours inférieur ou égal à $\min(I, J)$. Cette propriété n'est plus vérifiée dans le cas tensoriel (Kolda, 2001; De Lathauwer *et al.*, 2000). En effet, le rang d'un tenseur $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \times \cdots I_N}$ peut être :

$$R > \min(I_1, \ldots, I_N). \tag{2.23}$$

Le rang d'un tenseur peut être typique ou générique. Une propriété est dite typique si elle est vrai sur un ensemble de volumes non nulle, alors qu'elle est dite générique si elle est vrai presque partout. En d'autres termes, une propriété générique est typique, mais l'inverse n'est pas vrai (Comon et Berge, 2008). De ce fait, on peut dire qu'un rang est typique si sa probabilité est non nulle. Le rang générique d'un tenseur est celui obtenu en choisissant ses éléments aléatoirement selon une loi continue. S'il y a un seul rang typique, il peut être appelé le rang générique.

2.5 conclusion

Ce chapitre fournit les éléments fondamentaux de l'algèbre multilinéaire qui seront utilités par la suite. Dans un premier temps, nous avons introduit les notations qui seront utilisées tout au long de ce manuscrit. Ensuite, nous avons donné la définition d'un tenseur, ainsi que quelques propriètés mathématiques. Dans une deuxième partie, nous avons défini les représentations et les opérations les plus importantes concernant les tenseurs, ainsi que leurs proprietés.

Nous avons mis en évidence le fait que la généralisation des outils bilinéaires aux outils multilinéaires n'est pas triviale. La majorité des opérations entre les matrices se généralisent facilement pour les tenseurs. C'est le cas par exemple de la somme et de la multiplication d'un tenseur avec un réel. Pour le produit scalaire et la norme, la définition s'étend aussi naturellement. Cependant, il en existe d'autres où les matrices et les tenseurs se distinguent. L'une des principales distinctions est la non-extension du théorème d'ECKART-YOUNG (Eckart et Young, 1936) au cas tensoriel (tenseur d'ordre > 2). Il en résulte deux définitions de rangs distincts pour un tenseur : Le rang multilinéaire (n-modal), donné par Hitchcock, qui généralise le rang des vecteurs colonnes et des vecteurs lignes d'une matrice pour un tenseur d'ordre supérieur (HITCHCOCK, 1927a,b), et le

rang tensoriel R qui généralise le rang associé au produit extérieur d'une matrice(Carroll et Chang, 1970; R.Harshman, 1970; KRUSKAL, 1977; Kruskal, 1989).

Le rang n-modal R_n et le rang du tenseur R mènent donc à deux définitions des décompositions tensorielles, toutes les deux issues de la généralisation de la décomposition en valeurs singulières d'une matrice, notée SVD (*Singular Values Decomposition*). Ainsi, le rang multilinéaire est lié à la décomposition tensorielle qui généralise la modélisation matricielle de la SVD, appelée la décomposition de TUCKER3, alors que le rang tensoriel R, associé à la décomposition tensorielle, généralise la modélisation canonique de la SVD, appelée la décomposition de PARAFAC/CANDECOMP. Nous présenterons dans le chapitre suivant ces deux décompositions tensorielles ainsi que d'autres.

Chapitre

3 Décomposition tensorielle

3.1 Introduction

Dans un nombre croissant de domaines ; tel que le traitement du signal, l'analyse multivariée ou la programmation scientifique, il est nécessaire de manipuler des quantités à plusieurs indices (Chevalier *et al.*, 2005; Sidiropoulos *et al.*, 2000b; Moreau et de C. Luigi, 2003). Ces quantités sont appelées des tenseurs. L'analyse des tenseurs d'ordre supérieur à trois est dite algèbre multilinéaire. On s'intéresse aux tenseurs d'ordre supérieur à trois, car ils possèdent des propriétés qui ne sont pas valables pour les matrices ou les vecteurs.

De même qu'une matrice qui peut être décomposée de plusieurs façons (SVD, EVD, QR, etc) (Golub et Loan, 1996), un tenseur peut se décomposer de différentes manières. La décomposition tensorielle est un domaine de l'algèbre multilinéaire qui caractérise un tenseur comme une combinaison linéaire de produits extérieurs des vecteurs. Parmi les différentes décompositions tensorielles, deux types qui ont été largement étudiés dans la littérature, tout en faisant l'objectif de nombreuses applications dans des domaines différents. Ces différentes décompositions sont la décomposition PARAFAC (*Parallel Factor Analysis*) (R.Harshman, 1970; Harshman, 1972), la décomposition de TUCKER3, qui peut être interprétée comme une généralisation de l'analyse en composantes principales (ACP) à des ordres supérieurs (Tucker, 1966a). Cette décomposition est connue aussi sous le nom de la décomposition en valeurs singulières d'ordre supérieur (HOSVD) (Kolda, 2001), qui est une généralisation de la SVD matricielle. La décomposition Tucker-3 a été appliquée avec succès dans différents domaines tels que la chromatographie (Bro, 1998) et l'analyse de la perception des personnes (Kroonenberg, 1983b).

Nous nous intéressons ici à la décomposition PARAFAC des tenseurs. Cette décomposition a été introduite de manière indépendante sous le nom de CANDECOMP (*Canonical Decomposition*) en psychométrie (Carroll et Chang, 1970) et de PARAFAC (*Parallel Factor Analysis*) en phonétique (R.Harshman, 1970), mais la términologie la plus répandue est CP (*Canonical Polyadic Decomposition*) (Comon *et al.*, 2009). Elle a ensuite été utilisée dans des domaines variés, comme la chimiométrie (BRO, 1997), ou les télécommunications (Sidiropoulos *et al.*, 2000b,a). La caractèristique la plus intéressante de la décomposition CP est son originalité intrinsèque. Car, et contrairement à la décomposition matricielle, où le problème de liberté de rotation se manifeste, la décomposition PARARAC des tenseurs est essentiellement unique, à l'indétermination de l'échelle et de permutation (KRUSKAL, 1977; Stegeman et Sidiropoulos, 2007). La première preuve d'unicité a été fournie par Kruskal dans (KRUSKAL, 1977). Récemment, cette preuve a

été reformulée en utilisant l'algèbre linéaire de base (Stegeman et Sidiropoulos, 2007). Une preuve concise valable pour les tenseurs complexes a été donnée dans (Sidiropoulos *et al.*, 2000b). Contrairement à la décomposition CP, la décomposition de Tucker-3 ne possède pas une unicité intrinsèque et n'est pas intéressante d'un point de vue d'estimation des paramètres. Cependant, on peut parfois résoudre ce problème en utilisant des versions limitées du modèle de Tucker.

Dans ce chapitre, nous présenterons une nouvelle technique de la décomposition tensorielle d'un tenseur d'ordre 3, basée sur la décomposition CP. Les applications de cette décomposition seront abordées dans le chapitre 4.

Le reste de ce chapitre est organisé de la manière suivante : dans la section 3.2, quelques décompositions tensorielles seront présentées. Une plus grande attention sera accordée aux décompositions d'un tenseur de troisième ordre, puisque ça sera le cas dans la plupart des applications rencontrées dans cette thèse. Dans certains cas, la généralisation au $N^{\text{ème}}$ ordre sera également donnée. La section 3.3 sera dédiée à notre première contribution, dans laquelle nous proposerons une décomposition tensorielle basée sur la décomposition CP, et qui résout le problème de l'indétermination de l'échelle. Nous prouverons l'existence de cette décomposition et nous étudierons son unicité. Il est bien connu que les matrices facteurs sont identifiées à une indétermination d'échelle près. Cette indétermination est compliquée à prendre en compte, étant donné que le produit de tous les facteurs d'échelle doit être égal à la matrice identité. Pour cette raison, seuls les indices de performance approximatifs ont été utilisés jusqu'à présent en ignorant la dernière contrainte. Cependant, on peut se demander s'il est possible de calculer l'indice de performance exact ? C'est ce qui représentera notre deuxième contribution. Enfin, nous finirons ce chapitre par une conclusion donnée dans la section 3.4.

3.2 Les différentes techniques de la décomposition tensorielle

La présente section est dediée à la décomposition des tenseurs d'ordre supérieur. La décomposition tensorielle, aussi appelée analyse factorielle à voies multiples (*Multi-way factor analysis*), est un domaine d'algèbre multilinéaire, qui caractérise un tenseur comme une combinaison linéaire de produits extérieurs des vecteurs. Selon l'approche considérée, les décompositions de tenseurs peuvent être considérées comme des généralisations de l'analyse en composantes principales (ACP) ou de la décomposition en valeurs singulières (SVD) pour les ordres supérieures à deux. L'analyse d'un tenseur en terme de ses facteurs décomposés est utile dans les problèmes où différents facteurs d'un mélange multilinéaire doivent être identifiés à partir des données mesurées. Dans le contexte de cette thèse, le calcul d'une décomposition tensorielle d'un tenseur de données observées permet de séparer les signaux émis par les différentes sources.

Dans ce qui suit, quelques décompositions tensorielles seront présentées. Nous détaillerons la décomposition de Tucker-3 qui généralise l'analyse en composantes principales. Par la suite, nous présenterons la décomposition tensorielle de PARAFAC/CANDECOMP, dont l'acronyme utilisé dans ce manuscrit sera la décomposition CP. Une plus grande attention sera accordée aux décompositions tensorielles de troisième ordre, puisque ça concernera la plupart des applications rencontrées dans cette thèse.

3.2.1 La décomposition Tucker-3

La décomposition de Tucker-3 a été introduite pour la première fois par Tucker en 1963 (Tucker, 1963). Un affinement est apporté dans (Tucker, 1964, 1966b). Plusieurs noms ont été attribués dans la littérature pour désigner cette décomposition, les plus connus sont :

(i) *Three-mode factor analysis (3MFA/TUCKER3)* (Tucker, 1966b)
(ii) *Three-mode principal components analysis (3MPCA)* (Kroonenberg et Leeuw, 1980)
(iii) *N-mode principal components analysis* (A. Kapteyn et Wansbeek, 1986)
(iv) *Higher-order* SVD (HOSVD) (De Lathauwer *et al.*, 2000)

La décomposition Tucker-3 est générale, dans le sens où elle incorpore la plupart des autres décompositions tensorielles de troisième ordre comme des cas particuliers. Le modèle de Tucker-3 a été initialement élaboré pour effectuer une décomposition en valeurs singulières sur tous les modes du tenseur, indépendamment les uns des autres.

3.2.1.1 Le modèle de Tucker-3

La décomposition de Tucker-3 est une forme d'analyse en composantes principales. Elle factorise un tenseur $\mathcal{T}$ d'ordre N en un tenseur noyau (*core tensor*), multiplié par une matrice sur chacun de ses N modes. Pour un tenseur $\mathcal{T} \in \mathbb{C}^{I\times J\times K}$ d'ordre 3, la décomposition de Tucker-3 en forme scalaire peut être exprimée comme suit :

$$t_{ijk} = \sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N} a_{if}\, b_{jm}\, c_{kn}\, g_{fmn} \tag{3.1}$$

où a_{if}, b_{jm} et c_{kn} sont les éléments des trois matrices $\mathbf{A} \in \mathbb{C}^{I\times F}$, $\mathbf{B} \in \mathbb{C}^{J\times M}$ et $\mathbf{C} \in \mathbb{C}^{K\times N}$ respectivement, et g_{fmn} est l'élément du tenseur noyau $\mathcal{G} \in \mathbb{C}^{F\times M\times N}$ dont chaque indice indique le niveau d'intéraction entre les différents composants.

À partir de l'equation (3.1), nous pouvons dire que la décomposition du tenseur selon Tucker-3 est une combinaison linéaire des produits extérieurs, où le coefficient de chaque terme du produit extérieur représente l'élément scalaire correspondant du tenseur noyau. Une illustration de la décomposition Tucker-3 est donnée dans la Figure (3.1).

La décomposition de Tucker-3 peut être aussi exprimée par recours à la notation du produit n-modal, définie dans (3.2) :

$$\mathcal{T} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \tag{3.2}$$

La décomposition Tucker-3 d'ordre N

La généralisation de la décomposition de Tucker-3 (3.1) au $n^{\text{ième}}$ ordre est directe. Prenons un tenseur $\mathcal{T} \in \mathbb{C}^{I_1\times I_2\times\cdots\times I_N}$ d'ordre N. La décomposition Tucker-3 du tenseur $\mathcal{T}$ peut être exprimée comme :

$$t_{i_1\ldots i_N} = \sum_{j_1=1}^{J_1}\sum_{j_2=1}^{J_2}\cdots\sum_{j_N=1}^{J_N} x^{(1)}_{i_1 j_1}\, x^{(2)}_{i_2 j_2}\cdots x^{(N)}_{i_N j_N}\, g_{j_1 j_2\ldots j_N}, \tag{3.3}$$

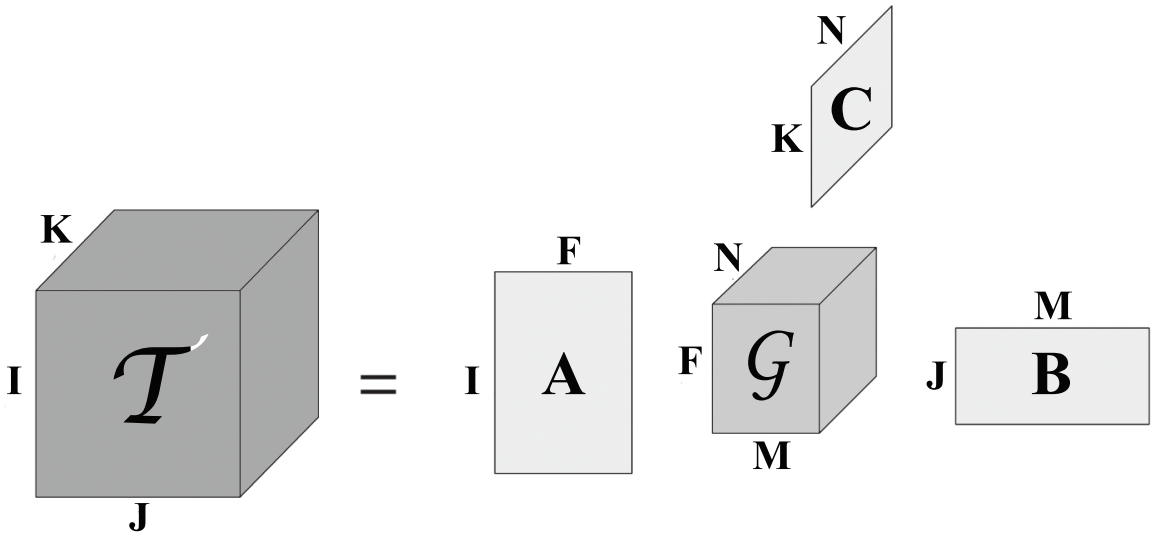

FIGURE 3.1 – Schéma de la décomposition de Tucker3 pour un tenseur d'ordre 3

où $x^{(n)}_{i_n j_n}$ sont les composants scalaires des matrices $\mathbf{X}^{(n)}$ $\forall n = 1, \cdots, N$, appelées aussi les matrices facteurs, et $g_{j_1 j_2 \cdots j_N}$ est le composant scalaire du tenseur noyau $\mathcal{G}$ d'ordre N. Autrement écrit, en utilisant le produit n-modal, le modèle (3.3) s'exprime par :

$$\mathcal{T} = \mathcal{G} \times_1 \mathbf{X}^{(1)} \times_2 \mathbf{X}^{(2)} \cdots \times_N \mathbf{X}^{(N)}, \tag{3.4}$$

Revenons au cas du troisième ordre. La décomposition de Tucker-3 n'est pas unique, à cause de l'infinité des solutions trouvées pour les matrices facteurs et le tenseur noyau menant au même tenseur $\mathcal{T}$. En d'autres termes, la décomposition de Tucker-3 permet d'avoir plusieurs transformations linéaires arbitraires sur les trois matrices facteurs (à condition que l'inverse de ces transformations soit appliqué au tenseur noyau) sans affecter le tenseur reconstruit $\mathcal{T}$. L'unicité complète des matrices facteurs et du tenseur noyau de la décomposition Tucker-3 n'est possible que dans certains cas spéciaux, où au moins deux matrices facteurs ont une structure spéciale qui permet une détermination unique des matrices de transformation. De plus, il a été montré que l'unicité partielle peut exister dans les cas où le tenseur noyau de la décomposition Tucker-3 est sous la contrainte d'avoir plusieurs éléments égaux à zéro (ten Berge et Smilde, 2002).

Comme évoqué préalablement, la décomposition de Tucker-3 est connue sous différents noms comme *Higher-order* SVD (HOSVD) que nous entamerons dans la sous section suivante.

3.2.1.2 HOSVD

La HOSVD permet de déterminer les matrices orthogonales $\mathbf{X}^{(n)}$, en effectuant la décomposition matricielle SVD sur chaque mode du tenseur indépendamment les uns des autres. Les matrices $\mathbf{X}^{(n)}$ sont donc les vecteurs propres de la matrice de covariance de la dépliante dans le n-mode de $\mathcal{T}$. Les deux étapes nécessaires à la détermination de la décomposition de Tucker-3 sont donc :

1. $\forall n = 1, \cdots, N$,

   (a) Déploiement de $\mathcal{T} \rightarrow \mathbf{T}_n$
   (b) Calcul de la SVD de $\mathbf{T}_n = \mathbf{X}^{(n)}\Sigma\mathbf{V}^{(n)} \rightarrow \mathbf{X}^{(n)}$
2. Calcul du tenseur noyau $\rightarrow \mathcal{G} = \mathcal{T} \times_1 \mathbf{X}^{(1)^T} \times_1 \cdots \times_N \mathbf{X}^{(N)^T}$

La troncature de la HOSVD revient à ne conserver que les J_n premiers vecteurs propres, associés aux J_n valeurs propres de la matrice de covariance n-modale (les J_n premières colonnes de la matrice $\mathbf{X}^{(n)}$).

Dans le cas d'une matrice $\mathbf{A} \in \mathbb{C}^{I \times J}$, les vecteurs singuliers gauches et droites sont assimilés aux vecteurs singuliers 1 et 2-modaux respectivement. Par conséquent, la SVD d'une matrice $\mathbf{A}$ peut s'écrire en écriture tensorielle comme :

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T = \Sigma \times_1 \mathbf{U} \times_2 \mathbf{V} \tag{3.5}$$

où Σ est la matrice diagonale, nommée matrice noyau, contenant les valeurs singulières, $\mathbf{U} \in \mathbb{C}^{I \times I}$ et $\mathbf{V} \in \mathbb{C}^{J \times J}$ sont les matrices unitaires contenant les vecteurs singuliers gauches et droits de la matrice $\mathbf{A}$.

Même si la HOSVD semble très similaire à la SVD, il existe des distinctions entre les deux. Ces différences concernent essentiellement la matrice noyau Σ impliquée dans la SVD, et aussi le tenseur noyau $\mathcal{G}$ impliqué dans la HOSVD (De Lathauwer *et al.*, 2000).

3.2.1.3 Applications de la décomposition Tucker-3

La décomposition de Tucker-3 est utilisée dans des domaines très variés (Kolda et Bader, 2009). Historiquement, cette décomposition trouve sa première utilité en analyse chimique, initiée par Herion (Henrion, 1994). De nombreux exemples d'applications en psychométrie ont été réalisés par Kiers et Van Mechelen (Kiers et Mechelen, 2001). Ensuite, De Lathauwer et al. ont utilisé la décomposition de Tucker-3 en traitement du signal. Muti et Bourennane (Muti, 2004) l'ont appliqué pour étendre le filtre de Wiener pour le débruitage d'images couleurs et des signaux sismiques. Cette technique est utilisée pour modéliser les expressions du visages et la compression d'images (Wang et Ahuja, 2003). Finalement, la décomposition de Tucker-3 trouve un intérêt dans l'extraction des connaissances, où Savas et Eldén (Savas, 2003) l'exploitent pour la reconnaissance de l'écriture manuscrite.

3.2.2 La décomposition CP

3.2.2.1 Généralités

Comme pour le modèle de Tucker-3, la décomposition CP est connue sous plusieurs noms dans la littérature. On la retrouve en 1927 sous le nom de *Polyadic Form of a Tensor*, instaurée par Hitchcock (HITCHCOCK, 1927a). Puis c'est en 1944 que Cattell (Cattell, 1944) a décrit le principe du *Parrallel proportional analysis* et le concept d'axes multiples pour l'analyse, autrement dit PARAFAC qui est l'abréviation de *Parallel Factor Analysis*. Cette idée est devenue populaire en 1970, en psychométrie grâce à Harshman (R.Harshman, 1970). Elle a été ensuite reprise indépendamment par Carroll et Chang sous la terminologie *Canonical Decomposition* (CanDecomp) (Carroll et Chang, 1970). Cette décomposition est alors répandue sous le nom de *Canonical Polyadic Decomposition* (CP) (Comon

et al., 2009). Alors que le modèle de Tucker-3 généralise la décomposition matricielle de la SVD, la décomposition CP généralise cette représentation sous sa forme canonique.

Récemment, cette décomposition a trouvé plusieurs applications dans des domaines aussi variés que la psychométrie (Carroll et Chang, 1970) et la phonétique (R.Harshman, 1970). Elle connaît également un grand intérêt dans la séparation de sources, l'identification aveugle et dans l'analyse en composantes indépendantes (ACI) (Comon, 1994; Lathauwer *et al.*, 1995; Comon et Jutten, 2010) ainsi qu'en télécommunications pour les signaux CDMA (*Code Division Multiple Access*) et les technologies MIMO (*Multi-Input Multi-Output*) (M. Castella et Pesquet, 2007; Dimitri, 2007; Sidiropoulos *et al.*, 2000a,b; Almeida *et al.*, 2007). La décomposition CP décrit également la structure de base des cumulants d'ordre supérieur des données multivariées, sur laquelle toutes les méthodes algébriques pour l'analyse en composantes indépendantes sont fondées (Zarzoso et Comon, 2006; Henrion, 1994).

3.2.2.2 Le modèle de la décomposition CP

Considèrons un tenseur d'ordre trois, admettant une décomposition sous forme d'une somme de tenseurs de rang 1. Ainsi, pour un tenseur $\mathcal{T} \in \mathbb{C}^{I\times J\times K}$, on a :

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{3.6}$$

où les 3 matrices $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_R] \in \mathbb{C}^{I\times R}$, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_R] \in \mathbb{C}^{J\times R}$ et $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_R] \in \mathbb{C}^{K\times R}$ sont les matrices de facteurs dont les colonnes sont appelées les facteurs. Lorsque le nombre entier R est minimal, qui est le nombre de tenseurs de rang 1 nécessaires au maintien de l'égalité (3.6), on parle alors de la décomposition canonique polyadique (CP), où R représente le rang du tenseur $\mathcal{T}$. La Figure (3.2) illustre la décomposition CP pour un tenseur d'ordre 3.

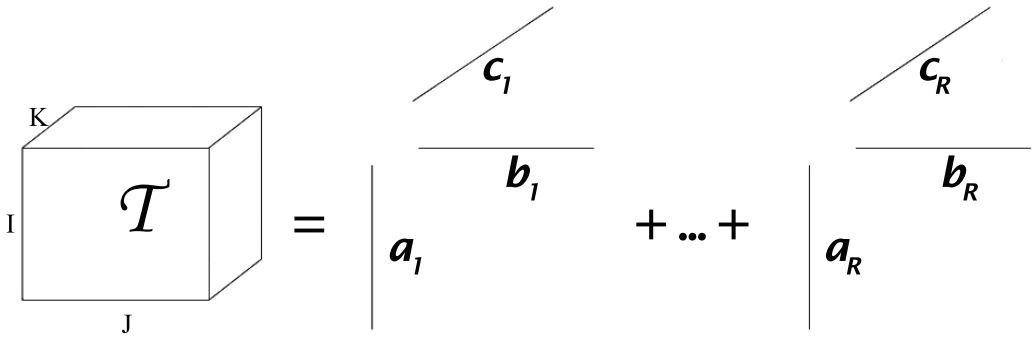

FIGURE 3.2 – Schéma de la décomposition CP d'un tenseur d'ordre 3

En faisant intervenir les composantes des matrices facteurs, la réecriture de l'égalité (3.6) est comme suit :

$$t_{ijk} = \sum_{r=1}^{R} a_{ir}\, b_{jr}\, c_{kr}, \tag{3.7}$$

Une écriture matricielle, couramment utilisée, considère les tranches du tenseur représentées dans la Figure (2.2). Notamment, la décomposition CP pour un tenseur $\mathcal{T}$ d'ordre

3 selon les tranches frontales est donnée par :

$$\begin{aligned} \mathcal{T}_{i,:,:} &= \mathbf{B}D_i(\mathbf{A})\mathbf{C}^T, \\ \mathcal{T}_{:,j,:} &= \mathbf{C}D_j(\mathbf{B})\mathbf{A}^T, \\ \mathcal{T}_{:,:,k} &= \mathbf{A}D_k(\mathbf{C})\mathbf{B}^T, \end{aligned} \tag{3.8}$$

où $D_i(\mathbf{A})$ est la matrice diagonale de la $i^{\text{ème}}$ ligne de la matrice $\mathbf{A} \in \mathbb{C}^{I \times R}$.

En terme de représentations matricielles du tenseur $\mathcal{T}$, la décomposition CP peut s'écrire comme :

$$\begin{aligned} \mathbf{T}_{(1)}^{I,KJ} &= (\mathbf{C} \odot \mathbf{B})\mathbf{A}^T, \\ \mathbf{T}_{(2)}^{J,KI} &= (\mathbf{C} \odot \mathbf{A})\mathbf{B}^T, \\ \mathbf{T}_{(3)}^{K,JI} &= (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T, \end{aligned} \tag{3.9}$$

En pratique, on préfère souvent s'adapter à un modèle multi-linéaire de rang inférieur, $R < rang\{\mathcal{T}\}$, fixé à l'avance, afin de faire face à un problème d'approximation. Plus précisément, il vise à minimiser une fonction objectif de la forme :

$$\Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \|\mathcal{T} - \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F^2 \tag{3.10}$$

Malheureusement, le problème de l'approximation n'est pas toujours bien posé. En fait, comme l'a souligné Harshman (Kruskal *et al.*, 1989; Bro, 2004), la borne inférieure ne peut jamais être atteinte, car cette approximation peut être de rang plus élevé que R. Selon la classification proposée par Richard Harshman (Comon *et al.*, 2009), ce phénomène est appelé *CP-degeneracies*. Il peut se produire lorsqu'on tente d'approximer un tenseur par un autre de rang inférieur. Plusieurs exemples qui expliquent ce phénomène sont donnés dans (Comon *et al.*, 2008; de Silva et Lim, 2008), en particulier le cas d'une séquence de tenseur de rang 2 convergeant vers un tenseur de rang 4. Ici, nous donnons un exemple un peu plus compliqué proposé par Comon dans (Comon *et al.*, 2009).

Soit $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$ trois vecteurs indépendants dans $\mathbb{R}^3$. Alors, la séquence du tenseur symétrique de rang 3 suivant :

$$\mathcal{T}(n) = n^2(\mathbf{x} + \frac{1}{n^2}\mathbf{y} + \frac{1}{n}\mathbf{z})^{\circ 3} + n^2(\mathbf{x} + \frac{1}{n^2}\mathbf{y} - \frac{1}{n}\mathbf{z})^{\circ 3} - 2n^2\mathbf{x}^{\circ 3} \tag{3.11}$$

converge vers le tenseur ci-dessous, qui est de rang 6 quand n tend vers l'infini :

$$\mathcal{T}(\infty) = \mathbf{x} \circ \mathbf{x} \circ \mathbf{y} + \mathbf{x} \circ \mathbf{y} \circ \mathbf{x} + \mathbf{y} \circ \mathbf{x} \circ \mathbf{x} + \mathbf{x} \circ \mathbf{z} \circ \mathbf{z} + \mathbf{z} \circ \mathbf{x} \circ \mathbf{z} + \mathbf{z} \circ \mathbf{z} \circ \mathbf{x} \tag{3.12}$$

Pour conclure, un tenseur $\mathcal{T}(\infty)$ peut être approché arbitrairement par des tenseurs $\mathcal{T}(n)$ de rang 3, mais la limite n'est jamais atteinte, car il est de rang 5. Afin de garantir l'existence d'un minimum, le sous-ensemble des tenseurs de rang R doit être fermé, ce qui n'est pas le cas des tenseurs avec des entrées libres, sauf si $R \leq 1$ (Comon *et al.*, 2008). Si ce n'est pas le cas, on peut effectuer la minimisation soit sur un sous-ensemble fermé (par exemple, le cône de tenseurs avec des entrées positives (Lim et Comon, 2009)), ou sur un sur-ensemble fermé.

3.2.2.3 Unicité de la décomposition CP

Une des propriétés intéressantes propres aux tenseurs d'ordre supérieur à deux, est que leur décomposition tensorielle CP est souvent unique, ce qui n'est pas le cas des décompositions matricielles (R.Harshman, 1970; Kolda et Bader, 2009; Sidiropoulos et Bro, 2000). La décomposition d'une matrice en une somme de matrices de rang un existe elle aussi, mais elle n'est pas unique, sauf si l'on impose certaines contraintes fortes sur les matrices de la décomposition, par exemple une condition d'orthogonalité.

L'unicité signifie qu'il n'y a qu'une seule combinaison de tenseurs de rang un qui par sommation génèrent le tenseur $\mathcal{T}$, à l'exception de l'indétermination d'échelle et de permutation. Lorsque la décomposition est unique à ces deux indéterminations près, on dit qu'elle est essentiellement unique.

Le résultat le plus reconnu d'unicité est présenté par Kruskal (KRUSKAL, 1977; Kruskal, 1989). L'étude de la condition d'unicité de la décomposition CP est basée sur un concept fondamental, qui est le concept de k-rang (*Kruskal-rank*), plus restreinte que la notion usuelle du rang matriciel. Le concept k-rang a été proposé par Kruskal dans son article fondateur (KRUSKAL, 1977), bien que le terme *Kruskal-rank* a été d'abord utilisé par Harshman et Lundy (R.Harshman, 1970). Le concept k-rang a été largement utilisé comme un concept clé indiquant l'unicité de CP.

Rang de Kruskal

Le rang d'une matrice $\mathbf{A} \in \mathbb{C}^{I \times J}$, noté $r_{\mathbf{A}}$, est égal à r si et seulement si $\mathbf{A}$ contient au moins un ensemble de r colonnes linéairement indépendantes. Le rang de Kruskal de $\mathbf{A}$ (ou le k-rang de $\mathbf{A}$), noté $k_{\mathbf{A}}$, est le nombre maximum k de colonnes de $\mathbf{A}$, tel que toute sous-matrice de $\mathbf{A}$ de k colonnes soit de rang plein. Notez que le k-rang d'une matrice est toujours inférieur ou égal à son rang, et nous avons :

$$k_{\mathbf{A}} \leq r_{\mathbf{A}} \leq \min(I, J), \tag{3.13}$$

Exemple 1 : Soit la matrice $\mathbf{A}$ de taille 3×3 définie par :

$$\mathbf{A} = \begin{pmatrix} 1 & 4 & 2 \\ 4 & 2 & 8 \\ 3 & 7 & 6 \end{pmatrix} \tag{3.14}$$

Le rang matriciel de $\mathbf{A}$ est $r_{\mathbf{A}} = 2$, puisque les vecteurs associés à la $1^{\text{ère}}$ et à la $3^{\text{ème}}$ colonne sont colinéaires. Le rang de Kruskal de la matrice $\mathbf{A}$ est $k_{\mathbf{A}} = 1$, vu qu'il n'y a que les ensembles d'une colonne qui sont tous linéairement indépendants.

Exemple 2 :

$$\mathbf{A} = \begin{pmatrix} 1 & 4 & 3 \\ 2 & 4 & 2 \\ 3 & 7 & 4 \end{pmatrix} \tag{3.15}$$

Le rang matriciel de $\mathbf{A}$ est 2, puisque la $2^{\text{ème}}$ colonne est la somme de la $1^{\text{ère}}$ et la $3^{\text{ème}}$ colonne. Donc le rang de la matrice elle même n'est pas plein. Par contre, tous les ensembles de 2 colonnes sont linéairement indépendants. Ainsi, le rang de Kruskal de $\mathbf{A}$ est 2.

Le résultat présenté par Kruskal (KRUSKAL, 1977; Kruskal, 1989), est fortement lié au concept du rang tensoriel. Il a imposé une condition suffisante qui implique l'unicité

de la décomposition CP pour un tenseur d'ordre 3. Cette condition est donnée par le théorème :

Théorème 1. *Soit un tenseur d'ordre* 3 *et de rang R, qui se décompose sous la forme (3.6). Si*

$$2R + 2 \leq k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \tag{3.16}$$

alors cette décomposition est essentiellement unique. $k_{\mathbf{A}}$*,* $k_{\mathbf{B}}$ *et* $k_{\mathbf{C}}$ *représentent le rang de Kruskal des trois matrices* $\mathbf{A}$*,* $\mathbf{B}$ *et* $\mathbf{C}$ *respectivement.*

Nous ferons référence à la partie droite de l'inéquation sous le nom de *borne de Kruskal*. La preuve de ce théorème a d'abord été donnée à l'ordre 3 pour des tenseurs réels (KRUSKAL, 1977), puis plus tard à l'ordre 3 pour des tenseurs complexes (Sidiropoulos *et al.*, 2000b), et à tout ordre pour des tenseurs complexes (Sidiropoulos et Bro, 2000).
La décomposition CP d'un tenseur $\mathcal{T} \in \mathbb{C}^{I_1,\ldots,I_N}$ peut s'exprimer ainsi :

$$t_{i_1,i_2,\ldots,i_N} = \sum_{r=1}^{R} u^{(1)}_{i_1 r} u^{(2)}_{i_2 r} \cdots u^{(N)}_{i_N r} = \sum_{r=1}^{R} \prod_{n=1}^{N} u^{(n)}_{i_n r}, \tag{3.17}$$

tel que $u^{(n)}_{i_n r}$ sont les éléments des matrices $\mathbf{U}^{(n)}$, $i_n = 1, \ldots, I_n$, $n = 1, \ldots, N$. Dans (Sidiropoulos et Bro, 2000), Sidiropoulos et Bro ont imposé une condition suffisante qui implique l'unicité de la décomposition CP pour un tenseur d'ordre N :

$$2R + (N-1) \leq \sum_{n=1}^{N} k_{\mathbf{A}^{(n)}} \tag{3.18}$$

On remarque bien que pour le $N^{\text{ème}}$ ordre, les conditions nécessaires découlent directement de celles du troisième ordre données dans l'équation (3.16).

3.3 La décomposition CP et les indéterminations d'échelle et de permutation

De nombreuses décompositions tensorielles sont présentes dans la littérature. Elles divergent en fonction des données pour lesquelles elles sont prédestinées. Dans notre cas, la décomposition tensorielle qui nous interessera est la CP. Comme présentée dans la section 3.2, l'unicité de la décomposition CP peut être prouvée, mais qu'à deux indéterminations près.

Dans cette section, nous présenterons les deux indéterminations de la décomposition tensorielle CP, ainsi que les solutions existantes dans la littérature jusqu'à nos jours. Ensuite, nous introduirons la décomposition tensorielle CP avec isolation du facteur d'échelle. Ce facteur sera calculé de deux manières différentes : Dans un premier temps, nous calculerons le facteur d'échelle comme étant le produit des normes des matrices facteurs. La deuxième solution que nous proposerons consiste à trouver le facteur d'échelle optimal qui minimise la fonction objectif (3.10). Après le nouveau conditionnement du problème, nous prouverons l'existance et l'unicité de la décomposition CP avec le facteur d'échelle optimal.

3.3.1 Indéterminations d'échelle et de permutation

La décomposition CP (3.6) n'est unique qu'à deux indéterminations près : l'indétermination d'échelle et l'indétermination de permutaion. En effet, la décomposition (3.6) peut aussi s'écrire :

$$\mathcal{T} = \sum_{r=1}^{R} (\frac{1}{\lambda_r(\mathbf{A})\lambda_r(\mathbf{B})\lambda_r(\mathbf{C})})(\lambda_r(\mathbf{A})\mathbf{a}_r \circ \lambda_r(\mathbf{B})\mathbf{b}_r \circ \lambda_r(\mathbf{C})\mathbf{c}_r) \tag{3.19}$$

où λ_r est le facteur d'échelle. Si $\lambda_r(\mathbf{A})\lambda_r(\mathbf{B})\lambda_r(\mathbf{C}) = 1$, les vecteurs $\lambda_r(\mathbf{A})\mathbf{a}_r$, $\lambda_r(\mathbf{B})\mathbf{b}_r$ et $\lambda_r(\mathbf{C})\mathbf{c}_r$) sont donc des solutions au même titre que les vecteurs $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$. Il s'agit de l'ambiguïté d'échelle. De plus, l'ordre des termes $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$ est arbitraire. Rien ne garantit que cet ordre sera respecté pendant la reconstruction de celles-ci. Il s'agit de l'ambiguïté de permutation.

En d'autres termes, les matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ sont uniques à une permutation près, de même qu'à un facteur d'échelle près des colonnes des matrices (Sidiropoulos *et al.*, 2000b). Cela signifie que toutes matrices $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ et $\tilde{\mathbf{C}}$ satisfaisant l'équation (3.8) sont liées à $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ par :

$$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1, \qquad \tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2, \qquad \tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3 \tag{3.20}$$

où $\mathbf{\Pi}$ est la matrice des permutation et $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ et $\mathbf{\Lambda}_3$ sont les matrices d'échelle diagonales, tel que $\mathbf{\Lambda} = \text{Diag}\{\lambda_1, \ldots, \lambda_R\}$ satisfaisant la condition :

$$\Lambda_1\Lambda_2\Lambda_3 = \mathbf{I} \tag{3.21}$$

avec $\mathbf{I}$ la matrice d'identité de taille $R \times R$. Dans la littérature des tenseurs, l'optimisation de la décomposition CP (3.6) est effectuée sans isoler le facteur d'échelle $\mathbf{\Lambda}$, qui est généralement compris dans l'une des matrices facteur, de telle sorte à avoir $\mathbf{\Lambda} = \mathbf{I}$. Notre première proposition est de mettre les facteurs λ_r à l'extérieur du produit. Cela nous permet de contrôler le conditionnement du problème. Les indéterminations d'échelle sont alors nettement réduites au norme unité, mais ne sont pas complètement fixées, d'où la difficulté d'estimer l'erreur d'identification des matrices facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$. Notre deuxième proposition est de calculer les $3R$ phases complexes (réduit au signes dans le cas réel).

3.3.2 Existence et unicité de la décomposition CP en isolant le facteur d'échelle

L'objectif est d'identifier tous les paramètres du côté droit de l'équation (3.8), étant donné le tenseur $\mathcal{T}$. Selon les résultats existants (Sidiropoulos *et al.*, 2000a; Berge et Sidiropoulos, 2002; KRUSKAL, 1977), un tenseur du troisième ordre et de rang R peut être représenté de façon unique comme somme de R tenseurs de rang 1 sous certaines conditions. Comme mentionné dans la section 3.2.2.3, Kruskal a démontré que la condition $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2$ est suffisante pour garantir l'unicité de la décomposition CP (KRUSKAL, 1977). Cela signifie que sous la condition de Kruskal, les matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ sont uniques à une permutation près et à un facteur d'échelle de ses colonnes.

Pour l'unicité, Harshman a montré qu'il est suffisant d'avoir $\mathbf{A}$ et $\mathbf{B}$ de rang plein, et $\mathbf{C}$ de k-rang ≥ 2 (R.Harshman, 1970). Lorsque $1 < R \leq 2$, les conditions de Kruskal et de

Harshman sont équivalentes. Si $R > 2$, la condition de Kruskal peut être satisfaite même si celle de Harshman ne l'est pas, et cette condition est prétendue être juste suffisante pour $R > 3$ (Berge et Sidiropoulos, 2002). Cependant, les observations sont effectivement corrompues par le bruit, de sorte que l'équation (3.6) ne tient pas exactement.

3.3.2.1 Approximation de rang inférieur

Calculer la décomposition CP de $\mathcal{T}$ consiste à estimer les matrices $\mathbf{A}$,$\mathbf{B}$ et $\mathbf{C}$, ce qui peut être fait par la minimisation de la fonction de coût suivante :

$$\Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{\Lambda}) = \|\mathcal{T} - (\mathbf{A}, \mathbf{B}, \mathbf{C}).\mathbf{\Lambda}\|_F^2 \tag{3.22}$$

tel que $\mathbf{\Lambda}$ denote un tenseur diagonal de taille $R \times R \times R$, dont les éléments diagonaux sont λ_r, et $\{(\mathbf{A}, \mathbf{B}, \mathbf{C}).\mathbf{\Lambda}\}$ est un ersatz d'un tenseur de trois dimensions de coordonnées : $\sum_r \lambda_r a_{ir} b_{jr} c_{kr}$. Jusqu'à maintenant, le choix de la norme n'a pas été spécifié, mais nous allons utiliser par la suite la norme de Frobenius pour les matrices ou les tenseurs : $\|\mathcal{T}\|_F^2 = \sum_{ijk} |T_{ijk}|^2$, et la norme $\mathcal{L}^2$ pour les vecteurs.

Minimiser l'erreur quadratique (3.22) revient à trouver le meilleur rang R approximatif de $\mathcal{T}$ et sa décomposition CP (Comon et Lim, 2011). La fonction de coût (3.22) peut aussi s'écrire sous trois formes compactes équivalentes relativement aux trois matrices facteurs :

$$\begin{aligned} \Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{\Lambda}) &= \|\mathbf{T}_1^{I,KJ} - \mathbf{A}\mathbf{\Lambda}(\mathbf{C} \odot \mathbf{B})^T\|_F^2, \\ &= \|\mathbf{T}_2^{J,KI} - \mathbf{B}\mathbf{\Lambda}(\mathbf{C} \odot \mathbf{A})^T\|_F^2, \\ &= \|\mathbf{T}_3^{K,JI} - \mathbf{C}\mathbf{\Lambda}(\mathbf{B} \odot \mathbf{A})^T\|_F^2. \end{aligned} \tag{3.23}$$

3.3.2.2 Conditionnement du problème

Dans (Kolda et Bader, 2009), Kolda et al ont proposé de réduire les indéterminations en normalisant les vecteurs et stocker les normes en $\mathbf{\Lambda}$. Notre première proposition est de tirer les facteurs λ_R à l'extérieur du produit, et calculer la valeur optimale du facteur d'échelle, ce qui permet de contrôler le conditionnement du problème.

Supposant que les matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ sont données, nous allons calculer la valeur optimale du facteur d'échelle $\mathbf{\Lambda}$. Cela peut être fait en développant la norme de Frobenius dans l'équation (3.22), qui est une forme quadratique des entrées λ_r :

$$\Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{\Lambda}) = \|\mathcal{T}\|^2 - \sum_p \lambda_p f_p^* - \sum_q \lambda_q^* f_q + \sum_{pq} \lambda_p \lambda_q^* g_{pq}.$$

et en annulant le gradient par rapport à λ_r (voir détails dans l'Annexe A.1). Ensuite, le système linéaire suivant est obtenu :

$$\mathbf{G}\lambda = \mathbf{f}, \tag{3.24}$$

où $\mathbf{f}$ est un vecteur de dimension $1 \times R$ défini par la contraction : $f_r = \sum_{ijk} t_{ijk} a_{ir} b_{jr} c_{kr}$, et $\mathbf{G}$ représente la matrice de Gram de dimension $R \times R$ définie par :

$$g_{pq} = (\mathbf{a}_p \otimes \mathbf{b}_p \otimes \mathbf{c}_p)^H (\mathbf{a}_q \otimes \mathbf{b}_q \otimes \mathbf{c}_q) \tag{3.25}$$

Rappelons la définition d'une matrice de Gram.

Matrice de Gram Soit E un espace vectoriel euclidien et $\mathfrak{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ une famille de n vecteurs, la matrice de Gram de cette famille est la matrice formée par tous les produits scalaires $\langle \mathbf{u}_i, \mathbf{u}_j \rangle$. Son déterminant noté $Gram(\mathfrak{u})$ est appelé le déterminant de Gram de $\mathfrak{u}$.

Introduisons aussi le concept de cohèrence que nous utiliserons dans la suite de conditionnement de problème. La notion de cohérence a reçu différents noms dans la littérature : l'incohérence mutuelle de deux dictionnaires (Donoho et Elad, 2003), la cohérence mutuelle de deux dictionnaires (Candes et Romberg, 2007), la cohérence d'une projection de sous-espace (Candès et Terence, 2010), etc. La version ici suit celle de (Gribonval et Nielsen, 2003).

Cohérence la cohérence d'un ensemble de vecteurs de normes unitaires est définie comme la valeur maximale du module de produits scalaires croisées :

$$\mu_{\mathbf{A}} = \sup_{p \neq q} |\mathbf{a}_p^H \mathbf{a}_q| \tag{3.26}$$

Nous définissons de cette façon les cohérences $\mu_{\mathbf{A}}$, $\mu_{\mathbf{B}}$ et $\mu_{\mathbf{C}}$ associées aux matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, respectivement, avec $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$ leurs colonnes.

Au vue de la matrice $\mathbf{G}$, nous pouvons voir que les cohérences jouent un rôle dans le conditionnement du problème. À partir des équations (3.24) et (3.25), et puisque les entrées diagonales de $\mathbf{G}$ sont égales à 1, il est clair que le fait d'imposer aux produits scalaires croisés $\mathbf{a}_p^H \mathbf{a}_q$ d'avoir un module strictement inférieur à 1, conduira à un conditionnement acceptable (Comon *et al.*, 2013; Lim et Comon, 2010). Il est à noter également que les produits scalaires n'apparaissent pas individuellement (3.25), mais à travers leurs produits, puisque les entrées de $\mathbf{G}$ peuvent également être écrites comme : $g_{pq} = \mathbf{a}_p^H \mathbf{a}_q \mathbf{b}_p^H \mathbf{b}_q \mathbf{c}_p^H \mathbf{c}_q$. Cela a des implications très importantes, en particulier dans l'existence et l'unicité de la solution du problème (3.22), que nous élaborerons par la suite.

3.3.2.3 Existence

La fonction objectif (3.22) n'est pas coercive, ce qui explique pourquoi le minimum ne peut pas exister. En effet, ce phénomène où un tenseur ne parvient pas à avoir une meilleure approximation de rang r est beaucoup plus répandu qu'on ne l'imagine. Il se produit sur un large éventail de dimensions, des ordres et des rangs, et indépendamment du choix de la norme utilisée. L'exemple suivant représente mieux ce phénomène dans le cas d'un tenseur de rang 3.

Exemple 3.3.1. Soit $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{C}^J$, $i = 1, 2, 3$, $\mathcal{T} = \mathbf{a}_1 \circ \mathbf{a}_2 \circ \mathbf{b}_3 + \mathbf{a}_1 \circ \mathbf{b}_2 \circ \mathbf{a}_3 + \mathbf{b}_1 \circ \mathbf{a}_2 \circ \mathbf{a}_3$, alors $\forall n \in \mathbb{N}$:

$$\mathcal{T}_n = n(\mathbf{a}_1 + \frac{1}{n}\mathbf{b}_1) \circ (\mathbf{a}_2 + \frac{1}{n}\mathbf{b}_2) \circ (\mathbf{a}_3 + \frac{1}{n}\mathbf{b}_3) - n\mathbf{a}_1 \circ \mathbf{a}_2 \circ \mathbf{a}_3$$

On peut montrer que le rang de $\mathcal{T}$ est égal à 3 si $\mathbf{a}_i$ et $\mathbf{b}_i$ sont linéairement indépendants. Comme il est clair que le rang de $\mathcal{T}_n$ est inférieur à 2 par construction, et $\lim_{n\to\infty} \mathcal{T}_n = \mathcal{T}$. Le tenseur $\mathcal{T}$ de rang 3 n'a pas de meilleur approximation de rang 2 (Lim et Comon, 2010). Un tel tenseur est dit de rang dans la frontière 2 (*border rank 2*).

Notre but est d'empêcher le phénomène observé dans l'exemple 3.3.1 de se produire, et cela, en imposant des faibles contraintes sans réduire la recherche à un ensemble compact. Rappelons qu'une fonction réelle f avec un domaine non borné $dom(f)$ et dont la limite $\lim_{x\in dom(f),\|x\|\to+\infty} = +\infty$ est appelée coercive (ou 0 coercive). Un avantage principal de ces fonctions est que l'existence d'un minimum global est garantie.

Il est clair que la fonction objectif (3.22) n'est pas coercive, ce qui explique pourquoi le minimum ne peut pas exister. Mais avec une condition supplémentaire sur la cohérence, nous serons en mesure de prouver l'existence grâce à la coercivité. Le théorème suivant montre que la solution au problème (3.22) existe toujours en ajoutant la condition sur la cohérence.

Théorème 2. *Définissons trois cohérences $\mu_{\mathbf{A}}$, $\mu_{\mathbf{B}}$ et $\mu_{\mathbf{C}}$ associées aux matrices* **A**, **B** *et* **C**.

Si :

$$\mu_{\mathbf{A}}\mu_{\mathbf{B}}\mu_{\mathbf{C}} \leq \frac{1}{R-1}, \tag{3.27}$$

alors, la borne inférieure de l'équation (3.22) est atteinte.

Avec cette condition, l'existence d'une meilleure approximation de rang inférieur est prouvée (voir la démonstaration dans l'annexe A.2). On peut constater que cette condition donne déjà une limite quantitative pour le conditionnement de l'équation (3.24), en raison des cohérences qui limitent les entrées extra diagonales de la matrice **G**, possèdant des 1 sur sa diagonale (Comon *et al.*, 2013).

3.3.2.4 Unicité

Afin de relier l'unicité et la minimalité des décompositions multilinéaires à la cohérence, nous avons besoin d'une simple observation sur la notion de rang de Kruskal introduite dans la section 3.2.2.3.
Selon le Lemme proposé dans (Lim et Comon, 2010; Gribonval et Nielsen, 2003), on peut observer que $k_{\mathbf{A}} \geq \frac{1}{\mu_{\mathbf{A}}}$, aussi longtemps que $k_{\mathbf{A}}$ est strictement inférieur au rang des colonnes de **A**. Inclure cette inégalité dans l'équation (3.16), conduit à la condition unique et suffisante suivante pour un tenseur d'ordre 3 :

$$\mu_{\mathbf{A}}^{-1} + \mu_{\mathbf{B}}^{-1} + \mu_{\mathbf{C}}^{-1} \geq 2(R+1), \tag{3.28}$$

Une généralisation de cette condition pour des tenseurs d'ordre d sera comme suit :

$$\sum_{k=1}^{d} \mu_k^{-1} \geq 2R + d - 1 \tag{3.29}$$

3.3.3 Critère de performance proposé

La décomposition tensorielle CP en isolant la matrice des facteurs d'échelles $\mathbf{\Lambda}$ a deux conséquences majeures :

(1) Le problème de conditionnement se présente explicitepment, et pourrait être contrôlé par une contrainte sur les soi-disant cohérences,

(2) Un critère de performance concernant les matrices facteurs peut être exactement calculé, et est plus réaliste que celui utilisé dans la littérature.

Comme montré dans la section précédente, l'isolation de la matrice des facteurs d'échelle dans la décomposition CP a des conséquences profondes sur le conditionnement du problème dans lequel la cohérence joue un rôle majeur, spécialement sur l'existence et l'unicité de la solution du problème (3.22).
Deuxièmement, une mesure de performance qui ne concerne que les matrices facteurs peut être exactement calculée, et ne présente pas un biais optimiste de l'erreur minimale généralement utilisée dans la littérature.
En effet, il est bien connu que les matrices facteur sont identifiées à un facteur d'échelle près. Cette indétermination est compliquée à prendre en compte en raison de la contrainte suivante :

Le produit de toutes les matrices d'échelle doit être égal à l'identité.

