Géométrie et optimisation riemannienne pour la diagonalisation conjointe : application à la séparation de sources d'électroencéphalogrammes

Florent Bouchard

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **CIA – ingénierie de la cognition, de l'interaction, de l'apprentissage et de la création**

Arrêté ministériel : 25 mai 2016

Présentée par
Florent BOUCHARD

Thèse dirigée par **Marco CONGEDO**, CNRS, GIPSA-lab
et codirigée par **Jérôme MALICK**, CNRS, LJK
préparée au sein du **GIPSA-lab**
dans **l'école doctorale ISCE**

Géométrie et optimisation riemannienne pour la diagonalisation conjointe : application à la séparation de sources d'électroencéphalogrammes

Thèse soutenue publiquement le **22 novembre 2018**,
devant le jury composé de :

Pierre-Antoine ABSIL
Université catholique de Louvain, Rapporteur
Antoine SOULOUMIAC
CEA de Saclay, Rapporteur
Christian JUTTEN
UGA, GIPSA-lab, Examinateur, Président du jury
Jean-François CARDOSO
CNRS, Institut d'astrophysique de Paris, Examinateur
Salem SAID
CNRS, IMS Bordeaux, Examinateur
Marco CONGEDO
CNRS, GIPSA-lab, Directeur de thèse
Jérôme MALICK
CNRS, LJK, co-Directeur de thèse

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE ISCE

T H È S E

pour obtenir le titre de

docteur de la Communauté Université de Grenoble Alpes

Spécialité : CIA – ingénierie de la cognition, de l'interaction, de l'apprentissage et de la création

Présentée et soutenue par

Florent BOUCHARD

Géométrie et optimisation riemannienne pour la diagonalisation conjointe : application à la séparation de sources d'électroencéphalogrammes

Thèse dirigée par Marco CONGEDO et Jérôme MALICK

préparée au GIPSA-lab

soutenue le 22 novembre 2018

Jury :

Rapporteurs :	Pierre-Antoine ABSIL	-	Université catholique de Louvain
	Antoine SOULOUMIAC	-	CEA de Saclay
Directeur :	Marco CONGEDO	-	CNRS, GIPSA-lab
co-Directeur :	Jérôme MALICK	-	CNRS, LJK
Président :	Christian JUTTEN	-	UGA, GIPSA-lab
Examinateurs :	Jean-François CARDOSO	-	CNRS, Institut d'astrophysique de Paris
	Salem SAID	-	CNRS, IMS Bordeaux

Remerciements

Je tiens d'abord à remercier mes directeurs de thèse Marco Congedo et Jérôme Malick. Vous m'avez fourni des conditions de travail formidables et un accompagnement idéal tout au long de cette thèse. Marco, depuis mon stage de fin d'études d'école d'ingénieur, tu m'as fait découvrir le monde de la recherche et tu m'as notamment appris à écrire convenablement. Jérôme, même si tu as été moins présent que Marco, tu m'as beaucoup apporté et un bon nombre d'idées dans ce travail trouvent leur origine dans des discussions que nous avons eu. Ce fut un grand plaisir de travailler avec vous et j'espère que nous continuerons à collaborer ensemble.

Je veux également remercier Christian Jutten pour m'avoir donné l'opportunité de faire mon stage de fin d'études au Gipsa-lab, qui m'a donné envie de continuer dans la recherche et faire un doctorat. Je souhaite aussi te remercier de m'avoir associé au projet Chess et d'avoir par exemple financé mon voyage à Budapest. Enfin, je te remercie d'avoir accepté d'être président de mon jury. Merci aussi aux autres membres de mon jury, c'est à dire Pierre-Antoine Absil, Antoine Souloumiac, Jean-François Cardoso et Salem Said. Vos rapports et les discussions que nous avons eu sur ces travaux ont significativement contribués à faire évoluer ma vision du sujet et améliorer la version finale de ce manuscrit.

Je remercie de plus les chercheurs que j'ai rencontré au cours de ma thèse et avec qui j'ai pu collaborer, en particulier Bijan Afsari, Guillaume Ginohlac, Arnaud Breloy, Alexandre Renaux et Romain Couillet. C'était un plaisir d'élargir un peu mes horizons de recherche avec vous et de découvrir d'autres domaines. Un gros clin d'œil aux autres étudiants de Marco avec qui j'ai échangé et travaillé, notamment Louis Korczowski, Pedro Rodrigues et Grégoire Cattan. Merci également aux autres personnes de mon entourage avec qui j'ai travaillé, en particulier Victor Maurandi et Octave Curmi.

Je remercie mes parents pour leur soutien et les corrections ortographiques qu'ils ont faites sur ce manuscrit (j'ai effectué des modifications dans cette version finale et des erreurs se sont certainement glissées). Enfin, cette thèse concluant mes études, je souhaite remercier les quelques professeurs qui m'ont marqués comme Frédéric Normandin et Laure Pauliat.

Pour conclure ces remerciements, merci à mes proches, qui se reconnaîtrons, pour leur soutien et tous les moment qu'on a partagé et qui m'ont permis de me rafraîchir les idées.

Table des matières

Introduction

Le cerveau et son fonctionnement restent relativement peu connus et sont parmi les grands défis de la recherche actuelle. Les avancées dans ce domaine sont très dépendantes de notre capacité à observer l'ensemble des phénomènes issus du cerveau, à les localiser, à en comprendre l'origine et à en mesurer la dynamique. La recherche sur le cerveau et son fonctionnement est donc intrinsèquement liée aux avancées méthodologiques, notamment pour la mesure et l'analyse de l'activité cérébrale. Différents moyens d'observation de l'activité cérébrale, appelés modalités de neuroimagerie, ont été développés. La plus précise pour enregistrer les signaux électriques générés par le cerveau est l'électroencéphalographie intracranienne (ECoG) où des électrodes sont implantées directement dans le cerveau avec pour limitation d'être très invasive et de ne pouvoir observer que l'activité locale. Deux modalités permettent de mesurer l'activité électrique du cerveau de manière non-invasive et globale : l'électroencéphalographie (EEG) et la magnétoencéphalographie (MEG). L'électroencéphalographie consiste à enregistrer l'activité électrique à la surface du cuir chevelu avec des électrodes. La magnétoencéphalographie mesure les champs magnétiques à la surface du cuir chevelu par le biais de magnétomètres. Ces deux modalités de neuroimagerie ont une bonne résolution temporelle qui permet de bien capturer la dynamique des phénomènes observés, mais elles ont un mauvais rapport signal sur bruit et une mauvaise résolution spatiale, ce qui empêche de les localiser précisémment. D'autres modalités, comme par exemple l'imagerie à résonance magnétique fonctionnelle (IRMf) ou la spectroscopie proche infrarouge fonctionnelle (fNIRS), estiment l'activité cérébrale de façon indirecte en mesurant la quantité d'oxygène consommée par chaque partie du cerveau. Ces modalités localisent précisémment les différents phénomènes au niveau spatial, mais ils ont une mauvaise résolution temporelle et ne permettent pas de bien caractériser leur dynamique.

Dans cette thèse, nous considérons des problématiques reliées à l'électroencéphalographie. Cette technologie, proposée pour la première fois dans [17], est simple à utiliser, peu coûteuse, peu encombrante et a une haute résolution temporelle. Elle a plusieurs applications, notamment dans le diagnostic et le traitement de pathologies, et dans les interfaces cerveau-ordinateur, où l'activité cérébrale est utilisée pour intéragir avec un ordinateur. Dans l'ensemble de ces cas, on est limité par le rapport signal sur bruit et par la résolution spatiale des enregistrements. En effet, ces deux limitations rendent difficile l'analyse et la classification des phénomènes observés, comme expliqué par exemple dans [40]. Pour remédier à ces problèmes, des méthodes d'analyse des signaux électroencéphalographiques sont requises. Une méthode qui permet en théorie de retrouver les signaux générés par le cerveau à partir de ceux enregistrés par les électrodes est la séparation aveugle de sources [39], utilisée pour l'électroencéphalographie par exemple dans [40, 41]. Cette technique exploite les statistiques des observations pour retrouver les sources qui en sont à l'origine, en faisant l'hypothèse que celles-ci sont statistiquement indépendantes. Nous nous concentrons sur une méthode de résolution du problème de séparation aveugle de sources : la diagonalisation conjointe approximée de matrices symétriques positives définies contenant les statistiques des observations, méthode introduite dans [33, 51]. Mathématiquement, la diagonalisation conjointe approximée se formule comme un problème d'optimisation sur l'ensemble des matrices inversibles avec trois composantes :

le choix du critère à minimiser, la contrainte imposée pour éviter les solutions dégénérées et l'algorithme de résolution, qui est systématiquement itératif. Les approches existantes considèrent principalement deux critères : le critère des moindres carrés introduit dans [33, 34], venant du raisonnement pratique qu'il faut annuler les termes hors-diagonaux, et le critère log-vraissemblance, obtenu à partir d'un modèle théorique de la séparation aveugle de sources dans [95]. Les méthodes précédemment proposées sont de plus spécifiques à une contrainte de non-dégénérescence et se restreignent aussi à un seul type d'algorithme de résolution. De ce fait, il est difficile de modifier l'une des trois composantes et d'estimer leur influence lorsqu'on compare plusieurs méthodes.

Cette thèse s'inscrit en traitement du signal et mathématiques appliquées avec des applications pour l'ananlyse d'enregistrements électroencéphalographiques. Nous adoptons une approche exploitant des outils de géométrie et d'optimisation riemannienne, qui unifient des approches émergeant de plusieurs domaines tels que la physique quantique, la théorie de l'information et le transport optimal. Le travail effectué peut se découper selon deux axes. Le premier thème de recherche porte sur la modélisation du problème de diagonalisation conjointe approximée d'un ensemble de matrices symétrique positives définies par une approche géométrique. Dans notre formulation, la diagonalisation conjointe approximée est conçue comme un problème de minimisation d'un critère mesurant la positon relative des matrices à diagonaliser par rapport à l'ensemble des matrices diagonales positives définies. En particulier, on peut utiliser des divergences, qui généralisent la notion de distance. Cette approche prend en compte les critères usuels (moindres carrés et log-vraissemblance). Elle permet aussi de définir des critères qui n'ont pas été étudiés, notamment basés sur la divergence de Kullback-Leibler, la divergence log-det α, la distance riemannienne naturelle, la distance log-euclidienne et la distance de Wasserstein. Notre second sujet de recherche concerne le développement d'outils d'optimisation permettant de définir des méthodes modulaires où l'on peut faire varier les trois composantes du problème de diagonalisation conjointe indépendamment. Cet objectif nous conduit à employer l'optimisation riemannienne [2], particulièrement attractive car elle permet de prendre en compte n'importe quel critère, inclut naturellement les contraintes et offre un large panel d'algorithmes de résolution génériques. Nous considérons trois géométries différentes pour transformer l'ensemble des matrices inversibles en variété riemannienne et nous traitons l'ensemble des contraintes précédemment utilisées pour la diagonalisation conjointe.

Après cette introduction, ce manuscrit comporte quatre chapitres, des conclusions et perspectives, ainsi que quatre annexes. Dans le chapitre 1, nous détaillons le contexte du sujet de cette thèse et nous faisons une revue bibliographique. Nous y présentons l'électroencéphalographie, la séparation aveugle de sources, la diagonalisation conjointe approximée, la géométrie et l'optimisation riemannienne, et les variétés que nous utilisons dans la suite. Dans le chapitre 2, nous établissons notre modèle géométrique pour la diagonalisation conjointe approximée. Après avoir défini le modèle des critères, nous proposons trois stratégies d'optimisation, puis nous présentons les divergences sur lesquelles les critères sont construits. Nous étudions aussi les propriétés souhaitables pour la diagonalisation conjointe. Dans le chapitre 3, nous proposons un cadre d'optimisation riemannienne générique pour la diagonalisation conjointe, où nous considérons trois géométries différentes pour l'ensemble des matrices inversibles. Pour

deux de ces géométries, nous considérons cette variété directement, comme décrit dans le chapitre 1 ; pour la troisième, nous proposons d'exploiter la décomposition polaire et nous construisons la variété produit correspondante. Nous traitons les trois contraintes utilisées pour la diagonalisation conjointe en distinguant celles qui donnent des sous-variétés de celle aboutissant sur une variété quotient. Dans le chapitre 4, nous illustrons les résultats théoriques des deux chapitres précédents par des expériences numériques sur des données simulées et sur des enregistrements électroencéphalographiques. Ces expériences montrent que notre approche par optimisation riemannienne est viable et donne des résultats compétitifs par rapport aux méthodes existantes. Elles indiquent également que les deux critères traditionnels (moindres carrés et log-vraissemblance) ne sont pas les meilleurs dans toutes les situations. Nous tirons ensuite des conclusions et nous dessinons les perspectives que ce travail apporte. Les annexes contiennent les parties les plus techniques des résultats théoriques que nous avons obtenus. Les annexes A, B et C sont en lien avec le chapitre 2. L'annexe A propose des méthodes pour estimer la matrice diagonale la plus proche d'une matrice selon les divergences pour lesquelles un algorithme itératif est requis. L'annexe B comporte les gradients et les hessiennes de l'ensemble des critères. L'annexe C présente des méthodes pour évaluer numériquement le logarithme matriciel et sa dérivée première et contient une contribution originale concernant l'estimation de la dérivée seconde du logarithme matriciel. Enfin, l'annexe D est composée des preuves de certaines propositions du chapitre 3.

Ce manuscrit reprend, améliore et amplifie les travaux que nous avons réalisé pendant cette thèse et qui sont présentés dans les articles publiés [23, 24, 26] et dans l'article sur le point d'être soumis [25]. L'article [24], qui comprend le contenu de [23, 26], propose le modèle géométrique pour la diagonalisation conjointe et l'optimisation riemannienne sur la variété polaire associée aux contraintes oblique et intrinsèque. Les résultats théoriques de cet article sont repris dans les chapitres 2 et 3, qui apportent également du contenu original : des résultats théoriques, des explications et des simplifications, notamment au niveau de l'intégration des contraintes dans la variété polaire. Dans [25], nous considérons la variété des matrices inversibles équipée des métriques invariantes à gauche et à droite. Nous développons les outils d'optimisation pour les contraintes oblique et non-holonomique, qui sont dans le chapitre 3. Nous ajoutons ici l'intégration de la contrainte intrinsèque. Plusieurs travaux effectués au cours de cette thèse ne sont pas présentés dans ce manuscrit car ils sont à la marge de nos problématiques principales. Dans [27], nous investiguons la réduction de dimension dans le cadre de la diagonalisation conjointe et de la séparation aveugle de sources. La réduction de dimension est aussi traitée dans [44, 100, 101] pour la classification de signaux électroencéphalographiques. Dans [73], nous proposons un modèle bilinéaire pour la séparation aveugle de sources dans l'objectif d'analyser des signaux qui possèdent une structure temporelle. Enfin, dans [29, 30], nous déterminons la distance de Fisher associée aux distributions elliptiques complexes.

Chapitre 1

Contexte et revue bibliographique

Sommaire

Dans ce chapitre, nous présentons le contexte de ce travail de thèse et nous proposons une revue bibliographique des sujets au centre de nos intérêts. Dans la section 1.1, nous introduisons l'électroencéphalographie, qui est une méthode de mesure de l'activité cérébrale avec plusieurs avantages et applications, mais aussi avec des limitations. Le développement de méthodes d'analyses est donc nécessaire pour arriver à dépasser ces limitations. Dans la section 1.2, nous exposons la séparation aveugle de sources, une famille de techniques statistiques d'analyse de données, qui peut notamment être utilisée pour l'analyse d'électroencéphalogrammes afin d'obtenir les signaux des sources cérébrales à l'origine des enregistrements. Dans la section 1.3, nous présentons la diagonalisation conjointe approximée qui joue un rôle central dans la résolution du problème de séparation aveugle de sources. Beaucoup d'algorithmes ont été proposés pour diagonaliser conjointement un ensemble de matrices. Cependant, la grande majorité de ces méthodes sont spécifiques et il reste à unifier, généraliser et étendre les différentes approches développées. C'est l'objectif de cette thèse où nous adoptons un point de vue géométrique du modèle de diagonalisation conjointe et proposons un cadre d'optimisation générique adapté à ce problème. Dans la section 1.4, nous rappelons les notions de la géométrie

et de l'optimisation riemannienne que nous utilisons dans la suite pour répondre à ces besoins. Finalement, dans la section 1.5, nous présentons trois variétés qui sont particulièrement importantes par rapport à notre travail : la variété des matrices symétriques positives définies, celle des matrices orthogonales et celle des matrices inversibles.

1.1 Électroencéphalographie : principe et enjeux

La méthode de mesure de l'activité cérébrale, appelée modalité de neuroimagerie, qui nous intéresse dans ce travail est l'électroencéphalographie (EEG). Cette modalité non invasive consiste à enregistrer les potentiels électriques, de l'ordre du microvolt, provenant du cerveau d'un sujet en plaçant des électrodes au niveau du cuir chevelu. L'introduction de cette technique est généralement attribuée à Hans Berger, qui produisit le premier électroencéphalogramme au cours des années 1920 [17]. Pour comparer plus facilement différents enregistrements, le positionnement des électrodes sur le cuir chevelu est standardisé à partir des années 1950 [62]. Le signal mesuré par les électrodes est généré par la sommation de potentiels postsynaptiques de 10^8 à 10^9 neurones pyramidaux du cortex (organisés en colonnes et orientés dans la même direction) qui s'activent de manière synchrone [86, 88]. Ces colonnes de neurones peuvent alors être identifiées à des dipôles électriques, caractérisés par leur position et leur orientation [86, 88]. Nous pouvons considérer qu'il n'y a pas de retards entre les signaux enregistrés et les dipôles du moment que ces derniers ne bougent pas [86] et donc que l'activité instantanée enregistrée correspond bien à l'activité instantanée réelle. Ces sources cérébrales oscillent à des fréquences où les équations de Maxwell quasi-statiques s'appliquent et pour lesquelles les effets capacitifs peuvent être négligés. De ce fait, les données observées sont en phase avec les dipôles réels et un modèle de conduction électrique linéaire peut donc être employé [88]. La figure 1.1 présente une illustration du principe de fonctionnement de l'électroencéphalographie, un schéma du système international 10-20 de placement des électrodes et un exemple d'électroencéphalogramme sur huit électrodes.

Son coût relativement faible, sa simplicité, son faible encombrement (la situation reste proche du naturel) et sa haute résolution temporelle (qui permet de bien capturer la dynamique de l'activité cérébrale) ont fait la popularité de l'électroencéphalographie et ont permis son utilisation dans un certain nombre d'applications. Cette technique est en effet utilisée massivement en recherche fondamentale sur le cerveau et reste aujourd'hui incontournable dans l'étude de certains phénomènes tels que le sommeil [49] et l'épilepsie [85]. Une autre application de l'électroencéphalographie est le neurofeedback où l'objectif est d'apprendre à moduler sa propre activité cérébrale pour en améliorer le contrôle [14]. Le neurofeedback a notamment été proposé pour le traitement des troubles de l'attention et de l'hyperactivité. Grâce à ses caractéristiques, l'électroencéphalographie est aussi la modalité de neuroimagerie privilégiée pour les interfaces cerveau-ordinateur, où le sujet intéragit avec un ordinateur par le biais des signaux caractérisant son activité cérébrale. Les interfaces cerveau-ordinateur peuvent par exemple être utilisées dans l'assistance de personnes souffrant de handicaps moteurs aigus [78] ou pour les jeux vidéo [46].

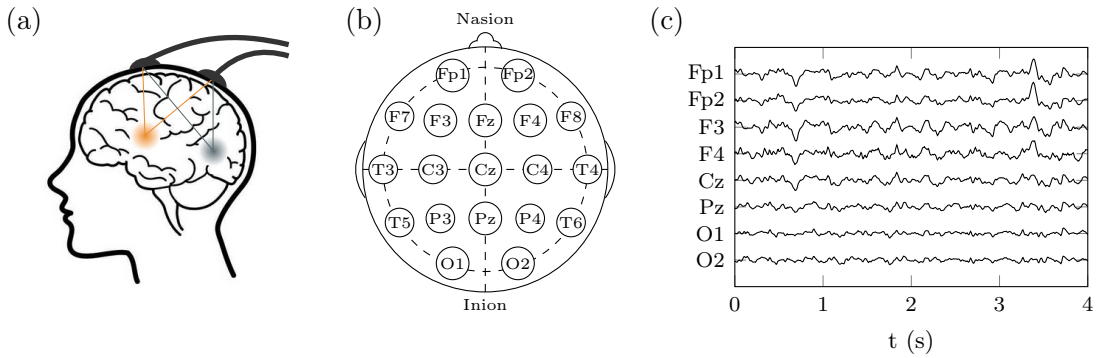

FIGURE *1.1 – Illustration de l'électroencéphalographie (EEG). (a) : illustration du principe de fonctionnement de l'EEG. (b) : schéma du système international 10-20 de placement des électrodes. (c) : exemple d'électroencéphalogramme sur huit électrodes*

L'électroencéphalographie présente deux limitations majeures : sa faible résolution spatiale et son faible rapport signal sur bruit. Le signal électrique d'une source cérébrale (qui traverse du tissu cérébral, le liquide céphalo-rachidien, le crâne et la peau) contribue aux potentiels enregistrés par plusieurs électrodes, ce qui nuit à sa localisation et explique la faible résolution spatiale. De plus, comme plusieurs dipôles cérébraux sont simultanément actifs, on observe un mélange linéaire instantanné de leurs signaux au niveau des électrodes [40]. Ce phénomène explique en partie le faible rapport signal sur bruit du fait que l'activité cérébrale d'intérêt est en partie masquée par l'activité cérébrale de fond qui n'est pas liée aux signaux qu'on souhaite observer. En effet, le cerveau est actif et génère des oscillations même au repos [31]. Lors de l'étude d'un phénomène, on veut donc différencier les oscillations de repos, qui sont alors considérées comme du bruit intra-cérébral, de celles liées à la tâche étudiée. Au mélange des signaux des dipôles cérébraux, s'ajoute également du bruit extra-cérébral de différentes natures. Le premier type est le bruit environnemental, qui correspond par exemple au rayonnement du secteur électrique à 50 Hz ou à des potentiels générés par des mouvements à proximité des électrodes. Le deuxième type est le bruit instrumental, qui est par exemple dû à des problèmes d'impédance, de mauvais contacts d'électrodes ou de mouvements des électrodes et des fils. Notons que le bruit d'amplification est généralement négligeable. On traite ces deux types de bruit soit par filtrage, soit par rejection de certaines portions de l'enregistrement pour certaines électrodes. Leur élimination est un défi technique au niveau du matériel d'enregistrement. Le troisième type est le bruit biologique, qui correspond à des artéfacts générés par exemple par des mouvements oculaires ou musculaires qui peuvent avoir une puissance supérieure à celle de l'activité cérébrale et viennent perturber les mesures. Tout comme l'activité cérébrale, ces artéfacts se mélangent de manière linéaire et instantanée au niveau des électrodes et peuvent donc être directement intégrés dans ce modèle de génération des signaux électroencéphalographiques [40].

Un des enjeux de la recherche méthodologique en électroencéphalographie est donc d'arriver à dépasser ces limitations en développant des méthodes d'analyses qui permettent de retrouver les signaux des sources cérébrales d'intérêt et de les localiser. Une première possibilité est d'utiliser une solution inverse de l'électroencéphalographie [40, 90] qui se base sur

l'étude de la propagation des ondes électriques dans la tête. Une fois qu'un modèle de conduction est défini, le but est de rétropropager les signaux des électrodes afin d'estimer les dipôles à l'origine des observations. La séparation aveugle de sources [39] est également considérée pour l'électroencéphalographie par exemple dans [40, 41]. Cette famille de méthodes, sur laquelle nous nous focalisons pour cette thèse, permet de "démélanger" les signaux des électrodes et d'estimer les signaux des sources en utilisant les statistiques des données. Notons que ces deux approches sont en fait complémentaires puisqu'il est possible d'utiliser les méthodes de solutions inverses sur les sources estimées par séparation aveugle [40].

1.2 Séparation aveugle de sources

Dans cette section, nous présentons la séparation aveugle de sources (*blind source separation*, BSS) initiée dans les articles fondateurs [38, 68]. Un traitement détaillé et complet de ce sujet (théorie et applications) se trouve dans le livre de référence [39]. L'objectif est de démélanger des signaux enregistrés par un ensemble de capteurs pour en extraire des composantes cachées, *i.e.*, les sources à l'origine des phénomènes observés. Cette analyse est dite aveugle car elle repose uniquement sur l'exploitation des statistiques des observations. La séparation aveugle de sources s'est imposée comme un outil majeur du traitement du signal et de l'analyse de données avec de nombreux domaines d'applications tels que les communications, le traitement d'image, l'analyse de signaux audio et biomédicaux.

La section 1.2.1 introduit le problème classique dans le cas du modèle de mélange linéaire instantané. La section 1.2.2 contient les différentes approches qui ont été considérées pour réaliser la séparation aveugle de sources. Finalement, la section 1.2.3 présente l'état de l'art en ce qui concerne l'application de ces méthodes à l'électroencéphalographie.

1.2.1 Modèle et formulation du problème

Nous considérons ici le problème de séparation aveugle de sources, basé sur le modèle de mélange linéaire instantané

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}(t), \qquad 1 \leq t \leq T, \tag{1.1}$$

où t représente l'index de temps (échantillon), $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ est la matrice de mixage (supposée inversible), $\boldsymbol{x}(t)$, $\boldsymbol{s}(t)$ et $\boldsymbol{n}(t)$ sont des vecteurs de $\mathbb{R}^n$ qui correspondent respectivement aux signaux observés, aux sources réelles et à du bruit additif (qui comprend les erreurs du modèle). Étant donné T réalisations de $\boldsymbol{x}(t)$, l'objectif est de trouver des estimations $\widehat{\boldsymbol{A}}$ et $\widehat{\boldsymbol{s}}(t)$ du mixage $\boldsymbol{A}$ et des sources $\boldsymbol{s}(t)$ en faisant uniquement des hypothèses sur les statistiques des sources. Aucune hypothèse n'est donc faite sur le mixage réalisé par $\boldsymbol{A}$ et nous exploitons seulement les informations statistiques contenues dans les observations $\boldsymbol{x}(t)$. La solution $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{s}}(t))$ de ce problème de séparation n'est pas unique : pour toute matrice de permutation $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ et toute matrice diagonale inversible $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, $(\widehat{\boldsymbol{A}}\boldsymbol{P}^T\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Sigma}\boldsymbol{P}\widehat{\boldsymbol{s}}(t))$ est en effet une solution équivalente. Les sources peuvent donc seulement être retrouvées à permutation et échelle

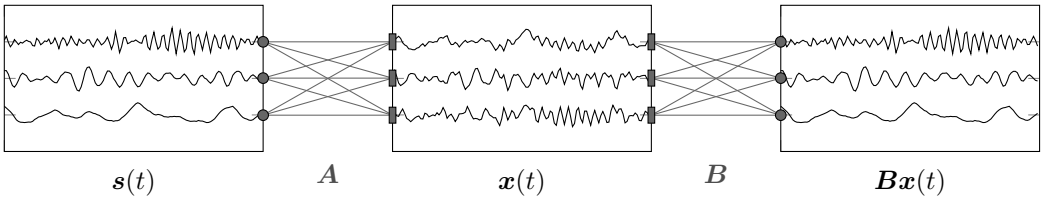

FIGURE *1.2 – Illustration du principe de séparation aveugle de sources. Les sources* $\boldsymbol{s}(t)$ *sont mélangées par la matrice* $\boldsymbol{A}$ *pour former les observations* $\boldsymbol{x}(t)$*. L'objectif est de trouver une matrice de démixage* $\boldsymbol{B}$ *qui permet de retrouver les sources d'origine dans* $\boldsymbol{Bx}(t)$*.*

près. L'amplitude réelle des sources reste donc inconnue et seule leur forme est accessible [1]. Le problème de séparation aveugle de sources revient à rechercher une matrice inversible $\boldsymbol{B} \in \mathbb{R}^{n\times n}$, dite de démixage, telle que $\widehat{\boldsymbol{A}} = \boldsymbol{B}^{-1}$ et $\widehat{\boldsymbol{s}}(t) = \boldsymbol{Bx}(t)$. La figure 1.2 illustre le modèle et le principe de la séparation aveugle de sources.

La séparation de sources se base sur l'hypothèse que les sources $\boldsymbol{s}(t)$ sont indépendantes les unes des autres pour certaines statistiques. On cherche donc $\boldsymbol{B}$ de façon à minimiser la dépendance statistique entre les composantes de $\boldsymbol{Bx}(t)$. Le choix des statistiques utilisées est crucial puisqu'elles doivent permettre de retrouver les sources. Il faut capturer les informations statistiques relatives aux sources présentes dans les observations qui permettent de les discriminer les unes par rapport aux autres, *i.e.*, qui révèlent leurs différences. Selon le modèle considéré pour les sources, différents types de statistiques peuvent être exploités. Dans la formulation originelle du problème, qui correspond à l'analyse en composantes indépendantes (*independent component analysis*, ICA), aucune hypothèse n'est faite sur l'éventuelle structure temporelle des sources, *i.e*, l'hypothèse de travail est que les sources sont temporellement indépendantes et identiquement distribuées (i.i.d.). Pour être séparables, les sources doivent alors être non-gaussiennes. Le caractère non-gaussien est exploité par l'utilisation de statistiques d'ordre supérieur à deux (appelées statistiques d'ordre supérieur). Alternativement, on peut choisir de considérer que les sources possèdent une structure temporelle, *i.e.* non temporellement i.i.d.. Les sources peuvent alors être gaussiennes et on exploite leur non-stationnarité ou leur coloration (spectre en fréquence) pour les séparer. L'utilisation de statistiques d'ordre deux est suffisante dans ce cas.

1.2.2 Méthodes de résolution

Différentes approches ont été considérées pour résoudre le problème de séparation aveugle de sources. L'ensemble des méthodes sont basées sur l'optimisation de critères, appelés fonctions de contraste, qui sont caractérisés par leur capacité à identifier le mélange et séparer les sources. Cette notion de fonction de contraste ainsi que les propriétés de séparabilité à vérifier

1. Notons qu'il est tout de même possible d'accéder à la contribution de la i-ème source estimée $\widehat{s}_i(t)$ de $\widehat{\boldsymbol{s}}(t)$ dans $\boldsymbol{x}(t)$ en la reprojetant dans l'espace des capteurs avec $\widehat{\boldsymbol{a}}_i\widehat{s}_i(t)$, où $\widehat{\boldsymbol{a}}_i \in \mathbb{R}^n$ est la i-ème colonne de la matrice de mixage estimée $\widehat{\boldsymbol{A}}$.

ont été introduites dans [38] et le chapitre 3 de [39] y est consacré.

Dans le cadre de l'analyse en composantes indépendantes, *i.e.*, pour les sources non-gaussiennes, une première solution est de minimiser l'information mutuelle, définie au moyen de la divergence de Kullback-Leibler (ou entropie relative) sur l'espace des densités de probabilité, voir *e.g.*, [38, 91] et [39, chapitre 2]. Une autre classe de méthodes repose sur le maximum de vraisemblance, où l'objectif est de maximiser la fonction de vraisemblance pour le modèle de mélange sous hypothèse d'indépendance des sources, voir par exemple [32, 97] et [39, chapitre 4]. Les approches basées sur l'information mutuelle et sur le maximum de vraisemblance sont très liées comme montré dans [32] et dans les chapitres 3 et 4 de [39]. Finalement, le troisième type de méthodes exploite la diagonalisation conjointe approximée de cumulants des données, voir, entre autres, [33, 38, 83] et [39, chapitre 3].

Pour les sources avec une structure temporelle, il est possible d'utiliser le taux d'information mutuelle (voir [92, 94, 95] et [39, chapitre 2]), ce qui amène à un problème de diagonalisation conjointe approximée avec le critère log-vraisemblance de [93] lorsque les sources sont de plus supposées gaussiennes [39, 92, 95]. On peut également encore utiliser le maximum de vraisemblance comme décrit dans [95] et [39, chapitre 4]. Cette approche se transforme aussi en problème de diagonalisation conjointe avec le critère log-vraisemblance. Cette situation est donc traitée en résolvant un problème de diagonalisation conjointe et il reste à discuter du choix des matrices employées. Si on souhaite exploiter la non-stationnarité, une solution est par exemple d'utiliser les matrices de covariances instantanées définies dans [95]. Pour la coloration, on peut choisir des matrices de corrélation symétrisées comme dans [16] ou des cospectres de Fourier [41]. Notons que le chapitre 7 de [39] contient une revue des méthodes qui exploitent les statistiques d'ordre deux et la coloration.

1.2.3 Application à l'électroencéphalographie

Pour choisir quelles statistiques utiliser afin d'effectuer la séparation aveugle de sources de signaux électroencéphalographiques, il convient d'étudier les caractéristiques des différents processus que nous voulons discriminer, *i.e.*, les artéfacts extra-cérébraux et les potentiels électriques des dipôles cérébraux. Nous résumons ici les arguments présentés dans [41]. En ce qui concerne la détection des artéfacts provenant des mouvements oculaires, la coloration et la non-gaussianité ont toutes deux été considérées. Une conclusion claire ne peut cependant pas être faite sur le choix le plus avantageux et il semblerait en fait que le modèle de séparation de sources qu'on considère n'est pas optimal dans ce cas de figure. Par contre, la séparation des artéfacts musculaires est plus facile et ils sont retrouvés par l'ensemble des méthodes, quel que soit le choix des statistiques. Du côté de l'activité cérébrale, on peut distinguer deux types de phénomènes. Le premier, l'activité dite spontanée, est caractérisé par les différents rythmes observés au niveau de l'électroencéphalogramme [31]. De ce fait, les sources cérébrales correspondantes sont fortement colorées et, comme elles ne sont pas perpétuellement actives, la non-stationnarité peut également être exploitée. Leurs statistiques peuvent donc être précisément modélisées en utilisant uniquement l'ordre deux. Le second type de phénomènes correspond aux activités transitoires telles que les pointes qui sont espacées par des

périodes d'inactivité. Il est possible que des statistiques d'ordre deux permettent de les détecter (non-stationnarité), cependant leur nature non-gaussienne fait qu'elles sont probablement mieux modélisées par des statistiques d'ordre supérieur.

En résumé, exploiter la structure temporelle des données électroencéphalographiques par des statistiques d'ordre deux semble approprié pour analyser un large panel de signaux observés même si les statistiques d'ordre supérieur peuvent être avantageuses pour l'étude de certains phénomènes. Il est donc préconisé par [41] d'exploiter la coloration et la non-stationnarité de l'électroencéphalographie. Pour ce faire, [41] propose de diagonaliser conjointement des matrices cospectrales de Fourier (coloration) des différentes conditions expérimentales (non-stationnarité). Notons qu'une matrice cospectrale de Fourier est la matrice de covariance des données pour une fréquence choisie et à condition d'avoir suffisamment d'échantillons, ce qui est nécessaire pour leur estimation, ces matrices sont symétriques positives définies. L'objectif étant de créer un profil unique pour les sources, les cospectres de Fourier capturent les signatures fréquentielles des différents rythmes et permettent de discriminer les sources dont les spectres sont non proportionnels. De plus, considérer les cospectres de différentes conditions expérimentales permet de différencier les sources qui ne sont pas actives de la même façon, *i.e.*, dont les énergies sont non proportionnelles dans les différentes conditions.

1.3 Diagonalisation conjointe approximée

Dans cette section, nous présentons la diagonalisation conjointe approximée (*approximate joint diagonalization*, AJD) qui est donc une méthode essentielle pour résoudre le problème de séparation aveugle de sources. La section 1.3.1 expose le modèle de diagonalisation conjointe approximée qui aboutit sur un problème d'optimisation sous contraintes. La section 1.3.2 propose une revue des différentes solutions algorithmiques qui ont été proposées. Enfin, la section 1.3.3 liste les limites actuelles et les enjeux de la recherche sur ce sujet.

1.3.1 Un problème d'optimisation sous contraintes

En supposant que les données suivent le modèle (1.1), nous observons un ensemble $\{\boldsymbol{C}_k\}$ de K matrices symétriques et nous supposons qu'elles sont générées selon le modèle

$$\boldsymbol{C}_k = \boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{A}^T + \boldsymbol{N}_k, \qquad 1 \leq k \leq K, \tag{1.2}$$

où $\boldsymbol{A}$ est la matrice de mixage inversible, les matrices $\boldsymbol{\Lambda}_k$ sont diagonales et correspondent aux sources et les matrices $\boldsymbol{N}_k$ contiennent le bruit et les erreurs de modélisation. Selon ce modèle, deux sources peuvent être séparées l'une de l'autre à condition que leurs profils selon la dimension k soient non proportionnels comme montré dans [4]. La diagonalisation conjointe approximée a pour objectif de trouver une matrice inversible $\boldsymbol{B} \in \mathbb{R}^{n\times n}$, appelée diagonalisateur conjoint, telle que l'ensemble $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\}$ contienne des matrices aussi diagonales que possible selon une certaine mesure de diagonalité. Autrement dit, étant donné un critère

$f(\boldsymbol{B}, \{\boldsymbol{C}_k\})$ qui mesure le degré de diagonalité de l'ensemble $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\}$, le diagonalisateur conjoint $\boldsymbol{B}$ est défini comme la solution du problème d'optimisation

$$\underset{\boldsymbol{B}\in \mathrm{GL}_n}{\operatorname{argmin}} \quad f(\boldsymbol{B}, \{\boldsymbol{C}_k\}), \tag{1.3}$$

où GL_n correspond à l'ensemble des matrices inversibles (groupe général linéaire). Dans la suite, $f(\boldsymbol{B}, \{\boldsymbol{C}_k\})$ est simplement noté $f(\boldsymbol{B})$ lorsque préciser quelles matrices $\{\boldsymbol{C}_k\}$ sont utilisées n'est pas utile. Dès lors que $K > 2$, le problème de diagonalisation conjointe approximée n'admet pas de solution analytique dans le cas général pour l'ensemble des fonctions de coût considérées et il faut employer des algorithmes d'optimisation itératifs.

Jusqu'à présent, l'écrasante majorité des études sur la diagonalisation conjointe approximée a utilisé soit le critère basé sur la distance de Frobenius (moindres carrés) introduit dans [33, 34], soit le critère log-vraisemblance initié par [51, 93]. Le critère associé à la distance de Frobenius est

$$f_{\mathrm{F}}(\boldsymbol{B}) = \sum_k w_k \left\| \boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T - \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T) \right\|_F^2, \tag{1.4}$$

où les scalaires w_k sont des poids positifs, $\|\cdot\|_F^2$ est la norme de Frobenius et $\operatorname{ddiag}(\cdot)$ annule les éléments hors de la diagonale de son argument. Le critère log-vraisemblance (*log-likelihood*) est quant à lui

$$f_{\ell\ell}(\boldsymbol{B}) = \sum_k w_k \log \frac{\det(\operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T))}{\det(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T)}, \tag{1.5}$$

où $\det(\cdot)$ est le déterminant. Contrairement au critère de Frobenius (1.4) qui est défini pour l'ensemble des matrices symétriques, les matrices $\boldsymbol{C}_k$ doivent être symétriques positives définies pour que le critère log-vraisemblance (1.5) soit défini.

Comme pour la séparation de sources, la solution $\boldsymbol{B} \in \mathrm{GL}_n$ n'est pas unique. En effet, la solution du problème de diagonalisation conjointe est l'ensemble de la classe d'équivalence

$$\{\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{B} : \boldsymbol{P} \in \mathbb{P}_n, \boldsymbol{\Sigma} \in \mathcal{D}_n^*\}, \tag{1.6}$$

où $\mathbb{P}_n$ est l'ensemble des matrices de permutations et $\mathcal{D}_n^*$ est celui des matrices diagonales inversibles. L'ambiguité de permutation n'est pas problématique en pratique, cependant l'ambiguité d'échelle peut avoir un impact important sur la recherche de la solution. En effet, comme il est possible de construire une suite $(\boldsymbol{\Sigma}_i)$ de matrices dans $\mathcal{D}_n^*$ qui converge vers la matrice nulle, toutes les classes d'équivalences (1.6) contiennent des solutions dégénérées qui doivent être évitées. De ce fait, il est en général préférable d'ajouter des contraintes sur $\boldsymbol{B}$ par-delà celle d'inversibilité. Dans les premiers travaux sur le sujet (*e.g.*, [33, 34, 51]), l'orthogonalité est imposée au diagonalisateur conjoint, justifiée par le fait que les données peuvent être orthogonalisées au préalable par blanchiment. Cependant, cette approche induit des erreurs dans l'estimation de $\boldsymbol{B}$ et il est préférable de rechercher un diagonalisateur non-orthogonal comme expliqué dans [93, 113]. Des contraintes alternatives peuvent être envisagées. Une première possibilité consiste à fixer la norme des lignes de $\boldsymbol{B}$, comme par exemple dans [1, 107] avec la contrainte oblique

$$\operatorname{ddiag}\left(\boldsymbol{B}\boldsymbol{B}^T\right) = \boldsymbol{I}_n, \tag{1.7}$$

où $\boldsymbol{I}_n$ est la matrice identité de $\mathbb{R}^{n\times n}$. Un autre choix de normalisation des lignes, conçu dans [42, 96] pour être adapté aux matrices $\boldsymbol{C}_k$ à diagonaliser, est la contrainte dite intrinsèque

$$\sum_{k=1}^{K} \left(\operatorname{ddiag}\left(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\right)\right)^2 = \boldsymbol{I}_n. \tag{1.8}$$

Finalement, la dernière contrainte évoquée dans ce manuscrit est la contrainte dite non-holonomique qui a été introduite pour la séparation aveugle de sources dans [12] puis utilisée pour la diagonalisation conjointe approximée dans [4, 5, 116]. L'idée derrière cette contrainte est d'exploiter la géométrie des classes d'équivalences (1.6) et d'annuler l'action des matrices diagonales plutôt que de se contenter de sélectionner un représentant comme le font les deux contraintes précédentes (1.7) et (1.8).

1.3.2 Tour d'horizon des solutions algorithmiques

Un grand nombre d'algorithmes a été développé dans le but de résoudre le problème de diagonalisation conjointe approximée. Ces algorithmes diffèrent les uns des autres, d'une part au niveau du choix du critère et de la contrainte, et d'autre part sur la méthode utilisée pour minimiser le critère et sur la façon d'assurer le respect des contraintes.

Au niveau des méthodes d'optimisation, les articles pionniers [33, 34, 51] proposent des algorithmes de type Jacobi. Cette famille de méthodes est également beaucoup utilisée dans le cas non-orthogonal, voir *e.g.*, [55, 77, 93, 103]. On trouve aussi des algorithmes exploitant des multiplicateurs de Lagrange [42, 108] et d'autres utilisant des méthodes de gradient [66], Newton [65, 67] ou quasi-Newton [96]. Une autre approche très efficace développée dans [107] optimise (1.4) de façon indirecte avec un algorithme de point fixe et des itérations de Gauss. Notons finalement que des méthodes utilisant l'optimisation riemannienne ont été proposées dans le cas orthogonal [61, 98, 106] et pour la contrainte oblique [1], mais cette dernière méthode n'assure pas l'inversibilité de la solution. L'approche du gradient naturel développé pour la séparation aveugle de sources dans [11, 12] et appliqué à la diagonalisation conjointe approximée dans [4, 5, 114] se situe aussi dans le cadre de l'optimisation riemannienne. L'idée est d'utiliser un gradient riemannien obtenu en équipant GL_n avec une métrique qui a un lien intrinsèque avec le modèle de mélange du problème. Cependant, le reste de la mise à jour de la matrice de démixage dans ces méthodes est euclidienne.

Pour imposer les contraintes, différentes possibilités ont été utilisées. La première correspond au cas où le respect des contraintes est assuré de façon intrinsèque par la technique d'optimisation employée. Par exemple, les transformations appliquées dans [33, 34, 51] conservent l'orthogonalité. Un autre cas particulièrement intéressant est [116] où l'inversibilité est imposée grâce à l'exploitation du théorème des matrices à diagonale dominante. Les contraintes peuvent également être appliquées à la fin du processus de mise à jour comme par exemple dans [96, 107]. Une autre façon de faire est d'ajouter un terme de pénalité dans la fonction de coût comme dans [66, 67]. Finalement, [1] utilise l'optimisation riemannienne pour imposer la contrainte. Cette approche permet de prendre naturellement en compte les contraintes tout en offrant un large panel de méthodes génériques d'optimisation [2].

1.3.3 Limites actuelles et enjeux

Lorsqu'on met de côté les travaux sur le choix des matrices à diagonaliser, la recherche sur les modèles de la diagonalisation conjointe approximée s'est principalement concentrée autour des deux critères (1.4) et (1.5). Le critère de Frobenius (1.4) vient du raisonnement pratique qu'il faut annuler les termes hors-diagonaux et le critère log-vraisemblance (1.5) est très utilisé du fait de ses liens avec l'information mutuelle et le maximum de vraisemblance. Les travaux [9, 45] suggèrent que les avancées récentes sur la géométrie des matrices symétriques positives définies [15, 20, 21, 36, 50, 81, 104] peuvent être exploitées pour la diagonalisation conjointe approximée. En effet, [45] fait le parallèle entre la diagonalisation conjointe approximée et la recherche de centre de masses et considère notamment les propriétés que les critères de diagonalisation conjointe devraient posséder. De son côté, [9] propose des mesures de diagonalité basées sur plusieurs divergences mais la solution pratique proposée se limite à la divergence log-det α (définie dans [36]) et n'optimise pas la mesure de diagonalité optimale correspondante. Il reste donc à formaliser, généraliser et unifier le modèle de diagonalisation conjointe approximée selon l'approche géométrique ainsi qu'à développer des stratégies pratiques pour résoudre le problème dans ce cadre général.

Les algorithmes développés jusqu'ici sont spécifiques à un critère, une contrainte et une méthode d'optimisation. Dans la plupart des cas, il n'y a pas de façon simple de transposer une méthode spécifique vers d'autres cas. Ceci a pour effet d'entraver la compréhension de l'origine des différences observées dans les résultats de différentes méthodes. Par exemple, [93] et [107] diffèrent à la fois de par leur critère, leur contrainte et leur technique d'optimisation et on ne peut pas vraiment estimer l'impact de chacun de ces constituants dans les différences entre les solutions trouvées par ces deux algorithmes. Un enjeu de la recherche est donc de développer des outils qui permettent d'unifier les différentes approches possibles et d'offrir une grande modularité au niveau du choix des constituants de la diagonalisation conjointe approximée. Par les liens possibles avec le modèle de diagonalisation conjointe, la gestion naturelle des contraintes et les méthodes d'optimisation génériques disponibles, l'optimisation riemannienne apparaît comme un bon candidat pour développer de tels outils.

1.4 Géométrie et optimisation riemannienne

Dans cette section, nous introduisons les concepts de géométrie et d'optimisation riemannienne sur les variétés matricielles qui sont utilisés dans ce travail de thèse. Au niveau des objets de géométrie riemannienne, nous présentons l'espace tangent, la métrique, la connection de Levi-Civita, les géodésiques, l'exponentielle et le logarithme riemanniens, la distance et le transport parallèle. Pour les outils spécifiques à l'optimisation, nous nous concentrons sur le gradient, la hessienne, la rétraction et le transport de vecteurs. L'ensemble de ces notions se trouve dans les livres de références [2, 52, 75]. Un tel sujet ne pouvant être détaillé de façon concise, il est recommandé de consulter un ouvrage de référence comme [2] pour une présentation précise et pédagogique des outils introduits dans cette section. Dans le cadre de l'optimisation riemannienne, ces notions suffisent pour l'utilisation d'un large panel d'al-

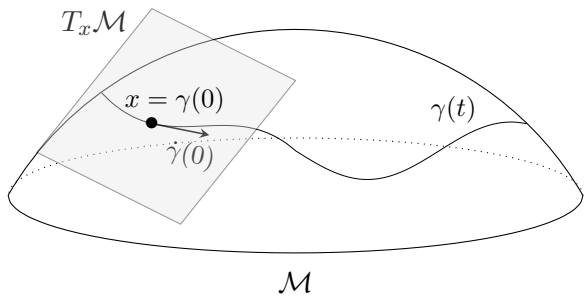

FIGURE *1.3 – Schéma d'illustration d'une variété lisse $\mathcal{M}$ et de l'espace tangent $T_x\mathcal{M}$ à $x \in \mathcal{M}$. γ est une courbe lisse sur $\mathcal{M}$ telle que $\gamma(0) = x$. Sa dérivée $\dot{\gamma}(0)$ est donc un vecteur de $T_x\mathcal{M}$.*

gorithmes d'optimisation génériques, *e.g.*, la descente de gradient, le gradient conjugué, la méthode de Newton et des méthodes de quasi-Newton [2]. Un avantage important de l'optimisation riemannienne est qu'il existe des preuves de convergence des algorithmes d'optimisation dans le cas général, voir [2]. Dans la section 1.4.1, nous traitons le cadre général des variétés lisses. Dans la section 1.4.2, nous exposons le cadre particulier des sous-variétés. Enfin, dans la section 1.4.3, nous présentons celui des variétés quotientes.

1.4.1 Variété lisse

Géométrie

De manière analytique, une variété lisse $\mathcal{M}$ est un ensemble localement difféomorphe à un espace linéaire qui admet une structure différentielle, *i.e.*, tout point $x \in \mathcal{M}$ a un espace tangent $T_x\mathcal{M}$ dont les éléments, appelés vecteurs tangents, généralisent l'idée de dérivées directionnelles. Soit $\gamma : \mathbb{R} \to \mathcal{M}$, une courbe lisse dans $\mathcal{M}$ qui passe par le point $x \in \mathcal{M}$ en 0 (*i.e.*, $\gamma(0) = x$). On peut alors définir la dérivée $\dot{\gamma}(0)$ de γ à 0 par la formule classique

$$\dot{\gamma}(0) = \lim_{t \to 0} \frac{\gamma(t) - \gamma(0)}{t}. \tag{1.9}$$

L'espace tangent $T_x\mathcal{M}$ de x dans $\mathcal{M}$ est l'ensemble $\{\dot{\gamma}(0) : \gamma \text{ courbe lisse dans } \mathcal{M}, \gamma(0) = x\}$. Il peut être vu comme une linéarisation locale de la variété $\mathcal{M}$ autour de x (voir la figure 1.3 pour une illustration). La variété $\mathcal{M}$ devient une variété riemannienne lorsqu'on l'équipe d'une métrique riemannienne, qui définit un produit scalaire sur tous les espaces tangents.

Nous sommes régulièrement amenés à manipuler des champs de vecteurs sur $\mathcal{M}$, *i.e.*, des fonctions qui associent un vecteur tangent de $T_x\mathcal{M}$ à chaque point $x \in \mathcal{M}$. Les dérivées directionnelles des champs de vecteurs sont généralisées par les connections affines. Une connection affine a un rôle particulier sur une variété riemannienne $\mathcal{M}$. La connection de Levi-Civita ∇ associée à la métrique $\langle \cdot, \cdot \rangle$. de $\mathcal{M}$ est l'unique connection affine sur $\mathcal{M}$ qui vérifie l'équation

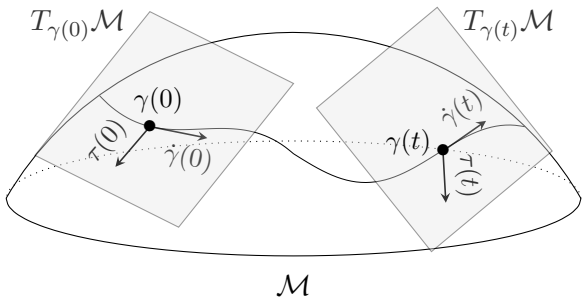

FIGURE *1.4 – Schéma d'illustration du transport parallèle τ sur $\mathcal{M}$ le long de la courbe γ. Le vecteur tangent $\tau(0)$ de $T_{\gamma(0)}\mathcal{M}$ correspond au vecteur tangent $\tau(t)$ dans $T_{\gamma(t)}\mathcal{M}$.*

de Koszul

$$2\langle\nabla_{\xi_x}\eta_x,\nu_x\rangle_x = \mathrm{D}\langle\eta_x,\nu_x\rangle_x[\xi_x] + \mathrm{D}\langle\xi_x,\nu_x\rangle_x[\eta_x] - \mathrm{D}\langle\xi_x,\eta_x\rangle_x[\nu_x] + \langle\nu_x,[\xi_x,\eta_x]\rangle_x + \langle\eta_x,[\nu_x,\xi_x]\rangle_x - \langle\xi_x,[\eta_x,\nu_x]\rangle_x, \quad (1.10)$$

où ξ_x, η_x et ν_x sont des champs de vecteurs sur $\mathcal{M}$ évalués en x, $\mathrm{D}\,f(x)[\xi_x]$ est la dérivée directionnelle d'une fonction f à x dans la direction ξ_x et $[\cdot,\cdot]$ est le crochet de Lie.

La connection de Levi-Civita permet de définir les géodésiques sur $\mathcal{M}$ associées à la métrique $\langle\cdot,\cdot\rangle$.. Les géodésiques permettent de généraliser le concept de lignes droites pour les variétés riemanniennes. Ce sont en effet les courbes $\gamma : I \subset \mathbb{R} \to \mathcal{M}$ dont l'accélération est nulle, ce qui se traduit dans notre contexte par

$$\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0. \quad (1.11)$$

Notons qu'une variété riemanniene est dite géodésiquement complète si le domaine de définition I de ses géodésiques γ est toute la ligne $\mathbb{R}$. Chaque géodésique γ est caractérisée par le choix d'un point initial $\gamma(0) = x \in \mathcal{M}$ et d'une direction initiale $\dot\gamma(0) = \xi \in T_x\mathcal{M}$. Notons qu'il est possible (et équivalent) de choisir un second point $\gamma(1) = y \in \mathcal{M}$ plutôt que $\dot\gamma(0)$ pour caractériser γ. La définition de la distance sur $\mathcal{M}$ associée à la métrique $\langle\cdot,\cdot\rangle$. s'en suit : étant donné deux points x et y de $\mathcal{M}$ liés par la géodésique γ (telle que $\gamma(0) = x$ et $\gamma(1) = y$), la distance $\delta(x,y)$ entre x et y est définie par

$$\delta^2(x,y) = \int_0^1 \langle\dot\gamma(t),\dot\gamma(t)\rangle_{\gamma(t)}\, dt. \quad (1.12)$$

Un objet intimement lié aux géodésiques est l'exponentielle riemannienne. À un point $x \in \mathcal{M}$, c'est l'application de $T_x\mathcal{M}$ vers $\mathcal{M}$ qui, à $\xi \in T_x\mathcal{M}$, associe $\exp_x(\xi) = \gamma(1)$, où γ est la géodésique tel que $\gamma(0) = x$ et $\dot\gamma(0) = \xi$. On peut également définir la réciproque de cette fonction : le logarithme riemannien. À $x \in \mathcal{M}$, c'est cette fois l'application de $\mathcal{M}$ vers $T_x\mathcal{M}$ qui, à $y \in \mathcal{M}$ associe le vecteur $\log_x(y) = \xi \in T_x\mathcal{M}$ tel que $\exp_x(\xi) = y$.

Le dernier objet que nous présentons ici est le transport parallèle sur $\mathcal{M}$, dont un schéma illustratif est donné figure 1.4. Étant donné une courbe γ sur $\mathcal{M}$, l'idée est de transporter un

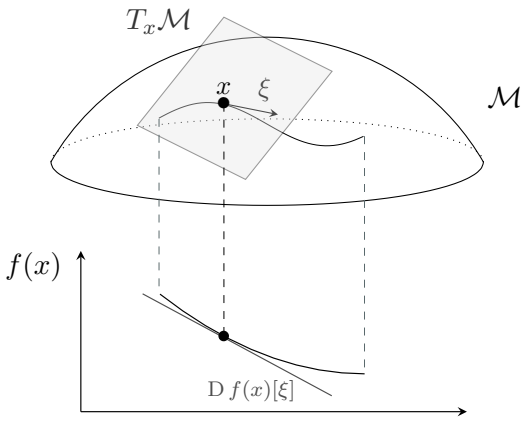

FIGURE *1.5 – Schéma d'illustration de l'optimisation riemannienne sur une variété $\mathcal{M}$. Pour minimiser la fonction f, on construit un chemin sur $\mathcal{M}$ tel que, à x, on suit une direction de descente $\xi \in T_x\mathcal{M}$ telle que $\mathrm{D}\, f(x)[\xi] < 0$ (*i.e.*, qui entraîne une diminution de f).*

vecteur tangent $\tau(0)$ de $T_{\gamma(0)}\mathcal{M}$ le long de γ de sorte d'obtenir le vecteur tangent correspondant $\tau(t)$ dans $T_{\gamma(t)}\mathcal{M}$. Le vecteur $\tau(t)$ de $T_{\gamma(t)}\mathcal{M}$ correspond au vecteur $\tau(0)$ de $T_{\gamma(0)}\mathcal{M}$ dans le sens où le transport le long de γ est parallèle par rapport à la métrique $\langle \cdot, \cdot \rangle_{\cdot}$ de $\mathcal{M}$, ce qui se traduit de la façon suivante : étant donné $\tau(0) \in T_{\gamma(0)}\mathcal{M}$, le transport parallèle τ le long de γ est solution de l'équation

$$\nabla_{\dot{\gamma}(t)}\, \tau(t) = 0. \tag{1.13}$$

Optimisation

Considérons maintenant une fonction objectif $f : \mathcal{M} \to \mathbb{R}$ (supposée lisse) que nous souhaitons optimiser. Le gradient riemannien $\mathrm{grad}_{\mathcal{M}}\, f(x)$ de f à $x \in \mathcal{M}$ est défini au travers de la métrique $\langle \cdot, \cdot \rangle_{\cdot}$ comme l'unique vecteur tangent de $T_x\mathcal{M}$ tel que pour tout $\xi \in T_x\mathcal{M}$

$$\langle \mathrm{grad}_{\mathcal{M}}\, f(x), \xi \rangle_x = \mathrm{D}\, f(x)[\xi]. \tag{1.14}$$

Remarquons que le gradient riemannien $\mathrm{grad}_{\mathcal{M}}\, f$ est un champ de vecteurs sur $\mathcal{M}$. De ce fait, la hessienne riemannienne $\mathrm{hess}_{\mathcal{M}}\, f(x)$ de f à $x \in \mathcal{M}$ est définie au travers de la connection de Levi-Civita ∇ comme l'application linéaire de $T_x\mathcal{M}$ dans $T_x\mathcal{M}$ telle que pour tout $\xi \in T_x\mathcal{M}$

$$\mathrm{hess}_{\mathcal{M}}\, f(x)[\xi] = \nabla_{\xi}\, \mathrm{grad}_{\mathcal{M}}\, f(x). \tag{1.15}$$

Le gradient riemannien, et éventuellement la hessienne, de f à $x \in \mathcal{M}$ sont utilisés pour obtenir une direction de descente qui est un vecteur tangent ξ de $T_x\mathcal{M}$ tel que $\mathrm{D}\, f(x)[\xi] < 0$ (voir la figure 1.5 pour une illustration). Pour obtenir un nouveau point sur $\mathcal{M}$ à partir de ξ, nous avons besoin d'une rétraction à x, qui est une application de $T_x\mathcal{M}$ sur $\mathcal{M}$. Toute variété riemannienne possède une rétraction naturelle : l'exponentielle riemannienne introduite précédemment. Seulement, l'exponentielle riemannienne peut être compliquée à obtenir ou

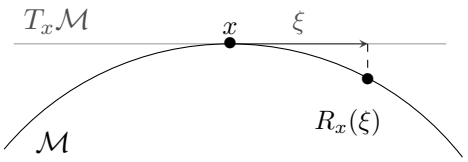

FIGURE *1.6 – Schéma d'illustration d'une rétraction R sur une variété $\mathcal{M}$.*

trop coûteuse à évaluer. Dans de tels cas, on peut utiliser une rétraction alternative R qui correspond à une approximation au premier ordre de l'exponentielle riemannienne, c'est à dire telle que

$$R_x(\xi) = x + \xi + o(\|\xi\|). \tag{1.16}$$

La figure 1.6 contient une illustration de cet objet. L'itération de la descente de gradient de f est alors définie comme suit : étant donné l'itéré $x_i \in \mathcal{M}$ et le pas $t_i > 0$, on définit

$$x_{i+1} = R_{x_i}(-t_i \operatorname{grad}_{\mathcal{M}} f(x_i)). \tag{1.17}$$

Enfin, certains algorithmes d'optimisation tels que le gradient conjugué ou les méthodes de quasi-Newton utilisent l'information donnée par la direction de descente d'un ou plusieurs itérés précédents. Dans le contexte de l'optimisation riemannienne, il faut donc être en mesure de transporter un vecteur tangent d'un point de $\mathcal{M}$ dans l'espace tangent à un autre point. C'est le rôle du transport de vecteurs $\mathcal{T}$ sur $\mathcal{M}$ qui, à partir d'un point $x \in \mathcal{M}$ et de vecteurs tangents $\xi, \eta \in T_x\mathcal{M}$, transporte le vecteur η dans l'espace tangent du point obtenu par la rétraction de ξ à x. Il est possible de définir un transport de vecteurs $\mathcal{T}$ sur $\mathcal{M}$ à partir du transport parallèle par $\mathcal{T}(x, \xi, \eta) = \tau(1)$, où τ est le transport parallèle le long du géodésique γ avec $\tau(0) = \eta$, $\gamma(0) = x$ et $\dot{\gamma}(0) = \xi$. Comme pour les rétractions, on peut aussi définir des transports de vecteurs alternatifs avec des hypothèses moins fortes comme montré pour les cas particuliers des sous-variétés et des variétés quotient.

1.4.2 Sous-variété

Géométrie

Nous traitons maintenant le cas d'une sous-variété $\mathcal{M}$ d'une variété riemannienne $\overline{\mathcal{M}}$ équipée d'une métrique $\langle \cdot, \cdot \rangle_\cdot$. L'espace tangent $T_x\mathcal{M}$ à x dans $\mathcal{M}$ est alors un sous-espace de l'espace tangent $T_x\overline{\mathcal{M}}$. Une sous-variété $\mathcal{M}$ est généralement définie à travers un ensemble de contraintes dans $\overline{\mathcal{M}}$ et $T_x\mathcal{M}$ peut être obtenu en les différenciant. $\mathcal{M}$ hérite simplement de la métrique $\langle \cdot, \cdot \rangle_\cdot$ de $\overline{\mathcal{M}}$ pour devenir une sous-variété riemannienne. On peut alors définir le projecteur orthogonal selon la métrique $\langle \cdot, \cdot \rangle_\cdot$, *i.e.*, défini de $T_x\overline{\mathcal{M}}$ sur $T_x\mathcal{M}$ et tel que pour tout vecteur $\xi \in T_x\overline{\mathcal{M}}$

$$\langle \xi - P_x(\xi), P_x(\xi) \rangle_x = 0. \tag{1.18}$$

Un schéma illustratif de cette situation est donné figure 1.7. Ce projecteur est particulièrement utile pour définir certains des objets de géométrie et d'optimisation requis.

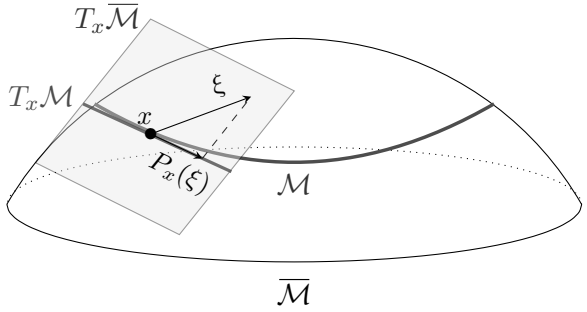

FIGURE *1.7 – Schéma d'illustration d'une sous-variété $\mathcal{M}$ d'une variété $\overline{\mathcal{M}}$ et de son espace tangent $T_x\mathcal{M}$ qui est un sous-espace de $T_x\overline{\mathcal{M}}$. Un vecteur tangent $\xi \in T_x\overline{\mathcal{M}}$ est projeté sur $T_x\mathcal{M}$ avec la projection orthogonale P_x (selon la métrique $\langle \cdot, \cdot \rangle$.).*

Un premier exemple de ce fait concerne la connection de Levi-Civita. Notons $\overline{\nabla}$ la connection de Levi-Civita de $\overline{\mathcal{M}}$. Il s'en suit que la connection de Levi-Civita ∇ de $\mathcal{M}$ est définie, pour $x \in \mathcal{M}$ et les champs de vecteurs ξ_x, η_x sur $\mathcal{M}$ évalués en x, par

$$\nabla_{\xi_x} \eta_x = P_x \left(\overline{\nabla}_{\xi_x} \eta_x \right). \tag{1.19}$$

En général, les géodésiques de $\mathcal{M}$ ne sont pas obtenues simplement à partir de ceux de $\overline{\mathcal{M}}$. Il en va de même pour la distance et le transport parallèle.

Optimisation

Considérons la fonction objectif $f : \overline{\mathcal{M}} \to \mathbb{R}$. Par abus de notation, nous notons également f sa restriction à la sous-variété $\mathcal{M}$. Le gradient riemannien $\operatorname{grad}_{\mathcal{M}} f(x)$ de f dans $\mathcal{M}$ est donné pour tout $x \in \mathcal{M}$ par

$$\operatorname{grad}_{\mathcal{M}} f(x) = P_x \left(\operatorname{grad}_{\overline{\mathcal{M}}} f(x) \right), \tag{1.20}$$

et la hessienne riemannienne est obtenue en injectant (1.19) et (1.20) dans (1.15). Finalement, étant donné une rétraction R sur $\mathcal{M}$, on peut définir un transport de vecteurs $\mathcal{T}(x, \xi, \eta)$ à $x \in \mathcal{M}$ et $\xi, \eta \in T_x\mathcal{M}$ par

$$\mathcal{T}(x, \xi, \eta) = P_{R_x(\xi)}(\eta). \tag{1.21}$$

Notons que P correspond ici par abus à la projection depuis l'espace ambiant, c'est à dire l'espace matriciel euclidien dans lesquels $\overline{\mathcal{M}}$ et $\mathcal{M}$ sont définis.

1.4.3 Variété quotient

Géométrie

La notion de variété quotient est plus abstraite. Les éléments d'une variété quotient $\mathcal{M}$ d'une variété $\overline{\mathcal{M}}$ sont en effet des classes d'équivalences sur $\overline{\mathcal{M}}$, *i.e.*, des ensembles de points

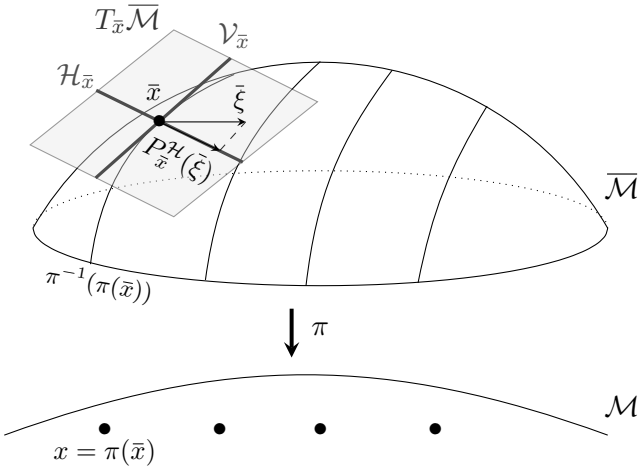

FIGURE *1.8 – Schéma d'illustration d'une variété quotient $\mathcal{M}$ d'une variété $\overline{\mathcal{M}}$. L'espace tangent $T_{\bar{x}}\overline{\mathcal{M}}$ se décompose en deux sous-espaces complémentaires : l'espace vertical $\mathcal{V}_{\bar{x}} = T_{\bar{x}}\pi^{-1}(\pi(\bar{x}))$ et l'espace horizontal $\mathcal{H}_{\bar{x}}$ qui fournit des représentations des vecteurs tangents de $T_x\mathcal{M}$ à $x = \pi(\bar{x})$. La projection horizontale $P^{\mathcal{H}}_{\bar{x}}$ permet de projeter $\bar{\xi} \in T_{\bar{x}}\overline{\mathcal{M}}$ sur $\mathcal{H}_{\bar{x}}$.*

de $\overline{\mathcal{M}}$ qui sont équivalents. Pour manipuler les éléments de la variété quotient $\mathcal{M}$, la technique usuelle est d'utiliser la projection canonique $\pi : \overline{\mathcal{M}} \to \mathcal{M}$ qui associe $x = \pi(\bar{x}) \in \mathcal{M}$ à tout $\bar{x} \in \overline{\mathcal{M}}$. Remarquons que la classe d'équivalence de $\bar{x} \in \overline{\mathcal{M}}$ est alors obtenue sur $\overline{\mathcal{M}}$ par $\pi^{-1}(\pi(\bar{x}))$. Chaque élément x de $\mathcal{M}$ peut donc être représenté par tout élément $\bar{x}$ de $\overline{\mathcal{M}}$ tel que $x = \pi(\bar{x})$. Plus généralement, l'ensemble des objets géométriques de $\mathcal{M}$ qui nous intéresse peut être caractérisé par des représentations au niveau matriciel.

L'espace tangent $T_x\mathcal{M}$ de $x = \pi(\bar{x}) \in \mathcal{M}$ peut être décrit par un sous-espace de l'espace tangent de $\bar{x}$ dans $\overline{\mathcal{M}}$. En effet, $T_{\bar{x}}\overline{\mathcal{M}}$ se décompose en deux sous-espaces complémentaires : l'espace vertical $\mathcal{V}_{\bar{x}}$ et l'espace horizontal $\mathcal{H}_{\bar{x}}$. L'espace vertical $\mathcal{V}_{\bar{x}}$ est l'espace tangent $T_{\bar{x}}\pi^{-1}(\pi(\bar{x}))$ de la classe d'équivalence $\pi^{-1}(\pi(\bar{x}))$ à $\bar{x}$. Il contient l'ensemble des éléments de $T_{\bar{x}}\overline{\mathcal{M}}$ qui induisent un déplacement le long de $\pi^{-1}(\pi(\bar{x}))$. L'espace horizontal $\mathcal{H}_{\bar{x}}$ est quant à lui défini comme le complément orthogonal à $\mathcal{V}_{\bar{x}}$ dans $T_{\bar{x}}\overline{\mathcal{M}}$ selon la métrique $\langle\cdot,\cdot\rangle_{\bar{x}}$ de $\overline{\mathcal{M}}$. Les éléments de $\mathcal{H}_{\bar{x}}$ procurent des représentations des vecteurs tangents de $T_x\mathcal{M}$: pour tout $\xi \in T_x\mathcal{M}$, il existe un unique vecteur horizontal $\bar{\xi} \in \mathcal{H}_{\bar{x}}$ tel que $\mathrm{D}\,\pi(\bar{x})[\bar{\xi}] = \xi$. Notons finalement qu'il est possible de définir la projection $P^{\mathcal{H}}_{\bar{x}}$ de $T_{\bar{x}}\overline{\mathcal{M}}$ sur $\mathcal{H}_{\bar{x}}$. Un schéma illustratif d'une variété quotient et de ces notions est disponible figure 1.8.

L'espace horizontal dépend donc du choix de la métrique $\langle\cdot,\cdot\rangle_{\bar{x}}$ sur $\overline{\mathcal{M}}$. Pour que $\mathcal{M}$ soit une variété quotient riemannienne correctement définie, il faut que $\langle\cdot,\cdot\rangle_{\bar{x}}$ induise une métrique riemannienne sur $\mathcal{M}$. Pour que ce soit le cas, la métrique $\langle\cdot,\cdot\rangle_{\bar{x}}$ doit être invariante le long de chaque classe d'équivalence. Prenons $x \in \mathcal{M}$, $\xi, \eta \in T_x\mathcal{M}$ et $\bar{x}, \bar{y} \in \pi^{-1}(x)$. Soient $\bar{\xi}_{\bar{x}}, \bar{\eta}_{\bar{x}} \in \mathcal{H}_{\bar{x}}$ et $\bar{\xi}_{\bar{y}}, \bar{\eta}_{\bar{y}} \in \mathcal{H}_{\bar{y}}$ les vecteurs horizontaux représentant les vecteurs tangents ξ et η (*i.e.*, tels que $\xi = \mathrm{D}\,\pi(\bar{x})[\bar{\xi}_{\bar{x}}] = D\pi(\bar{y})[\bar{\xi}_{\bar{y}}]$ et idem pour η). L'invariance de $\langle\cdot,\cdot\rangle_{\bar{x}}$ le long de la classe d'équivalence $\pi^{-1}(x)$ se traduit alors mathématiquement par

$$\langle\bar{\xi}_{\bar{x}}, \bar{\eta}_{\bar{x}}\rangle_{\bar{x}} = \langle\bar{\xi}_{\bar{y}}, \bar{\eta}_{\bar{y}}\rangle_{\bar{y}}. \tag{1.22}$$

Notons $\overline{\nabla}$ la connection de Levi-Civita sur $\overline{\mathcal{M}}$. Soient $x = \pi(\bar{x}) \in \mathcal{M}$ et les champs de vecteurs ξ_x, η_x évalués en x correspondant à $\bar{\xi}_{\bar{x}}, \bar{\eta}_{\bar{x}}$ dans $\mathcal{H}_{\bar{x}}$. La représentation de la connection de Levi-Civita $\nabla_{\xi_x} \eta_x$ sur $\mathcal{M}$ dans $\mathcal{H}_{\bar{x}}$ est donnée par $P^{\mathcal{H}}_{\bar{x}}(\overline{\nabla}_{\bar{\xi}_{\bar{x}}} \bar{\eta}_{\bar{x}})$, *i.e.*,

$$\nabla_{\xi_x} \eta_x = \mathrm{D}\,\pi(\bar{x})[P^{\mathcal{H}}_{\bar{x}}(\overline{\nabla}_{\bar{\xi}_{\bar{x}}} \bar{\eta}_{\bar{x}})]. \tag{1.23}$$

Il est possible d'obtenir les géodésiques sur $\mathcal{M}$ à partir de géodésiques sur $\overline{\mathcal{M}}$. En effet, si $\bar{\gamma}$ est une géodésique complète de $\overline{\mathcal{M}}$ qui reste horizontale (*i.e.*, sa dérivée $\dot{\bar{\gamma}}(t)$ à $t \in \mathbb{R}$ est dans $\mathcal{H}_{\bar{\gamma}(t)}$), alors $\gamma = \pi \circ \bar{\gamma}$ est une géodésique complète de $\mathcal{M}$. De même, si $\bar{\gamma}$ est une courbe sur $\overline{\mathcal{M}}$ qui reste horizontale (donc induit la courbe $\gamma = \pi \circ \bar{\gamma}$ sur $\mathcal{M}$) et que le transport parallèle $\bar{\tau}$ le long de $\bar{\gamma}$ reste également horizontal (*i.e.*, $\bar{\tau}(t) \in \mathcal{H}_{\bar{\gamma}(t)}$ pour tout t), alors $\tau(t) = \mathrm{D}\,\pi(\bar{\gamma}(t))[\bar{\tau}(t)]$ défini un transport parallèle le long de γ sur $\mathcal{M}$.