Pour cette raison, seuls les indices de performance approximatifs ont été utilisés jusqu'à présent, en ignorant tout simplement la dernière contrainte (Kolda et Bader, 2009; Comon *et al.*, 2009).

Notre première proposition était de calculer la valeur optimale du facteur d'échelle, ce qui permet de contrôler le conditionnement du problème. Les indéterminations d'échelle sont alors nettement réduites au module unitaire, mais ne sont pas complètement fixées. Il y a toujours une indétermination dans la représentation des tenseurs décomposables, caractérisée dans la décomposition CP par $3R$ nombres complexes de module unité. D'où la difficulté d'estimer l'erreur d'identification des matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$. Notre deuxième proposition est de calculer les $3R$ phases complexes.

Afin de mieux comprendre ce problème, soit $\mathbf{a}$, $\mathbf{b}$ et $\mathbf{c}$ les $r^{\text{ème}}$ colonnes des matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ respectivement, avec $1 \leq r \leq R$. De plus, $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$ et $\hat{\mathbf{c}}$ désignent les colonnes des matrices d'entrées estimées de la décomposition CP. Nous cherchons à minimiser la distance :

$$\delta(\mathbf{x};\hat{\mathbf{x}}) = \min_{\varphi,\psi,\chi}\{||\mathbf{a} - e^{\jmath\varphi}\hat{\mathbf{a}}||^2 + ||\mathbf{b} - e^{\jmath\psi}\hat{\mathbf{b}}||^2 + ||\mathbf{c} - e^{\jmath\chi}\hat{\mathbf{c}}||^2\}, \tag{3.30}$$

Dans la littérature, seul le critère de performance approximatif a été utilisé jusqu'à présent en négligeant la relation entre les angles φ, ψ et χ, donnée par :

$$\exp(\jmath(\varphi + \psi + \chi)) = 1,$$

Notre contribution consiste ici à trouver la distance minimale exacte (3.30) sous cette contrainte angulaire, en calculant les $3R$ phases optimales affectant les colonnes des matrices facteur estimées. Pour plus de commodité, $\mathbf{x}$ désigne le vecteur $[\mathbf{a}^T;\mathbf{b}^T;\mathbf{c}^T]^T$. Il s'avère que cette distance peut être exactement calculée. En fait, en raison de la contrainte, la fonction (3.30) comporte deux angles φ et ψ, puisque à partir de la contrainte angulaire on a :

$$\chi = -\varphi - \psi[2\pi] \tag{3.31}$$

donc, l'équation (3.30) peut être réécrite comme :

$$\begin{aligned}\delta \;=\; & ||\mathbf{a}||^2 + ||\hat{\mathbf{a}}||^2 + ||\mathbf{b}||^2 + ||\hat{\mathbf{b}}||^2 + ||\mathbf{c}||^2 + ||\hat{\mathbf{c}}||^2 \\ & -2\rho_{\mathbf{a}}\cos(\varphi - \alpha) - 2\rho_{\mathbf{b}}\cos(\psi - \beta) \\ & -2\rho_{\mathbf{c}}\cos(\varphi + \psi + \gamma),\end{aligned} \tag{3.32}$$

où $\mathbf{a}^H\hat{\mathbf{a}} \overset{\text{def}}{=} \rho_{\mathbf{a}}\, e^{\jmath\alpha}$, $\mathbf{b}^H\hat{\mathbf{b}} \overset{\text{def}}{=} \rho_{\mathbf{b}}\, e^{\jmath\beta}$ et $\mathbf{c}^H\hat{\mathbf{c}} \overset{\text{def}}{=} \rho_{\mathbf{c}}\, e^{\jmath\gamma}$.

La dérivée de l'équation (3.32) par rapport à φ et ψ donne un système de deux équations. Trouver les $3R$ phases revient à résoudre le système de deux équations suivant :

$$\begin{cases} \rho_{\mathbf{a}} \sin x + \rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) &= 0, \\ \rho_{\mathbf{b}} \sin y + \rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) &= 0. \end{cases} \tag{3.33}$$

tel que $x = \varphi - \alpha$ et $y = \psi - \beta$.

La première simplification est obtenue en notant que :

$$\sin y = \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin x,$$

Maintenant, et en utilisant des identités trigonométriques, on peut réécrire la première équation du système (3.33) comme :

$$\rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) = \rho_{\mathbf{c}} \sin(x + y + \alpha + \beta + \gamma) = -\rho_{\mathbf{a}} \sin x,$$

Après quelques manipulations trigonométriques décrites à l'annexe A.3, les solutions peuvent être obtenues en résolvant un polynôme de degré 6 en une seule variable φ qui a la forme suivante :

$$\begin{aligned} &c_0 + c_1 \cos(2x) + c_2 \cos^2(2x) + c_3 \cos^3(2x) \\ &+c_4 \cos^4(2x) + c_5 \cos^5(2x) + c_6 \cos^6(2x) = 0, \end{aligned} \tag{3.34}$$

avec $c_0, c_1, c_2, c_3, c_4, c_5, c_6$, les sept coefficients du polynôme dont leurs valeurs sont données en détail dans l'annexe A.3. La résolution de l'équation (3.34) de degré six nous donnera x, et donc la première variable φ. En remplaçant les valeurs admissibles de φ dans le système (3.33), nous obtenons les valeurs correspondantes de ψ. Le calcul de la distance minimale (3.30) est fait pour toutes les permutations possibles. Finalement, nous nous retrouvons avec le critère de performance suivant :

$$\mathcal{E}(\mathcal{T}; \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{\Lambda}) = \min_{\pi \in \Pi} \sum_{r=1}^{R} \delta(\mathbf{x}_r; \hat{\mathbf{x}}_{\pi(r)}), \tag{3.35}$$

où π est l'ensemble des permutations de $\{1, 2, \ldots, R\}$. Lorsque la permutation agit en très grande dimension, l'utilisation de l'algorithme de Glouton, que nous rappelons son principe par la suite, est possible pour limiter la recherche exhaustive dans l'ensemble des permutations.

Algorithme de Glouton Pour un problème d'optimisation, un algorithme Glouton est un algorithme qui cherche à construire une solution optimale pas à pas, sans jamais revenir sur ses décisions, en prenant à chaque étape la solution qui semble la meilleure localement.

Lorsqu'un problème peut se résoudre à l'aide d'un algorithme Glouton, la solution rendue par l'algorithme n'est pas forcément optimale. Pour assurer que l'algorithme Glouton donne une solution optimale pour un problème, il faut montrer que ce problème a les propriétés suivantes :

1. *Propriété du choix glouton* : Il existe toujours une solution optimale commençant par un choix glouton, c'est-à-dire qu'un choix optimal local peut mener à une solution optimale globale.
2. *Propriété de sous-structure optimale* : Trouver une solution optimale contenant le premier choix glouton se réduit à trouver une solution optimale pour un sous-probème de même nature.

Dans les cas des tenseurs d'ordre N, il suffit de résoudre un système de polynôme de degré 2 de N variables.

Notons $\pi(i)$ les permutations de Π, avec $1 \leq i \leq R!$. L'algorithme de calul de nouveau critère de performance se résume comme suit :

1. Pour $1 \leq i \leq R!$, faire :
2. Calculer $3R$ phases optimales affectant les colonnes des matrices de facteur estimées :
   (a) Permuter les colonnes des trois matrices estimées selon la permutation $\pi(i)$: $\widehat{\mathbf{A}}_{\pi(i)}$, $\widehat{\mathbf{B}}_{\pi(i)}$, et $\widehat{\mathbf{C}}_{\pi(i)}$;
   (b) Pour chaque rth colonnes de matrices de facteur et matrices estimées, faire :
   – Mettre $x = \varphi - \alpha$ et $y = \psi - \beta$ et résoudre le polynôme de degré 6 en une seule variable φ :

$$\begin{aligned} c_0 \;&+\; c_1\cos(2x) + c_2\cos^2(2x) + c_3\cos^3(2x) \\ &+\; c_4\cos^4(2x) + c_5\cos^5(2x) + c_6\cos^6(2x) = 0. \end{aligned}$$

   – Remplacer x dans $\sin y = \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\sin x$, obtenir y et par conséquent ψ ;
   – Calculer χ dans : $\exp(\jmath(\varphi + \psi + \chi)) = 1$;
   – Calculer la distance minimale δ :

$$\delta(\mathbf{x}; \hat{\mathbf{x}}) = \min_{\varphi,\psi,\chi}\{||\mathbf{a} - e^{\jmath\varphi}\hat{\mathbf{a}}||^2 + ||\mathbf{b} - e^{\jmath\psi}\hat{\mathbf{b}}||^2 + ||\mathbf{c} - e^{\jmath\chi}\hat{\mathbf{c}}||^2\}.$$

   – Sauvegarder les résultats : distance$(i) = \delta$, phase$_\varphi(i) = \varphi$, phase$_\psi(i) = \psi$ et phase$_\chi(i) = \chi$;
3. fin pour.
4. Choisir les $3R$ angles qui retournent la plus petite distance δ.

3.4 conclusion

Dans ce chapitre, nous avons fourni quelques éléments fondamentaux de l'algèbre multilinéaires et des décompositions tensorielles. Ces outils mathématiques offrent une meilleure capacité de modélisation pour des données tridimensionnelles.

Dans la première partie de ce chapitre, nous avons présenté plusieurs décompositions tensorielles qui sont importantes dans le contexte de cette thèse. Ces décompositions ont été formulées à la fois en scalaire (multi-indexé) et en tenseur (produit exterieur ou un produit mode-n) formes. La propriété d'unicité de ces décompositions a été discutée, et ses conditions ont été présentées pour les différentes décompositions.

Le chapitre contient des contributions originales qui sont le développement d'une nouvelle décomposition tensorielle basée sur la décomposition CP, nommée la décomposition

CP avec isolement du facteur d'échelle. Le principe de cette décomposition, comme décrit dans ce chapitre, consiste à calculer la valeur optimale du facteur d'échelle que nous isolons, et qui est mis en dehors des matrices facteurs dans la décomposition CP. Nous avons montré que dans la décomposition tensorielle CP, le conditionnement du calcul de la matrice de facteur d'échelle optimale $\mathbf{\Lambda}$ dépend de la cohérence via une matrice de Gram. Cela a des implications très importantes sur l'existance et l'unicité d'une décomposition CP approximative du tenseur $\mathcal{T}$. En effet, en introduisant une condition sur la cohérence, nous garantissons l'existence d'une décomposition approximative de rang inférieur. De plus, nous avons donné une condition suffisante qui prouve l'unicité de la décomposition tensorielle.

L'isolement de la matrice de facteur d'échelle permet de réduire les indéterminations d'échelle au module unitaire et non pas les fixer complètement, d'où la difficulté d'estimer l'erreur d'identification des matrices facteurs. Notre deuxième contribution dans ce chapitre était de calculer l'indice de performance exacte plus réaliste que les critères de performance utilisées dans la littérature qui sont optimistes par construction. Comme souligné dans le chapitre, l'indétermination est caractérisée dans la décomposition CP par $3R$ nombres complexes de module unitaire. Notre contribution consistait à trouver la distance minimale exacte sous une contrainte angulaire, en calculant les $3R$ phases optimales.

Chapitre

4 Les algorithmes d'optimisation de la décomposition CP

4.1 Introduction

En pratique, et comme présenté dans le chapitre précédent, on préfère souvent s'adapter à un modèle multi-linéaire de rang inférieur, $R < rangT$, fixé à l'avance, afin de faire face à un problème d'approximation (3.22). Il est utile de trouver des algorithmes d'optimisation qui minimisent le problème (3.22), tout en prenant en compte le facteur d'échelle optimal Λ. En effet, comme présenté dans (Comon *et al.*, 2013; Rouijel *et al.*, 2014a,b), l'isolement du facteur d'échelle dans la décomposition CP a plusieurs implications, spécialement sur l'existence et l'unicité de la solution de (3.22). Donc, il faut trouver des algorithmes d'optimisation qui prennent en compte deux contraintes : une contrainte d'égalité sur les colonnes des matrices de facteurs, et une autre d'inégalité sur les cohérences. Dans ce chapitre, nous proposerons plusieurs algorithmes de calcul des décompositions tensorielles.

Dans la section 4.2, une étude bibliographique sur les problèmes d'optimisation sera faite, ainsi que les différents types des méthodes d'optimisation seront étudiés. Ensuite, nous étudierons l'algorithme des moindres carrés alternés ALS (*Alternating Least Squares*), fréquement utilisé dans la littérature pour optimiser la décomposition tensorielle CP (BRO, 1997; Sidiropoulos *et al.*, 2000a; Comon *et al.*, 2009).
Àtravers le présent chapitre, on s'attachera à une description plus spécifique des algorithmes itératifs (ou méthodes itératives) qui ont été implémentés et qui permettent la résolution des problèmes d'optimisation. Il convient de souligner que la plupart des algorithmes d'optimisation, avec contrainte ou non, fonctionnent selon un schéma général consistant, à chaque itération, à se rapprocher du minimum par la résolution d'un sous-problème de minimisation. Dans la section 4.4, nous considérerons les méthodes permettant de résoudre un problème d'optimisation sans contraintes (appelées aussi parfois méthodes d'optimisation directes), que nous décriverons. Parmi lesquelles :

1. Les méthodes basées sur le gradient
2. Les méthodes de Newton
3. Les méthodes utilisant des directions conjuguées

Ces méthodes utilisent des dérivées, à l'exception des méthodes de directions conjuguées (sauf dans le cas particulier de la méthode du gradient conjugué).
La section 4.7 est dédiée aux algorithmes que nous proposerons pour minimiser la fonction (3.22). Dans un premier temps, nous proposerons des algorithmes basés sur la descente du gradient pour minimiser la fonction du coût sous une contrainte d'égalité. Par la suite,

nous introduirons un algorithme d'optimisation qui minimisera cette fois-ci notre fonction de coût sous deux contraintres, d'égalité et d'inégalité.

4.2 Étude bibliographique

Un problème d'optimisation est usuellement formulé comme un problème de minimisation, et écrit sous la forme :

$$\begin{cases} \min_x f(x), & \\ tel\ que, & \\ g_i(x) \leq 0\,, & i = 1, \ldots, m, \\ h_j(x) = 0\,, & j = 1, \ldots, p, \\ x \in \mathcal{S} \subset \mathbb{R}^n, & \end{cases} \tag{4.1}$$

où f est la fonction (scalaire) à minimiser, appelée fonction coût ou fonction objectif, x représente le vecteur des variables d'optimisation, g_i sont les contraintes d'inégalité et h_j les contraintes d'égalité, et $\mathcal{S}$ est l'espace des variables (appelé aussi espace de recherche). $\mathcal{S}$ indique le type de variables considérées : réelles, entières, mixtes (réelles et entières dans un même problème), discrètes, continues, bornées, etc.

Un point x_A est appelé un point admissible si $x_A \in \mathcal{S}$ et si les contraintes d'optimisation sont satisfaites : $g_i(x_A) \leq 0\,, i = 1, \ldots, m$ et $h_j(x_A) = 0\,, j = 1, \ldots, p$. La solution de (4.1) est l'ensemble des optima $\{x^*\}$. Dans ce mémoire, on s'intéressera plutôt aux problèmes d'optimisation en variables réelles et complexes, avec des contraintes d'égalité et d'inégalité.

x^* est un minimum global de f si et seulement si $f(x^*) \leq f(x)\ \forall x \in \mathcal{S}$, et x^* est un minimum local de f si et seulement si $f(x^*) \leq f(x)\ \forall x \in \mathcal{S}\ /\ \|x - x^*\| \leq \epsilon$, $\epsilon > 0$. La Figure (4.1) présente un exemple d'une fonction à une variable, avec des minima locaux et un minimum global. Parmi les minima locaux, celui qui possède la plus petite valeur de f est le minimum global.

Une fonction multimodale présente plusieurs minima (locaux et globaux), alors qu'une fonction unimodale n'a qu'un minimum, le minimum global. La Figure (4.2) illustre une fonction multimodale à deux variables.

On appelle méthode (ou algorithme ou recherche) locale celle qui converge vers un minimum local. Les recherches locales partent usuellement d'un point initial x_0 avec un pas initial ρ_0. Ces paramètres vont conditionner la descente d'une des vallées de la fonction.

La méthode d'optimisation est conditionnée par des paramètres de contrôle et des conditions initiales (valeurs initiales des variables de conception et des paramètres de contrôle). Elle peut être caractérisée selon le modèle illustré en Figure (4.3).

L'efficacité d'une méthode d'optimisation est liée à la sensibilité et à la robustesse par rapport aux paramètres de contrôle et aux conditions initiales. Lorsque les variables de conception doivent prendre une valeur bien précise pour que la méthode de résolution converge vers l'optimum d'une fonction donnée, la méthode est dite sensible aux conditions initiales. Une méthode d'optimisation est robuste si pour une même valeur des paramètres de contrôle et des conditions initiales, elle est capable de trouver l'optimum de différentes

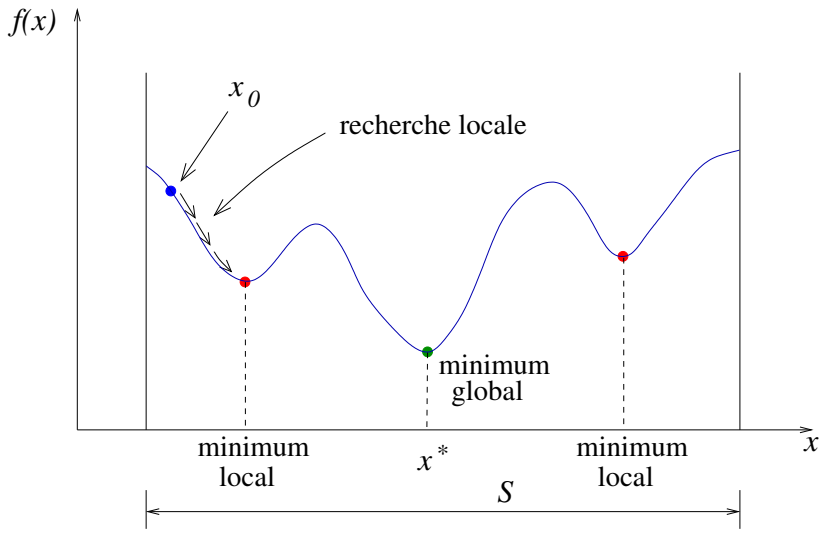

FIGURE 4.1 – Minima locaux et minimum global d'une fonction à une variable.

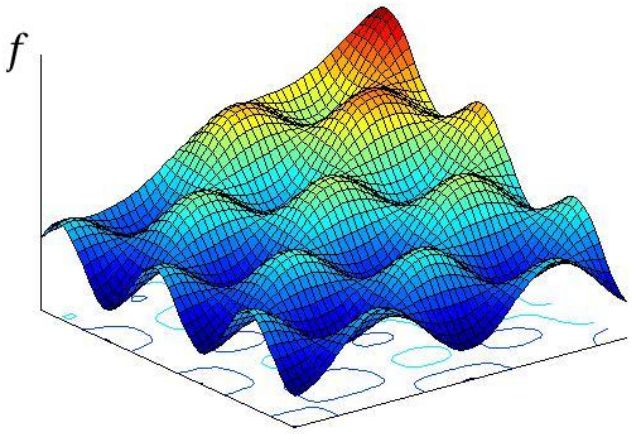

FIGURE 4.2 – Exemple de fonction multimodale à deux variables.

fonctions. Une méthode parfaite devrait être totalement insensible aux conditions initiales et aux variables de conception, et converger vers l'optimum quelles que soient la fonction objectif et les contraintes.

4.2.1 Opérateurs de recherche fondamentaux

En général, la recherche de l'optimum d'une fonction est réalisée à l'aide de deux opérateurs fondamentaux : l'exploration et l'exploitation. Le premier opérateur, qui est l'exploration, permet une localisation imprécise de l'optimum global, alors que le deuxième opérateur, l'exploitation, affine cette solution en augmentant la précision de l'optimum.

Le succès et l'efficacité d'une technique de résolution dépendent la plupart du temps d'un compromis entre l'exploration et l'exploitation. Certaines méthodes toutefois n'utilisent qu'un seul de ces opérateurs pour parvenir à l'optimum. Ainsi, les méthodes déterministes, exploitant les dérivées de la fonction objectif et des contraintes pour atteindre rapidement et précisément le minimum local le plus proche du point de départ, privilégient l'exploitation au détriment de l'exploration.

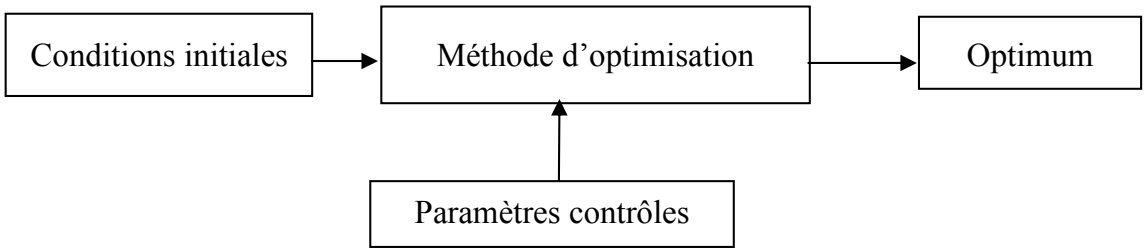

FIGURE 4.3 – Modèle des méthodes d'optimisation.

Tout algorithme d'optimisation doit utiliser ces deux stratégies pour trouver l'optimum global : l'exploration pour la recherche de régions inexplorées de l'espace de recherche, et l'exploitation pour exploiter la connaissance acquise aux points déjà visités et ainsi trouver des points meilleurs. Ces deux exigences peuvent paraître contradictoires, mais un bon algorithme de recherche doit trouver le bon compromis entre les deux. Une recherche purement aléatoire est bonne pour l'exploration, mais pas pour l'exploitation, alors que la recherche dans le voisinage est une bonne méthode d'exploitation mais pas d'exploration.

4.2.2 Ordre d'une méthode d'optimisation

Les méthodes d'optimisation peuvent être classées à partir de leur ordre selon leurs nécessitées ou non du calcul des dérivées de la fonction objectif et des contraintes par rapport aux paramètres.

Une méthode est dite d'ordre zéro si elle utilise uniquement la connaissance de la fonction elle-même. Elle est d'ordre un si elle requiert le calcul des dérivées premières et d'ordre deux s'il lui faut aussi accéder aux dérivées secondes. Les méthodes d'ordre zéro sont en général peu précises et convergent plus lentement vers l'optimum. En revanche, elles offrent l'avantage d'éviter le calcul du gradient, ce qui est intéressant lorsque la fonction n'est pas différentiable, ou si le calcul de son gradient représente un coût important. C'est notamment le cas des modèles à éléments finis.

Les méthodes d'ordre un permettent d'accélérer la localisation de l'optimum, puisque le gradient donne l'information sur la direction de l'amélioration (direction de recherche). Par contre, elles sont applicables seulement aux problèmes où les fonctions objectifs et les contraintes sont différentiables.

4.2.3 Classification des méthodes d'optimisation

Les méthodes d'optimisations sont classées, selon le mode de recherche de l'optimum, en deux grands groupes : les méthodes déterministes et les méthodes non déterministes.

4.2.3.1 Les méthodes déterministes

Les méthodes déterministes sont généralement efficaces quand l'évaluation de la fonction est très rapide, ou quand la forme de la fonction est connue a priori. Les cas les plus complexes (temps de calcul important, nombreux optima locaux, fonctions non-dérivables,

fonctions fractales, fonctions bruitées) seront souvent traités plus efficacement par des méthodes non-déterministes.

Ces méthodes peuvent être subdivisées en plusieurs sous classes : les méthodes heuristiques, les méthodes statistiques, les méthodes Branch et Bound, les méthodes mathématiques, et les méthodes d'apprentissage automatique. Cette classification est illustrée en table (4.1). La classe qui nous intéresse dans ce manuscrit est celle des méthodes mathématiques.

Classe	**Algorithme**	**Ordre**	**Glocal \ Local**
Méthodes heuristiques	Simplex	0	Local
Plans d'expériences		0	Global
Méthode Branch et Bound		0	Global
Méthodes mathématiques	Direction conjuguée	0	Local
-	Plus grande pente	≥ 1	Local
-	Gradient conjugué	≥ 1	Local
-	Quasi Newton	≥ 1	Local
Méthodes d'apprentissage automatique	Réseaux de neurones	0	Local

TABLE 4.1 – Méthodes d'optimisation déterministes

La recherche des extrema d'une fonction f à n variables $(x_1, x_2, \ldots, x_n)$ revient à résoudre un système de n équations à n inconnus, linéaires ou non linéaires :

$$\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = 0 \tag{4.2}$$

tel que $i = 1, \ldots, n$.

En général, les techniques classiques de résolution sont des méthodes simples à comprendre et faciles à mettre en œuvre, mais leur utilisation nécessite comme étape préliminaire la localisation des extremas. La fonction objectif dans ce cas doit être continue et dérivable.

Cette exigence n'est pas toujours vérifiée dans le cas de problèmes réels. La nature dynamique et discontinue des problèmes rend l'utilisation des méthodes classiques limitée. Parmi les méthodes classiques, on peut citer la méthode du gradient, les méthodes Quasi-Newton, la méthode du gradient conjugué...etc, que nous détaillerons dans la suite de ce chapitre.

4.2.3.2 Les méthodes non déterministes

Les méthodes d'optimisation non déterministes, appelées aussi méthodes stochastiques, s'appuient sur des mécanismes de transition probabilistes et aléatoires. Cette caractéristique indique que plusieurs exécutions successives de ces méthodes peuvent conduire à des résultats différents, pour une même configuration initiale d'un problème d'optimisation.

Ces méthodes ont une grande capacité à trouver l'optimum global du problème. Contrairement à la plupart des méthodes déterministes, elles ne nécessitent ni point de départ, ni la connaissance du gradient de la fonction objectif pour atteindre la solution

optimale. Elles sont d'ordre zéro. Cependant, elles demandent un nombre important d'évaluations de la fonction objectif. La Figure (4.4) présente les méthodes stochastiques les plus utilisées.

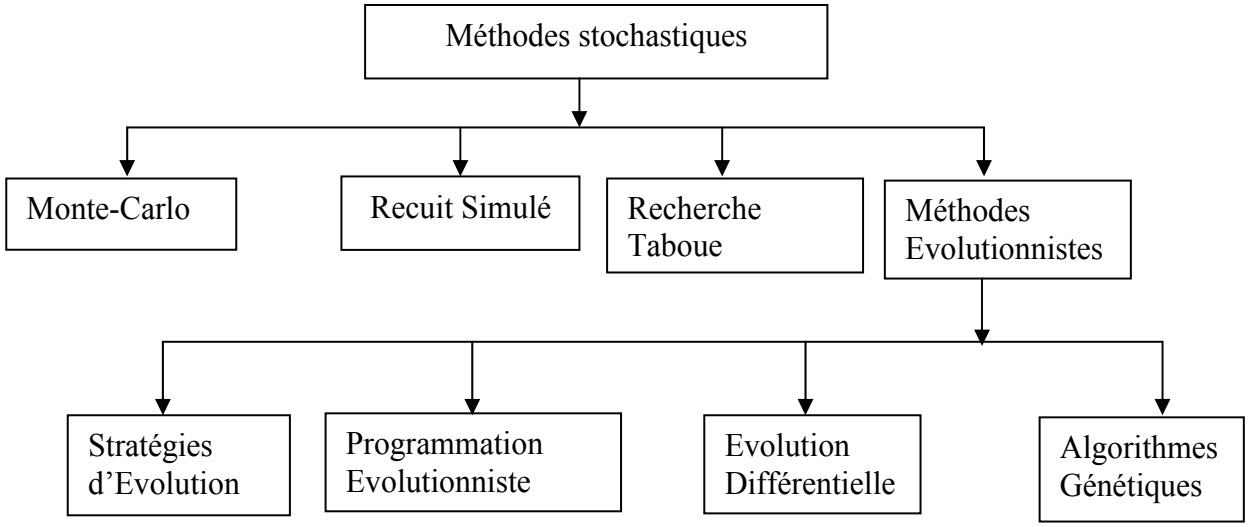

FIGURE 4.4 – Les méthodes d'optimisation non-déterministes.

Dans les problèmes d'optimisation, on cherche à minimiser une fonction qui peut être complexe, coûteuse à estimer, et dont les dérivées ne sont pas toujours disponibles. À ces difficultés s'ajoutent la prise en compte des contraintes non linéaires, et parfois l'aspect multi-objectifs, d'où l'existence de deux types de méthode ou algorithme d'optimisation :

- Algorithmes d'optimisation sans contraintes
- Algorithmes d'optimisation avec contraintes

Dans ce manuscrit, nous nous intéresserons aux deux types d'algorithmes pour optimiser le calcul de la décomposition tensorielle CP.

4.3 Algorithmes des moindres carrés altérnés ALS

L'évaluation des trois matrices de la décomposition CP est généralement effectuée en minimisant la fonction de coût quadratique suivante :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C}) = \|\mathcal{T} - \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F^2 \tag{4.3}$$

Il existe plusieurs algorithmes d'optimisation qui ont été utilisés pour optimiser la décomposition tensorielle (4.3) sans contrainte. Dans cette section, nous discuterons les différentes méthodes ou algorithmes pour résoudre le problème de l'optimisation sans contrainte (4.3).

4.3.1 Algorithme ALS et la décomposition CP

Le principe général de cet algorithme est assez simple. Soit $\mathcal{T}$ un tenseur d'ordre N suivant une décomposition donnée, dont les N matrices de facteurs inconnues sont

$\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}$. Notons $\mathbf{T}^{(n)}$, $n = 1, \ldots, N$, les représentations matricielles du tenseur $\mathcal{T}$. Pour chaque $\mathbf{T}^{(n)}$, il existe une matrice $\mathbf{Z}^{(n)}$, construite à partir de $n-1$ matrices de l'ensemble $S = \{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}\} - \{\mathbf{A}^{(n)}\}$, qui vérifie :

$$\mathbf{T}^{(n)} = \mathbf{Z}^{(n)}\mathbf{A}^{(n)} \tag{4.4}$$

Étant donné que seul le tenseur $\mathcal{T}$ est connu, l'algorithme ALS consiste à exploiter les N équations du type (4.5) pour estimer les matrices $\mathbf{A}^{(n)}$ d'une manière alternée. Ainsi, une fois l'estimation de toutes les matrices de l'ensemble S faite, l'estimation de $\mathbf{A}^{(n)}$ au sens des moindres carrés est obtenue par :

$$\hat{\mathbf{A}}^{(n)} = (\mathbf{Z}^{(n)})^{\dagger}\mathbf{T}^{(n)} \tag{4.5}$$

Dans le cas d'un tenseur d'ordre 3, l'algorithme ALS est la solution classique pour minimiser cette fonction de coût (4.3) (BRO, 1997; Sidiropoulos *et al.*, 2000b; Smilde *et al.*, 2004). Il s'agit d'un algorithme itératif qui alterne entre l'estimation de $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$. Le principe de l'algorithme ALS est de mettre à jour de manière alternée chacune des matrices en gardant les deux autres fixées. Si deux matrices parmi les trois sont fixées, le système à résoudre devient alors un problème simple des moindres carrés (LS).

Chaque itération de l'algorithme ALS se compose de trois étapes d'estimation LS. À chaque étape, une matrice de facteurs, par exemple, $\mathbf{A}$ est mise à jour tandis que les deux autres ($\mathbf{B}$ et $\mathbf{C}$) sont fixées à leurs valeurs obtenues dans les étapes d'estimation précédentes. L'algorithme ALS utilise les factorisations *Khatri-Rao* des matrices dépliées $\mathbf{T}_1^{I,KJ}$, $\mathbf{T}_2^{J,KI}$ et $\mathbf{T}_3^{K,JI}$ données dans l'équation (3.9), et représentées dans la Figure 2.4. Chaque représentation permet d'estimer l'une des trois matrices au sens des moindres carrées. Ainsi, chaque mise à jour des matrices est obtenue par une simple inversion matricielle. Un synopsis de l'algorithme ALS est donné dans la table 4.2.

Algorithme ALS
1. Mettre $k = 0$, Initialiser $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)})$.
2. Pour $k \geq 1$ et jusqu'à vérification du critère d'arrê, faire :
– Estimer $\mathbf{A}^{(k)}$: $\hat{\mathbf{A}} = \mathbf{T}^{I,KJ}((\mathbf{C}^{(k-1)} \odot \mathbf{B}^{(k-1)})^{\dagger})^T$
– Mettre à jour $\mathbf{A}^{(k)} = \hat{\mathbf{A}}$
– Estimer $\mathbf{B}^{(k)}$: $\hat{\mathbf{B}} = \mathbf{T}^{J,KI}((\mathbf{C}^{(k-1)} \odot \mathbf{A}^{(k)})^{\dagger})^T$
– Mettre à jour $\mathbf{B}^{(k)} = \hat{\mathbf{B}}$
– Estimer $\mathbf{C}^{(k)}$: $\hat{\mathbf{C}} = \mathbf{T}^{K,JI}((\mathbf{B}^{(k)} \odot \mathbf{A}^{(k)})^{\dagger})^T$
– Mettre à jour $\mathbf{C}^{(k)} = \hat{\mathbf{C}}$

TABLE 4.2 – Les étapes de l'algorithme ALS

La convergence à l'itération k est déclarée lorsque l'erreur entre le vrai tenseur et sa version reconstruite, à partir des matrices de facteurs estimées, ne change pas de manière significative entre les itérations k et $k+1$. L'erreur de reconstruction à la $k^{\text{ème}}$ itération peut être calculée à partir de la formule suivante :

$$\Upsilon(k) = \|\mathbf{T}_{(1)}^{I,KJ} - (\mathbf{C}^{(k)} \odot \mathbf{B}^{(k)})(\mathbf{A}^{(k)})^T\|_F \tag{4.6}$$

On peut dire que l'algorithme converge à l'itération k lorsque $\|\Upsilon(k+1) - \Upsilon(k)\| < \epsilon$, tel que ϵ et un seuil dont sa valeur est fixée d'avance. La mise à jour conditionnelle d'une matrice donnée peut être soit améliorée ou maintenue, mais ne peut pas aggraver la forme actuelle. L'algorithme converge toujours de manière monotone, au moins, vers un minimum local. Cependant, l'algorithme ALS est fortement dépendant de l'initialisation, et sa convergence vers le minimum global peut parfois être lente. De plus, la convergence de l'algorithme peut, dans certains cas, tomber dans les régions de marais (*swamps*), au cours desquelles la vitesse de convergence est très faible et l'erreur entre deux itérations consécutives ne diminue pas. Dans ce cas, pour éviter un arrêt anticipé de l'algorithme, une pratique courante est d'imposer une valeur minimale acceptable de l'erreur $\Upsilon(k)$, au-dessus de laquelle la convergence globale n'est pas supposée encore atteinte.

Plusieurs variantes de l'algorithme ALS ont été proposées dans la littérature. Afin d'atténuer les problèmes de convergence lente causés par une initialisation aléatoire de l'algorithme, une solution d'analyse propre peut être utilisée (Sanchez et Kowalski, 1990; Sidiropoulos *et al.*, 2000b; Leurgans *et al.*, 1993). Cette solution est également connue comme la décomposition trilinéaire directe (Sanchez et Kowalski, 1990). Elle consiste à obtenir une première estimation des matrices de facteurs de la décomposition, par la construction d'un problème de valeurs propres généralisées (ou un problème de diagonalisation conjointe) à partir de deux tranches de tenseur. L'initialisation à base de l'analyse propre, en plus de se limiter à des tenseurs avec seulement deux tranches dans l'un des modes, exige que les deux matrices de facteurs sont de rang colonne plein, et que la troisième ne contient pas des éléments zéro. Dans (De Lathauwer, 2006), une généralisation de la solution d'analyse propre pour des tenseurs avec plus de deux tranches a été proposée, en liant l'estimation des matrices de facteurs de la décomposition CP au problème de diagonalisation matricielle simultanée.

Un autre moyen d'améliorer la vitesse de l'algorithme ALS est basé sur la méthode de compression Tucker-3 (Andersson et Bro, 1998; Bro et Andersson, 1998). Cette méthode est utile lorsque les dimensions du tenseur sont grandes. Dans (Sidiropoulos *et al.*, 2000b), un algorithme qui accélére la convergence de ALS est proposée. Cet algorithme applique la méthode de compression Tucker-3 suivie d'une initialisation basée sur l'analyse propre.

4.3.2 Ajout de recherche linéaire optimisée ELS à l'algorithme ALS

La convergence de l'algorithme ALS peut également être améliorée par le biais de ce qu'on appele la méthode de la ligne de recherche améliorée (ELS) (Rajih et Comon, 2005b). Cette méthode a montré son utilité lorsque la décomposition du tenseur est affectée par des facteurs dégénérescences (*Factor degeneracies*). La méthode ELS a également été utilisée pour estimer les facteurs des décompositions tensorielles en blocs (Rajih *et al.*, 2008; Rajih et Comon, 2005a).

L'idée de la méthode ELS consiste à rechercher le facteur de relaxation optimal R_{LS}, qui conduit à la solution finale d'un cycle donné en une seule étape. Par l'itération (k), on définit $\mathbf{L}_{\mathbf{A}}^{(k)} = \mathbf{A}^{(k-1)} - \mathbf{A}^{(k-2)}$ comme étant la direction d'estimation de la matrice $\mathbf{A}$. De même, nous définissons les deux directions $\mathbf{L}_{\mathbf{B}}^{(k)}$ et $\mathbf{L}_{\mathbf{C}}^{(k)}$. Au lieu de fixer une seule valeur de R_{LS} pour les trois modes comme utilisé dans la méthode LS (*Line search*), nous pouvons

chercher le triplet optimal $(R_{\mathbf{A}}, R_{\mathbf{B}}, R_{\mathbf{C}})$ qui minimise :

$$\Upsilon_{ELS} = \|\mathbf{T}_{(1)}^{I,KJ} - (\mathbf{A}^{(k-2)} + R_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}^{(k)})((\mathbf{C}^{(k-2)} + R_{\mathbf{C}}\mathbf{L}_{\mathbf{C}}^{(k)}) \odot (\mathbf{B}^{(k-2)} + R_{\mathbf{B}}\mathbf{L}_{\mathbf{B}}^{(k)}))^T\|_F^2 \quad (4.7)$$

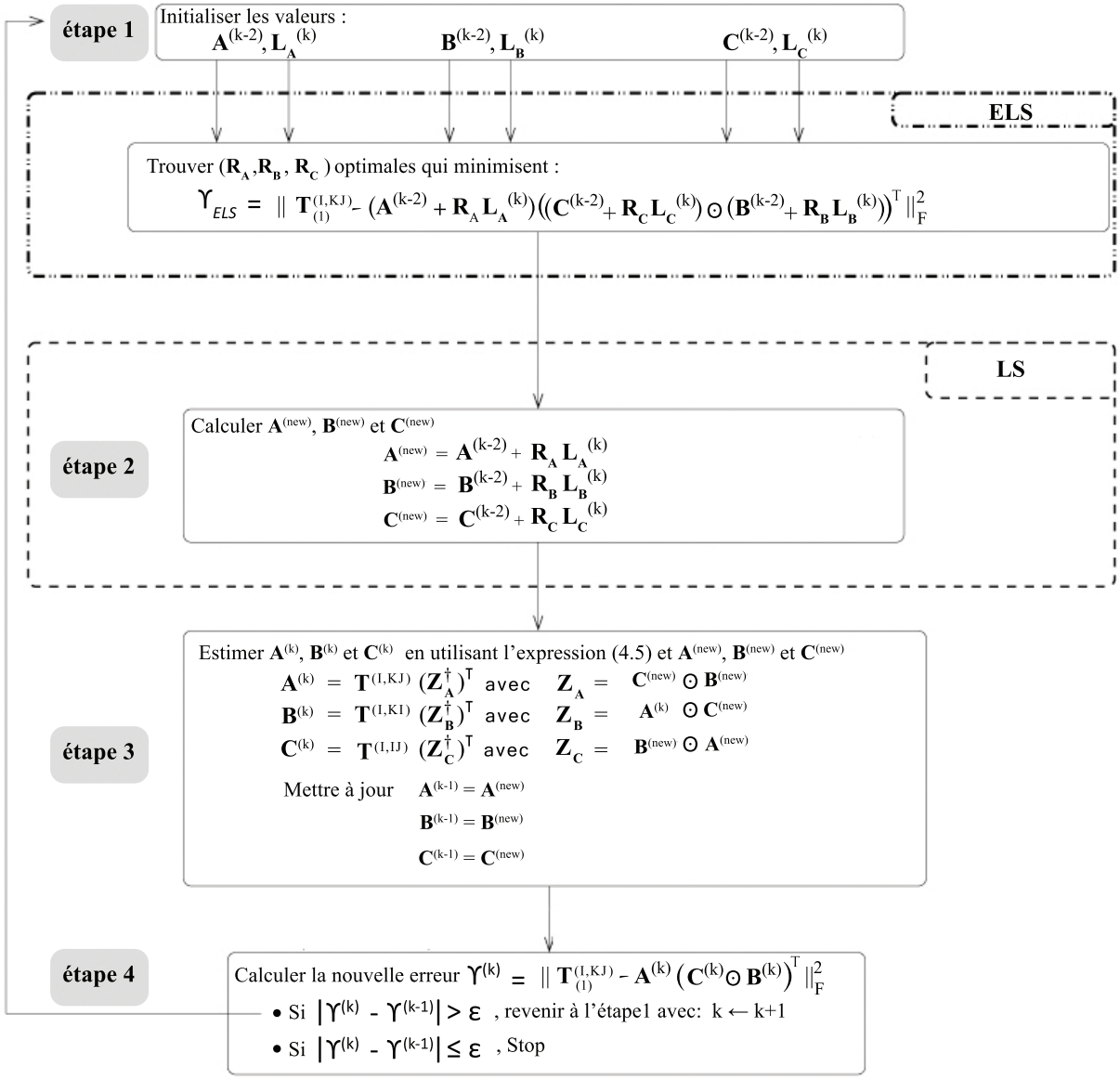

FIGURE 4.5 – Les étapes de l'algorithme ALS avec ELS.

La méthode ELS est exécutée au début de l'algorithme ALS comme representée dans la Figure 4.5, où l'étape 1 correspond à la partie ELS. Les facteurs de relaxation appliqués aux matrices de facteurs sont calculés à l'étape 1 de la Figure 4.5, en tant que valeurs optimales qui fournissent la plus petite erreur Υ_{ELS}. À l'étape 3, et après estimation des matrices de facteurs de l'itération (k), nous mettons à jour celles de l'itération $(k-1)$ pour avoir les nouvelles matrices $\mathbf{A}^{(new)}$, $\mathbf{B}^{(new)}$ et $\mathbf{C}^{(new)}$. Les matrices de facteurs des deux itérations $(k-1)$ et (k) sont ensuite utilisées dans l'itération suivante si l'algorithme n'atteint pas encore la convergence.

4.3.3 Simulations

Dans cette section, nous évaluerons les performances de l'algorithme ALS sous différents scénarios. On notera par ALS+ELS l'algorithme ALS avec recherche de ligne améliorée décrit dans la section 4.3.1 et représenté dans la Figure 4.5. Chaque résultat obtenu est une moyenne de 100 réalisations indépendantes de Monte-Carlo. La tolérance est fixée à $\epsilon_{ALS} = 10^{-10}$. Les éléments des matrices de facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ suivent une loi gaussienne de moyenne nulle et de variance un.

Dans le premier scénario, nous considérons un tenseur d'ordre 3 et de taille $I = 6$, $J = 5$, $K = 4$ et de rang $R = 5$. Nous évaluons l'erreur de la reconstruction du tenseur $\mathcal{T}$ par rapport au nombre d'itérations demandés en utilisant la formule suivante : $\frac{\|\mathcal{T}-\hat{\mathcal{T}}\|_F^2}{\|\mathcal{T}\|_F^2}$, où $\mathcal{T}$ est le tenseur original et $\hat{\mathcal{T}}$ est le tenseur reconstruit à partir des matrices de facteurs estimées. La Figure 4.6 représente la convergence de l'erreur de reconstruction du tenseur par rapport au nombre d'itérations pour l'algorithme ALS avec/sans la méthode ELS. Nous remarquons sur la Figure 4.6 que la méthode ELS réduit le nombre d'itérations nécessaires pour atteindre le critère de convergence. Ainsi, le nombre d'itérations diminue de 1000 itérations à environ 150. De plus, même si le ALS atteint une valeur d'erreur de 10^{-10}, cela nécessite plus d'itérations et donc plus de temps. En revanche, le ALS + ELS converge vers des valeurs d'erreur inférieures à 10^{-30} pour 500 itérations.

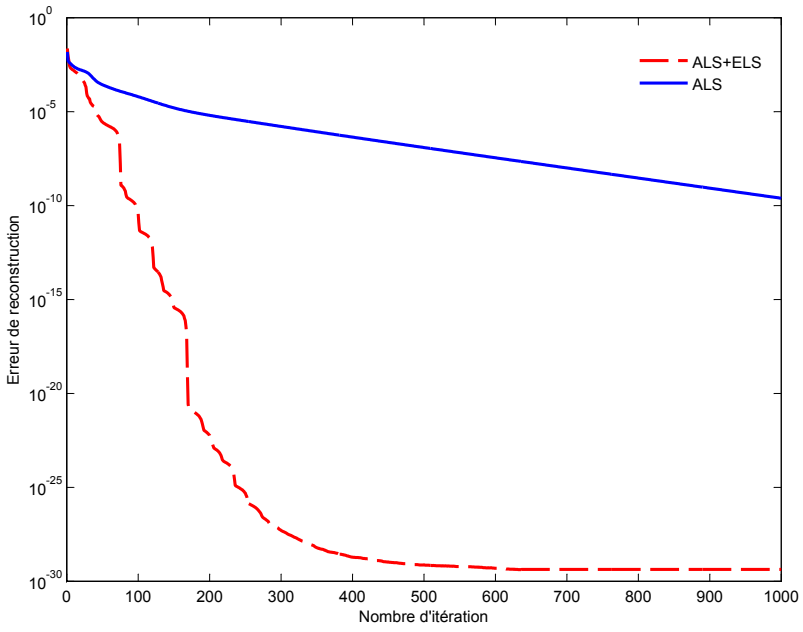

FIGURE 4.6 – Performance de l'algorithme ALS et ALS+ELS en fonction du nombre d'itérations, pour un tenseur de rang 5 et de dimensions $6 \times 5 \times 4$.

Dans la Figure 4.7, nous avons tracé l'erreur d'estimation d'une matrice de facteurs (EEM), telle que l'erreur estimée est égale au nombre d'éléments de $\mathbf{A} - \hat{\mathbf{A}}$ différents de 0, divisé par le nombre total des éléments de $\mathbf{A}$ (c'est-à-dire $I \times R$), où $\hat{\mathbf{A}}$ désigne l'estimée de la matrice $\mathbf{A}$ après décision. La Figure 4.7 montre que l'ELS est très utile pour réduire le nombre d'itérations nécessaires afin d'estimer la matrice de facteurs $\mathbf{A}$. En utilisant l'ELS, le nombre d'itérations diminue de 1000 à 500 pour atteindre une erreur d'estimation de 10^{-10}. Pour le deuxième scénario, nous considérons un tenseur d'ordre

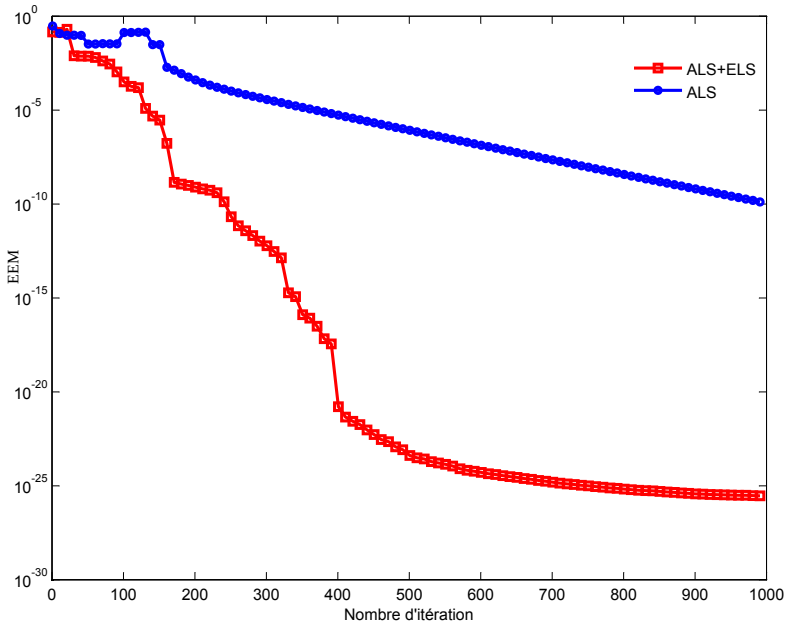

FIGURE 4.7 – Erreur d'estimation de la matrice de facteurs en utilisant les algorithmes ALS et ALS+ELS.

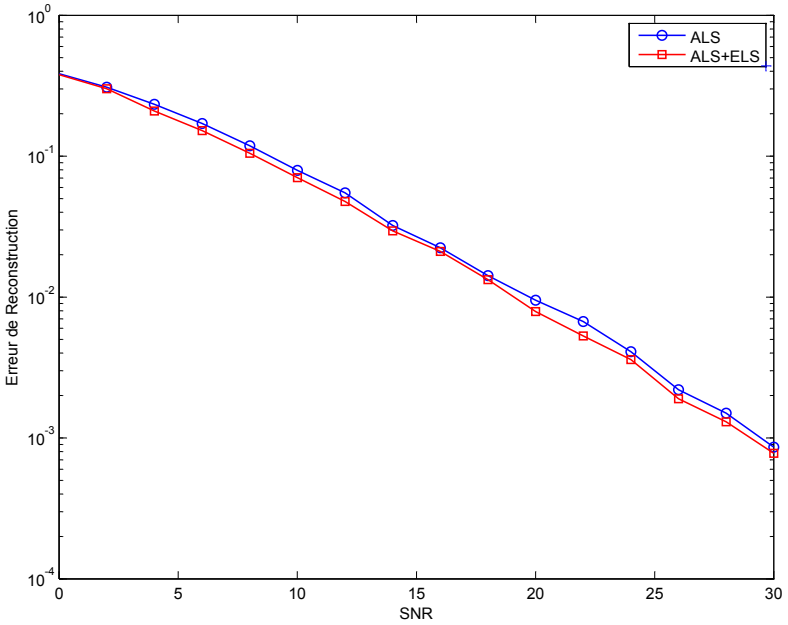

FIGURE 4.8 – Performance de l'algorithme ALS et ALS+ELS en fonction de SNR.

3, de taille $I = J = K = 10$ et de rang $R = 4$. Dans ce scénario, nous comparons la convergence de l'algorithme ALS avec/sans ELS par rapport à la valeur du rapport signal sur bruit SNR(*Signal Noise Ratio*). Sur la Figure 4.8, nous avons tracé la convergence de l'erreur de reconstruction du tenseur $\mathcal{T}$ par rapport au SNR. À partir de cette figure, nous remarquons bien que l'allure des deux courbes est presque similaire surtout pour des faibles valeurs de SNR. Une petite différence entre les deux courbes (ALS, ALS+ELS), est observée lorsque les valeurs de SNR deviennent de plus en plus grandes.

4.4 Algorithmes de descente

Partant d'un point $x^{(0)}$ arbitrairement choisi, un algorithme de descente cherche à générer une suite d'itérés $(x^{(k)})_{k\in\mathbb{N}}$ telle que :

$$\forall k \in \mathbb{N}, \quad f^{(k+1)} \leq f^{(k)}. \tag{4.8}$$

Dans le cas de notre fonction objectif (4.3), l'équation (4.9) deviendra comme suit :

$$\forall k \in \mathbb{N}, \quad \Upsilon(\mathbf{A}^{(k+1)}, \mathbf{B}^{(k+1)}, \mathbf{C}^{(k+1)}) \leq \Upsilon(\mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{C}^{(k)}). \tag{4.9}$$

Commençons par définir plus précisément la notion de descente. Le gradient joue un rôle essentiel en optimisation. Dans le cadre des méthodes d'optimisation, il sera également important d'analyser le comportement de la fonction objectif dans certaines directions.