Optimisation

Soit $\bar{f}$ une fonction objectif sur $\overline{\mathcal{M}}$. La fonction $\bar{f}$ induit une fonction f sur $\mathcal{M}$ si, pour tous $\bar{x} \in \overline{\mathcal{M}}$ et $\bar{y} \in \pi^{-1}(\pi(\bar{x}))$, $\bar{f}(\bar{x}) = \bar{f}(\bar{y})$. On a alors $\bar{f} = f \circ \pi$. Le gradient riemannien $\mathrm{grad}_{\mathcal{M}}\, f(x)$ de f à $x = \pi(\bar{x}) \in \mathcal{M}$ est simplement représenté par le gradient de $\bar{f}$ à $\bar{x}$ qui est un élément de $\mathcal{H}_{\bar{x}}$, *i.e.*,

$$\mathrm{grad}_{\mathcal{M}}\, f(x) = \mathrm{D}\,\pi(\bar{x})[\mathrm{grad}_{\overline{\mathcal{M}}}\, \bar{f}(\bar{x})]. \tag{1.24}$$

De plus, la hessienne riemannienne $\mathrm{hess}_{\mathcal{M}}\, f(x)[\xi]$ de f à $x = \pi(\bar{x})$ et $\xi = \mathrm{D}\,\pi(\bar{x})[\bar{\xi}]$ ($\bar{\xi} \in \mathcal{H}_{\bar{x}}$) est donnée par

$$\mathrm{hess}_{\mathcal{M}}\, f(x)[\xi] = \mathrm{D}\,\pi(\bar{x})[P^{\mathcal{H}}_{\bar{x}}(\mathrm{hess}_{\overline{\mathcal{M}}}\, \bar{f}(\bar{x})[\bar{\xi}])]. \tag{1.25}$$

Une rétraction sur $\overline{\mathcal{M}}$ induit une rétraction sur $\mathcal{M}$ si, pour tous les éléments d'une même classe, les rétractions de vecteurs horizontaux correspondants donnent de nouveaux éléments qui appartiennent tous à la même classe. Soient $x \in \mathcal{M}$, $\xi \in T_x\mathcal{M}$ et prenons $\bar{x}, \bar{y} \in \pi^{-1}(x)$ dont les vecteurs horizontaux correspondant à ξ sont notés $\bar{\xi}_{\bar{x}} \in \mathcal{H}_{\bar{x}}$ et $\bar{\xi}_{\bar{y}} \in \mathcal{H}_{\bar{y}}$. La rétraction $\overline{R}$ de $\overline{\mathcal{M}}$ induit une rétraction R sur $\mathcal{M}$ si $\overline{R}_{\bar{x}}(\bar{\xi}_{\bar{x}}) = \overline{R}_{\bar{y}}(\bar{\xi}_{\bar{y}})$ et on a alors

$$R_x(\xi) = \pi\left(\overline{R}_{\bar{x}}(\bar{\xi}_{\bar{x}})\right). \tag{1.26}$$

Une illustration de la rétraction dans le cas des variétés quotient est proposée dans la figure 1.9.

Intéressons nous pour terminer au transport de vecteurs et prenons une rétraction $\overline{R}$ sur $\overline{\mathcal{M}}$ qui induit une rétraction sur $\mathcal{M}$. Étant donné $x = \pi(\bar{x}) \in \mathcal{M}$ et $\xi, \eta \in T_x\mathcal{M}$ dont les vecteurs horizontaux sont $\bar{\xi}$ et $\bar{\eta}$ dans $\mathcal{H}_{\bar{x}}$, un transport de vecteurs approprié $\mathcal{T}(x, \xi, \eta)$ sur $\mathcal{M}$ est représenté par $P^{\mathcal{H}}_{\overline{R}_{\bar{x}}(\bar{\xi})}(\bar{\eta})$, *i.e.*,

$$\mathcal{T}(x, \xi, \eta) = \mathrm{D}\,\pi(\overline{R}_{\bar{x}}(\bar{\xi}))[P^{\mathcal{H}}_{\overline{R}_{\bar{x}}(\bar{\xi})}(\bar{\eta})]. \tag{1.27}$$

Notons qu'ici aussi, $P^{\mathcal{H}}$ correspond par abus à la projection depuis l'espace ambiant de $\overline{\mathcal{M}}$.

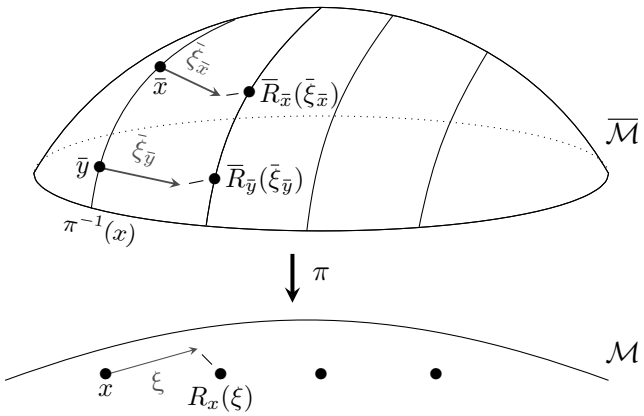

FIGURE *1.9 – Schéma d'illustration d'une rétraction $\overline{R}$ sur $\overline{\mathcal{M}}$ qui induit une rétraction R sur le quotient $\mathcal{M}$. Étant donné $x \in \mathcal{M}$ et $\xi \in T_x\mathcal{M}$, pour tous les éléments de la classe d'équivalence $\pi^{-1}(x)$ ($\bar{x}$ et $\bar{y}$ sur la figure), les rétractions avec $\overline{R}$ des vecteurs horizontaux correspondants à ξ ($\bar{\xi}_{\bar{x}}$ et $\bar{\xi}_{\bar{y}}$ sur la figure) doivent appartenir à la même classe d'équivalence pour que $R_x(\xi)$ soit correctement défini.*

1.5 Variétés riemanniennes d'intérêt

Dans cette section, nous présentons trois variétés riemanniennes que nous utilisons dans la suite. Dans la section 1.5.1, nous traitons la variété des matrices symétriques positives définies de taille $n \times n$ notée $\mathcal{S}_n^{++}$. La géométrie de cette variété joue un rôle important dans le modèle de diagonalisation conjointe approximée comme nous considérons que les matrices à diagonaliser sont dans $\mathcal{S}_n^{++}$ dans ce travail. Cette variété est aussi importante pour nous au niveau de l'optimisation sur l'ensemble des matrices inversibles lorsque nous exploitons la décomposition polaire. Dans la section 1.5.2, nous exposons la variété des matrices orthogonales de taille $n \times n$ notée $\mathcal{O}_n$. Comme pour $\mathcal{S}_n^{++}$, nous utilisons cette variété lorsque nous exploitons la décomposition polaire pour résoudre des problèmes d'optimisation sur l'ensemble des matrices inversibles. Enfin, dans la section 1.5.3, nous présentons la variété des matrices inversibles de taille $n \times n$ qui correspond au groupe général linéaire noté GL_n et qu'on utilisera pour la résolution de problèmes d'optimisation.

1.5.1 Variété des matrices symétriques positives définies $\mathcal{S}_n^{++}$

Commençons par noter qu'un traitement détaillé de cette variété est par exemple disponible dans le livre [20]. L'ensemble des matrices symétriques positives définies est

$$\mathcal{S}_n^{++} = \left\{ \boldsymbol{S} \in \mathcal{S}_n : \boldsymbol{z}^T \boldsymbol{S} \boldsymbol{z} > 0 \text{ pour tout } \boldsymbol{z} \in \mathbb{R}^n, \boldsymbol{z} \neq \boldsymbol{0} \right\}, \tag{1.28}$$

où $\mathcal{S}_n$ est l'ensemble des matrices symétriques de taille $n \times n$ et $\boldsymbol{0}$ est le vecteur nul. La variété $\mathcal{S}_n^{++}$ est ouvert dans $\mathcal{S}_n$ donc son espace tangent $T_{\boldsymbol{S}}\mathcal{S}_n^{++}$ à n'importe quel point $\boldsymbol{S}$ peut être

identifié à $\mathcal{S}_n$. Notons que la projection sur $\mathcal{S}_n$ depuis l'espace ambient $\mathbb{R}^{n\times n}$ est, pour tout $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$,

$$P_{\boldsymbol{S}}^{\mathcal{S}_n^{++}}(\boldsymbol{Z}) = \mathrm{sym}(\boldsymbol{Z}) = \frac{\boldsymbol{Z}+\boldsymbol{Z}^T}{2}. \tag{1.29}$$

Le choix classique de métrique riemannienne sur $\mathcal{S}_n^{++}$ est, pour tous $\boldsymbol{S} \in \mathcal{S}_n^{++}$, $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{S}_n$,

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{S}}^{\mathcal{S}_n^{++}} = \mathrm{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\xi}\boldsymbol{S}^{-1}\boldsymbol{\eta}\right), \tag{1.30}$$

où $\mathrm{tr}(\cdot)$ est l'opérateur trace. Cette métrique est associée à la connection de Levi-Civita dérivée dans [79, annexe B]. Elle est définie, pour $\boldsymbol{S} \in \mathcal{S}_n^{++}$ et les champs de vecteurs $\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\eta}_{\boldsymbol{S}}$ évalués en $\boldsymbol{S}$, par

$$\nabla_{\boldsymbol{\xi}_{\boldsymbol{S}}}^{\mathcal{S}_n^{++}} \boldsymbol{\eta}_{\boldsymbol{S}} = \mathrm{D}\,\boldsymbol{\eta}_{\boldsymbol{S}}[\boldsymbol{\xi}_{\boldsymbol{S}}] - \mathrm{sym}\left(\boldsymbol{\eta}_{\boldsymbol{S}}\boldsymbol{S}^{-1}\boldsymbol{\xi}_{\boldsymbol{S}}\right). \tag{1.31}$$

La géodésique γ telle que $\gamma(0) = \boldsymbol{S} \in \mathcal{S}_n^{++}$ et $\dot{\gamma}(0) = \boldsymbol{\xi} \in \mathcal{S}_n$ est définie pour tout $t \in \mathbb{R}$ par

$$\gamma(t) = \boldsymbol{S}^{1/2} \exp\left(t\boldsymbol{S}^{-1/2}\boldsymbol{\xi}\boldsymbol{S}^{-1/2}\right)\boldsymbol{S}^{1/2}, \tag{1.32}$$

où $\exp(\cdot)$ est l'exponentielle matricielle classique. Notons qu'on peut aussi définir γ telle que $\gamma(0) = \boldsymbol{S}$ et $\gamma(1) = \boldsymbol{C}$ avec

$$\gamma(t) = \boldsymbol{S}^{1/2}\left(\boldsymbol{S}^{-1/2}\boldsymbol{C}\boldsymbol{S}^{-1/2}\right)^t\boldsymbol{S}^{1/2}, \tag{1.33}$$

où $(\cdot)^t$ est la fonction puissance définie avec l'exponentielle et le logarithme matriciels classiques. Il s'en suit que la distance riemannienne naturelle [2] entre $\boldsymbol{S}$ et $\boldsymbol{C}$ dans $\mathcal{S}_n^{++}$ est

$$\delta_{\mathrm{R}}^2(\boldsymbol{S}, \boldsymbol{C}) = \left\|\log\left(\boldsymbol{S}^{-1/2}\boldsymbol{C}\boldsymbol{S}^{-1/2}\right)\right\|_{\mathrm{F}}^2. \tag{1.34}$$

L'exponentielle riemannienne sur $\mathcal{S}_n^{++}$, notée $\exp^{\mathcal{S}_n^{++}}(\cdot)$, découle de la définition des géodésiques et le logarithme riemannien à $\boldsymbol{S} \in \mathcal{S}_n^{++}$ est donné, pour tout $\boldsymbol{C} \in \mathcal{S}_n^{++}$, par

$$\log_{\boldsymbol{S}}^{\mathcal{S}_n^{++}}(\boldsymbol{C}) = \boldsymbol{S}^{1/2}\log\left(\boldsymbol{S}^{-1/2}\boldsymbol{C}\boldsymbol{S}^{-1/2}\right)\boldsymbol{S}^{1/2}. \tag{1.35}$$

Le transport parallèle τ le long de la géodésique γ avec $\tau(0) = \boldsymbol{\eta}$, $\gamma(0) = \boldsymbol{S}$ et $\dot{\gamma}(0) = \boldsymbol{\xi}$ dérivé dans [63] est

$$\tau(t) = \gamma(t)\boldsymbol{S}^{-1}\boldsymbol{\eta}\boldsymbol{S}^{-1}\gamma(t). \tag{1.36}$$

En optimisation, on peut utiliser le transport de vecteurs qui découle de ce transport parallèle ou il est aussi possible d'utiliser simplement $\mathcal{T}(\boldsymbol{S}, \boldsymbol{\xi}, \boldsymbol{\eta}) = \boldsymbol{\eta}$. Considérons enfin une fonction objectif $f : \mathcal{S}_n^{++} \to \mathbb{R}$. Étant donné le gradient euclidien $\mathrm{grad}_{\mathcal{E}}\, f(\boldsymbol{S})$ à $\boldsymbol{S}$ (gradient de f sur $\mathbb{R}^{n\times n}$), le gradient riemannien $\mathrm{grad}_{\mathcal{S}_n^{++}} f(\boldsymbol{S})$ est

$$\mathrm{grad}_{\mathcal{S}_n^{++}} f(\boldsymbol{S}) = \boldsymbol{S}\,\mathrm{sym}\left(\mathrm{grad}_{\mathcal{E}}\, f(\boldsymbol{S})\right)\boldsymbol{S}. \tag{1.37}$$

La hessienne riemannienne $\mathrm{hess}_{\mathcal{S}_n^{++}} f(\boldsymbol{S})[\boldsymbol{\xi}]$ est quant à elle

$$\mathrm{hess}_{\mathcal{S}_n^{++}} f(\boldsymbol{S})[\boldsymbol{\xi}] = \boldsymbol{S}\,\mathrm{sym}\left(\mathrm{hess}_{\mathcal{E}}\, f(\boldsymbol{S})[\boldsymbol{\xi}]\right)\boldsymbol{S} + \mathrm{sym}(\boldsymbol{\xi}\,\mathrm{sym}\left(\mathrm{grad}_{\mathcal{E}}\, f(\boldsymbol{S})\right)\boldsymbol{S}), \tag{1.38}$$

où $\mathrm{hess}_{\mathcal{E}}\, f(\boldsymbol{S})[\boldsymbol{\xi}] = \mathrm{D}\,\mathrm{grad}_{\mathcal{E}}\, f(\boldsymbol{S})[\boldsymbol{\xi}]$ est la hessienne euclidienne de f.

2. Cette distance est considérée comme la distance riemannienne naturelle sur $\mathcal{S}_n^{++}$ car elle correspond à la moyenne dite géométrique [20], très étudiée du fait des propriétés qu'elle possède. Elle a de plus un rôle particulier en traitement du signal et en géométrie de l'information où elle correspond à la métrique de Fisher entre deux distributions normales multivariées [53, 102].

1.5.2 Groupe orthogonal $\mathcal{O}_n$

L'ensemble des informations fournies dans cette section se trouve par exemple dans le livre [2]. La variété des matrices orthogonales est définie par

$$\mathcal{O}_n = \left\{ \boldsymbol{U} \in \mathbb{R}^{n\times n} : \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_n \right\}. \tag{1.39}$$

Cette variété peut être traitée comme une sous-variété de l'espace euclidien $\mathbb{R}^{n\times n}$. Son espace tangent à $\boldsymbol{U} \in \mathcal{O}_n$ est

$$T_{\boldsymbol{U}}\mathcal{O}_n = \left\{ \boldsymbol{\xi} = \boldsymbol{U}\boldsymbol{\Omega} : \boldsymbol{\Omega} \in \mathbb{R}^{n\times n}, \boldsymbol{\Omega}^T = -\boldsymbol{\Omega} \right\}. \tag{1.40}$$

Le groupe orthogonal $\mathcal{O}_n$ hérite de la métrique euclidienne, *i.e*, pour tous $\boldsymbol{\xi}, \boldsymbol{\eta} \in T_{\boldsymbol{U}}\mathcal{O}_n$,

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{U}}^{\mathcal{O}_n} = \operatorname{tr}(\boldsymbol{\xi}^T\boldsymbol{\eta}). \tag{1.41}$$

On peut en déduire le projecteur orthogonal sur $T_{\boldsymbol{U}}\mathcal{O}_n$, donné, pour tout $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$, par

$$P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{U}\operatorname{sym}(\boldsymbol{U}^T\boldsymbol{Z}). \tag{1.42}$$

La connection de Levi-Civita (obtenue par (1.19)) est, pour $\boldsymbol{U} \in \mathcal{O}_n$ et les champs de vecteurs $\boldsymbol{\xi}_{\boldsymbol{U}}, \boldsymbol{\eta}_{\boldsymbol{U}}$ évalués en $\boldsymbol{U}$,

$$\nabla_{\boldsymbol{\xi}_{\boldsymbol{U}}}^{\mathcal{O}_n} \boldsymbol{\eta}_{\boldsymbol{U}} = P_{\boldsymbol{U}}^{\mathcal{O}_n}(\mathrm{D}\,\boldsymbol{\eta}_{\boldsymbol{U}}[\boldsymbol{\xi}_{\boldsymbol{U}}]). \tag{1.43}$$

Il s'en suit que la géodésique γ sur $\mathcal{O}_n$ telle que $\gamma(0) = \boldsymbol{U}$ et $\dot\gamma(0) = \boldsymbol{\xi}$ est définie par

$$\gamma(t) = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{\xi}\end{bmatrix} \exp\left(t \begin{bmatrix} \boldsymbol{U}^T\boldsymbol{\xi} & -\boldsymbol{\xi}^T\boldsymbol{\xi} \\ \boldsymbol{I}_n & \boldsymbol{U}^T\boldsymbol{\xi} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0}_n \end{bmatrix} \exp(-t\boldsymbol{U}^T\boldsymbol{\xi}), \tag{1.44}$$

où $\boldsymbol{0}_n$ est la matrice nulle de taille $n \times n$. Pour l'optimisation, on peut donc utiliser l'exponentielle riemannienne, notée $\exp^{\mathcal{O}_n}(\cdot)$, induite par ces géodésiques comme rétraction. Des rétractions alternatives ont également été proposées, comme

$$R_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{\xi}) = \operatorname{qf}(\boldsymbol{U} + \boldsymbol{\xi}), \qquad \text{ou} \qquad R_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{\xi}) = \operatorname{uf}(\boldsymbol{U} + \boldsymbol{\xi}), \tag{1.45}$$

où $\operatorname{qf}(\cdot)$ et $\operatorname{uf}(\cdot)$ retournent la partie orthogonale de la décomposition QR et polaire, respectivement. Le transport de vecteurs sur $\mathcal{O}_n$ est directement donné par (1.21). Enfin, considérons une fonction objectif $f : \mathcal{O}_n \to \mathbb{R}$. Le gradient riemannien $\operatorname{grad}_{\mathcal{O}_n} f(\boldsymbol{U})$ à $\boldsymbol{U}$ est simplement obtenu par (1.20), *i.e.*, en projetant le gradient euclidien $\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{U})$ sur l'espace tangent de $\boldsymbol{U}$. La hessienne riemannienne est quant à elle

$$\operatorname{hess}_{\mathcal{O}_n} f(\boldsymbol{U})[\boldsymbol{\xi}] = P_{\boldsymbol{U}}^{\mathcal{O}_n}\left(\operatorname{hess}_{\mathcal{E}} f(\boldsymbol{U})[\boldsymbol{\xi}] - \boldsymbol{\xi}\operatorname{sym}\left(\boldsymbol{U}^T \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{U})\right)\right), \tag{1.46}$$

où $\operatorname{hess}_{\mathcal{E}} f(\boldsymbol{U})[\boldsymbol{\xi}]$ est la hessienne euclidienne de f.

1.5.3 Groupe général linéaire GL_n

Les résultats présentés dans cette section proviennent de [13, 76, 80, 109, 115]. L'ensemble des matrices inversibles, ou groupe général linéaire GL_n, est ouvert dans $\mathbb{R}^{n\times n}$ donc l'espace tangent de tout $\boldsymbol{B} \in \mathrm{GL}_n$ peut être identifié à $\mathbb{R}^{n\times n}$. Nous équipons GL_n soit avec une métrique invariante à gauche, soit avec une métrique invariante à droite qui sont respectivement définies pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^{n\times n}$ par

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle^{\ell}_{\boldsymbol{B}} = \operatorname{tr}\left(\boldsymbol{B}^{-1}\boldsymbol{\xi}(\boldsymbol{B}^{-1}\boldsymbol{\eta})^T\right) \qquad \text{et} \qquad \langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle^{r}_{\boldsymbol{B}} = \operatorname{tr}\left(\boldsymbol{\xi}\boldsymbol{B}^{-1}(\boldsymbol{\eta}\boldsymbol{B}^{-1})^T\right). \tag{1.47}$$

Les connections de Levi-Civita associées à ces deux métriques sont, pour $\boldsymbol{B} \in \mathrm{GL}_n$ et les champs de vecteurs $\boldsymbol{\xi}_{\boldsymbol{B}}, \boldsymbol{\eta}_{\boldsymbol{B}}$ évalués en $\boldsymbol{B}$,

$$\begin{aligned}\nabla^{\ell}_{\boldsymbol{\xi}_{\boldsymbol{B}}} \boldsymbol{\eta}_{\boldsymbol{B}} = \mathrm{D}\,\boldsymbol{\eta}_{\boldsymbol{B}}[\boldsymbol{\xi}_{\boldsymbol{B}}] - \frac{1}{2}\left(\boldsymbol{\eta}_{\boldsymbol{B}}\boldsymbol{B}^{-1}\boldsymbol{\xi}_{\boldsymbol{B}} + \boldsymbol{\xi}_{\boldsymbol{B}}\boldsymbol{B}^{-1}\boldsymbol{\eta}_{\boldsymbol{B}}\right) \\ + \boldsymbol{B}\operatorname{sym}\left(\boldsymbol{B}^{-1}\boldsymbol{\xi}_{\boldsymbol{B}}(\boldsymbol{B}^{-1}\boldsymbol{\eta}_{\boldsymbol{B}})^T - (\boldsymbol{B}^{-1}\boldsymbol{\xi}_{\boldsymbol{B}})^T\boldsymbol{B}^{-1}\boldsymbol{\eta}_{\boldsymbol{B}}\right),\end{aligned} \tag{1.48}$$

et

$$\begin{aligned}\nabla^{r}_{\boldsymbol{\xi}_{\boldsymbol{B}}} \boldsymbol{\eta}_{\boldsymbol{B}} = \mathrm{D}\,\boldsymbol{\eta}_{\boldsymbol{B}}[\boldsymbol{\xi}_{\boldsymbol{B}}] - \frac{1}{2}\left(\boldsymbol{\eta}_{\boldsymbol{B}}\boldsymbol{B}^{-1}\boldsymbol{\xi}_{\boldsymbol{B}} + \boldsymbol{\xi}_{\boldsymbol{B}}\boldsymbol{B}^{-1}\boldsymbol{\eta}_{\boldsymbol{B}}\right) \\ + \operatorname{sym}\left((\boldsymbol{\xi}_{\boldsymbol{B}}\boldsymbol{B}^{-1})^T\boldsymbol{\eta}_{\boldsymbol{B}}\boldsymbol{B}^{-1} - \boldsymbol{\xi}_{\boldsymbol{B}}\boldsymbol{B}^{-1}(\boldsymbol{\eta}_{\boldsymbol{B}}\boldsymbol{B}^{-1})^T\right)\boldsymbol{B}.\end{aligned} \tag{1.49}$$

Les géodésiques correspondantes γ_ℓ et γ_r sont définies pour tout $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\xi} \in \mathbb{R}^{n\times n}$ par

$$\gamma_\ell(t) = \boldsymbol{B}\exp\left(t(\boldsymbol{B}^{-1}\boldsymbol{\xi})^T\right)\exp\left(t\left(\boldsymbol{B}^{-1}\boldsymbol{\xi} - (\boldsymbol{B}^{-1}\boldsymbol{\xi})^T\right)\right), \tag{1.50}$$

et

$$\gamma_r(t) = \exp\left(t\left(\boldsymbol{\xi}\boldsymbol{B}^{-1} - (\boldsymbol{\xi}\boldsymbol{B}^{-1})^T\right)\right)\exp\left(t(\boldsymbol{\xi}\boldsymbol{B}^{-1})^T\right)\boldsymbol{B}. \tag{1.51}$$

Les exponentielles riemanniennes qui correspondent à ces géodésiques sont notées $\exp^{\ell}$ et $\exp^{r}$, respectivement. Pour le transport de vecteurs, on peut se contenter de prendre dans les deux cas $\mathcal{T}(\boldsymbol{B}, \boldsymbol{\xi}, \boldsymbol{\eta}) = \boldsymbol{\eta}$ pour tout $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^{n\times n}$. Considérons enfin une fonction objectif f sur GL_n. Étant donné le gradient euclidien $\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})$ à $\boldsymbol{B}$, les gradients riemanniens associés aux métriques invariantes à gauche ou à droite sont

$$\operatorname{grad}_{\ell} f(\boldsymbol{B}) = \boldsymbol{B}\boldsymbol{B}^T\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}) \qquad \text{et} \qquad \operatorname{grad}_{r} f(\boldsymbol{B}) = \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})\boldsymbol{B}^T\boldsymbol{B}. \tag{1.52}$$

Les hessiennes riemanniennes sont quant à elles

$$\begin{aligned}\operatorname{hess}_{\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] = \boldsymbol{B}\boldsymbol{B}^T\operatorname{hess}_{\mathcal{E}} f(\boldsymbol{B})[\boldsymbol{\xi}] + \operatorname{sym}(\boldsymbol{\xi}\boldsymbol{B}^T)\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}) \\ + \boldsymbol{B}\operatorname{sym}(\boldsymbol{B}^{-1}\boldsymbol{\xi}\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})^T\boldsymbol{B}) - \boldsymbol{B}\operatorname{sym}(\boldsymbol{B}^T\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}))\boldsymbol{B}^{-1}\boldsymbol{\xi},\end{aligned} \tag{1.53}$$

et

$$\begin{aligned}\operatorname{hess}_{r} f(\boldsymbol{B})[\boldsymbol{\xi}] = \operatorname{hess}_{\mathcal{E}} f(\boldsymbol{B})[\boldsymbol{\xi}]\boldsymbol{B}^T\boldsymbol{B} + \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})\operatorname{sym}(\boldsymbol{B}^T\boldsymbol{\xi}) \\ + \operatorname{sym}(\boldsymbol{B}\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})^T\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{B} - \boldsymbol{\xi}\boldsymbol{B}^{-1}\operatorname{sym}(\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B})\boldsymbol{B}^T)\boldsymbol{B},\end{aligned} \tag{1.54}$$

où $\operatorname{hess}_{\mathcal{E}} f(\boldsymbol{B})[\boldsymbol{\xi}]$ est la hessienne euclidienne de f.

Chapitre 2

Modèle géométrique de la diagonalisation conjointe approximée

Sommaire

Dans ce chapitre, nous considérons le problème de la diagonalisation conjointe approximée d'un ensemble de matrices symétriques positives définies. Commençons par rappeler qu'étant donné un ensemble $\{\boldsymbol{C}_k\}$ de K matrices symétriques positives définies, l'objectif est de trouver un diagonalisateur conjoint $\boldsymbol{B}$ dans GL_n tel que les matrices de l'ensemble $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\}$ soient aussi diagonales que possible. Il faut donc avant tout définir un critère qui permet de mesurer le degré de diagonalité des matrices de $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\}$. Les deux choix classiques sont les critères de Frobenius (1.4) et log-vraisemblance (1.5) présentés dans le chapitre 1. L'article [9] montre qu'une approche géométrique est possible et propose plusieurs mesures de diagonalités basées sur des divergences de $\mathcal{S}_n^{++}$. Une divergence mesure la dissimilarité entre deux points ; elle généralise la notion de distance, *i.e.*, c'est une fonction positive qui s'annule si et seulement si les deux points sont égaux, mais il n'est pas nécessaire qu'elle soit symétrique et vérifie l'inégalité triangulaire. Notons que les distances au carré sont des divergences. Dans la section 2.1, nous formalisons et généralisons le travail de [9] et nous adoptons un point de vue géométrique pour proposer un modèle des critères du problème de la diagonalisation conjointe approximée de matrices symétriques positives définies. Dans la section 2.2, nous présentons trois stratégies d'optimisation différentes pour résoudre en pratique le problème de diagonalisation conjointe approximée modélisé par un critère suivant le modèle de la section 2.1 : une

stratégie directe classique, une nouvelle stratégie indirecte et une généralisation de l'approche proposée dans [107]. Dans la section 2.3, nous présentons les différentes divergences que nous considérons dans ce travail pour construire des critères de diagonalisation conjointe approximée. Il s'agit de la distance de Frobenius, la divergence de Kullback-Leibler, la divergence log-det α, la distance riemannienne naturelle, la distance log-euclidienne et la distance de Wasserstein. Tous les critères de cette section peuvent être optimisés avec les stratégies de la section 2.2, ce qui offre un large panel de méthodes de diagonalisation conjointe approximée inédites. Dans l'ensemble des travaux précédents, on trouve des solutions algorithmiques seulement pour la distance de Frobenius, la log-vraisemblance qui est liée à la divergence de Kullback-Leibler et la divergence log-det α, mais dans des conditions non optimales. Dans la section 2.4, nous poursuivons le travail de [45] en étudiant quatre propriétés importantes pour les critères de diagonalisation conjointe approximée. La première est liée à la définition de mesure de diagonalité, la deuxième est en rapport avec le diagonalisateur conjoint et les deux dernières concernent les effets de transformations appliquées aux matrices à diagonaliser.

2.1 Modèle des critères

D'un point de vue géométrique, pour mesurer le degré de diagonalité des matrices de $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\}$, il convient de s'intéresser à la position relative des matrices $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ par rapport à l'ensemble des matrices diagonales positives définies noté $\mathcal{D}_n^{++}$, qui est une sous-variété riemannienne géodésiquement fermée de $\mathcal{S}_n^{++}$. Un critère approprié f est donc de la forme

$$f(\boldsymbol{B}, \{\boldsymbol{C}_k\}) = \sum_k w_k \, g\left(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \mathcal{D}_n^{++}\right), \tag{2.1}$$

où les w_k sont des poids positifs et $g(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \mathcal{D}_n^{++})$ mesure la position relative de $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ par rapport à $\mathcal{D}_n^{++}$, dont le minimum est atteint lorsque $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T \in \mathcal{D}_n^{++}$. Pour définir la position relative de $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ par rapport à $\mathcal{D}_n^{++}$, on sélectionne une matrice $\boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{C}_k)$ dans $\mathcal{D}_n^{++}$ et on regarde la position relative de $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ par rapport à $\boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{C}_k)$. Dans la suite, $\boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{C}_k)$ est simplement noté $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ lorsque ce n'est pas ambigu. Pour que le choix de $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ soit satisfaisant, il est impératif que $\boldsymbol{\Lambda}_k(\boldsymbol{B}) = \boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ lorsque $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T \in \mathcal{D}_n^{++}$. On peut alors réécrire la fonction f comme

$$f(\boldsymbol{B}, \{\boldsymbol{C}_k\}) = \sum_k w_k \, h\left(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{\Lambda}_k(\boldsymbol{B})\right), \tag{2.2}$$

où $h(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{\Lambda}_k(\boldsymbol{B}))$ mesure la position relative de $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ par rapport à $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ dont le minimum est atteint lorsque $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T = \boldsymbol{\Lambda}_k(\boldsymbol{B})$. Un schéma illustrant les influences respectives du choix de la fonction $h(\cdot,\cdot)$ et des matrices diagonales cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ se trouve figure 2.1. Le choix naturel pour la fonction $h(\cdot,\cdot)$ est de prendre une divergence $d(\cdot,\cdot)$ sur $\mathcal{S}_n^{++}$. Notons que d'autres possibilités sont envisageables : comme illustré dans la figure 2.2, on peut par exemple choisir $h(\cdot,\cdot)$ comme la mesure de l'angle entre la géodésique reliant $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ à l'identité $\boldsymbol{I}_n$ et celle reliant $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ à l'identité. Étant donné $h(\cdot,\cdot)$, des considérations géométriques donnent un choix naturel pour les matrices diagonales cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$. C'est en effet la matrice $\boldsymbol{\Lambda}$ de

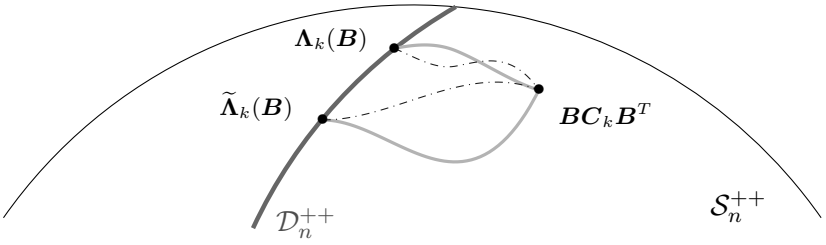

FIGURE *2.1 – Schéma d'illustration de l'approche géométrique du modèle de diagonalisation conjointe approximée et de l'influence du choix de la fonction $h(\cdot,\cdot)$ et des matrices diagonales cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ dans* (2.2)*. La matrice diagonale cible ($\boldsymbol{\Lambda}_k(\boldsymbol{B})$ ou $\widetilde{\boldsymbol{\Lambda}}_k(\boldsymbol{B})$ sur la figure) détermine la destination à atteindre dans $\mathcal{D}_n^{++}$. La fonction $h(\cdot,\cdot)$ (longueur des courbes reliant $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ et $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ ou $\widetilde{\boldsymbol{\Lambda}}_k(\boldsymbol{B})$ sur la figure) définit comment se rapprocher de la matrice diagonale cible. Les courbes continues et en pointillés représentent deux choix possibles pour $h(\cdot,\cdot)$.*

$\mathcal{D}_n^{++}$ qui minimise $h(\cdot,\cdot)$ avec $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$, *i.e.*,

$$\boldsymbol{\Lambda}_k(\boldsymbol{B}) = \underset{\boldsymbol{\Lambda}\in\mathcal{D}_n^{++}}{\operatorname{argmin}} \quad h(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{\Lambda}). \tag{2.3}$$

Nous terminons avec la proposition 2.1 qui contient un résultat théorique concernant la dérivation d'un critère f de la forme (2.2) pour lequel les matrices diagonales cibles sont définies dans (2.3). Cette proposition permet de simplifier le calcul du gradient de f, voire de le rendre possible dans certains cas compliqués où nous n'avons pas de formule en forme fermée pour $\boldsymbol{\Lambda}_k(\boldsymbol{B})$.

Proposition 2.1 (Gradient d'un critère de diagonalisation conjointe approximée)
Lors du calcul du gradient à $\boldsymbol{B} \in \mathrm{GL}_n$ d'un critère f de la forme (2.2) *pour lequel les matrices diagonales cibles sont définies dans* (2.3)*, on peut traiter les matrices $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ comme des constantes, il n'est donc pas nécessaire de les dériver. En effet, étant donné $\boldsymbol{B} \in \mathrm{GL}_n$, fixons $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_k(\boldsymbol{B})$ et définissons $\widehat{f}(\widehat{\boldsymbol{B}}) = \sum_k w_k\, h(\widehat{\boldsymbol{B}}\boldsymbol{C}_k\widehat{\boldsymbol{B}}^T, \boldsymbol{\Lambda}_k)$, alors*

$$\operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}) = \operatorname{grad}_{\mathcal{E}} \widehat{f}(\boldsymbol{B}).$$

Démonstration. Soient $\boldsymbol{M}_k(\boldsymbol{B}) = \boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ et $\pi_k(\boldsymbol{B}) = (\boldsymbol{M}_k(\boldsymbol{B}), \boldsymbol{\Lambda}_k(\boldsymbol{B}))$, alors $f(\boldsymbol{B}) = \sum_k w_k\, h \circ \pi_k(\boldsymbol{B})$. Avec la métrique euclidienne $\langle\cdot,\cdot\rangle^{\mathcal{E}}$, on a pour tout $\boldsymbol{\xi} \in \mathbb{R}^{n\times n}$

$$\langle \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}), \boldsymbol{\xi}\rangle^{\mathcal{E}} = \mathrm{D}\, f(\boldsymbol{B})[\boldsymbol{\xi}] = \sum_k w_k\ \mathrm{D}\, h(\pi_k(\boldsymbol{B}))\left[\mathrm{D}\, \pi_k(\boldsymbol{B})[\boldsymbol{\xi}]\right],$$

par linéarité de la dérivée directionnelle et règle de dérivation des compositions. De plus,

$$\begin{aligned}\mathrm{D}\, \pi_k(\boldsymbol{B})[\boldsymbol{\xi}] &= (\mathrm{D}\, \boldsymbol{M}_k(\boldsymbol{B})[\boldsymbol{\xi}], \mathrm{D}\, \boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}]) \\ &= (\mathrm{D}\, \boldsymbol{M}_k(\boldsymbol{B})[\boldsymbol{\xi}], \boldsymbol{0}_n) + (\boldsymbol{0}_n, \mathrm{D}\, \boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}]).\end{aligned}$$

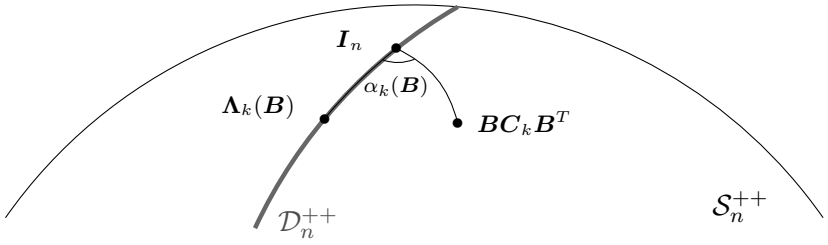

FIGURE *2.2 – Schéma d'illustration du choix d'une mesure d'angle pour la fonction* $h(\cdot,\cdot)$ *de* (2.2) *qui mesure la similarité entre* $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ *et* $\boldsymbol{\Lambda}_k(\boldsymbol{B})$, i.e., $h(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{\Lambda}_k(\boldsymbol{B})) = \alpha_k(\boldsymbol{B})$.

Donc par linéarité,

$$\begin{aligned}\langle \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}), \boldsymbol{\xi}\rangle^{\mathcal{E}} = &\sum_k w_k \, \mathrm{D}\, h(\pi_k(\boldsymbol{B}))\,[(\mathrm{D}\,\boldsymbol{M}_k(\boldsymbol{B})[\boldsymbol{\xi}], \boldsymbol{0}_n)] \\ &+ \sum_k w_k \, \mathrm{D}\, h(\pi_k(\boldsymbol{B}))\,[(\boldsymbol{0}_n, \mathrm{D}\,\boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}])]\,.\end{aligned}$$

On peut montrer que

$$\sum_k w_k \, \mathrm{D}\, h(\pi_k(\boldsymbol{B}))\,[(\mathrm{D}\,\boldsymbol{M}_k(\boldsymbol{B})[\boldsymbol{\xi}], \boldsymbol{0}_n)] = \mathrm{D}\,\widehat{f}(\boldsymbol{B})[\boldsymbol{\xi}] = \langle \operatorname{grad}_{\mathcal{E}} \widehat{f}(\boldsymbol{B}), \boldsymbol{\xi}\rangle^{\mathcal{E}}.$$

Étant donné $\boldsymbol{B}$, en définissant $\bar{h}_k(\boldsymbol{\Lambda}) = h(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{\Lambda})$ pour $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$, on a

$$\begin{aligned}\sum_k w_k \, \mathrm{D}\, h(\pi_k(\boldsymbol{B}))\,[(\boldsymbol{0}_n, \mathrm{D}\,\boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}])] &= \sum_k w_k \, \mathrm{D}\, \bar{h}_k(\boldsymbol{\Lambda}_k(\boldsymbol{B}))\,[\mathrm{D}\,\boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}]] \\ &= \sum_k w_k \, \langle \operatorname{grad}_{\mathcal{E}} \bar{h}_k(\boldsymbol{\Lambda}_k(\boldsymbol{B})), \mathrm{D}\,\boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}]\rangle^{\mathcal{E}}.\end{aligned}$$

Comme $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ correspond par définition au minimum de $\bar{h}_k$, on a $\operatorname{grad}_{\mathcal{E}} \bar{h}_k(\boldsymbol{\Lambda}_k(\boldsymbol{B})) = 0$ pour tout $1 \leq k \leq K$. On a donc finalement

$$\langle \operatorname{grad}_{\mathcal{E}} f(\boldsymbol{B}), \boldsymbol{\xi}\rangle^{\mathcal{E}} = \langle \operatorname{grad}_{\mathcal{E}} \widehat{f}(\boldsymbol{B}), \boldsymbol{\xi}\rangle^{\mathcal{E}},$$

et le résultat est obtenu par identification. □

2.2 Stratégies d'optimisation

Dans cette section, nous présentons trois différentes stratégies pour résoudre le problème de la diagonalisation conjointe approximée dans un cadre général d'optimisation riemannienne. On suppose ici qu'on est en mesure d'obtenir une direction de descente d'une fonction objectif à un point donné à partir de son gradient (et éventuellement hessienne) et qu'on dispose d'une rétraction sur la variété choisie. Ces trois stratégies partagent les premières étapes de

Algorithme 2.1 : Diagonalisation conjointe approximée riemannienne

Input : matrices $\{\boldsymbol{C}_k\}$ dans $\mathcal{S}_n^{++}$, point initial $\boldsymbol{B}_0$ for $\boldsymbol{B}$
Output : itérés $\boldsymbol{B}_i$ pout le diagonalisateur conjoint estimé

1 Calculer les matrices $\{\boldsymbol{B}_0\boldsymbol{C}_k\boldsymbol{B}_0^T\}$ et définir $i = 0$.
2 **while** *convergence non atteinte* **do**
3 Calculer $\boldsymbol{B}_{i+1}$ en utilisant la mise à jour (a), (b) ou (c) sur $\boldsymbol{B}_i$.
4 $i \leftarrow i + 1$

Mise à jour (a): stratégie directe

1 Obtenir une direction de descente $\boldsymbol{\xi}_i$ à partir du gradient (et éventuellement hessienne) de (2.2) à $\boldsymbol{B} = \boldsymbol{B}_i$.
2 Calculer $\boldsymbol{B}_{i+1}$ par la rétraction de $t_i\boldsymbol{\xi}_i$ à $\boldsymbol{B}_i$, où t_i est le pas.

Mise à jour (b): stratégie indirecte

1 Calculer les matrices $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_k(\boldsymbol{B}_i)$.
2 Obtenir une direction de descente $\boldsymbol{\xi}_i$ à partir du gradient (et éventuellement hessienne) de (2.4) à $\widehat{\boldsymbol{B}} = \boldsymbol{B}_i$.
3 Calculer $\boldsymbol{B}_{i+1}$ par la rétraction de $t_i\boldsymbol{\xi}_i$ à $\boldsymbol{B}_i$, où t_i est le pas.

Mise à jour (c): stratégie indirecte inverse

1 Calculer les matrices $\boldsymbol{\Lambda}_k(\boldsymbol{B}_i)$ et définir $\boldsymbol{A}_0 = \boldsymbol{I}_n$.
2 Obtenir une direction de descente $\boldsymbol{\xi}_i$ à partir du gradient (et éventuellement hessienne) de (2.5) à $\boldsymbol{A} = \boldsymbol{A}_0$.
3 Calculer $\boldsymbol{A}_i$ par la rétraction de $t_i\boldsymbol{\xi}_i$ à $\boldsymbol{A}_0$, où t_i est le pas.
4 $\boldsymbol{B}_{i+1} \leftarrow \boldsymbol{A}_i^{-1}\boldsymbol{B}_i$.

l'algorithme 2.1, puis diffèrent dans la façon dont la solution estimée $\boldsymbol{B}$ est mise à jour et dans la manière de traiter la dépendance des matrices diagonales cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ par rapport à $\boldsymbol{B}$. Le critère d'arrêt de l'algorithme 2.1 peut être définie de diverses manières et nous précisons le choix que nous faisons dans nos expériences numériques dans le chapitre correspondant.

Une fois qu'on a choisi une fonction $h(\cdot,\cdot)$ et des matrices diagonales cibles dans (2.2), la façon naturelle et intuitive de résoudre le problème de diagonalisation conjointe approximée est d'optimiser le critère (2.2) directement, en suivant la règle de mise à jour (a) de l'algorithme 2.1. Comme illustré dans la figure 2.3, la mise à jour de $\boldsymbol{B}$ dépend à la fois de la façon dont $h(\cdot,\cdot)$ lie $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ et $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ et des modifications engendrées dans les cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$. Notons que dans le cas où les matrices $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ sont définies par (2.3), la proposition 2.1 peut s'interpréter comme suit : la façon dont $h(\cdot,\cdot)$ lie $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ et $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ suffit pour prédire les changements de $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ au niveau du gradient. On a donc seulement besoin de différencier les fonctions $\boldsymbol{B} \mapsto \boldsymbol{\Lambda}_k(\boldsymbol{B})$ pour calculer la hessienne alors que dans le cas général, il faut les différencier dès le calcul du gradient. Cependant, la différentiation de $\boldsymbol{B} \mapsto \boldsymbol{\Lambda}_k(\boldsymbol{B})$ peut s'avérer compliquée dans certains cas et si on veut tester différentes possibilités pour $\boldsymbol{\Lambda}_k(\boldsymbol{B})$, il faut calculer le gradient et la hessienne de (2.2) dans chacun des cas, ce qui peut s'avérer fastidieux.

Un moyen de dépasser ces limitations est de développer des méthodes où les cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$

sont considérées fixes à chaque itération. Bien qu'elles soient mises à jour à chaque itération, elles sont traitées comme des constantes au moment de dériver le gradient et la hessienne. Une première possibilité pour appliquer ce principe est, étant donné $\boldsymbol{B} \in \mathrm{GL}_n$, de fixer $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_k(\boldsymbol{B})$ et de considérer le sous-problème d'optimisation dont la fonction objectif est

$$\widehat{f}(\widehat{\boldsymbol{B}}) = \sum_k w_k \, h(\widehat{\boldsymbol{B}} \boldsymbol{C}_k \widehat{\boldsymbol{B}}^T, \boldsymbol{\Lambda}_k). \tag{2.4}$$

En partant de $\widehat{\boldsymbol{B}}_0 = \boldsymbol{B}$, on trouve une direction de descente de $\widehat{f}$ qui permet d'obtenir un nouveau point $\widehat{\boldsymbol{B}}$ puis on met $\boldsymbol{B}$ à jour avec $\boldsymbol{B} \leftarrow \widehat{\boldsymbol{B}}$. Cette stratégie d'optimisation correspond à la règle de mise à jour (b) de l'algorithme 2.1. Comme illustré dans la figure 2.3, à chaque itération, on rapproche les matrices $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ de $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ selon $h(\cdot,\cdot)$ mais sans tenir compte du mouvement des cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ le long de $\mathcal{D}_n^{++}$. Notons que dans le cas où les matrices $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ sont choisies comme les matrices diagonales les plus proches des $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ selon $h(\cdot,\cdot)$, *i.e.*, définies par (2.3), la proposition 2.1 montre que les méthodes de gradient employant la stratégie (b) sont identiques à celles employant la stratégie (a). Cela n'est plus vrai à partir du moment où on exploite l'information fournie par la hessienne.

Une autre solution est d'utiliser l'approche proposée dans [107] et de l'adapter au contexte générique de l'optimisation riemannienne. À chaque itération, on fixe $\boldsymbol{B} \in \mathrm{GL}_n$ et on considère le sous-problème d'optimisation dont la fonction objectif est

$$\widetilde{f}(\boldsymbol{A}) = \sum_k w_k \, h(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T, \boldsymbol{A}\boldsymbol{\Lambda}_k(\boldsymbol{B})\boldsymbol{A}^T). \tag{2.5}$$

En partant de $\boldsymbol{A}_0 = \boldsymbol{I}_n$, on définit une direction de descente de $\widetilde{f}$ qui permet d'otenir un nouveau point $\boldsymbol{A}$ puis on met $\boldsymbol{B}$ à jour avec $\boldsymbol{B} \leftarrow \boldsymbol{A}^{-1}\boldsymbol{B}$, comme décrit dans la règle de mise à jour (c) de l'algorithme 2.1. Dans ce cas de figure aussi, les matrices $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ se rapprochent des cibles $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ en négligeant le mouvement de ces dernières dans $\mathcal{D}_n^{++}$. La différence est que nous commençons par déplacer les matrices $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ dans $\mathcal{S}_n^{++}$ en direction des $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ avec la matrice $\boldsymbol{A}$ puis nous inversons la transformation pour que chaque $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ se rapproche de $\boldsymbol{\Lambda}_k(\boldsymbol{B})$, comme montré dans la figure 2.3. Avec cette stratégie d'optimisation, on peut en pratique effectuer l'optimisation sur GL_n directement. L'ajout de contraintes supplémentaires n'est en effet pas nécessaire du fait que $\boldsymbol{A}$ est initialisée avec l'identité puis inversée à chaque itération, ce qui permet de conserver un conditionnement de $\boldsymbol{B}$ satisfaisant et d'éviter les solutions dégénérées. Par contre, comme $\boldsymbol{A}$ est réinitialisée à chaque itération, il n'est pas possible d'utiliser des méthodes d'optimisation où les directions de descente des itérés précédents sont utilisées, *e.g.*, gradient conjugué ou quasi-Newton.

Il est souhaitable de pouvoir prouver la convergence de l'algorithme 2.1 pour les trois règles de mise à jour et de déterminer si les solutions obtenues sont des minimums globaux (ou seulement locaux) du problème. Ces questions ne sont cependant pas triviales, en particulier pour les règles (b) et (c) où l'on considère des sous-problèmes d'optimisation à chaque itération [1].

1. Les règles (b) et (c) sont peu conventionnelles d'un point de vue optimisation. Nous les considérons tout de même car, comme montré dans [107], la règle (c) permet d'obtenir de meilleures performances avec le critère (1.4) basé sur la distance de Frobenius. Notre but est ici de montrer que cette approche peut être généralisée pour l'ensemble des critères considérés. De plus, la règle (b) nous sert à montrer qu'en pratique, considérer des matrices cibles fixes à chaque itération est aussi possible de façon plus directe.

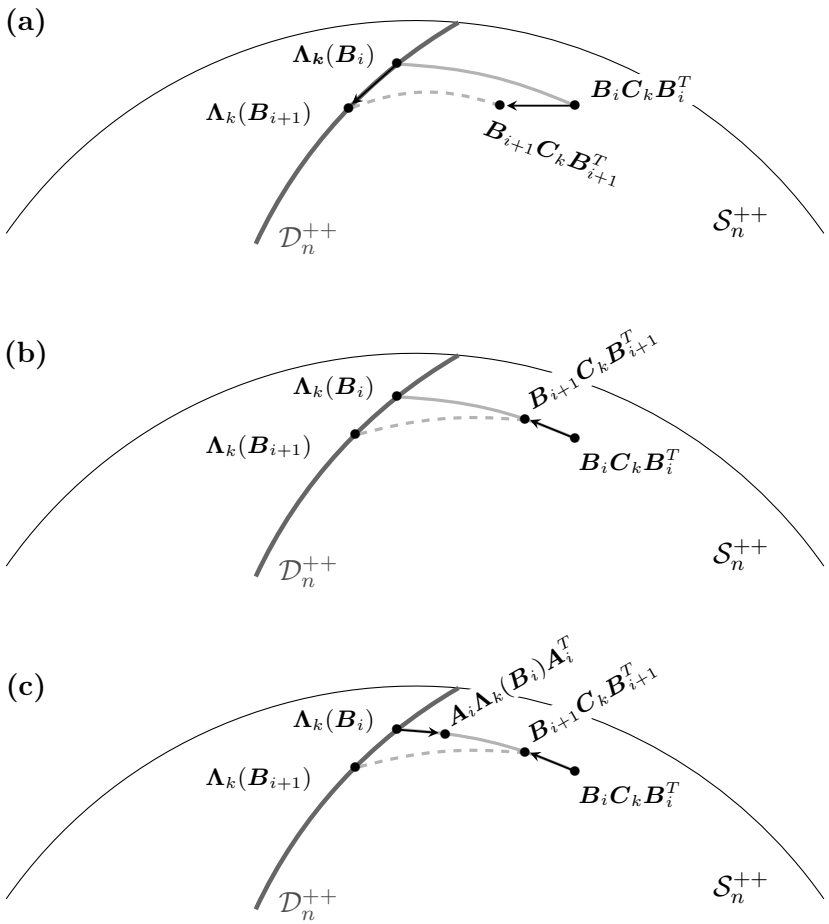

FIGURE *2.3 – Schéma d'illustration des différentes stratégies d'optimisation de l'algorithme 2.1. (a) correspond à la mise à jour (a), pour laquelle on prend en compte la façon dont* $\boldsymbol{B}_i\boldsymbol{C}_k\boldsymbol{B}_i^T$ *et* $\mathbf{\Lambda}_k(\boldsymbol{B}_i)$ *sont liés par* $h(\cdot,\cdot)$ *et les mouvements des matrices diagonales cibles pour calculer le nouvel itéré* $\boldsymbol{B}_{i+1}$*. (b) représente la mise à jour (b) où* $\boldsymbol{B}_i$ *est modifié de sorte que les matrices* $\boldsymbol{B}_{i+1}\boldsymbol{C}_k\boldsymbol{B}_{i+1}^T$ *se rapprochent des cibles* $\mathbf{\Lambda}_k(\boldsymbol{B}_i)$ *selon* $h(\cdot,\cdot)$*, sans tenir compte des changements résultants dans les matrices diagonales cibles. (c) illustre la mise à jour (c) où l'on commence par définir* $\boldsymbol{A}_i$ *tel que les matrices* $\boldsymbol{A}_i\mathbf{\Lambda}_k(\boldsymbol{B}_i)\boldsymbol{A}_i^T$ *se rapprochent de* $\boldsymbol{B}_i\boldsymbol{C}_k\boldsymbol{B}_i^T$ *puis nous prenons* $\boldsymbol{B}_{i+1} = \boldsymbol{A}_i^{-1}\boldsymbol{B}_i$ *pour se rapprocher des cibles* $\mathbf{\Lambda}_k(\boldsymbol{B}_i)$*. Ici aussi les modifications des matrices diagonales cibles ne sont pas prises en compte.*

Même dans le cas le plus simple de la règle (a), ces questions sont liées à des propriétés des fonctions objectifs dans la variété utilisée comme la convexité géodésique et requièrent donc une étude spécifique poussée des fonctions $h(\cdot,\cdot)$, des matrices diagonales cibles et des géodésiques sur GL_n considérées, ce qui sort du cadre de ce travail. Notons tout de même que la convergence numérique est observée pour les différents critères que nous considérons. Remarquons que dans le cas général, ces trois règles de mise à jour peuvent générer des itérés différents et donc en principe atteindre des solutions différentes, ce qui est aussi observé en pratique.

2.3 Divergences considérées

Pour construire des critères de la forme (2.2), (2.4) et (2.5) qui vont nous permettre de résoudre le problème de diagonalisation conjointe approximée en pratique, nous considérons plusieurs divergences, dont des distances au carré, pour définir $h(\cdot,\cdot)$. Pour chacune d'entre elles, nous donnons sa définition, expliquons son intérêt et définissons la matrice diagonale la plus proche, *i.e.*, celle qui correspond à la solution de (2.3). Dans l'ensemble des cas, le critère (2.2) utilisé avec la stratégie d'optimisation directe est défini avec les matrices diagonales cibles les plus proches. Pour simplifier la lecture, les détails techniques de l'étude des gradients et des hessiennes de tous les critères sont rassemblés dans l'annexe B.

2.3.1 Distance de Frobenius

La distance de Frobenius entre les matrices $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ est définie par

$$\delta_{\mathrm{F}}^2(\boldsymbol{M},\boldsymbol{\Lambda}) = \|\boldsymbol{M}-\boldsymbol{\Lambda}\|_{\mathrm{F}}^2. \tag{2.6}$$

C'est la distance que l'on obtient lorsqu'on équipe $\mathcal{S}_n^{++}$ avec la métrique euclidienne. Elle est donc définie sur tout $\mathcal{S}_n$. Comme expliqué dans le chapitre 1, cette distance donne le critère (1.4) de diagonalisation conjointe le plus étudié du fait de sa simplicité. Ce critère vient en effet du raisonnement pratique suivant : pour diagonaliser des matrices, il suffit de minimiser leurs éléments hors-diagonaux. La matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de la matrice $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est

$$\boldsymbol{\Lambda} = \mathrm{ddiag}(\boldsymbol{M}). \tag{2.7}$$

Les critères de la forme (2.2), (2.4) et (2.5) exploitant la distance de Frobenius (au carré) sont rexpectivement notés f_{F}, $\widehat{f}_{\mathrm{F}}$ et $\widetilde{f}_{\mathrm{F}}$. Le critère $\widetilde{f}_{\mathrm{F}}$ associé à la stratégie d'optimisation indirecte inverse (règle de mise à jour (c) de l'algorithme 2.1) correspond à la méthode proposée dans [107] au détail près de la technique d'optimisation employée.

2.3.2 Divergence de Kullback-Leibler

La divergence de Kullback-Leibler entre deux distributions gaussiennes avec les matrices de covariances $\boldsymbol{P}$ et $\boldsymbol{S}$ est définie par

$$d_{\mathrm{KL}}(\boldsymbol{P},\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{S}^{-1} - \boldsymbol{I}_n) - \log\det(\boldsymbol{P}\boldsymbol{S}^{-1}). \tag{2.8}$$

L'intérêt de cette divergence provient des informations statistiques qu'elle contient et de ses liens avec l'information mutuelle et la vraisemblance (voir par exemple [39, 95] et le chapitre 1). Du fait que cette divergence n'est pas symétrique par rapport à ses arguments, elle est à l'origine de plusieurs mesures de diagonalité. La première, qu'on appelle divergence de Kullback-Leibler gauche, est définie pour les matrices $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ par

$$d_{\ell\mathrm{KL}}(\boldsymbol{M},\boldsymbol{\Lambda}) = d_{\mathrm{KL}}(\boldsymbol{M},\boldsymbol{\Lambda}). \tag{2.9}$$

Dans ce cas, [9] a montré que la matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est

$$\boldsymbol{\Lambda} = \mathrm{ddiag}(\boldsymbol{M}). \tag{2.10}$$

Les critères de la forme (2.2), (2.4) et (2.5) exploitant la divergence de Kullback-Leibler gauche sont notés $f_{\ell\mathrm{KL}}$, $\widehat{f}_{\ell\mathrm{KL}}$ et $\widetilde{f}_{\ell\mathrm{KL}}$. Remarquons que $f_{\ell\mathrm{KL}}$ est en fait le critère log-vraissemblance (1.5) historique [2]. La deuxième est la divergence de Kullback-Leibler droite, qui est définie pour les matrices $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ par

$$d_{r\mathrm{KL}}(\boldsymbol{M},\boldsymbol{\Lambda}) = d_{\mathrm{KL}}(\boldsymbol{\Lambda},\boldsymbol{M}). \tag{2.11}$$

Pour cette mesure, [9] a montré que la matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est

$$\boldsymbol{\Lambda} = \mathrm{ddiag}(\boldsymbol{M}^{-1})^{-1}. \tag{2.12}$$

Les critères de la forme (2.2), (2.4) et (2.5) exploitant la divergence de Kullback-Leibler droite sont notés $f_{r\mathrm{KL}}$, $\widehat{f}_{r\mathrm{KL}}$ et $\widetilde{f}_{r\mathrm{KL}}$. Il est également possible de considérer une version symétrisée de cette divergence, comme fait par exemple dans [82], qu'on peut simplement définir pour les matrices $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ par

$$d_{s\mathrm{KL}}(\boldsymbol{M},\boldsymbol{\Lambda}) = \frac{1}{2}(d_{\mathrm{KL}}(\boldsymbol{M},\boldsymbol{\Lambda}) + d_{\mathrm{KL}}(\boldsymbol{\Lambda},\boldsymbol{M})). \tag{2.13}$$

La matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$, trouvée dans [9], est cette fois-ci

$$\boldsymbol{\Lambda} = \mathrm{ddiag}(\boldsymbol{M})^{1/2}\,\mathrm{ddiag}(\boldsymbol{M}^{-1})^{-1/2}. \tag{2.14}$$

Il est intéressant de noter qu'elle correspond à la moyenne géométrique des matrices diagonales les plus proches selon les divergences de Kullback-Leibler gauche et droite [3]. Les critères de la forme (2.2), (2.4) et (2.5) exploitant la divergence de Kullback-Leibler symétrisée sont notés $f_{s\mathrm{KL}}$, $\widehat{f}_{s\mathrm{KL}}$ et $\widetilde{f}_{s\mathrm{KL}}$.

2. Il est intéressant de noter que les deux critères usuels (Frobenius et Kullback-Leibler gauche), sont les deux seuls critères considérés ici pour lesquels les matrices diagonales les plus proches de $\boldsymbol{M}$ sont données simplement par la partie diagonale de $\boldsymbol{M}$.

3. Il existe une relation similaire pour les moyennes d'un ensemble de matrices symétriques positives définies : la moyenne pour la divergence de Kullback-Leibler symétrisée est la moyenne géométrique des moyennes pour les divergences de Kullback-Leibler gauche et droite, qui correspondent respectivement à la moyenne arithmétique et harmonique [82].

2.3.3 Divergence log-det α

La divergence log-det α [36] entre $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ est

$$d_{\alpha\mathrm{LD}}(\boldsymbol{M},\boldsymbol{\Lambda}) = \frac{4}{1-\alpha^2}\log\frac{\det(\frac{1-\alpha}{2}\boldsymbol{M}+\frac{1+\alpha}{2}\boldsymbol{\Lambda})}{\det(\boldsymbol{M})^{\frac{1-\alpha}{2}}\det(\boldsymbol{\Lambda})^{\frac{1+\alpha}{2}}}, \tag{2.15}$$

pour $\alpha \in]-1,1[$. Notons que cette divergence n'est symétrique que pour $\alpha = 0$, mais nous avons la relation $d_{\alpha\mathrm{LD}}(\boldsymbol{M},\boldsymbol{\Lambda}) = d_{-\alpha\mathrm{LD}}(\boldsymbol{\Lambda},\boldsymbol{M})$ et il n'est donc pas nécessaire de séparer différents cas comme pour la divergence de Kullback-Leibler dans la section précédente. Cette divergence a récemment été utilisée pour la diagonalisation conjointe approximée dans [9]. La propriété très intéressante de cette divergence est qu'on a un continuum en α et que pour $\alpha \to -1$ et $\alpha \to 1$, elle coïncide avec les mesures de Kullback-Leibler gauche et droite présentées précédemment. De plus, le cas $\alpha = 0$ donne la distance de Bhattacharyya [36, 82], aussi appelée S-divergence [104]. Cette distance est particulièrement importante car elle est proche de la distance riemannienne naturelle de $\mathcal{S}_n^{++}$ [82, 104], tout en étant numériquement moins coûteuse à évaluer. Ce lien est particulièrement bien illustré par la proposition 4 de [82] qui établit qu'on peut retrouver la métrique classique (1.30) de $\mathcal{S}_n^{++}$ en dérivant la distance de Bhattacharyya, *i.e.*, pour tous $\boldsymbol{M} \in \mathcal{S}_n^{++}$, $\boldsymbol{\xi},\boldsymbol{\eta} \in \mathcal{S}_n$,

$$\frac{\partial^2}{\partial s \partial t} d_{0\mathrm{LD}}(\boldsymbol{M},\boldsymbol{M}+t\boldsymbol{\xi}+s\boldsymbol{\eta})\Big|_{t=s=0} = \frac{1}{4}\operatorname{tr}(\boldsymbol{M}^{-1}\boldsymbol{\xi}\boldsymbol{M}^{-1}\boldsymbol{\eta}), \tag{2.16}$$

ce qui montre que la distance de Bhattacharyya se comporte comme la distance riemannienne naturelle pour des matrices suffisamment proches. Comme montré dans [9], la matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est l'unique solution de

$$\operatorname{ddiag}\left(\left(\frac{1-\alpha}{2}\boldsymbol{M}+\frac{1+\alpha}{2}\boldsymbol{\Lambda}\right)^{-1}\right) = \boldsymbol{\Lambda}^{-1}. \tag{2.17}$$

Une méthode pour résoudre cette équation est donnée dans l'annexe A. Les critères de la forme (2.2), (2.4) et (2.5) exploitant la divergence log-det α sont notés $f_{\alpha\mathrm{LD}}$, $\widehat{f}_{\alpha\mathrm{LD}}$ et $\widetilde{f}_{\alpha\mathrm{LD}}$. Nous ne donnons pas de méthode pour calculer en pratique la hessienne de $f_{\alpha\mathrm{LD}}$ et nous nous contentons donc des méthodes d'optimisation exploitant le gradient pour ce critère.

2.3.4 Distance riemannienne naturelle

Comme introduit dans le chapitre 1, la distance riemannienne naturelle sur $\mathcal{S}_n^{++}$ entre $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ est définie par

$$\delta_{\mathrm{R}}^2(\boldsymbol{M},\boldsymbol{\Lambda}) = \left\|\log(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}\boldsymbol{\Lambda}^{-1/2})\right\|_{\mathrm{F}}^2. \tag{2.18}$$

Cette distance est obtenue en adoptant un point de vue purement géométrique (voir *e.g.*, [20, 50, 81]), où elle correspond à la longueur du chemin le plus court reliant deux points selon la métrique (1.30), et du point de vue de la géométrie de l'information dans le cas des distributions normales multivariées en utilisant la métrique de Fisher [53, 102], introduite dans les articles

fondateurs [10, 99]. Cette distance a donc un intérêt majeur lorsqu'on manipule des matrices symétriques positives définies, d'autant plus dans le cadre de méthodes d'analyse de données qui exploitent les statistiques d'ordre deux. La matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$, dérivée dans [9], est l'unique solution de

$$\operatorname{ddiag}(\log(\boldsymbol{M}^{-1}\boldsymbol{\Lambda})) = \mathbf{0}_n. \tag{2.19}$$

Une méthode pour résoudre cette équation est donnée dans l'annexe A. Les critères de la forme (2.2), (2.4) et (2.5) exploitant la distance riemannienne naturelle sont notés f_{R}, $\widehat{f}_{\mathrm{R}}$ et $\widetilde{f}_{\mathrm{R}}$. Nous ne proposons pas de méthode pour calculer en pratique la hessienne de f_{R} et nous nous limitons donc aux méthodes de gradient pour optimiser ce critère. Pour calculer cette distance et les gradients des critères basés dessus, on a recours au logarithme matriciel. De plus, pour les hessiennes, on a besoin de la dérivée première du logarithme matriciel. Des méthodes pour évaluer numériquement le logarithme matriciel et sa dérivée première sont présentées dans l'annexe C.

2.3.5 Distance log-euclidienne

La distance log-euclidienne [15, 50] entre $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ est définie par

$$\delta_{\mathrm{LE}}^2(\boldsymbol{M}, \boldsymbol{\Lambda}) = \|\log(\boldsymbol{M}) - \log(\boldsymbol{\Lambda})\|_{\mathrm{F}}^2. \tag{2.20}$$

Si $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ commutent, la distance log-euclidienne est équivalente à la distance riemannienne naturelle. Elle peut être vue comme une linéarisation de la distance riemannienne naturelle autour de l'identité $\boldsymbol{I}_n$. En effet, elle correspond à la distance de Frobenius de la projection de $\boldsymbol{M}$ et de $\boldsymbol{\Lambda}$ dans l'espace tangent de $\boldsymbol{I}_n$. Comme cet espace tangent est une linéarisation de $\mathcal{S}_n^{++}$ autour de l'identité, sa distance naturelle est celle de Frobenius. La matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est

$$\boldsymbol{\Lambda} = \exp(\operatorname{ddiag}(\log(\boldsymbol{M}))). \tag{2.21}$$

Les critères de la forme (2.2), (2.4) et (2.5) exploitant la distance log-euclidienne sont notés f_{LE}, $\widehat{f}_{\mathrm{LE}}$ et $\widetilde{f}_{\mathrm{LE}}$. Il est intéressant de remarquer que f_{LE} peut également être obtenue en projetant $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ dans l'espace tangent de $\boldsymbol{I}_n$, ce qui est équivalent à projeter $\boldsymbol{C}_k$ dans l'espace tangent de $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$, puis d'utiliser la distance de Frobenius. Pour calculer cette distance, nous utilisons le logarithme matriciel. Pour les gradients des critères basés dessus, on a besoin de la dérivée première du logarithme matriciel. Sa dérivée seconde est également requise pour les hessiennes. Dans l'annexe C, nous présentons des méthodes estimer le logarithme matriciel et sa dérivée première, et nous proposons une nouvelle méthode pour évaluer numériquement sa dérivée seconde.

2.3.6 Distance de Wasserstein

Enfin, la distance de Wasserstein entre $\boldsymbol{M}$ et $\boldsymbol{\Lambda}$ est donnée par

$$\delta_{\mathrm{W}}^2(\boldsymbol{M}, \boldsymbol{\Lambda}) = \operatorname{tr}\left(\frac{1}{2}(\boldsymbol{M} + \boldsymbol{\Lambda}) - (\boldsymbol{\Lambda}^{1/2}\boldsymbol{M}\boldsymbol{\Lambda}^{1/2})^{1/2}\right). \tag{2.22}$$

Cette distance a l'avantage d'être définie sur l'ensemble des matrices symétriques positives semi-définies. Elle joue un rôle important dans le cadre du transport optimal [48, 89, 111] et s'est montrée utile dans l'étude des matrices de covariance [87, 105] et pour l'analyse spectrale de séries temporelles [64]. Cette distance correspond à une structure riemannienne sur la variété $\mathcal{S}_n^{++}$, qui est décrite de manière complète dans [21]. La matrice diagonale $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ la plus proche de $\boldsymbol{M} \in \mathcal{S}_n^{++}$ est l'unique solution de

$$\operatorname{ddiag}\left((\boldsymbol{\Lambda}^{1/2}\boldsymbol{M}\boldsymbol{\Lambda}^{1/2})^{1/2}\right) = \boldsymbol{\Lambda}. \tag{2.23}$$

La preuve de ce résultat et une méthode pour résoudre cette équation sont donnés dans l'annexe A. Les critères de la forme (2.2), (2.4) et (2.5) exploitant la distance de Wasserstein sont notés f_{W}, $\widehat{f}_{\mathrm{W}}$ et $\widetilde{f}_{\mathrm{W}}$. Nous ne donnons pas de méthode pour calculer en pratique la hessienne de f_{W} et nous nous contentons donc des méthodes de gradient pour l'optimisation de ce critère. Bien que cette distance soit définie sur l'ensemble des matrices symétriques positives semi-définies, nous considérons que les matrices à diagonaliser sont positives définies pour calculer les gradients et les hessiennes de f_{W}, $\widehat{f}_{\mathrm{W}}$ et $\widetilde{f}_{\mathrm{W}}$.

2.4 Propriétés d'intérêt des critères

Une conséquence intéressante de l'approche géométrique du problème de diagonalisation conjointe approximée est qu'elle permet d'inférer plusieurs propriétés souhaitables selon le critère utilisé. Nous présentons quatre propriétés d'intérêt dans ce manuscript. La première est liée à la définition de mesure de diagonalité d'une matrice, la deuxième concerne le diagonalisateur conjoint $\boldsymbol{B}$ et les deux dernières sont en rapport avec les effets de transformations appliquées à l'ensemble $\{\boldsymbol{C}_k\}$ dans $\mathcal{S}_n^{++}$. Pour chacune d'entre elles, nous commençons par l'énoncer puis nous déterminons les conditions pour qu'un critère f de la forme (2.2) la vérifie et nous montrons enfin comment construire un critère qui la vérifie à partir d'un critère quelconque.

2.4.1 Rééchelonnement des matrices d'entrée

La première propriété est issue de la constatation qu'une mesure de diagonalité ne devrait pas dépendre de l'échelle des matrices d'entrée $\boldsymbol{C}_k$ étant donné que celle-ci s'applique uniformément sur l'ensemble des éléments diagonaux et hors diagonaux d'une matrice.

Propriété 2.1 (Invariance par rééchelonnement des matrices d'entrée)
Le critère f est dit invariant par rééchelonnement des matrices d'entrée $\boldsymbol{C}_k$ si pour tous scalaires a_k strictement positifs, on a

$$f(\boldsymbol{B}, \{\boldsymbol{C}_k\}) = f(\boldsymbol{B}, \{a_k\boldsymbol{C}_k\}).$$

Une illustration de cette propriété est proposée dans la figure 2.4. En remarquant que les

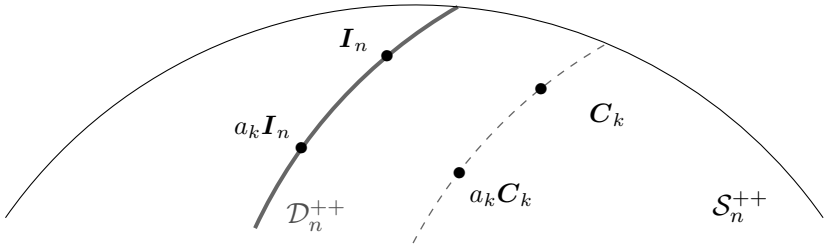

FIGURE *2.4 – Schéma d'illustration de la propriété 2.1 d'invariance par rééchelonnement des matrices d'entrée. Le changement de* $\boldsymbol{C}_k$ *à* $a_k\boldsymbol{C}_k$*, où* $a_k > 0$*, induit un déplacement parallèle à* $\mathcal{D}_n^{++}$ *où* $\boldsymbol{I}_n$ *est remplacé par* $a_k\boldsymbol{I}_n$*. La position relative des matrices d'entrée par rapport à* $\mathcal{D}_n^{++}$ *reste donc inchangée.*

matrices diagonales cibles vérifient $\boldsymbol{\Lambda}(\boldsymbol{B}, a_k\boldsymbol{C}_k) = a_k\boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{C}_k)$, un critère f de la forme (2.2) possède cette propriété si, pour tous $\boldsymbol{M} \in \mathcal{S}_n^{++}$, $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ et $a > 0$,

$$h(a\boldsymbol{M}, a\boldsymbol{\Lambda}) = h(\boldsymbol{M}, \boldsymbol{\Lambda}). \tag{2.24}$$

Étant donné un critère f, on peut construire simplement un critère assuré de posséder la propriété 2.1 par

$$\bar{f}(\boldsymbol{B}, \{\boldsymbol{C}_k\}) = f(\boldsymbol{B}, \{a(\boldsymbol{C}_k)\boldsymbol{C}_k\}), \tag{2.25}$$

où $a(\cdot)$ est une fonction qui permet de fixer l'échelle de $a(\boldsymbol{C}_k)\boldsymbol{C}_k$, par exemple l'inverse de la trace ou de la norme de $\boldsymbol{C}_k$. Une telle normalisation ad hoc des matrices à diagonaliser peut s'avérer nécessaire dans certaines applications quand on utilise un critère qui ne vérifie pas cette propriété, comme c'est le cas pour l'électroencéphalographie avec le critère de Frobenius [41].