Définition 3. *Soient $f \in \mathbb{R}^n \to \mathbb{R}$ et $x \in \mathbb{R}^n$. Le vecteur $d \in \mathbb{R}^n$ est une direction de descente pour f à partir du point x si $t \mapsto f(x^{(k)} + \rho d^{(k)})$ est décroissante en $\rho = 0$, c'est-à-dire s'il existe $\eta > 0$ tel que :*

$$\forall \rho \in]0, \eta], \quad f(x^{(k)} + \rho d^{(k)}) < f(x^{(k)}) \tag{4.10}$$

Proposition 4. *Soient $f : \mathbb{R}^n \to \mathbb{R}$ différentiable et $x^{(k)} \in \mathbb{R}^n$ tel que : $\nabla f(x^{(k)}) \neq 0$. Le vecteur $d^{(k)} \in \mathbb{R}^n$ est une direction de descente pour f à partir du point $x^{(k)}$ si et seulement si la dérivée directionnelle de f en $x^{(k)}$ dans la direction d vérifie :*

$$df(x^{(k)}; d^{(k)}) = \nabla f(x^{(k)})^T d^{(k)} < 0. \tag{4.11}$$

De plus pour tout $\beta < 1$, il existe $\bar{\eta} > 0$ tel que :

$$\forall \rho \in]0, \bar{\eta}], \quad f(x^{(k)} + \rho d^{(k)}) < f(x^{(k)}) + \rho\beta \nabla f(x^{(k)})^T d^{(k)}. \tag{4.12}$$

Autrement dit, la relation (4.12) affirme que la décroissance de la fonction objectif f en faisant un pas de taille ρ dans la direction $d^{(k)}$, est égale au moins au pas multiplié par une fraction β de la pente.

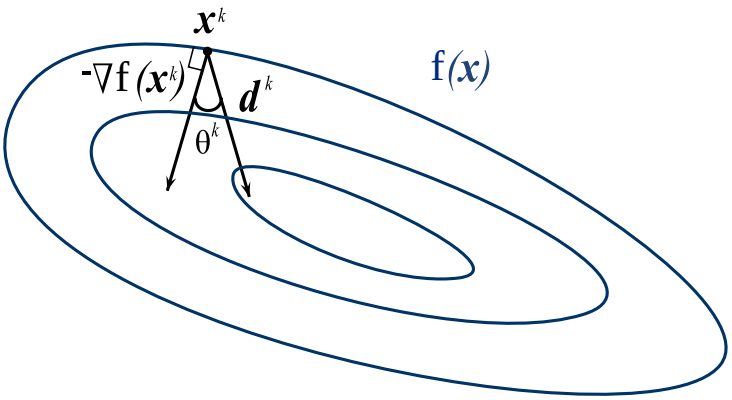

FIGURE 4.9 – Représentation de l'angle $\theta^{(k)}$ formé par $d^{(k)}$ et $-\nabla f(x^{(k)})$

Parmi toutes les directions de descente existantes en un point $x^{(k)}$ donné, une des plus remarquables est celle où la pente est la plus forte, c'est-à-dire que $d^{(k)}$ fait avec $-\nabla f(x^{(k)})$ un angle $\theta^{(k)}$ strictement plus petit que 90 degrés, comme l'illustre la Figure 4.9. Pour le démontrer, il suffit de comparer les dérivées directionnelles :

Théorème 5 (Direction de plus forte descente)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable. Soit $x^{(k)} \in \mathbb{R}^n$. Alors pour toute direction $d^{(k)}$ de norme constante égale à $\|d^{(k)}\| = \|\nabla f(x^{(k)})\|$, on a :*

$$(-\nabla f(x^{(k)}))^T \nabla f(x^{(k)}) \leq (d^{(k)})^T \nabla f(x^{(k)}). \tag{4.13}$$

La direction $(d^\star)^{(k)} = \nabla f(x^{(k)})$ est appelée direction de plus forte descente.

L'algorithme de descente utilise cette propriété pour minimiser la fonction f. La suite $\{x^{(k)}\}_{k\geq 0}$ est construite par la récurrence :

$$x^{(k+1)} = x^{(k)} + \rho^{(k)} d^{(k)} \tag{4.14}$$

où $\rho^{(k)} > 0$ est le pas et $d^{(k)}$ est une direction de descente (Figure 4.10). Une telle méthode consiste à alterner la construction de $d^{(k)}$ et la détermination de $\rho^{(k)}$. L'arrêt de la récurrence est contrôlé par un test portant généralement sur les faibles valeurs du gradient $\nabla f(x^{(k)})$. C'est en effet ce que suggère la condition d'optimalité $\nabla f(x^\star) = 0$. Bien que l'annulation du gradient puisse intervenir en un maximum ou en un point selle de f, le choix de $(\rho^{(k)}, d^{(k)})$ va empêcher l'algorithme de converger vers un tel point.

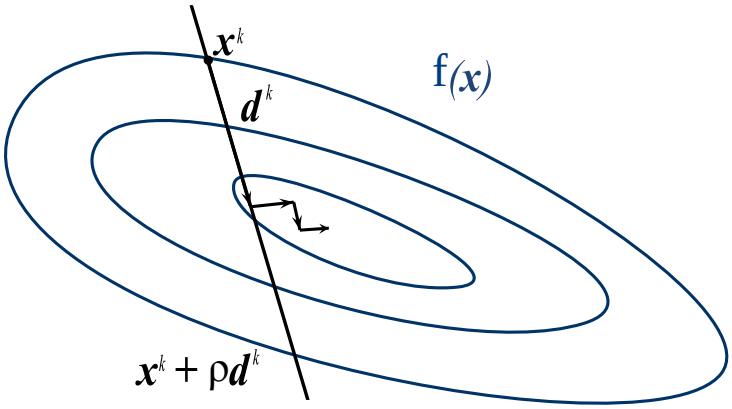

FIGURE 4.10 – La méthode de descente itérative et la minimisation approchée de f le long de la direction $d^{(k)}$ par une stratégie de recherche linéaire.

Revenons à notre fonction objectif sous sa forme matricielle (équation 3.23). Pour pouvoir utiliser d'autres types d'algorithmes d'optimisation que le ALS, la différentielle $d\Upsilon$ de la fonction Υ doit être calculée. De cette façon, les matrices $\nabla_{\mathbf{A}}\Upsilon$, $\nabla_{\mathbf{B}}\Upsilon$ et $\nabla_{\mathbf{C}}\Upsilon$ de taille $I \times R$,$J \times R$ et $K \times R$ respectivement, peuvent être évaluées. Dans les calculs suivants, nous considérons le cas où $\mathbf{\Lambda} = \mathbf{I}$ est une matrice identité de taille $R \times R$, comme été abondamment traité dans la littérature (Royer *et al.*, 2010), et $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$ et $\mathbf{C} \in \mathbb{R}^{K\times R}$. On a :

$$\begin{aligned}
\nabla_{\mathbf{A}}\Upsilon &= \frac{\partial}{\partial \mathbf{A}}(\|\mathbf{T}_1^{I,KJ} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T\|_F^2) \\
&= 2[-\mathbf{T}_1^{I,KJ} + \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T](\mathbf{C}\odot\mathbf{B}) \\
&= 2(-\mathbf{T}_1^{I,KJ}(\mathbf{C}\odot\mathbf{B}) + \mathbf{A}(\mathbf{C}^T\mathbf{C}) \boxdot (\mathbf{B}^T\mathbf{B})),
\end{aligned} \tag{4.15}$$

$$\begin{aligned}\nabla_{\mathbf{B}}\Upsilon &= \frac{\partial}{\partial \mathbf{B}}(\|\mathbf{T}_2^{J,KI} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T\|_F^2) \\ &= 2[-\mathbf{T}_2^{J,KI} + \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T](\mathbf{C}\odot\mathbf{A}) \\ &= 2(-\mathbf{T}_2^{J,KI}(\mathbf{C}\odot\mathbf{A}) + \mathbf{B}(\mathbf{C}^T\mathbf{C})\boxdot(\mathbf{A}^T\mathbf{A})), \end{aligned} \tag{4.16}$$

$$\begin{aligned}\nabla_{\mathbf{C}}\Upsilon &= \frac{\partial}{\partial \mathbf{C}}(\|\mathbf{T}_3^{K,JI} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T\|_F^2) \\ &= 2[-\mathbf{T}_3^{K,JI} + \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T](\mathbf{B}\odot\mathbf{A}) \\ &= 2(-\mathbf{T}_3^{K,JI}(\mathbf{B}\odot\mathbf{A}) + \mathbf{C}(\mathbf{B}^T\mathbf{B})\boxdot(\mathbf{A}^T\mathbf{A})), \end{aligned} \tag{4.17}$$

Dans l'optique d'estimer les trois matrices de facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, nous suggérons d'optimiser les fonctions de coût considérées (3.23) simultanément vis à vis les trois matrices de facteurs et non pas de façon alternée, comme dans l'ALS. Nous allons donc rappeler le principe des différents algorithmes d'optimisation à base de descente du gradient.

Dans la suite de ce chapitre, et pour avoir une simple écriture des équations, nous metterons les différentes matrices de facteurs dans une seule matrice $\mathbf{X}$, et les différents gradients partiels de Υ dans une matrice $\mathbf{G}$ comme montré dans les équations (4.18)(4.19) :

$$\mathbf{X}^{(k)} = \begin{pmatrix} \mathbf{A}^{(k)} \\ \mathbf{B}^{(k)} \\ \mathbf{C}^{(k)} \end{pmatrix} \tag{4.18}$$

$$\mathbf{G}^{(k)} = \begin{pmatrix} \nabla_{\mathbf{A}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)}) \\ \nabla_{\mathbf{B}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)}) \\ \nabla_{\mathbf{C}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)}) \end{pmatrix} \tag{4.19}$$

ou encore dans des vecteur $\mathbf{x}$ et $\mathbf{g}$ de taille $(I+J+K)R \times 1$:

$$\mathbf{x}^{(k)} = \begin{pmatrix} \text{vec}\{\mathbf{A}^{(k)}\} \\ \text{vec}\{\mathbf{B}^{(k)}\} \\ \text{vec}\{\mathbf{C}^{(k)}\} \end{pmatrix} \tag{4.20}$$

$$\mathbf{g}^{(k)} = \begin{pmatrix} \text{vec}\{\nabla_{\mathbf{A}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)})\} \\ \text{vec}\{\nabla_{\mathbf{B}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)})\} \\ \text{vec}\{\nabla_{\mathbf{C}}\Upsilon(\mathbf{A}^{(k)},\mathbf{B}^{(k)},\mathbf{C}^{(k)})\} \end{pmatrix} \tag{4.21}$$

Dans l'approche classique du gradient, la variable $\mathbf{X}$ est mise à jour à chaque itération suivant la règle d'adaptation suivante :

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \rho^{(k)}\mathbf{G}^{(k)} \tag{4.22}$$

On peut également utiliser cette règle sous la forme vectorielle suivante :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)} \tag{4.23}$$

4.4.1 Algorithme du gradient

La méthode du gradient fait partie des classes de méthodes dites de descente. Considérons un point de départ $\mathbf{x}^{(0)}$, et cherchons à minimiser la fonction de coût Υ. Puisqu'on veut atteindre $\mathbf{x}^{\star}$, nous cherchons à avoir : $\Upsilon(\mathbf{x}^{(1)}) < \Upsilon(\mathbf{x}^{(0)})$. Une forme particulièrement simple est de chercher $\mathbf{x}^{(1)}$ tel que le vecteur $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ soit colinéaire à une direction de descente $d^{(0)} \neq 0$. Nous le noterons : $\mathbf{x}^{(1)} - \mathbf{x}^{(0)} = \rho^{(0)} d^{(1)}$, où $\rho^{(0)}$ est le pas de descente de la méthode et $d^{(0)}$ est la direction de descente. La direction et le pas de descente peuvent être fixes ou changer à chaque itération. On retrouve alors la règle d'adaptation du gradient simple, à savoir, pour le cas matriciel :

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \rho^{(k)} \mathbf{D}^{(k)} \tag{4.24}$$

ou encore pour le cas vectoriel :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho^{(k)} \mathbf{d}^{(k)} \tag{4.25}$$

donnée en (4.23), avec $\mathbf{d}^{(k)} = -\mathbf{g}^{(k)}$ et $\mathbf{D}^{(k)} = -\mathbf{G}^{(k)}$.

Lorsque l'on travaille sur une résolution numérique d'un problème, on se donne en général un critère d'arrêt de la forme :

$$\|\Upsilon(\mathbf{x}^{(k+1)}) - \Upsilon(\mathbf{x}^{(k)})\| < \epsilon \tag{4.26}$$

De plus, puisque la convergence n'est pas toujours assurée, une règle de base est de fixer un nombre maximum d'itérations k_{max}. On obtient alors l'algorithme présenté dans la table 4.3 et dit du gradient.

Algorithme de Gradient
1. Mettre $k = 0$, Initialiser $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)})$.
2. Faire :
Calculer la direction de descente comme gradient par rapport à $\mathbf{X}$:
$\mathbf{D}^{(k)} = -\mathbf{G}^{(k)} = -\nabla\Upsilon(\mathbf{X}^{(k)})$
Calculer le pas $\rho^{(k)}$
Mettre à jour $\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)} + \rho^{(k)}\mathbf{D}^{(k)}$
Mettre à jour k : $k = k + 1$
3. Tant que $k \geq 1$ et jusqu'à vérification du critère d'arrêt : $\|\Upsilon(\mathbf{x}^{(k)}) - \Upsilon(\mathbf{x}^{(k-1)})\| < \epsilon$

TABLE 4.3 – Les étapes de l'algorithme du Gradient

Même si ces méthodes sont conceptuellement très simples et qu'elles peuvent être programmées directement, elles sont souvent lentes dans la pratique. Elles convergent mais sous des conditions de convergence souvent complexes. Par conséquent, on utilise plutôt la direction de gradient conjugué (CG).

4.4.2 Algorithme du Gradient Conjugué (CG)

À l'origine, l'algorithme du CG était conçu pour résoudre l'équation linéaire $\mathbf{A}x = y$, où la matrice $\mathbf{A}$ est symétrique et définie positive. L'extension de l'algorithme au cas

non-linéaire (NLCG) a été proposée par Fletcher et Reeves (Fletcher et Reeves, 1964) qui définissent :

$$\mathbf{d}^{(k)} = \begin{cases} -\mathbf{g}^{(k)}, & si \quad k = 0\,; \\ -\mathbf{g}^{(k)} + \beta^{(k)}\mathbf{d}^{(k-1)}, & si \quad k \geq 1. \end{cases} \tag{4.27}$$

avec le coefficient de conjugaison :

$$(\beta^{FR})^{(k)} = \frac{(\mathbf{g}^T)^{(k)}\mathbf{g}^{(k)}}{(\mathbf{g}^T)^{(k-1)}\mathbf{g}^{(k-1)}}. \tag{4.28}$$

Lorsque la fonction à minimiser est quadratique, le coefficient $\beta^{(k)}$ assure la conjugaison entre $\mathbf{d}^{(k)}$ et $\mathbf{d}^{(k-1)}$, et la méthode doit converger en au plus n itérations, où n est la dimension du vecteur d'inconnus. Dans le cas non-quadratique, il existe de nombreuses variantes de la méthode de Fletcher et Reeves, qui différent par le choix du paramètre $\beta^{(k)}$ (Hager et Zhang, 2006). De plus, lorsque $\beta^{(k)}$ n'assure pas que $\mathbf{d}^{(k)}$ soit une direction de descente, cette dernière est remplacée par son opposé $-\mathbf{d}^{(k)}$. Pour un problème non-linéaire quelconque, il est conseillé dans (Nocedal et Wright, 1999) de choisir la formule de (Gilbert et Nocedal, 1992) :

$$(\beta^{PRP+})^{(k)} = \max\{(\beta^{PRP})^{(k)}, 0\}. \tag{4.29}$$

où $(\beta^{PRP})^{(k)}$ est le coefficient de Polak-Ribière (Polak, 1997) :

$$(\beta^{PRP})^{(k)} = \frac{(\mathbf{g}^T)^{(k)}(\mathbf{g}^{(k)} - \mathbf{g}^{(k-1)})}{\|\mathbf{g}^{(k-1)}\|} \tag{4.30}$$

En pratique, la direction du gradient conjugué non-linéaire est souvent bien plus efficace que la direction de plus grande pente, en terme de vitesse de convergence et en simplicité de calcul (Nocedal et Wright, 1999).

4.4.3 Algorithme Quasi Newton

Les méthodes Quasi-Newton consistent à imiter la méthode de Newton, où l'optimisation d'une fonction est obtenue à partir des minimisations successives de son approximation au second ordre. En effet, l'inconvénient de la direction de Newton est qu'elle nécessite la connaissance du Hessien de la fonction objectif. Le calcul de cette matrice peut être compliqué et coûteux. De plus, pour certains problèmes, le critère n'est pas deux fois différentiable. Ceci a motivé l'apparition des méthodes de Quasi-Newton qui définissent la direction par :

$$\mathbf{d}^{(k)} = -(\mathbf{B}^{-1})^{(k)}\mathbf{g}^{(k)} \tag{4.31}$$

où $\mathbf{B}^{(k)}$ est une approximation du Hessien générée itérativement par une formule de mise à jour (Dennis et Moré, 1977).

La première méthode de quasi-Newton a été suggérée par Davidon (Davidon, 1991) en 1959, puis améliorée par Fletcher et Powell (Fletcher et Powell, 1963) en 1963, connue sous la dénomination de formule DFP. La formule de quasi-Newton la plus robuste est celle de BFGS, indépendamment suggérée par Broyden (BROYDEN, 1970), Fletcher (Fletcher,

1970), Goldfarb (Goldfarb, 1970) et Shanno (Shanno, 1970) en 1970. La formule de mise à jour s'écrit comme suit :

$$\mathbf{B}^{(k+1)} = \mathbf{B}^{(k)} - \frac{\mathbf{B}^{(k)}\mathbf{z}^{(k)}(\mathbf{z}^T)^{(k)}\mathbf{B}^{(k)}}{(\mathbf{z}^T)^{(k)}\mathbf{B}^{(k)}\mathbf{z}^{(k)}} + \frac{\delta^{(k)}(\delta^T)^{(k)}}{(\delta^T)^{(k)}\mathbf{z}^{(k)}}. \tag{4.32}$$

avec :

$$\mathbf{z}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \qquad \delta^{(k)} = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}.$$

La direction de Quasi-Newton nécessite de calculer la matrice $\mathbf{M}^{(k)} = (\mathbf{B}^{(-1)})^{(k)}$. Cette matrice s'obtient simplement par récurrence :

$$\mathbf{M}^{(k+1)} = \mathbf{M}^{(k)} + \varrho[1 + \varrho(\delta^T)^{(k)}\mathbf{M}^{(k)}\delta^{(k)}]\mathbf{z}^{(k)}(\mathbf{z}^T)^{(k)} - \varrho\mathbf{z}^{(k)}(\delta^T)^{(k)}\mathbf{M}^{(k)} - \varrho\mathbf{M}^{(k)}\delta^{(k)}(\mathbf{z}^T)^{(k)} \tag{4.33}$$

La règle d'adaptation de l'algorithme BFGS sera alors :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho^{(k)}\mathbf{M}^{(k)}\mathbf{g}^{(k)} \tag{4.34}$$

4.4.4 Algorithme de Levenberg-Marquardt

Cette méthode est connue depuis déjà de nombreuses années, proposée pour la première fois par Levenberg en 1944 (Levenberg, 1944). Il a fallu ensuite attendre 20 ans pour voir l'amélioration de Marquardt (Marquardt, 1963). Comme pour la méthode de Gauss-Newton (voir la sous-section précédente), la méthode de Levenberg-Marquardt est une méthode itérative fondée sur un schéma newtonien qui prend en compte l'approximation du hessien.

Dans certaines applications, par exemple lorsque l'approximation du Hessien n'est pas définie positivement, la direction $\mathbf{d}^{(k)}$ trouvée par l'algorithme de Gauss-Newton ne produit pas une réduction suffisante de la fonction objectif. Dans ce cas, la méthode de Levenberg-Marquardt permet de forcer la convergence en rapprochant la direction de descente $\mathbf{d}^{(k)}$ de la direction $-\mathbf{g}^{(k)}$. Cette méthode consiste à approcher le Hessien de la fonction objectif en ajoutant un multiple de la matrice identité à la matrice $\mathbf{B}$ avant son inversion. L'équation (4.34) deviendra alors :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho^{(k)}(\mathbf{B}^{(k)} + \alpha\mathbf{I})^{-1}\mathbf{g}^{(k)} \tag{4.35}$$

avec α est le coefficient de relaxation et $\mathbf{I}$ la matrice identité de taille $(I + J + K)R$. Si on pose $\alpha = 0$ dans l'équation (4.35), on retombe sur l'algorithme BFGS (4.34). De même, en posant $\mathbf{M} = \mathbf{I}$ ou encore, en considérant α suffisamment grand par rapport aux valeurs de $\mathbf{M}$, l'algorithme du Gradient est retrouvé (équation 4.24).

4.5 Méthodes de choix de pas de recherche

Dans les algorithmes de descente itératifs, le schéma d'optimisation se formule comme la détermination, à chaque itération, d'un pas minimisant le critère dans un sous-espace généré par une ou plusieurs directions de recherche. Dans cette section, nous étudierons les techniques employées pour déterminer le pas, appelée recherche ou stratégie de pas. Avoir une bonne stratégie de pas est fondamental car cela influe de façon importante sur

les propriétés de convergence de l'algorithme. Il est bien sûr possible d'utiliser un pas fixe tout au long des itérations, mais ce choix risque d'être synonyme d'une faible vitesse de convergence si le pas est choisi trop petit. De plus, si on choisit un pas trop grand, l'algorithme d'optimisation risque de diverger et on se trouve dans l'incapacité d'estimer d'une manière correcte les matrices de facteurs. Lorsque le sous-espace se limite à une unique direction, la recherche du pas consiste en la minimisation approchée d'une fonction scalaire. Nous présentons dans cette section, les stratégies usuelles pour le calcul du pas scalaire, également appelées stratégies de recherche linéaire, en se basant principalement sur (Bertsekas, 1999; Boyd et Vandenberghe, 2004).

4.5.1 Pas exactes

La détermination du pas exact d'une fonction f correspond à la règle de Cauchy. Il permet, comme son nom l'indique, de déterminer le pas de recherche qui minimse au maximum la fonction f à l'étape k, en fonction de la direction de descente calculée. Lorsque la fonction f est multimodale, il peut être plus simple de rechercher le minimum dans un intervalle restreint (règle de Cauchy limitée) ou bien de déterminer le plus petit point stationnaire de f (règle de Curry) (Bertsekas, 1999). Ces règles, qualifiées d'exactes, ne sont utilisées que lorsque le minimum s'exprime de façon analytique.

1 Règle de Cauchy :

$$\rho^{(k)} = argmin_{\rho \in \mathbb{R}^+}\{f(\mathbf{x}^{(k)} + \rho \mathbf{d}^{(k)})\} \tag{4.36}$$

2 Règle de Cauchy limitée :

$$\rho^{(k)} = argmin_{\rho \in [0,s]}\{f(\mathbf{x}^{(k)} + \rho \mathbf{d}^{(k)})\} \tag{4.37}$$

3 Règle de Curry :

$$\rho^{(k)} = \min\{\rho > 0 | \nabla f(\mathbf{x}^{(k)} + \rho \mathbf{d}^{(k)})^T \mathbf{d}^{(k)} = 0\} \tag{4.38}$$

Cependant, la résolution exacte peut demander un temps de calcul trop grand, et ne peut de toute façon pas être faite avec une précision infinie. De plus, l'efficacité supplémentaire, éventuellement apportée à l'algorithme par un pas exact permet rarement de compenser le temps consacré à déterminer un tel pas. Par conséquent, les stratégies usuelles pour le calcul du pas sont basées sur la vérification de conditions moins restrictives, qui permettent toutefois de garantir la convergence des algorithmes.

4.5.2 Pas approchés

Le calcul du pas exacte n'implique pas forcément des performances optimales de l'algorithme. De plus, elle n'est pas nécessaire pour assurer la convergence, dès lors que des conditions sont assurées par le pas utilisé. De plus, le calcul du pas exacte (ou pas optimal) est une solution coûteuse en terme du temps de calcul, qui consiste à calculer les racines d'un polynôme de degré élevé dans le cas de notre fonction objectif (4.3). Dans des problèmes de grandes dimensions (des matrices de facteurs de grandes tailles), à chaque itération, le calcul du pas exact représente la plus grande partie de la complexité algorithmique totale. Une condition naturelle est de demander au pas d'entraîner une

décroissance suffisante de la fonction f. Cela se traduit le plus souvent par une inégalité de la forme :

$$f(\mathbf{x}^{(k)} + \rho^{(k)}\mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + \mathcal{C} \tag{4.39}$$

avec $\mathcal{C}$ un terme inférieur strictement à zero. Cette forme est appelée condition de descente suffisante. Parmi les différentes méthodes de recherche du pas inexacte ou approché, on retrouve une qui est très simple et très efficace, appelée recherche du pas par marche arrière (*Backtracking*). Cette méthode permet à moindre coût d'obtenir une bonne approximation du pas $\rho^{(k)}$. Elle dépend de deux constantes α et β, avec $0 < \alpha < 0,5$ et $0 < \beta < 1$.

Cette recherche de pas est appelée marche arrière, car il commence avec un pas ρ suffisamment grand au début, par exemple un pas unité, puis cette valeur est réduite par le facteur β tel que $\rho = \rho\beta$, jusqu'à ce que la condition suivante sera verifiée :

$$f(x + \rho\mathbf{d}) < f(x) + \alpha\rho\nabla f(x)^T\mathbf{d}, \tag{4.40}$$

Cette condition est appelée la condition d'Armijo (Boyd et Vandenberghe, 2004; Luenberger et Ye, 2008). La variable d est la direction de descente et $\nabla f(x)$ est le gradient vu dans les sections précédentes. Dans le cas de l'algorithme du gradient $\mathbf{d} = -\mathbf{g}$, alors que $\mathbf{d} = -\mathbf{Mg}$ pour les algorithmes newtoniens (BFGS par exemple). La constante α peut être interprétée comme la fraction de la diminution de la fonction f prédite par extrapolation linéaire acceptée. Le paramètre α est typiquement choisi entre $0,01$ et $0,3$, ce qui signifie que l'on accepte une diminution de f comprise entre 1% et 30% de la prédiction par extrapolation linéaire. Le paramètre β est souvent choisi entre $0,1$ et $0,8$ (Boyd et Vandenberghe, 2004).

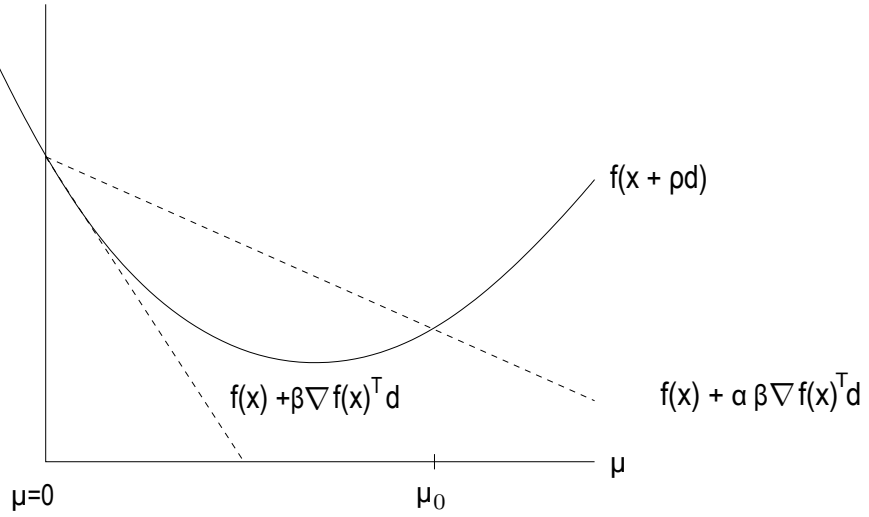

FIGURE 4.11 – Le principe de la méthode *Backtracking* (Boyd et Vandenberghe, 2004)

La Figure 4.11 illustre le principe de la méthode marche arrière (*Backtracking*). La courbe en pointillée inférieure représente l'extrapolation linéaire de f, et la courbe en pointillée supérieure correspond à la droite de pente α plus faible. On remarque bien que la condition d'Armijo est verifiée pour des valeurs de β comprises entre 0 et μ_0, où la fonction f est en dessous de la droite pointillée supérieure.

En reprenant nos notations, et en considérant notre fonction de coût $\Upsilon(.)$ (4.3), la condition d'Armijo s'écrit :

$$\Upsilon(\mathbf{x} + \rho\mathbf{d}) < \Upsilon(\mathbf{x}) + \alpha\rho\mathbf{g}^T\mathbf{d} \tag{4.41}$$

Un résumé de l'algorithme marche arrière est donné dans la table 4.4.

Algorithme de marche arrière
1. Étant donné une direction de descente $\mathbf{d}$ pour Υ, $\alpha \in [0, 0.5]$, $\beta \in [0, 1]$.
2. Initialiser : $\rho = 1$.
3. Faire :
$\rho = \beta\rho$.
Tant que : $\Upsilon(\mathbf{x} + \rho\mathbf{d}) < \Upsilon(\mathbf{x}) + \alpha\rho\mathbf{g}^T\mathbf{d}$.

TABLE 4.4 – Résumé de l'algorithme backtracking

4.6 Comparaison des différents algorithmes

Dans cette section, nous allons présenter plusieurs simulations pour les différents algorithmes d'optimisation de la décomposition CP d'un tenseur d'ordre 3. Nous allons illustrer le comportement des différents algorithmes présentés auparavant, afin de mettre en évidence l'impact de plusieurs facteurs sur leurs performances.

Pour la première simulation, nous considérons un tenseur $\mathcal{T}_1$ non bruité d'ordre 3, de taille $15 \times 14 \times 16$ et de rang 5. Donc, nous somme menés à estimer $(14 + 15 + 16) \times 5$ variables. Le tenseur estimé sera noté $\hat{\mathcal{T}}_1$, avec $\hat{\mathcal{T}}_1 = \sum_{r=1}^{R} \hat{\mathbf{a}}_r \circ \hat{\mathbf{b}}_r \circ \hat{\mathbf{c}}_r$ et $\hat{\mathbf{a}}_r$, $\hat{\mathbf{b}}_r$ et $\hat{\mathbf{c}}_r$ les estimés de la $r^{\text{ème}}$ colonne des différentes matrices de facteurs pour $r = 1, \ldots, R$.

Pour avoir la possibilité de comparer entre les différents algorithmes, nous avons besoin d'un indice de performance. Pour ce premier scénario, nous choisissons d'utiliser un indice d'erreur, défini comme :

$$E = ||\mathcal{T}_1 - \hat{\mathcal{T}}_1||_F^2 \tag{4.42}$$

Lorsque l'indice de performance E est proche de 0, les résultats obtenus sont dites meilleurs. Donc nous devons donner un seuil à partir duquel on considère que le minimum global de notre fonction est atteint, par exemple $\epsilon = 10^{-10}$. De ce fait, le critère d'arrêt pour les algorithmes itératifs sera $E < \epsilon$. Dans le cas des données non bruitées et si la décomposition est unique, la norme des résidus est nulle lorsque le minimum global est atteint, ce qui interprète bien le critère (4.42). Toutefois, lorsqu'un palier est rencontré, l'indice de performance E stagne de telle sorte à ce que le critère (4.42) ne peut pas être satisfait dans un délai de temps. Pour cette raison, le critère (4.42) sera couplé à un autre critère qui permet d'éviter les cas extrêmes, qui est un nombre d'itérations maximum fixe.

Dans le cas des données bruitées, la norme des résidus est non nulle lorsque le minimum global de notre fonction est atteint. De ce fait, le critère (4.42) ne peut pas être utilisé, mais nous utiliserons plutôt le critère suivant :

$$|E^{(k)} - E^{(k-1)}| < \epsilon \tag{4.43}$$

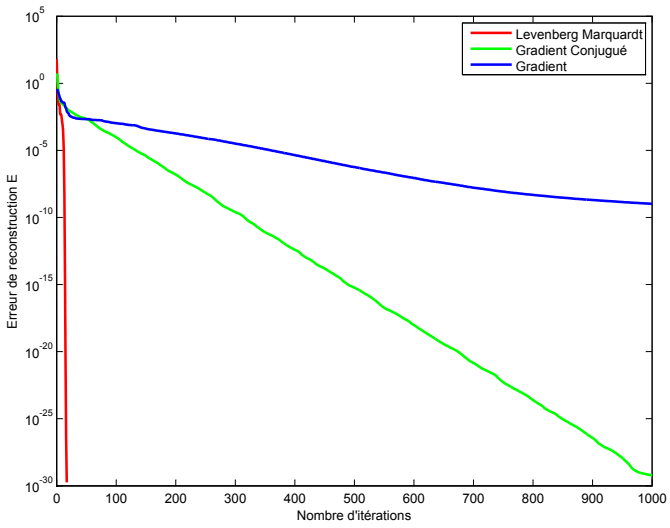

FIGURE 4.12 – Erreur de reconstruction par rapport au nombre d'itération demandé pour le tenseur $\mathcal{T}_1$ d'ordre 3 et de taille $I = 15$, $J = 14$ et $K = 16$

Ce critère (4.43) sera utilisé en plus des deux autres mentionnés précédemment (Le nombre d'itérations maximum et le critère (4.42)).

Après avoir générer les trois matrices de facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ à partir des distributions gaussiennes, nous avons testé chaque algorithme d'optimisation étudié dans les sections précédentes. La Figure 4.12 illustre une comparaison entre les différents algorithmes d'optimisation (Gradient, Gradient Conjugé et Levenberg-Marquardt). À partir des résultats obtenus dans cette figure, on remarque bien que l'algorithme Levenberg-Marquardt converge vers la valeur de 10^{-10} pour un nombre d'itérations petit, 15 itérations, alors que c'est l'algorithme du Gradient qui demande plus d'itérations pour converger. Toutefois, cela n'est pas suffisant pour conclure que l'algorithme Levenberg-Marquardt est le meilleur.

Pour faire une bonne comparaison entre les algorithmes d'optimisation étudiés, nous avons évalué la complexité algorithmique de ces derniers, ainsi que le temps de calcul. Le tableau 4.5 représente le temps pris par chaque algorithme pour converger vers le minimum de la fonction de coût (4.3). D'après le tableau 4.5, on constate que l'algorithme de Gradient nécessite moins de temps pour converger vers le minimum en le comparant avec les deux autres algorithmes. Par contre, l'algorithme de Levenberg-Marquardt, qui demande un nombre petit d'itérations (4.12), prend plus de temps pour converger vers le minimum.

Algorithme	Temps de Calcul (s)
Gradient	40.62
Gradient Conjugué (CG)	48.95
Levemberg Marquardt (LM)	80.61

TABLE 4.5 – Le temps d'exécution pour les trois algorithmes : Gradient, Gradient Conjugué, Levemberg Marquardt

Le troisième indice de performance évalué dans cette section est la complexité algo-

rithmique, qui représente un bon moyen de comparaison du temps de calcul des différents algorithmes indépendamment, du processeur de calcul utilisé.
Commençant d'abord par l'algorithme ALS. Pour une seule itération, et pour estimer la matrice de facteurs $\mathbf{A}$, on a un produit de Khatri Rao de complexité $O(JKR)$, le calcul du pseudo inverse de complexité $O(7JKR^2 + \frac{11}{3}R^3)$, ainsi que l'expression de l'estimation de la matrice $\mathbf{A}$ de complexité $O(IJKR+IR^2+IR)$. Donc, la complexité de l'algorithme ALS pour une seule itération est de : $O(3IJKR + 11R^3 + (R + R^2)(I + J + K) + (R + 7R^2)(IJ+IK+JK))$. Concernant l'algorithme Gradient, il y a trois composantes de gradient à calculer. Dans chaque composante, on a un produit de Khatri Rao, et 4 produits matriciels. Pour la première composante (4.15), la complexité du produit de Khatri Rao est $O(JKR)$ et celle des 4 produits matriciels (Le produit entre le produit de Khatri Rao et la matrice dépliante du tenseur, le produit entre $\mathbf{C}^T$ et $\mathbf{C}$, le produit entre $\mathbf{B}^T$ et $\mathbf{B}$ et le produit entre la matrice $\mathbf{A}$ et le produit de Hadamard) est $O(IJKR+IR^2+JR^2+KR^2)$. De même pour les deux autres composantes du gradient en respectant les matrices de facteurs. La complexité algorithmique du Gradient, pour une seule itération, est de : $O(3IJKR+3R^2(I+J+K)+R(IJ+IK+JK))$. Pour l'algorithme du Gradient Conjugué GC, on trouve le calcul des composantes de Gradient ainsi que le coefficient β (4.28) pour trouver la direction de descente. La complexité de cet algorithme sera alors la somme des complexités de l'algorithme Gradient et celle du calcul du coefficient β. Dans le calcul de ce dernier, on a deux produits matriciels entre les matrices $\mathbf{G}$, de taille $(I + J + K)R$, qui contiennent les composantes du gradient. Par conséquence, la complexité de l'algorithme Gradient Conjugué CG est de : $O(3IJKR+5R^2(I+J+K)+R(IJ+IK+JK))$. En ce qui concerne l'algorithme Quasi Newton BFGS (4.34), en plus du calcul des composantes du gradient, nous aurons besoin de calculer l'approximé de la matrice Hessienne. Cette dernière demande 5 produits matriciels et une inversion matricielle. De ce fait, la complexité algorithmique par itération est de l'ordre de : $O((I + J + K)^3R^3 + 4(I + J + K)^2R^2 + 3IJKR+3R^2(I+J+K)+R(IJ+IK+JK))$. Pour l'algorithme de Levenberg Marquardt, il a la même complexité algorithmique que BFGS, vu que la seule différence de calcul est l'ajout d'un multiple de la matrice identité à la matrice $\mathbf{B}$ avant son inversion. Le tableau 4.6 représente un résumé de complexité algorithmique des différents algorithmes étudiés dans cette section. À partir de l'ensemble des résultats résumés dans le tableau 4.6, on peut bien constaté que les algorithmes Gradient et Gradient Conjugé ont une faible complexité algorithmique, alors que les algorithmes Levenberg-Marquardt et BFGS ont des complexités plus grande que les deux premiers.

Algorithme	Complexité par itération
ALS	$3IJKR + 11R^3 + (R + R^2)(I + J + K) + (R + 7R^2)(IJ + IK + JK)$
Gradient	$3IJKR + 3R^2(I + J + K) + R(IJ + IK + JK)$
CG	$3IJKR + 5R^2(I + J + K) + R(IJ + IK + JK)$
BFGS	$(I + J + K)^3R^3 + 4(I + J + K)^2R^2 + 3IJKR + 3R^2(I + J + K)$ $+R(IJ + IK + JK)$
LM	$(I + J + K)^3R^3 + 4(I + J + K)^2R^2 + 3IJKR + 3R^2(I + J + K)$ $+R(IJ + IK + JK)$

TABLE 4.6 – La complexité algorithmique des différentes méthodes

4.7 Algorithmes proposés

Notre problème d'optimisation consiste à minimiser l'erreur quadratique (3.22) sous une contrainte d'égalité. C'est une contrainte qui est toujours active dans le problème d'optimisation. Elle se définie comme un ensemble de vecteur unitaire qui appartient à un espace de Hilbert $\mathbb{H}$.

Le problème d'optimisation que nous allons traité est le suivant :

$$\begin{aligned} \min \quad & \|\mathcal{T} - \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F^2 \\ S.C \quad & ||\mathbf{a}_r|| = ||\mathbf{b}_r|| = ||\mathbf{c}_r|| = 1 \end{aligned} \tag{4.44}$$

La difficulté qui se pose dans cette optimisation est de savoir comment se déplacer dans l'ensemble des contraintes. Donc, les algorithmes d'optimisation vus aux sections 4.3 et 4.4 ne sont plus valables pour le problème (4.44). Une élaboration des méthodes d'optimisation permettant la prise en compte des contraintes sera faite dans cette section.

4.7.1 Algorithmes avec contrainte d'égalité

Généralement, un problème d'optimisation avec contrainte d'égalité est de cette forme :

$$\begin{cases} \min_x & f(x), \\ S.C & h(x) = 0 \end{cases} \tag{4.45}$$

La première question est celle de l'existence du minimum global de la fonction f sur $\mathcal{C}$. Il existe principalement deux théorèmes qui permettent de répondre à cette question. Le premier affirme l'existence d'un point de minimum lorsque l'ensemble des contraintes est fermé borné. Le second est son équivalent pour les ensembles de contraintes fermés mais non bornés. Étant donnée une direction de recherche, comment garantir que l'on reste dans l'ensemble des contraintes $\mathcal{C}$. Pour cela, on introduit la notion de direction admissible.

Définition 6 (**Direction admissible**). *Soit $x \in \mathbb{R}^n$ un point admissible du problème (4.44). Une direction $d \in \mathbb{R}^n$ sera dite admissible en x s'il existe $\eta > 0$ tel que $x + \rho d$ soit admissible quel que soit $\rho \in]0; \eta]$.*

Rappelons que, dans le cas sans contrainte, les algorithmes de descente s'écrivent sous la forme générique :

$$\begin{cases} x^{(0)} \text{donné}, \\ x^{(k+1)} = x^{(k)} + \rho^{(k)} d^{(k)} \end{cases} \tag{4.46}$$

où $\rho^{(k)}$ et $d^{(k)}$ sont choisis de telle sorte que $f(x^{(x+1)}) \leq f(x^{(x)})$. Lorsqu'on minimise sous un ensemble de contraintes $\mathcal{C}$, il n'est pas sûr que $x^{(k)}$ reste sur $\mathcal{C}$. Il est donc nécessaire de se ramener sur $\mathcal{C}$. On réalise cette dernière opération grâce à une projection sur $\mathcal{C}$. Ceci donne lieu alors naturellement aux algorithmes de descente projetés, parmi lesquels nous citons :

– Gradient projeté
– Gradient Conjugué projeté

Construisons maintenant un algorithme de descente projeté qui résout le problème (4.44). Cet algorithme doit trouver trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ de colonnes de norme 1 qui minimisent la fonction (4.44). En d'autre terme, trouver les trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ sur une sphère de rayon 1.

La méthode du gradient projeté s'inspire des méthodes de gradient décrites dans la section 4.4. L'idée de base consiste à suivre la direction de plus profonde descente, comme dans le cas sans contrainte :

$$\begin{aligned}\mathbf{A}^{(k+1)} &= \mathbf{A}^{(k)} - \rho^{(k)}\nabla\Upsilon_{\mathbf{A}} \\ \mathbf{B}^{(k+1)} &= \mathbf{B}^{(k)} - \rho^{(k)}\nabla\Upsilon_{\mathbf{B}} \\ \mathbf{C}^{(k+1)} &= \mathbf{C}^{(k)} - \rho^{(k)}\nabla\Upsilon_{\mathbf{C}}\end{aligned}$$

où $\rho^{(k)} > 0$ est choisi de sorte que : $\Upsilon(\lambda^{(k+1)}, \mathbf{A}^{(k+1)}, \mathbf{B}^{(k+1)}, \mathbf{C}^{(k+1)}) < \Upsilon(\lambda^{(k)}, \mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{C}^{(k)})$. Toutefois, si $(\lambda^{(k)}, \mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{C}^{(k)}) \in \mathcal{C}$, rien ne garantit que $(\lambda^{(k+1)}, \mathbf{A}^{(k+1)}, \mathbf{B}^{(k+1)}, \mathbf{C}^{(k+1)})$ appartienne également à $\mathcal{C}$. Dès que l'on obtient un point non admissible, on projette celui-ci sur l'ensemble de contrainte $\mathcal{C}$. Le principe de l'algorithme est le suivant : Soit $\mathbf{x}^{(k)}$ l'itéré courant. On génère à l'itération suivante le point admissible :

$$\mathbf{y}^{(k)} = p_{\mathcal{C}}(\mathbf{x}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)}) \tag{4.47}$$

où $p_{\mathcal{C}}$ désigne l'opérateur de projection sur l'ensemble $\mathcal{C}$, $\mathbf{x}^{(k)}$ est le vecteur qui contient les éléments des matrices facteurs et $\mathbf{g}^{(k)}$ est le vecteur qui contient les composantes du nouveau gradient calculées par la suite (4.48).

$$\begin{aligned}\frac{\partial\Upsilon(\Lambda,\mathbf{A},\mathbf{B},\mathbf{C})}{\partial a_{\alpha\beta}} &= -2\sum_{ijk}(T_{ijk} - \sum_{r=1}^{R}\lambda_r a_{ir}b_{jr}c_{kr})\lambda_\beta b_{j\beta}c_{k\beta} \\ \frac{\partial\Upsilon(\Lambda,\mathbf{A},\mathbf{B},\mathbf{C})}{\partial b_{\alpha\beta}} &= -2\sum_{ijk}(T_{ijk} - \sum_{r=1}^{R}\lambda_r a_{ir}b_{jr}c_{kr})\lambda_\beta a_{i\beta}c_{k\beta} \\ \frac{\partial\Upsilon(\Lambda,\mathbf{A},\mathbf{B},\mathbf{C})}{\partial c_{\alpha\beta}} &= -2\sum_{ijk}(T_{ijk} - \sum_{r=1}^{R}\lambda_r a_{ir}b_{jr}c_{kr})\lambda_\beta a_{i\beta}b_{j\beta}\end{aligned} \tag{4.48}$$

La forme matricielle des composantes du gradient $\nabla_{\mathbf{A}}\Upsilon$, $\nabla_{\mathbf{B}}\Upsilon$ et $\nabla_{\mathbf{C}}\Upsilon$ de taille $I \times R$, $J \times R$ et $K \times R$ respectivement sont :

$$\begin{aligned}\nabla_{\mathbf{A}}\Upsilon &= \frac{\partial}{\partial\mathbf{A}}(\|\mathbf{T}_1^{I,KJ} - \mathbf{A}\Lambda(\mathbf{C}\odot\mathbf{B})^T\|_F^2) \\ &= 2[-\mathbf{T}_1^{I,KJ} + \mathbf{A}\Lambda(\mathbf{C}\odot\mathbf{B})^T](\mathbf{C}\odot\mathbf{B})\Lambda \\ &= 2(-\mathbf{T}_1^{I,KJ}(\mathbf{C}\odot\mathbf{B})\Lambda + \mathbf{A}\Lambda(\mathbf{C}^T\mathbf{C})\boxdot(\mathbf{B}^T\mathbf{B})\Lambda),\end{aligned} \tag{4.49}$$

$$\begin{aligned}\nabla_{\mathbf{B}}\Upsilon &= \frac{\partial}{\partial\mathbf{B}}(\|\mathbf{T}_2^{J,KI} - \mathbf{B}\Lambda(\mathbf{C}\odot\mathbf{A})^T\|_F^2) \\ &= 2[-\mathbf{T}_2^{J,KI} + \mathbf{B}\Lambda(\mathbf{C}\odot\mathbf{A})^T](\mathbf{C}\odot\mathbf{A})\Lambda \\ &= 2(-\mathbf{T}_2^{J,KI}(\mathbf{C}\odot\mathbf{A})\Lambda + \mathbf{B}\Lambda(\mathbf{C}^T\mathbf{C})\boxdot(\mathbf{A}^T\mathbf{A})\Lambda),\end{aligned} \tag{4.50}$$

$$\begin{aligned}\nabla_{\mathbf{C}}\Upsilon &= \frac{\partial}{\partial \mathbf{C}}(\|\mathbf{T}_3^{K,JI} - \mathbf{C}\Lambda(\mathbf{B}\odot\mathbf{A})^T\|_F^2)\\ &= 2[-\mathbf{T}_3^{K,JI} + \mathbf{C}\Lambda(\mathbf{B}\odot\mathbf{A})^T](\mathbf{B}\odot\mathbf{A})\Lambda\\ &= 2(-\mathbf{T}_3^{K,JI}(\mathbf{B}\odot\mathbf{A})\Lambda + \mathbf{C}\Lambda(\mathbf{B}^T\mathbf{B})\boxdot(\mathbf{A}^T\mathbf{A})\Lambda), \end{aligned} \tag{4.51}$$

Si la direction $\mathbf{d}^{(k)} = \mathbf{y}^{(k)} - \mathbf{x}^{(k)}$ est non nulle, alors c'est une direction de descente de Υ en $\mathbf{x}^{(k)}$. La direction $\mathbf{d}^{(k)}$ possède les propriétés suivantes :

1. Si $\mathbf{d}^{(k)} = 0$, alors : $p_{\mathcal{C}}(\mathbf{x}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)}) = \mathbf{x}^{(k)}$. Cela signifie que la direction choisie par l'algorithme de gradient est orthogonale à l'ensemble des contraintes $\mathcal{C}$ en $\mathbf{x}^{(k)}$.
2. Supposons $\mathbf{d}^{(k)} \neq 0$. Alors $\mathbf{x}^{(k)}$ et $p_{\mathcal{C}}(\mathbf{x}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)})$ sont des points admissibles du problème (4.44).

Dans les sous sections suivantes, nous allons proposer deux versions de l'algorithme d'optimisation Gradient projeté de la fonction 4.44 avec deux façons de calcul du facteur d'échelle Λ.

4.7.1.1 Algorithme1

Dans les derniers travaux sur la décomposition tensorielle CP, l'optimisation de la fonction objectif (4.44) était faite sans prendre en considération le facteur Λ. Alors qu'on trouve que le facteur d'échelle Λ est une matrice d'identité dans les travaux (Comon *et al.*, 2009; Rajih *et al.*, 2008), il est mis dans l'une des trois matrices dans (Royer *et al.*, 2010), ce qui rend ces dernières non-libres. La première solution proposée consiste à appliquer l'algorithme de Gradient projeté tout en calculant le Λ comme produit des normes des colonnes des trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ après leurs normalisations (4.52) (Rouijel *et al.*, 2014a; Comon *et al.*, 2013).

$$\Lambda^{(k+1)} = \Lambda^{(k)}\Lambda_{\mathbf{A}}\Lambda_{\mathbf{B}}\Lambda_{\mathbf{C}} \tag{4.52}$$

avec $\Lambda_{\mathbf{A}} = diag\{[\|\mathbf{a}_1\|, \cdots, \|\mathbf{a}_R\|]^T\}$. De même pour $\Lambda_{\mathbf{B}}$ et $\Lambda_{\mathbf{C}}$. Cette approche que nous nommerons "Algorithme1" a le schéma présenté dans la table 4.7.

4.7.1.2 Algorithme 2

Dans notre deuxième solution, nous avons appliqué l'algorithme du gradient projeté sur le problème (4.44), mais cette fois en calculant le Λ contracté où encore le Λ optimal.

Comme présenté dans la section 3.3 du chapitre précédent, le fait d'isoler le facteur d'échelle et de calculer sa valeur optimale, permet de réduire les indéterminations d'échelle à des colonnes de norme unitaire. Trouver la valeur optimale de Λ revient à résoudre le système des équations (3.24), de tel sorte à avoir :

$$\Lambda = \mathbf{G}^{-1}\mathbf{f} \tag{4.53}$$

Le principe de notre deuxième algorithme, que nous nommerons Algorithme 2, consiste à trouver à chaque itération un point admissible que nous projeterons après sur l'ensemble des contraintes. En d'autre termes, à chaque itération, et après avoir trouvé les trois matrices facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ qui minimisent la fonction de coût (4.44), nous calculons la valeur optimale du facteur d'échelle Λ, puis nous normalisons les trois matrices de facteurs (Rouijel *et al.*, 2014a; Comon *et al.*, 2013).