2.4.2 Changement d'échelle diagonale

La deuxième propriété d'intérêt vient du fait que la solution du problème de diagonalisation conjointe approximée est toute la classe d'équivalence (1.6), comme expliqué dans le chapitre 1 : les matrices $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\Sigma}\boldsymbol{B}$ sont des solutions équivalentes pour toute matrice diagonale inversible $\boldsymbol{\Sigma}$, dont l'ensemble est noté $\mathcal{D}_n^*$, et il est donc souhaitable que le critère f prenne les mêmes valeurs à $\boldsymbol{B}$ et $\boldsymbol{\Sigma}\boldsymbol{B}$, ce qui se traduit par la propriété suivante (qui est illustrée dans la figure 2.5) :

Propriété 2.2 (Invariance par changement d'échelle diagonale)
Le critère f *est dit invariant par changement d'échelle diagonale si, pour tous* $\boldsymbol{B} \in \mathrm{GL}_n$ *et* $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$*, on a*

$$f(\boldsymbol{B}) = f(\boldsymbol{\Sigma}\boldsymbol{B}).$$

En remarquant que les matrices diagonales cibles vérifient $\boldsymbol{\Lambda}_k(\boldsymbol{\Sigma}\boldsymbol{B}) = \boldsymbol{\Sigma}\boldsymbol{\Lambda}_k(\boldsymbol{B})\boldsymbol{\Sigma}$, un critère f de la forme (2.2) possède cette propriété si on a, pour tous $\boldsymbol{M} \in \mathcal{S}_n^{++}$, $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ et $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$,

$$h(\boldsymbol{\Sigma}\boldsymbol{M}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma}) = h(\boldsymbol{M}, \boldsymbol{\Lambda}). \tag{2.26}$$

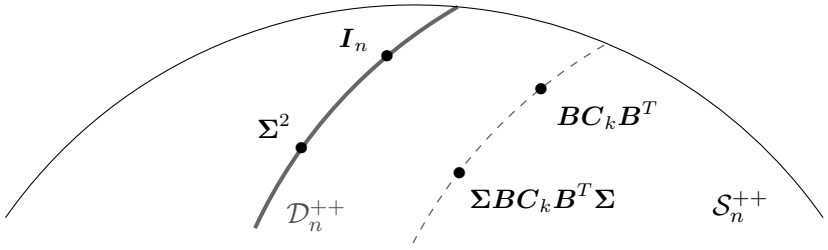

FIGURE *2.5 – Schéma d'illustration de la propriété 2.2 d'invariance par changement d'échelle diagonale. Le changement de* $\boldsymbol{B}$ *à* $\boldsymbol{\Sigma}\boldsymbol{B}$*, où* $\boldsymbol{B} \in \mathrm{GL}_n$ *et* $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$*, induit un déplacement parallèle à* $\mathcal{D}_n^{++}$ *où* $\boldsymbol{I}_n$ *est remplacé par* $\boldsymbol{\Sigma}^2$*. Un tel changement ne permet donc pas de se rapprocher de* $\mathcal{D}_n^{++}$*.*

Remarquons que dans le cas des critères suivant le modèle (2.2), la propriété 2.2 permet d'assurer que la propriété 2.1 est vérifiée. Étant donné un critère f de la forme (2.2), on peut construire un critère assuré de posséder la propriété 2.2 de deux façons différentes. La première possibilité, qui est une généralisation de la modification proposée dans [4] pour le critère de Frobenius, est de définir

$$\bar{f}(\boldsymbol{B}) = \sum_k w_k \, h(\boldsymbol{C}_k, \boldsymbol{B}^{-1}\boldsymbol{\Lambda}_k(\boldsymbol{B})\boldsymbol{B}^{-T}). \tag{2.27}$$

Les fonctions de coût de la forme (2.27) sortent du cadre du modèle (2.2). Comme illustré dans la figure 2.6, on peut les interpréter de façon géométrique comme suit : l'objectif est de trouver le sous-espace $\boldsymbol{B}^{-1}\mathcal{D}_n^{++}\boldsymbol{B}^{-T}$ qui correspond le mieux à l'ensemble de matrices d'entrées $\{\boldsymbol{C}_k\}$. Plutôt que faire bouger les matrices à diagonaliser dans $\mathcal{S}_n^{++}$, on choisit dans ce cas de faire bouger $\mathcal{D}_n^{++}$ pour coller aux données. Notons que dans cette situation, la propriété 2.2 ne permet pas de garantir que la propriété 2.1 est vérifiée. Bien que très intéressant, nous n'étudions pas plus en détail ce modèle pour la diagonalisation conjointe approximée dans ce travail de thèse. La seconde modification du critère f possible, qui généralise celle du critère de Frobenius dans [9], est quant à elle

$$\bar{f}(\boldsymbol{B}) = \sum_k w_k \, h(\boldsymbol{\Lambda}_k(\boldsymbol{B})^{-1/2}\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T\boldsymbol{\Lambda}_k(\boldsymbol{B})^{-1/2}, \boldsymbol{I}_n). \tag{2.28}$$

Cette fois-ci, la modification proposée dans (2.28) reste dans le modèle (2.2). Notons que dans les deux cas, le choix naturel pour les matrices diagonales cibles est susceptible d'être modifié. Il est également intéressant de remarquer que le domaine de définition par rapport aux matrices $\boldsymbol{C}_k$ de la fonction modifiée dans (2.27) reste identique à celui de f. En contraste, les matrices $\boldsymbol{C}_k$ doivent nécessairement être positives définies pour (2.28), même si le domaine de définition de f est moins restrictif.

2.4.3 Transformations dans $\mathcal{S}_n^{++}$: inversion et congruence

Pour les deux dernières propriétés, nous nous intéressons aux effets de deux types de transformations des matrices $\boldsymbol{C}_k$ dans $\mathcal{S}_n^{++}$: l'inversion et la congruence. Du côté de l'inversion,

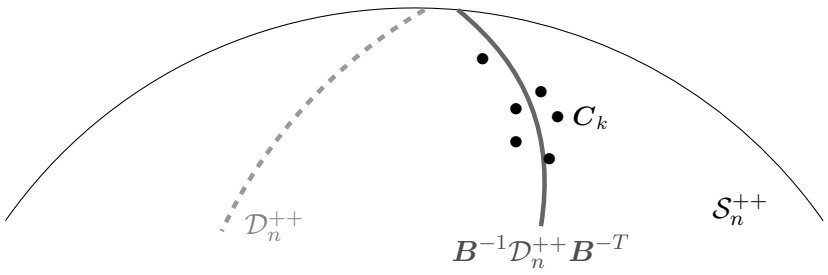

FIGURE *2.6 – Schéma d'illustration du modèle de diagonalisation conjointe approximée basé sur le modèle* (2.27)*. Dans cette situation, on cherche le sous-espace* $\boldsymbol{B}^{-1}\mathcal{D}_n^{++}\boldsymbol{B}^{-T}$ *qui se rapproche le plus de l'ensemble* $\{\boldsymbol{C}_k\}$ *des matrices à diagonaliser, qui restent fixes dans* $\mathcal{S}_n^{++}$*.*

observons que si $\boldsymbol{B}$ est le diagonalisateur conjoint de l'ensemble $\{\boldsymbol{C}_k\}$, il apparaît logique que le diagonalisateur conjoint de l'ensemble $\{\boldsymbol{C}_k^{-1}\}$ soit $\boldsymbol{B}^{-T}$, ce qui donne la propriété 2.3, illustrée dans la figure 2.7. En ce qui concerne la congruence, remarquons que si $\boldsymbol{B}$ est le diagonalisateur conjoint de l'ensemble $\{\boldsymbol{W}\boldsymbol{C}_k\boldsymbol{W}^T\}$ pour $\boldsymbol{W} \in \mathrm{GL}_n$, alors le diagonalisateur conjoint de l'ensemble $\{\boldsymbol{C}_k\}$ devrait être $\boldsymbol{B}\boldsymbol{W}$, ce qui aboutit à la propriété 2.4, illustrée dans la figure 2.8.

Propriété 2.3 (Invariance par inversion)
Le critère f est dit invariant par inversion des matrices d'entrée $\boldsymbol{C}_k$ si, pour tout $\boldsymbol{B}$, on a

$$f(\boldsymbol{B}^{-T}, \{\boldsymbol{C}_k^{-1}\}) = f(\boldsymbol{B}, \{\boldsymbol{C}_k\}).$$

Un critère f de la forme (2.2) a cette propriété si, pour tous $\boldsymbol{M} \in \mathcal{S}_n^{++}$, $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$,

$$h(\boldsymbol{M}^{-1}, \boldsymbol{\Lambda}^{-1}) = h(\boldsymbol{M}, \boldsymbol{\Lambda}), \tag{2.29}$$

et pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{C}_k \in \mathcal{S}_n^{++}$,

$$\boldsymbol{\Lambda}(\boldsymbol{B}^{-T}, \boldsymbol{C}_k^{-1}) = \boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{C}_k)^{-1}. \tag{2.30}$$

Notons que si h vérifie (2.29), alors les matrices diagonales cibles de (2.3) correspondantes vérifient (2.30). Étant donné un critère f, on peut par exemple simplement construire

$$\bar{f}(\boldsymbol{B}, \{\boldsymbol{C}_k\}) = \frac{1}{2}\left(f(\boldsymbol{B}, \{\boldsymbol{C}_k\}) + f(\boldsymbol{B}^{-T}, \{\boldsymbol{C}_k^{-1}\})\right), \tag{2.31}$$

qui possède alors la propriété 2.3.

Propriété 2.4 (Invariance par congruence)
Le critère f est dit invariant par congruence des matrices d'entrée $\boldsymbol{C}_k$ si, pour tous $\boldsymbol{B}, \boldsymbol{W} \in \mathrm{GL}_n$, on a

$$f(\boldsymbol{B}, \{\boldsymbol{W}\boldsymbol{C}_k\boldsymbol{W}^T\}) = f(\boldsymbol{B}\boldsymbol{W}, \{\boldsymbol{C}_k\}).$$

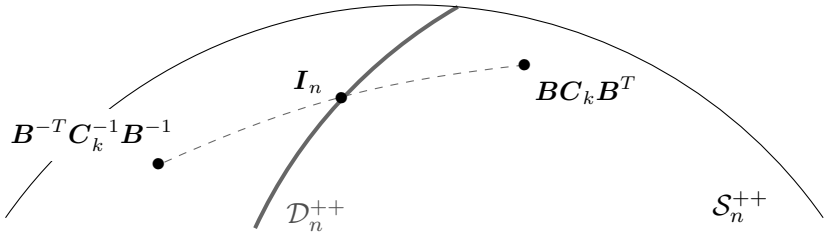

FIGURE *2.7 – Schéma d'illustration de la propriété 2.3 d'invariance par inversion. L'inversion de* $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ *peut être vue comme une symétrie par rapport à l'identité* $\boldsymbol{I}_n$*. En effet,* $\boldsymbol{I}_n$ *se trouve au milieu du géodésique reliant* $\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$ *et son inverse.*

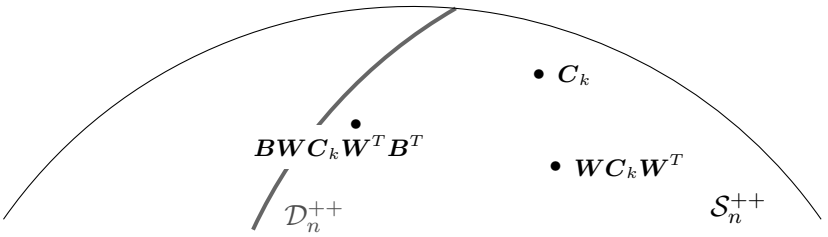

FIGURE *2.8 – Schéma d'illustration de la propriété 2.4 d'invariance par congruence. Que l'on considère* $\boldsymbol{B}$ *pour* $\boldsymbol{W}\boldsymbol{C}_k\boldsymbol{W}^T$ *ou* $\boldsymbol{B}\boldsymbol{W}$ *pour* $\boldsymbol{C}_k$*, on obtient le même élément de* $\mathcal{S}_n^{++}$*.*

En supposant que $\boldsymbol{\Lambda}(\boldsymbol{B}, \boldsymbol{W}\boldsymbol{C}_k\boldsymbol{W}^T) = \boldsymbol{\Lambda}(\boldsymbol{B}\boldsymbol{W}, \boldsymbol{C}_k)$, ce qui devrait toujours être vrai, la propriété 2.4 est toujours vérifiée pour tout critère f de la forme (2.2). En contraste, ce n'est pas toujours le cas pour les critères qui suivent le modèle (2.27).

Pour conclure cette partie, nous présentons le tableau suivant, qui indique quelles propriétés sont vérifiées par les critères de diagonalisation conjointe approximée de la forme (2.2) associés aux divergences de la section précédente.

	f_{F}	$f_{\ell\mathrm{KL}}$	$f_{r\mathrm{KL}}$	$f_{s\mathrm{KL}}$	$f_{\alpha\mathrm{LD}}$ $\alpha \neq 0$	$f_{\alpha\mathrm{LD}}$ $\alpha = 0$	f_{R}	f_{LE}	f_{W}
propriété 2.1		✓	✓	✓	✓	✓	✓	✓	
propriété 2.2		✓	✓	✓	✓	✓	✓		
propriété 2.3				✓		✓	✓	✓	
propriété 2.4	✓	✓	✓	✓	✓	✓	✓	✓	✓

On peut remarquer que seuls les critères basés sur la mesure de Kullback-Leibler symétrique, la divergence log-det α pour $\alpha = 0$ et la distance riemannienne naturelle possèdent l'ensemble de ces propriétés. Ceci est un argument fort en faveur de ces trois critères, qui semblent donc à préférer, au moins d'un point de vue théorique. C'est en contradiction avec l'ensemble des travaux précédents sur la diagonalisation conjointe approximée où des solutions algorithmiques ont été proposées seulement pour la distance de Frobenius, la mesure de Kullback-Leibler

gauche et la divergence log-det α, mais sans considérer les matrices diagonales cibles qui permettent d'assurer l'ensemble des propriétés.

Chapitre 3

Optimisation riemannienne pour la diagonalisation conjointe

Sommaire

Dans ce chapitre, nous développons un cadre d'optimisation riemannienne adapté à la diagonalisation conjointe approximée. C'est un problème d'optimisation sur la variété des matrices inversibles GL_n pour lequel des contraintes additionnelles sont généralement requises, comme expliqué dans le chapitre 1. Nous considérons la contrainte oblique (1.7), la contrainte intrinsèque (1.8) et la contrainte non-holonomique, qui exploite la géométrie des classes d'équivalences (1.6). Nous devons d'abord définir une structure riemannienne pour l'optimisation sur GL_n et ensuite y intégrer les contraintes. Pour l'optimisation riemannienne sur GL_n, nous considérons trois possibilités :

- GL_n équipée de la *métrique invariante à gauche* présentée dans la section 1.5.3, dont l'utilisation est inédite dans le contexte de la diagonalisation conjointe approximée et qui présente l'intérêt majeur d'être invariante le long des classes d'équivalences (1.6).
- GL_n avec la *métrique invariante à droite* également présentée dans la section 1.5.3, qui possède un lien naturel avec le modèle de la diagonalisation conjointe approximée comme montré dans [4, 11, 12].
- La variété produit qui découle de la décomposition polaire sur GL_n, qu'on nomme *variété polaire* et qu'on note $\mathcal{P}_n$. Cette variété correspond au produit de deux variétés classiques particulièrement intéressantes : celle des matrices symétriques positives définies $\mathcal{S}_n^{++}$ et celle des matrices orthogonales $\mathcal{O}_n$, présentées dans la section 1.5.

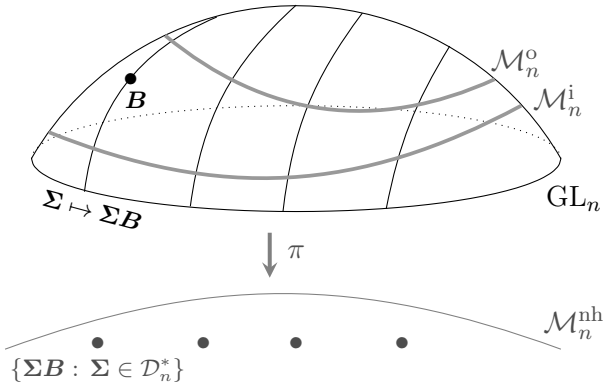

FIGURE *3.1 – Schéma d'illustration de l'intégration dans* GL_n *des contraintes oblique, intrinsèque et non-holonomique qui forment respectivement les variétés* $\mathcal{M}_n^{\mathrm{o}}$, $\mathcal{M}_n^{\mathrm{i}}$ *et* $\mathcal{M}_n^{\mathrm{nh}}$.

En ce qui concerne l'intégration des contraintes, la contrainte oblique (1.7) et la contrainte intrinsèque (1.8) forment des sous-variétés de GL_n respectivement notées $\mathcal{M}_n^{\mathrm{o}}$ et $\mathcal{M}_n^{\mathrm{i}}$ et la contrainte non-holonomique engendre une variété quotient de GL_n notée $\mathcal{M}_n^{\mathrm{nh}}$. La figure 3.1 propose une illustration géométrique de l'intégration de ces trois contraintes dans GL_n.

Le but de ce chapitre est de poser les fondations de notre cadre d'optimisation riemannienne, son contenu est donc technique. Dans la section 3.1, nous définissons la variété polaire $\mathcal{P}_n$ qui permet d'optimiser sur GL_n grâce à la décomposition polaire. Nous étudions les liens de cette variété avec GL_n et nous montrons notamment qu'il est possible de rendre en pratique son utilisation transparente à celle de GL_n, *i.e.*, il suffit à l'utilisateur de définir la fonction objectif sur GL_n et de fournir son gradient euclidien (et éventuellement hessienne) sur GL_n. Dans la section 3.2, nous traitons les cas de la contrainte oblique et de la contrainte intrinsèque, qui sont intégrées dans des sous-variétés riemanniennes de GL_n équipé des métriques invariantes à gauche et à droite et de la variété polaire $\mathcal{P}_n$. Pour terminer, dans la section 3.3, nous étudions la contrainte non-holonomique, qui forme une variété quotient de GL_n. Nous montrons que la métrique invariante à gauche permet de définir une variété quotient riemannienne. Nous étudions également la possibilité d'optimiser sur cette variété quotient alors que GL_n est équipé de la métrique invariante à droite. Nous considérons finalement la possibilité d'exploiter la géométrie de cette variété pour optimiser des fonctions objectifs sur GL_n qui n'induisent pas de fonction sur le quotient.

3.1 Optimisation riemannienne sur GL_n grâce à la décomposition polaire

Toute matrice $\boldsymbol{B} \in \mathrm{GL}_n$ admet une décomposition polaire unique $\boldsymbol{B} = \boldsymbol{SU}$, où $\boldsymbol{S} \in \mathcal{S}_n^{++}$ et $\boldsymbol{U} \in \mathcal{O}_n$. La fonction $\Gamma : \mathcal{P}_n = \mathcal{S}_n^{++} \times \mathcal{O}_n \to \mathrm{GL}_n$ telle que $\Gamma(\boldsymbol{S}, \boldsymbol{U}) = \boldsymbol{SU}$ est un difféomorphisme. La minimisation d'une fonction objectif f sur GL_n correspond donc à celle

de $f \circ \Gamma$ sur $\mathcal{P}_n$. Étant le produit des variétés $\mathcal{S}_n^{++}$ et $\mathcal{O}_n$, $\mathcal{P}_n$ est une variété qu'on peut équiper d'une structure riemannienne et qu'on nomme variété polaire. Dans la section 3.1.1, nous présentons l'ensemble des outils de géométrie et d'optimisation riemannienne sur $\mathcal{P}_n$ dont nous avons besoin pour construire des méthodes de résolution du problème de diagonalisation conjointe approximée. Dans la section 3.1.2, nous montrons qu'il est possible d'obtenir certains des ingrédients pour l'optimisation sur le produit $\mathcal{P}_n$ à partir d'objets définis au niveau de GL_n, ce qui augmente sa modularité et simplifie son utilisation.

3.1.1 La variété polaire

Nous utilisons les objets de $\mathcal{S}_n^{++}$ et de $\mathcal{O}_n$ donnés dans la section 1.5 pour définir ceux de $\mathcal{P}_n$. L'ensemble des preuves des résultats de cette partie provient directement des propriétés des produits de variétés [2]. L'espace tangent à $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$ est simplement $T_{\boldsymbol{S}}\mathcal{S}_n^{++} \times T_{\boldsymbol{U}}\mathcal{O}_n$:

$$T_{\mathcal{B}}\mathcal{P}_n = \left\{ (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{U}\boldsymbol{\Omega}) : \boldsymbol{\xi}_{\boldsymbol{S}} \in \mathcal{S}_n, \boldsymbol{\Omega} \in \mathbb{R}^{n\times n}, \boldsymbol{\Omega}^T = -\boldsymbol{\Omega} \right\}. \tag{3.1}$$

Le choix naturel pour la métrique riemannienne d'une variété produit est de prendre la somme des métriques des variétés du produit. Nous équipons donc $\mathcal{P}_n$ avec la métrique définie, pour tous $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$, $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}), H = (\boldsymbol{\eta}_{\boldsymbol{S}}, \boldsymbol{\eta}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{P}_n$, par

$$\langle \Xi, H \rangle_{\mathcal{B}}^{\mathcal{P}_n} = \langle \boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\eta}_{\boldsymbol{S}} \rangle_{\boldsymbol{S}}^{\mathcal{S}_n^{++}} + \langle \boldsymbol{\xi}_{\boldsymbol{U}}, \boldsymbol{\eta}_{\boldsymbol{U}} \rangle_{\boldsymbol{U}}^{\mathcal{O}_n}, \tag{3.2}$$

où les métriques $\langle \cdot, \cdot \rangle_{\cdot}^{\mathcal{S}_n^{++}}$ et $\langle \cdot, \cdot \rangle_{\cdot}^{\mathcal{O}_n}$ sont définies dans (1.30) et (1.41). La projection à $\mathcal{B} \in \mathcal{P}_n$ sur $T_{\mathcal{B}}\mathcal{P}_n$ depuis l'espace ambient $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ est donnée, pour tout $\mathcal{Z} = (\boldsymbol{Z}_{\boldsymbol{S}}, \boldsymbol{Z}_{\boldsymbol{U}})$, par

$$P_{\mathcal{B}}^{\mathcal{P}_n}(\mathcal{Z}) = \left(P_{\boldsymbol{S}}^{\mathcal{S}_n^{++}}(\boldsymbol{Z}_{\boldsymbol{S}}), P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{Z}_{\boldsymbol{U}}) \right). \tag{3.3}$$

La connection de Levi-Civita sur $\mathcal{P}_n$ associée à la métrique (3.2) est, pour $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$ et les champs de vecteurs $\Xi_{\mathcal{B}} = (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}), H_{\mathcal{B}} = (\boldsymbol{\eta}_{\boldsymbol{S}}, \boldsymbol{\eta}_{\boldsymbol{U}})$ évalués en $\mathcal{B}$,

$$\nabla_{\Xi_{\mathcal{B}}}^{\mathcal{P}_n} H_{\mathcal{B}} = \left(\mathrm{D}\, \boldsymbol{\eta}_{\boldsymbol{S}}[\Xi_{\mathcal{B}}] - \mathrm{sym}\left(\boldsymbol{\eta}_{\boldsymbol{S}} \boldsymbol{S}^{-1} \boldsymbol{\xi}_{\boldsymbol{S}} \right), P_{\boldsymbol{U}}^{\mathcal{O}_n}\left(\mathrm{D}\, \boldsymbol{\eta}_{\boldsymbol{U}}[\Xi_{\mathcal{B}}] \right) \right). \tag{3.4}$$

L'exponentielle riemannienne sur $\mathcal{P}_n$ est alors définie, pour tous $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$ et $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{P}_n$, par

$$\exp_{\mathcal{B}}^{\mathcal{P}_n}(\Xi) = \left(\exp_{\boldsymbol{S}}^{\mathcal{S}_n^{++}}(\boldsymbol{\xi}_{\boldsymbol{S}}), \exp_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{\xi}_{\boldsymbol{U}}) \right). \tag{3.5}$$

Notons qu'on peut aussi utiliser une rétraction sur $\mathcal{P}_n$ en remplaçant par exemple $\exp_{\cdot}^{\mathcal{O}_n}(\cdot)$ par une des rétractions définies dans (1.45). Pour le transport de vecteurs, on peut utiliser

$$\mathcal{T}^{\mathcal{P}_n}(\mathcal{B}, \Xi, H) = P^{\mathcal{P}_n}_{\exp_{\mathcal{B}}^{\mathcal{P}_n}(\Xi)}(H) = \left(\boldsymbol{\eta}_{\boldsymbol{S}}, P^{\mathcal{O}_n}_{\exp_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{\xi}_{\boldsymbol{U}})}(\boldsymbol{\eta}_{\boldsymbol{U}}) \right). \tag{3.6}$$

Ici encore, on peut remplacer $\exp_{\cdot}^{\mathcal{O}_n}(\cdot)$ par une des rétractions définies dans (1.45). Enfin, soit une fonction objectif f définie sur $\mathcal{P}_n$ dont le gradient euclidien dans $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ à $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$ est $\mathrm{grad}_{\mathfrak{E}^2} f(\mathcal{B}) = (\mathrm{grad}_{\mathfrak{E}^2} f(\boldsymbol{S}), \mathrm{grad}_{\mathfrak{E}^2} f(\boldsymbol{U}))$. Pour souligner le fait que le

gradient euclidien est dans $\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}$ et pas dans $\mathbb{R}^{n\times n}$, on le note avec l'indice $\mathfrak{E}^2$. Le gradient riemannien de f est donné par

$$\operatorname{grad}_{\mathcal{P}_n} f(\mathcal{B}) = \left(\boldsymbol{S}\operatorname{sym}(\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{S}))\boldsymbol{S}, P^{\mathcal{O}_n}_{\boldsymbol{U}}(\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{U}))\right). \tag{3.7}$$

La hessienne riemannienne $\operatorname{hess}_{\mathcal{P}_n} f(\mathcal{B})[\Xi] = (\operatorname{hess}_{\mathcal{P}_n} f(\boldsymbol{S})[\Xi], \operatorname{hess}_{\mathcal{P}_n} f(\boldsymbol{U})[\Xi])$ est

$$\begin{aligned}\operatorname{hess}_{\mathcal{P}_n} f(\boldsymbol{S})[\Xi] &= \boldsymbol{S}\operatorname{sym}\left(\operatorname{hess}_{\mathfrak{E}^2} f(\boldsymbol{S})[\Xi]\right)\boldsymbol{S} + \operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\operatorname{sym}\left(\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{S})\right)\boldsymbol{S}),\\ \operatorname{hess}_{\mathcal{P}_n} f(\boldsymbol{U})[\Xi] &= P^{\mathcal{O}_n}_{\boldsymbol{U}}\left(\operatorname{hess}_{\mathfrak{E}^2} f(\boldsymbol{U})[\Xi] - \boldsymbol{\xi}_{\boldsymbol{U}}\operatorname{sym}\left(\boldsymbol{U}^T\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{U})\right)\right),\end{aligned} \tag{3.8}$$

où $\operatorname{hess}_{\mathfrak{E}^2} f(\mathcal{B})[\Xi] = (\operatorname{hess}_{\mathfrak{E}^2} f(\boldsymbol{S})[\Xi], \operatorname{hess}_{\mathfrak{E}^2} f(\boldsymbol{U})[\Xi])$ est la hessienne euclidienne de f, définie dans $\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}$. Ce résultat est obtenu avec $\operatorname{hess}_{\mathcal{P}_n} f(\mathcal{B})[\Xi] = \nabla^{\mathcal{P}_n}_{\Xi}\operatorname{grad}_{\mathcal{P}_n} f(\mathcal{B})$. Pour le terme $\operatorname{hess}_{\mathcal{P}_n} f(\boldsymbol{U})[\Xi]$, on exploite $\mathrm{D}\, P^{\mathcal{O}_n}_{\boldsymbol{U}}(\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{U}))[\Xi] = P^{\mathcal{O}_n}_{\boldsymbol{U}}(\operatorname{hess}_{\mathfrak{E}^2} f(\boldsymbol{U})[\Xi]) + \mathrm{D}\, P^{\mathcal{O}_n}_{\boldsymbol{U}}[\boldsymbol{\xi}_{\boldsymbol{U}}](\operatorname{grad}_{\mathfrak{E}^2} f(\boldsymbol{U}))$ et $P^{\mathcal{O}_n}_{\boldsymbol{U}}(\mathrm{D}\, P^{\mathcal{O}_n}_{\boldsymbol{U}}[\boldsymbol{\xi}_{\boldsymbol{U}}](\boldsymbol{Z})) = -P^{\mathcal{O}_n}_{\boldsymbol{U}}(\boldsymbol{\xi}_{\boldsymbol{U}}\operatorname{sym}(\boldsymbol{U}^T\boldsymbol{Z}))$, où $\mathrm{D}\, P^{\mathcal{O}_n}_{\boldsymbol{U}}[\boldsymbol{\xi}_{\boldsymbol{U}}](\boldsymbol{Z})$ correspond à la dérivée directionnelle de $\boldsymbol{U}\mapsto P^{\mathcal{O}_n}_{\boldsymbol{U}}(\boldsymbol{Z})$.

3.1.2 Liens avec GL_n

La fonction $\Gamma : \mathcal{P}_n \to \mathrm{GL}_n$ telle que $\Gamma(\boldsymbol{S},\boldsymbol{U}) = \boldsymbol{S}\boldsymbol{U}$ est un difféomorphisme dont l'inverse $\Gamma^{-1} : \mathrm{GL}_n \to \mathcal{P}_n$ est donnée, pour tout $\boldsymbol{B}\in\mathrm{GL}_n$, par

$$\Gamma^{-1}(\boldsymbol{B}) = \left((\boldsymbol{B}\boldsymbol{B}^T)^{1/2}, (\boldsymbol{B}\boldsymbol{B}^T)^{-1/2}\boldsymbol{B}\right). \tag{3.9}$$

La dérivée directionnelle de Γ à $\mathcal{B} = (\boldsymbol{S},\boldsymbol{U})\in\mathcal{P}_n$ dans la direction $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}},\boldsymbol{\xi}_{\boldsymbol{U}})\in T_{\mathcal{B}}\mathcal{P}_n$ est

$$\mathrm{D}\,\Gamma(\mathcal{B})[\Xi] = \boldsymbol{S}\boldsymbol{\xi}_{\boldsymbol{U}} + \boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{U}. \tag{3.10}$$

Autrement dit, le vecteur tangent Ξ de $\mathcal{B}$ dans $\mathcal{P}_n$ correspond au vecteur tangent $\boldsymbol{\xi} = \boldsymbol{S}\boldsymbol{\xi}_{\boldsymbol{U}} + \boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{U}$ de $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{U}$ dans GL_n. On peut également inverser $\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]$, *i.e.*, retrouver le vecteur tangent $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}},\boldsymbol{\xi}_{\boldsymbol{U}})$ de $\mathcal{B} = \Gamma^{-1}(\boldsymbol{B})$ qui correspond à un vecteur tangent $\boldsymbol{\xi}\in\mathbb{R}^{n\times n}$ de $\boldsymbol{B}\in\mathrm{GL}_n$. En effet, étant donné $\boldsymbol{B}\in\mathrm{GL}_n$ et $\boldsymbol{\xi}\in\mathbb{R}^{n\times n}$, on a

$$\mathrm{D}\,\Gamma^{-1}(\boldsymbol{B})[\boldsymbol{\xi}] = \left(\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{S}^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{U})\right), \tag{3.11}$$

où $\boldsymbol{\xi}_{\boldsymbol{S}}\in\mathcal{S}_n$ est l'unique solution de l'équation de Sylvester [1]

$$\boldsymbol{S}\boldsymbol{\xi}_{\boldsymbol{S}} + \boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{S} = \boldsymbol{\xi}\boldsymbol{B}^T + \boldsymbol{B}\boldsymbol{\xi}^T. \tag{3.12}$$

De ce fait, on est en mesure de définir les points de $\mathcal{P}_n$ et leurs vecteurs tangents à partir des points de GL_n et de leurs vecteurs tangent dans $\mathbb{R}^{n\times n}$. La figure 3.2 contient une illustration de ces liens entre $\mathcal{P}_n$ et GL_n.

Une conséquence est, qu'étant donné une sous-variété $\mathcal{M}^{\mathrm{GL}_n}_n$ de GL_n, on peut définir la sous-variété $\mathcal{M}^{\mathcal{P}_n}_n$ correspondante, *i.e.*, telle que

$$\mathcal{M}^{\mathcal{P}_n}_n = \left\{\Gamma^{-1}(\boldsymbol{B}) :\ \boldsymbol{B}\in\mathcal{M}^{\mathrm{GL}_n}_n\right\}. \tag{3.13}$$

1. On est assuré que la solution existe, est unique et symétrique car $\boldsymbol{S}$ est symétrique positive définie et $\boldsymbol{\xi}\boldsymbol{B}^T + \boldsymbol{B}\boldsymbol{\xi}^T$ est symétrique.

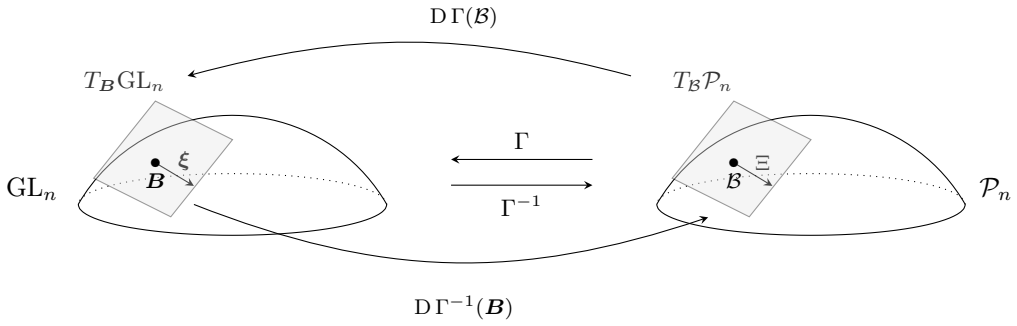

FIGURE *3.2 – Schéma d'illustration des liens entre $\mathcal{P}_n$ et GL_n. On peut passer d'une variété à l'autre en employant Γ, Γ^{-1}, $\mathrm{D}\,\Gamma$ et $\mathrm{D}\,\Gamma^{-1}$.*

Son espace tangent à $\Gamma^{-1}(\boldsymbol{B})$ est alors par définition

$$T_{\Gamma^{-1}(\boldsymbol{B})}\mathcal{M}_n^{\mathcal{P}_n} = \left\{\mathrm{D}\,\Gamma^{-1}(\boldsymbol{B})[\boldsymbol{\xi}] : \boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{GL}_n}\right\}. \tag{3.14}$$

Réciproquement, si $\mathcal{M}_n^{\mathcal{P}_n}$ est une sous-variété de $\mathcal{P}_n$, on peut définir la sous-variété $\mathcal{M}_n^{\mathrm{GL}_n}$ de GL_n correspondante avec Γ et pour tout $\mathcal{B} \in \mathcal{M}_n^{\mathcal{P}}$, l'espace tangent à $\Gamma(\mathcal{B})$ est obtenu avec $\mathrm{D}\,\Gamma(\mathcal{B})$.

Dans la suite de cette section, $\mathcal{M}_n^{\mathrm{GL}_n}$ désigne indistinctement GL_n ou une de ses sous-variétés et $\mathcal{M}_n^{\mathcal{P}_n}$ est la variété qui correspond à $\mathcal{M}_n^{\mathrm{GL}_n}$, qui peut être $\mathcal{P}_n$ ou une de ses sous-variétés. Nous donnons maintenant deux propositions qui montrent les liens qu'il existe entre deux objets de $\mathcal{M}_n^{\mathrm{GL}_n}$ et de $\mathcal{M}_n^{\mathcal{P}_n}$. En effet, la proposition 3.1 montre qu'une rétraction de $\mathcal{M}_n^{\mathrm{GL}_n}$ engendre une rétraction sur $\mathcal{M}_n^{\mathcal{P}_n}$, et réciproquement. La proposition 3.2 donne la relation entre le gradient euclidien et la hessienne euclidienne d'une fonction objectif f définie sur $\mathcal{M}_n^{\mathrm{GL}_n}$ et ceux de la fonction objectif correspondante $\overline{f} = f \circ \Gamma$ définie sur $\mathcal{M}_n^{\mathcal{P}_n}$.

Proposition 3.1 (Rétraction)
Si R^{GL_n} est une rétraction sur $\mathcal{M}_n^{\mathrm{GL}_n}$, alors la fonction définie pour tous $\mathcal{B} \in \mathcal{M}_n^{\mathcal{P}_n}$ et $\Xi \in T_{\mathcal{B}}\mathcal{M}_n^{\mathcal{P}_n}$ par

$$R_{\mathcal{B}}^{\mathcal{P}_n}(\Xi) = \Gamma^{-1}\left(R_{\Gamma(\mathcal{B})}^{\mathrm{GL}_n}(\mathrm{D}\,\Gamma(\mathcal{B})[\Xi])\right),$$

est une rétraction sur $\mathcal{M}_n^{\mathcal{P}_n}$. De la même façon, si $R^{\mathcal{P}_n}$ est une rétraction sur $\mathcal{M}_n^{\mathcal{P}_n}$, on peut définir une rétraction sur $\mathcal{M}_n^{\mathrm{GL}_n}$ par

$$R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi}) = \Gamma\left(R_{\Gamma^{-1}(\boldsymbol{B})}^{\mathcal{P}_n}(\mathrm{D}\,\Gamma^{-1}(\boldsymbol{B})[\boldsymbol{\xi}])\right),$$

pour tous $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{GL}_n}$ et $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{GL}_n}$.