Algorithme 1

1. Initialisation :
($\mathbf{A}^{(0)}$,$\mathbf{B}^{(0)}$,$\mathbf{C}^{(0)}$ point initial arbitrairement choisi, $\varepsilon > 0$ précision demandée, $\Lambda^{(0)} = \mathbf{I}$.
2. $k := 1$
3. Normaliser les colonnes des matrices $\mathbf{A}(k-1)$, $\mathbf{B}(k-1)$ et $\mathbf{C}(k-1)$
4. Tant que critère d'arrêt non satisfait,
   - Direction de descente : $d^{(k)} = -\nabla\Upsilon(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$
   - Recherche d'un pas $\rho^{(k)}$ tel que :

$\Upsilon(\mathbf{A}^{(k-1)} + \rho^{(k)} d_{\mathbf{A}}^{(k)}, \mathbf{B}^{(k-1)} + \rho^{(k)} d_{\mathbf{B}}^{(k)}, \mathbf{C}^{(k-1)} + \rho^{(k)} d_{\mathbf{C}}^{(k)}, \Lambda(k-1)) < \Upsilon(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$.

   - Mise à jour :
     - $\mathbf{A}^{(k)} = \mathbf{A}^{(k-1)} + \rho^{(k)} d_{\mathbf{A}}^{(k)}$,
     - $\mathbf{B}^{(k)} = \mathbf{B}^{(k-1)} + \rho^{(k)} d_{\mathbf{B}}^{(k)}$,
     - $\mathbf{C}^{(k)} = \mathbf{C}^{(k-1)} + \rho^{(k)} d_{\mathbf{C}}^{(k)}$,
     - $k = k + 1$;
     - Normaliser les nouveaux $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$
     - mettre à jour le Λ :$\Lambda^{(k)} = \Lambda^{(k-1)} \Lambda_{\mathbf{A}} \Lambda_{\mathbf{B}} \Lambda_{\mathbf{C}}$.
5. Retourner $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$.

TABLE 4.7 – Résumé de l'algorithme 1 basé sur le gradient projeté

Le schéma de notre algorithme d'optimisation "Algorithme2" est donné dans la table 4.8.

Algorithme 2

1. Initialisation :
$\mathbf{A}^{(0)}$,$\mathbf{B}^{(0)}$,$\mathbf{C}^{(0)}$ point initial arbitrairement choisi, $\varepsilon > 0$ précision demandée.
2. $k := 1$
3. Normaliser les colonnes des matrices $\mathbf{A}(k-1)$, $\mathbf{B}(k-1)$ et $\mathbf{C}(k-1)$
4. Calculer le $\Lambda^{(0)}$: $\Lambda^{(0)} = (\mathbf{G}^{(0)})^{-1}\mathbf{f}^{(0)}$,
5. Tant que critère d'arrêt non satisfait,
   - Direction de descente : $d^{(k)} = -\nabla\Upsilon(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$
   - Recherche d'un pas $\rho^{(k)}$ tel que :

$\Upsilon(\mathbf{A}^{(k-1)} + \rho^{(k)} d_{\mathbf{A}}^{(k)}, \mathbf{B}^{(k-1)} + \rho^{(k)} d_{\mathbf{B}}^{(k)}, \mathbf{C} - \mathbf{1}^{(k)} + \rho^{(k)} d_{\mathbf{C}}^{(k)}, \Lambda(k-1)) < \Upsilon(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$.

   - Mise à jour :
     - $\mathbf{A}^{(k)} = \mathbf{A}^{(k-1)} + \rho^{(k)} d_{\mathbf{A}}^{(k)}$,
     - $\mathbf{B}^{(k)} = \mathbf{B}^{(k-1)} + \rho^{(k)} d_{\mathbf{B}}^{(k)}$,
     - $\mathbf{C}^{(k)} = \mathbf{C}^{(k-1)} + \rho^{(k)} d_{\mathbf{C}}^{(k)}$,
     - $k = k + 1$;
     - Normaliser les nouveaux matrices $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$
     - mettre à jour le Λ : $\Lambda^{(k)} = (\mathbf{G}^{(k)})^{-1}\mathbf{f}^{(k)}$.
5. Retourner $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$.

TABLE 4.8 – Résumé de l'algorithme 2 basé sur le gradient projeté

4.7.1.3 Simulations

Dans cette section, nous allons présenter les simulations réalisées, pour illustrer le comportement et le fonctionnement des trois algorithmes présentés antérieurement :

- Algorithme de Gradient Projeté avec $\Lambda = \mathbf{I}$, que nous l'appelons dans ce manuscrit Algorithme 0,
- Algorithme 1 avec Λ pris comme produit des normes des trois matrices de facteurs,
- Algorithme 2 avec Λ optimal.

Deux tenseurs différents sont considerés, $\mathcal{T}_1$ et $\mathcal{T}_2$. Le premier tenseur est de rang $R = 2$ et de taille $2 \times 3 \times 4$, alors que le deuxième est de rang $R = 4$ et taille $3 \times 3 \times 7$. Pour établir une comparaison entre les différents algorithmes, nous avons besoin d'un indice d'erreur. Nous avons choisi de tracer l'erreur quadratique absolue $E = \frac{\|\mathcal{T}-\hat{\mathcal{T}}\|_F^2}{\|\mathcal{T}\|_F^2}$ pour illustrer les performances des trois Algorithmes.

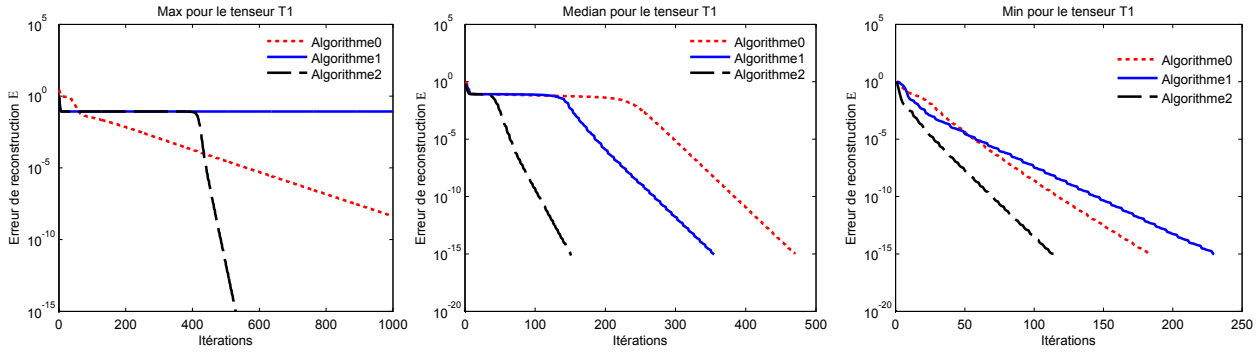

FIGURE 4.13 – Erreur de reconstruction du tenseur $\mathcal{T}_1$ de taille $2 \times 3 \times 4$ et de rang 2 en fonction du nombre d'itérations

Dans la première simulation, nous considérons le tenseur $\mathcal{T}_1$. La recherche du pas est faite par la méthode "Backtracking" et l'erreur à atteindre est de 10^{-15}. L'initialisation des trois matrices à estimer $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ est faite 21 fois. La Figure 4.13 présente la convergence des trois algorithmes par rapport au nombre d'itérations. Cela est fait pour la plus mauvaise, la moyenne et la meilleure initialisation, représentées par "Max", "Median" et "Min" respectivement. À partir de la Figure 4.13, on remarque bien que quelque soient les valeurs initiales des trois matrices, l'algorithme 2 reste toujours le premier qui converge avec un nombre d'itérations minimal.

Dans notre deuxième simulation, nous considérons le tenseur $\mathcal{T}_2$ et nous comparons la convergence des trois algorithmes par rapport à un seuil d'itérations précis. La Figure 4.14 présente la convergence de ces algorithmes par rapport au seuil qui est $Seuil_{Iter} = 200$ itérations. En comparant les trois courbes obtenues dans cette figure, on remarque bien que pour 200 itérations, l'algorithme 0 tend vers une erreur de reconstruction d'un peu près 10^{-4}, l'algorithme 1 tend vers 10^{-5}, alors que notre deuxième proposition, qui est l'algorithme 2 atteint une erreur de 10^{-6}. Tous ces résultats nous laisse conclure que l'algorithme 2 avec le Λ optimal (4.8) est plus performant que les deux autres.

Les simulations faites jusqu'à maintenant étaient sur des données non bruitées. Considérons maintenant un tenseur $\mathcal{T}_3$ de dimensions $I = 4$, $J = 7$ et $K = 10$. Les matrices de

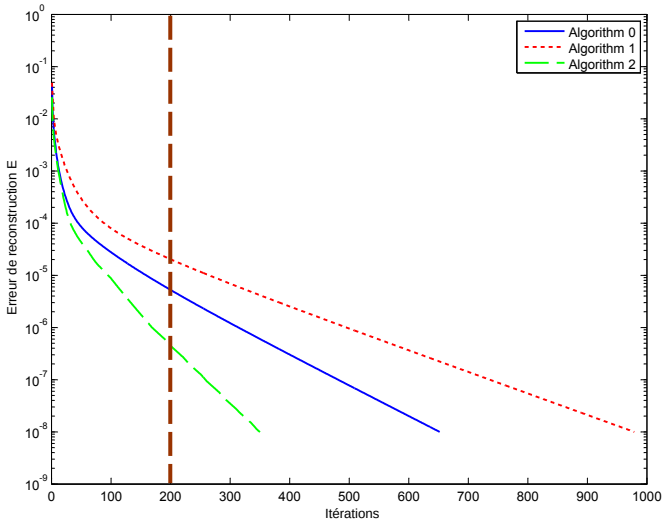

FIGURE 4.14 – Erreur de la reconstruction du tenseur $\mathcal{T}_2$ en fonction du nombre d'itérations.

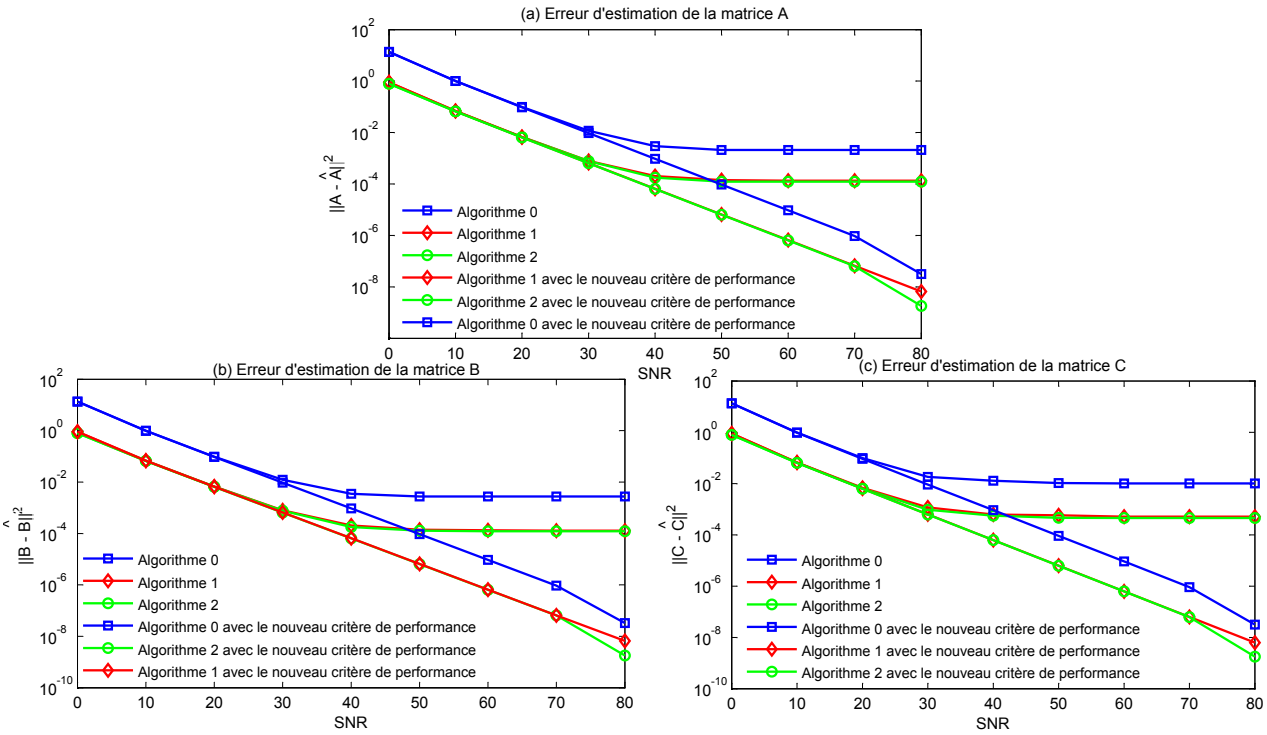

FIGURE 4.15 – Erreur d'estimation de la matrice, pour un tenseur de $4 \times 7 \times 10$ et de rang 4 en fonction du SNR

facteurs sont initialisées aléatoirement avec 4 colonnes. Les résultats sont obtenus après 100 itérations de Monte Carlo. À chaque itération, et pour chaque valeur du SNR, un nouveau bruit est ajouté à notre tenseur $\mathcal{T}_3$.

Dans les Figures 4.15 et 4.16, une comparaison entre les 3 algorithmes en terme de l'erreur d'estimation des trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ est faite, ainsi qu'en terme de critère de performance présenté dans le chapitre précédent. Nous avons tracé l'erreur d'estimation

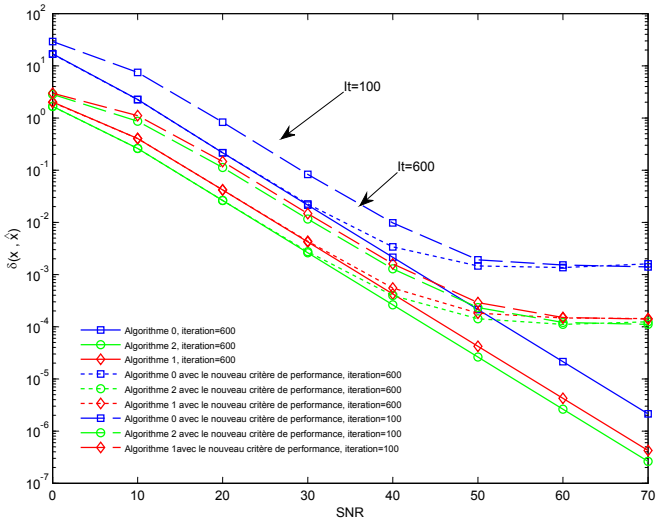

FIGURE 4.16 – La somme des erreurs d'estimation des matrices de facteurs (3.30) en fonction de SNR. Notez que l'asymptote dépend de nombre maximum d'itérations exécutés.

des matrices de facteurs en fonction de la valeur du SNR.

Premièrement, et à partir de la Figure 4.15, il est bien clair que la performance de l'algorithme 0 est très faible en comparaison avec l'algorithme 1 et 2 (courbes avec étoiles et ceux avec cercles). Cela confirme l'idée que nos algorithmes d'isolement de la matrice d'échelle sont attractifs. De plus, si on analyse le comportement des algorithmes selon le critère de performance utilisé (Figure 4.16), on remarque bien que lorsqu'on ne considère pas les contraintes de phase lors du calcul de performance, les résultats sont nettement plus optimistes que ceux avec les contraintes sur les phases, en particulier pour les grandes valeurs du SNR (voir les courbes de la même couleur et symbole). Cela consolide l'intérêt d'utiliser notre indice de performance défini dans l'équation (3.30).

4.7.2 Proposition d'un algorithme avec contrainte d'inégalité

L'isolation de la matrice d'échelle λ dans la décomposition CP a plusieurs implications, parmi lesquelles l'existance et l'unicité de la décomposition CP. Comme présenté dans le chapitre précédent, on a bien montré que le fait d'imposer la contrainte (4.54), permet de garantir l'unicité de la décomposition CP.

$$\mu_{\mathbf{A}}^{-1} + \mu_{\mathbf{B}}^{-1} + \mu_{\mathbf{C}}^{-1} \geq 2(R+1), \tag{4.54}$$

Malgré l'utilisation de cette contrainte, la minimisation de la fonction (4.3) souleve d'autres problèmes, comme l'absence de la différentiabilité de (4.54). C'est pourquoi plusieurs simplifications sont appliquées sur la contrainte (4.54) (voir l'annexe B.1) pour avoir la formule suivante :

$$\mathcal{C}_2 \overset{\text{def}}{=} \sum_{j<k} |\mathbf{a}_j^T \mathbf{a}_k|^2 + \sum_{\ell<m} |\mathbf{b}_\ell^T \mathbf{b}_m|^2 + \sum_{n<q} |\mathbf{c}_n^T \mathbf{c}_q|^2 - 3\left(\frac{3}{2R+2}\right)^2 \leq 0 \tag{4.55}$$

Notre objectif sera alors d'optimiser la décomposition tensorielle CP sous deux contraintes, une d'égalité et une autre d'inégalité. Ce qui veut dire, résoudre le problème d'optimisation suivant :

$$\begin{aligned} \min \Upsilon(\Lambda, \mathbf{A}, \mathbf{B}, \mathbf{C}) = \|\mathcal{T} &- \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \mathbf{b}_r \mathbf{c}_r)\|_F^2 \\ \mathcal{S}.\mathcal{C} \ \|\mathbf{a}_r\| &= \|\mathbf{b}_r\| = \|\mathbf{c}_r\| = 1 \\ \sum_{i=1}^{3} \mu_i &\leq 3\left(\frac{3}{2R+2}\right)^2 \end{aligned} \tag{4.56}$$

Les algorithmes qu'on a proposé jusqu'à maintenant traitent l'optimisation d'une décomposition tensorielle sous une contrainte d'égalité, donc ils ne peuvent pas être appliqués au le problème (4.56).

4.7.2.1 Algorithme de Barrière

Nous allons maintenant décrire les méthodes de pénalité qui permettent de résoudre le problème d'optimisation (4.56). Cette section traite des méthodes de pénalité intérieure, aussi appelées méthodes de barrière ou méthode de point intérieur. Le principe de ces méthodes réside dans la transformation d'un problème avec contrainte en une séquence de problèmes sans contraintes, en ajoutant au coût une pénalité en cas de violation de celles-ci. La solution d'un tel sous-problème est trouvée à chaque itération en utilisant une méthode de pénalité. L'appellation "pénalité intérieure" ou point intérieur est employée car le minimum est approché depuis l'intérieur de $\mathcal{C}$. Les méthodes de barrière s'appliquent aux problèmes dont l'ensemble admissible $\mathcal{C}$ est défini uniquement par une collection d'inégalités :

$$\begin{aligned} \min \quad & f(x) \\ \mathcal{S}.\mathcal{C} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, n \end{aligned} \tag{4.57}$$

En effet, ces méthodes utilisent des fonctions dites de barrière, définies uniquement à l'intérieur de $\mathcal{C}$. Si des contraintes sous forme d'égalités étaient introduites, l'intérieur de cet ensemble, c'est-à-dire son sous-ensemble tel qu'en chacun de ses points aucune contrainte n'est active, serait clairement vide (Boyd et Vandenberghe, 2004). C'est donc le domaine de définition de la fonction de barrière qui serait vide, rendant l'utilisation de la méthode impossible. L'intérieur de l'ensemble $\mathcal{C}$ défini par les g_i est le suivant :

$$\mathcal{C}_I = \{x \mid g_i(x) < 0, \forall i \in \{1, \ldots, n\}\}. \tag{4.58}$$

La fonction de barrière, notée $B(x)$, est ajoutée au coût $f(x)$ (Nesterov et Nemirovsky, 1994; Roos *et al.*, 1997) ; elle est continue sur $\mathcal{C}_I$ et sa valeur tend vers l'infini lorsque la frontière de $\mathcal{C}$ est approchée par l'intérieur, c'est-à-dire lorsque l'un des $g_i(x)$ approche zéro par les valeurs négatives. Une itération de la méthode consiste ensuite à minimiser la fonction $f(x)+\eta B(x)$ (où η est un paramètre réel strictement positif) à l'aide d'algorithmes de minimisation directe, comme par exemple ceux décrits au section 4.4. Une fonction $B(x)$ et un η convenablement choisis assurent que cette minimisation ne puisse nous mener à des points situés hors de $\mathcal{C}_I$. La suite du processus consiste à réduire progressivement η afin de diminuer la pénalité et d'autoriser les algorithmes de minimisation directe à se rapprocher peu à peu de la frontière de $\mathcal{C}$.

Le tableau (4.9) et la Figure (4.17) présentent quelques exemples de fonctions $B(x)$ qui sont des barrières associées au domaine $\mathcal{D}_B = [0, +\infty[$. Les barrières logarithmique (Fiacco et McCormick, 1968) et inverse sont des barrières strictes. En effet, elles tendent vers l'infini en $x = 0$. À l'inverse, les barrières entropique et hyperbolique sont définies en 0 et valent 0. Il est important de noter que, si tous les g_i sont convexes, ces deux

Nom	Barrière $B(x)$
Barrière logarithmique	$-\log(x)$
Barrière inverse	$\frac{1}{x}$
Barrière entropique	$x\log(x)$
Barrière hyperbolique	$-\sqrt{x}$

TABLE 4.9 – Exemples des fonctions barrières

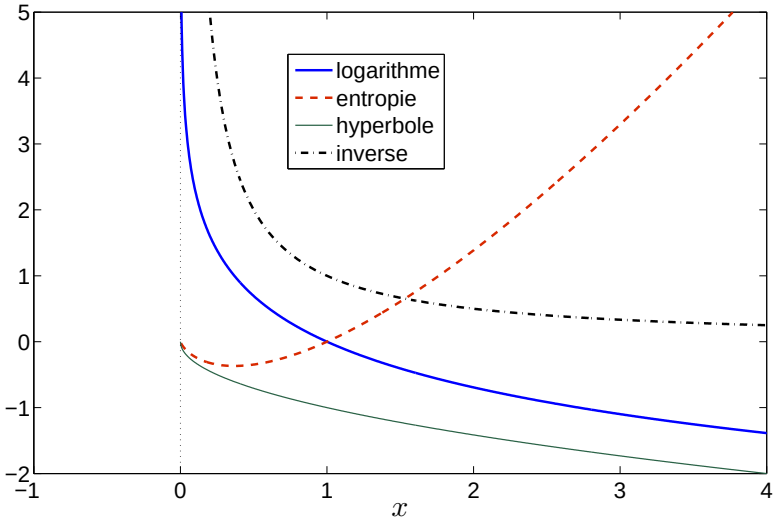

FIGURE 4.17 – Exemples de fonctions barrières.

fonctions de barrière (barrières entropique et hyperbolique) le sont également (Boyd et Vandenberghe, 2004).

Remarquons qu'une nécessité, pour pouvoir appliquer une telle méthode, est de disposer d'un point initial situé à l'intérieur de $\mathcal{C}$. La méthode de barrière est définie en introduisant la séquence de paramètres η^k ; $k = 0, 1, \ldots$ avec $0 < \eta^{k+1} < \eta^k$ et $\eta^k \to 0$ lorsque $k \to \infty$.

Le fait que $B(x)$ ne soit défini que dans l'intérieur $\mathcal{C}_I$ et tende vers l'infini au fur et à mesure que l'on se rapproche des bords de $\mathcal{C}$ assure que, même avec un algorithme de minimisation directe, le point obtenu à chaque itération appartient lui aussi à $\mathcal{C}_I$. Le fait que η^k tend vers zéro implique que le terme $f(x) + \eta^k B(x)$ tend vers $f(x)$ lorsque k tend vers l'infini (Boyd et Vandenberghe, 2004).

Le comportement de la méthode dépend fortement du choix du paramètre initial η^0 et du facteur, disons t, satisfaisant $0 < t < 2$, utilisé pour faire décroître η^k à chaque itération par la formule $\eta^{k+1} = t\eta^k$. Il n'existe pas de règle universelle permettant d'obtenir un bon choix de η^0 et t. Cela dépend fortement du problème à résoudre, ainsi que du point initial

x^0. L'utilisateur d'une méthode de barrière sera souvent condamné à exécuter la méthode plusieurs fois avec différentes valeurs de ces paramètres jusqu'à obtenir une convergence satisfaisante.

Les faits suivants peuvent néanmoins être relevés : si η^k est trop petit, et à plus forte raison si x^k est proche des bords du domaine, le terme de barrière peut se révéler trop faible, ne parvenant pas à empêcher une sortie de $\mathcal{C}$ (il faut donc prendre garde, durant l'exécution, et vérifier qu'on ne sort pas de cet ensemble, et, si cela est tout de même le cas, recommencer avec un η^0 ou un t plus grand). Dans le cas contraire, si η^k est trop grand, l'algorithme de minimisation directe ne pourra s'approcher suffisamment des bords du domaine, conduisant à une convergence globale lente. Ces paramètres doivent donc être soigneusement choisis selon le problème et le point initial. Revenons maintenant à notre problème d'optimisation (4.56). L'idée est de modifier la fonction objectif Υ de (4.56) à l'intérieur de l'ensemble des contraintes en ajoutant la fonction : $\log(\mathcal{C}_2)$, tel que $\mathcal{C}_2$ représente la contrainte sur la cohérence, qui a la forme suivante :

$$\mathcal{C}_2 = 3\left(\frac{3}{2R+2}\right)^2 - \sum_{j<k} |\mathbf{a}_j^H \mathbf{a}_k|^2 - \sum_{\ell<m} |\mathbf{b}_\ell^H \mathbf{b}_m|^2 - \sum_{n<q} |\mathbf{c}_n^H \mathbf{c}_q|^2 \geq 0 \tag{4.59}$$

La fonction du coût du problème (4.56) prendra alors la forme :

$$\Upsilon_b(\mathbf{A}, \mathbf{B}, \mathbf{C}, \Lambda) = \Upsilon - \eta \log(\mathcal{C}_2) \tag{4.60}$$

et donc, nous devons résoudre notre nouveau problème d'optimisation qui a la forme suivante :

$$\begin{aligned} \min \quad & \Upsilon_b(\mathbf{A}, \mathbf{B}, \mathbf{C}, \Lambda) = \Upsilon - \eta \log(\mathcal{C}_2) \\ \mathcal{S}.\mathcal{C} \quad & \|\mathbf{a}_r\| = \|\mathbf{b}_r\| = \|\mathbf{c}_r\| = 1 \end{aligned} \tag{4.61}$$

Dans la littérature, les problèmes d'optimisation avec des contraintes d'inégalité sont résolus en utilisant la méthode barrière avec l'algorithme de Newton pour la recherche du point optimal de chaque itération. Dans cette partie, nous proposons comme solution au problème (4.61) la méthode Barrière avec l'algorithme du Gradient Projeté au lieu de l'algorithme Newton, dans l'objectif de diminuer la complexité du calcul.

4.7.2.2 Algorithme proposé Barrière-GP

En vue de la fonction de coût de notre nouveau problème d'optimisation, il est nécessaire de calculer le gradient de la nouvelle fonction avec Barrière.

Cas réel :

On veut calculer le gradient de la fonction de coût (4.61) dans le cas réel, c'est à dire, dans le cas où les trois matrices de facteurs $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ appartiennent à $\mathbb{R}^{I\times R}$, $\mathbb{R}^{J\times R}$ et $\mathbb{R}^{K\times R}$ respectivement. Calculons le gradient de Υ_b par rapport aux éléments $a_{\alpha\beta}$, $b_{\alpha\beta}$ et

$c_{\alpha\beta}$ de $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ respectivement. On a :

$$\frac{\partial \Upsilon_b}{\partial a_{\alpha\beta}} = \frac{\partial \Upsilon}{\partial a_{\alpha\beta}} - \eta \frac{\partial \log(\mathcal{C}_2)}{\partial a_{\alpha\beta}} \tag{4.62}$$

$$\frac{\partial \Upsilon_b}{\partial b_{\alpha\beta}} = \frac{\partial \Upsilon}{\partial b_{\alpha\beta}} - \eta \frac{\partial \log(\mathcal{C}_2)}{\partial b_{\alpha\beta}} \tag{4.63}$$

$$\frac{\partial \Upsilon_b}{\partial c_{\alpha\beta}} = \frac{\partial \Upsilon}{\partial c_{\alpha\beta}} - \eta \frac{\partial \log(\mathcal{C}_2)}{\partial c_{\alpha\beta}} \tag{4.64}$$

avec $\mathcal{C}_2$ la contrainte d'inégalité définie par l'équation (4.55). Calculons d'abord les composantes du gradient de Υ, ensuite ceux du gradient de $\log(C_2)$. Dans la section 4.7.1, nous avons calculé les composantes du gradient de Υ par rapport aux éléments $a_{\alpha\beta}$, $b_{\alpha\beta}$ et $c_{\alpha\beta}$ (4.48), ainsi que par rapport aux matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ (4.49)(4.50)(4.51). Ce qui nous reste à calculer dans cette section sont les composantes du gradient de $\log(C_2)$.

$$\frac{\partial \log(\mathcal{C}_2)}{\partial \mathbf{a}_\beta} = \frac{2\sum_{\beta\neq q}(\mathbf{a}_q\mathbf{a}_q^T)\mathbf{a}_\beta}{\mathcal{C}_2} \tag{4.65}$$

$$\frac{\partial \log(\mathcal{C}_2)}{\partial \mathbf{b}_\beta} = \frac{2\sum_{\beta\neq m}(\mathbf{b}_m\mathbf{b}_m^T)\mathbf{b}_\beta}{\mathcal{C}_2} \tag{4.66}$$

$$\frac{\partial \log(\mathcal{C}_2)}{\partial \mathbf{c}_\beta} = \frac{2\sum_{\beta\neq s}(\mathbf{c}_s\mathbf{c}_s^T)\mathbf{c}_\beta}{\mathcal{C}_2} \tag{4.67}$$

Cas complexe :

Dans le cas complexe, le calcul des composantes du gradient de Υ_b se fait par rapport à une matrice complexe $\mathbf{A}$. Cela nous pousse à réécrire les deux parties de la fonction (4.60) sous l'une des formes présentées dans (Comon, 1986) (A.Hjorungnes et D.Gesbert, 2007). Commençons d'abord par la première partie, qui est Υ. On a :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) = \sum_{ijk}(T_{ijk} - \sum_{r=1}^{R}\lambda_r\, a_{ir}b_{jr}c_{kr})(T^*_{ijk} - \sum_{q=1}^{R}\lambda^*_q\, a^*_{iq}b^*_{jq}c^*_{kq}) \tag{4.68}$$

$$\begin{aligned}\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) &= ||T_{ijk}||^2 + \sum_{ijk}\sum_{r=1}^{R}\sum_{q=1}^{R}\lambda_r\lambda^*_q\, a_{ir}a^*_{iq}\, b_{jr}b^*_{jq}\, c_{kr}c^*_{kq} \\ &- 2\sum_{ijk}\Re\{T^*_{ijk}\sum_{r=1}^{R}\lambda_r\, a_{ir}b_{jr}c_{kr}\}\end{aligned}$$

Posons maintenant $M_{rq} = \sum_{jk}\lambda_r\lambda^*_q\, b_{jr}b^*_{jq}c_{kr}c^*_{kq}$ et $N_{ir} = \sum_{jk}T_{ijk}\,\lambda^*_r\, b^*_{jr}c^*_{kr}$. La nouvelle expression de Υ sera :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) = ||T_{ijk}||^2 + \sum_{i}\sum_{r=1}^{R}\sum_{q=1}^{R}a_{ir}M_{rq}a^*_{iq} - \sum_i(\sum_{q=1}^{R}N_{iq}a^*_{iq} + \sum_{r=1}^{R}a_{ir}N^*_{ir})$$

Le gradient de Υ par rapport à $\mathbf{A}$ sera :

$$\frac{\partial \Upsilon}{\partial \mathbf{A}} = 2\mathbf{A}\mathbf{M} - 2\mathbf{N} \tag{4.69}$$

De même, nous allons réecrire la deuxième partie de l'équation (4.60), $\log(\mathcal{C}_2)$, pour pouvoir calculer son gradient par rapport à $\mathbf{A}$.

$$\mathcal{C}_2 = 3(\frac{3}{2R+2})^2 - \sum_{r<q} \mathbf{a}_r^H(\mathbf{a}_q\mathbf{a}_q^H)\mathbf{a}_r - \sum_{r<q} \mathbf{b}_r^H(\mathbf{b}_q\mathbf{b}_q^H)\mathbf{b}_r - \sum_{r<q} \mathbf{c}_r^H(\mathbf{c}_q\mathbf{c}_q^H)\mathbf{c}_r \tag{4.70}$$

D'où le gradient par rapport à chaque colonne $\mathbf{a}_\ell$ de $\mathbf{A}$:

$$\frac{\partial \mathcal{C}_2}{\partial \mathbf{a}_\ell} = -2\sum_{q\neq\ell}(\mathbf{a}_q\mathbf{a}_q^H)\mathbf{a}_\ell = -2\left(\mathbf{A}\mathbf{A}^H - ||\mathbf{a}_\ell||^2\mathbf{I}\right)\mathbf{a}_\ell$$

et celui de $\log(\mathcal{C}_2)$ par rapport à la matrice $\mathbf{A}$, si ses colonnes sont de norme unité :

$$\frac{\partial \log(\mathcal{C}_2)}{\partial \mathbf{A}} = -\frac{2}{\mathcal{C}_2}\left(\mathbf{A}\,\mathbf{A}^H - \mathbf{I}\right)\mathbf{A} \tag{4.71}$$

Le gradient de la fonction (4.60) par rapport à $\mathbf{A}$ est finalement :

$$\frac{\partial \Upsilon_b}{\partial \mathbf{A}} = 2\mathbf{A}\mathbf{M} - 2\mathbf{N} - \frac{2\eta}{\mathcal{C}_2}\left(\mathbf{A}\,\mathbf{A}^H - \mathbf{I}\right)\mathbf{A} \tag{4.72}$$

Le calcul du gradient de Υ_b par rapport à $\mathbf{B}$ et $\mathbf{C}$ est similaire, en tenant compte du fait que les matrices $\mathbf{M}$ et $\mathbf{N}$ sont différentes.

Le principe de notre algorithme d'optimisation du problème (4.61) est donné dans la table 4.10. Cet algorithme est basé sur l'algorithme de point intérieur et Gradient projeté, et nous le nommerons "Algorithme Barrière-GP".

4.7.2.3 Simulations

Dans cette section, nous allons présenter les simulations réalisées pour illustrer le comportement de l'algorithme Barrière présenté antérieurement. Deux tenseurs différents sont simulés, $\mathcal{T}_1$ et $\mathcal{T}_2$. Le premier tenseur est de rang $R = 2$ et de taille $3 \times 5 \times 5$, alors que le deuxième et de rang $R = 3$ et de taille $13 \times 13 \times 13$. Le choix de la dimension du tenseur dépend de son rang et de la contrainte $\mathcal{C}_2$.

Pour illustrer la performance de notre algorithme, nous avons choisi de tracer l'erreur quadratique absolue $E = \frac{\|\mathcal{T}-\hat{\mathcal{T}}\|_F^2}{\|\mathcal{T}\|_F^2}$ appelée "Erreur sans barrière", ainsi qu'une deuxième erreur quadratique qui depend de la barrière $E' = \frac{\|\mathcal{T}-\hat{\mathcal{T}}\|_F^2 - \eta\log(\mathcal{C}_2)}{\|\mathcal{T}\|_F^2}$ appelée "Erreur avec Barrière". Tracer ces deux erreurs quadratiques nous permettra de savoir quand est ce que la contrainte de cohérence est active.

La Figure 4.18 représente la convergence de l'algorithme Barrière-GP par rapport au nombre d'itérations. Dans cette simulation, le tenseur utilisé est $\mathcal{T}_1$, tout en ajoutant une perturbation aux trois matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ ($SNR = 0db$). La recherche du pas est faite par la méthode "Backtracking" et l'erreur à atteindre est de 10^{-6}. L'initialisation des trois matrices à estimer $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ est faite arbitrairement.

La Figure 4.18 montre que, grâce à la contrainte $\mathcal{C}_2$, les solutions trouvées à chaque itération ($\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$) sont incitées à rester dans la région faisable, où l'existence est garantie. De plus, l'algorithme d'optimisation converge rapidement, tel que pour un nombre d'itération qui ne dépasse pas les 60, l'algorithme peut atteindre une erreur de

Algorithme Barrière-GP

1. Initialisation :
($\mathbf{A}_0$,$\mathbf{B}_0$,$\mathbf{C}_0$) point initial choisi arbitrairement, $\varepsilon > 0$ précision demandée,
$\eta^0 > 0$ et $t = 1.5$ paramètres de la barrière,
2. $k := 1$
3. Normaliser les colonnes des matrices $\mathbf{A}(k-1)$, $\mathbf{B}(k-1)$ et $\mathbf{C}(k-1)$
4. Calculer le $\Lambda^{(0)} = (\mathbf{G}^{(0)})^{-1}\mathbf{f}^{(0)}$
5. Test sur la contrainte $\mathcal{C}_2$:
   - Si $\mathcal{C}_2 \leq 0$, retourner à l'initialisation
   - Sinon : Tant que critère d'arrêt non satisfait,faire :
     - Direction de descente : $d^{(k)} = -\nabla\Upsilon_b(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$
     - Recherche d'un pas $\rho^{(k)}$ tel que :
$\Upsilon_b(\mathbf{A}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{A}}, \mathbf{B}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{B}}, \mathbf{C}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{C}}, \Lambda(k-1)) < \Upsilon_b(\mathbf{A}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{C}^{(k-1)}, \Lambda(k-1))$.
     - Mise à jour :
       - $\mathbf{A}^{(k)} = \mathbf{A}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{A}}$,
       - $\mathbf{B}^{(k)} = \mathbf{B}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{B}}$,
       - $\mathbf{C}^{(k)} = \mathbf{C}^{(k-1)} + \rho^{(k)}d^{(k)}_{\mathbf{C}}$,
       - $k = k + 1$;
       - Normaliser les nouveaux $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$
       - mettre à jour le $\Lambda^{(k)} = (\mathbf{G}^{(k)})^{-1}\mathbf{f}^{(k)}$.
     - Mettre à jour le $\eta^{(k)}$: $\eta^{(k)} = \eta^{(k-1)}/t$
6. Retourner $\mathbf{A}^{(k)}$, $\mathbf{B}^{(k)}$ et $\mathbf{C}^{(k)}$.

TABLE 4.10 – Résumé de l'algorithme barrière utilisant le gradient projeté

reconstruction de 10^{-7}. De plus, nous remarquons bien que la contrainte est très active dans les première itérations, ce qui pousse l'algorithme à décroître la valeur du paramètre η pour ramener le problème à l'intérieur de la contrainte $\mathcal{C}_2$.

Pour illustrer le comportement de l'algorithme Barrière-GP, nous considérons cette fois-ci le tenseur $\mathcal{T}_2$. À chaque valeur de SNR, un bruit blanc additif gaussien est ajouté au tenseur. La Figure 4.19 représente la convergence de l'algorithme Barrière-GP en fonction des valeurs du SNR.

4.8 Conclusion

Dans ce chapitre, nous avons présenté les différents algorithmes d'optimisation qui minimisent la décomposition approximative du tenseur $\mathcal{T}$.

En premier lieu, un état de l'art des méthodes d'optimisation mathématiques a été dressé. Ces méthodes peuvent être réunies en deux différents groupes : les méthodes déterministes et les méthodes stochastiques. Les méthodes déterministes peuvent trouver le minimum global de la fonction sous certaines hypothèses, comme la convexité et la différentiabilité. En d'autres termes, si la fonction objectif remplit ces hypothèses dans une région locale contenant le minimum désiré, et si la configuration initiale est quelque part à l'intérieur de cette région, les méthodes déterministes convergent très rapidement vers ce minimum.

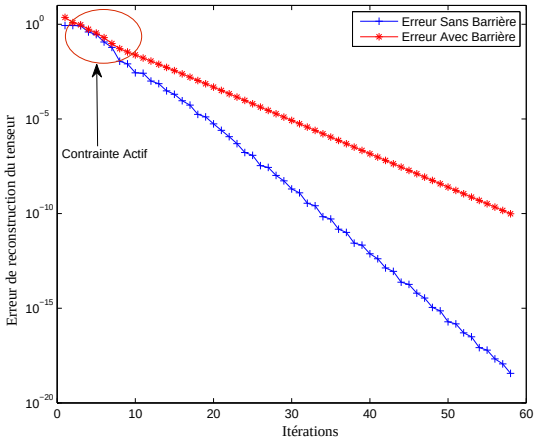

FIGURE 4.18 – Erreur de reconstruction du tenseur $\mathcal{T}_1$ en fonction de nombre d'itérations avec et sans l'utilisation de la fonction barrière.

Une étude a porté sur les aspects fondamentaux des algorithmes déterministes standard. Plus précisemment, les méthodes mathèmatiques (Gradient, Gradient conjugué, Levenberg Marquardt, Quasi Newton, BFGS ...), et des améliorations ont été proposées pour les rendre plus performants. L'étude de ces algorithmes mathématiques nous a montré leurs non-fonctionnement sur des problèmes d'optimisation sans contraintes.

En vu de notre problème d'optimisation qui est sous deux contraintes : une contrainte d'égalité imposée sur les colonnes des matrices de facteurs, et une contrainte d'inégalité sur les cohérences pour garantir l'existence et l'unicité de la solution ; il s'est avéré important de trouver des nouveaux algorithmes d'optimisation permettant de minimiser notre nouvelle fonction de coût. Trois nouvelles méthodes d'optimisation plus performantes ont été développées : "Algorithme 1", "Algorithme 2" et "Algorithme Barrière-GP". les deux premiers algorithmes ("Algorithme 1" et "Algorithme 2") sont des algorithmes d'optimisation dédiés à la minimisation des problème avec contrainte d'égalité. Ils sont basés sur la méthode de gradient projeté tout en prenant en compte l'isolement du facteur d'échelle. Le troisième algorithme proposé, qui est l'algorithme Barrière-GP, est conçu pour la résolution de notre fonction de coût sous les contraintes d'inégalité. Cet algorithme permet la transformation d'un problème d'optimisation avec contrainte d'inégalité en un problème sans contrainte, ce qui facilite la recherche de la solution.

Dans le chapitre suivant, nous allons présenter une méthode de séparation aveugle des signaux des systèmes radio mobiles, tout en utilisant les algorithmes présentés dans ce chapitre.

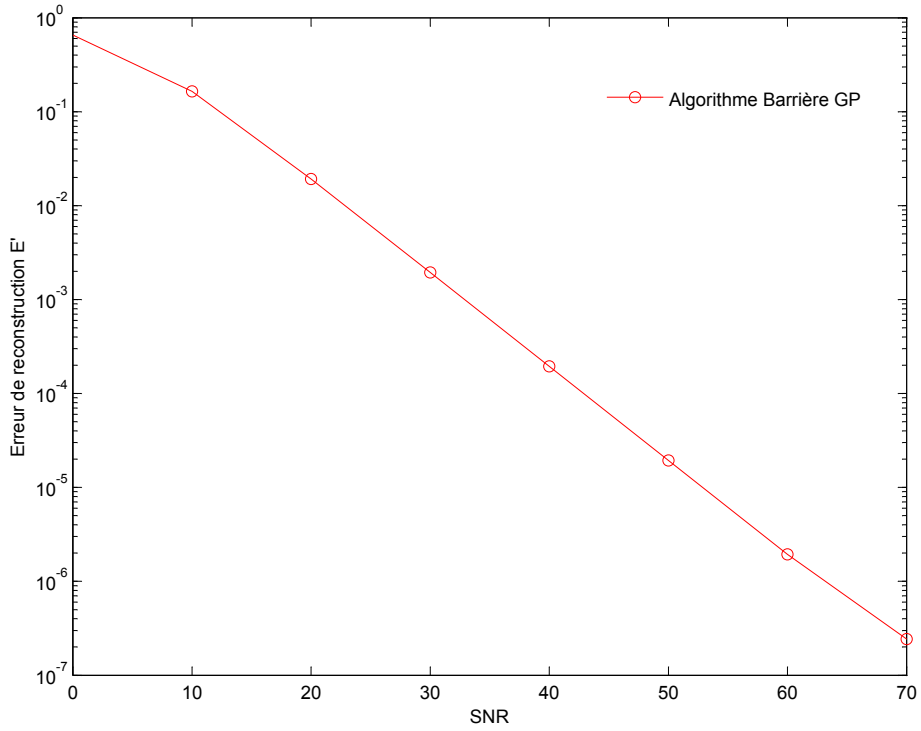

FIGURE 4.19 – Erreur de reconstruction du tenseur $\mathcal{T}_2$ en fonction de valeur de SNR en utilisant l'algorithme de Barrière GP.

Chapitre

5 Application aux systèmes radio mobiles

5.1 Introduction

Actuellement, nous assistons à l'avènement de la visiophonie et du visionnage de signaux audiovisuels sur des appareils de téléphonie mobile. Ces nouvelles fonctionnalités nécessitent des transmissions de plus en plus rapides, garantissant à la fois une flexibilité et une impérieuse efficacité au niveau de la qualité de service. Pour ce faire, les nouveaux réseaux de télécommunications doivent garantir un accès simultané des utilisateurs, de plus en plus nombreux, aux multiples services proposés par les différents opérateurs de téléphonie mobile. Plusieurs techniques de multiplexage ont successivement été proposées dans le but d'utiliser à bon escient la bande de fréquence disponible. La technique d'accès multiple à répartition en fréquence, communément connue sous le nom de FDMA (*Frequency Division Multiple Access*) est la méthode de partage de ressource spectrale la plus ancienne. Elle consiste à allouer à chaque utilisateur une bande de fréquence différente pour permettre des transmissions simultanées. Un autre procédé permet à tous les utilisateurs d'émettre sur l'ensemble de la bande de fréquence successivement dans le temps. Il s'agit de la technique d'accès multiple à répartition en temps, appelée plus communément TDMA (*Time Division Multiple Access*). Généralement, les techniques FDMA et TDMA sont combinées pour une meilleure exploitation de la bande de fréquence. En effet, le principal inconvénient de ces deux techniques est qu'il est difficile de gérer l'ensemble de la bande de fréquence de façon optimale. D'une part, avec la technique FDMA, si au cours du temps un utilisateur n'émet pas de signal, alors la bande de fréquence qui lui est allouée n'est pas utilisée. La combinaison des deux techniques rend l'allocation de ressources plus flexible. Cependant, ces deux techniques conservent une certaine rigidité qui nuit à l'utilisation optimale des ressources de transmission.

C'est pourquoi, une autre technique a fait son apparition dans les standards de télécommunication. Il s'agit de la technique d'accès multiple à répartition en code, appelée plus communément CDMA (*Code Division Multiple Access*). Cette technique permet à tous les utilisateurs de transmettre simultanément dans une même bande de fréquence au cours du temps. Ainsi, toutes les ressources disponibles sont exploitées de manière optimale. Par contre, la difficulté réside dans la séparation des signaux des différents utilisateurs. Pour ce faire, un code d'étalement spécifique est alloué à chaque utilisateur. Lors de la réception, ces codes servent à la distinction des signaux émis par les différents utilisateurs. Cette technique de multiplexage est désormais privilégiée par les nouveaux standards de téléphonie mobile. Ainsi, des normes radio-mobiles récentes telles que la

norme américaine CDMA 2000 et la norme européenne UMTS (*Universal Mobile Telecommunications System*) l'ont retenu comme technique d'accès. Au niveau technologique, les performances optimales d'un système mono-utilisateur sont atteintes par des systèmes multi-utilisateurs de type CDMA, en assignant aux différents utilisateurs des codes orthogonaux entre eux.

En 2002, l'équipe de Li Ping a proposé une nouvelle technique d'accès multiple baptisée IDMA (*Interleave Division Multiple Acces*) (Ping *et al.*, 2002). Cette technique s'avère être en fait un cas particulier de la technique CDMA. En effet, les utilisateurs sont distingués à l'aide d'entrelaceurs et non plus de codes orthogonaux. La technique IDMA bénéficie alors de plusieurs des avantages de la technique CDMA, notamment, des qualités de l'étalement de spectre. En effet, l'un des avantages de la technique d'étalement de spectre est sa robustesse face aux différents types de brouillage. De plus, ses propriétés d'autocorrélation permettent de tirer parti au mieux de la diversité des canaux multi-trajets à évanouissements. La principale caractéristique de la technique IDMA est la possibilité d'utiliser un même code d'étalement pour tous les utilisateurs. La distinction des signaux des différents utilisateurs est donc assurée par des entrelaceurs spécifiques. À la différence des codes pour la technique CDMA, la définition des différents entrelaceurs n'est pas fondamentalement un problème. Ainsi, un système IDMA constitué d'entrelaceurs générés de façon aléatoire présente des performances proches de la limite théorique d'un système multi-utilisateurs (Ping *et al.*, 2002). C'est la raison pour laquelle la technique IDMA a aussitôt attiré l'attention de la communauté scientifique. Les derniers travaux publiés sur le sujet laissent à penser qu'il s'agit d'une technique d'accès prometteuse pour les futurs systèmes de télécommunication.

Dans ce chapitre, nous proposerons d'appliquer les différents algorithmes de décomposition tensorielle présentés dans le chapitre précédent. Nous considérerons des scénarios de propagation et nous montrerons que les expressions analytiques des signaux reçus peuvent s'exprimer algébriquement à l'aide des formes de décompositions tensorielles des chapitres 3 et 4. Les approches que nous utiliserons dans ce chapitre pour détecter et séparer les signaux envoyés par les utilisateurs sont purement déterministes, et se basent sur l'exploitation de la structure algébrique des signaux reçus. Notre technique de détection et de séparation des signaux ne nécessite pas des longues trames d'observationn contrairement aux techniques statistiques, ce qui atténuera la stationnarité du canal. De plus, la contrainte d'orthogonalité des codes d'étalement n'est plus posée dans le cas de séparation aveugle par les méthodes tensorielles.

Dans les sections 5.2 et 5.3, nous donnerons respectivement un bref rappel de la chaîne de communication et de ses composantes, ainsi que les différentes techniques d'accès multiple. Dans la section 5.4, nous rappellerons le modèle de transmission des signaux CDMA dans le cas coopératif. Ensuite, nous montrerons que le modèle analytique de ces signaux peut s'écrire sous la forme d'un tenseur des observations dont sa décomposition est proposée dans le chapitre précédent. Puis, nous estimerons ces signaux d'une manière aveugle en s'appuyant sur les algorithmes d'optimisations proposés. La section 5.5 sera consacrée à une étude bibliographique sur la technique multi-utilisateurs IDMA. Dans cette section, nous donnerons la structure de l'émetteur ainsi que celle du récepteur dans le cas coopératif. Dans la section 5.7, nous proposerons une nouvelle technique d'accès multiple que

nous baptiserons OWDM-IDMA. Cette technique sera une combinaison de la technique IDMA et celle de multiplexage OWDM. Ensuite, nous donnerons une représentation algébrique des signaux IDMA dans la section 5.6. En effet, la repésentation tensorielle des signaux IDMA, proposée dans cette section, nous permettera d'estimer les signaux reçus sans avoir aucune informations sur ces derniers. La solution peut être obtenue alors par décomposition du tenseur des observations en utilisant les algorithmes proposé et présenté auparavant, et qui joueront ici le rôle du récepteur aveugle.

5.2 Chaîne de communication numérique

Les systèmes de transmission numérique véhiculent de l'information d'une source à un destinataire en utilisant un support physique (câble, fibre optique,...), ou encore la propagation sur un canal radioélectrique. Les signaux transportés peuvent être soit directement d'origine numérique, comme dans les réseaux de données, soit d'origine analogique (parole, image, etc.) mais convertis sous une forme numérique. Le propos de notre étude n'étant pas la numérisation de la source, le message délivré par cette dernière sera considéré d'origine numérique. La tâche du système de transmission est d'acheminer l'information de la source vers le destinataire avec la plus grande fiabilité possible.

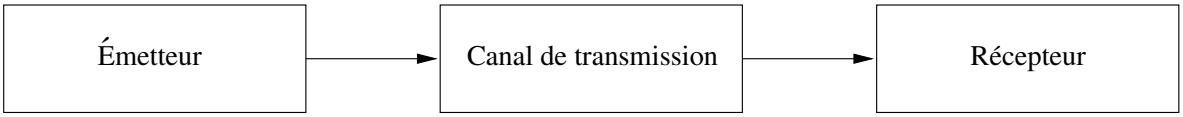

FIGURE 5.1 – Modèle de la chaîne de transmission numérique

Le schéma de principe d'une chaîne de transmission numérique est représenté sur la Figure 5.1. Elle peut se décomposer en trois blocs :

- l'émetteur,
- le milieu de transmission,
- le récepteur.

Dans un système de communication numérique, l'information à émettre est une séquence binaire. Dans le cas idéal, cette séquence doit être la plus courte possible. À l'émission, le codage de source est alors effectué dans le but de réduire le volume des données à transmettre. Celui-ci consiste à compresser les données tout en conservant l'information utile, permettant ainsi d'augmenter l'efficacité de la transmission et d'optimiser l'utilisation des ressources du système. Afin de transmettre le message numérique, l'émetteur doit associer au message une représentation physique sous la forme d'un signal qui s'adapte au canal de transmission. Le récepteur réalise l'opération inverse en reconstituant le message émis à partir du signal reçu. Ces deux opérations sont appelées respectivement modulation à l'émission et démodulation à la réception.

Le canal de transmission est un lien physique entre l'émetteur et le récepteur ; il diffère selon le type d'applications envisagées. Dans une étude algorithmique en communication numérique, il inclut généralement le milieu de transmission (fibre optique, espace libre, câble bifilaire, câble coaxial,...), le bruit (interne et externe) et éventuellement, les filtres

d'émission/réception, ou encore les antennes d'émission et réception. Dans le cadre de cette thèse, nous nous plaçons dans un contexte de communications sans fil. Quel que soit le milieu de transmission, le canal introduit des perturbations affectant le signal émis. C'est pourquoi, le signal reçu est altéré par du bruit et des distorsions qui peuvent induire le récepteur en erreur.

Sachant que le signal émis a subi des perturbations lors de sa propagation à travers le canal de transmission, l'objectif est de récupérer l'information transmise avec le moins d'erreurs possibles. Pour cela, le récepteur doit traiter le signal reçu en fonction des perturbations subies. Il est alors nécessaire de connaître le type du canal de transmission. En pratique, il faudra disposer d'un modèle de canal qui permet de définir au mieux l'environnement réel.

5.2.1 Canal de transmission

Dans ce chapitre, un canal à bruit blanc additif gaussien (AWGN) sera considéré lors des études théoriques et de la faisabilité d'un système. Étant considéré le contexte des communications sans fil, le modèle du canal multi-trajets sera également traité. Dans cette étude, nous représenterons la chaîne de transmission numérique par un modèle à temps discret, pour lequel les données que ça soient émises ou reçues s'écrivent sous forme d'une suite d'échantillons indexés par le temps.

Le canal AWGN est le modèle du canal le plus fréquemment utilisé pour la simulation de transmission numérique. Il représente le signal reçu comme étant la somme du signal émis et d'un bruit blanc additif gaussien. Par ailleurs, en plus de l'influence du bruit blanc additif, la puissance du signal émis peut être affectée par une atténuation, appelée aussi évanouissement, dont l'amplitude peut varier lentement ou rapidement dans le temps selon le contexte de transmission. C'est le cas d'une transmission à travers un canal de Rayleigh blanc. L'atténuation du signal est principalement due à un environnement de propagation riche en échos et donc caractérisé par des trajets multiples. Elle peut être également provoquée par le mouvement relatif de l'émetteur et du récepteur, entraînant des variations temporelles du canal. Ces phénomènes sont courants dans l'environnement radio-mobile.

5.2.1.1 Canal multi trajets

Les trajets multiples sont engendrés par les phénomènes physiques comme la réflexion et la diffraction, causées par le milieu de propagation (immeubles, collines, voitures, murs, ...), comme présenté dans le scénario sur la Figure 5.2. Dans le processus de modélisation du canal, seuls les trajets significatifs sont pris en compte, même si le nombre des trajets empruntés par un signal peut être important. Ainsi, pour un canal comportant L trajets significatifs, le récepteur reçoit L répliques du signal émis provenant de diverses directions, avec différents retards et atténuations. Les retards sont calculés par rapport au premier trajet détecté par le récepteur. Les échantillons du signal reçu $\mathbf{y}_j$ peuvent alors s'écrire comme la somme des échantillons du même signal émis, retardé et atténué, suivant L

trajets différents, avec des échantillons de bruit blanc additif gaussien :

$$\mathbf{y}_j(t) = \sum_{l=0}^{L} h_{j,l}\mathbf{x}_j(t-\tau_l) + n_j(t). \tag{5.1}$$

avec, $\mathbf{x}_j$ les échantillons du signal émis pour un bloc d'information de taille J, $h_{j,l}$ les coefficients d'atténuation et n_j désignent les échantillons du bruit blanc gaussien, de moyenne nulle et de variance $\sigma_n^2 = N_0$.

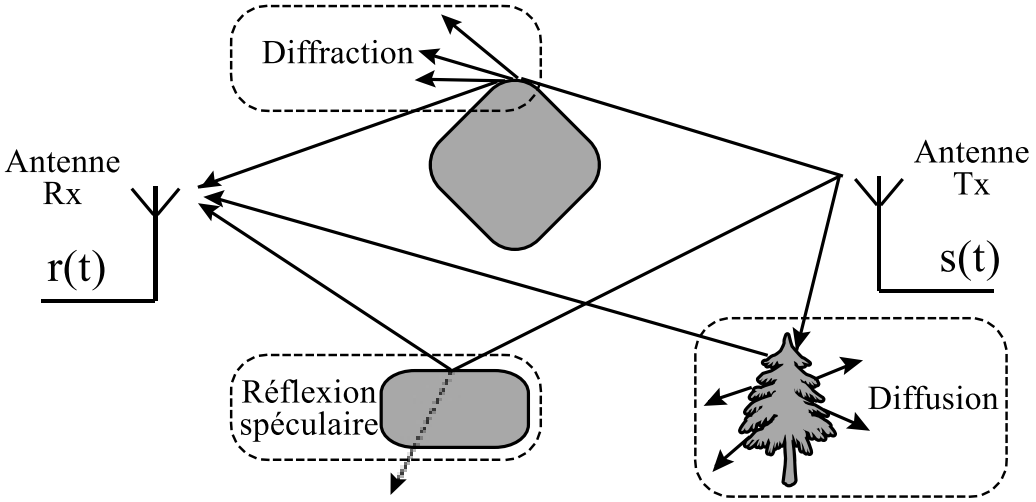

FIGURE 5.2 – Une transmission radio-mobile à travers un canal à trajets multiples

Par rapport au canal à trajet unique, le canal multi-trajets engendre le phénomène de variations du canal, caractérisées par : l'étalement temporel et l'effet Doppler. L'étalement temporel est le délai maximal des retards des trajets, noté τ_{max}. Il est défini par la différence entre le plus long et le plus court retard. Pour quantifier l'étalement temporel du canal, il est souvent préferable d'utiliser l'étalement moyen au sens des moindres carrés (Root Mean Square, RMS). Sa valeur est donnée par la relation (Deneire, 2003) :

$$\tau_{RMS} = \sqrt{\frac{\sum_{l=0}^{L-1}\tau_l^2\Omega_l}{\sum_{l=0}^{L-1}\Omega_l} - \bar{\tau}^2}. \tag{5.2}$$

où : Ω_l désigne la puissance du $l^{\text{ème}}$ trajet, défini par $\Omega_l = |h_{j,l}|^2$ et $\bar{\tau}$ représente le délai moyen défini par : $\bar{\tau} = \frac{\sum_{l=0}^{L-1}\tau_l\Omega_l}{\sum_{l=0}^{L-1}\Omega_l}$.

L'effet Doppler est un phénomène causé par le déplacement de l'émetteur/récepteur, ou encore par le déplacement des obstacles entre l'émetteur et le récepteur. Il est généralement modélisé par un décalage constant en fréquence, proportionnel à la fréquence porteuse et à la vitesse de déplacement. La fréquence Doppler peut être exprimée par la relation suivante :

$$f_{Doppler} = \frac{\upsilon f_c \cos(\alpha)}{c}. \tag{5.3}$$

où υ est la vitesse du déplacement du terminal, c est la célérité (la vitesse de la lumière), f_c est la fréquence porteuse et α désigne l'angle d'arrivée du signal.

La capacité du canal désigne la quantité d'information pouvant être transmise sans erreur sur le canal. En présence d'un bruit blanc additif gaussien et pour une entrée de

type gaussien, la capacité C du canal est défini par la relation suivante (Shannon, 1948) :

$$C = B \log(1 + \frac{P_S}{P_b}) \tag{5.4}$$

où B la bande occupée par le signal émis en Hertz (Hz), P_S et P_b désignent respectivement la puissance du signal émis et du bruit en Watt (W). On constate bien qu'à partir de l'équation (5.4), deux approches permettant d'augmenter la capacité peuvent être énoncées. La première consiste à utiliser une bande étroite avec un rapport $\frac{P_S}{P_b}$ important, alors que la seconde permet d'exploiter une bande large avec un faible rapport $\frac{P_S}{P_b}$. La technique d'étalement de spectre que nous allons aborder dans la section suivante est basée sur la seconde approche.

5.3 L'accès multiple

Afin d'obtenir une utilisation efficace des ressources disponibles, les utilisateurs des systèmes de communication, de plus en plus nombreux, sont amenés à cohabiter. Le problème posé par cette cohabitation, encore appelée accès multiple, consiste alors à examiner comment organiser l'accès d'un nombre important d'usagers à une ressource commune. Selon les ressources disponibles, il existe trois méthodes classiques permettant de gérer l'accès multiple : la technique TDMA, la technique FDMA et la technique CDMA (Figure 5.3).

La technique FDMA est la technique de partage de ressources spectrales la plus ancienne. En FDMA, la répartition est faite en découpant le spectre en canaux de largeur suffisante, en attribuant l'un de ces canaux à chaque utilisateur qui désire établir une communication. Cette technique d'accès multiple présente l'avantage d'être facilement implémentée puisqu'en réception, la dissociation des utilisateurs se fait par des opérations de filtrage. En revanche, l'inconvénient majeur de cette technique est le nombre maximal d'utilisateurs devront partager la bande totale. En effet, la largeur de la bande allouée à chaque utilisateur, diminuant avec l'accroissement du nombre des utilisateurs, ne doit pas être trop réduite afin d'éviter qu'à un instant donné, toutes les composantes spectrales d'un signal ne soient fortement atténuées.

La technique TDMA est une technique d'accès multiple basée sur la répartition des ressources dans le temps. En TDMA, on attribue aux utilisateurs des courts intervalles de temps, appelés fenêtres temporelles, pendant lesquels ils peuvent communiquer sur le canal. Un utilisateur se voit affecter une ou plusieurs fenêtres temporelles pour la durée de la communication. Généralement cette technique est plus difficile à implémenter que la FDMA puisqu'elle nécessite une synchronisation parfaite entre tous les émetteurs et les récepteurs.

Parfois, les techniques FDMA et TDMA sont combinées pour exploiter au mieux la ressource disponible. En effet, le principal inconvénient pour ces deux techniques est qu'il est difficile de gérer les ressources de façon optimale. D'une part, avec la technique FDMA, si à un certain moment un utilisateur n'émet pas de signal, alors la bande de fréquence qui lui est allouée n'est pas utilisée. D'autre part, avec la technique TDMA, si un utilisateur n'émet pas durant l'instant qui lui est réservé, alors l'intervalle de temps

qui lui est accordé n'est pas utilisé. Bien qu'il existe différentes méthodes d'allocation de ressources, ces deux techniques conservent une certaine rigidité, qui peut nuire à la capacité en nombre d'utilisateurs du système.

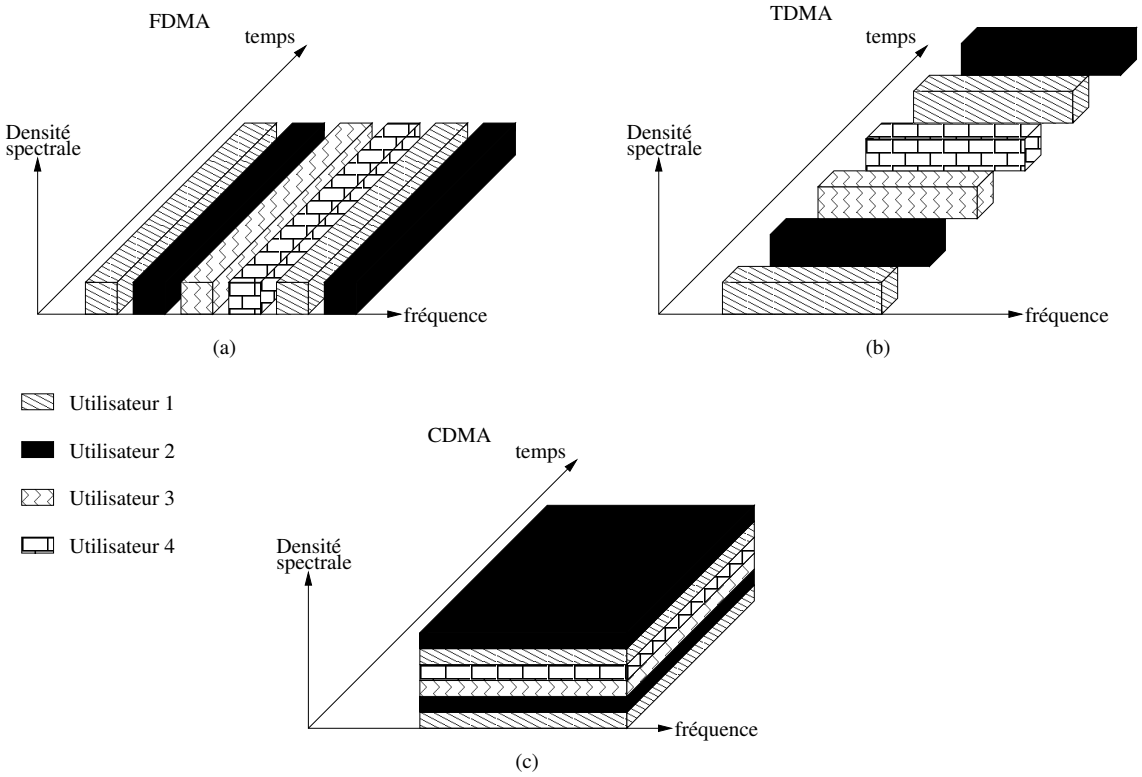

FIGURE 5.3 – Les trois principales techniques d'accès multiple : (a) FDMA, (b) TDMA, (c) CDMA

L'étalement de spectre exploité par la technique d'accès multiple à répartition en code CDMA permet de résoudre ce problème. En CDMA, un usager émet en permanence sur toute la bande en utilisant une technique d'étalement de spectre. Pour que cela soit possible, il faut que les signaux émis par les différents utilisateurs possèdent certaines propriétés permettant de les dissocier. Chaque utilisateur se voit donc affecter, pour la durée de la communication, un code spécifique appelé aussi signature. De plus, les codes de tous les utilisateurs doivent être orthogonaux ou du moins faiblement corrélés entre eux. À la réception, ces codes permettront de séparer les signaux émis par les différents utilisateurs. Cette technique est privilégiée par les nouveaux réseaux de la $3G$. En effet, des normes radio-mobiles récentes telles que la norme américaine CDMA 2000 et la norme européenne UMTS l'ont choisi comme technique d'accès. Les systèmes basés sur la technique CDMA ont été considérés comme des candidats potentiels pour l'interface air de la $4G$.

5.4 La technique CDMA : cas coopératif vs. cas aveugle

La technique CDMA utilise le bon comportement des signaux à étalement de spectre en présence d'autres signaux de même type. Une diversité des méthodes existent pour réaliser l'étalement de spectre, mais dans cette étude, nous allons nous intéresser seulement

à l'étalement de spectre par séquence directe, ce qui est connu sous le nom de la technique CDMA.

L'étalement de spectre par séquence directe consiste à multiplier chaque symbole à transmettre par un code ou une séquence à un débit nettement supérieur à celui du signal d'origine. Afin de rendre le signal étalé aléatoire, le code d'étalement $c = \{c_i; i = 0, \ldots, I-1\}$ composé de I chips est généralement une séquence pseudo-aléatoire notée PN pour Pseudo-Noise. Il possède les propriétés suivantes :

– une moyenne approximativement nulle

$$\sum_{i=0}^{I-1} c_i \approx 0 \tag{5.5}$$

– une propriété d'auto-corrélation donnée par

$$\rho(t) = \sum_{i=0}^{I-1} c_i c_{i+t} \approx \begin{cases} 1, & \text{si } t = 0 \\ 0, & \text{sinon} \end{cases}$$

avec $c_i = 0$ si $i > I - 1$

Si T_c correspond à la durée d'un *chip* et T_i à celle de la séquence d'étalement, le gain d'étalement est défini par :

$$I = \frac{T_i}{T_c} = \frac{B^{'}}{B} \tag{5.6}$$

où $B = \frac{1}{T_i}$ est la largeur de bande du signal non étalé et $B^{'} = \frac{1}{T_c}$ est celle du signal étalé. Le paramètre I est également appelé facteur d'étalement.

5.4.1 Émetteur CDMA

Soit $\mathbf{s}(r)$ l'information provenant de l'utilisateur r. Cette information est modulée en une séquence de chips $\mathbf{x}^{(r)}$ par un code pseudo-aléatoire $\mathbf{c}^{(r)}$. Étant normalisé, le code $\mathbf{c}^{(r)}$ correspondant à un utilisateur r prend sa valeur dans la base $B_i = \{\frac{+1}{\sqrt{I}}, \frac{-1}{\sqrt{I}}\}$. Les codes doivent être différents pour chaque utilisateur et orthogonaux entre eux. Ainsi, chaque utilisateur possède un code notée $\mathbf{c}^{(r)}(t)$ de durée T_i. Le signal reçu pour les R utilisateurs après échantillonage s'écrira comme suit :

$$\mathbf{y}_i(t) = \sum_{r=1}^{R} \mathbf{x}_{ir} + \mathbf{w}_i \tag{5.7}$$

tel que $\mathbf{x}_{ir}$ désigne le $i^{\text{ème}}$ *chip* transmis par le $r^{\text{ème}}$ utilisateur, $\mathbf{w}_i$ représentent les échantillons du bruit blanc additif gaussien de moyenne nulle et de variance $\sigma^2 = \frac{N_0}{2}$. La Figure 5.4 représente la structure de l'émetteur CDMA conventionnel pour R utilisateurs, émettant de manière parfaitement synchrone leurs messages.

5.4.2 Récepteur CDMA

A la réception, pour récupérer les messages envoyés par l'utilisateur m, $m = 1, \ldots, R$, le récepteur effectue un produit scalaire entre le signal reçu $\mathbf{y}$ et le code d'étalement

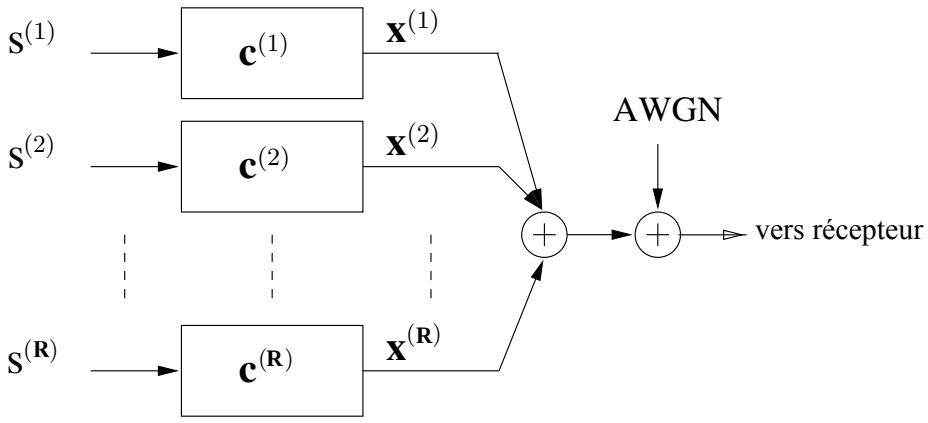

FIGURE 5.4 – Structure de l'émetteur CDMA

correspondant à l'utilisateur m.

$$\begin{aligned}
\langle \mathbf{y}^{(\mathbf{j})}, \mathbf{c}^{(m)} \rangle &= \frac{1}{I} \sum_{i=1}^{I} y_{ij}, c_{im} \\
&= \frac{1}{I} \sum_{i=1}^{I} \sum_{r=1}^{R} x_{jir} c_{im} \\
&= \frac{1}{I} \sum_{i=1}^{I} \sum_{r=1}^{R} s_{jr} c_{ir} c_{im} \\
&= \sum_{r=1}^{R} s_{jr} \langle \mathbf{c}_r, \mathbf{c}_m \rangle \\
&= s_{jm}
\end{aligned} \tag{5.8}$$

Le récepteur retrouve donc le symbole s_{jm} en utilisant la propriété d'orthogonalité des séquences d'étalement $\mathbf{c}_r$.

Les familles de codes d'étalement les plus couramment utilisées dans les systèmes CDMA sont (Mahafeno, 2007) :

- les codes de Gold qui sont bien adaptés aux systèmes synchrones ; ceux de longueur $2^{18} - 1$ et $2^{41} - 1$ ont d'ailleurs été retenus pour l'UMTS,
- les codes de Kasami, qui ont été choisis comme codes de scrambling pour la liaison montante de certains standards du système $3G$ (CDMA 2000, UTRA (UMTS Terrestrial Radio Access)),
- les codes de Walsh Hadamard utilisés dans les systèmes CDMA synchrones, en particulier dans les liaisons descendantes des communications radio-mobiles.

5.4.3 les codes d'étalement

Dans les systèmes CDMA, afin de restituer correctement les informations numériques relatives à chaque utilisateur, il est important que les signaux des différents usagers soient les plus décorrélés possible les uns aux autres. Pour cela, on a plus de libérté dans le choix judicieux des codes d'étalement, qui seront attribués par la suite aux différents

utilisateurs en fonction des caractéristiques de communication. Ainsi, en présence d'une communication synchrone sur un canal AWGN non sélectif, les performances optimales peuvent être obtenues par l'utilisation de codes orthogonaux, tels que les codes de Walsh-Hadamard ou encore les codes de type OVSF (*Orthogonal Variable Spreading Factor*) (Dell'Amico *et al.*, 2002). En revanche, en présence d'un canal sélectif en fréquence ou en temps, l'utilisation d'autres familles de codes permet de se rapprocher des performances optimales. Parmi ces familles de codes, on peut notamment citepr les codes de Gold, les codes de Kasami, les codes de Zadoff-Chu, etc (Faqihi, 2009).

5.4.3.1 Les codes de Walsh-Hadamard

Les codes de Walsh-Hadamard sont générés à partir de la matrice de transformation de Sylvester-Hadamard. Plus exactement, ils correspondent aux lignes ou aux colonnes orthogonales de cette matrice composée de $(+,-)1$. La matrice de transformation de Sylvester-Hadamard de taille $2^n \times 2^n$ satisfait la condition suivante :

$$H_m H_m^T = m I_m \tag{5.9}$$

où H_m^T est la matrice transposée de la matrice de Sylvester-Hadamard de taille $m \times m$ et I_m est la matrice identité de taille $m \times m$. Ainsi, d'après cette définition, les lignes ou les colonnes sont mutuellement orthogonales. Le fait d'interchanger les lignes ou les colonnes n'affecte en rien les propriétés d'une telle matrice.

La matrice de transformation de Sylvester-Hadamard de taille $L \times L$ peut être construite récursivement de la manière suivante :

$$\begin{cases} H_1 = +1 \\ H_L = \begin{pmatrix} H_{L/2} & H_{L/2} \\ H_{L/2} & -H_{L/2} \end{pmatrix} \end{cases}$$

5.4.3.2 Les codes de Golay

Tout comme les codes de Walsh-Hadamard, les codes de Golay sont obtenus à partir d'une matrice construite récursivement. En effet, les codes de Golay correspondent aux lignes de la matrice CG_L de taille $L \times L$ (avec L $= 2^n$ et $n \neq 0$) définie par :

$$\begin{cases} CG_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = [A_2 B_2] \\ CG_L = [A_L B_L] \end{cases}$$

avec :

$$\begin{cases} A_L = \begin{pmatrix} A_{L/2} & B_{L/2} \\ A_{L/2} & B_{L/2} \end{pmatrix} \\ B_L = \begin{pmatrix} A_{L/2} & -B_{L/2} \\ -A_{L/2} & B_{L/2} \end{pmatrix} \end{cases}$$

où les matrices A_L et B_L sont de tailles $L \times L/2$. Par exemple, en posant $L = 4$, on obtient :

$$CG_4 = \begin{pmatrix} +1 & +1 & +1 & -1 \\ +1 & -1 & +1 & +1 \\ +1 & +1 & -1 & +1 \\ +1 & -1 & -1 & -1 \end{pmatrix}$$

Les codes de Golay, étant orthogonaux comme les codes de Walsh-Hadamard, sont bien adaptés aux systèmes de transmission synchrones. Ils ont également la particularité d'être complémentaires deux à deux. Deux codes SC_i et SC_j sont dits complémentaires si et seulement si :

$$\Gamma_{SC_i}(k) + \Gamma_{SC_j}(k) = 2L\delta(k) \tag{5.10}$$

5.4.4 Simulations dans le cas coopératif

Dans le but de vérifier la faisabilité du système CDMA, un canal AWGN et une conversion bit/symbole BPSK (*Binary Phase Shift Keying*) sont considérés, ainsi, les codes de Walsh Hadamard seront utilisés dans un premier temps.

Nous considérons $R = 16$ utilisateurs qui transmettent simultanément leurs messages de taille $J = 100$, ces derniers seront étalés en utilisant des codes de Walsh Hadamard de taille $I = 6$. La Figure 5.5 présente les performances du système CDMA sur un canal AWGN. Sur cette figure, nous avons tracé le BER (*Bit Error Rate*) en fonction du rapport signal à bruit SNR tout en prenant la courbe du BER théorique (en bleu) comme référence. Sur la Figure 5.5, on remarque bien que la courbe des performances du système CDMA converge vers celle théorique, ce qui nous laisse dire que le système CDMA atteint les limites théoriques.

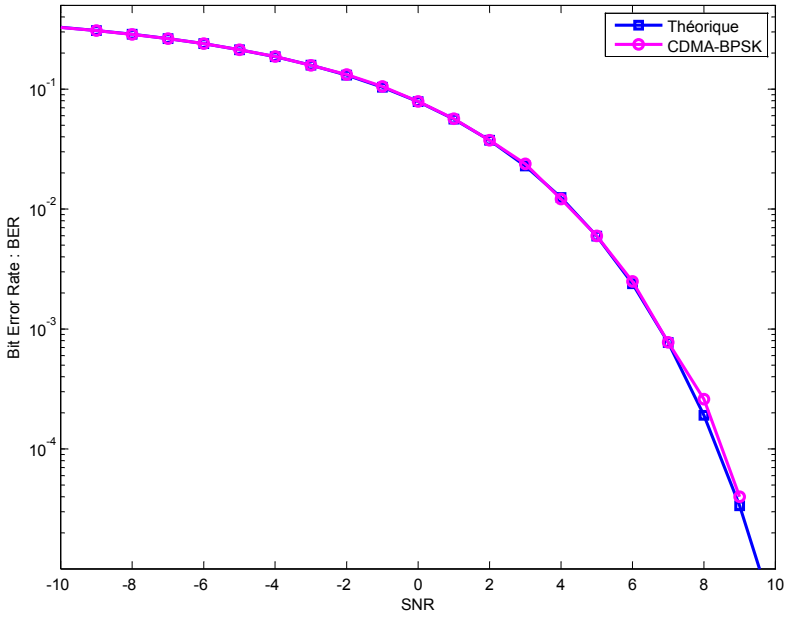

FIGURE 5.5 – Performances du système CDMA sans codage sur canal AWGN pour 16 utilisateurs et une conversion bit/symbole BPSK

Dans la deuxième simulation, nous considérons la conversion bit/symbole QPSK (*Quadrature Phase Shift Keying*), et nous reprenons les mêmes conditions de la première simulation.

À partir des résultats présentés sur la figure , on peut dire que le système CDMA avec une modulation QPSK converge aussi bien que le système CDMA avec une modulation BPSK, sauf que pour cette nouvelle simulation, les performances de ce système s'éloignent un peut de la limite théorique pour des valeurs de SNR supérieures à $6db$.

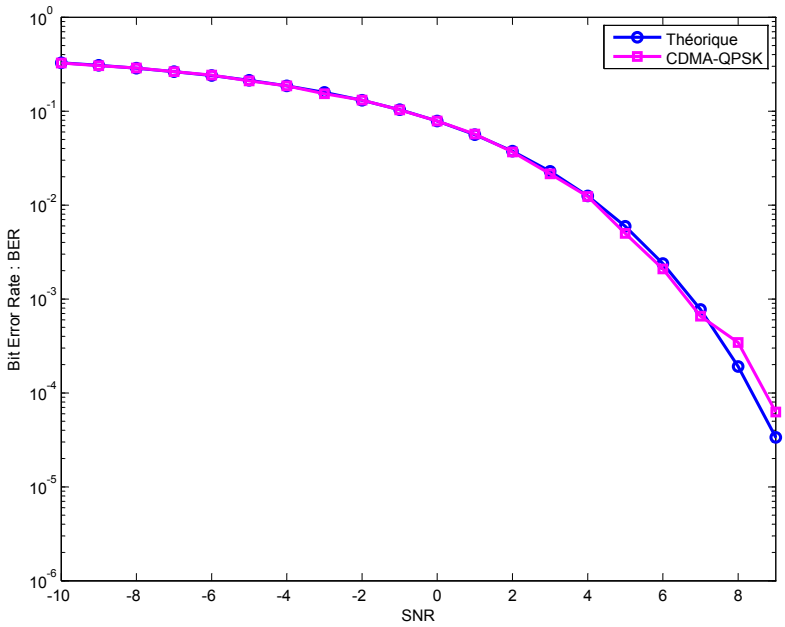

FIGURE 5.6 – Performances du système CDMA sans codage sur canal AWGN pour un nombre d'utilisateurs = 16 et une conversion bit/symbole QPSK

Dans le cas coopératif, où l'émetteur coopére avec le récepteur en lui envoyant des informations, les hypothèses choisies sont très fortes : le récepteur doit connaître les séquences d'apprentissage, qui doivent être orthogonales (ou pseudo-orthogonales). Dans le cas aveugle que nous allons aborder maintenant, on suppose évidemment inconnues les séquences d'étalement, mais nous allons aussi supposer qu'elles ne sont pas nécessairement orthogonales.

5.4.5 Le système CDMA dans le cas aveugle (Rouijel *et al.*, 2014a)

Dans les systèmes de coopération civil actuels, la connaissance du code d'étalement est exploitée à la réception afin d'extraire le signal d'un utilisateur particulier, notamment par l'annulation des interférences des autres utilisateurs. Toutefois, dans le cas d'une écoute discrète, le signal intercepté résulte d'une somme de tous les signaux transmis simultanément dans la même bande de base, sans connaissance des codes d'étalement CDMA ni des séquences d'apprentissage (Nion, 2007). Dans cette section, nous nous plaçons dans un contexte aveugle. On suppose que le récepteur n'a aucune information sur les codes d'étalement, ni sur les signaux émis par les différents utilisateurs. Classiquement, les techniques aveugles en télécommunications sont basées sur des propriétés connues a priori, telles les propriétés temporelles des signaux transmis ainsi que les propriétés spatiales du récepteur (Moulines *et al.*, 1995; Godard, 1980; der Veen et Paulraj, 1996). Indépendamment des propriétés utilisées, les problèmes aveugles sont classiquement formulés en termes du calcul matriciel, et impliquent la résolution d'un problème de décomposition de matrice de la forme : $\mathbf{Y} = \mathbf{H}.\mathbf{S}$, lorsque $\mathbf{H}$ caractérise par exemple le canal de propagation et $\mathbf{S}$ contient les symboles numériques transmis (Nion, 2007).

Plus récemment, les méthodes d'algèbre multilinéaire ont retenu une attention particulière dans le domaine de traitement du signal (Comon et Jutten, 2010). Il s'avère

également que les outils d'algèbre multilinéaire sont souvent plus puissants que leurs équivalents matriciels. Sidiropoulos et al. sont les premiers à adopter le tenseur dans le secteur des télécommunications (Sidiropoulos *et al.*, 2000a). Ils ont observé que les échantillons d'un signal CDMA reçu par un réseau d'antennes peuvent être stockés dans un cube, chaque dimension correspondant à une diversité (diversité de codage, temporelle et spatiale). Ainsi, ils ont montré que le problème de la séparation aveugle des signaux CDMA déterministe peut être résolu par la décomposition PARAFAC (de Almeida *et al.*, 2007). Résoudre le problème de séparation de sources revient alors à déterminer les paramètres de la décomposition.

Dans cette section, nous proposerons d'appliquer les algorithmes d'optimisation de la décomposition CP comme décrites dans le chapitre précédent, en adoptant le nouveau critère de performance pour la technique CDMA. Une comparaison avec l'algorithme d'ALS est ensuite effectuée.

5.4.5.1 Modélisation tensorielle

Nous considérons R utilisateurs avec une antenne d'émission. Ces utilisateurs transmettent simultanément leurs signaux à un ensemble de K antennes réceptrices. En d'autres termes, nous considérons un système de communication de type "CDMA-SIMO" (Rouijel *et al.*, 2014a).

Par exemple, supposons que l'utilisateur r transmet les symboles $\mathbf{s}_r$ de taille J. Ces symboles sont étalés par un code d'étalement $\mathbf{c}_r$ de longueur I, alloué uniquement à l'utilisateur r. Après l'étalement, la séquence est transmise sur un canal sans mémoire, et reçu par le réseau des antennes sous l'angle d'arrivée θ_r. Chacune des K antennes reçoit alors un signal $\mathbf{y}_k(t)$ de longueur $J \times I$. Notre approche pour la détection et la séparation des signaux reçus est d'exploiter la structure algébrique multilinéaire de ces signaux en utilisant notre nouveau critère de performance. Durant un laps de temps de durée $J.T_s$, nous observons les signaux $\mathbf{y}_k(t)$, où T_s est la période d'un symbole. Ces signaux sont échantillonés à la période de *chip* $T_c = T_s/I$, où I désigne le facteur d'étalement. Donc, chaque antenne fournit une matrice contenant $I \times J$ échantillons, et la concaténation de ces K matrices selon la troisième dimension permet de construire un tenseur des observations d'ordre 3 et de dimension $I \times J \times K$.

Reprenons le modèle présenté en (5.7), mais cette fois en considérant les K antennes :

$$y_{ijk} = \sum_{r=1}^{R} a_{kr} s_{jr} c_{ir} \quad i \in [1, I] \quad j \in [1, J] \quad k \in [1, K]. \tag{5.11}$$

L'élément y_{ijk} correspond à l'échantillon du signal global reçu par la $k^{\text{ème}}$ antenne au $i^{\text{ème}}$ instant d'échantillonnage de la $j^{\text{ème}}$ période symbole. Il peut être vu comme l'élément d'indice ijk d'un tenseur $\mathcal{Y}$ de taille $I \times J \times K$:

$$\mathcal{Y} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{s}_r \circ \mathbf{c}_r \tag{5.12}$$

où le scalaire $a_{kr} = \mathbf{a}_k(\theta_r)$ est la réponse de l'antenne k selon l'angle d'incidence θ_r du trajet provenant de l'utilisateur r. $\mathbf{a}_r$, $\mathbf{s}_r$, $\mathbf{c}_r$ représentent respectivement le vecteur des

coefficients d'antenne, le vecteur d'information de l'utilisateur r et la séquence d'étalement allouée à l'utilisateur r.

L'équation (5.12) est une décomposition CP du tenseur $\mathcal{Y}$. La structure CP des données CDMA a été relevée pour la première fois dans (Sidiropoulos *et al.*, 2000b). En effet, Sidiropoulos *et al* ont montré que pour un canal de propagation sans mémoire, le modèle algébrique vu dans l'équation (5.12) correspond à la décomposition CP du tenseur $\mathcal{Y} \in \mathbb{C}^{I\times J\times K}$. Chaque dimension du tenseur $\mathcal{Y}$ correspond à une diversité à la réception. En effet, la première dimension correspond à la divesité de codage I, la deuxième dimension représente la diversité temporelle J et la troisième à la diversité spatiale K. Ainsi, la séparation des signaux reçus est alors équivalente à la décomposition du tenseur $\mathcal{Y}$ en une somme de R contributions, où R représente le nombre d'utilisateurs actifs dans le système.

Pour calculer la décomposition CP de $\mathcal{Y}$, nous proposons l'Algorithme 2 présenté dans la section 3.3.3 en prenant en compte le nouveau critère de performance (3.30). La détection et la séparation de la matrice des symboles transmis $\mathbf{S}$ seront faites, en utilisant la fonction de coût suivante :

$$\Upsilon(\mathbf{A},\mathbf{C},\mathbf{S},\Lambda) = \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{c}_r \circ \mathbf{s}_r\|_F^2. \tag{5.13}$$

L'équation (5.13) nous permettra d'éliminer l'ambiguïté d'échelle, et cela en normalisant les trois matrices $\mathbf{A}$, $\mathbf{C}$ et $\mathbf{S}$, et en calculant la valeur exacte du facteur d'échelle Λ. Pour corriger les phases des trois matrices estimées ainsi que les ambiguïtées de permutation, nous allons utiliser l'indice de performance exacte (5.14) proposé dans la section 3.3.3 (Rouijel *et al.*, 2014a) :

$$\mathcal{E}(\mathcal{Y};\mathbf{A},\mathbf{C},\mathbf{S},\Lambda) = \min_{\pi\in\Pi} \sum_{r=1}^{R} (\min_{\varphi,\psi,\chi} \{||\mathbf{a}_r - e^{\jmath\varphi}\hat{\mathbf{a}}_{\pi(r)}||^2 + ||\mathbf{c}_r - e^{\jmath\psi}\hat{\mathbf{c}}_{\pi(r)}||^2 + ||\mathbf{s}_r - e^{\jmath\chi}\hat{\mathbf{s}}_{\pi(r)}||^2\}). \tag{5.14}$$

Supposons maintenant que les signaux après étalement se propagent selon plusieurs trajets avant d'atteindre le récepteur. Pour ne pas faire la distinction entre les angles d'arrivée des trajets issus d'un même utilisateur, nous supposons que les réflecteurs sont situés seulement dans l'environnement proche des utilisateurs. Cela signifie que l'étalement des angles d'arrivée est négligeable car les antennes ne visualisent qu'un seul trajet provenant de chaque utilisateur. Le modèle (5.12) reste valable dans le cas où il y a de l'interférence entre chips (ICI) mais pas d'interférence entre symboles (ISI) (Sidiropoulos *et al.*, 2000b). Il suffit alors de remplacer dans (5.12) c_r par h_r, où h_r correspond au produit de convolution entre la séquence d'étalement du $r^{\text{ème}}$ utilisateur et la réponse impulsionnelle du canal correspondant. Le modèle s'écrit maintenant :

$$\mathcal{Y} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{s}_r \circ \mathbf{h}_r \tag{5.15}$$

Donc, en présence des interférences ISI, détecter et séparer les signaux des différents utilisateurs revient à estimer les trois matrices $\mathbf{A}$, $\mathbf{S}$ et $\mathbf{H}$, en minimisant en premier

temps la fonction de coût suivante :

$$\Upsilon(\mathbf{A}, \mathbf{C}, \mathbf{S}, \Lambda) = \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{h}_r \circ \mathbf{s}_r\|_F^2. \tag{5.16}$$

ensuite nous corrigeons les phases des trois matrices estimées par le critère de performance proposé.

5.4.5.2 Simulations dans le cas aveugle

Afin d'illustrer le comportement et les performances de notre récepteur aveugle proposé (Algorithme 2), nous présentons dans cette section deux simulations différentes. Dans la première simulation, nous nous plaçons dans le cas d'un canal sans mémoire. Nous considérons $K = 4$ antennes réceptrices, $R = 4$ utilisateurs, chacun de ces derniers possède un code d'étalement de taille $I = 10$ et émet un message de $J = 20$. Les symboles de la matrices S sont modulés suivant une constellation QPSK, et les angles d'arrivée des utilisateurs pris dans cette simulation sont présentés dans la table 5.1.

	angles d'arrivée
$(\theta_1, \theta_2, \theta_3, \theta_4)$	$(-60, -30, 0, 20)$

TABLE 5.1 – Angles d'arrivée pour 4 utilisateurs

La Figure 5.7 illustre les performances de notre récepteur aveugle "Algorithme 2", utilisant le nouveau critère de performance, et ceux du récepteur ALS. Ces résultats indiquent que la performance du récepteur proposé "Algorithm2" est meilleure que celle d'ALS. De plus, nous pouvons voir que le BER de l'Algorithme 2 sans l'utilisation de notre critère de performance est très optimiste. Cela est dû à la négligence de contrainte sur les angles dans le critère de performance classique utilisé dans la littérature, ce qui prouve l'intérêt de notre étude.

Reprenons la même simulation, mais cette fois pour 6 utilisateurs, avec les nouvelles valeurs des angles d'arrivées décrites sur la table 5.2. La Figure 5.8 illustre l'estimation des angles d'arrivées en fonction du SNR. À partir de cette figure, nous pouvons voir que l'algorithme 2 permet une estimation exacte des angles d'arrivée pour des valeurs du SNR supérieures à zéro.

	angles d'arrivée
$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$	$(-90, -60, -30, 0, 30, 60)$

TABLE 5.2 – Angles d'arrivée pour les 6 utilisateurs.

L'impact du facteur $\frac{K}{R}$, où K représente le nombre des antennes réceptrices, et R le nombre des utilisateurs, est illustré sur la Figure 5.9. Sur cette figure, la première courbe (avec des losanges) représente le cas où le nombre des antennes est plus grand que le nombre des utilisateurs, $K = 4$ et $R = 2$. Pour la deuxième courbe (avec des cercles), on a $K = 4$ et $R = 4$, alors que pour la troisième courbe (avec des carrés), $K = 2$ et $R = 5$. les valeurs des angles d'arrivées sont données dans la table 5.3.

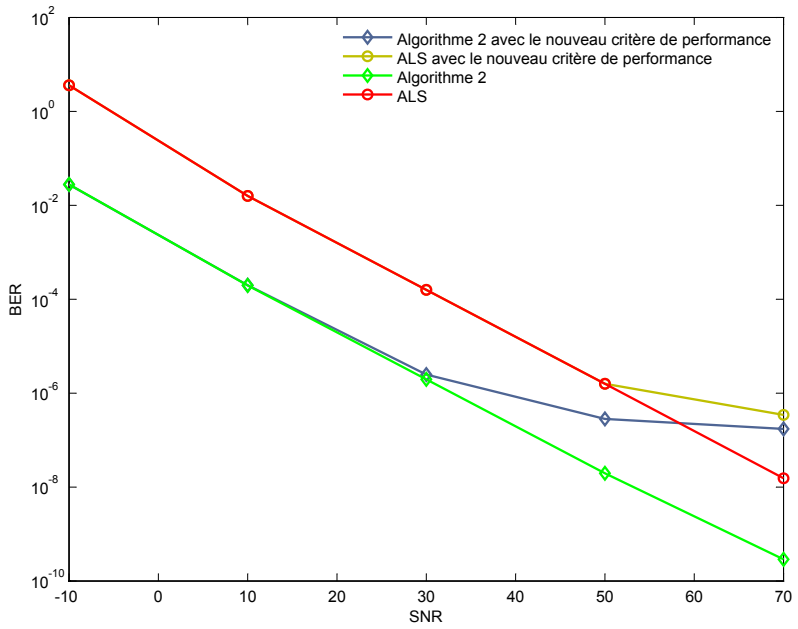

FIGURE 5.7 – BER en fonction du SNR pour le scénario : R = 4, K = 4, I = 10, J = 20 et constellation= QPSK

	angles of arrival
Curve 1	(-10, 20)
Curve 2	(-60, -40, 0, 20)
Curve 3	(-60,-50, 10, 40, 80)

TABLE 5.3 – Les angles d'arrivées pour les 3 scénarios

À partir de cette figure, on voit bien que la performance globale du système est améliorée lors de l'augmentation de ce facteur, ce qui indique l'importance de la diversité spatiale.

Dans la littérature, différentes méthodes de détection multi-utilisateurs ont été proposées (Patel et Holtzman, 1994; Madhow et Honig, 1994; Lupas et Verdu, 1989) pour le cas coopératif. Ces détecteurs donnent de bonnes performances pour une complexité raisonnable quand le taux de charge n'est pas élevé (en général $\tau_c < 100\%$). Rappelons que τ_c définit le rapport entre le nombre d'utilisateurs et le facteur d'étalement :

$$\tau_c = \frac{R}{I} \tag{5.17}$$

Par contre, lorsque les interférences MAI augmentent de façon significative, ou lorsque les interférences ISI s'ajoutent à celles-ci, les performances de ces détecteurs se dégradent rapidement. Une technique itérative qui permet de prendre en compte les interférences MAI, en associant à la détection un codage de canal et un entrelacement, est alors introduite par Li ping et son équipe (Ping *et al.*, 2002, 2006). De plus, cette technique permet d'atteindre un taux de charge qui dépassera les 100%, comme nous allons présenter dans la section suivante.

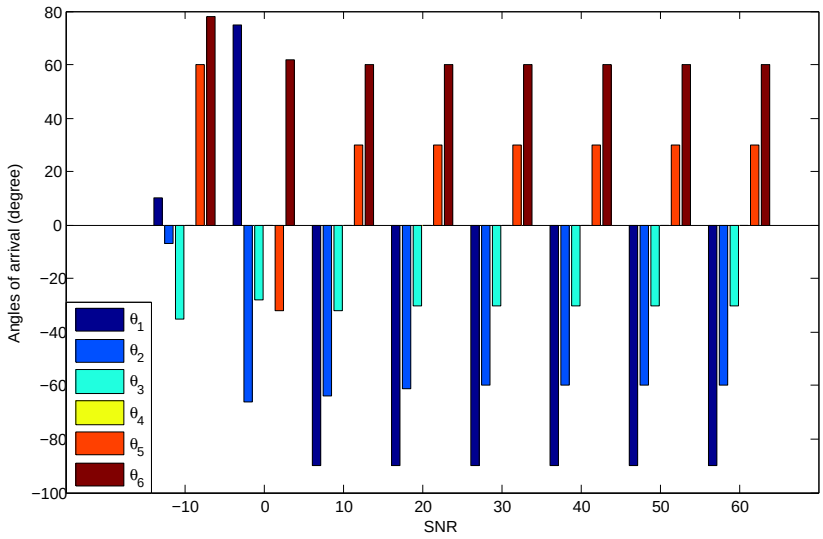

FIGURE 5.8 – Estimation des angles d'arrivées pour le scénario : R = 6, K = 6, I = 10, J = 20

5.5 La technique multi utilisateurs IDMA

Le système CDMA à entrelacement des *chips* cI CDMA (*chip Interleaved Code Division Multiple Access*), proposé en 2002 par R. H. Mahadevappa et J. G. Proakis (Mahadevappa et Proakis, 2002), permet d'éliminer conjointement les interférences MAI et les interférences ISI, en présence d'un canal multi-trajets.

Pour ce faire, un traitement itératif est utilisé à la réception. La détection multi-utilisateurs correspondante repose à la fois sur une technique d'annulation des interférences, et sur une technique de combinaison MRC (*Maximum Ratio Combining*). Les performances de ce système, sans codage de canal, s'approchent des performances mono-utilisateur sans interférences ISI même quand le nombre des utilisateurs R est égal au facteur d'étalement I, ce qui correspond à un taux de charge de 100% (Mahadevappa et Proakis, 2002).

Parallèlement, Li Ping et. al. ont proposé une nouvelle technique d'accès multiple qu'ils ont baptisé IDMA. Cette technique s'avère être un cas particulier du cI CDMA, pour lequel les utilisateurs sont distingués à l'aide de différents entrelaceurs. La technique IDMA bénéficie de différents avantages de la technique CDMA, notamment, des bonnes propriétés d'étalement du spectre évoquées dans les sections précédentes de ce chapitre. Mais contrairement au système CDMA, tous les utilisateurs peuvent avoir le même code d'étalement. En effet, les entrelaceurs sont le seul moyen de distinguer entre les différents utilisateurs. Par ailleurs, dans les études initialement publiées (Ping *et al.*, 2002)(Ping *et al.*, 2006), les performances du système IDMA sont proches de la limite théorique d'un système multi-utilisateurs, bien qu'aucune optimisation ne soit effectuée lors de la conception des entrelaceurs. Ces derniers sont en effet générés de façon aléatoire. C'est pourquoi, la technique IDMA a attiré l'attention de la communauté scientifique, tout d'abord par l'innovation qu'elle apporte, mais également parce que les premières publications sur cette technique montrent qu'elle semble prometteuse pour l'interface air de la $4^{\text{ème}}$ génération de communications mobiles.

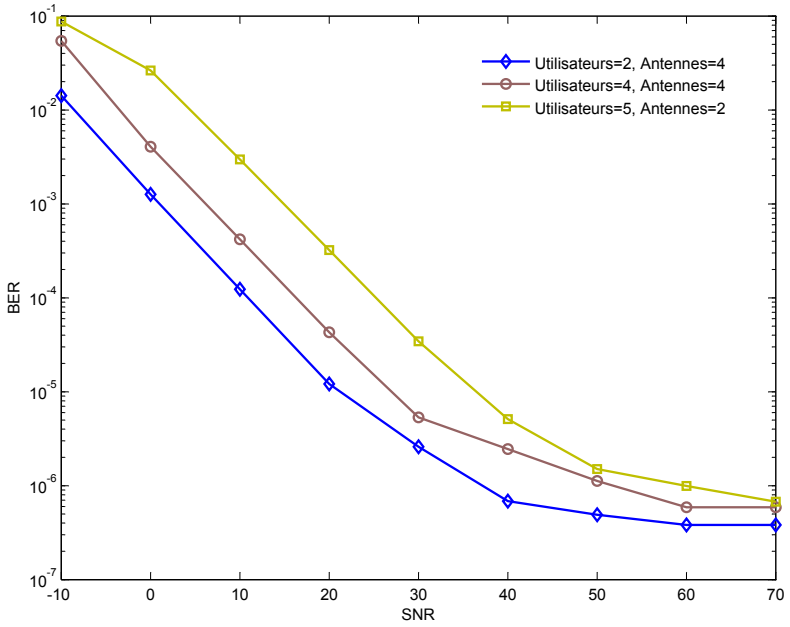

FIGURE 5.9 – Performance du récepteur en fonction du SNR pour : {2 utilisateurs, 4 antennes}, {4 utilisateurs, 4 antennes}, {5 utilisateurs, 2 antennes}

Au cours de ces dernières années, d'autres équipes de recherche ont commencé à s'intéresser à la technique IDMA. La plupart des investigations se sont focalisées sur l'étude des performances du système IDMA dans le contexte des réseaux de communications actuels et futurs. Ainsi, dès 2004, un système à base de la technique IDMA, pouvant répondre aux futurs besoins des systèmes de la 4G sur la voie montante a été proposé par une équipe de recherche de l'université de Kiel (Schoeneich et Hoeher, 2006). En 2005, Zhou et. al. ont quant à eux proposé le système IDMA comme une technique pouvant améliorer la capacité d'accès multiple d'une communication sur la voie descendante au delà de la 3G (Zhou *et al.*, 2005). En 2006, une étude d'un système IDMA dans le cadre de réseau Ad Hoc a été publiée par une équipe de DoCoMo Eurolabs (Kusume et Bauch, 2006).