Démonstration. Étant donné une rétraction R^{GL_n} sur $\mathcal{M}_n^{\mathrm{GL}_n}$ et $\mathcal{B} \in \mathcal{M}_n^{\mathcal{P}_n}$, on a

$$R_{\mathcal{B}}^{\mathcal{P}_n}(0) = \Gamma^{-1}\left(R_{\Gamma(\mathcal{B})}^{\mathrm{GL}_n}(0)\right) = \Gamma^{-1}(\Gamma(\mathcal{B})) = \mathcal{B}.$$

De plus, pour $\Xi \in T_{\mathcal{B}}\mathcal{M}_n^{\mathcal{P}_n}$, on a

$$\begin{aligned}
\mathrm{D}\, R_{\mathcal{B}}^{\mathcal{P}_n}(0)[\Xi] &= \mathrm{D}\left(\Gamma^{-1}\left(R_{\Gamma(\mathcal{B})}^{\mathrm{GL}_n}(0)\right)\right)[\Xi] \\
&= \mathrm{D}\,\Gamma^{-1}\left(R_{\Gamma(\mathcal{B})}^{\mathrm{GL}_n}(0)\right)\left[\mathrm{D}\, R_{\Gamma(\mathcal{B})}^{\mathrm{GL}_n}(0)\left[\mathrm{D}\,\Gamma(\boldsymbol{B})[\Xi]\right]\right] \\
&= \mathrm{D}\,\Gamma^{-1}(\Gamma(\mathcal{B}))[\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]] \\
&= \Xi.
\end{aligned}$$

On peut donc conclure que $R^{\mathcal{P}_n}$ est bien une rétraction sur $\mathcal{M}_n^{\mathcal{P}_n}$. La preuve dans l'autre sens, *i.e.*, qu'une rétraction sur $\mathcal{M}_n^{\mathcal{P}_n}$ induit une rétraction sur $\mathcal{M}_n^{\mathrm{GL}_n}$, est la même en intervertissant Γ^{-1} avec Γ et R^{GL_n} avec $R^{\mathcal{P}_n}$. □

Proposition 3.2 (Gradient euclidien et hessiennes euclidienne)
Soit une fonction objectif f dont le gradient euclidien et la hessienne euclidienne dans GL_n *sont notés* $\mathrm{grad}_{\mathcal{E}} f$ *et* $\mathrm{hess}_{\mathcal{E}} f$. *Étant donné $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{P}_n$ et $\Xi = (\boldsymbol{\xi_S}, \boldsymbol{\xi_U})$, le gradient et la hessienne euclidiennes de $\overline{f}$ dans $\mathcal{P}_n$ sont*

$$\begin{aligned}
\mathrm{grad}_{\mathfrak{E}^2}\,\overline{f}(\mathcal{B}) &= \left(\mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))\boldsymbol{U}^T, \boldsymbol{S}\,\mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))\right), \\
\mathrm{hess}_{\mathfrak{E}^2}\,\overline{f}(\boldsymbol{S})[\Xi] &= \mathrm{hess}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))[\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]]\boldsymbol{U}^T + \mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))\boldsymbol{\xi}_{\boldsymbol{U}}^T, \\
\mathrm{hess}_{\mathfrak{E}^2}\,\overline{f}(\boldsymbol{U})[\Xi] &= \boldsymbol{S}\,\mathrm{hess}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))[\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]] + \boldsymbol{\xi_S}\,\mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B})).
\end{aligned}$$

Démonstration. Pour le gradient, on a par définition

$$\begin{aligned}
\langle \mathrm{grad}_{\mathfrak{E}^2}\,\overline{f}(\mathcal{B}), \Xi\rangle^{\mathfrak{E}^2} &= \mathrm{tr}\left(\mathrm{grad}_{\mathfrak{E}^2}\,\overline{f}(\boldsymbol{S})\boldsymbol{\xi}_{\boldsymbol{S}}^T\right) + \mathrm{tr}\left(\mathrm{grad}_{\mathfrak{E}^2}\,\overline{f}(\boldsymbol{U})\boldsymbol{\xi}_{\boldsymbol{U}}^T\right) \\
&= \mathrm{D}\,\overline{f}(\mathcal{B})[\Xi] \\
&= \mathrm{D}\, f(\Gamma(\mathcal{B}))[\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]] \\
&= \langle \mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B})), \mathrm{D}\,\Gamma(\mathcal{B})[\Xi]\rangle^{\mathcal{E}} \\
&= \mathrm{tr}\left(\mathrm{grad}_{\mathcal{E}}\, f(\Gamma(\mathcal{B}))(\boldsymbol{\xi_S}\boldsymbol{U} + \boldsymbol{S}\boldsymbol{\xi_U})^T\right).
\end{aligned}$$

Des manipulations basiques de la trace donnent le résultat. La hessienne est ensuite obtenue par dérivation de la formule du gradient. □

3.2 Contraintes intégrées dans des sous-variétés de GL_n

Dans cette section, nous traitons la contrainte oblique (1.7) et celle intrinsèque (1.8) qui peuvent être intégrées dans des sous-variétés de GL_n, notées $\mathcal{M}_n^{\mathrm{o}}$ et $\mathcal{M}_n^{\mathrm{i}}$. Comme montré dans la section 3.1.1, $\mathcal{M}_n^{\mathrm{o}}$ et $\mathcal{M}_n^{\mathrm{i}}$ définissent des sous-variétés de $\mathcal{P}_n$, obtenues avec (3.13) et notées $\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ et $\mathcal{M}_n^{\mathrm{i},\mathcal{P}_n}$. Pour chaque contrainte, nous considérons trois structures riemanniennes qui proviennent de (*i*) GL_n équipé de la métrique invariante à gauche définie dans (1.47), (*ii*) GL_n équipé de la métrique invariante à droite également définie dans (1.47) et (*iii*) la variété polaire $\mathcal{P}_n$. Rappelons que lorsqu'on s'intéresse à l'optimisation riemannienne sur une sous-variété, traitée dans la section 1.4.2, il est suffisant de définir son espace tangent, le projecteur orthogonal, une rétraction et de caractériser la hessienne riemannienne des fonctions objectifs. Les autres objets requis, *i.e.*, la connection de Levi-Civita, le gradient riemannien et le transport de vecteurs, sont obtenus avec (1.19), (1.20) et (1.21). Pour plus de clarté, les démonstrations de certaines propositions de cette section sont reportées dans l'annexe D.

3.2.1 Contrainte oblique

La contrainte oblique (1.7) est intégrée dans la sous-variété de GL_n définie par

$$\mathcal{M}_n^{\mathrm{o}} = \{\boldsymbol{B} \in \mathrm{GL}_n : \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{B}^T) = \boldsymbol{I}_n\}, \tag{3.15}$$

dont l'espace tangent à $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ est

$$T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}} = \{\boldsymbol{\xi} \in \mathbb{R}^{n\times n} : \operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^T) = \boldsymbol{0}_n\}. \tag{3.16}$$

Proposition 3.3 (Projecteurs orthogonaux)
Pour GL_n *équipé de la métrique invariante à gauche, le projecteur orthogonal est défini, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ *et* $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$*, par*

$$P_{\boldsymbol{B}}^{\mathrm{o},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B}, \qquad \text{avec} \qquad \operatorname{diag}(\boldsymbol{\Lambda}_{\mathrm{o},\ell}) = \left(\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{B}\boldsymbol{B}^T\right)^{-1} \operatorname{diag}(\boldsymbol{Z}\boldsymbol{B}^T),$$

où $\odot$ *est le produit de Hadamard. Dans le cas où* GL_n *est équipé de la métrique invariante à droite, le projecteur orthogonal est, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ *et* $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$*,*

$$P_{\boldsymbol{B}}^{\mathrm{o},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B}, \qquad \text{avec} \qquad \boldsymbol{\Lambda}_{\mathrm{o},r} = \operatorname{ddiag}(\boldsymbol{Z}\boldsymbol{B}^T) \operatorname{ddiag}\left((\boldsymbol{B}\boldsymbol{B}^T)^2\right)^{-1}.$$

Enfin, pour la variété polaire $\mathcal{P}_n$*, le projecteur orthogonal est donné, pour tous* $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ *et* $\mathcal{Z} = (\boldsymbol{Z}_{\boldsymbol{S}}, \boldsymbol{Z}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{P}_n$*, par*

$$P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}(\mathcal{Z}) = \left(\boldsymbol{Z}_{\boldsymbol{S}} - \boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{Z}_{\boldsymbol{U}}\right),$$
$$\text{avec} \qquad \operatorname{diag}(\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}) = 2\left(\boldsymbol{S}^3 \odot \boldsymbol{S} + \boldsymbol{S}^2 \odot \boldsymbol{S}^2\right)^{-1} \operatorname{diag}(\boldsymbol{Z}_{\boldsymbol{S}}\boldsymbol{S}).$$

Démonstration. La preuve de cette proposition est donnée dans l'annexe D. □

Proposition 3.4 (Rétraction)
Étant donné une rétraction R^{GL_n} *sur* GL_n*, on peut définir une rétraction sur* $\mathcal{M}_n^{\mathrm{o}}$*, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ *et* $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$*, par*

$$R_{\boldsymbol{B}}^{\mathrm{o}}(\boldsymbol{\xi}) = \operatorname{ddiag}\left(R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})^T\right)^{-1/2} R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi}).$$

Démonstration. On peut montrer que $R_{\boldsymbol{B}}^{\mathrm{o}}(\boldsymbol{\xi}) \in \mathcal{M}_n^{\mathrm{o}}$ en vérifiant que $\operatorname{ddiag}(R_{\boldsymbol{B}}^{\mathrm{o}}(\boldsymbol{\xi})R_{\boldsymbol{B}}^{\mathrm{o}}(\boldsymbol{\xi})^T) = \boldsymbol{I}_n$. Comme $R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi}) = \boldsymbol{B} + \boldsymbol{\xi} + o(\|\xi\|)$, on a

$$\operatorname{ddiag}\left(R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})^T\right) = \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{B}^T) + 2\operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^T) + o(\|\xi\|).$$

Par définition, $\operatorname{ddiag}(\boldsymbol{B}\boldsymbol{B}^T) = \boldsymbol{I}_n$ et $\operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^T) = \boldsymbol{0}_n$, donc

$$\operatorname{ddiag}\left(R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})^T\right) = \boldsymbol{I}_n + o(\|\xi\|).$$

De ce fait, $R_{\boldsymbol{B}}^{\mathrm{o}}(\boldsymbol{\xi}) = \boldsymbol{B} + \boldsymbol{\xi} + o(\|\xi\|)$, ce qui complète la preuve. □

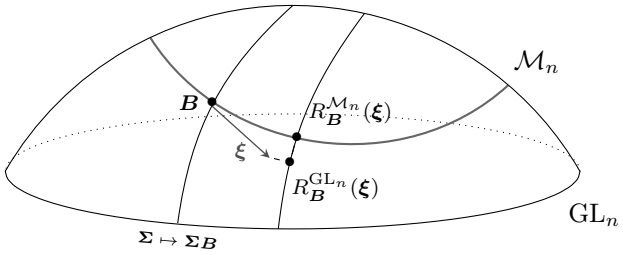

FIGURE *3.3 – Schéma d'illustration des rétractions définies dans les propositions 3.4 et 3.7 pour les sous-variétés $\mathcal{M}_n^{\mathrm{o}}$ et $\mathcal{M}_n^{\mathrm{i}}$ (représentées par $\mathcal{M}_n$ sur la figure). Pour effectuer la rétraction à $\boldsymbol{B}$ de $\boldsymbol{\xi}$ sur $\mathcal{M}_n$, on commence par faire la rétraction de $\boldsymbol{\xi}$ sur GL_n puis on se déplace le long de la classe d'équivalence de $R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})$ pour appliquer la contrainte.*

Remarquons que la fonction $\boldsymbol{X} \mapsto \operatorname{ddiag}(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}\boldsymbol{X}$ définit une projection de GL_n sur $\mathcal{M}_n^{\mathrm{o}}$. Une rétraction issue de la proposition 3.4 (voir la figure 3.3 pour une illustration) s'inscrit donc dans le contexte de [3]. Dans la proposition 3.4, nous choisissons (*i*) $R^{\mathrm{GL}_n} = \exp^{\ell}$ définie à partir de (1.50) lorsque GL_n est équipé de la métrique invariante à gauche. Quand GL_n est équipé de la métrique invariante à droite, nous prenons (*ii*) $R^{\mathrm{GL}_n} = \exp^{r}$ définie à partir de (1.51). Pour la variété polaire $\mathcal{P}_n$, on emploie (*iii*) la proposition 3.1 pour définir R^{GL_n} à partir de $\exp^{\mathcal{P}_n}$ donnée dans (3.5) et on réutilise cette proposition pour définir une rétraction sur $\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ à partir de R^{o}.

Proposition 3.5 (hessiennes riemanniennes)
Étant donné un critère f défini sur GL_n, la hessienne riemannienne sur $\mathcal{M}_n^{\mathrm{o}}$ dans le cas où GL_n est équipé de la métrique invariante à gauche est donnée, pour tous $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ et $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$, par

$$\operatorname{hess}_{\mathrm{o},\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] = P_{\boldsymbol{B}}^{\mathrm{o},\ell}\left(\operatorname{hess}_{\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] - \operatorname{sym}(\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B} - \boldsymbol{B}\operatorname{sym}(\boldsymbol{B}^{-1}\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B})\right),$$

où $\boldsymbol{\Lambda}_{\mathrm{o},\ell}$ est défini dans la proposition 3.3 pour $\boldsymbol{Z} = \operatorname{grad}_{\ell} f(\boldsymbol{B})$. Lorsque GL_n est équipé de la métrique invariante à droite, la hessienne riemannienne sur $\mathcal{M}_n^{\mathrm{o}}$ est, pour tous $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$ et $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$,

$$\begin{aligned}\operatorname{hess}_{\mathrm{o},r} f(\boldsymbol{B})[\boldsymbol{\xi}] = {} & P_{\boldsymbol{B}}^{\mathrm{o},r}\left(\operatorname{hess}_{r} f(\boldsymbol{B})[\boldsymbol{\xi}] + \boldsymbol{\xi}\boldsymbol{B}^{-1}\operatorname{sym}(\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},r})\boldsymbol{B}\right) \\ & - P_{\boldsymbol{B}}^{\mathrm{o},r}\left(\boldsymbol{\Lambda}_{\mathrm{o},r}\left(\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{B}\operatorname{sym}(\boldsymbol{\xi}^T\boldsymbol{B})\right) + \operatorname{sym}(\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{B}\right)\end{aligned}$$

où $\boldsymbol{\Lambda}_{\mathrm{o},r}$ est défini dans la proposition 3.3 pour $\boldsymbol{Z} = \operatorname{grad}_{r} f(\boldsymbol{B})$. Enfin, la hessienne riemannienne sur $\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ est donnée, pour tous $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ et $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$, par

$$\begin{aligned}\operatorname{hess}_{\mathrm{o},\mathcal{P}_n} f(\mathcal{B})[\Xi] = {} & P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}\left(\operatorname{hess}_{\mathcal{P}_n} f(\mathcal{B})[\Xi]\right) \\ & - P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}\left(\boldsymbol{S}\operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S} + \operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}), \boldsymbol{0}_n\right),\end{aligned}$$

où $\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}$ est défini dans la proposition 3.3 pour $\mathcal{Z} = \operatorname{grad}_{\mathcal{P}_n} f(\mathcal{B})$.

Démonstration. La preuve de cette proposition est donnée dans l'annexe D. □

3.2.2 Contrainte intrinsèque

Étant donné l'ensemble de matrices symétriques positives définies $\{\boldsymbol{C}_k\}$, la contrainte (1.8) donne la sous-variété de GL_n définie par

$$\mathcal{M}_n^{\mathrm{i}} = \left\{ \boldsymbol{B} \in \mathrm{GL}_n : \sum_k \mathrm{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T)^2 = \boldsymbol{I}_n \right\}, \tag{3.17}$$

dont l'espace tangent à $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{i}}$ est

$$T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{i}} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{n\times n} : \mathrm{ddiag}(\boldsymbol{\xi}\boldsymbol{Q}) = \boldsymbol{0}_n \right\}, \tag{3.18}$$

où $\boldsymbol{Q} = \sum_k \boldsymbol{C}_k\boldsymbol{B}^T \mathrm{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T)$.

Proposition 3.6 (Projecteurs orthogonaux)
Lorsque GL_n *est équipé de la métrique invariante à gauche, le projecteur orthogonal est défini, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{i}}$ *et* $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$*, par*

$$P_{\boldsymbol{B}}^{\mathrm{i},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{Q}^T, \qquad \text{avec} \qquad \mathrm{diag}(\boldsymbol{\Lambda}_{\mathrm{i},\ell}) = \left(\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{Q}^T\boldsymbol{Q}\right)^{-1} \mathrm{diag}(\boldsymbol{Z}\boldsymbol{Q}).$$

Dans le cas où GL_n *est équipé de la métrique invariante à droite, le projecteur orthogonal est, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{i}}$ *et* $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$*,*

$$P_{\boldsymbol{B}}^{\mathrm{i},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}, \qquad \text{avec} \qquad \boldsymbol{\Lambda}_{\mathrm{i},r} = \mathrm{ddiag}(\boldsymbol{Z}\boldsymbol{Q})\,\mathrm{ddiag}(\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Q})^{-1}.$$

Enfin, pour la variété polaire $\mathcal{P}_n$*, le projecteur orthogonal est donné, pour tous* $\mathcal{B} = (\boldsymbol{S}, \boldsymbol{U}) \in \mathcal{M}_n^{\mathrm{i},\mathcal{P}_n}$ *et* $\mathcal{Z} = (\boldsymbol{Z}_{\boldsymbol{S}}, \boldsymbol{Z}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{P}_n$*, par*

$$P_{\mathcal{B}}^{\mathrm{i},\mathcal{P}_n}(\mathcal{Z}) = \Big(\boldsymbol{Z}_{\boldsymbol{S}} - \boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{Z}_{\boldsymbol{U}} - P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T)\Big),$$
$$\text{avec} \qquad \mathrm{diag}(\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}) = 2\left(\boldsymbol{S} \odot \boldsymbol{Q}^T\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{U}\boldsymbol{Q} + \boldsymbol{S}^2 \odot \boldsymbol{Q}^T\boldsymbol{Q}\right)^{-1} \mathrm{diag}(\mathrm{D}\,\Gamma(\mathcal{B})[\mathcal{Z}]\boldsymbol{Q}),$$

où $\boldsymbol{Q}$ *est défini en prenant* $\boldsymbol{B} = \Gamma(\mathcal{B})$*.*

Démonstration. La preuve de cette proposition est donnée dans l'annexe D. □

Proposition 3.7 (Rétraction)
Étant donné une rétraction R^{GL_n} *sur* GL_n*, on peut définir une rétraction sur* $\mathcal{M}_n^{\mathrm{i}}$*, pour tous* $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{i}}$ *et* $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{i}}$*, par*

$$R_{\boldsymbol{B}}^{\mathrm{i}}(\boldsymbol{\xi}) = \left(\sum_k \mathrm{ddiag}(R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})\boldsymbol{C}_k R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi})^T)^2\right)^{-1/4} R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\boldsymbol{\xi}).$$

Démonstration. La preuve est très semblable à celle de la proposition 3.4. On peut montrer que $R^{\mathrm{i}}_{\boldsymbol{B}}(\boldsymbol{\xi}) \in \mathcal{M}^{\mathrm{i}}_n$ en vérifiant que $\sum_k \operatorname{ddiag}(R^{\mathrm{i}}_{\boldsymbol{B}}(\boldsymbol{\xi})\boldsymbol{C}_k R^{\mathrm{i}}_{\boldsymbol{B}}(\boldsymbol{\xi})^T)^2 = \boldsymbol{I}_n$. Comme $R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi}) = \boldsymbol{B} + \boldsymbol{\xi} + o(\|\xi\|)$, on a

$$\operatorname{ddiag}\left(R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi})\boldsymbol{C}_k R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi})^T\right) = \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T) + 2\operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{C}_k\boldsymbol{B}^T) + o(\|\xi\|).$$

Donc,

$$\begin{aligned}\sum_k \operatorname{ddiag}\left(R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi})\boldsymbol{C}_k R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi})^T\right)^2 &= \sum_k \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T)^2 + 4\operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{Q}) + o(\|\xi\|)\\ &= \boldsymbol{I}_n + o(\|\xi\|).\end{aligned}$$

De ce fait, $R^{\mathrm{i}}_{\boldsymbol{B}}(\boldsymbol{\xi}) = \boldsymbol{B} + \boldsymbol{\xi} + o(\|\xi\|)$, ce qui suffit pour conclure. □

Dans ce cas également, une rétraction issue de la proposition 3.7 (voir la figure 3.3 pour une illustration) s'inscrit dans le contexte de [3]. Dans la proposition 3.7, nous choisissons R^{GL_n} de la même façon que dans la proposition 3.4 pour (*i*) GL_n équipé de la métrique invariante à gauche, (*ii*) GL_n équipé de la métrique invariante à droite et (*iii*) la variété polaire $\mathcal{P}_n$.

Proposition 3.8 (hessiennes riemanniennes)
Étant donné un critère f défini sur GL_n, la hessienne riemannienne sur $\mathcal{M}^{\mathrm{i}}_n$ dans le cas où GL_n est équipé de la métrique invariante à gauche est donnée, pour tous $\boldsymbol{B} \in \mathcal{M}^{\mathrm{i}}_n$ et $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}^{\mathrm{i}}_n$, par

$$\begin{aligned}\operatorname{hess}_{\mathrm{i},\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] = P^{\mathrm{i},\ell}_{\boldsymbol{B}}&\left(\operatorname{hess}_\ell f(\boldsymbol{B})[\boldsymbol{\xi}] + \boldsymbol{B}\operatorname{sym}(\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{B})\boldsymbol{B}^{-1}\boldsymbol{\xi}\right)\\ &- P^{\mathrm{i},\ell}_{\boldsymbol{B}}\left(\operatorname{sym}(\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{Q}^T + \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{i},\ell}\dot{\boldsymbol{Q}}^T + \boldsymbol{B}\operatorname{sym}(\boldsymbol{B}^{-1}\boldsymbol{\xi}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{B})\right),\end{aligned}$$

où $\boldsymbol{\Lambda}_{\mathrm{i},\ell}$ est défini dans la proposition 3.6 pour $\boldsymbol{Z} = \operatorname{grad}_\ell f(\boldsymbol{B})$ et

$$\dot{\boldsymbol{Q}} = \mathrm{D}\,\boldsymbol{Q}[\boldsymbol{\xi}] = \sum_k \boldsymbol{C}_k\left(\boldsymbol{\xi}^T\operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T) + 2\boldsymbol{B}^T\operatorname{ddiag}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T)\right).$$

Lorsque GL_n est équipé de la métrique invariante à droite, la hessienne riemannienne sur $\mathcal{M}^{\mathrm{i}}_n$ est, pour tous $\boldsymbol{B} \in \mathcal{M}^{\mathrm{i}}_n$ et $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}^{\mathrm{i}}_n$,

$$\begin{aligned}\operatorname{hess}_{\mathrm{i},r} f(\boldsymbol{B})[\boldsymbol{\xi}] = P^{\mathrm{i},r}_{\boldsymbol{B}}&\left(\operatorname{hess}_r f(\boldsymbol{B})[\boldsymbol{\xi}] + \boldsymbol{\xi}\boldsymbol{B}^{-1}\operatorname{sym}(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},r})\boldsymbol{B}\right)\\ &- P^{\mathrm{i},r}_{\boldsymbol{B}}\left(\boldsymbol{\Lambda}_{\mathrm{i},r}\left(\dot{\boldsymbol{Q}}^T\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{Q}^T\operatorname{sym}(\boldsymbol{\xi}^T\boldsymbol{B})\right) + \operatorname{sym}(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{B}\right)\end{aligned}$$

où $\boldsymbol{\Lambda}_{\mathrm{i},r}$ est défini dans la proposition 3.6 pour $\boldsymbol{Z} = \operatorname{grad}_r f(\boldsymbol{B})$. Enfin, la hessienne riemannienne sur $\mathcal{M}^{\mathrm{i},\mathcal{P}_n}_n$ est donnée, pour tous $\mathcal{B} = (\boldsymbol{S},\boldsymbol{U}) \in \mathcal{M}^{\mathrm{i},\mathcal{P}_n}_n$ et $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}},\boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{M}^{\mathrm{i},\mathcal{P}_n}_n$, par

$$\begin{aligned}\operatorname{hess}_{\mathrm{i},\mathcal{P}_n} f(\mathcal{B})[\Xi] = P^{\mathrm{i},\mathcal{P}_n}_{\mathcal{B}}&\left(\operatorname{hess}_{\mathcal{P}_n} f(\mathcal{B})[\Xi]\right)\\ &- P^{\mathrm{i},\mathcal{P}_n}_{\mathcal{B}}\left(\boldsymbol{S}\operatorname{sym}\left((\boldsymbol{\xi}_{\boldsymbol{U}}\boldsymbol{Q} + \boldsymbol{U}\dot{\boldsymbol{Q}})\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\right)\boldsymbol{S} + \operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\operatorname{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}), \boldsymbol{0}_n\right)\\ &\quad - P^{\mathrm{i},\mathcal{P}_n}_{\mathcal{B}}\left(\boldsymbol{0}_n, P^{\mathcal{O}_n}_{\boldsymbol{U}}\left(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T + \boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\dot{\boldsymbol{Q}}^T - \boldsymbol{\xi}_{\boldsymbol{U}}\operatorname{sym}(\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T)\right)\right),\end{aligned}$$

où $\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}$ est défini dans la proposition 3.6 pour $\mathcal{Z} = \operatorname{grad}_{\mathcal{P}_n} f(\mathcal{B})$, $\boldsymbol{Q}$ et $\dot{\boldsymbol{Q}}$ sont définis en prenant $\boldsymbol{B} = \Gamma(\mathcal{B})$ et $\boldsymbol{\xi} = \mathrm{D}\,\Gamma(\mathcal{B})[\Xi]$.

Démonstration. La preuve de cette proposition est donnée dans l'annexe D. □

3.3 Contrainte non-holonomique : variété quotient de GL_n

Nous nous concentrons à présent sur la contrainte non-holonomique pour laquelle on exploite la géométrie des classes d'équivalence (1.6) et qui est intégrée dans une variété quotient de GL_n notée $\mathcal{M}_n^{\mathrm{nh}}$. Dans la section 3.3.1, nous décrivons la variété quotient $\mathcal{M}_n^{\mathrm{nh}}$ et nous étudions comment l'équiper d'une structure riemannienne. Rappelons que le cas général des variétés quotientes riemanniennes est traité dans la section 1.4.3. Dans la section 3.3.2, nous appliquons ces résultats lorsque GL_n est équipé de la métrique invariante à gauche, qui engendre une variété quotient riemannienne correctement définie. Dans la section 3.3.3, nous discutons de la possibilité d'optimiser sur la variété quotient $\mathcal{M}_n^{\mathrm{nh}}$ non-riemannienne dans le cas où GL_n est équipé de la métrique invariante à droite. Pour terminer, dans la section 3.3.4, nous expliquons comment exploiter la géométrie de $\mathcal{M}_n^{\mathrm{nh}}$ pour minimiser une fonction objectif sur GL_n qui n'induit pas de fonction objectif proprement définie sur $\mathcal{M}_n^{\mathrm{nh}}$.

3.3.1 Structure de variété quotient riemannienne

La contrainte non-holonomique entraîne la variété quotient $\mathcal{M}_n^{\mathrm{nh}} = \mathrm{GL}_n/\mathcal{D}_n^*$ définie par

$$\mathcal{M}_n^{\mathrm{nh}} = \{\{\boldsymbol{\Sigma B} : \boldsymbol{\Sigma} \in \mathcal{D}_n^*\} : \boldsymbol{B} \in \mathrm{GL}_n\}, \tag{3.19}$$

qui est une variété quotient lisse de dimension $n^2 - n$ [75, théorème 7.10]. On note π la projection canonique de GL_n sur $\mathcal{M}_n^{\mathrm{nh}}$. L'espace vertical $\mathcal{V}_{\boldsymbol{B}}$ à $\boldsymbol{B} \in \mathrm{GL}_n$ est

$$\mathcal{V}_{\boldsymbol{B}} = \{\boldsymbol{\Delta B} : \boldsymbol{\Delta} \in \mathcal{D}_n\}. \tag{3.20}$$

Pour induire une métrique riemannienne sur $\mathcal{M}_n^{\mathrm{nh}}$, il faut que la métrique $\langle \cdot, \cdot \rangle_{\cdot}$ de GL_n satisfasse (1.22), ce qui se traduit par

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{B}} = \langle \boldsymbol{\Sigma\xi}, \boldsymbol{\Sigma\eta} \rangle_{\boldsymbol{\Sigma B}}, \tag{3.21}$$

pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$ et $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^{n \times n}$. L'espace horizontal à $\boldsymbol{B} \in \mathrm{GL}_n$ est alors

$$\mathcal{H}_{\boldsymbol{B}} = \{\boldsymbol{\xi} \in \mathbb{R}^{n \times n} : \langle \boldsymbol{\xi}, \boldsymbol{\Delta B} \rangle_{\boldsymbol{B}} = 0,\ \boldsymbol{\Delta} \in \mathcal{D}_n\}. \tag{3.22}$$

Une rétraction R^{GL_n} sur GL_n induit une rétraction sur $\mathcal{M}_n^{\mathrm{nh}}$ définie avec (1.26) si, pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$ et $\boldsymbol{\xi} \in \mathcal{H}_{\boldsymbol{B}}$, il existe $\widetilde{\boldsymbol{\Sigma}} \in \mathcal{D}_n^*$ tel que

$$R^{\mathrm{GL}_n}_{\boldsymbol{\Sigma B}}(\boldsymbol{\Sigma\xi}) = \widetilde{\boldsymbol{\Sigma}} R^{\mathrm{GL}_n}_{\boldsymbol{B}}(\boldsymbol{\xi}). \tag{3.23}$$

Une fonction objectif f sur GL_n induit une fonction objectif $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$ telle que $f = \overline{f} \circ \pi$ si elle est invariante le long de chaque classe d'équivalence, *i.e.*, si $f(\boldsymbol{B}) = f(\boldsymbol{\Sigma B})$ pour tous $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$. Cette condition correspond à la propriété 2.2 introduite dans la section 2.4.

3.3.2 Métrique invariante à gauche

La métrique invariante à gauche de GL_n définie dans (1.47) induit une métrique riemannienne sur $\mathcal{M}_n^{\mathrm{nh}}$, qui devient alors une variété quotient riemannienne. Pour être en mesure d'optimiser sur $\mathcal{M}_n^{\mathrm{nh}}$ dans ce cas, il suffit de définir l'espace horizontal, le projecteur orthogonal et une rétraction. Les autres objets requis, *i.e.*, la connection de Levi-Civita, le gradient riemannien, la hessienne riemannienne et le transport de vecteurs, sont obtenus avec (1.23), (1.24), (1.25) et (1.27). L'espace horizontal et son projecteur orthogonal sont donnés dans la proposition 3.9. Pour la rétraction, nous pouvons utiliser l'exponentielle riemannienne sur $\mathcal{M}_n^{\mathrm{nh}}$ associée aux géodésiques, qui sont donnés dans la proposition 3.10.

Proposition 3.9 (Espace horizontal et projecteur orthogonal)
L'espace horizontal $\mathcal{H}_{\boldsymbol{B}}$ à $\boldsymbol{B} \in \mathrm{GL}_n$ équipé de la métrique invariante à gauche est

$$\mathcal{H}_{\boldsymbol{B}} = (\mathcal{V}_{\boldsymbol{B}})^{\perp,\ell} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{n\times n} : \operatorname{ddiag}\left((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{\xi}\boldsymbol{B}^T\right) = \mathbf{0}_n \right\}.$$

Le projecteur orthogonal à $\boldsymbol{B} \in \mathrm{GL}_n$ de $\mathbb{R}^{n\times n}$ sur $\mathcal{H}_{\boldsymbol{B}}$ est défini, pour tout $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$, par

$$P_{\boldsymbol{B}}^{\mathrm{nh},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{nh},\ell}\boldsymbol{B}, \text{ avec } \quad \operatorname{diag}(\boldsymbol{\Lambda}_{\mathrm{nh},\ell}) = \left((\boldsymbol{B}\boldsymbol{B}^T)^{-1} \odot \boldsymbol{B}\boldsymbol{B}^T\right)^{-1} \operatorname{diag}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{Z}\boldsymbol{B}^T),$$

où $\boldsymbol{\Lambda}_{\mathrm{nh},\ell} \in \mathcal{D}_n$.

Démonstration. L'ensemble $\mathcal{H}_{\boldsymbol{B}}$ est de dimension $n^2 - n$ et, pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\Lambda} \in \mathcal{D}_n$ et $\boldsymbol{\xi} \in \mathbb{R}^{n\times n}$, on a

$$\langle \boldsymbol{\xi}, \boldsymbol{\Lambda}\boldsymbol{B} \rangle_{\boldsymbol{B}}^{\ell} = \operatorname{tr}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{\Lambda}) = \operatorname{tr}(\operatorname{ddiag}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}).$$

De ce fait, $\langle \boldsymbol{\xi}, \boldsymbol{\Lambda}\boldsymbol{B} \rangle_{\boldsymbol{B}}^{\ell} = 0$ pour tout $\boldsymbol{\Lambda} \in \mathcal{D}_n$ si et seulement si $\operatorname{ddiag}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{\xi}\boldsymbol{B}^T) = \mathbf{0}_n$. Pour le projecteur orthogonal, on sait que $P_{\boldsymbol{B}}^{\mathrm{nh},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{nh},\ell}\boldsymbol{B}$ et l'équation

$$\operatorname{ddiag}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}P_{\boldsymbol{B}}^{\mathrm{nh},\ell}(\boldsymbol{Z})\boldsymbol{B}^T) = \mathbf{0}_n$$

peut être vectorisée par

$$\left((\boldsymbol{B}\boldsymbol{B}^T)^{-1} \odot \boldsymbol{B}\boldsymbol{B}^T\right) \operatorname{diag}(\boldsymbol{\Lambda}_{\mathrm{nh},\ell}) = \operatorname{diag}((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{Z}\boldsymbol{B}^T).$$

Comme $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{B}\boldsymbol{B}^T$ et $(\boldsymbol{B}\boldsymbol{B}^T)^{-1}$ sont positives définies et le théorème du produit de Schur assure que $((\boldsymbol{B}\boldsymbol{B}^T)^{-1} \odot \boldsymbol{B}\boldsymbol{B}^T)$ est inversible, ce qui permet de conclure. □

Proposition 3.10 (Géodésiques)
Pour tous $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\xi} \in \mathcal{H}_{\boldsymbol{B}}$, $\gamma_{\mathrm{nh},\ell} : \mathbb{R} \to \mathcal{M}_n^{\mathrm{nh}}$ défini par

$$\gamma_{\mathrm{nh},\ell}(t) = \pi(\gamma_{\ell}(t)),$$

où γ_ℓ est défini dans (1.50)*, est un géodésique complet sur $\mathcal{M}_n^{\mathrm{nh}}$.*

Démonstration. La dérivée de $\gamma_\ell(t)$ est donnée par

$$\dot{\gamma}_\ell(t) = \boldsymbol{B} \exp\left(t(\boldsymbol{B}^{-1}\boldsymbol{\xi})^T\right) \boldsymbol{B}^{-1}\boldsymbol{\xi} \exp\left(t(\boldsymbol{B}^{-1}\boldsymbol{\xi} - (\boldsymbol{B}^{-1}\boldsymbol{\xi})^T)\right).$$

Il suit que, pour tout $t \in \mathbb{R}$,

$$\operatorname{ddiag}\left((\gamma_\ell(t)\gamma_\ell(t)^T)^{-1}\dot{\gamma}_\ell(t)\gamma_\ell(t)^T\right) = \operatorname{ddiag}\left((\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{\xi}\boldsymbol{B}^T\right) = \mathbf{0}_n.$$

De ce fait, $\dot{\gamma}_\ell(t) \in \mathcal{H}_{\gamma_\ell(t)}$, ce qui montre que la courbe $\gamma_\ell(t)$ reste horizontale dans GL_n équipé de la métrique invariante à gauche (cela est garanti par la théorie des submersions riemanniennes [2]). Comme $\gamma_\ell(t)$ est un géodésique complet dans GL_n, $\pi(\gamma_\ell(t))$ est un géodésique complet dans $\mathcal{M}_n^{\mathrm{nh}}$ [52, proposition 2.109]. □

3.3.3 Métrique invariante à droite

Considérons le cas où GL_n est équipé de la métrique invariante à droite définie dans (1.47). Cette métrique ne vérifie pas (3.21) et n'induit donc pas de métrique riemannienne sur la variété quotient $\mathcal{M}_n^{\mathrm{nh}}$. Soit une fonction objectif f définie sur GL_n qui induit une fonction objectif $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$. Nous nous intéressons à la possibilité de définir un algorithme d'optimisation pour minimiser $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$ même si le gradient de $\overline{f}$ n'est pas défini sur la variété quotient non-riemannienne $\mathcal{M}_n^{\mathrm{nh}}$. En particulier, nous définissons des objets ressemblant à des objets riemanniens (gradient, rétraction et transport de vecteurs) qui ne possèdent pas les bonnes propriétés individuellement, mais qui permettent de définir une itération d'optimisation sur $\mathcal{M}_n^{\mathrm{nh}}$. À notre connaissance, c'est la première fois que de tels objets sont introduits et étudiés. Notons que nous prenons uniquement en compte les méthodes de type gradient dans cette section, *i.e.*, descente de gradient, gradient conjugué et quasi-Newton.