5.5.1 Émetteur IDMA

La technique d'accès multiple IDMA permet d'allouer une même bande de fréquences à plusieurs utilisateurs, tout en leur permettant d'émettre simultanément. Cette méthode d'accès est alors semblable à celle du système CDMA. La différence entre les deux techniques réside dans la manière de distinguer les différents utilisateurs et de séparer leurs signaux. Pour un système IDMA, la signature d'un utilisateur est caractérisée par un entrelaceur spécifique et non par un code d'étalement particulier comme pour la technique CDMA.

La Figure 5.10 représente la structure de l'émetteur IDMA pour un utilisateur r quelconque, ou $r = 1, \ldots, R$ et R est le nombre maximal des utilisateurs considérés. Cette structure est similaire à celle du système cI CDMA pour R utilisateurs.

Soit $\mathbf{s}_r = \{s_r(j); j = 1, \ldots, J\}$ l'information provenant de l'utilisateur r. J représente la taille de chaque bloc d'information. Celle-ci est codée en une séquence $\mathbf{d}_r = \{d_r(j); j =$

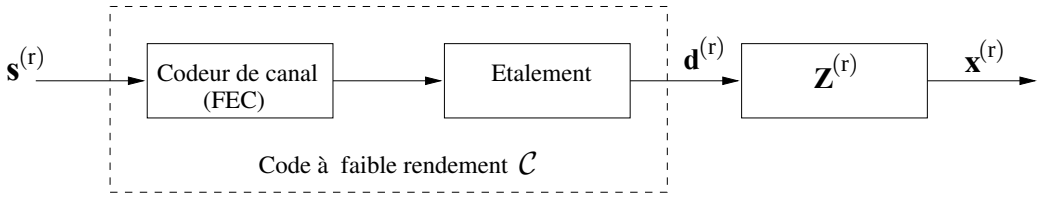

FIGURE 5.10 – Structure de l'émetteur IDMA pour un utilisateur r.

$1, \ldots, J\}$ en utilisant un code à faible rendement, noté $\mathcal{C}$. Ce code peut être identique ou différent pour chaque utilisateur, ce qui différencie le système IDMA du système cI CDMA. En effet, ce dernier utilise un code d'étalement spécifique à chaque utilisateur. Dans le cas présent, le code $\mathcal{C}$ peut être, soit un code correcteur d'erreurs ou FEC (*Forward Error Correction*), un code d'étalement, ou une combinaison des deux. Son principe sera décrit en détail dans la sous-section 5.5.1.1.

Le codeur est suivi d'un entrelaceur, noté $\mathbf{Z}_r$ pour un utilisateur r quelconque. Celui-ci transforme la séquence d'information étalée d_r en sa version entrelacée $\mathbf{x}_r = \{x_r(j)\}$. En respectant la convention pour un système CDMA, nous appelons chips les données étalées ou codées $\{\mathbf{d}_r\}$, en sortie du code à faible rendement $\mathcal{C}$. L'entrelacement est appliqué au niveau chips comme pour le système cI CDMA. Les entrelaceurs $\{\mathbf{Z}_r; r = 1, \ldots, R\}$ sont donc des entrelaceurs chips qui doivent être spécifiques à chaque utilisateur. Ces entrelaceurs sont en effet le seul moyen de distinguer entre les différents utilisateurs. Les entrelaceurs $\mathbf{Z}_r$ sont générés indépendamment et sont fixés durant toute la transmission. Dans notre étude, nous considérons des entrelaceurs générés de façon aléatoire.

Le signal reçu correspondant à R utilisateurs peut s'écrire sous la forme suivante :

$$y(j) = \sum_{r=1}^{R} h_r(j)x_r(j) + n(j) \qquad j = 1, 2, ..., J \tag{5.18}$$

où h_r est le coefficient du canal pour l'utilisateur r et $n(j)$ sont des échantillons du bruit AWGN, de moyenne nulle et de variance $\sigma_2 = N_0/2$. Nous supposons que les coefficients h_r du canal sont connus a priori.

5.5.1.1 Principe de codage

Sur la Figure 5.10, le code $\mathcal{C}$ est constitué d'une concaténation série d'un code correcteur d'erreurs et d'un code d'étalement. Toutefois, il peut également n'être constitué que de l'un ou de l'autre. Le système est dit non-codé si le code $\mathcal{C}$ est constitué uniquement d'un code d'étalement, tandis qu'un système codé utilisera toujours la combinaison d'un code correcteur d'erreurs et d'un code d'étalement. Par ailleurs, nous avons étudié les performances du système IDMA sans utilisation d'un code correcteur d'erreur. Cette étude sera abordée dans la section 5.5.5. Dans ce système, l'étalement est une forme particulière de codage. Il consiste à multiplier la séquence d'information par un code d'étalement. À la différence d'un système CDMA, l'utilisation des codes orthogonaux n'est pas a priori nécessaire, puisque pour un système IDMA, les utilisateurs sont distingués à l'aide des différents entrelaceurs.

Considérons un étalement binaire, chaque donnée d'information $s_r(j)$ à émettre est étalée dans le temps à l'aide d'un code à répétition, $c_r = \{c_r(i); i = 1, \ldots, I\}$ de taille I. I étant le facteur d'étalement défini dans les sections précédentes. Un *chip* correspondant à la donnée d'information émise par le $r^{\text{ème}}$ utilisateur peut alors s'écrire comme suit :

$$\mathbf{x}_r = \mathbf{s}_r \times \mathbf{c}_r \tag{5.19}$$

Une conversion bit/symbole BPSK étant considérée, le *chip* émis $x_r(j)$ prend alors sa valeur dans la base $B = \{+1, -1\}$. Dans cette étude, chaque utilisateur utilise le même code d'étalement, construit à partir d'une séquence de $+1$ et de -1 uniformément répartie (Liu *et al.*, 2003)(Ping *et al.*, 2002)(Ping *et al.*, 2006).

5.5.2 Récepteur IDMA

Afin de récupérer les données d'information émises par chacun des utilisateurs, le système fait appel à un récepteur itératif de type turbo. Celui-ci utilise une technique de détection et de décodage conjoints. Le récepteur proposé dans (Ping *et al.*, 2002) utilise une stratégie de détection *chip* par *chip*. Pour un utilisateur considéré, celle-ci consiste à créer une estimation séparée des interférences engendrées par les autres utilisateurs, afin de les soustraire au signal reçu. La Figure 5.11 représente la structure du récepteur IDMA proposé dans (Ping *et al.*, 2002).

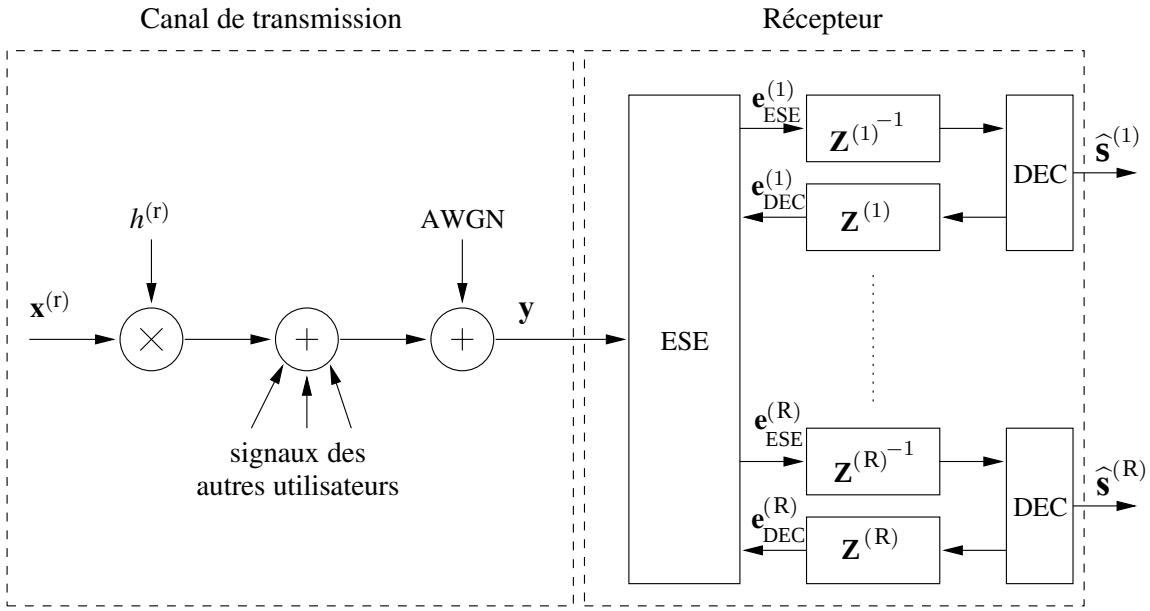

FIGURE 5.11 – Structure du récepteur IDMA pour R utilisateurs

Le récepteur est constitué d'un détecteur de *chip* gaussien que nous appellerons ESE (*Elementary Signal Estimator*), et de décodeurs DEC APP (*A Posteriori Probability*). Dans le processus itératif de récupération des données d'information, le détecteur ESE et les décodeurs DEC s'échangent une information sur les bits émis de manière turbo. Après un nombre suffisant d'itérations, les décodeurs DEC prennent des décisions strictes sur les bits d'information émis. Celles-ci sont notées $\hat{\mathbf{s}}_r$ pour un utilisateur r considéré.

Le récepteur permet d'estimer les bits à partir du signal reçu $\mathbf{Y} = \{\mathbf{y}_j\}$ et des coefficients du canal $\{\mathbf{h}_r\}$. Le récepteur optimal qui traite conjointement le code $\mathcal{C}$ et les entrelaceurs pour un canal à accès multiple, a une complexité calculatoire qui augmente exponentiellement avec les paramètres, rendant sa réalisation difficile. Considérer une approche sous-optimale permet de réduire la complexité. Elle consiste à effectuer des traitements séparés du canal à accès multiple et du code $\mathcal{C}$, en se basant sur une ou plusieurs conditions, puis à combiner par la suite les résultats en utilisant un processus itératif.

La sortie du détecteur ESE, $e_{ESE}(x_r) = \{e_{ESE}(x_r(j))\}$, est définie par le logarithme du rapport de vraisemblance extrinsèque, noté LLRExt (LLR : *Logarithm Likelihood Ratio*). Or, d'après le critère MAP (*Maximum A Posteriori*), et en utilisant la loi de Bayes, la probabilité a posteriori peut être exprimée comme suit (Proakis, 1995) :

$$P(x_r(j)) = \frac{P(\mathbf{Y}|x_r(j)) \times P(x_r(j))}{P(\mathbf{Y})} \tag{5.20}$$

où $P(x_r(j))$ est la probabilité a priori. Sachant que le dénominateur de l'équation 5.20 est indépendant du signal émis, cette équation peut se simplifier comme suit (Ping *et al.*, 2002) :

$$P(x_r(j)) = P(\mathbf{Y}|x_r(j)) \times P(x_r(j)) \tag{5.21}$$

D'où

$$LLR_{APP} = LLR_{Ext} + LLR_{AP} \tag{5.22}$$

Ainsi :

$$\begin{aligned} e_{ESE}(x_r(j)) &= LLR_{APP} - LLR_{AP} \\ &= \log \frac{P(\mathbf{Y}|x_r(j) = +1, \mathbf{h})}{P(\mathbf{Y}|x_r(j) = -1, \mathbf{h})} \qquad \forall r, j \end{aligned} \tag{5.23}$$

Dans le cas d'un canal mono-trajet sans mémoire, l'équation 5.23 peut être réécrite comme suit :

$$e_{ESE}(x_r(j)) = log\frac{P(\mathbf{y}(j)|x_r(j) = +1, \mathbf{h})}{P(\mathbf{y}(j)|x_r(j) = -1, \mathbf{h})} \qquad \forall r, j \tag{5.24}$$

Passant au décodeur, il est constitué de R décodeurs APP locaux. Le $r^{\text{ème}}$ décodeur effectue un décodage APP du code pour l'utilisateur r, en utilisant la version désentrelacée de $e_{ESE}(x_r)$. Sa sortie est définie par le LLR extrinsèque suivant :

$$e_{DEC}(x_r(j)) = log(\frac{P(x_r(j) = +1|e_{ESE}(x_r), C)}{P(x_r(j) = -1|e_{ESE}(x_r), C))}) - e_{ESE}(x_r(j) \qquad \forall r, j \tag{5.25}$$

Le LLR $e_{DEC}(x_r) = e_{DEC}(x_r(j))$ est l'information de retour ou *feedback* pour le $r^{\text{ème}}$ utilisateur. Il sera utilisé lors de la prochaine itération par le détecteur ESE comme le montre la Figure 5.11.

5.5.3 Détecteur multi-utilisateurs ESE

Le détecteur multi-utilisateurs ESE permet de générer des estimations pour les $x_r(j)$. La contrainte $\mathcal{C}$ du code étant ignorée (Ping *et al.*, 2002), le $j^{\text{ème}}$ *chip* du $r^{\text{ème}}$ utilisateur $x_r(j)$ est considéré comme une variable aléatoire. Le LLR extrinsèque $e_{DEC}(x_r(j))$ fourni

par le $r^{\text{ème}}$ décodeur APP est utilisé comme LLR a priori au niveau du détecteur ESE. Il s'écrit alors :

$$e_{DEC}(x_k(j)) = \log(\frac{P(x_k(j)=+1)}{P(x_k(j)=-1)}) \tag{5.26}$$

À la première itération, aucun LLR extrinsèque n'est disponible, le LLR a priori est alors initialisé à zéro pour le premier calcul du détecteur ESE. À partir de l'équation 5.26, le *chip* $x_r(j)$ est estimé en calculant sa moyenne $\mu_r(j)$ et sa variance $\vartheta_r(j)$, données par les équations suivantes :

$$\mu_r(j) = E(x_r(j)) \approx \frac{exp(e_{DEC}(x_r(j))) - 1}{exp(e_{DEC}(x_r(j))) + 1} = \tanh \frac{e_{DEC}(x_r(j))}{2} \tag{5.27}$$

$$\vartheta_r(j) = Var(x_r(j)) \approx 1 - (\mu_r(j))^2 \tag{5.28}$$

Le $j^{\text{ème}}$ *chip* du signal reçu peut s'écrire sous la forme :

$$\mathbf{y}(j) = \mathbf{h}_r x_r(j) + \xi_r(j) \tag{5.29}$$

Et cela en considérant l'équation 5.18 et en notant :

$$\xi_r(j) = \mathbf{y}(j) - \mathbf{h}_r x_r(j) \tag{5.30}$$

où $\xi_r(j)$ correspond au terme d'interférence plus bruit. Il peut être approché par une variable aléatoire gaussienne dont la moyenne et la variance s'expriment comme suit :

$$E(\xi_r(j)) = E(\mathbf{y}(j)) - \mathbf{h}_r \mu_r(j) \tag{5.31}$$

$$Var(\xi_r(j)) = Var(\mathbf{y}(j)) - (\mathbf{h}_r)^2 \vartheta_k(j) \tag{5.32}$$

En tenant compte de l'équation 5.18, le LLR dans l'équation 5.24 peut se réécrire sous la forme :

$$e_{ESE}(x_r(j)) = 2\mathbf{h}_r \times \frac{\mathbf{y}(j) - E(\xi_r(j))}{Var(\xi_r(j))} \tag{5.33}$$

Bien qu'issue d'un calcul de LLR donné par la relation 5.23, l'équation 5.33 met en évidence une annulation des interférences MAI. Le résumé de l'algorithme de détection *chip* est donné dans l'annexe C.

Après cette dernière opération, la donnée d'information $e_{ESE}(x_k(j))$ nécessaire au décodage est obtenue. Le décodeur calcule alors le LLR extrinsèque $e_{DEC}(x_k(j))$. Une fois le décodage terminé, les calculs de l'équation C.2 à l'équation C.8 sont réitérés.

5.5.4 Décodeur à probabilité à posteriori du système IDMA

Le codeur peut utiliser soit un code correcteur d'erreurs, un code à répétition, ou une combinaison des deux. Selon le type de codage, le décodage peut être un désétalement plus un décodage du canal, ou l'un des deux. Dans notre étude nous nous intéressons seulement au désétalement.

5.5.4.1 Décodage avec simple désétalement

Dans le cas ou le codeur correspond à un simple code à répétition, le processus de décodage consiste en un désétalement, c'est-à-dire une simple somme donnée par l'équation 5.34, où $c_r(i)$ représente l'élément du code à répétition correspondant à la donnée d'information émise. Le schéma bloc du récepteur correspondant à un utilisateur r quelconque est représenté par la Figure 5.13.

$$L_k^{'} = \sum_{i=0}^{I-1} c_r(i) e_{ESE}(x_r(j)) \tag{5.34}$$

Les $\{L_r^{'}\}$ ainsi obtenus sont donc les LLR a posteriori du décodeur. Ils sont ensuite réétalés pour pouvoir calculer les informations extrinsèques $\{e_{DEC}(x_r(j))\}$. Ces dernières sont alors données par :

$$e_{DEC}(x_r(j)) = (\sum_{i^{'}=0}^{I-1} c_r(i^{'}) e_{ESE}(x_r(j)))c(i) - e_{ESE}(x_r(j)) \qquad \forall j \tag{5.35}$$

Elles sont ensuite entrelacées avant d'être traitées par le détecteur ESE.

L'estimation des interférences s'affine au fur et à mesure des itérations. Pour ce faire, deux procédés de traitement de l'information peuvent être effectués : le traitement parallèle ou le traitement série.

En ce qui concerne le premier traitement, l'estimation et la suppression des interférences s'effectuent de façon parallèle, ce qui veut dire que lors de chaque itération, les opérations données par les équations C.2 à C.8 (Voir Annexe C) sont effectuées simultanément pour tous les utilisateurs. Tous les $\{e_{ESE}(x_r(j))\}$ de chaque utilisateur r sont alors obtenus en même temps. Les correspondants sont alors mis à jour simultanement pour l'itération suivante. La Figure 5.12 représente La structure du récepteur IDMA avec un traitement parallèle (Mahafeno, 2007).

Pour le traitement en série, l'estimation et la suppression des interférences sont réalisées successivement pour chaque utilisateur (Mahafeno, 2007). Pour ce faire, le processus peut commencer par l'utilisateur le plus puissant pour que les performances du système convergent rapidement au fur et à mesure des itérations. Dans notre étude, tous les utilisateurs ont la même puissance. C'est pourquoi, nous avons choisi de commencer arbitrairement le traitement par le premier utilisateur. Dans ce cas, à chaque itération, les opérations sur (ESE, DEC1) correspondant au 1$^{\text{er}}$ utilisateur sont tout d'abord effectuées, puis celles sur (ESE, DEC2) correspondant au 2$^{\text{ème}}$, ainsi de suite jusqu'au $R^{\text{ème}}$ utilisateur. Cela signifie que lors de chaque itération, $e_{ESE}(x_1(j))$ est obtenu en premier, en effectuant les calculs donnés par les équations C.2 à C.8. Celui-ci va être utilisé pour la mise à jour de $e_{DEC}(x_1(j))$. Ensuite, $e_{ESE}(x_2(j))$ est calculé avec la nouvelle valeur de $e_{DEC}(x_1(j))$. Il est par la suite utilisé pour la mise à jour de $e_{DEC}(x_2(j))$, et ainsi de suite jusqu'à l'obtention de $e_{ESE}(x_R(j))$. À l'itération suivante, ce procédé est réitéré jusqu'à la dernière itération. Le processus itératif correspondant à ce traitement est donné dans l'Annexe C. La Figure 5.13 représente la structure du récepteur IDMA avec un traitement série. Lors du traitement série, le système doit a priori présenter une meilleure convergence que lors d'un traitement parallèle. En effet, au cours de l'itération courante, chaque

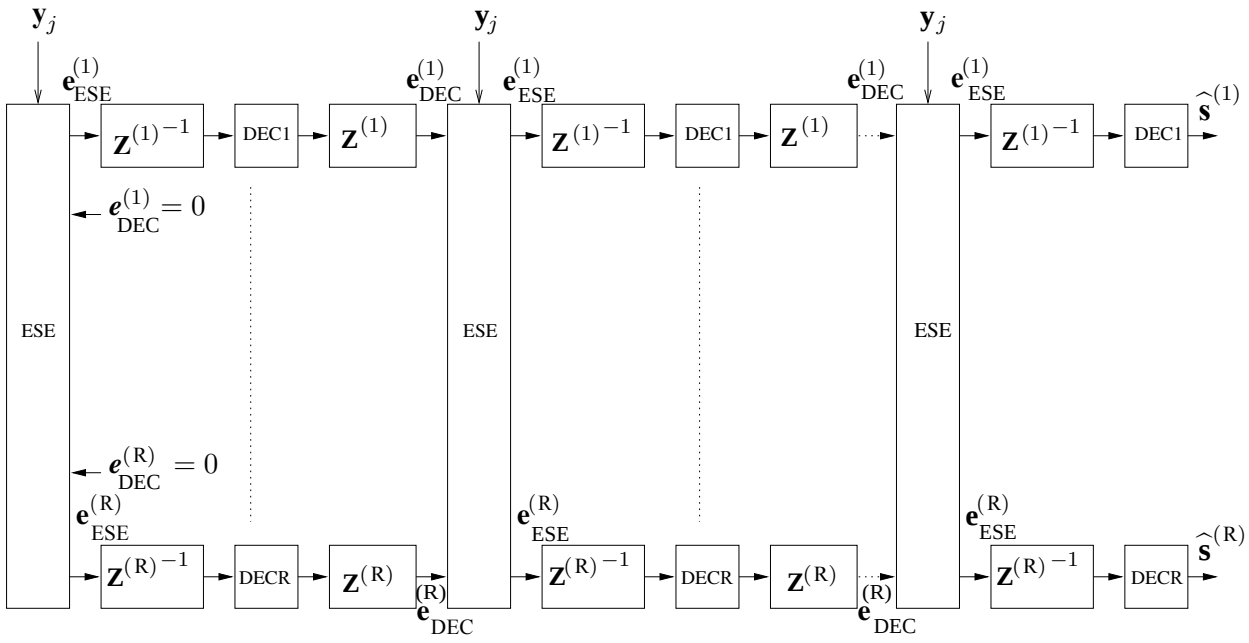

FIGURE 5.12 – Structure du récepteur IDMA utilisant un traitement parallèle

utilisateur bénéficie des informations des utilisateurs traités préalablement, ce qui n'est pas le cas lors du traitement parallèle. Par contre, le traitement parallèle possède moins de latence que le traitement série, du fait que les utilisateurs sont traités simultanément.

5.5.5 Performances du système IDMA dans le cas d'un canal mono-trajet

Les résultats de simulation présentés dans cette section ont pour objectif de valider la technique IDMA proposée dans (Ping *et al.*, 2002). Ils mettent en évidence l'intérêt du système IDMA en terme des performances. Pour ce faire, des canaux mono-trajets sont considérés. Par la suite, des performances montrant l'intérêt du code à répétition sont présentées.

Tout d'abord, quelques précisions sont nécessaires à la compréhension des résultats et des courbes de performances obtenues :

- La qualité de transmission est une des performances du système IDMA, elle est évaluée en terme du taux d'erreur binaire BER. Ce dernier est mesuré en utilisant le paramètre E_b/N_0 pour un utilisateur.
- Pour un système IDMA sans codage du canal, le rendement global est $R_G = 1/\text{I}$, le taux de charge a pour expression :

$$\tau_c = \frac{R}{I}$$

Dans le but de vérifier la faisabilité du système IDMA, un canal AWGN et une conversion bit/symbole BPSK sont considérés dans un premier temps. Les résultats de simulation sont obtenus en faisant varier différents paramètres : le nombre d'utilisateurs R, la taille du bloc d'information J, la taille du code à répétition I. le codage utilisé ici consiste simplement en une multiplication par la séquence $[+1; -1; +1; -1; \ldots; +1; -1]$ de taille I. Les simulations ont été effectuées en utilisant les paramètres suivants pour chaque utilisateur :

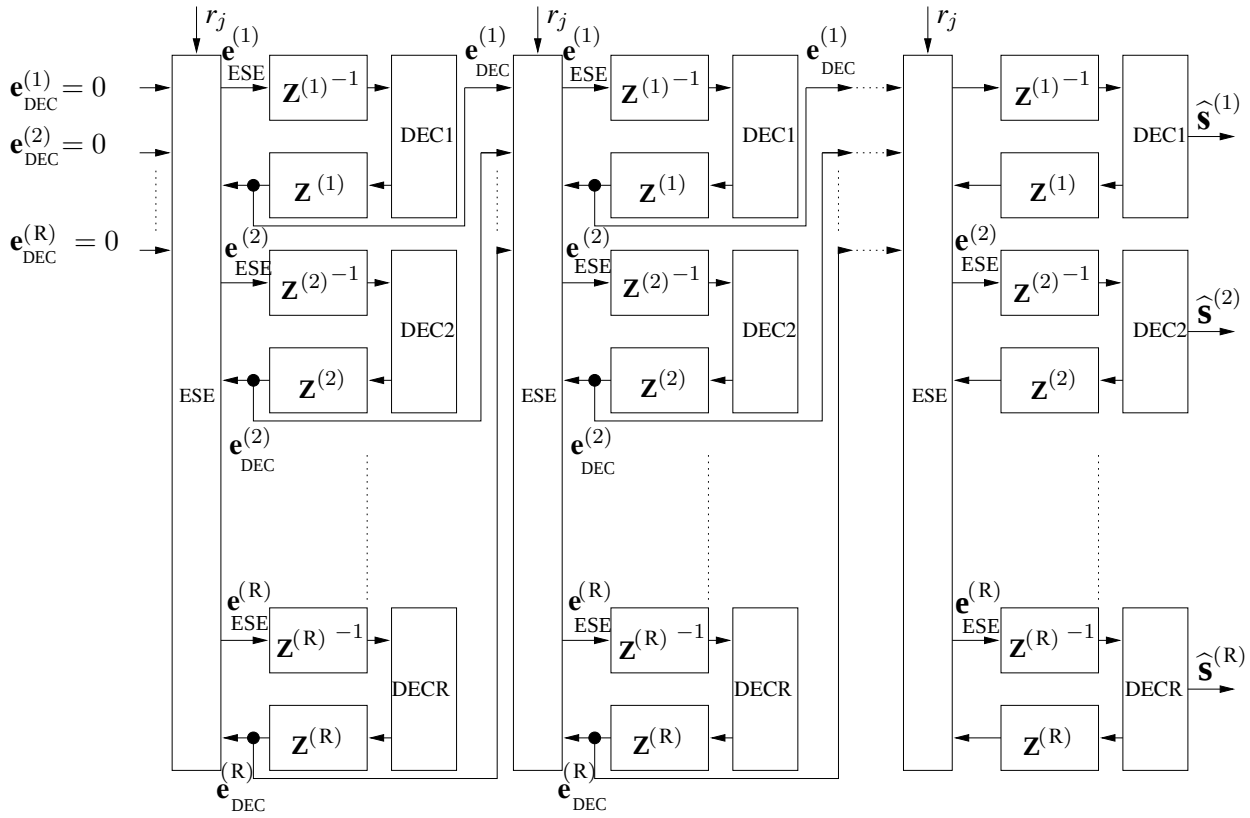

FIGURE 5.13 – Structure du récepteur IDMA utilisant un traitement série

- $J = 256$ bits, taille de l'information.
- $I = 64$ chips, correspondant à rendement global de $R_G = 1/64$;
- $K = 16384$ est le nombre de chips par bloc d'information, c'est aussi la taille de chaque entrelaceur ;
- $it = \{5, 10, 20, 30, 40, 50\}$ itérations.

La Figure 5.14 donne les performances du système IDMA en fonction du nombre d'utilisateurs $R = \{1, 16, 32, 64, 100\}$, et en adoptant à la réception un traitement série pour la récupération des données d'information émises.

Sur cette figure, la courbe du BER d'un système mono-utilisateur sur un canal AWGN, utilisant une conversion bit/symbole BPSK, est prise comme référence. Les courbes de performances du système IDMA convergent vers celles des performances mono-utilisateur pour $R = \{8, 16, 32, 64, 100\}$. Pour 64 utilisateurs (soit $R = I$), correspondant à un taux de charge $\tau_c = 100\%$, ces performances sont atteintes à partir d'un rapport signal à bruit SNR de $7dB$. Pour obtenir ces performances, cinq itérations sont nécessaires pour que les courbes de performances du système convergent vers celles des performances génie lorsque $R \leq I$. Au fur et à mesure que le nombre d'utilisateurs augmente, le problème des interférences MAI devient crucial. Plus d'itérations sont alors nécessaires afin de récupérer les données d'information émises. Dix itérations sont en effet considérées lorsque le système possède 100 utilisateurs, alors que cinq itérations suffisent quand $R = 64$.

La Figure 5.15 donne les performances du système IDMA en fonction du rendement du codeur qui est constitué juste d'un code d'étalement, $I = \{8, 16, 24, 32, 64\}$, pour un

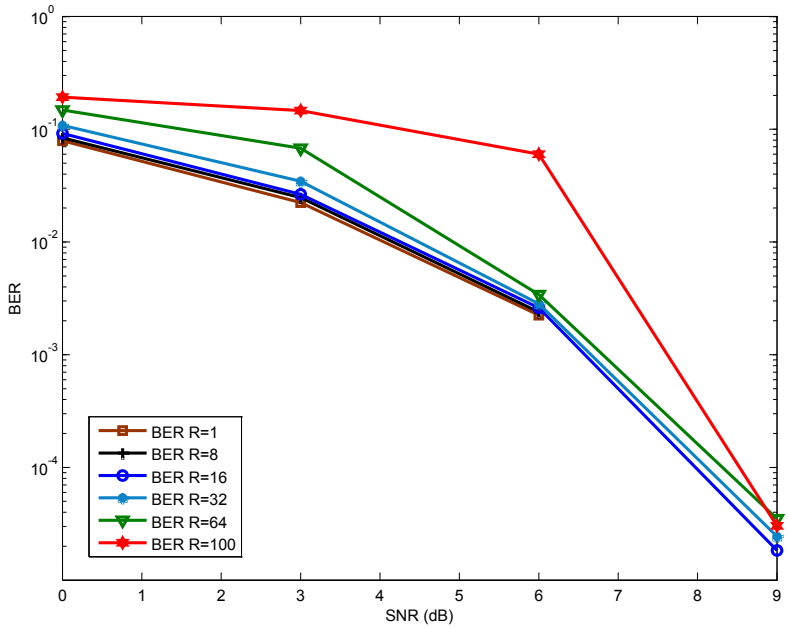

FIGURE 5.14 – Performances du système IDMA sans codage sur canal AWGN pour S = 64 et une conversion bit/symbole de type BPSK

nombre d'utilisateur qui est égal à 16 et une taille de l'information égale à 512 bits. Le résultat obtenu devient de plus en plus important lorsque la valeur du rendement $1/I$, est plus petite.

En effectuant un traitement parallèle pour la récupération des données d'information, les performances tracées sur la Figure 5.15 sont obtenues avec un nombre d'itérations significativement plus grand.

En utilisant un traitement série, la convergence est plus rapide quand $R \geq I$. Par contre, le traitement parallèle a moins de latence car les utilisateurs sont traités simultanément.

Pour conclure cette étude, il est à noter qu'avec un simple code à répétition, les performances du système IDMA s'approchent de la limite théorique du système multi-utilisateurs. Plusieurs itérations sont nécessaires quand le nombre d'utilisateurs augmente. Le nombre d'itérations varie selon le type du traitement (série ou parallèle) utilisé par le système pour la récupération des données d'information à la réception. En effet, celui-ci converge beaucoup plus vite quand un traitement série est considéré.

5.6 Le système IDMA avec récepteur aveugle

Comme présenté auparavant, le système IDMA est un cas particulier de CDMA. En effet, les utilisateurs sont distingués à l'aide d'entrelaceurs et non plus de codes orthogonaux. La technique IDMA bénéficie alors de plusieurs des avantages de la technique CDMA, notamment, des avantages de l'étalement de spectre. En effet, l'un des avantages de la technique d'étalement de spectre est le fait qu'elle soit robuste face aux différents types de brouillage. De plus, ses propriétés d'auto-corrélation permettent de tirer partie au mieux de la diversité des canaux multi-trajets à évanouissements. La principale ca-

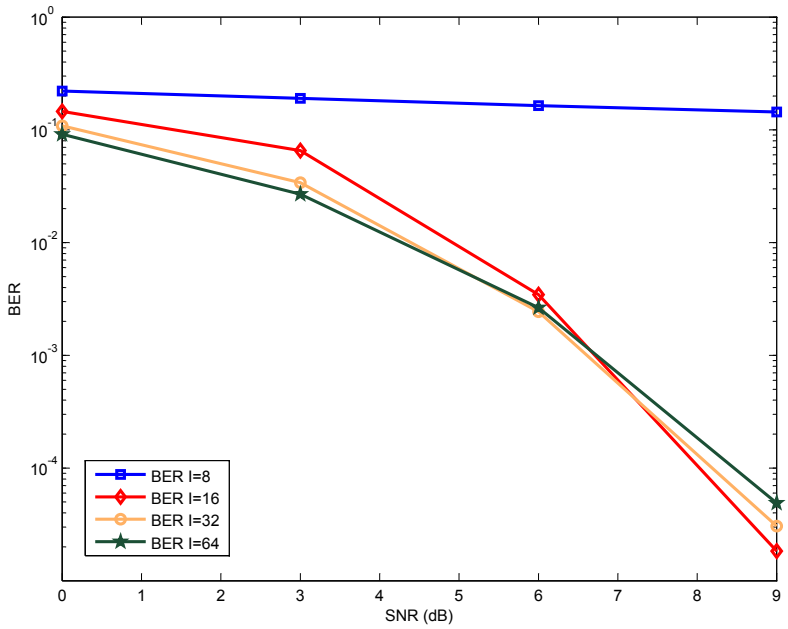

FIGURE 5.15 – Performances du système IDMA sans codage sur canal AWGN, pour K=16 et une conversion bit/symbole de type BPSK.

ractéristique de la technique IDMA est la possibilité d'utiliser pour tous les utilisateurs un même code d'étalement. Un système IDMA, constitué d'entrelaceurs générés de façon aléatoire, présente des performances proches de la limite théorique d'un système multi-utilisateurs (Ping *et al.*, 2002). C'est la raison pour laquelle la technique IDMA a tout de suite attiré l'attention de la communauté scientifique.

Dans le cas coopératif, la connaissance des entrelaceurs à la réception est prémordiale pour détecter et séparer les signaux des différents utilisateurs. Si on se place maintenant dans un cas aveugle, la détection et la séparation des signaux des différents utilisateurs doivent se faire d'une manière aveugle. C'est à dire, il faut trouver les $\mathbf{s}_r$ à partir des observations reçues $\mathbf{Y}$, sans connaissance des entrelaceurs ni de la séquence d'étalement.

Pour résoudre le problème de la détection et la sépartion aveugles des signaux IDMA reçus, nous proposons une approche algébrique multilinéaire qui permet d'effectuer ces deux opérations conjointement.

5.6.1 Modélisation tensorielle du système IDMA

Pour effectuer la séparation et l'égalisation aveugles conjointes des signaux reçus par un récepteur IDMA, nous exploiterons la structure algébrique multilinéaire de ces signaux. La Figure 5.16 synthétise le concept global de notre approche.

Dans les travaux publiés sur la technique IDMA, les entrelaceurs $\mathbf{Z}_r$ sont généralement générés aléatoirement (Ping *et al.*, 2007, 2002). Or, la question qui se pose est comment implémenter ces entrelaceurs aléatoires ? Dans la littérature, les entrelaceurs sont générés à partir d'une fonction qui génère des nombres aléatoires (Ping *et al.*, 2006). L'équation

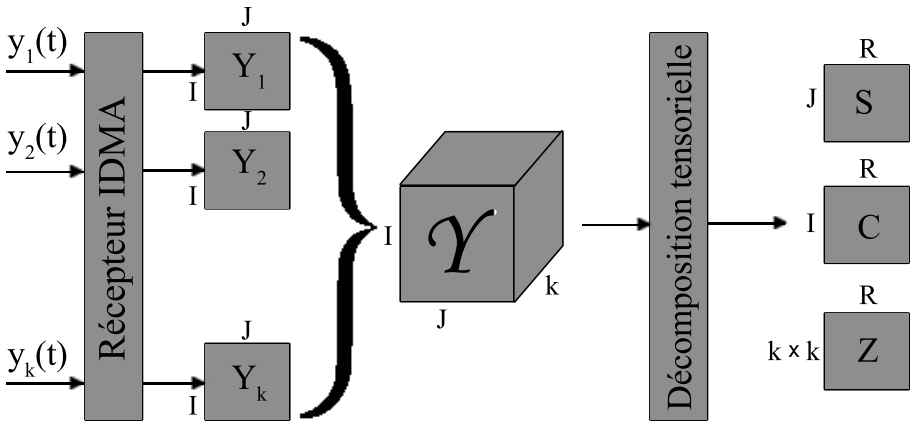

FIGURE 5.16 – Modèle de transmission en bande de base des signaux IDMA.

(5.36) donne un exemple d'entrelaceur aléatoire utilisé dans le cas coopératif.

$$\mathbf{Z} = \begin{pmatrix} 4 & 2 & 3 & 1 \\ 3 & 1 & 4 & 2 \\ 1 & 3 & 2 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \tag{5.36}$$

Dans cette étude, nous proposons d'appliquer au système IDMA des entrelaceurs sous forme de matrice de permutation. L'écriture des entrelaceurs sous cette forme nous permettra d'établir une modélisation tensorielle du système IDMA. En effet, l'utilisation des entrelaceurs aléatroires générés par une fonction, qui retourne un ensemble de nombres aléatoires comme présenté dans l'exemple (5.36) dans le système IDMA, ne permet pas d'écrire l'équation du signal transmis sous une forme tensorielle.

Notre proposition consistera alors, à utiliser des matrices de permutation, ce qui nous permettra de modéliser le signal IDMA $\mathbf{y}(t)$ sous forme d'un produit tensoriel entre les symboles envoyés par l'utilisateur r, le code d'étalement et l'entrelaceur attribué à cet utilisateur. Les entrelaceurs que nous proposons ici sont des matrices carrées de taille $K \times K$, dont la dimension K dépendra de la taille des symboles des utilisateurs ainsi que celle du code d'étalement, de tel sorte à avoir : $K = I \times J$. Chaque ligne de cette matrice contiendra un seul 1 et $K - 1$ coefficients égals à 0, et la position du 1 dans la ligne dépendra de la permutation choisie. Un exemple d'entrelaceur proposé est donné dans l'équation (5.37), où $\mathbf{d} = [d_1, d_2, \ldots, d_6]$ représente une séquence de *chips* après

étalement.

$$\begin{aligned}\mathbf{d}^T\ \mathbf{Z} &= \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{pmatrix}^T \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} d_3 \\ d_5 \\ d_4 \\ d_6 \\ d_1 \\ d_2 \end{pmatrix}^T \end{aligned} \tag{5.37}$$

L'utilisation de ces entrelaceurs permet donc de construire le tenseur des observations d'ordre 3, noté $\mathcal{Y} \in \mathbb{C}^{I\times J\times K\times K}$. Chaque dimension du tenseur des observations correspend à une diversité disponible à la réception : la première dimension correspond à la diversité de codage, la deuxième à la diversité temporelle alors que la troisième correspond à la diversité d'entrelacement. La solution que nous proposons pour résoudre le problème de séparation et d'égalisation aveugles consiste à décomposer le tenseur des observations $\mathcal{Y}$ en une somme de R contributions, où R représente ici le nombre d'utilisateurs actifs dans le système IDMA.

Considérons un système IDMA, où R utilisateurs émettent simultanément des séquences d'information de longueur J dans la même bande passante. Ces séquences sont ensuite étalées à l'aide d'un code d'étalement de longueur I. Chaque utilisateur se voit attribué de manière unique un entrelaceur de taille $K \times K$. À rappeler que dans un système IDMA, les codes d'étalement ne sont pas nécessairement orthogonaux, puisque les utilisateurs sont distingués à l'aide de différents entrelaceurs. On notera S_{jr} le $j^{\text{ème}}$ symbole d'information à transmettre par le $r^{\text{ème}}$ utilisateur, C_{ir} le $i^{\text{ème}}$ symbole de sa séquence d'étalement et $\mathbf{Z}_r$ est l'entrelaceur qui correspond au $r^{\text{ème}}$ utilisateur. Supposons que les signaux envoyés par les R utilisateurs se propagent selon un trajet unique avant d'atteindre le récepteur. Le canal est dans ce cas sans mémoire, et le mélange produit par ce dernier est linéaire et instantané. Le modèle analytique du signal en bande de base $y(t)$ reçu par le récepteur s'écrit :

$$y(t) = \sum_{r=1}^{R} \mathbf{s}_r \mathbf{c}_r \mathbf{Z}_r, \tag{5.38}$$

L'échantillon Y_{ijkl} du signal $y(t)$ reçu à l'instant d'échantillonnage $(jI+i)T_c$ s'écrit comme suit :

$$y_{ijkl} = \sum_{r=1}^{R} s_{jr} c_{ir} z_{kl}^r \quad i \in [1, I],\ j \in [1, J],\ k \in [1, K],\ l \in [1, K], \tag{5.39}$$

Le modèle analytique (5.39) n'est valable que dans le cas où il y a de l'interférence entre *chips*. Dans le cas d'interférence entre symboles (IES), il suffit de remplacer c_{ir} par le produit de convolution entre la séquence d'étalement du $r^{\text{ème}}$ utilisateur et la réponse impulsionnelle du canal correspondant, notée h_{ir}.

Le modèle tensoriel équivalent à (5.39) correspond à la décomposition CP du tenseur des observations $\mathcal{Y} \in \mathbb{C}^{I\times J\times K\times K}$, dans lequel les échantillons y_{ijkl} sont stockés. De ce fait, et dans le cas d'un canal linéaire instantané, la contribution de chaque utilisateur peut être modélisée par un tenseur de rang 1. La décomposition CP du tenseur $\mathcal{Y}$ consiste à estimer les matrices $\mathbf{S} \in \mathbb{C}^{J\times R}$, $\mathbf{C} \in \mathbb{C}^{I\times R}$ et le tenseur $\mathcal{Z} \in \mathbb{C}^{K\times K\times R}$; contenant respectivement les symboles d'information, le code d'étalement et les matrices d'entrelacement. Le modèle tensoriel équivalent à (5.39) s'écrira alors comme suit :

$$\mathcal{Y} = \sum_{r=1}^{R} \mathbf{s}_r \circ \mathbf{c}_r \circ \mathbf{Z}_r \tag{5.40}$$

Afin de simplifier la décomposition du tenseur des observations $\mathcal{Y}$, nous effectuerons une transformation du tenseur des entrelaceurs $\mathcal{Z}$ en une matrice $\mathbf{Z}$ de taille $(K^2) \times R$. De ce fait, le tenseur $\mathcal{Y} \in \mathbb{C}^{I\times J\times K\times K}$ d'ordre 4, deviendra un tenseur $\mathcal{Y} \in \mathbb{C}^{I\times J\times (K^2)}$ d'ordre 3. L'équation (5.39) deviendra alors :

$$y_{ijk} = \sum_{r=1}^{R} s_{jr} c_{ir} z_{kr} \quad i \in [1, I],\ j \in [1, J],\ k \in [1, K^2], \tag{5.41}$$

Dans la pratique, la décomposition CP d'un tenseur peut exister mais avec un rang très grand. Donc, il est préférable de s'adapter à un modèle multi-linéaire de rang inférieur fixé à l'avance, c'est à dire $R < rang(\mathcal{Y})$, afin que nous ayons à faire face à un problème d'approximation. Pour estimer les paramètres de la décomposition, et plus précisement les symboles d'information $\mathbf{S}$, nous devons minimiser la fonction de coût suivante :

$$\Upsilon(\mathbf{S}, \mathbf{C}, \mathbf{Z}, \Lambda) = \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \mathbf{s}_r \circ \mathbf{c}_r \circ \mathbf{z}_r\|_F^2. \tag{5.42}$$

où $\mathbf{s}_r$, $\mathbf{c}_r$ et $\mathbf{z}_r$ sont des vecteurs normalisés. L'ambiguité d'échelle présente dans les symboles estimés $\mathbf{s}_r$ est éliminée en normalisant chaque séquence de symbole par la norme de ce dernier, et en calculant le facteur d'échelle exacte Λ.

Pour corriger les phases des trois matrices estimées, nous allons utiliser l'indice de performance exacte que nous avons détaillé dans la section 3.3.3 :

$$\mathcal{E}(\mathcal{Y}; \mathbf{S}, \mathbf{C}, \mathbf{Z}, \Lambda) = \min_{\pi\in\Pi} \sum_{r=1}^{R} (\min_{\varphi,\psi,\chi} \{||\mathbf{s}_r - e^{\jmath\varphi}\hat{\mathbf{s}}_{\pi(r)}||^2 + ||\mathbf{c}_r - e^{\jmath\psi}\hat{\mathbf{c}}_{\pi(r)}||^2 + ||\mathbf{z}_r - e^{\jmath\chi}\hat{\mathbf{z}}_{\pi(r)}||^2\}). \tag{5.43}$$

Supposons maintenant que le signal transmis par chaque utilisateur est soumis à une propagation à trajets multiples, et arrive au récepteur via L trajets. Dans ce modèle, nous supposons que tous les utilisateurs ont le même nombre de trajets afin de simplifier la notation mathématique et la présentation du modèle. Pour l'utilisateur r, nous notons h_r la réponse impulsionnelle du canal global équivalent, résultant d'un produit de convolution entre la réponse impulsionnelle du canal effectif et la séquence d'étalement du $r^{\text{ème}}$ utilisateur. Le modèle tensoriel équivalent s'écrira comme suit :

$$\Upsilon(\mathbf{S}, \mathbf{H}, \mathbf{Z}, \Lambda) = \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \mathbf{s}_r \circ \mathbf{h}_r \circ \mathbf{z}_r\|_F^2. \tag{5.44}$$

5.6.2 Simulations et performances des récepteurs aveugles proposés

Considérons dans cette section un tenseur des observations $\mathcal{Y}$. Dans un premier temps, nous allons illustrer le comportement des différents algorithmes proposés, afin de mettre en évidence l'impact de plusieurs facteurs sur leurs performances. Dans toutes les simulations, les résultats sont obtenus à partir de 100 itérations Monte Carlo. À chaque itération et pour chaque valeur du SNR, une nouvelle réalisation du bruit est établie. On peut définir plusieurs critères d'arrêt pour les algorithmes itératifs proposés selon le cas étudié. Si le tenseur des observations $\mathcal{Y}$ est non bruité, alors le critère d'arrêt sera comme suit : $\Upsilon^{(n)} < \epsilon$. Dans le cas d'un tenseur $\mathcal{Y}$ bruité, le critère d'arrêt sera alors : $|\Upsilon^{(n)} - \Upsilon^{(n-1)}| < \epsilon$. Le critère d'arrêt que nous utiliserons dans nos simulations sera une combinaison du deuxième critère et d'un autre qui permet d'éviter les cas extrêmes, comme pour un nombre d'itérations maximal. Dans les simulations qui suivent, nous prenons : $\epsilon = 10^{-6}$.

Dans ces simulations, nous présenterons les performances des algorithmes de réception ("Algorithme 1" et "Algorithme 2" avec le nouveau critère de performance) qui estiment aveuglement les symbole $\mathbf{S}$. Nous considérons $R = 4$ utilisateurs qui communiquent simultanément sur la même bande passante. Chaque utilisateur transmet une séquence de $J = 10$ symboles consécutifs et se voit attribuer une séquence d'étalement de longueur $I = 6$. Dans ce système, les codes d'étalement peuvent être similaires puisque seuls les entrelaceurs distinguent entre les utilisateurs. Les symboles envoyés par les utilisateurs sont générés d'une manière indépendante et identiquement distribués (iid) et modulés en utilisant une modulation QPSK (*Pseudo Random Quaternary Phase Shift Keying*). Après étalement des symboles des différents utilisateurs, les entrelaceurs sont utilisés pour désordonner les symboles. Les entrelaceurs utilisés dans cette simulation sont des matrices de permutation carrées de taille $K \times K$, avec $K = I \times J$, et donc la taille de chaque matrice de permutation pour chaque utilisateur est de 60×60. Le signal de tous les utilisateurs est ensuite envoyé via un canal sans mémoire.

Dans la première simulation, nous allons montrer l'impact de l'initialisation des algorithmes proposés sur leur vitesse de convergence. Pour cela, nous avons généré des matrices aléatoires $\mathbf{S}$, $\mathbf{C}$ et $\mathbf{Z}$ qui contiennent respectivement les symboles à envoyer, les codes d'étalement et les entrelaceurs, et un tenseur $\mathcal{Y}$ bruité.

La Figure 5.17 présente les courbes de convergence des deux algorithmes proposés pour le scénario présenté auparavant. L'erreur de reconstruction du tenseur $\mathcal{Y}$ est tracée en fonction du nombre d'itérations pour diffétentes valeurs du SNR. On remarque bien que la convergence des deux algorithmes devient de plus en plus significative lorsque les valeurs du SNR augmentent.

La Figure 5.18 illustre la convergence des algorithmes "Algorithme 1" et "Algorithme 2" en fonction du nombre d'itérations. Dans cette simulation, nous avons adopté le même scénario de la première simulation, sauf le tenseur $\mathcal{Y}$ qui est non bruité. À partir de cette figure, il est bien claire que l'"Algorithme 2" converge plus rapidement que l'"Algorithme 1". En effet, l'algorithme 2 n'a besoin que de 8 itérations pour atteindre une erreur de reconstruction du tenseur des observation $\mathcal{Y}$ de 10^{-6}, alors que l'algorithme 1 atteint la même erreur pour 14 itérations.

La Figure 5.19 (b) représente les résultats obtenus pour la moins bonne initialisation.

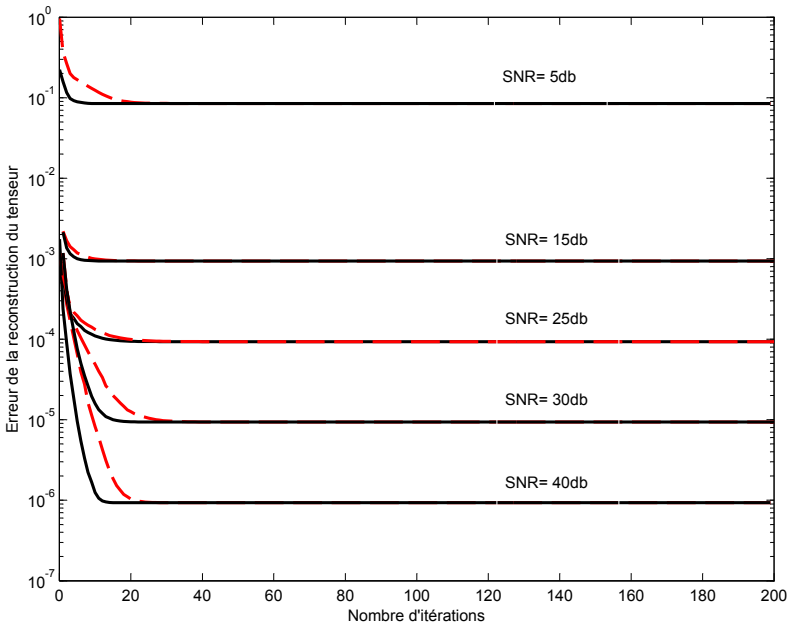

FIGURE 5.17 – Convergence des deux récepteurs aveugles proposés (“Algorithme 2” et “Algorithme 1”) en fonction du nombre d’itérations pour différentes valeurs du SNR.