Commençons par noter que pour tout $\boldsymbol{B} \in \mathrm{GL}_n$, on peut définir le complémentaire orthogonal $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ à l'espace vertical $\mathcal{V}_{\boldsymbol{B}}$ selon la métrique invariante à droite et le projecteur orthogonal associé. Ils sont tous deux donnés dans la proposition 3.11. Remarquons que $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ est isomorphe à l'espace tangent $T_{\pi(\boldsymbol{B})}\mathcal{M}_n^{\mathrm{nh}}$ de $\pi(\boldsymbol{B})$ dans $\mathcal{M}_n^{\mathrm{nh}}$.

Proposition 3.11 (Complémentaire orthogonal à l'espace vertical et projecteur orthogonal) *Pour tout $\boldsymbol{B} \in \mathrm{GL}_n$ équipé de la métrique invariante à droite, le complémentaire orthogonal à l'espace vertical $\mathcal{V}_{\boldsymbol{B}}$ est*

$$(\mathcal{V}_{\boldsymbol{B}})^{\perp,r} = \left\{\boldsymbol{\xi} \in \mathbb{R}^{n\times n} : \operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^{-1}) = \mathbf{0}_n\right\}.$$

Le projecteur orthogonal à $\boldsymbol{B} \in \mathrm{GL}_n$ de $\mathbb{R}^{n\times n}$ sur $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ est défini, pour tout $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$, par

$$P_{\boldsymbol{B}}^{\mathrm{nh},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \operatorname{ddiag}(\boldsymbol{Z}\boldsymbol{B}^{-1})\boldsymbol{B}.$$

Démonstration. L'ensemble $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ est de dimension $n^2 - n$ et, pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\Delta} \in \mathcal{D}_n$ et $\boldsymbol{\xi} \in \mathbb{R}^{n\times n}$, on a

$$\langle \boldsymbol{\xi}, \boldsymbol{\Delta}\boldsymbol{B}\rangle_{\boldsymbol{B}}^r = \operatorname{tr}(\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Delta}) = \operatorname{tr}(\operatorname{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{\Delta}).$$

De ce fait, $\langle \boldsymbol{\xi}, \boldsymbol{\Delta B} \rangle_{\boldsymbol{B}}^{r} = 0$ pour tout $\boldsymbol{\Delta} \in \mathcal{D}_n$ si et seulement si $\operatorname{ddiag}(\boldsymbol{\xi B}^{-1}) = \mathbf{0}_n$. Il reste à déterminer le projecteur orthogonal sur $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$. On sait que $P_{\boldsymbol{B}}^{\mathrm{nh},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda B}$ avec $\boldsymbol{\Lambda} \in \mathcal{D}_n$ et résoudre l'équation $\operatorname{ddiag}(P_{\boldsymbol{B}}^{\mathrm{nh},r}(\boldsymbol{Z})\boldsymbol{B}^{-1}) = \mathbf{0}_n$ donne le résultat. □

Pour optimiser $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$ alors que GL_n est équipé de la métrique invariante à droite, il faut définir un opérateur qui remplit le rôle du gradient, *i.e.*, un champ de vecteurs sur $\mathcal{M}_n^{\mathrm{nh}}$ qui permet d'obtenir une direction de descente de $\overline{f}$.

Proposition 3.12 (Pseudo-gradient)
Étant donné une fonction $\overline{f}$ définie sur $\mathcal{M}_n^{\mathrm{nh}}$ (induite par $f = \overline{f} \circ \pi$ sur GL_n), un opérateur approprié pour remplir un rôle similaire au gradient est défini, à $\mathcal{B} = \pi(\boldsymbol{B}) \in \mathcal{M}_n^{\mathrm{nh}}$, par $\operatorname{grad}_r f(\boldsymbol{B}) \in (\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$, i.e.,

$$\widetilde{\operatorname{grad}}\overline{f}(\mathcal{B}) = \mathrm{D}\,\pi(\boldsymbol{B})[\operatorname{grad}_r f(\boldsymbol{B})],$$

où $\operatorname{grad}_r f(\boldsymbol{B})$ *est défini dans* (1.52).

Démonstration. On peut tout d'abord montrer que $\operatorname{grad}_r f(\boldsymbol{B}) \in (\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ en remarquant que $\operatorname{ddiag}(\operatorname{grad}_r f(\boldsymbol{B})\boldsymbol{B}^{-1}) = \mathbf{0}_n$. Comme $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ est isomorphe à $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{nh}}$ pour $\mathcal{B} = \pi(\boldsymbol{B})$, $\operatorname{grad}_r f(\boldsymbol{B})$ permet de définir $\widetilde{\operatorname{grad}}\overline{f}(\mathcal{B})$ de façon unique. Enfin, comme $\operatorname{grad}_r f(\boldsymbol{B})$ permet de définir une direction de descente de $f = \overline{f} \circ \pi$ sur GL_n, $\widetilde{\operatorname{grad}}\overline{f}(\mathcal{B})$ permet d'obtenir une direction de descente de $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$. □

Étant donné une fonction objectif f sur GL_n qui induit une fonction objectif $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$, on peut montrer que, pour tous $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$, on a la relation

$$\operatorname{grad}_r f(\boldsymbol{\Sigma B}) = \boldsymbol{\Sigma}^{-1} \operatorname{grad}_r f(\boldsymbol{B})\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}. \tag{3.24}$$

De ce fait, lorsqu'il s'agit de définir une direction de descente à partir du pseudo-gradient sur $\mathcal{M}_n^{\mathrm{nh}}$ défini dans la proposition 3.12, le vecteur $\boldsymbol{\xi}$ dans $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ à $\boldsymbol{B}$ correspond au vecteur $\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}$ dans $(\mathcal{V}_{\boldsymbol{\Sigma B}})^{\perp,r}$ à $\boldsymbol{\Sigma B}$. En conséquence, une rétraction R^{GL_n} sur GL_n qui satisfait (3.23) ne semble pas appropriée car

$$\pi\left(R_{\boldsymbol{\Sigma B}}^{\mathrm{GL}_n}(\operatorname{grad}_r f(\boldsymbol{\Sigma B}))\right) \neq \pi\left(R_{\boldsymbol{B}}^{\mathrm{GL}_n}(\operatorname{grad}_r f(\boldsymbol{B}))\right). \tag{3.25}$$

Dans le but d'obtenir des algorithmes qui ne dépendent pas du choix du représentant $\boldsymbol{B}$ d'une classe d'équivalence, nous pouvons utiliser un opérateur similaire à une rétraction à la place, comme proposé dans la proposition 3.13.

Proposition 3.13 (Pseudo-rétraction)
Dans le contexte de l'optimisation sur $\mathcal{M}_n^{\mathrm{nh}}$ avec le pseudo-gradient de la proposition 3.12, une pseudo-rétraction appropriée $\widetilde{R}$ sur GL_n, qui induit une pseudo-rétraction sur $\mathcal{M}_n^{\mathrm{nh}}$, est définie, pour tous $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{\xi} \in (\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$, par

$$\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi}) = \exp\left(\boldsymbol{\Lambda}(\boldsymbol{B})\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Lambda}(\boldsymbol{B})^{-1} - (\boldsymbol{\xi}\boldsymbol{B}^{-1})^T\right)\exp\left((\boldsymbol{\xi}\boldsymbol{B}^{-1})^T\right)\boldsymbol{B},$$

où $\boldsymbol{\Lambda}(\boldsymbol{B}) = \operatorname{ddiag}(\boldsymbol{B}\boldsymbol{B}^T)$.

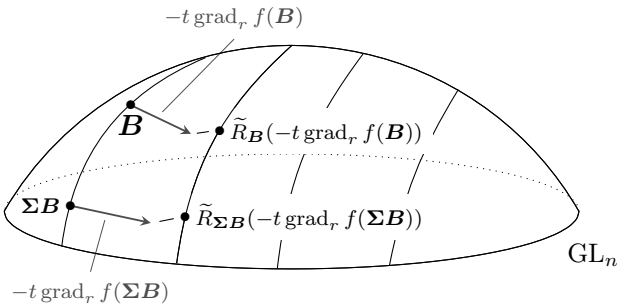

FIGURE *3.4 – Schéma d'illustration des outils développés pour optimiser sur le quotient* $\mathcal{M}_n^{\mathrm{nh}}$ *dans le cas de la métrique invariante à droite. La pseudo-rétraction de la proposition 3.13 permet d'obtenir la même classe d'équivalence à partir du pseudo-gradient de la proposition 3.12 pour tous les éléments de la classe d'équivalence* $\pi(\boldsymbol{B})$.

Démonstration. Comme le vecteur $\boldsymbol{\xi}$ de $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ correspond au vecteur $\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}$ dans $(\mathcal{V}_{\boldsymbol{\Sigma B}})^{\perp,r}$, l'opérateur $\widetilde{R}$ doit vérifier

$$\pi\left(\widetilde{R}_{\boldsymbol{\Sigma B}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B})\right) = \pi\left(\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})\right),$$

pour tous $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{\xi} \in (\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ et $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$. En effet, cette propriété permet d'assurer que différents représentants d'une même classe de $\mathcal{M}_n^{\mathrm{nh}}$ engendrent tous la même classe d'équivalence. On peut montrer que

$$\widetilde{R}_{\boldsymbol{\Sigma B}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}) = \boldsymbol{\Sigma}\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi}),$$

ce qui permet de conclure. □

Par conséquent, étant donné $\mathcal{B}_i = \pi(\boldsymbol{B}_i)$ dans $\mathcal{M}_n^{\mathrm{nh}}$, l'itération de descente de gradient "pseudo-riemannienne" du critère $\overline{f}$ sur $\mathcal{M}_n^{\mathrm{nh}}$ induit par $f = \overline{f} \circ \pi$ sur GL_n est

$$\mathcal{B}_{i+1} = \pi\left(\widetilde{R}_{\boldsymbol{B}_i}(-t_i\,\mathrm{grad}_r\, f(\boldsymbol{B}_i))\right). \tag{3.26}$$

Une illustration est proposée dans la figure 3.4. Pour utiliser des méthodes d'optimisation plus sophistiquées, telles qu'un gradient conjugué ou une méthode quasi-Newton, il nous reste à définir un opérateur similaire au transport de vecteurs sur $\mathcal{M}_n^{\mathrm{nh}}$. La proposition 3.14 définit un tel opérateur à travers sa représentation sur GL_n.

Proposition 3.14 (Pseudo-transport de vecteur)
Un opérateur approprié pour remplir le rôle du transport de vecteur sur GL_n *est défini, pour tous* $\boldsymbol{B} \in \mathrm{GL}_n$ *et* $\boldsymbol{\xi}, \boldsymbol{\eta} \in (\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$*, par*

$$\widetilde{\mathcal{T}}(\boldsymbol{B}, \boldsymbol{\xi}, \boldsymbol{\eta}) = P^{\mathrm{nh},r}_{\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})}\left(\boldsymbol{\eta}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})^T\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})\right).$$

Démonstration. $\widetilde{\mathcal{T}}(\boldsymbol{B}, \boldsymbol{\xi}, \boldsymbol{\eta})$ vise à transporter $\boldsymbol{\eta}$ depuis $(\mathcal{V}_{\boldsymbol{B}})^{\perp,r}$ dans $(\mathcal{V}_{\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})})^{\perp,r}$. Étant donné $\boldsymbol{\Sigma} \in \mathcal{D}_n^*$, les vecteurs tangents $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ au point $\boldsymbol{B}$ correspondent aux vecteurs tangents

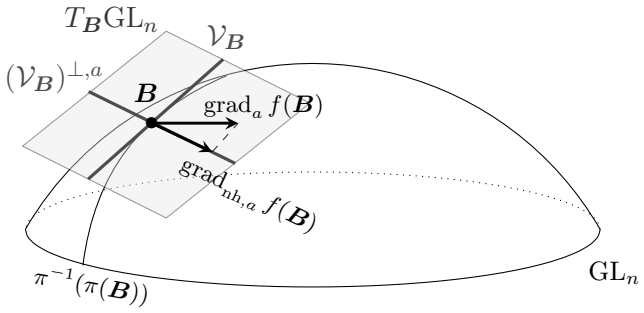

FIGURE *3.5 – Schéma d'illustration de la modification* $\mathrm{grad}_{\mathrm{nh},a}\, f(\boldsymbol{B})$ *appliquée au gradient riemannien* $\mathrm{grad}_a\, f(\boldsymbol{B})$ *d'une fonction objectif* f *à* $\boldsymbol{B} \in \mathrm{GL}_n$*, qui n'est pas invariante le long de la classe d'équivalence* $\pi^{-1}(\pi(\boldsymbol{B}))$*, dans le but d'annuler l'action des matrices diagonales. Le caractère* a *correspond à* ℓ *ou* r *selon la métrique choisie.*

$\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}$ et $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}$ au point $\boldsymbol{\Sigma}\boldsymbol{B}$. On a aussi $\widetilde{R}_{\boldsymbol{\Sigma}\boldsymbol{B}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}) = \boldsymbol{\Sigma}\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})$. Il suit que $\widetilde{\mathcal{T}}(\boldsymbol{\Sigma}\boldsymbol{B}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B})$ doit être le vecteur de $(\mathcal{V}_{\boldsymbol{\Sigma}\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})})^{\perp,r}$ qui correspond à $\widetilde{\mathcal{T}}(\boldsymbol{B}, \boldsymbol{\xi}, \boldsymbol{\eta})$. Cela se traduit par la condition

$$\widetilde{\mathcal{T}}(\boldsymbol{\Sigma}\boldsymbol{B}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\boldsymbol{B}^{-1}\boldsymbol{\Sigma}^2\boldsymbol{B}) = \boldsymbol{\Sigma}^{-1}\widetilde{\mathcal{T}}(\boldsymbol{B}, \boldsymbol{\xi}, \boldsymbol{\eta})\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi})^{-1}\boldsymbol{\Sigma}^2\widetilde{R}_{\boldsymbol{B}}(\boldsymbol{\xi}).$$

On peut vérifier que $\widetilde{\mathcal{T}}$ possède bien cette propriété, ce qui justifie son utilisation. □

3.3.4 Fonctions objectifs mal définies

Pour conclure notre étude de la contrainte non-holonomique, nous considérons la possibilité d'exploiter la géométrie de $\mathcal{M}_n^{\mathrm{nh}}$ pour optimiser une fonction objectif f définie sur GL_n qui n'induit pas de fonction objectif sur $\mathcal{M}_n^{\mathrm{nh}}$, *i.e.*, qui n'est pas invariante le long des classes d'équivalence. Dans cette section, nous nous limitons également aux méthodes d'optimisation de type gradient. Pour la métrique invariante à gauche comme pour celle invariante à droite, toutes deux définies dans (1.47), les gradients riemanniens $\mathrm{grad}_\ell\, f(\boldsymbol{B})$ et $\mathrm{grad}_r\, f(\boldsymbol{B})$ de f à $\boldsymbol{B} \in \mathrm{GL}_n$ ont des composantes non nulles sur l'espace vertical $\mathcal{V}_{\boldsymbol{B}}$. Dans le but de neutraliser l'action de $\mathcal{D}_n^*$, l'idée est de modifier les gradients riemanniens en annulant leur composante sur l'espace vertical $\mathcal{V}_{\boldsymbol{B}}$ en utilisant les projecteurs orthogonaux définis dans les propositions 3.9 et 3.11, comme illustré sur la figure 3.5. Les gradients riemanniens modifiés à $\boldsymbol{B} \in \mathrm{GL}_n$ sont alors définis par

$$\mathrm{grad}_{\mathrm{nh},\ell}\, f(\boldsymbol{B}) = P^{\mathrm{nh},\ell}(\mathrm{grad}_\ell\, f(\boldsymbol{B})) \qquad \text{et} \qquad \mathrm{grad}_{\mathrm{nh},r}\, f(\boldsymbol{B}) = P^{\mathrm{nh},r}(\mathrm{grad}_r\, f(\boldsymbol{B})). \quad (3.27)$$

Pour définir des algorithmes d'optimisation dans cette situation, nous utilisons ces gradients modifiés et nous les combinons avec les autres outils de GL_n équipé de la métrique invariante à gauche ou à droite. Notons que dans les deux cas, ces gradients modifiés ne sont pas invariants par rapport au choix du représentant $\boldsymbol{B}$ d'une classe d'équivalence $\pi^{-1}(\pi(\boldsymbol{B}))$ en général. De ce fait, dans cette section, les algorithmes d'optimisation "pseudo-riemanniens" résultants dépendent toujours du choix du représentant d'une classe d'équivalence.

Chapitre 4

Illustrations numériques

Sommaire

Ce chapitre contient l'étude des performances des méthodes de diagonalisation conjointe approximée obtenues à partir des résultats théoriques proposés dans les chapitres 2 et 3 de ce manuscrit. Dans la section 4.1, nous présentons les algorithmes testés : nous indiquons comment nous les désignons et nous précisons l'ensemble des paramètres utilisés pour ces illustrations numériques. Dans la section 4.2, nous étudions les performances des méthodes de diagonalisation conjointe approximée sur des matrices symétriques positives définies synthétiques pour lesquelles la solution est connue. Nous comparons nos algorithmes entre eux et avec les algorithmes de l'état de l'art NOJoB [42], uwedge [107] et jadiag [93]. Dans la section 4.3, nous nous intéressons à l'analyse de données électroencéphalographiques. Nous exploitons nos méthodes de diagonalisation conjointe approximée pour effectuer la séparation de sources des données recueillies sur trois sujets. Nous comparons les résultats obtenus pour les critères construits à partir des divergences de la section 2.3 du chapitre 2 grâce à la définition d'une mesure objective de la performance de la séparation.

4.1 Algorithmes testés

Nos algorithmes sont obtenus par l'optimisation des critères construits à partir des divergences de la section 2.3 avec les stratégies d'optimisation pour la diagonalisation conjointe de la section 2.2 en utilisant un cadre d'optimisation riemannienne du chapitre 3. Pour désigner un algorithme, on utilise le système suivant :

- On indique d'abord la divergence utilisée avec F pour la distance de Frobenius, ℓKL, rKL et sKL pour les mesures de Kullback-Leibler gauche, droite et symétrisée, αLD pour la divergence log-det α, R pour la distance riemannienne naturelle, LE pour la distance log-euclidienne et W pour la distance de Wasserstein.

- On précise ensuite la variété avec $\mathcal{P}$ pour la variété polaire $\mathcal{P}_n$, GL^{ℓ} et GL^{r} pour GL_n équipé des métriques invariantes à gauche et à droite, respectivement.
- On note la contrainte avec o pour la contrainte oblique, i pour la contrainte intrinsèque et nh pour la contrainte non-holonomique.
- On indique enfin la stratégie d'optimisation exploitée par a pour la stratégie directe (a), b pour celle indirecte (b) et c pour celle indirecte inverse (c).

Par exemple, l'algorithme qui optimise le critère de Frobenius avec la stratégie directe (a) sur la variété polaire avec la contrainte oblique est noté F-$\mathcal{P}$-o-a. Rappelons que pour la stratégie indirecte inverse (c), nous n'ajoutons pas de contrainte supplémentaire. L'algorithme qui optimise le critère de Frobenius avec la stratégie indirecte inverse (c) sur la variété polaire est donc noté F-$\mathcal{P}$-c. Pour la divergence log-det α, nous considérons uniquement la valeur $\alpha = 0$ dans ce chapitre. Avec les stratégies indirecte (b) et indirecte inverse (c), nous prenons la matrice diagonale la plus proche selon la divergence utilisée par le critère à optimiser.

L'optimisation est effectuée avec la librairie *manopt* [28] dans Matlab. Pour l'ensemble de ces expériences numériques, le critère d'arrêt pour l'itéré $\boldsymbol{B}_i$ est défini comme $\|\boldsymbol{B}_{i-1}\boldsymbol{B}_i^{-1} - \boldsymbol{I}_n\|_{\mathrm{F}}^2/n$ et sa tolérance est fixée à $\varepsilon = 10^{-9}$. Mis à part lorsque c'est indiqué, les alogrithmes basés sur les stratégies directe (a) et indirecte (b) utilisent l'algorithme d'optimisation RLBFGS développé dans [61] et ceux basés sur la stratégie indirecte inverse (c) utilisent l'algorithme d'optimisation à région de confiance décrit dans [2] (l'algorithme RLBFGS ne peut être utilisé dans ce cas là comme expliqué dans le chapitre 2). Dans le cas des critères pour lesquels nous n'avons pas de formule analytique explicite pour la matrice diagonale la plus proche, *i.e.*, pour la divergence log-det α, la distance riemannienne naturelle et la distance de Wasserstein, nous l'évaluons par le biais d'une méthode riemannienne de gradient conjugué sur $\mathcal{D}_n^{++}$.

4.2 Expériences avec des données simulées

Dans cette section, nous testons les méthodes de diagonalisation conjointe approximée sur des matrices symétriques positives définies simulées. La section 4.2.1 contient le modèle de simulation des données et la mesure de performance des algorithmes en terme de précision. La section 4.2.2 présente les résultats obtenus. Nous comparons tout d'abord les performances de nos algorithmes avec celles des algorithmes de l'état de l'art NOJoB, uwedge et jadiag. Nous comparons ensuite les différents cadres d'optimisation riemannienne proposés dans le chapitre 3, les trois stratégies d'optimisation pour la diagonalisation conjointe considérées dans la section 2.2 et les critères construits à partir des divergences de la section 2.3.

4.2.1 Modèle et mesure de performance

Nous simulons des ensembles de $K = 50$ matrices $\boldsymbol{C}_k$ symétriques positives définies de taille $n \times n$, avec $n = 16$, selon le modèle

$$\boldsymbol{C}_k = \boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{A}^T + \frac{1}{\sigma}\boldsymbol{E}_k\boldsymbol{\Delta}_k\boldsymbol{E}_k^T, \tag{4.1}$$

où les matrices $\boldsymbol{A}$ et $\boldsymbol{E}_k$ sont des matrices aléatoires dont les éléments sont indépendamment et identiquement distribués (i.i.d.) et générés depuis la distribution normale standard. On s'assure de plus que le conditionnement de $\boldsymbol{A}$ par rapport à l'inversion [54] est dans l'intervalle $[1, 10]$, ce qui définit un conditionnement très bon. Les matrices diagonales $\boldsymbol{\Lambda}_k$ et $\boldsymbol{\Delta}_k$, dont les éléments sont i.i.d., contiennent les informations des sources de signal et de bruit, respectivement. Leur $p^{\text{ème}}$ élément est généré depuis la distribution χ^2 d'espérance 1 et divisé par p. Cela simule des sources avec une grande variabilité d'énergie, ce qui est souvent le cas des données réelles, notamment pour l'électroencéphalographie. Enfin, le paramètre $\sigma \in \{10, 50, 100, 500, 1000\}$ définit le rapport signal sur bruit des données. Avant d'effectuer la diagonalisation conjointe approximée, nous faisons un blanchiment des matrices $\boldsymbol{C}_k$ avec l'inverse de la racine carré de leur moyenne arithmétique et nous fixons l'ensemble des poids w_k à 1. L'ensemble des algorithmes est initialisé avec la matrice identité.

Pour évaluer la performance des algorithmes, nous utilisons une mesure standard de la précision pour les méthodes de diagonalisation conjointe approximée, le critère de Moreau-Amari [84], qui est donné par

$$I_{\text{M-A}}(M) = \frac{1}{2n(n-1)} \sum_{p=1}^{n} \left(\frac{\sum_{q=1}^{n} |M_{pq}|}{\max\limits_{1 \leq q \leq n} |M_{pq}|} + \frac{\sum_{q=1}^{n} |M_{qp}|}{\max\limits_{1 \leq q \leq n} |M_{qp}|} - 2 \right), \tag{4.2}$$

avec $\boldsymbol{M} = \boldsymbol{B}\boldsymbol{A}$, où $\boldsymbol{B}$ est le diagonalisateur conjoint estimé et $\boldsymbol{A}$ est la vraie matrice de mixage du signal dans (4.1). Le critère de performance $I_{\text{M-A}}$ est une mesure dans $[0, 1]$, où 0 correspond au cas où on a parfaitement retrouvé la solution. Une valeur de $I_{\text{M-A}}$ inférieure à -17 dB est généralement considérée très satisfaisante. Pour analyser les résultats, lorsque plus de deux méthodes sont à comparer simultanément, on effectue d'abord un test de Friedmann, une analyse de la variance à mesures répétées non-paramétrique, qui permet de vérifier si les critères de Moreau-Amari des méthodes dans un groupe de comparaisons sont équivalents. Si le test de Friedman est significatif, ce qui signifie que la tendance centrale d'au moins une des méthodes est statistiquement dominante, on réalise toutes les comparaisons *post hoc* par paires avec le test non-paramétrique de Wilcoxon. Afin de prendre en compte le nombre de tests effectués, nous appliquons une correction de Bonferroni. Pour les tests de Friedman, l'erreur de première espèce, qui définit le seuil de significativité de la valeur-p, est fixée à $2,0 \cdot 10^{-3}$. Cela correspond à $\alpha/25$, où α est l'erreur de première espèce tolérée, fixée partout à 5%, et 25 est le nombre d'analyses de la variance. Pour les tests de Wilcoxon, la correction de Bonferroni dépend du nombre de comparaisons par paires réalisées dans un groupe. Ce nombre est variable et on précise le seuil fixé dans chaque cas. Les corrections de Bonferroni garantissent que la probabilité de rejeter ne serait-ce qu'une seule hypothèse nulle parmi toutes les hypothèses testées simultanément est inférieure à α [112].

4.2.2 Résultats

Comparaisons avec l'état de l'art Nous commençons par vérifier que les outils d'optimisation riemannienne que nous proposons sont appropriés pour résoudre le problème de diagonalisation conjointe approximée. Pour ce faire, nous comparons les performances des

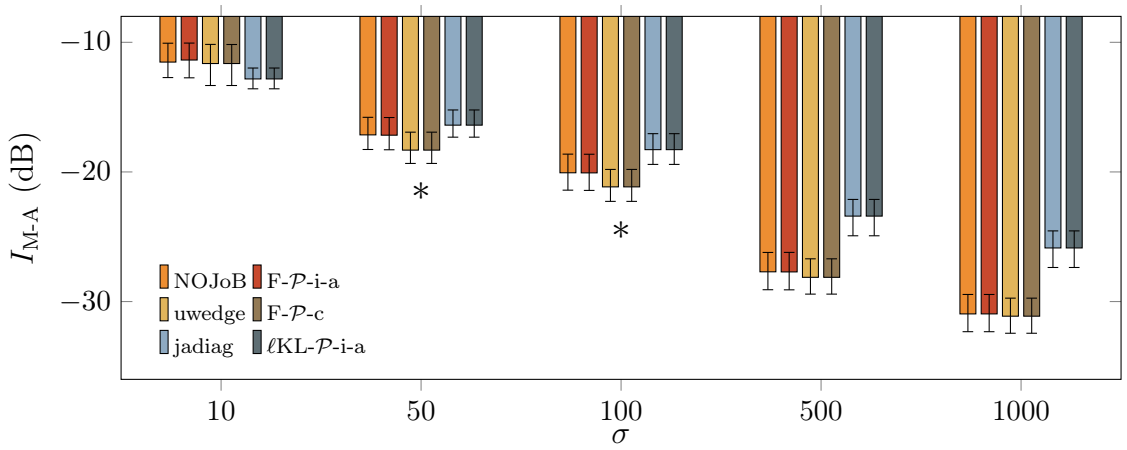

FIGURE *4.1 – Comparaisons avec l'état de l'art : médianes, premiers et neuvièmes déciles (barres d'erreurs) estimés sur* 50 *essais des critères de Moreau-Amari en fonction du rapport signal sur bruit* σ *pour les méthodes de l'état de l'art NOJoB, uwedge et jadiag et celles proposées correspondantes F-*$\mathcal{P}$*-i-a, F-*$\mathcal{P}$*-c et* ℓ*KL-*$\mathcal{P}$*-i-a. Le signe* $*$ *indique une différence significative (*$p < 3,3 \cdot 10^{-3}$*) selon le test de Wilcoxon entre NOJoB et F-*$\mathcal{P}$*-i-a.*

algorithmes de l'état de l'art NOJoB, uwedge et jadiag avec celles des algorithmes correspondants F-$\mathcal{P}$-i-a, F-$\mathcal{P}$-c et ℓKL-$\mathcal{P}$-i-a. La figure 4.1 présente les médianes et les premiers et neuvièmes déciles des critères de Moreau-Amari obtenus en fonction des cinq rapports signal sur bruit. Comme nous comparons chaque algorithme de l'état de l'art avec son équivalent riemannien, nous pouvons directement effectuer les tests de Wilcoxon pour chaque valeur du rapport signal sur bruit. Dans ce cas, le seuil de significativité est fixé à $3,3 \cdot 10^{-3}$, qui correspond à $\alpha/15$, où 15 est le nombre de tests faits. Les tests de Wilcoxon ne révèlent pas de différences entre les performances de F-$\mathcal{P}$-c et de uwedge et entre celles de ℓKL-$\mathcal{P}$-i-a et de jadiag. En revanche, ces tests montrent des différences significatives entre F-$\mathcal{P}$-i-a et NOJoB pour $\sigma = 50$ et 100. Dans ces deux cas, F-$\mathcal{P}$-i-a conduit plus régulièrement à des solutions plus précises. Cependant, les différences entre les performances de ces deux méthodes sont relativement faibles, *i.e.*, on observe des différences inférieures à 0,1 dB entre les médianes et les déciles des deux algorithmes. On peut donc conclure que l'optimisation riemannienne est une solution viable pour traiter le problème de diagonalisation conjointe approximée.

Comparaisons des cadres d'optimisation Nous comparons ensuite les différents cadres d'optimisation pour la diagonalisation conjointe approximée proposés dans le chapitre 3. La figure 4.2 donne les médianes et les premiers et neuvièmes déciles des mesures de Moreau-Amari en fonction du rapport signal sur bruit pour les algorithmes qui optimisent les critères de Frobenius (en haut de la figure) et de Kullback-Leibler gauche (en bas de la figure) avec la stratégie d'optimisation directe (a). On prend ces deux critères du fait que ce sont ceux majoritairement utilisés jusque là. Pour chaque critère et chaque valeur du rapport signal sur bruit, nous commençons par effectuer un test de Friedman sur les huit méthodes. Dans le cas où ce dernier est significatif, nous faisons toutes les comparaisons *post hoc* deux à deux avec le test de Wilcoxon. Son seuil de significativité est fixé à $3,6 \cdot 10^{-4}$, qui correspond à $\alpha/(5 \times 28)$,

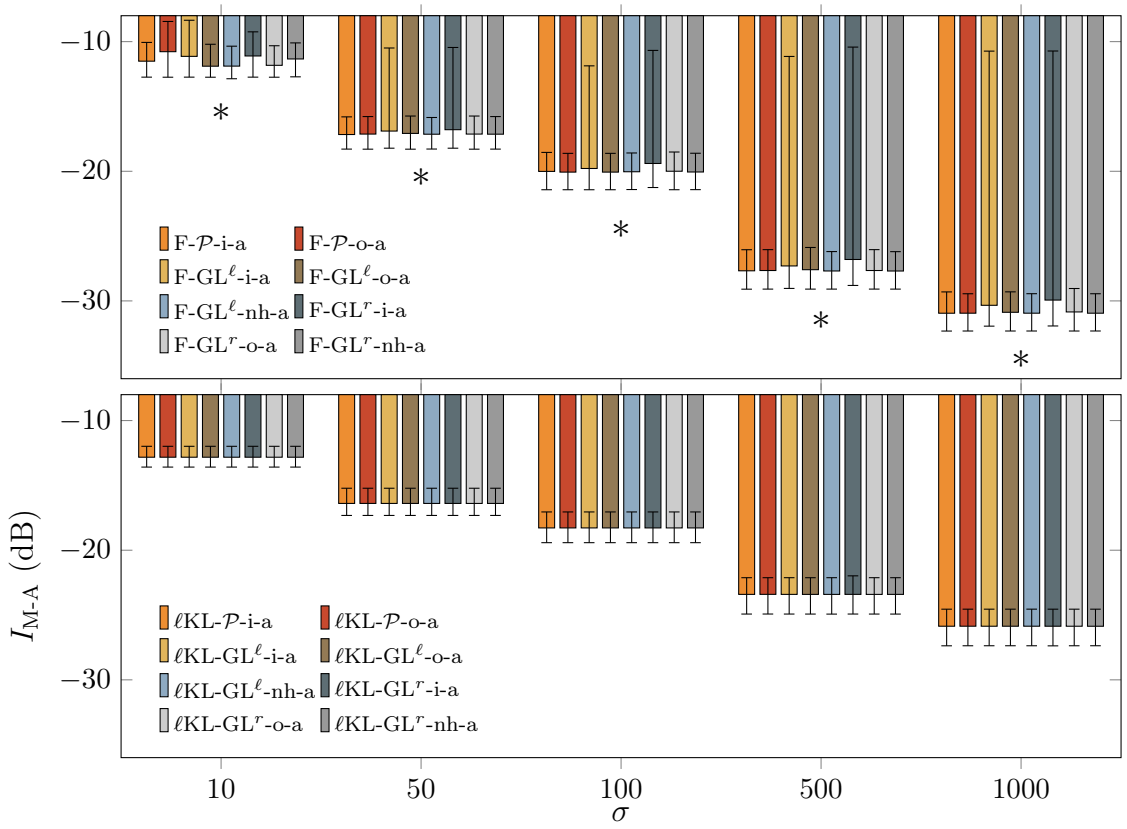

FIGURE *4.2 – Comparaisons des cadres d'optimisation : médianes, premiers et neuvièmes déciles (barres d'erreurs) estimés sur* 50 *essais des critères de Moreau-Amari en fonction du rapport signal sur bruit* σ *pour l'ensemble des algorithmes qui exploitent la stratégie d'optimisation directe (a) et qui optimisent le critère de Frobenius (en haut) et celui de Kullback-Leibler gauche (en bas). Le signe* $*$ *indique une différence significative* ($p < 2,5 \cdot 10^{-3}$) *selon le test de Friedman.*

où 5 est le nombre de tests de Friedman et 28 est le nombre de comparaisons par paires. Remarquons tout d'abord que les tests de Friedman ne font pas apparaître de différences entre les méthodes qui exploitent la mesure de Kullback-Leibler gauche. Pour ce critère, le choix du cadre d'optimisation n'a donc pas d'influence sur les performances obtenues. En ce qui concerne les algorithmes basés sur le critère de Frobenius, les tests de Friedman montrent des différences significatives pour toutes les valeurs de σ. Les tests des rangs signés de Wilcoxon révèlent que :

- Pour $\sigma = 10$, F-GL$^{\ell}$-nh-a est préférable à tous les algorithmes ; F-GL$^{\ell}$-o-a et F-GL^{r}-o-a sont plus efficaces que F-$\mathcal{P}$-o-a, F-GL$^{\ell}$-i-a et F-GL^{r}-nh-a.
- F-$\mathcal{P}$-i-a est meilleur que F-$\mathcal{P}$-o-a, F-GL$^{\ell}$-o-a, F-GL^{r}-o-a et F-GL^{r}-nh-a pour $\sigma \geq 50$, que F-GL$^{\ell}$-nh-a pour $\sigma \geq 100$ et que F-GL^{r}-i-a pour $\sigma = 1000$.
- F-GL$^{\ell}$-nh-a est désavantageux par rapport à F-$\mathcal{P}$-o-a, F-GL$^{\ell}$-o-a et F-GL^{r}-nh-a pour $\sigma \geq 500$ et par rapport à F-GL^{r}-o-a pour $\sigma = 500$.
- F-GL^{r}-nh-a a des performances inférieures à celles de F-$\mathcal{P}$-o-a et F-GL$^{\ell}$-o-a pour $\sigma = \{100, 500\}$, et à celles F-GL^{r}-o-a pour $\sigma = 500$.

Pour $\sigma \geq 50$, les différences entre les performances des algorithmes mises en évidence par les tests de Wilcoxon sont relativement faibles : les différences entre les médianes sont inférieures à $0,1$ dB et celles entre les centiles ne dépassent pas $0,4$ dB. On observe également que les neuvièmes déciles de F-GL$^{\ell}$-i-a et F-GL^{r}-i-a sont aberrants pour $\sigma \geq 50$, ce qui indique que ces deux méthodes convergent vers des solutions très peu satisfaisantes pour certains essais. Cependant, les tests de Wilcoxon ne montrent pas de différences significatives pour ces algorithmes dans la grande majorité des cas.

En résumé, le choix du cadre d'optimisation n'a pas d'influence sur le résultat obtenu pour le critère de Kullback-Leibler gauche. En revanche, pour le critère de Frobenius, les solutions trouvées ne sont pas toujours équivalentes en fonction du cadre d'optimisation sélectionné. Il apparaît que dans la plupart des cas étudiés, l'algorithme F-$\mathcal{P}$-i-a donne le plus régulièrement la meilleure solution, bien que les différences dans les performances sont souvent relativement faibles. La contrainte intrinsèque n'est cependant pas à préférer lorsqu'on équipe GL_n des métriques invariantes à gauche ou à droite. La contrainte oblique donne des résultats consistants dans l'ensemble des situations considérées. Pour sa part, la contrainte non-holonomique est régulièrement légèrement moins efficace. Enfin, on observe que la variété polaire $\mathcal{P}_n$ est compétitive par rapport à GL_n équipé des métriques invariantes à gauche ou à droite.