Comparativement à la Figure 5.19(b), la meilleure initialisation des algorithmes sur la Figure 5.19(a) nous a permit d’obtenir une bonne estimation de la matrice des symboles en fonction du rapport signal à bruit. En effet, sur la Figure 5.19(b) la mauvaise initialisation des algorithmes “Algorithme 1” et “Algorithme 2” a mené à des matrices estimées mal conditionnées, ce qui explique la mauvaise convergence. L’impact de la mauvaise initialisation est plus remarquable sur l’“Algorithme 1”.

La figure 5.20 illustre l’évolution du BER global, moyenné sur tous les utilisateurs, en fonction du SNR pour les algorithmes aveugles “Algorithme 1”, “Algorithme 2” dans le cas d’un canal multi-trajets. Les résultats sont obtenus pour un nombre moyen d’initialisations réussies sur 15 essayes aléatoires et différentes. En effet, pour chaque itération de Monte-Carlo, la sélection de la meilleure initialisation permet d’éviter des minima locaux dans le calcul du BER. À chaque itération, les réponses du canal effectif sont générées aléatroirement et convoluées au séquences d’étalement.

À partir de la figure 5.20, on peut noter que la valeur de BER est améliorée lorsque le nombre de trajets L s’augmente de 3 à 4, ce qui indique la bonne exploitation de la diversité multi-trajets. Dans le cas coopératif, les détecteurs multi-utilisateurs itérative Turbo (*Turbotype iterative Multi-User Detection MUD*) ont été largement étudiés afin de contourner les MAI et ISI, ce qui a donné lieu à des progrès significatifs dans ce sens. Plus récemment, l’OFDM (*Orthogonal Frequency Division Multiple*) et l’IDMA (Liu *et al.*, 2006; Ping *et al.*, 2006) peuvent être combinés, donnant lieu au schéma OFDM-IDMA, ce qui permet d’éliminer les ISI à travers l’OFDM, et de supprimer les interférences MAI par l’utilisation de la technique IDMA (Ping *et al.*, 2007). Cette technique utilise une forme d’onde rectangulaire comme filtre de mise en forme. Cette configuration présente l’avantage d’une implémentation efficace de la modulation/démodulation à travers la transformée de Fourier discrète DFT (*discrete Fourier transform algorithms*). Cepen-

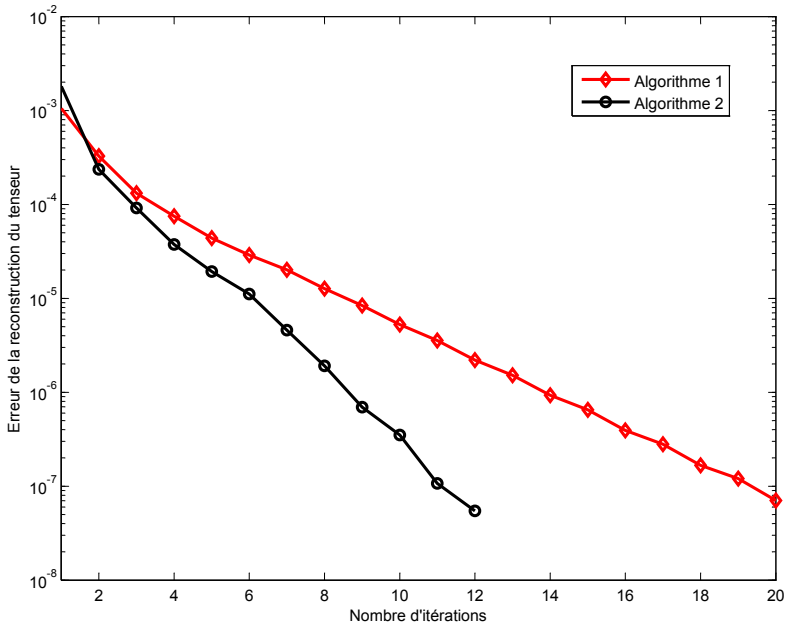

FIGURE 5.18 – Comparaison de la convergence des deux récepteurs, Algorithme 2 et Algorithme1 pour un tenseur d'observation $\mathcal{Y}$ non bruité.

dant, cette forme d'onde est non-optimale pour la transmission radioélectrique. Ceci est principalement dû à sa faible localisation fréquentielle, ainsi que sa sensibilité à la dispersion temporelle du canal de propagation. L'utilisation du préfixe cyclique permet de contourner ce problème, mais cette solution n'est pas gratuite vu qu'elle cause une perte au niveau de l'efficacité spectrale.

Parmi les autres limites de l'OFDM est la large fluctuation de l'enveloppe du signal, qui peut dégrader l'efficacité des amplificateurs de puissance des émetteurs, surtout lorsqu'ils opèrent sur des puissances moyennes assez faibles (Armstrong, 2002). Ce phénomène peut être quantifié à travers la valeur du PAPR (*Peak to Average Power Ratio*), qui résulte de la superposition d'un large nombre de composantes indépendantes, qui peuvent générer des combinaisons constructives, entraînant des pics assez élevés. La consommation d'énergie de l'amplificateur dépend principalement de la puissance de crête plus que celle moyenne. Ainsi, des pics importants provoquent une faible efficacité de puissance.

Dans la suite de ce chapitre, nous proposerons un système combinant la technique de multiplexage OWDM (*Orthogonal Wavelength-Division Multiplexing*) avec la technique IDMA, que nous avons baptisé OWDM-IDMA. Cette technique est proposée afin de faire face au problème de la forme d'onde de la modulation OFDM. En effet, à partir de la définition des ondelettes et de leurs applications dans les paquets d'ondelettes et les bancs de filtre, il est possible de construire une base orthogonale en temps/fréquence, dont les propriétés peuvent servir dans le développement d'un système de communication à porteuses multiples, utilisant le spectre disponible d'une manière plus optimale. En plus, l'utilisation de la technique IDMA peut donner au système plus de robustesse vis à vis les MAI, ce qui permet d'améliorer le rendement d'une manière significative.

Les principes de l'émetteur et du récepteur du système OWDM-IDMA seront tout d'abord décrits. Ensuite, l'étude des performances du système dans un contexte de com-

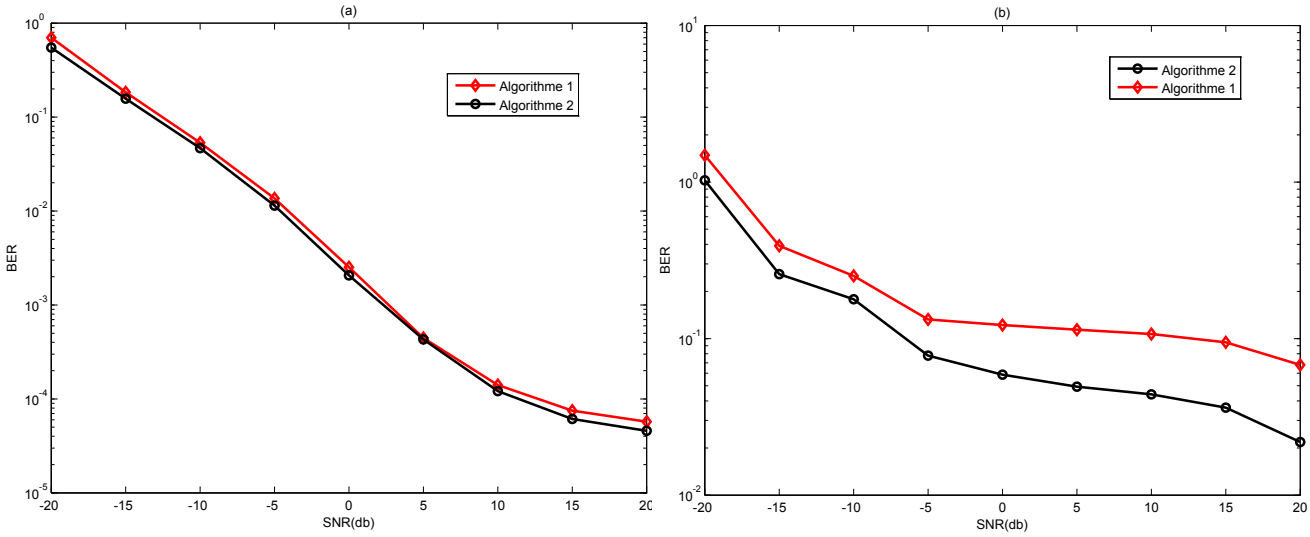

FIGURE 5.19 – L'impact de l'initialisation sur la convergence des récepteurs : (a) Meilleure initialisation (b) Mauvaise initialisation

munication sans fil seront présentées.

5.7 Le système OWDM-IDMA

Depuis les années 80, le secteur des télécommunications a connu une forte croissance grâce aux progrès technologiques réalisés dans plusieurs domaines scientifiques. Ce progrès se manifeste clairement dans le domaine des communications radio mobiles, avec l'émergence des différentes générations de la téléphonie mobile. Ainsi, d'autres applications qui peuvent bénéficier de cette technologie ont continué de se diversifier. L'OFDM est considérée comme une technique prometteuse pour les standards de communication sans fil, grâce à sa résistance envers les évanouissements causés par la propagation en trajets multiples. Ainsi, l'OFDM a été adoptée dans plusieurs standard comme le WIFI (802.11), WIMAX (802.16), ADSL, et plein d'autres.

Afin de faire face aux nouvelles exigences multimédia, transmission des vidéos haute qualité, et d'autres applications gourmandes en bande passante, la tendance est d'augmenter les débits de transmission pour répondre à ces besoins. Toutefois, la disponibilité du spectre limitée ainsi que les contraintes sur les puissances d'émission donnent lieu à un besoin en nouvelles approches de signalisation. Comme prémentionné, nous proposons une nouvelle technique hybride que nous baptisons OWDM-IDMA. Elle se présente comme une technique permettant de satisfaire les exigences présentées auparavant.

La nouvelle technique OWDM-IDMA consiste à combiner la technique de multiplexage OWDM et celle d'accès multiple IDMA. L'introduction de l'OWDM dans ce schéma offre de meilleures propriétés. La première idée d'utilisation de la transformée en ondelettes dans le secteur des communications était introduite dans les techniques des signaux multidimensionnels (Jain et Myers, 2003). La forme d'onde des paquets d'ondelettes possède de bonnes propriétés de localisation temps/fréquence(Jain et Myers, 2003). En utilisant les propriétés de localisation temporelle de ces formes d'ondes, une modulation à porteuses multiples IDMA basée sur l'utilisation des paquets d'ondelettes peut être conçue

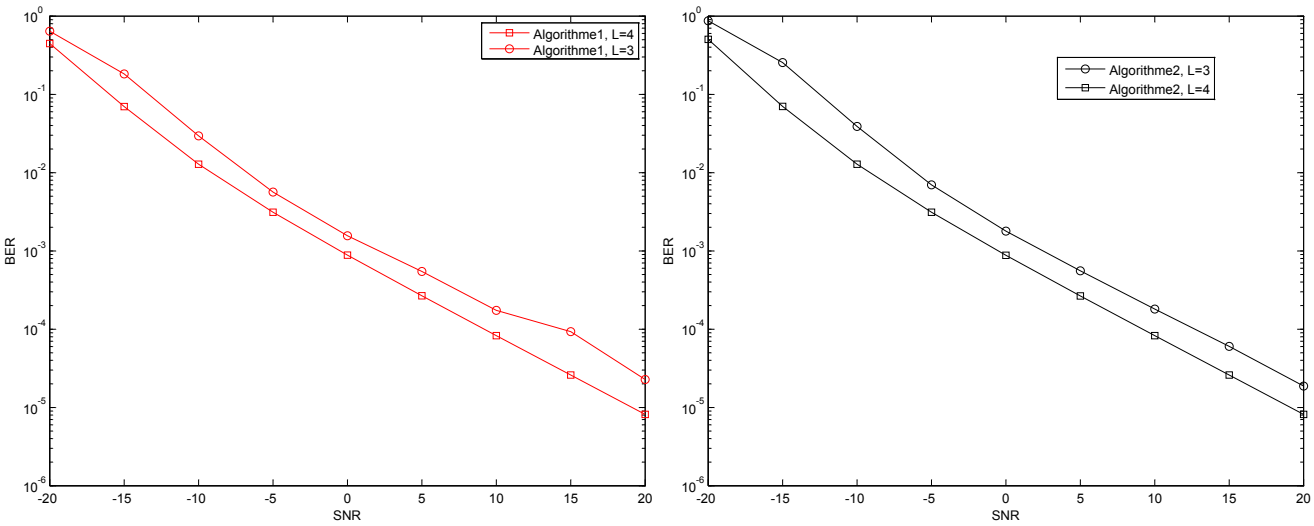

FIGURE 5.20 – L'influence de l'augmentation de nombre de trajets L sur le BER : (a) Algorithme 1 (b) Algorithme 2

afin d'offrir plus de diversité fréquentielle et temporelle. De plus, l'OWDM-IDMA peut limiter les ISI à travers la technique OWDM, et atténuer les interférences d'accès multiple grâce à l'IDMA (Rouijel *et al.*, 2010), qui utilise une technique de détection MUD itérative à faible complexité. Cette technique s'applique aussi pour les systèmes disposant d'un grand nombre d'utilisateurs.

5.7.1 La transformée en ondelettes discrète DWT

La transformée de Fourier continue est un des outils de traitement du signal les plus utilisés. Elle possède cependant certaines limites qui concernent l'analyse des signaux non-stationnaires. Le problème majeur de cette transformation est la non prise en compte de l'information véhiculée par la structure temporelle du signal, d'où la difficulté de déterminer, et plus encore de localiser les discontinuités du signal. À cause de ces limitations, de nouvelles méthodes d'analyse des signaux ont vu le jour, comme la transformée de Fourier fenêtrée ou la transformée en ondelette continue. Mais la solution réalisant le meilleur compromis possible entre la résolution temporelle et fréquentielle de l'analyse du signal, est la transformée en ondelette continue.

5.7.1.1 Des ondelettes continues aux ondelettes discrètes

A. La transformée en ondelettes continues CWT

La transformée en ondelettes continue réalise une projection sur un ensemble de fonctions appelées classiquement ondelettes. Partant d'une fonction Ψ de $L^2(\mathbb{R})$ de moyenne nulle appelée ondelette mère, la famille des fonctions analysantes $\psi_{(a,b)}$ est obtenue par action conjointe des opérateurs de dilation en échelle et de translation en temps :

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}}\Psi(\frac{t-b}{a}) \tag{5.45}$$

Dans cette expression, a est le facteur d'échelle et b le paramètre de translation. La variable a joue le rôle de l'inverse de la fréquence : plus a est petite, moins l'ondelette créée est

étendue temporellement. Le facteur multiplicatif $\frac{1}{\sqrt{a}}$ permet de conserver l'énergie lors du changement d'échelle.

L'ondelette doit être une fonction de moyenne nulle ; en d'autres termes, Ψ doit être une onde, Ce qui s'écrit mathématiquement :

$$\int_{-\infty}^{+\infty} \Psi(t)dt = 0 \tag{5.46}$$

Le choix de l'ondelette est donc en principe très ouvert, il faut cependant noter que la robustesse et la vitesse de convergence de l'algorithme de reconstruction sont très dépendantes du choix de l'ondelette.

La transformation donnée par la transformée continue est redondante, donc un échantillonnage de la transformée continue permet d'obtenir une transformation non-redondante.

B. La transformée en ondelettes continues DWT

La transformée en ondelettes discrète DWT (*Discrete Wavelet Transform*) est donc obtenue par échantillonnage des coefficients d'échelle et de temps. Pour la transformée dyadique à l'échelle 2^l, on obtient les coefficients $DWT_\Psi x(l,p)$ donnés par :

$$DWT_\Psi x(l,p) = 2^{-\frac{l}{2}} \int_{-\infty}^{+\infty} x(t)\psi(2^{-l} - p)dt \tag{5.47}$$

La transformé en ondelettes discrète DWT (*Discrete Wavelet Transform*), basée sur le codage par sous-bandes permet de minimiser le temps de calcul de la transformée en ondelettes. Elle est plus facile à implémenter, et dispose d'un temps de calcul et de ressources requises plus réduites. Dans le cas de la DWT, une représentation sur l'échelle du temps du signal numérique est obtenue en utilisant des techniques de filtrage numérique. La DWT est calculée par des filtrages successifs passe-bas/passe-haut du signal temporel discret comme sur la Figure 5.21.

Le signal est dénoté par $x[j]$, où j est un nombre entier. Le filtrage passe-bas est noté par G_0, alors que le filtrage passe-haut est H_0. L'opérateur $\downarrow 2$ est la fonction de sous-échantillonnage, qui consiste à éliminer des échantillons du signal (Gopinath et Burrus, 1991). Le nombre maximal des niveaux de filtrage dépend de la taille du signal.

La Figure 5.21 montre le processus de reconstruction du signal original à partir des coefficients d'ondelettes. D'une manière basique, la reconstruction est le processus inverse de la décomposition. G_0 et H_0 seront remplacés par les filtre de synthèse G_1 et H_1, alors que l'opérateur $\uparrow 2$ est le processus d'insertion des zéros dans le signal (sur-échantillonnage) (Lindsey, 1997).

Après étude de la théorie des ondelettes, il est possible de définir les fonctions d'ondelettes pour une modulation à porteuses multiples. On peut alors définir le principe de la modulation multiporteuses utilisant les paquets d'ondelettes.

Les principes de base de la modulation multiporteuses utilisant les paquets d'ondelettes sont illustrés sur la Figure 5.22. Du côté de l'émetteur, les symboles sont transformés du domaine des ondelettes vers le domaine temporel par le biais de la IDWPT (*Inverse Discrete Wavelet Packet Transform*) ; à la réception, le signal reçu est transformé du domaine temporel à celui des ondelettes par une DWPT (*Discrete Wavelet Packet Transform*)(Gautier et Lienard, 2007; Chan, 1994).

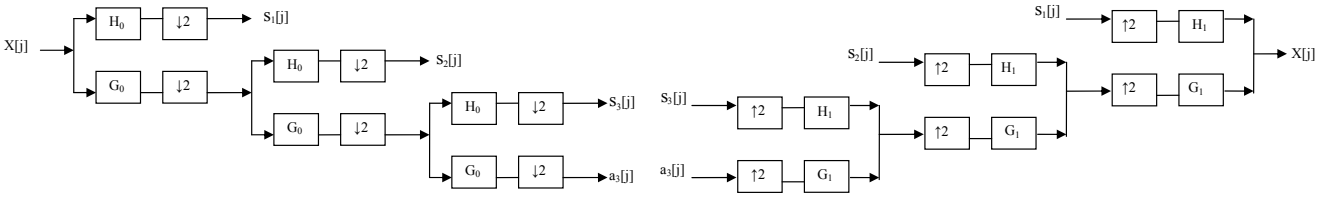

FIGURE 5.21 – la (a) décomposition et (b) reconstruction par ondelettes à 3 niveaux.

Dans la Figure 5.22, les blocs S et A représentent les blocs de décomposition et de reconstruction respectivement, et sont appelés le schéma d'Analyse/synthèse.

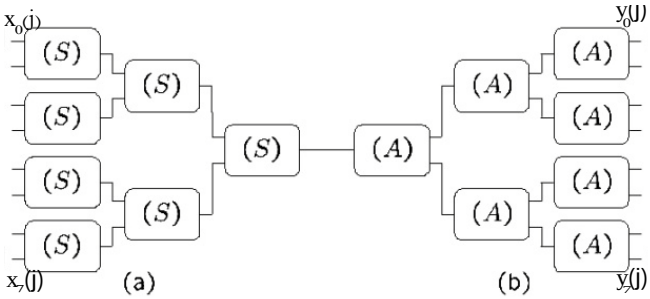

FIGURE 5.22 – Principe de la (a) modulation (IDWPT) et la (b) démodulation (DWPT) multiporteuses utilisant les ondelettes.

C. Les familles d'ondelettes

Il existe un certain nombre de fonctions de base qui peuvent être utilisées en tant qu'ondelette mère pour la transformation par ondelettes. Etant donné que l'ondelette mère produit toutes les fonctions d'ondelettes utilisés dans la transformation par translation et mise à l'échelle, elle détermine les caractéristiques de la transformée en ondelettes obtenue. Par conséquent, les détails de l'application particulière doivent être prises en compte et l'ondelette mère appropriée doit être choisie de façon à utiliser la transformée en ondelettes efficacement.

La Figure 5.23 illustre certaines fonctions des ondelettes couramment utilisées. L'ondelette de Haar est l'une des plus ancienne et la plus simple d'ondelettes. Par conséquent, toute discussion des ondelettes commence avec l'ondelette de Haar. les ondelettes de Daubechies sont les plus populaires. Elles représentent les bases du traitement du signal en ondelettes et sont utilisées dans de nombreuses applications. Elles sont également appelées ondelettes de Maxflat puisque leurs réponses en fréquence ont une planitude maximale à des fréquences 0 et π, une propriété très souhaitable dans certaines applications.

Les ondelettes de Haar, Daubechies, Symlets et Coiflets construisent des ondelettes orthogonales de support compact. Ces ondelettes avec celles de Meyer sont capables de faire une reconstruction parfaite. Les ondelettes de Meyer, Morlet et ceux de Chapeau Mexican sont symétriques en forme. Les ondelettes sont choisies en fonction de leurs formes et de leurs capacités à analyser le signal dans une application particulière.

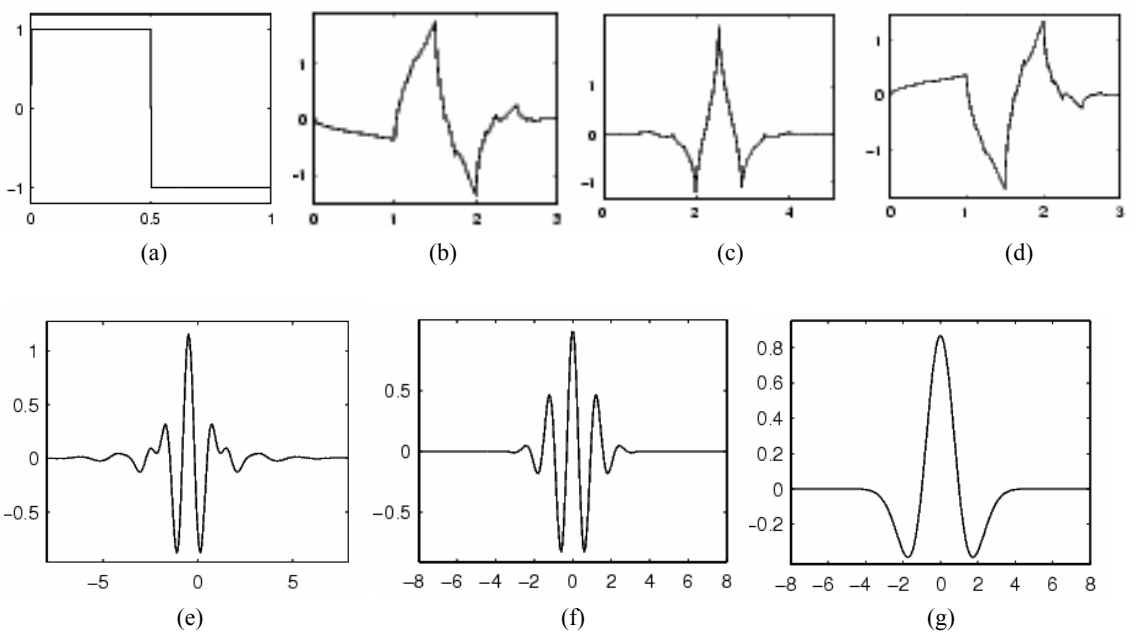

FIGURE 5.23 – Les familles d'ondelettes : (a) Haar (b) Daubechies4 (c) Coiflet1 (d) Symlet2 (e) Meyer (f) Morlet (g) Chapeau Mexican.

5.7.2 Principe de l'émetteur/récepteur du système OWDM-IDMA

La Figure 5.24 présente la structure de l'émetteur/récepteur du système OWDM-IDMA proposé pour R utilisateurs.

La structure de l'émetteur du système OWDM-IDMA ressemble à celle du système OFDM-IDMA. La différence réside dans l'application de la modulation OWDM aux symboles complexes QPSK.

Pour l'émetteur du $r^{\text{éme}}$ utilisateur, l'information est codée en une séquence d_r. La séquence étalée d_r sera ensuite entrelacée par un entrelaceur $\mathbf{Z}_r$ spécifique à l'utilisateur r. Ainsi, le signal résultant, exprimé par $\mathbf{x}_r$, est modulé en utilisant l'IDWPT afin d'avoir le signal $\mathbf{Y}_r$. Le signal reçu est égal à la somme des signaux reçus à partir des différents utilisateurs (Rouijel *et al.*, 2010, 2011b, 2012, 2011a). Le signal multi-utilisateurs reçu peut être écrit sous la forme :

$$\mathbf{y}^{''}(j) = \sum_{r=1}^{R} \mathbf{h}_r(j)\mathbf{y}_r^{'}(j) + n(j) \tag{5.48}$$

avec :

$$\mathbf{y}_r^{'}(j) = \sum_{j} \sum_{m=0}^{M-1} \mathbf{x}_r(j)\varphi_{(m,j)}(t) \tag{5.49}$$

où $\mathbf{h}_r$ sont les coefficients du canal pour le $r^{\text{éme}}$ utilisateur, $n(j)$ sont les échantillons du bruit AWGN avec une variance $\sigma^2 = N_0/2$, $\mathbf{y}_r^{'}(j)$ la séquence des *chips* du $r^{\text{ème}}$ utilisateur aprés modulation OWDM, M le nombre des sous-canaux associés à la transmission, et j représente le temps de transmission.

Comme sur la Figure 5.24, la structure du récepteur OWDM-IDMA est constituée d'un DWPT pour la démodulation OWDM, un estimateur élémentaire du signal ESE

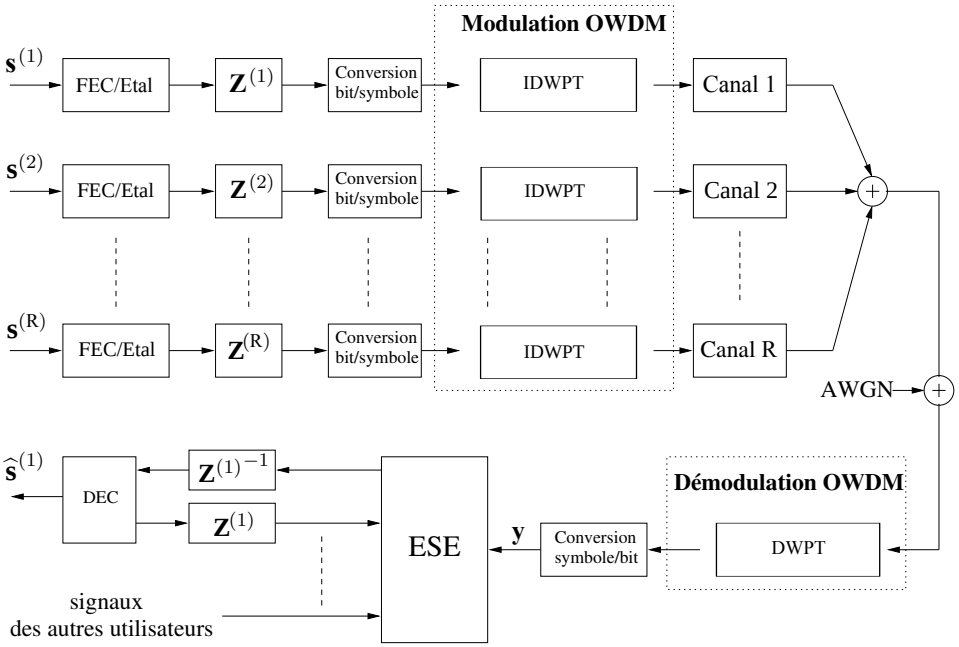

FIGURE 5.24 – Structure de l'émetteur/récepteur du système OWDM-IDMA pour R utilisateurs.

(*Elementary Signal Estimator*) et R décodeurs de probabilité a posteriori (APP-DEC). Après une démodulation OWDM par la DWPT, le processus itératif du récepteur IDMA peut être appliqué pour le signal reçu, afin d'extraire les informations relatives à chaque utilisateur :

$$\mathbf{y}(j) = \sum_{r=1}^{R} \mathbf{H}_r(j)\mathbf{x}_r(j) + n(j) \tag{5.50}$$

où $\mathbf{H}_r(j)$ fait référence au gain du canal du $j^{\text{ème}}$ *chip* du $r^{\text{ème}}$ utilisateur. L'opération d'estimation à travers l'ESE peut s'effectuer *chip-by-chip* (Ping *et al.*, 2002). Donc, le processus itératif n'inclut pas l'opération de démodulation OWDM. Si le nombre des trajets L est très grand, cette structure permet de réduire d'une manière significative la complexité de calcul du récepteur. L'information fournie au détecteur ESE est de même nature que le système multi-utilisateurs IDMA à travers un trajet unique. Ainsi, afin de retrouver la forme d'information de chaque utilisateur, le détecteur ESE tient compte juste des interférences MAI.

5.7.3 Performances du système OWDM-IDMA sur différents types de canaux

Pour vérifier les performances du système multi-utilisateur OWDM-IDMA, nous avons simulé sa chaine de transmission dans le scénario d'une liaison ascendante (*uplink*). Le nombre des ondelettes de Haar est égal à 64, et R représente le nombre des utilisateurs. Par souci de simplicité, nous considérons un système OWDM-IDMA sans codage de canal.

La Figure 5.25 montre le taux d'erreur binaire BER du système OWDM-IDMA, avec des entrelaceurs générés aléatoirement et une séquence d'étalement de longueur 16 pour tous les utilisateurs, à travers un canal AWGN, et en utilisant une modulation BPSK. Le nombre maximal des itérations est de 10, le message est d'une longueur $J = 3072$ bits

pour chaque utilisateur. On considère que les coefficients du canal sont toutes égales à $h_r = 1$, alors que le nombre des utilisateurs R varie entre $1, 4, 8, 16$ et 20. La contrainte unique qui s'impose dans la sélection de la séquence d'étalement du système OWDM-IDMA est quelle doit contenir un mélange équilibré des valeurs $+1$ et -1, afin d'assurer un caractère aléatoire des données. D'une manière simple on utilise $+1, -1, +1, -1$ pour tous les utilisateurs.

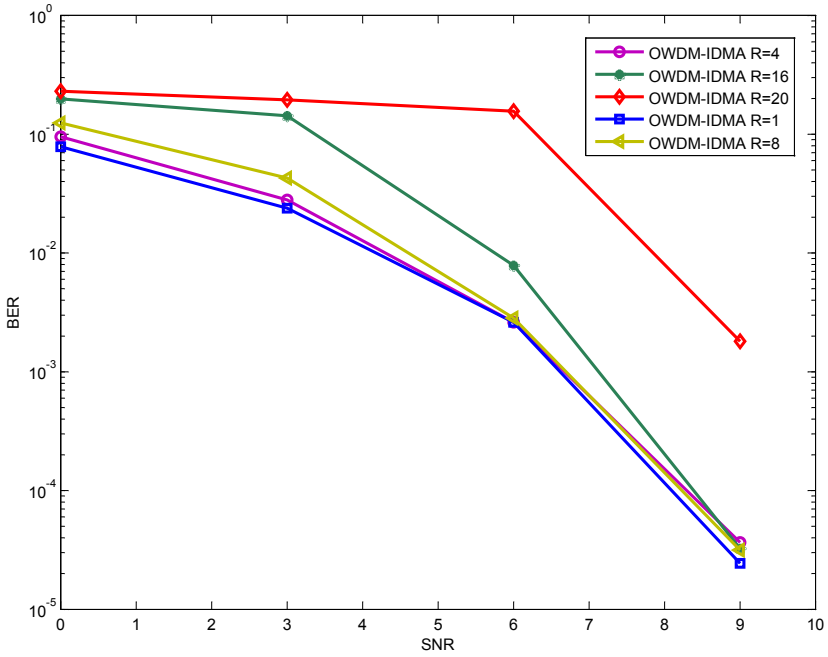

FIGURE 5.25 – Les courbes des performances simulées pour le système OWDM-IDMA.

On peut remarquer d'après les résultats de la Figure 5.25 que pour un grand nombre d'utilisateurs, on s'approche toujours des performances du système mono-utilisateur. Toutefois, les performances du système OWDM-IDMA commencent à se dégrader légèrement lorsque $R > 20$. Cette dégradation est causée par les MAI qui deviennent plus critique lorsque le nombre des utilisateurs augmente. D'un autre point de vue, l'OWDM-IDMA supporte plus d'utilisateurs que le facteur d'étalement ($R > I$), avec un taux de charge qui peut atteindre 125% pour $R = 20$ utilisateurs. Le seuil de convergence du système, avec un nombre d'utilisateurs presque deux fois plus grand que le code d'étalement, est situé à 6 dB.

La Figure 5.26 illustre la propriété de convergence du système OWDM-IDMA ci-dessus sur un canal AWGN, avec $R = 20$ et $I = 16$. Cette simulation est réalisée pour différents nombres d'itération et en fonction des valeurs du SNR. D'après les résultats de la Figure 5.26, on constate que la convergence est généralement atteinte à partir de 14 itérations (à noter que le seuil de convergence ici est fixé à une valeur de 10^{-3}).

Dans la deuxième expérimentation, nous avons étudié les performances du système OWDM-IDMA dans le cas d'un canal à trajets multiples. La Figure 5.27 montre les performances du système présenté ci-haut, en utilisant les ondelettes de Haar, un nombre d'utilisateurs $R = 4$ et des coefficients de canal $h = [0.407, 0.815, 0.407]$ (Stüber, 2001). L'égaliseur du canal utilisé ici est un Zero-Forcing, qui consiste à appliquer l'inverse de la réponse du canal au signal reçu, afin d'établir le signal d'origine (Stüber, 2001). On peut

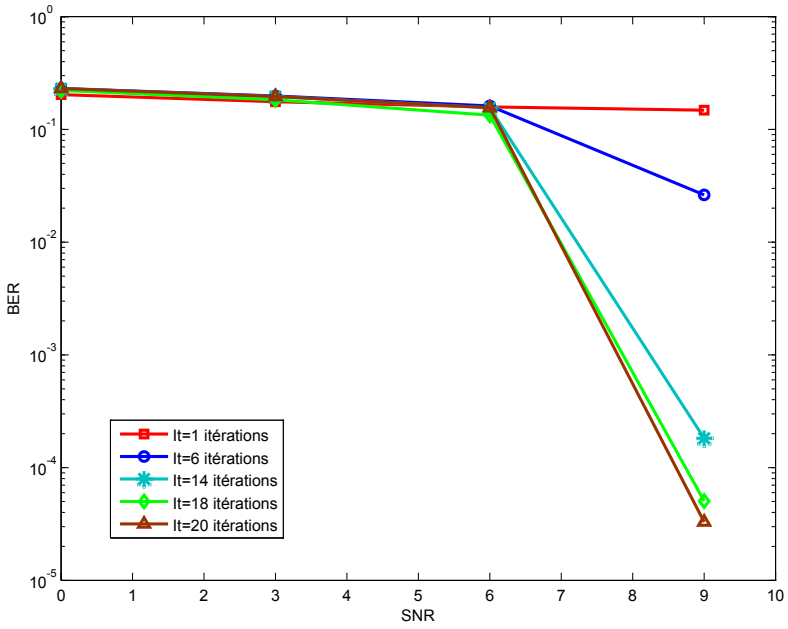

FIGURE 5.26 – Propriétés de convergence du système OWDM-IDMA dans un canal AWGN pour $R = 20$, $I = 16$, et $J = 3072$.

remarquer que les performances du système s'améliorent lorsqu'on augmente le nombre des coefficients de l'égaliseur et donc, on augmente par la même occasion la diversité du système.

Dans la troisième expérimentation, nous allons utiliser la distribution cumulative complémentaire (CCDF) de la valeur du PAPR afin d'évaluer le système proposé. Rappelons tout d'abord le principe du facteur de crête, couramment appelé PAPR. Le facteur de crête est un paramètre défini comme le rapport entre la puissance instantanée et moyenne d'un signal donné, et permettant de dimensionner ce dernier vis à vis de l'amplificateur de puissance (et vice versa). Ce paramètre apparaît dans la littérature sous plusieurs acronymes : PAPR, PMEPR (*Peak to Mean Enveloppe Power Ratio*) ou encore CF (*Crest Factor*). Devant cette multiplicité, il nous a paru nécessaire de synthétiser la notion de facteur de crête en introduisant celle du PAPR.

Un inconvénient majeur de la modulation à porteuses multiples repose sur la grande fluctuation d'enveloppe de leur signaux. La raison de ce phénomène est simple : étant donné qu'un signal à porteuses multiples est constitué d'un certain nombre de sous-porteuses modulées indépendamment, lorsque les sous-symboles pour chaque sous-porteuse sont additionnés de manière cohérente, la puissance instantanée maximale du signal de porteuses multiples pourrait être beaucoup plus grande que sa puissance moyenne (Merchan *et al.*, 1998). Typiquement, le PAPR est utilisé pour quantifier les excursions d'enveloppe des signaux multi-porteuses. Le PAPR de tout signal continu $y(t)$ est défini comme le rapport entre la puissance maximale et la puissance moyenne d'un signal (Tellado, 2000), c'est à dire :

$$PAPR(y) = \frac{\max |y_r|^2}{mean|y_r|^2} \tag{5.51}$$

où y_r est le symbole multi-porteuses discret. Un PAPR élevé se traduit directement par

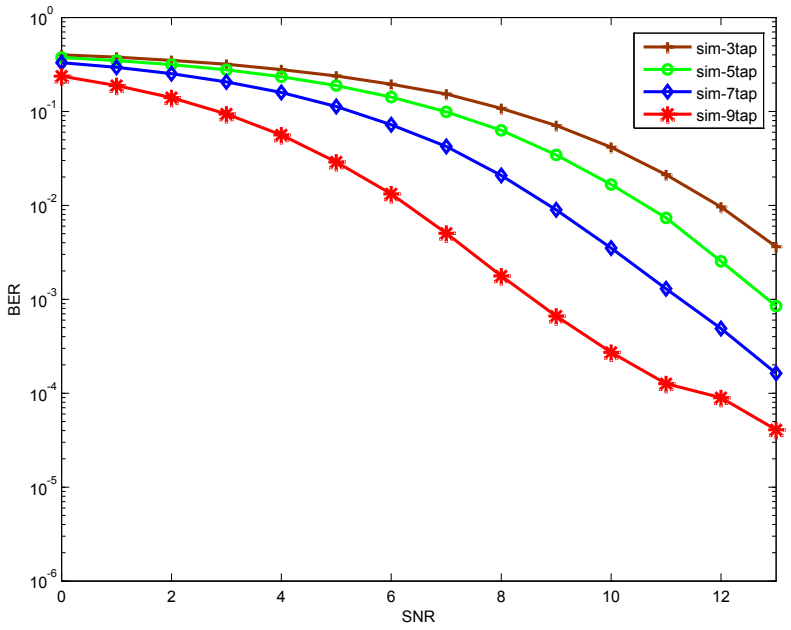

FIGURE 5.27 – Les performances du système OWDM-IDMA sur un canal AWGN à trajets multiples.

une puissance de crête élevée, qui peut dépasser l'intervalle dynamique linéaire de l'amplificateur de puissance. La distorsion non-linéaire affecte la qualité du signal. Ainsi, le récepteur peut avoir des difficultés à récupérer les données transmises. Pour contourner ce problème, un certain nombre de méthodes différentes ont été utilisées pour réduire les effets de distorsion non-linéaire par l'amplificateur. Pour que le signal ne subisse aucune distorsion par l'amplificateur, il est nécessaire que celui-ci reste dans la zone de fonctionnement linéaire, et que sa puissance maximale soit inférieure à celle correspondant au point de compression. Si le facteur de crête PAPR est élevé, il sera nécessaire de surdimensionner l'amplificateur, c'est à dire le choisir de telle sorte que la puissance de compression soit largement supérieure à la puissance moyenne du signal, le rapport entre les deux puissances étant égal au PAPR. Donc pour deux signaux de puissance moyenne égale, celui qui a le PAPR le plus élevé demandera un amplificateur avec une plus grande puissance de saturation. Mais plus on augmente la puissance de saturation de l'amplificateur, plus on augmente son coût et sa consommation énergétique. Ainsi, lorsque le signal transmis possède un facteur de crête élevé, il est souvent nécessaire de trouver un compromis entre puissance et distorsion de l'amplificateur .

La distribution cumulative complémentaire (CCDF) de la valeur du PAPR est exprimée par :

$$CCDF(PAPR_0) = Prob[PAPR > PAPR_0] \tag{5.52}$$

où $PAPR_0$ est le seuil choisi.

Dans la troisième expérimentation, les performances du système proposé sont comparés à un système OFDM-IDMA pour des symboles modulés en utilisant la constellation 4-QAM sur $M = 128$ sous-porteuses, comme indiqué sur la Figure 5.28, qui représente le CCDF de la technique proposée pour les ondelettes de Haar, Coifelet, Daubechies et

Symlet. Il est évident que le signal OFDM-IDMA possède des valeurs du PAPR qui excédent 12.4dB avec une probabilité de 0.4, contre 10 dB pour la technique proposée en utilisant l'ondelette de Haar avec la même probabilité. De même, cette valeur est égale à 9.8dB pour l'ondelette de Coiflet d'ordre 3, 11.3dB pour l'ondelette de Symlet à l'ordre 4 et 7.5dB pour l'ondelette Daubechies d'ordre 8 (Rouijel *et al.*, 2011b, 2012).

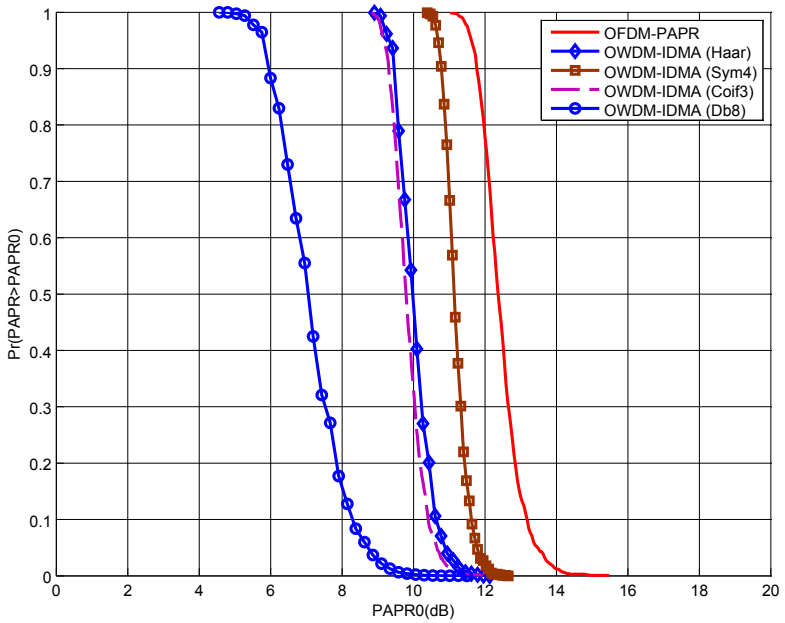

FIGURE 5.28 – Comparaison des valeurs du PAPR pour OFDM-IDMA Vs. OWDM-IDMA.

Une réduction de 4.9dB est atteinte lors de l'utilisation de l'ondelette de Daubechies d'ordre 8. Ce résultat est très intéressant en terme de réduction du PAPR. Toutefois, la complexité de cette technique est supérieure à celle de la modulation OFDM-IDMA. Cette réduction du PAPR est principalement due à la nature du signal OFDM-IDMA. Ce dernier peut être représenté sous forme d'une somme d'ondes modulées par des porteuses différentes, ce qui provoque des valeurs élevées du PAPR. Ainsi, ce signal devient plus sensible aux non-linéarités des amplificateurs à haute puissance, requises pour étaler le signal à travers le canal. En plus, les performances en terme de réduction de la valeur du PAPR sont de plus en plus meilleures lorsque l'ordre de l'ondelette augmente. Ceci vient du fait que lorsqu'on augmente l'ordre de l'ondelette, elle devient de plus en plus dispersée et présente plus de fluctuation sur l'axe des temps. On peut conclure que les performances du système OWDM-IDMA en terme de réduction du PAPR dépendent du choix de l'ondelette.

La Figure 5.29 illustre les performances en terme du BER du système OWDM-IDMA pour différentes ondelettes. Ces résultats montrent que le BER de l'ondelette de Haar est nettement meilleur que celui des autres ondelettes utilisées, et qui décroît en augmentant la valeur du SNR. Pour les trois autres ondelettes, Daubechies et Symlet d'ordre 4 possèdent la même longueur du support, qui est plus courte que celle de l'ondelette Coiflet d'ordre 4, mais leurs performances sont meilleures que ceux de Coiflet. Les propriétés de symétrie de l'ondelette de Symlet lui permettent d'avoir des performances meilleures que l'ondelette de Daubechies.

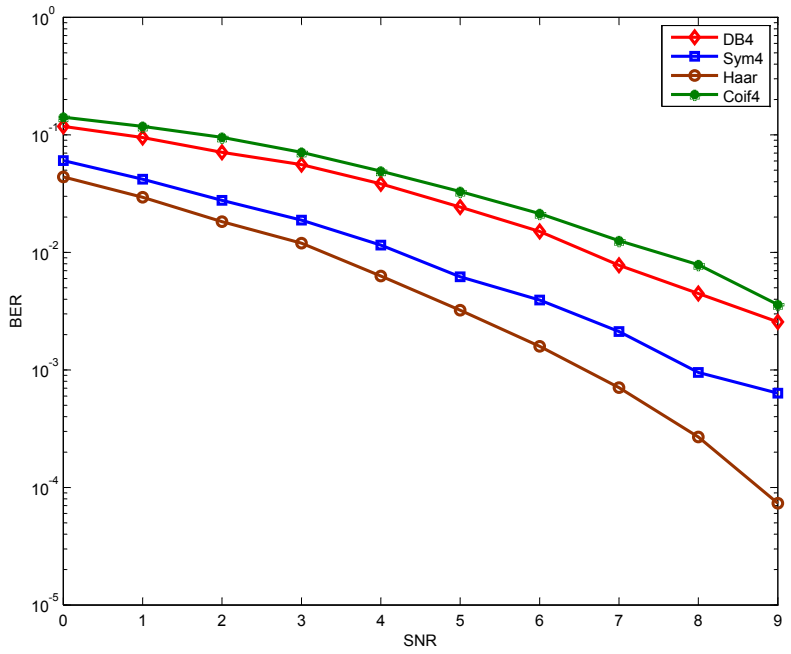

FIGURE 5.29 – Les performances en terme de BER du système OWDM-IDMA pour différentes ondelettes, sur un canal AWGN et pour R=4.

5.7.4 Modélisation tensorielle du signal OWDM-IDMA

Dans cette section, nous exploiterons la structure algébrique multilinéaire des signaux reçus par le récepteur OWDM-IDMA. Reprenons le modèle présenté en (5.48)

$$\begin{aligned} y_{ijk} &= \sum_{r=1}^{R} h_{ir} x_{jkr} \phi_{ir} \\ &= \sum_{r=1}^{R} s_{jr} c_{ir} z_{kr} h_{ir} \phi_{ir} \end{aligned} \tag{5.53}$$

Écrivons cette équation sous forme tensorielle. L'élément y_{ijk} peut être vu comme l'élément d'indice ijk d'un tenseur $\mathcal{Y}$ de taille $I \times J \times K$, et si on prend $s_{jr} = s_{jr}$, $a_{ir} = c_{ir}\phi_{ir}h_{ir}$ et $z_{kr} = z_{kr}$, l'équation (5.53) deviendra :

$$\mathcal{Y} = \sum_{r=1}^{R} \mathbf{s}_r \circ \mathbf{a}_r \circ \mathbf{z}_r \tag{5.54}$$

où $\mathbf{a}_r$, $\mathbf{z}_r$, $\mathbf{s}_r$ représentent respectivement le vecteur de séquence d'étalement convolué au canal et à l'ondelette, la séquence d'étalement d'entrelacement et le vecteur d'information de l'utilisateur r. Cette équation est une décomposition CP du tenseur $\mathcal{Y}$. Ce modèle algébrique que nous proposons pour le système OWDM-IDMA, permettra de résoudre le problème d'estimation aveugle des informations envoyées par les R utilisateurs.

En présence de bruit, on cherche souvent un modèle tensoriel approximé. Par exemple, l'erreur quadratique dans le cas d'un bruit blanc additif gaussien. Notre modèle sera alors comme suit :

$$\Upsilon(\mathbf{A}, \mathbf{S}, \mathbf{Z}, \Lambda) = \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \mathbf{s}_r \circ \mathbf{a}_r \circ \mathbf{z}_r\|_F^2 \tag{5.55}$$

où $\mathbf{s}_r$, $\mathbf{a}_r$ et $\mathbf{z}_r$ sont des vecteurs normalisés. Les ambiguïtés de mise en échelle des symboles estimés sont éliminées en normalisant chaque séquence de symboles par sa norme, et en calculant le facteur d'échelle exacte. Pour corriger les phases des trois matrices estimées, nous allons utilisé le critère de performance exacte vu dans le chapitre précédent. L'équation (5.55) deviendra alors :

$$\mathcal{E}(\mathcal{Y};\mathbf{A},\mathbf{S},\mathbf{Z},\Lambda)=\min_{\pi\in\Pi}\sum_{r=1}^{R}(\min_{\varphi,\psi,\chi}\{||\mathbf{s}_r-e^{\jmath\varphi}\hat{\mathbf{s}}_{\pi(r)}||^2+||\mathbf{a}_r-e^{\jmath\psi}\hat{\mathbf{a}}_{\pi(r)}||^2+||\mathbf{z}_r-e^{\jmath\chi}\hat{\mathbf{z}}_{\pi(r)}||^2\}). \tag{5.56}$$

5.8 Conclusion

Dans ce chapitre, nous avons présenté des méthodes de séparation de chacun des signaux CDMA, IDMA et OWDM-IDMA en utilisant les algorithmes proposés au chapitre précédent. Ces algorithmes permettent de réaliser conjointement la séparation et l'égalisation aveugles des signaux de chacune des trois applications.

Dans un premier temps, nous avons étudié le système CDMA dans le cas coopératif. Les principes de l'émetteur et le récepteur de ce système ont été détaillés et des simulations dans le cas coopératif ont été effectuées. Ensuite, Nous avons fait une étude sur la technique d'accès multiple IDMA dans le cas coopératif. Après une description détaillée du principe de l'émetteur IDMA, ainsi que celui du récepteur itératif correspondant, les performances du système IDMA ont été présentées. Ces performances obtenues par simulations pour un système multi-utilisateurs IDMA sur un canal AWGN sont proches des limites théoriques d'un système multi-utilisateurs. Cette étude a permis d'analyser la technique IDMA et les performances initialement publiées dans (Ping *et al.*, 2007). Pour combattre les interférence ISI dans le cas d'un canal multi-trajets, nous avons proposé une nouvelle technique hybride basée sur la technique IDMA, et consiste à combiner la technique IDMA avec la technique OWDM. Les principes de l'émetteur et du récepteur du système OWDM-IDMA résultant sont alors décrits, puis, les performances de ce système sont étudiées. Nous avons comparé la technique proposée à l'OFDM-IDMA et nous avons montré qu'elle permet d'obtenir une réduction significative de la valeur du PAPR pour un faible nombre de canaux. De plus, nous avons montré qu'il y a un compromis entre la complexité et la reduction du PAPR en utilisant les différentes familles d'ondelettes.

Dans un deuxième temps, nous avons proposé des récepteurs algébriques multilinéaires pour les trois systèmes CDMA, IDMA et OWDM-IDMA. Une modélisation tensorielle pour le signal reçu dans le cas de chaque système de communication sans fil a été proposée, à savoir, le système CDMA, IDMA et OWDM-IDMA. Les algorithmes pour le calcul des décompositions tensorielles proposés au chapitre précedent sont alors appliqués pour réaliser la séparation et l'égalisation des signaux du système CDMA et IDMA.

De manière général et à partir des performances obtenues, pour les deux systèmes CDMA et IDMA, par les différentes simulations, nous avons remarqué que l'algorithme 1 est plus sensible à l'initialisation que l'algorithme 2. Afin d'éviter les minima locaux vers lesquels les algorithmes proposés risquent de converger, on peut utiliser plusieurs initialisations aléatoires. C'est une solution qui permet un gain de performance considérable, au prix d'un coût calculatoire plus élevé.

Chapitre 6

Conclusion générale

Contributions

Une communication multi-utilisateurs est une transmission de données provenant de plusieurs utilisateurs sur un même canal de transmission. Les divers utilisateurs doivent alors se partager le canal de transmission. Dès lors, l'efficacité de l'occupation du canal de transmission par les différents utilisateurs s'avère primordiale. Plusieurs critères discriminants existent pour séparer les signaux provenant des différents utilisateurs. Parmi eux, la technique CDMA qui distingue entre les utilisateurs à l'aide des codes d'étalement, ainsi que la technique l'IDMA, proposée par l'équipe de Li Ping, qui discrimine les utilisateurs à l'aide des entrelaceurs.

Les travaux présentés dans ce document s'inscrivent dans le cadre des méthodes algébriques déterministes pour la séparation aveugle des signaux CDMA et IDMA reçus, dans un contexte multi-utilisateurs. Ces méthodes s'appuient sur la structure des données, qui peuvent s'écrire sous forme de tenseurs possédant une forme algébrique particulière. En effet, les problèmes de séparation de souces aveuglement sont classiquement formulés en termes d'algèbre matricielle. L'originalité de notre travail réside dans la modélisation tensorielle que nous avons proposé pour les signaux reçus du système CDMA, IDMA et la nouvelle technique proposée OWDM-IDMA. Nous nous sommes intéressé en particulier à une décomposition des tenseurs connue sous le nom de la décomposition CP. Cette décomposition propose de décomposer un tenseur de rang R en une somme de R tenseurs de rang un. D'où, la séparation des signaux consiste alors à décomposer ce tenseur d'observation en utilisant des outils d'algèbre multilinéaire, afin d'estimer la contribution de chaque utilisateur. Dans ce contexte, l'objectif de cette thèse a été de développer des méthodes qui séparent la contribution de chaque utilisateur sans que le récepteur ait des informations à priori. En effet, l'état de l'art des méthodes répondant à ces problématiques dans le cas coopératif, a montré l'existence d'une dichotomie entre le débit et les informations envoyées au récepteur pour la séparation et l'égalisations des signaux reçus. Dans ce but, nous avons proposé de considérer les données dans leur ensemble en les représentant par un tenseur d'ordre 3. Les traitements proposés sont alors basés sur l'algèbre multilinéaire et notamment sur des décompositions tensorielles.

Dans un premier temps, nous avons fourni quelques éléments fondamentaux de l'algèbre multilinéaires et des décompositions tensorielles. Ces outils mathématiques offrent une meilleure capacité de modélisation pour des données tridimensionnelles. Puis, nous avons présenté plusieurs décompositions tensorielles qui sont importantes dans le contexte

de cette thèse. Les propriété d'unicité et d'existence de ces décompositions ont été discutée.

D'un point de vu mathématique, deux contributions originales ont été proposées dans ce document :

- Une nouvelle décomposition tensorielle basée sur la décomposition CP, que nous avons baptisé "la décomposition CP avec isolement du facteur d'échelle".
- Le calcul d'un indice de performance exacte plus réaliste.

Pour la première contribution, le principe de cette nouvelle décomposition est de calculer la valeur optimale de facteur d'échelle que nous avons isolé et mis en dehors des matrices facteur dans la décomposition CP. Nous avons montré que dans la décomposition tensorielle CP, le conditionnement du calcul de la matrice de facteur d'échelle optimale Λ dépend de la cohérence via une matrice de Gram. Cela a des implications très importantes sur l'existence et l'unicité d'une décomposition CP approximative du tenseur $\mathcal{T}$. De plus, nous avons donné une condition suffisante qui prouve l'unicité de la décomposition tensorielle. L'isolement de la matrice de facteur d'échelle permet de réduire les indéterminations d'échelle en module unitaire et non pas les fixer complètement, d'où la difficulté d'estimer l'erreur d'identiffication des matrices facteur. Notre deuxième contribution consistait alors à calculer l'indice de performance exacte plus réaliste que les critères de performance utilisés dans la littérature qui sont optimistes par construction. Vu que l'indétermination a été caractérisée dans la décomposition CP par $3R$ nombres complexes de module unitaire, notre travail consistait alors à trouver la distance minimale exacte sous une contrainte angulaire, en calculant les $3R$ phases optimales.

D'un point de vu algorithmique, nous avons développé plusieurs techniques de calcul de la nouvelle décomposition. Une étude détaillée a été portée sur les aspects fondamentaux des algorithmes déterministes standard. Plus précisemment, les méthodes mathématiques (Gradient, Gradient conjugué, Levenberg Marquardt, Quasi Newton, BFGS ...), et des améliorations ont été proposées pour les rendre plus performants. L'étude de ces algorithmes mathématiques nous a montré leurs non-fonctionnement sur des problèmes d'optimisation avec contrainte et la sensibilité de quelques un au conditionnement et à l'initialisation. En vu de notre problème d'optimisation qui est sous une contrainte d'égalité imposée sur les colonnes des matrices de facteurs, il s'est avéré important de trouver des nouveaux algorithmes d'optimisation permettant de minimiser notre nouvelle fonction de coût. Deux nouvelles méthodes d'optimisation plus performantes ont été développées : "Algorithme 1" et "Algorithme 2". De plus, lorsque nous avons consideré aussi la contrainte d'inégalité sur les cohérences pour garantir l'existence et l'unicité de la nouvelle solution, nous avons développé un nouveau algorithme qui traite le problème d'optimisation avec contrainte d'inégalité, baptisé "Algorithme Barrière-GP". les deux premiers algorithmes à savoir "Algorithme 1" et "Algorithme 2" sont des algorithmes d'optimisation dédiés à la minimisation des problème avec contrainte d'égalité. Ils sont basés sur la méthode de gradient projeté tout en prenant en compte l'isolement du facteur d'échelle. Ces deux algorithmes de type Descente de gradient sont beaucoup moins sensible à l'initialisation que le ALS et offrent de meilleurs propriétés de convergence. Le troisième algorithme proposé, qui est l'algorithme Barrière-GP, a été conçu pour la résolution de notre fonction de coût sous la contrainte d'inégalité. Cet algorithme permet la transformation d'un problème

d'optimisation avec contrainte d'inégalité en un problème sans contrainte, ce qui facilite la recherche de la solution.

D'un point de vu applicatif, après avoir étudier les techniques d'accès multiple IDMA et CDMA dans un contexte coopératif, nous avons proposé une nouvelle technique d'accès multiple OWDM-IDMA qui combatte conjointement les interférences ISI et MAI. À la réception, l'extraction des contributions de chaque utilisateur à partir des signaux ou mélange reçus n'est possible que si le récepteur possède déjà des informations envoyées par l'émetteur lui permettant d'estimer le canal et de retrouver le message qui lui est destiné, ce qui engendre une perte au niveau du débit. Il s'est avéré alors intéressant d'être capable de retrouver les signaux émis sans connaissance a priori. Nous avons établi que la décomposition CP étudiée auparavant, permet de réaliser la séparation aveugle des signaux présents dans un mélange convolutif. Nous avons proposé différentes applications de cette technique en séparation de sources. La première application proposée concerne la séparation aveugle de signaux CDMA. En effet, les données reçues à chaque instant peuvent être vues comme des éléments d'un tenseur d'ordre trois. Ce tenseur s'écrit comme la somme de tenseurs de rang un, correspondant chacun à la contribution d'un seul utilisateur, et pouvant s'écrire comme le produit extérieur de trois vecteurs, le premier contenant les symboles d'information de cet utilisateur, le second contenant son code d'étalement, et le troisième contenant les coefficients du canal entre cet utilisateur et les antennes de réception.

La deuxième application proposée vise la séparation aveugle des signaux IDMA. Nous avons proposé un récepteur algébrique multilinéaire qui permet la séparation des signaux IDMA. De même que la première application, nous avons établi une modélisation des données reçues sous forme d'un tenseur d'ordre trois et qui s'écrit comme une somme de produit extérieur de trois vecteurs. Le premier vecteur contenant les symboles d'information de l'utilisateur, le deuxième contenant le produit de convolution entre le code d'étalement, qui est le même pour tous les utilisateurs, et la réponse impulsionnelle du canal correspondant, alors que le troisième contenant son entrelaceur. Enfin, la troisième application proposée était la séparation aveugle des signaux OWDM-IDMA, dont une modélisation tensorielle pour ces signaux a été conçu.

Perspectives

Dans l'immédiat, pour compléter l'étude menée dans ce manuscrit, il serait intéressant de continuer le travail commencé sur la séparation aveugles des signaux OWDM-IDMA et d'étudier ses performances dans différentes conditions. En effet, dans notre étude on s'est arreté à la modélisation tensorielle des signaux OWDM-IDMA reçus. Actuellement, nous continuons ce travail, en étudiant les performances de ce système pour différents types de canal.

De plus, nous avons considéré dans notre étude la liaison montante d'un système SIMO multi-utilisateurs, où chaque utilisateur ne dispose que d'une seule antenne émettrice. Une perspective envisageable serait alors de généraliser les modélisations proposées pour résoudre le problème de séparation aveugle dans un système MIMO multi-utilisateurs.

Finalement, même si les outils d'algèbre multilinéaire proposés trouvent des applica-

tions directes en traitement du signal pour les télécommunication, on trouve qu'ils se sont considérablement élargis depuis 1970 dans plusieurs d'autres applications. Dès lors, une perspective majeure sur le long terme serait d'appliquer les décompositions CP au système radar, dont on trouve plusieurs contraintes posées et différents problèmes à résoudre.

Annexes

ANNEXE

A

ANNEXE A

A.1 Détail de calcul de Λ optimale

Notre but est de calculer la valeur optimale de Λ qui minimise l'expression suivante :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) = \|\mathcal{T} - (\mathbf{A},\mathbf{B},\mathbf{C}).\Lambda\|_F^2. \tag{A.1}$$

En développant cette équation, elle conduit à :

$$\begin{aligned}\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) &= \|\mathcal{T}\|^2 - \sum_{ijk}\sum_{p} t^*_{ijk}\lambda_p a_{ip}b_{jp}c_{kp} - \sum_{ijk}\sum_{q} t_{ijk}\lambda^*_q a^*_{iq}b^*_{jq}c^*_{kq},\\ &+ \sum_{pq}\sum_{ijk}\lambda_p a_{ip}b_{jp}c_{kp}\lambda^*_q a^*_{iq}b^*_{jq}c^*_{kq},\\ &= \|\mathcal{T}\|^2 - \sum_{p}\lambda_p f^*_p - \sum_{q}\lambda^*_q f_q + \sum_{pq}\lambda_p\lambda^*_q G_{pq}. \end{aligned}\tag{A.2}$$

tel que $G_{pq} = \sum_{ijk} a_{ip}b_{jp}c_{kp}a^*_{iq}b^*_{jq}c^*_{kq}$ and $f_q = \sum_{ijk} t_{ijk}a^*_{iq}b^*_{jq}c^*_{kq}$.

Maintenant, et en annulant la dérivée de Υ par rapport à λ, nous pouvons trouver le système linéaire suivant :

$$\mathbf{G}\lambda = \mathbf{f}. \tag{A.3}$$

D'où :

$$\lambda = \mathbf{G}^{-1}\mathbf{f}. \tag{A.4}$$

A.2 Preuve du théorème 2

Soit $\mathcal{T} \in \mathbb{C}^{I\times J\times K}$, $\mathfrak{A} = \{\mathbf{A} \in \mathbb{C}^{I\times R} \mid \mu(\mathbf{A}) \leq \mu_{\mathbf{A}}\}$, $\mathfrak{B} = \{\mathbf{B} \in \mathbb{C}^{J\times R} \mid \mu(\mathbf{B}) \leq \mu_{\mathbf{B}}\}$, $\mathfrak{C} = \{\mathbf{C} \in \mathbb{C}^{K\times R} \mid \mu(\mathbf{C}) \leq \mu_{\mathbf{C}}\}$ et considérons la fonction du coût suivante :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) = \|\mathcal{T} - \sum_{r=1}^{R}\lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F^2. \tag{A.5}$$

Considérons aussi $\aleph = \mathbb{C}^R \times \mathfrak{A} \times \mathfrak{B} \times \mathfrak{C}$ un sous ensemble non compact de $\mathbb{C}^{R(1+I+J+K)}$, et soit la borne inférieure en question : $\eta := \inf\{\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) \mid (\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) \in \aleph\}$. Nous allons montrer que le sous-niveau de l'ensemble Υ restreint à $\aleph_\alpha$, définie comme $\aleph_\alpha = \{(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) \in \aleph \mid \Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) \leq \alpha\}$, est compact pour tout $\alpha > \eta$ et donc la borne inférieure de Υ sur $\aleph$ est atteint. L'ensemble $\aleph_\alpha = \aleph \cap \Upsilon^{-1}(-\infty,\alpha]$ est fermé puisque $\aleph$ est fermé et Υ est continue (Par la continuité de la norme). Il reste à montrer

que $\aleph_\alpha$ est borné.
Supposons le contraire, alors ils existe une séquence $((\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k)_{k=1}^{\infty} \subset \aleph$ avec $\|(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k\|_2 \to \infty$ mais $\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k \leq \alpha$ pour tout k. Il est claire que $\|(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k\|_2 \to \infty$ implique que $\|\lambda)^(k)\|_2 \to \infty$. Noter que :

$$\Upsilon(\mathbf{A},\mathbf{B},\mathbf{C},\Lambda) \geq [\|\mathcal{T}\|_F - \|\sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F]^2. \tag{A.6}$$

Nous avons :

$$\begin{aligned}
\|\sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r\|_F^2 &= \sum_{p,q=1}^{R} \lambda_p \overline{\lambda_q} \langle \mathbf{a}_p, \mathbf{a}_q\rangle \langle \mathbf{b}_p, \mathbf{b}_q\rangle \langle \mathbf{c}_p, \mathbf{c}_q\rangle \\
&\geq \sum_{p=1}^{R} |\lambda_p|^2 \|\mathbf{a}_p\|_2^2 \|\mathbf{b}_p\|_2^2 \|\mathbf{c}_p\|_2^2 - \sum_{p\neq q} |\lambda_p \overline{\lambda_q} \langle \mathbf{a}_p, \mathbf{a}_q\rangle \langle \mathbf{b}_p, \mathbf{b}_q\rangle \langle \mathbf{c}_p, \mathbf{c}_q\rangle| \\
&\geq \sum_{p=1}^{R} |\lambda_p|^2 - \mu_{\mathbf{A}}\mu_{\mathbf{B}}\mu_{\mathbf{C}} (\sum_{p\neq q} |\lambda_p \overline{\lambda_q}|) \\
&\geq \|\lambda\|_2^2 - \mu_{\mathbf{A}}\mu_{\mathbf{B}}\mu_{\mathbf{C}} \|\lambda\|_1^2 \geq (1 - R\mu_{\mathbf{A}}\mu_{\mathbf{B}}\mu_{\mathbf{C}}) \|\lambda\|_2^2
\end{aligned}$$

La dernière inégalité resulte du fait que $\|\lambda\|_1 \leq \sqrt{R}\lambda\|_2$ pour tout $\lambda \in \mathbb{C}^R$. À partir de l'hypotèse $1 - R\mu_{\mathbf{A}}\mu_{\mathbf{B}}\mu_{\mathbf{C}} > 0$ et $\|\lambda^{(k)}\|_2 \to \infty$, $\Upsilon((\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k) \to \infty$, ce qui contredit l'hypothèse $\Upsilon((\mathbf{A},\mathbf{B},\mathbf{C},\Lambda)_k) \leq \alpha$ pour tout k.

A.3 Le calcul du nouveau critère de performance

Dans cette annexe, nous expliquons de façon plus détaillée comment obtenir l'indice de performance δ, et en particulier la façon dont les phases (φ, ψ, χ) sont calculées.
Définissant $\chi = -\varphi - \psi$ $[2\pi]$, l'équation (3.30) peut être reécrire comme suit :

$$\begin{aligned}
\delta &= ||\mathbf{a}||^2 + ||\hat{\mathbf{a}}||^2 + ||\mathbf{b}||^2 + ||\hat{\mathbf{b}}||^2 + ||\mathbf{c}||^2 + ||\hat{\mathbf{c}}||^2 \\
&\quad -2\rho_{\mathbf{a}} \cos(\varphi - \alpha) - 2\rho_{\mathbf{b}} \cos(\psi - \beta) \\
&\quad -2\rho_{\mathbf{c}} \cos(\varphi + \psi + \gamma).
\end{aligned}$$

où $\mathbf{a}^H\hat{\mathbf{a}} \stackrel{\text{def}}{=} \rho_{\mathbf{a}}\, e^{\jmath\alpha}$, $\mathbf{b}^H\hat{\mathbf{b}} \stackrel{\text{def}}{=} \rho_{\mathbf{b}}\, e^{\jmath\beta}$ and $\mathbf{c}^H\hat{\mathbf{c}} \stackrel{\text{def}}{=} \rho_{\mathbf{c}}\, e^{\jmath\gamma}$. Les points stationaires sont donnés en résolvant le système trigonométrique dont les variables sont, par exemple, $x = \varphi - \alpha$ and $y = \psi - \beta$ inconnus :

$$\begin{aligned}
\rho_{\mathbf{a}} \sin x + \rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) &= 0, \\
\rho_{\mathbf{b}} \sin y + \rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) &= 0.
\end{aligned}$$

La première simplification est obtenue en notant que

$$\rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) = -\rho_{\mathbf{a}} \sin x = -\rho_{\mathbf{b}} \sin y.$$

ce qui implique

$$\sin y = \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin x.$$

Maintenant, en utilisant des identités trigonométriques, on peut réécrire la première équation du système trigonométrique

$$\rho_{\mathbf{c}} \sin(\varphi + \psi + \gamma) = \rho_{\mathbf{c}} \sin(x + y + \alpha + \beta + \gamma) = -\rho_{\mathbf{a}} \sin x.$$

comme suit :

$$\begin{aligned} \rho_{\mathbf{a}} \sin x &= -\rho_{\mathbf{c}} \big[\sin x \cos y \cos(\alpha + \beta + \gamma) \\ &\quad + \sin y \cos x \cos(\alpha + \beta + \gamma) \\ &\quad + \cos x \cos y \sin(\alpha + \beta + \gamma) \\ &\quad - \sin x \sin y \sin(\alpha + \beta + \gamma) \big]. \end{aligned}$$

En utilisant $\cos y = \sqrt{1 - \sin^2 y}$ et $\sin y = \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin x$, nous obtenons

$$\begin{aligned} \rho_{\mathbf{a}} \sin x &= -\frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin x \cos x \cos(\alpha + \beta + \gamma) \\ &\quad + \frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin^2 x \sin(\alpha + \beta + \gamma) \\ &\quad + \big[\rho_{\mathbf{c}} \sin x \sin(\alpha + \beta + \gamma) \\ &\quad - \rho_{\mathbf{c}} \cos x \sin(\alpha + \beta + \gamma) \big] \sqrt{1 - \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}} \sin^2 x}. \end{aligned}$$

L'objectif de la prochaine étape consiste à éliminer la racine carrée et de réécrire l'équation en termes de variables $\sin x$ or $\cos x$. Alors, nous allons élever au carré les deux côtés de cette équation et nous allons utilisé des identités trigonométriques telles que $\cos^2 x = \frac{1+\cos(2x)}{2}$, $\sin^2 x = \frac{1-\cos(2x)}{2}$, $\cos x \sin x = \frac{\sin(2x)}{2}$ et $\cos^2(2x) + \sin^2(2x) = 1$. Ainsi, après simplification nous obtenons :

$$\begin{aligned} &\frac{\rho_{\mathbf{b}}^2}{2} + \frac{1}{2}\big(\frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\big)^2 - \frac{\rho_{\mathbf{c}}^2}{2} + \Big[\frac{\rho_{\mathbf{b}}^2}{2} + \frac{1}{2}\big(\frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\big)^2 \\ &+ \frac{\rho_{\mathbf{c}}^2}{2} \cos^2(\alpha + \beta + \gamma) - \frac{\rho_{\mathbf{c}}^2}{2} \sin^2(\alpha + \beta + \gamma) \Big] \cos^2(2x) \\ &+ \Big[2\rho_{\mathbf{c}} \cos(\alpha + \beta + \gamma) \sin(\alpha + \beta + \gamma) \Big] \sqrt{1 - \cos^2(2x)} \\ &+ \Big[2\big(\frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}^2}{\rho_{\mathbf{b}}}\big) \cos(\alpha + \beta + \gamma) \sqrt{1 - \cos^2(2x)} \\ &- \big(\frac{\rho_{\mathbf{c}} \rho_{\mathbf{a}}^2}{\rho_{\mathbf{b}}}\big) \sin(\alpha + \beta + \gamma) \big(1 - \cos(2x)\big) \Big] \sqrt{\frac{1 - \cos(2x)}{2}} \\ &= 0. \end{aligned}$$

De la même manière que ci-dessus, nous avons élévé au carré les deux côtés de l'équation résultante deux fois pour éliminer les racines carrés. Enfin, nous obtenons une équation de degré six de la forme :

$$\begin{aligned} &c_0 + c_1 \cos(2x) + c_2 \cos^2(2x) + c_3 \cos^3(2x) \\ &+ c_4 \cos^4(2x) + c_5 \cos^5(2x) + c_6 \cos^6(2x) = 0. \end{aligned}$$

avec

$$\begin{cases} c_0 &= {c'}_1^2 - {c'}_5^2, \\ c_1 &= 2{c'}_1{c'}_4 - 2{c'}_5{c'}_7, \\ c_2 &= 2{c'}_1{c'}_3 + {c'}_4^2 - 2{c'}_5{c'}_6 - {c'}_7^2 + {c'}_5^2, \\ c_3 &= 2{c'}_1{c'}_2 + 2{c'}_3{c'}_4 - 2{c'}_6{c'}_7 + 2{c'}_5{c'}_7, \\ c_4 &= {c'}_3^2 + 2{c'}_2{c'}_4 - {c'}_6^2 + 2{c'}_5{c'}_6 + {c'}_7^2, \\ c_5 &= 2{c'}_2{c'}_3 + 2{c'}_6{c'}_7, \\ c_6 &= {c'}_2^2 + {c'}_6^2. \end{cases}$$

and

$$\begin{cases} {c'}_1 &= {c''}_1^2 + {c''}_3^2 - \frac{1}{2}{c''}_4^2 - \frac{1}{2}{c''}_5^2, \\ {c'}_2 &= -\frac{1}{2}{c''}_4^2 + \frac{1}{2}{c''}_5^2, \\ {c'}_3 &= {c''}_2^2 - {c''}_3^2 + \frac{1}{2}{c''}_4^2 - \frac{3}{2}{c''}_5^2, \\ {c'}_4 &= 2{c''}_1{c''}_2 + \frac{1}{2}{c''}_4^2 + \frac{4}{2}{c''}_5^2, \\ {c'}_5 &= -2{c''}_1{c''}_3 + {c''}_4{c''}_5, \\ {c'}_6 &= {c''}_4{c''}_5, \\ {c'}_7 &= -2{c''}_2{c''}_3 - 2{c''}_4{c''}_5. \end{cases}$$

and

$$\begin{cases} {c''}_1 &= \frac{1}{2}\rho_{\mathbf{a}}^2 + \frac{1}{2}\left(\frac{\rho_{\mathbf{c}}\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\right)^2 - \frac{1}{2}\rho_{\mathbf{c}}^2, \\ {c''}_2 &= -\frac{1}{2}\rho_{\mathbf{a}}^2 - \frac{1}{2}\left(\frac{\rho_{\mathbf{c}}\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\right)^2 + \rho_{\mathbf{c}}^2\cos^2(\alpha+\beta+\gamma) - \frac{1}{2}, \\ {c''}_3 &= -2\rho_{\mathbf{c}}^2\cos(\alpha+\beta+\gamma)\sin(\alpha+\beta+\gamma), \\ {c''}_4 &= -2\frac{\rho_{\mathbf{c}}\rho_{\mathbf{a}}^2}{\rho_{\mathbf{b}}}\cos(\alpha+\beta+\gamma), \\ {c''}_5 &= \frac{\rho_{\mathbf{c}}\rho_{\mathbf{a}}^2}{\rho_{\mathbf{b}}}\sin(\alpha+\beta+\gamma). \end{cases}$$

En résolvant l'équation de degré six, nous trouverons la variable x. En remplaçant x dans $\sin y = \frac{\rho_{\mathbf{a}}}{\rho_{\mathbf{b}}}\sin x$ permet de trouver y.

ANNEXE

B

ANNEXE B

B.1 Développement de la contrainte d'inégalité

La minimisation de la fonction objectif (4.3) sous la contrainte (4.54) soulève plusieurs problèmes, dont entre autres l'absence de différentiabilité de (4.54). C'est pourquoi plusieurs simplifications sont proposées. Au premier temps, mettant $\mu_1 = \mu_{\mathbf{A}}$, $\mu_2 = \mu_{\mathbf{B}}$ et $\mu_3 = \mu_{\mathbf{C}}$

Première approximation

La contrainte (4.54) fait apparaître la moyenne harmonique entre les μ_i, et peut se réécrire :

$$Harm(\mu) \stackrel{\text{def}}{=} \left(\frac{1}{3} \sum_{i=1}^{3} \frac{1}{\mu_i} \right)^{-1} \geq \frac{3}{2R+2} \tag{B.1}$$

Définissons les moyennes géométrique et quadratique :

$$Geom(\mu) \stackrel{\text{def}}{=} \left(\prod_{i=1}^{3} \mu_i \right)^{1/3} \text{ et } Quad(\mu) \stackrel{\text{def}}{=} \sqrt{\frac{1}{3} \sum_{i=1}^{3} \mu_i^2}$$

Alors nous savons que :

$$Min(\mu) \leq Harm(\mu) \leq Geom(\mu) \leq \frac{\sum_i \mu_i}{3} \leq Quad(\mu) \leq Max(\mu)$$

Donc, la condition (B.1) sera vérifiée si l'une des conditions suffisantes suivantes est vérifiée :

$$\prod_i \mu_i \leq \left(\frac{3}{2R+2} \right)^3 \tag{B.2}$$

$$\sum_i \mu_i \leq \frac{3}{2R+2} \tag{B.3}$$

$$\sum_i \mu_i^2 \leq 3 \left(\frac{3}{2R+2} \right)^2 \tag{B.4}$$

La contrainte (B.2) est plus proche de (B.1), mais (B.3) et (B.4) sont plus faciles à dériver. La contrainte (B.2) a été choisie dans (Lim et Comon, 2010; Comon et Lim, 2011), mais il est possible que nous nous orientions vers (B.4) pour simplifier les calculs.

Deuxième approximation

La cohérence définie en (4.54) doit vérifier une des contraintes de majoration définies dans la section B.1. Donc si un majorant de la cohérence la vérifie, ce sera une condition suffisante.

Pour cela, nous remarquons que nous avons toujours les inégalités suivantes entre les normes L^p :

$$\max_i |x_i| \leq \sqrt{\sum_i |x_i|^2} \leq \sum_i |x_i| \tag{B.5}$$

La norme L^2 étant la plus facile à dériver, nous proposons le majorant suivant pour la cohérence :

$$\mu_1^2 \stackrel{\text{def}}{=} \max_{p \neq q} |\mathbf{a}_p^T \mathbf{a}_q|^2 \leq \sum_{p<q} |\mathbf{a}_p^T \mathbf{a}_q|^2 \stackrel{\text{def}}{=} M_1 \tag{B.6}$$

On définit de manière similaire les majorants M_2 et M_3 pour μ_2^2 et μ_3^2. La contrainte pratique s'obtient en remplaçant μ_i par $\sqrt{M_i}$ dans les expressions de la section B.1. La contrainte (B.4) s'écrit donc :

$$\sum_{i=1}^{3} M_i \leq 3 \left(\frac{3}{2R+2} \right)^2$$

soit par conséquent la contrainte suivante, suffisante en pratique pour assurer (4.54) :

$$\mathcal{G} \stackrel{\text{def}}{=} \sum_{j<k} |\mathbf{a}_j^T \mathbf{a}_k|^2 + \sum_{\ell<m} |\mathbf{b}_\ell^T \mathbf{b}_m|^2 + \sum_{n<q} |\mathbf{c}_n^T \mathbf{c}_q|^2 - 3 \left(\frac{3}{2R+2} \right)^2 \leq 0 \tag{B.7}$$

ANNEXE

C

ANNEXE C

Algorithme de détection chip

L'algorithme de détection **chip by chip** peut être résumé comme suit :

Initialisation :

$$e_{DEC}(x_r(j)) = 0 \qquad \forall r, j \qquad \text{(C.1)}$$

Processus itératif :

$$\mu_r(j) = tanh\frac{e_{DEC}(x_r(j))}{2} \qquad \forall r, j \qquad \text{(C.2)}$$

$$\vartheta_r(j) = 1 - (\mu_r(j))^2 \qquad \forall r, j \qquad \text{(C.3)}$$

$$E(\mathbf{y}(j)) = \sum_{r=1}^{R} \mathbf{h}_r \mu_r(j) \qquad \forall j \qquad \text{(C.4)}$$

$$Var(\mathbf{y}(j)) = \sigma_n^2 + \sum_{r=1}^{R} (\mathbf{h}_r)^2 \vartheta_r(j) \qquad \forall j \qquad \text{(C.5)}$$

$$E(\xi_r(j)) = E(\mathbf{y}(j)) - \mathbf{h}_r \mu_r(j) \qquad \forall r, j \qquad \text{(C.6)}$$

$$Var(\xi_r(j)) = Var(\mathbf{y}(j)) - (\mathbf{h}_r)^2 \vartheta_r(j) \qquad \forall r, j \qquad \text{(C.7)}$$

$$e_{ESE}(x_r(j)) = 2\mathbf{h}_r \times \frac{\mathbf{y}(j) - E(\xi_r(j))}{Var(\xi_r(j))} \qquad \forall r, j \qquad \text{(C.8)}$$

Algorithme de detection de chip dans un traitement série

Le processus itératif correspondant à un traitement série est résumé comme suit :

Initialisation :
Mettre

- $\forall r, j$: $\mu_r(j) = (x_r(j)) = 0$, $\vartheta_r(j) = Var(x_r(j)) = 1$
- $\forall j$:

$$\begin{cases} E(\mathbf{y}(j)) = 0 \\ Var(\mathbf{y}(j)) = \sigma_n^2 + \sum_{r=1}^{R} (\mathbf{h}_r)^2 \end{cases}$$

- Itération =1, r=1

Processus itératif :

$$e_{ESE}(x_r(j)) = 2\mathbf{h}_r \times \frac{\mathbf{y}(j) - E(\mathbf{y}(j)) + \mathbf{h}_r \mu_r(j)}{Var(\xi_r(j)) - (\mathbf{h}_r)^2 \vartheta_r(j)} \quad \forall j \tag{C.9}$$

Désentrelacement des LLR avant désétalement et/ou décodage.
Mise à jour de l'information extrinsèque : l'information désétalée et/ou décodée à laquelle l'information à priori a été enlevée, est réentrelacée pour former l'information extrinsèque $e_{DEC}(x_1(j))$.

$$\mu_r(j) = \tanh \frac{e_{DEC}(x_r(j))}{2} \quad \forall j \tag{C.10}$$

$$\vartheta_r(j) = 1 - (\mu_r(j))^2 \quad \forall j \tag{C.11}$$

$$\triangle\mu_r(j) = \mu_r(j)|_{ite-courante} - \mu_r(j)|_{ite-precedante} \quad \forall j \tag{C.12}$$

$$E(\mathbf{y}(j)) = E(\mathbf{y}(j)) + \mathbf{h}_r \triangle\mu_r(j) \quad \forall j \tag{C.13}$$

$$\triangle\vartheta_r(j) = \vartheta_r(j)|_{ite-courante} - \vartheta_r(j)|_{ite-precedante} \quad \forall j \tag{C.14}$$

$$Var(\mathbf{y}(j)) = Var(\mathbf{y}(j)) + (\mathbf{h}_r)^2 \triangle\vartheta_r(j) \quad \forall j \tag{C.15}$$

Test du critère d'arrêt : $r = r + 1$

Si $r = R$, alors remettre $r = 1$.

Sinon, réitérer l'ensemble des calculs de l'équation C.9 à l'équation C.15.

Bibliographie

A. Kapteyn, H. N. et Wansbeek, T. (1986). An approach to n-mode components analysis. *Psychometrika*, 51(2):269–275.

A.Hjorungnes et D.Gesbert (2007). complex valued matrix differentiation : techniques and key results. *IEEE Transactions on Signal Processing*, 55(6):2740–2746.

Almeida, A. D., G.Favier et Mota, J. C. M. (2007). The constrained trilinear decomposition with application to mimo wireless communication systems. *In GRETSI*, Troyes.

Andersson, C. A. et Bro, R. (1998). Improving the speed of multi-way algorithms : : Part i. tucker3. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2):93 – 103.

Armstrong, J. (2002). Peak-to-average power reduction for ofdm by repeated clipping and frequency domain filtering. *Electronics Letters*, 38(5):246–247.

Bader, B. W. et Kolda, T. G. (2006). Algorithm 862 : MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F. et Moulines, E. (1997). A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444.

Berge, J. M. F. T. et Sidiropoulos, N. (2002). On uniqueness in Candecomp/Parafac,. *Psychometrika*, 67(3):399–409.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, 2nd édition.

Boyd, S. et Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

BRO, R. (1997). Parafac, tutorial and applications. *Chemometrics and Intellegent Laboratory Systems*, 38.

Bro, R. (1998). Multi-way analysis in the food industry - models, algorithms, and applications. Rapport technique, MRI, EPG and EMA, Proc ICSLP 2000, Amsterdam.

Bro, R. (2004). The problem and nature of degenerate solutions or decompositions of 3-way arrays. Rapport technique, Workshop On Tensor Decompositions, Palo Alto, California.

Bro, R. et Andersson, C. A. (1998). Improving the speed of multiway algorithms : Part ii : Compression. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2):105 – 113.

BROYDEN, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Candes, E. et Romberg, J. (2007). Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969.

Candès, E. et Terence, T. (2010). The power of convex relaxation : Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080.

Carroll, J. et Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319.

Cattell, R. B. (1944). Parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika*, 9(4):267–283.

Chan, Y. (1994). *Wavelet Basics*. Springer US.

Chevalier, P., Albera, L., Ferreol, A. et Comon, P. (2005). On the virtual array concept for higher order array processing. *IEEE Trans. Sig. Proc.*, 53(4):1254–1271.

Comon, P. (1986). Estimation multivariable complexe. *Traitement du Signal*, 3(2):97–101.

Comon, P. (1994). Independent Component Analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314. Special issue on Higher-Order Statistics. hal-00417283.

Comon, P. (2000). Tensor decompositions, state of the art and applications. *In IMA Conf. Mathematics in Signal Processing*, Warwick, UK. keynote address.

Comon, P. (2006). *Generic Properties of Symmetric Tensors*. TRICAP, Chania, Crete, Greece. invited.

Comon, P. et Berge, J. T. (2008). Generic and typical ranks of three-way arrays. *In Icassp'08*, pages 3313–3316, Las Vegas. hal-00327627.

Comon, P., Golub, G., Lim, L.-H. et Mourrain, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis Appl.*, 30(3):1254–1279. hal-00327599.

Comon, P. et Jutten, C., éditeurs (2010). *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, Oxford UK, Burlington USA. ISBN : 978-0-12-374726-6, 19 chapters, 830 pages. hal-00460653.

Comon, P. et Lim, L.-H. (2011). Sparse representations and low-rank tensor approximation. Research Report ISRN I3S//RR-2011-01-FR, I3S, Sophia-Antipolis, France.

Comon, P., Luciani, X. et Almeida, A. L. F. D. (2009). Tensor decompositions, alternating least squares and other tales. *Jour. Chemometrics*, 23:393–405. hal-00410057.

Comon, P., Minaoui, K., Rouijel, A. et Aboutajdine, D. (2013). Performance index for tensor polyadic decompositions. *In 21th EUSIPCO conference*, Marrakech, Morocco.

Davidon, W. (1991). Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17.

de Almeida, A. L., Favier, G. et Mota, J. C. M. (2007). Parafac-based unified tensor modeling for wireless communication systems with application to blind multiuser equalization. *Signal Processing*, 87(2):337 – 351. Tensor Signal Processing.

De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM J. Matrix Anal. Appl.*, 28(3): 642–666.

De Lathauwer, L., De Moor, B. et Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

de Silva, V. et Lim, L. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3): 1084–1127.

Dell'Amico, M., Merani, M. et Maffioli, F. (2002). Efficient algorithms for the assignment of ovsf codes in wideband cdma. *In Communications, 2002. ICC 2002. IEEE International Conference on*, volume 5, pages 3055–3060 vol.5.

Deneire, L. (2003). *Réseaux locaux et personnels sans fil.*

Dennis, Jr., J. et Moré, J. (1977). Quasi-newton methods, motivation and theory. *SIAM Review*, 19(1):46–89.

der Veen, A. J. V. et Paulraj, A. (1996). An analytical constant modulus algorithm. *IEEE Trans. Signal Proc*, 44:1136–1155.

Dimitri, N. (2007). *Méthodes PARAFAC généralisées pour l'extraction aveugle de sources : Application aux systèmes DS-CDMA*. Thèse de doctorat dirigée par Fijalkow, Inbar Traitement du signal et des images Cergy-Pontoise 2007.

Donoho, D. L. et Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l 1 minimization. *In Proc. Natl Acad. Sci. USA 100 2197-202.*

Eckart, C. et Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Faqihi, A. (2009). *Etude et Optimisation des Techniques MC-CDMA pour les Futures Générations des Transmissions Radio Mobiles*. Thèse de doctorat, Faculté des sciences - Université Mohammed-V, Rabat. Th. doct. : Sciences pour l'Ingénieur, Informatique et télécommunication, 2009.

Fiacco, A. V. et McCormick, G. P. (1968). *Nonlinear Programming : Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, New York.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322.

Fletcher, R. et Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2):163–168.

Fletcher, R. et Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154.

Gautier, M. et Lienard, J. (2007). Application of wavelet packet based multicarrier modulation to wireless transmissions. *ANNALES DES TELECOMMUNICATIONS-ANNALS OF TELECOMMUNICATIONS*, 62(7-8):871–893.

Gilbert, J. et Nocedal, J. (1992). Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2(1):21–42.

Godard, D. N. (1980). Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE Trans. Communications*, 28:1867–1875.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26.

Golub, G. et Loan, C. V. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Gopinath, R. A. et Burrus, C. S. (1991). Wavelets and filter banks. *In IN WAVELETS : A TUTORIAL IN THEORY AND APPLICATIONS*, pages 603–654. Academic Press.

Gribonval, R. et Nielsen, M. (2003). Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320–3325.

Hager, W. W. et Zhang, H. (2006). A survey of nonlinear conjugate gradient methods. *Pacific journal of Optimization*, 2(1):35–58.

Harshman, R. A. (1972). Determination and proof of minimum uniqueness conditions for PARAFAC1. *UCLA Working Papers in Phonetics*, 22:111–117.

Henrion, R. (1994). N-way principal component analysis theory, algorithms and applications. *Chemometrics and Intelligent Laboratory Systems*, 25(1):1–23.

HITCHCOCK, F. (1927a). The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.*, 6(1):164–189.

HITCHCOCK, F. (1927b). Multiple invariants and generalized rank of a p-way matrix or tensor. *J. Math. Phys.*, 7(1):39–79.

Jain, V. et Myers, B. (2003). Owss : a new signaling system for 100-150-mb/s wireless lans. *Wireless Communications, IEEE*, 10(4):16–24.

Kiers, H. et Mechelen, I. (2001). Three-way component analysis : Principles and illustrative application. *Psychological methods*, 6(1):84–110.

Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14(3):105–122.

Kolda, T. G. (2001). Orthogonal tensor decompositions. *SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS*, 23(1):243–255.

Kolda, T. G. et Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Kroonenberg, P. (1983a). *Three-mode Principal Component Analysis : Theory and Applications*. M & T series. DSWO Press.

Kroonenberg, P. et Leeuw, J. D. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97.

Kroonenberg, P. M. (1983b). Three-mode principal component analysis. theory and applications. *DSWO Press*, page 398.

KRUSKAL, J. B. (1977). Three-way arrays : Rank and uniqueness of trilinear decompositions. *Linear Algebra and Applications*, 18:95–138.

Kruskal, J. B. (1989). Multiway data analysis. chapitre Rank, Decomposition, and Uniqueness for 3-way and N-way Arrays, pages 7–18. Elsevier, North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands.

Kruskal, J. B., Harshman, R. A. et Lundy, M. E. (1989). Multiway data analysis. chapitre How 3-MFA Data Can Cause Degenerate Parafac Solutions, Among Other Relationships, pages 115–122. North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands.

Kusume, K. et Bauch, G. (2006). Some aspects of interleave division multiple access in ad hoc networks. *In Turbo Codes Related Topics ; 6th International ITG-Conference on Source and Channel Coding (TURBOCODING), 2006 4th International Symposium on*, pages 1–6.

Lathauwer, L. D., Comon, P. et Others (1995). Higher-order power method, application in Independent Component Analysis. *In NOLTA Conference*, volume 1, pages 91–96, Las Vegas.

Lathauwer, L. D., Moor, B. D. et Vandewalle, J. (2000). On the best rank-1 and rank-(r1,r2,. . .,rn) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342.

Leurgans, S., Ross, R. et Abel, R. (1993). A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. J. Appl. Maths.*, II(2):164–168.

Lim, L.-H. et Comon, P. (2009). Nonnegative approximations of nonnegative tensors. *Jour. Chemometrics*, 23:432–441.

Lim, L.-H. et Comon, P. (2010). Multiarray signal processing : Tensor decomposition meets compressed sensing. *Compte-Rendus Mécanique de l'Academie des Sciences*, 338(6):311–320. hal-00512271.

Lindsey, A. (1997). Wavelet packet modulation for orthogonally multiplexed communication. *Trans. Sig. Proc.*, 45(5):1336–1339.

Liu, L., Leung, W. K. et Ping, L. (2003). Simple iterative chip-by-chip multiuser detection for cdma systems. *In Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*, volume 3, pages 2157–2161 vol.3.

Liu, L., Tong, J. et Ping, L. (2006). Analysis and optimization of cdma systems with chip-level interleavers. *Selected Areas in Communications, IEEE Journal on*, 24(1):141–150.

Luenberger, D. et Ye, Y. (2008). *Linear and Nonlinear Programming*. Springer, third édition.

Lupas, R. et Verdu, S. (1989). Linear multiuser detectors for synchronous code-division multiple-access channels. *Information Theory, IEEE Transactions on*, 35(1):123–136.

M. Castella, S. Rhioui, E. M. et Pesquet, J. (2007). Quadratic higher order criteria for iterative blind separation of a mimo convolutive mixture of sources. *Signal Processing, IEEE Transactions on*, 55(1):218–232.

Madhow, U. et Honig, M. (1994). Mmse interference suppression for direct-sequence spread-spectrum cdma. *Communications, IEEE Transactions on*, 42(12):3178–3188.

Mahadevappa, R. et Proakis, J. (2002). Mitigating multiple access interference and intersymbol interference in uncoded cdma systems with chip-level interleaving. *Wireless Communications, IEEE Transactions on*, 1(4):781–792.

Mahafeno, I. (2007). *Etude de la technique d'accès multiple IDMA (Interleave Division Multiple Access)*. Thèse de doctorat, ELEC - Dépt. Electronique (Institut Mines-Télécom-Télécom Bretagne-UEB), UBS - Université de Bretagne Sud (UBS). Th. doct. : Sciences pour l'Ingénieur, UBS, Institut Mines-Télécom-Télécom Bretagne-UEB, 2007.

Mahafeno, I. M., Langlais, C. et Jego, C. (2006). Ofdm-idma versus idma with isi cancellation for quasistatic rayleigh fading multipath channels. *In Turbo Codes Related Topics ; 6th International ITG-Conference on Source and Channel Coding (TURBOCODING), 2006 4th International Symposium on*, pages 1–6.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

Merchan, S., Armada, A. et Garcia, J. (1998). Ofdm performance in amplifier nonlinearity. *Broadcasting, IEEE Transactions on*, 44(1):106–114.

Moreau, E. et de C. Luigi (2003). A least-squares approach to joint-diagonalization of tensor with application to source separation. *In Signal Processing and Its Applications. Proceedings. Seventh International Symposium on*, volume 2, pages 101–104.

Moulines, E., Duhamel, P., Cardoso, J. F. et Mayrargue, S. (1995). Subspace methods for the blind identification of multichannel fir filters. *IEEE Trans. Signal Processing*, 43:516–525.

Muti, D. (2004). *Traitement du Signal Tensoriel, application aux images en couleurs et aux signaux sismiques bruités*.

Nesterov, Y. et Nemirovsky, A. (1994). Interior-point polynomial methods in convex programming, volume 13 of studies in applied mathematics. *SIAM, Philadelphia, PA*.

Nion, D. (2007). *Méthodes PARAFAC généralisées pour l'extraction aveugle de sources. Application aux systèmes DS-CDMA*. Thèse de doctorat, Université de Cergy-Pontoise. Direction : I. Fijalkow, L. de Lathauwer.

Nocedal, J. et Wright, S. (1999). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer.

Patel, P. et Holtzman, J. (1994). Performance comparison of a ds/cdma system using a successive interference cancellation (ic) scheme and a parallel ic scheme under fading. *In Communications, 1994. ICC '94, SUPERCOMM/ICC '94, Conference Record, 'Serving Humanity Through Communications.' IEEE International Conference on*, pages 510–514 vol.1.

Ping, L., Guo, Q. et Tong, J. (2007). The ofdm-idma approach to wireless communication systems. *Wireless Communications, IEEE*, 14(3):18–24.

Ping, L., Liu, L., Wu, K. et Leung, W. K. (2006). Interleave division multiple-access. *Wireless Communications, IEEE Transactions on*, 5(4):938–947.

Ping, L., Wu, K., Liu, L. et Leung, W. (2002). A simple, unified approach to nearly optimal multiuser detection and space-time coding. *In Information Theory Workshop, 2002. Proceedings of the 2002 IEEE*, pages 53–56.

Polak, E. (1997). *Optimization : Algorithms and Consistent Approximation*. Applied Mathematical Sciences Series. Springer-Verlag.

Proakis, J. G. (1995). *Digital Communication*. 3rd edition édition.

Rajih, M. et Comon, P. (2005a). Enhanced line search : A novel method to accelerate Parafac. *In Eusipco'05*, Antalya, Turkey.

Rajih, M. et Comon, P. (2005b). Enhanced line search applied to blind channel identification : Identifiability conditions. *In IEEE SSP'05*, Bordeaux, France.

Rajih, M., Comon, P. et Harshman, R. (2008). Enhanced line search : A novel method to accelerate Parafac. *SIAM Journal on Matrix Analysis Appl.*, 30(3):1148–1171.

Rao, C. R. et Mitra, S. K. (1972). Generalized inverse of a matrix and its applications. *In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Theory of Statistics*, pages 601–620, Berkeley, Calif. University of California Press.

R.Harshman (1970). Foundations of the parafac precedure : model and conditions of an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonitics*, 16:1–84.

Roos, C., Terlaky, T. et Vial, J.-P. (1997). *Theory and algorithms for linear optimization : an interior point approach.* Wiley Chichester.

Rouijel, A., Minaoui, K., Comon, P. et Aboutajdine, D. (2014a). Cp decomposition approach to blind separation for ds-cdma system using a new performance index. *EURASIP Journal on Advances in Signal Processing*, 2014(1):128.

Rouijel, A., Minaoui, K., Comon, P. et Aboutajdine, D. (2014b). Short : Blind separation of the multiarray multisensor systems using the CP decomposition. *In Networked Systems - Second International Conference, NETYS 2014, Marrakech, Morocco, May 15-17, 2014. Revised Selected Papers*, pages 313–318.

Rouijel, A., Nsiri, B. et Aboutajdine, D. (2011a). Peak to average power ratio analysis for owdm-idma system. *In Multimedia Computing and Systems (ICMCS), 2011 International Conference on*, pages 1–4.

Rouijel, A., Nsiri, B. et Aboutajdine, D. (2011b). Peak to average power ratio analysis for owdm-idma system. *In Multimedia Computing and Systems (ICMCS), 2011 International Conference on*, pages 1–4.

Rouijel, A., Nsiri, B. et Aboutajdine, D. (2012). Performance analysis of a novel owdm-idma approach for wireless communication system. *Journal of Communications Software and Systems. JCOMSS*, 8(3).

Rouijel, A., Nsiri, B., Faqihi, A. et Aboutajdine, D. (2010). A new approach for wireless communication systems based on idma technique. *In I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on*, pages 1–4.

Royer, J.-P., Thirion-Moreau, N. et Comon, P. (2010). Computing the polyadic decomposition of nonnegative third order tensors. *Signal Processing.*

Sanchez, E. et Kowalski, B. R. (1990). Tensorial resolution : A direct trilinear decomposition. *Journal of Chemometrics*, 4(1):29–45.

Savas, B. (2003). *Analyses and Tests of Handwritten Digit Recognition Algorithms.* Linkoping University,, Sweden.

Schoeneich, H. et Hoeher, P. A. (2006). Iterative pilot-layer aided channel estimation with emphasis on interleave-division multiple access systems. *EURASIP Journal on Advances in Signal Processing*, 2006(1):081729.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sidiropoulos, N., Bro, R. et Giannakis, G. (2000a). Parallel factor analysis in sensor array processing. *In IEEE Transactions on signal processing.* VOL. 48, no.8.

Sidiropoulos, N. D. et Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14(3):229–239.

Sidiropoulos, N. D., Giannakis, G. B. et Bro, R. (2000b). Blind parafac receivers for ds-cdma systems. *IEEE Transactions on Signal Processing*, 48(3):810–823.

Smilde, A., Bro, R. et Geladi, P. (2004). *Multi-way Analysis : Applications in the Chemical Sciences.* Wiley.

Stegeman, A. et Sidiropoulos, N. D. (2007). On kruskal's uniqueness condition for the candecomp/parafac decomposition. *Linear Algebra and its Applications*, 420(2-3):540 – 552.

Stüber, G. (2001). *Principles of Mobile Communication.* Kluwer Academic.

Tellado, J. (2000). *Multicarrier Modulation with Low PAR : Applications to DSL and Wireless.* Kluwer Academic Publishers, Norwell, MA, USA.

ten Berge, J. M. F. et Smilde, A. K. (2002). Non-triviality and identification of a constrained tucker3 analysis. *Journal of Chemometrics*, 16(12):609–612.

Tucker, L. (1964). The extension of factor analysis to three-dimensional matrices. *H. Gulliksen, N. Frederiksen (Eds.), Contributions to Mathematical Psychology, Holt*, pages 109–127.

Tucker, L. (1966a). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Tucker, L. (1966b). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. pages 122–137. Madison : University of Wisconsin Press.

Wang, H. et Ahuja, N. (2003). Facial expression decomposition. *In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 958–965. IEEE.

Wang, P., Xiao, J. et Ping, L. (2006). Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems. *IEEE Vehic. Tech. Mag on*, 1(3):4–11.

Zarzoso, V. et Comon, P. (2006). Alphabet-based deflation for blind source extraction in underdetermined mixtures. *In Proc. ICA Research Network International Workshop*, pages 21–24, Liverpool, UK.

Zhou, S., Li, Y., Zhao, M., Xu, X., Wang, J. et Yao, Y. (2005). Novel techniques to improve downlink multiple access capacity for beyond 3g. *Communications Magazine, IEEE*, 43(1):61–69.