Comparaisons des stratégies d'optimisation pour la diagonalisation Nous étudions à présent les trois stratégies d'optimisation de la section 2.2. Pour ces expériences, les méthodes reposent toutes sur un algorithme d'optimisation à région de confiance. Ce choix est motivé par deux raisons : premièrement, cela permet d'éliminer le biais éventuel introduit par l'utilisation de différents algorithmes d'optimisation et deuxièmement, l'utilisation du second ordre permet de différencier les stratégies directe (a) et indirecte (b) lorsqu'on prend les matrices diagonales les plus proches comme expliqué dans la section 2.2. La figure 4.3 présente les médianes et les premiers et neuvièmes déciles des mesures de Moreau-Amari en fonction du rapport signal sur bruit pour les algorithmes qui optimisent les critères de Frobenius (en haut de la figure) et de Kullback-Leibler gauche (en bas de la figure) avec les trois stratégies d'optimisation sur

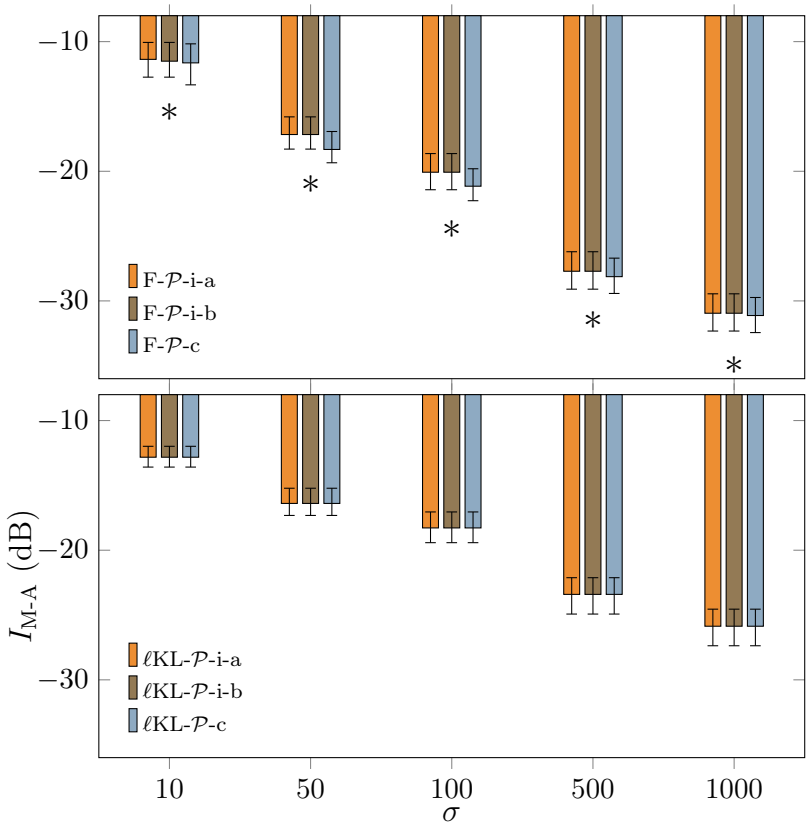

FIGURE *4.3 – Comparaisons des stratégies d'optimisation pour la diagonalisation : médianes, premiers et neuvièmes déciles (barres d'erreurs) estimés sur* 50 *essais des critères de Moreau-Amari en fonction du rapport signal sur bruit* σ *pour les algorithmes qui optimisent les critères de Frobenius (en haut) et de Kullback-Leibler gauche (en bas) avec les trois stratégies d'optimisation sur la variété polaire* $\mathcal{P}_n$ *avec la contrainte intrinsèque. Le signe* $*$ *indique une différence significative* ($p < 2,5 \cdot 10^{-3}$) *selon le test de Friedman.*

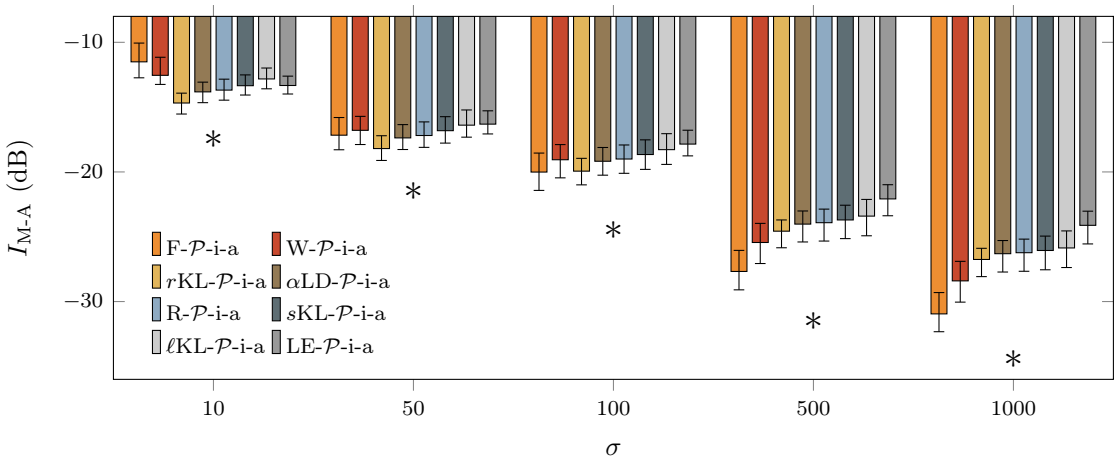

FIGURE *4.4 – Comparaisons des divergences : médianes, premiers et neuvièmes déciles (barres d'erreurs) estimés sur* 50 *essais des critères de Moreau-Amari en fonction du rapport signal sur bruit* σ *pour les algorithmes qui optimisent les critères basés sur l'ensemble des divergences considérées avec la stratégie d'optimisation directe sur la variété polaire* $\mathcal{P}_n$ *avec la contrainte intrinsèque. Le signe* $*$ *indique une différence significative* ($p < 2,5 \cdot 10^{-3}$) *selon le test de Friedman.*

la variété polaire $\mathcal{P}_n$ avec la contrainte intrinsèque. Pour chaque divergence et pour chaque valeur de rapport signal sur bruit, nous faisons un test de Friedman sur les trois algorithmes. Dans le cas où il est significatif, nous faisons toutes les comparaisons par paires avec le test de Wilcoxon. Son seuil de significativité est fixé à $3,0 \cdot 10^{-3}$, qui correspond à $\alpha/(5\times3)$, où 5 est le nombre de tests de Friedman et 3 est le nombre de comparaisons par paires. Remarquons tout d'abord que les tests de Friedman ne montrent pas de différences entre les méthodes qui exploitent la mesure de Kullback-Leibler gauche. Pour ce critère, le choix de la stratégie d'optimisation pour la diagonalisation conjointe n'a donc pas d'influence sur les performances obtenues. Pour les méthodes basées sur le critère de Frobenius, les tests de Friedman indiquent des différences significatives pour toutes les valeurs du rapport signal sur bruit. Les tests de Wilcoxon révèlent que F-$\mathcal{P}$-c est plus performant que F-$\mathcal{P}$-i-a et F-$\mathcal{P}$-i-b dans tous les cas. La stratégie indirecte inverse est donc à préférer pour ce critère.

Comparaisons des divergences Nous nous intéressons enfin aux différents critères de la section 2.3. Nous examinons l'influence de la divergence utilisée pour construire le critère de diagonalisation conjointe approximée. La figure 4.4 contient les médianes et les premiers et neuvièmes déciles des mesures de Moreau-Amari en fonction du rapport signal sur bruit pour les algorithmes qui optimisent les critères basés sur l'ensemble des divergences considérées avec la stratégie d'optimisation directe sur la variété polaire $\mathcal{P}_n$ avec la contrainte intrinsèque. Pour chaque valeur de rapport signal sur bruit, nous faisons un test de Friedman sur les huit algorithmes. Dans le cas où il est significatif, nous faisons toutes les comparaisons par paires avec le test de Wilcoxon. Son seuil est fixé à $3,6 \cdot 10^{-4}$, qui correspond à $\alpha/(5\times28)$, où 5 est le nombre de tests de Friedman et 28 est le nombre de comparaisons par paires. Dans cette

situation, les tests de Friedman montrent des différences significatives pour toutes les valeurs de σ. Les tests de Wilcoxon révèlent que :

- rKL-$\mathcal{P}$-i-a est plus performant que αLD-$\mathcal{P}$-i-a (pour $\alpha = 0$), qui est plus avantageux que R-$\mathcal{P}$-i-a, lui-même plus précis que sKL-$\mathcal{P}$-i-a, dont les performances sont meilleures que celles de ℓKL-$\mathcal{P}$-i-a pour toutes les valeurs de σ.
- LE-$\mathcal{P}$-i-a est entre R-$\mathcal{P}$-i-a et ℓKL-$\mathcal{P}$-i-a pour $\sigma = 10$, sans différences significatives avec sKL-$\mathcal{P}$-i-a. Pour $\sigma \geq 50$, LE-$\mathcal{P}$-i-a est moins précis que tous les autres algorithmes.
- F-$\mathcal{P}$-i-a est désavantageux par rapport à tous les autres algorithmes pour $\sigma = 10$. Pour $\sigma = 50$, F-$\mathcal{P}$-i-a est entre ℓKL-$\mathcal{P}$-i-a et rKL-$\mathcal{P}$-i-a, sans différences significatives avec sKL-$\mathcal{P}$-i-a, αLD-$\mathcal{P}$-i-a, R-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a. Pour $\sigma = 100$, F-$\mathcal{P}$-i-a est plus précis que tous les algorithmes sauf rKL-$\mathcal{P}$-i-a, avec lequel les différences ne sont pas significatives. Pour $\sigma \geq 500$, F-$\mathcal{P}$-i-a donne de meilleurs résultats que toutes les autres méthodes.
- W-$\mathcal{P}$-i-a est entre ℓKL-$\mathcal{P}$-i-a et F-$\mathcal{P}$-i-a pour $\sigma = 10$. Pour $\sigma = 50$, W-$\mathcal{P}$-i-a est entre R-$\mathcal{P}$-i-a et ℓKL-$\mathcal{P}$-i-a, tout en étant significativement moins bon que F-$\mathcal{P}$-i-a et sans différences significatives avec sKL-$\mathcal{P}$-i-a. Pour $\sigma = 100$, W-$\mathcal{P}$-i-a est moins bon que F-$\mathcal{P}$-i-a et rKL-$\mathcal{P}$-i-a, meilleur que sKL-$\mathcal{P}$-i-a et les différences avec αLD-$\mathcal{P}$-i-a et R-$\mathcal{P}$-i-a ne sont pas significatives. Pour $\sigma \geq 500$, W-$\mathcal{P}$-i-a est entre F-$\mathcal{P}$-i-a et rKL-$\mathcal{P}$-i-a.

Pour résumer, on trouve qu'étant donné une valeur du rapport signal sur bruit, les performances obtenues avec les différents critères de diagonalisation conjointe sont relativement ordonnées. Les performances des critères basés sur la mesure de Kullback-Leibler gauche et sur la distance log-euclidienne sont mauvaises sur ces données simulées. Le critère de Frobenius est très avantageux pour les grands rapports signal sur bruit, mais il n'est pas robuste par rapport à ce paramètre. On observe également ce comportement pour le critère de Wasserstein. De l'autre côté, le critère de Kullback-Leibler droite est plus résistant au bruit et produit des résultats avantageux pour des rapports signal sur bruit plus faibles, tout en étant relativement efficace lorsqu'il y a peu de bruit.

Conclusions Plusieurs points ressortent de cette analyse des simulations. Tout d'abord, l'ensemble des résultats présentés dans cette section sont spécifiques aux paramètres du modèle (4.1), *i.e.*, la taille des matrices et leur nombre, le modèle du profil des sources dans les matrices diagonales et le conditionnement des matrices de mixage. On ne peut donc pas s'attendre à ce qu'ils soient transposables en général. Dans ce contexte, les outils d'optimisation riemannienne que nous proposons dans ce travail de thèse permettent de définir des algorithmes de diagonalisation conjointe approximée compétitifs par rapport aux algorithmes de l'état de l'art. Pour le critère de Kullback-Leibler gauche, le cadre d'optimisation riemannienne et la stratégie d'optimisation pour la diagonalisation conjointe n'ont pas d'incidence sur la précision des résultats. La consistance des résultats pour ce critère informe sur sa robustesse et joue donc en sa faveur. Cela ne se généralise pas à l'ensemble des critères : pour le critère de Frobenius, la variété polaire $\mathcal{P}_n$ associée à la contrainte intrinsèque donne les meilleurs résultats avec la stratégie d'optimisation directe (a) et la stratégie d'optimisation indirecte inverse (c) est la plus efficace sur la variété polaire. En ce qui concerne le choix de la divergence, nous remarquons notamment que les distances de Frobenius et de Wasserstein sont efficaces lorsque le bruit est faible, mais ne sont pas robustes par rapport à cette variable. La mesure de Kullback-Leibler droite donne quant à elle des solutions avantageuses pour des

rapports signal sur bruit plus faibles et est également efficace pour des rapports signal sur bruit importants. Il est intéressant de remarquer que les deux critères traditionnellement utilisés (Frobenius et Kullback-Leibler gauche) ne sont pas les meilleurs dans toutes les situations considérées. Bien que spécifiques au modèle, ces résultats sont cohérents avec ceux obtenus sur les données réelles présentés dans la section suivante, où on trouve que les solutions obtenues par le biais des distances de Frobenius et de Wasserstein sont moins consistantes que les autres divergences.

4.3 Expériences avec des données électroencéphalographiques

Dans cette section, nous employons les méthodes de diagonalisation conjointe approximée pour effectuer la séparation aveugle de sources d'enregistrements électroencéphalographiques. L'objectif est d'étudier le comportement de nos algorithmes sur des données réelles. Par soucis de simplicité, nous nous contentons des algorithmes qui optimisent les critères basés sur l'ensemble des divergences considérées dans la section 2.3 avec la stratégie d'optimisation directe sur la variété polaire $\mathcal{P}_n$ avec la contrainte intrinsèque. Ces méthodes donnent des résultats satisfaisants sur les données simulées. Un des défis de l'analyse de données électroencéphalographiques est de pouvoir mesurer la qualité des résultats obtenus alors qu'on ne connaît pas la matrice de mélange réelle. Dans ce travail, nous considérons des données pour lesquelles nous avons des connaissances a priori qui permettent de définir une mesure de performance objective et donc de comparer différentes méthodes. Dans la section 4.3.1, nous présentons les données et nous décrivons notre analyse, *i.e.*, les pré-traitements, le calcul des matrices à diagonaliser, le choix de leur pondération et la mesure de performance. Dans la section 4.3.2, nous exposons les résultats obtenus, *i.e.*, nous décrivons les sources trouvées et nous comparons les performances des méthodes de diagonalisation conjointe.

4.3.1 Données et analyse

Nous nous intéressons à des données électroencéphalographiques qui contiennent des potentiels évoqués visuels à l'état d'équilibre (SSVEP, *steady state visually evoked potential*) [110] (une illustration est donnée dans la figure 4.5). Ce sont des signaux issus du cortex visuel qui correspondent à des réponses à des stimulations lumineuses qui oscillent à des fréquences fixes. En effet, lorsque la rétine est excitée par des stimuli oscillant à une fréquence fixe, le cortex visuel génère des signaux électriques qui oscillent à la même fréquence et/ou à des multiples de cette fréquence. L'avantage des données de ce paradigme est que nous sommes en mesure de caractériser le spectre attendu pour une des sources cérébrales ; celle qui sépare la source de SSVEP. Pour nos expériences numériques, nous utilisons les données utilisées dans l'étude [69]. Ce jeu de données est public. Il contient les électroencéphalogrammes de 12 sujets qui ont effectué entre 2 et 5 sessions. Les enregistrements sont échantillonnés à 256 Hz et acquis sur 8 électrodes situés au niveau occipital (électrodes Oz, O1 et O2) et pariéto-occipital (électrodes POz, PO3, PO4, PO7 et PO8). Chaque session comporte 32 essais de 3 secondes divisés en 4 classes (8 essais par classe) : la classe repos, pour laquelle il n'y a pas de stimuli, et les classes

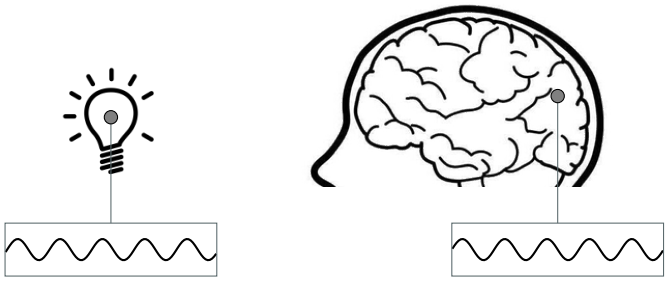

FIGURE *4.5 – Schéma d'illustration d'une expérience permettant d'induire des potentiels évoqués visuels à l'état d'équilibre (SSVEP). Lorsque la rétine est excitée par une source lumineuse oscillant à une fréquence fixe, le cortex visuel se synchronise avec ces stimuli et génère des signaux électriques à la même fréquence.*

13, 17 et 21 Hz, pour lesquelles le sujet regarde une source lumineuse dont la luminance oscille à 13, 17 et 21 Hz. Nous traitons uniquement la session 2 du sujet 3 et les sessions 1 et 5 du sujet 12. Ces enregistrements ont été choisis pour la qualité des signaux enregistrés et la présence claire de l'activité SSVEP, ce qui valide le protocole expérimental. Pour ces trois enregistrements, les figures 4.6, 4.7 et 4.8 montrent les moyennes des densités spectrales de puissance calculées sur chaque essai avec la méthode de Welch pour chaque classe et chaque électrode. Pour la méthode de Welch, on sélectionne les fréquences entre 1 et 43 Hz avec une résolution de $0,5$ Hz et on utilise des fenêtres glissantes de 2 secondes avec un recouvrement de $1,75$ secondes. On observe bien des pics aux fréquences d'intérêt présents uniquement dans les classes correspondantes. On remarque qu'ils sont plutôt au niveau des électrodes Oz, O1, O2 et POz et que les pics à 13 et 17 Hz sont plus puissants que celui à 21 Hz, qui est relativement faible.

Pour effectuer la séparation aveugle de sources de ces électroencéphalogrammes, dont le but est de retrouver la source de SSVEP, nous pré-traitons d'abord les données avec un filtre de Butterworth passe-bande d'ordre 4 entre 1 et 43 Hz. Il faut ensuite définir les matrices à diagonaliser. Pour ce faire, nous calculons pour chaque essai les cospectres entre 1 et 43 Hz avec une résolution de 1 Hz en utilisant la méthode de Welch avec des fenêtres glissantes de 1 seconde avec un recouvrement de $0,75$ seconde. En ce qui concerne les fréquences d'intérêt pour les SSVEP, nous considérons les fréquences de base de chaque classe et leur premier multiple : $\{13, 26\}$, $\{17, 34\}$ et $\{21, 42\}$ pour les classes 13, 17 et 21 Hz, respectivement. Pour toute fréquence $f \notin \{13, 17, 21, 26, 34, 42\}$, on définit $\boldsymbol{C}_f$ comme la moyenne arithmétique des cospectres à la fréquence f de l'ensemble des essais. Pour $f \in \{13, 17, 21, 26, 34, 42\}$, on définit $\boldsymbol{C}_f$ comme la moyenne arithmétique des cospectres à la fréquence f des essais n'étant pas dans la classe qui correspond à f. Enfin, pour $f \in \{13, 17, 21, 26, 34, 42\}$, on définit $\overline{\boldsymbol{C}}_f$ comme la moyenne arithmétique des cospectres à la fréquence f des essais étant dans la classe qui correspond à f. Comme proposé dans [41], nous normalisons la trace des matrices $\boldsymbol{C}_f$ et $\overline{\boldsymbol{C}}_f$, qui constituent alors l'ensemble des matrices à diagonaliser. De plus, pour les matrices $\boldsymbol{C}_f$, nous utilisons les poids w_f calculés à partir de la mesure de non-diagonalité de $\boldsymbol{C}_f$ définie dans [41] et nous les normalisons de sorte que leur somme soit 1. Pour les matrices $\overline{\boldsymbol{C}}_f$, nous prenons les poids $\overline{w}_f = 1/6$, dont la somme est également 1.

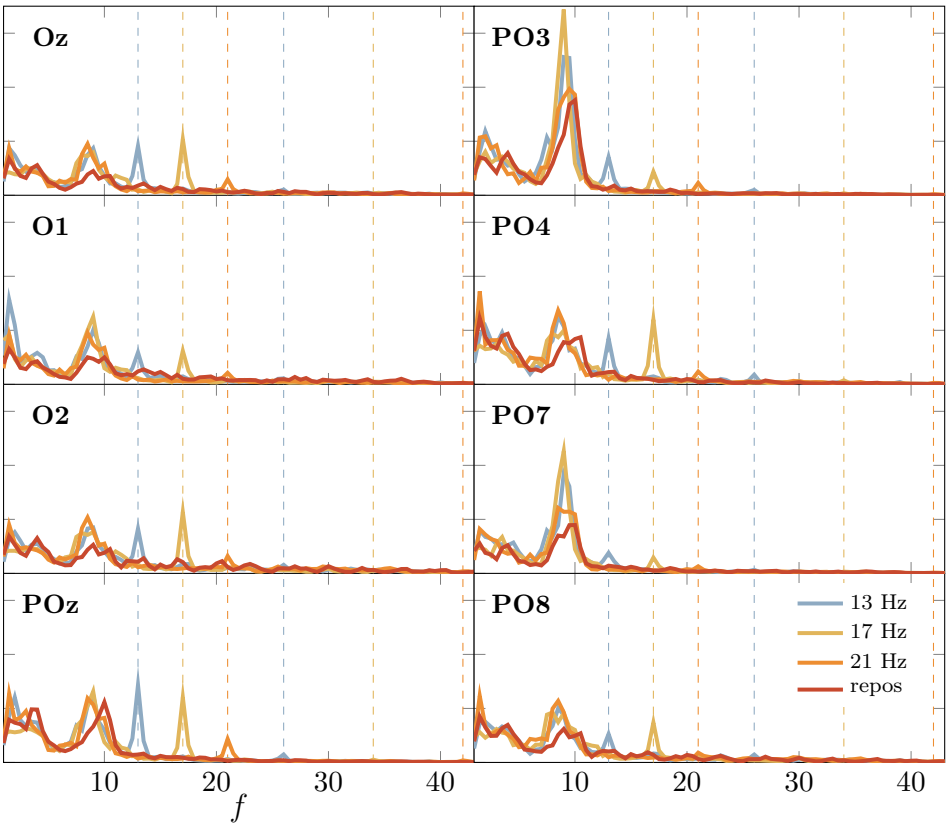

FIGURE *4.6 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) des huit électrodes calculées sur chaque essai avec la méthode de Welch pour la session 2 du sujet 3.*

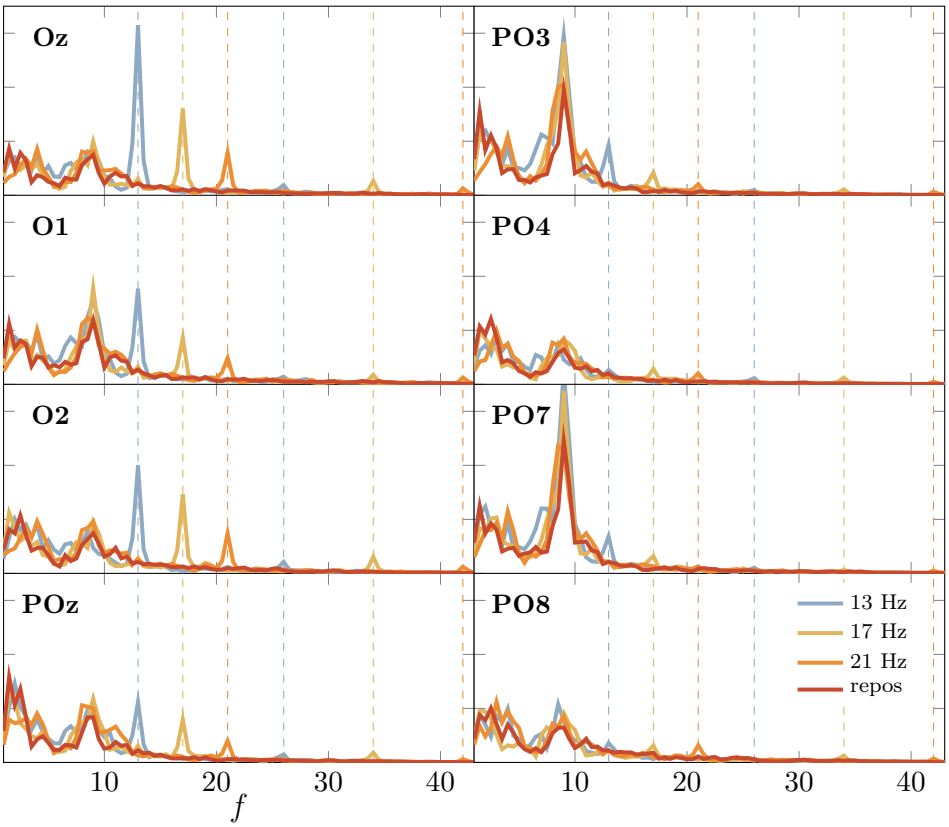

FIGURE *4.7 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) des huit électrodes calculées sur chaque essai avec la méthode de Welch pour la session 1 du sujet 12.*

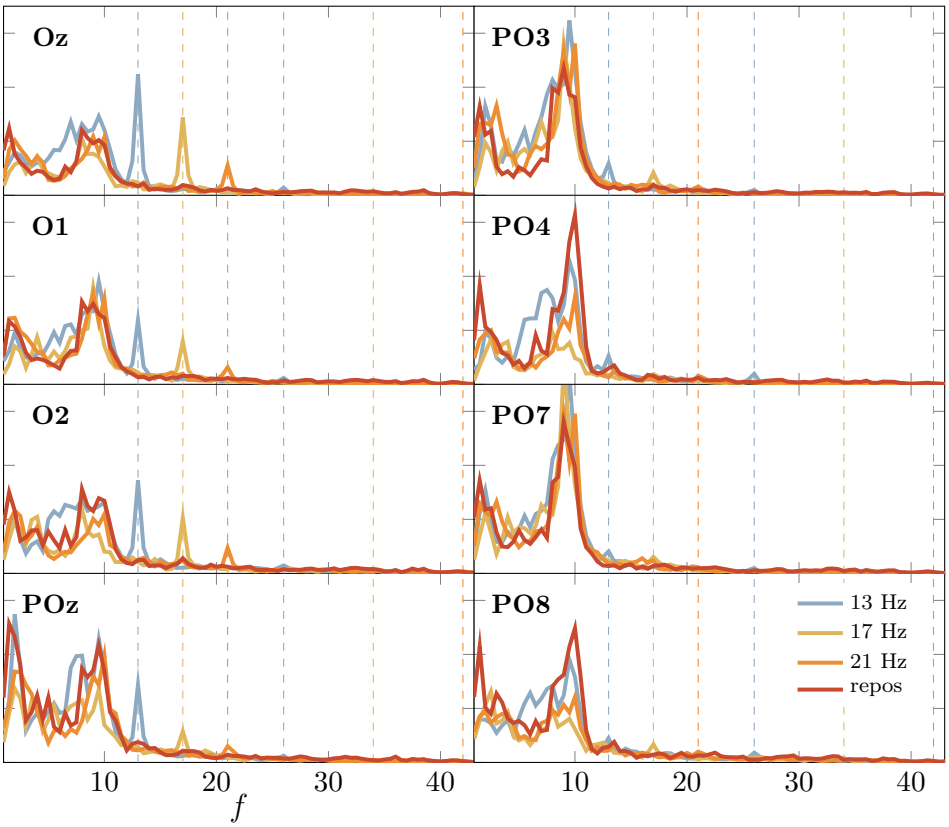

FIGURE *4.8 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) des huit électrodes calculées sur chaque essai avec la méthode de Welch pour la session 5 du sujet 12.*

Remarquons que pour obtenir la séparation de sources, nous utilisons d'une part la coloration des signaux électroencéphalographiques stationnaires avec les matrices $\boldsymbol{C}_f$ et d'autre part la non-stationnarité de la source de SSVEP entre les classes avec les matrices $\overline{\boldsymbol{C}}_f$. Avant d'effectuer la diagonalisation conjointe approximée, nous faisons un blanchiment des matrices à diagonaliser avec l'inverse de la racine carré de leur moyenne arithmétique. L'ensemble des algorithmes est ensuite initialisé avec la matrice identité.

Dans l'objectif d'estimer les performances de la séparation, nous proposons une mesure de qualité de la source de SSVEP. Étant donné la classe $f \in \{13, 17, 21\}$, nous définissons le critère de performance $I_{\text{SSVEP}}(f)$ pour la source de SSVEP estimée $\hat{\boldsymbol{s}}$ par

$$I_{\text{SSVEP}}(f) = \frac{1}{K_f} \sum_{k \in \mathcal{K}_f} \frac{P_{\hat{\boldsymbol{s}}}(k, f) + P_{\hat{\boldsymbol{s}}}(k, 2f)}{\sum_{\widetilde{f}=1}^{43} P_{\hat{\boldsymbol{s}}}(k, \widetilde{f})}, \tag{4.3}$$

où $P_{\hat{\boldsymbol{s}}}(k, \widetilde{f})$ est la densité spectrale de puissance de la source de SSVEP estimée $\hat{\boldsymbol{s}}$ pour l'essai k à la fréquence $\widetilde{f}$, $\mathcal{K}_f$ est l'ensemble des essais de la classe f et K_f est son cardinal. La mesure $I_{\text{SSVEP}}(f)$ indique la proportion de la densité spectrale de puissance à f et $2f$ par rapport à la puissance totale sur l'ensemble des essais de classe f. C'est donc une mesure dans $[0, 1]$ où 1 correspond à un signal mélangeant uniquement des oscillations à f et $2f$, ce qui est l'activité attendue pour une source de SSVEP idéale dans le domaine de fréquences qu'on considère. Remarquons que du fait de la normalisation, l'indétermination d'échelle de la source de SSVEP estimée n'a pas d'incidence sur la mesure $I_{\text{SSVEP}}(f)$.

4.3.2 Résultats

Analyse qualitative des sources trouvées Pour commencer, nous analysons qualitativement les sources estimées pour les trois enregistrements électroencéphalographiques. Avec les huit méthodes de diagonalisation conjointe approximée testées, on obtient des sources qui se correspondent d'une méthode à l'autre. Les figures 4.9, 4.10 et 4.11 présentent les moyennes pour chaque classe des densités spectrales de puissance des quatre sources les plus puissantes obtenues avec l'algorithme W-$\mathcal{P}$-i-a pour la session 2 du sujet 3, et avec l'algorithme rKL-$\mathcal{P}$-i-a pour les sessions 1 et 5 du sujet 12. Les densités spectrales de puissance sont calculées avec la méthode de Welch, pour laquelle on sélectionne les fréquences entre 1 et 43 Hz avec une résolution de $0,5$ Hz et on utilise des fenêtres glissantes de 2 secondes avec un recouvrement de $1,75$ secondes. Dans chacun des cas, les quatres sources non montrées ont des densités spectrales de puissances faibles par rapport aux sources représentées. Pour trier les sources en fonction de leur puissance totale, nous calculons d'abord la rétroprojection de la source au niveau des électrodes, *i.e.*, nous éliminons l'activité de toutes les autres sources sur les électrodes, puis nous calculons les densités spectrales de puissance totales des signaux résultants sur l'ensemble des électrodes. On observe sur les figures 4.9, 4.10 et 4.11 que la source 1 correspond à la source de SSVEP. En effet, pour les trois électroencéphalogrammes, elle est inactive dans la classe de repos et est active uniquement aux fréquences 13, 17 et 21 Hz dans les classes correspondantes. La source 2 est quant à elle principalement active autour de $9,5$ Hz dans les 4 classes. Elle correspond à une source du rythme occipital dominant (ondes

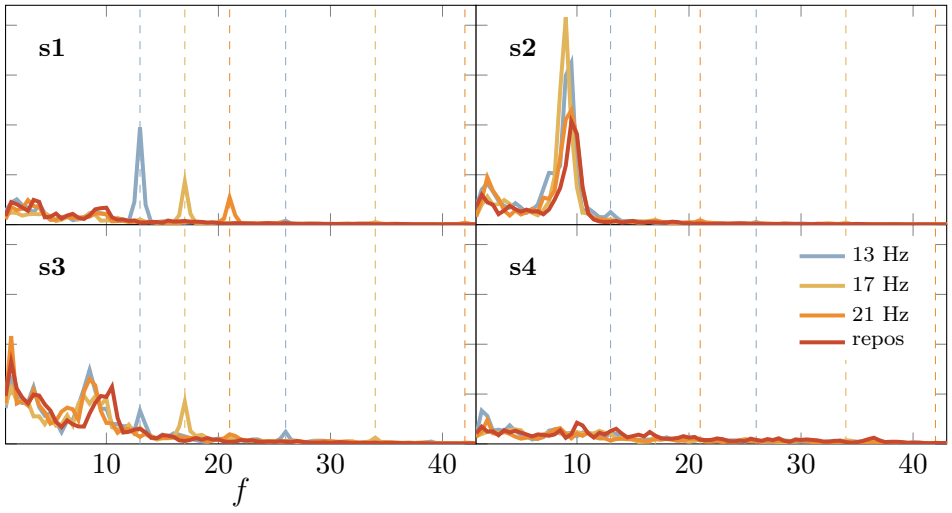

FIGURE *4.9 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) calculées sur chaque essai avec la méthode de Welch des quatre sources les plus puissantes (en moyenne) obtenues avec la méthode W-$\mathcal{P}$-i-a pour la session 2 du sujet 3. Les lignes pointillées verticales marquent les fréquences d'intérêt :* 13, 17, 21, 26, 34 *et* 42 *Hz.*

$I_{\text{SSVEP}}(f)$	F-$\mathcal{P}$-i-a	ℓKL-$\mathcal{P}$-i-a	rKL-$\mathcal{P}$-i-a	sKL-$\mathcal{P}$-i-a	αLD-$\mathcal{P}$-i-a	R-$\mathcal{P}$-i-a	LE-$\mathcal{P}$-i-a	W-$\mathcal{P}$-i-a
13 Hz	**0, 70**	0, 69	**0, 70**	0, 69	**0, 70**	**0, 70**	**0, 70**	**0, 70**
17 Hz	**0, 59**	0, 55	0, 56	0, 55	0, 56	0, 55	0, 56	**0, 59**
21 Hz	0, 38	**0, 41**	**0, 41**	**0, 41**	**0, 41**	**0, 41**	**0, 41**	0, 39

TABLE *4.1 – Scores obtenus avec la mesure de performance* (4.3) *pour les huit méthodes testées sur les classes* 13, 17 *et* 21 *Hz pour la session 2 du sujet 3.*

alpha), qui est bien connu dans le domaine de l'électroencéphalographie [31]. La source 3 est plus variable selon l'enregistrement. Elle présente de l'activité pour les basses fréquences jusqu'autour de 10 Hz. De plus, des pics sont visibles à 13, 17, 21, 26 et 34 Hz dans les classes qui correspondent à ces fréquences pour la session 2 du sujet 3 et légèrement pour la session 1 du sujet 12. Enfin, l'activité de la source 4 est faible pour toutes les fréquences dans toutes les configurations.

Comparaison des performances des algorithmes Nous comparons maintenant les scores obtenus par les sources de SSVEP estimées par les différentes méthodes pour chaque classe avec la mesure de performance définie dans (4.3), qui sont donnés dans les tables 4.1, 4.2 et 4.3 pour la session 2 du sujet 3, la session 1 du sujet 12 et la session 5 du sujet 12, respectivement. On distingue deux groupes de méthodes qui ont le même comportement au niveau des scores : F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a d'un côté, et les autres algorithmes de l'autre. Pour l'ensemble des algorithmes à l'exception de F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a, on observe très peu de différences dans les scores obtenues : les sources de SSVEP estimées par ces méthodes sont très similaires. On remarque également que la source de SSVEP estimée par W-$\mathcal{P}$-i-a est un peu meilleure que celle estimée par F-$\mathcal{P}$-i-a, en particulier pour la session 1 du sujet 12.

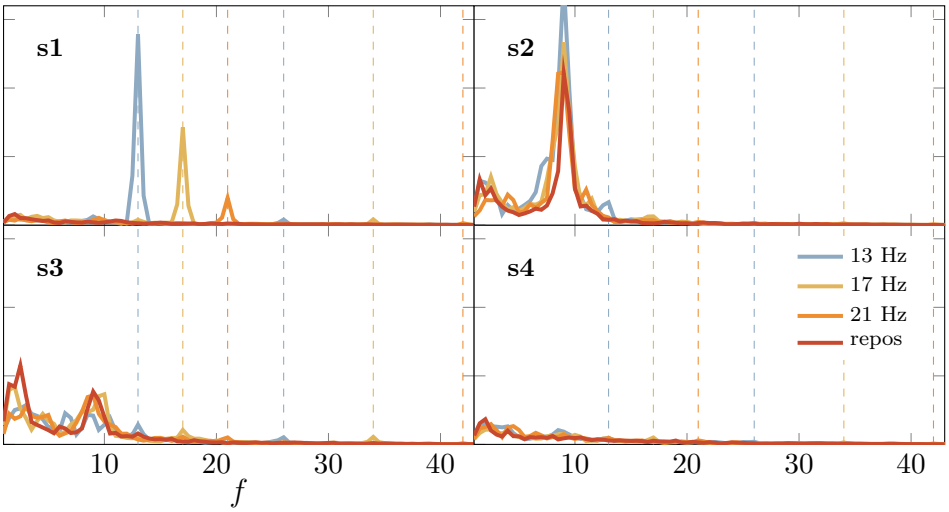

FIGURE *4.10 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) calculées sur chaque essai avec la méthode de Welch des quatre sources les plus puissantes (en moyenne) obtenues avec la méthode rKL-$\mathcal{P}$-i-a pour la session 1 du sujet 12. Les lignes pointillées verticales marquent les fréquences d'intérêt :* 13, 17, 21, 26, 34 *et* 42 *Hz.*

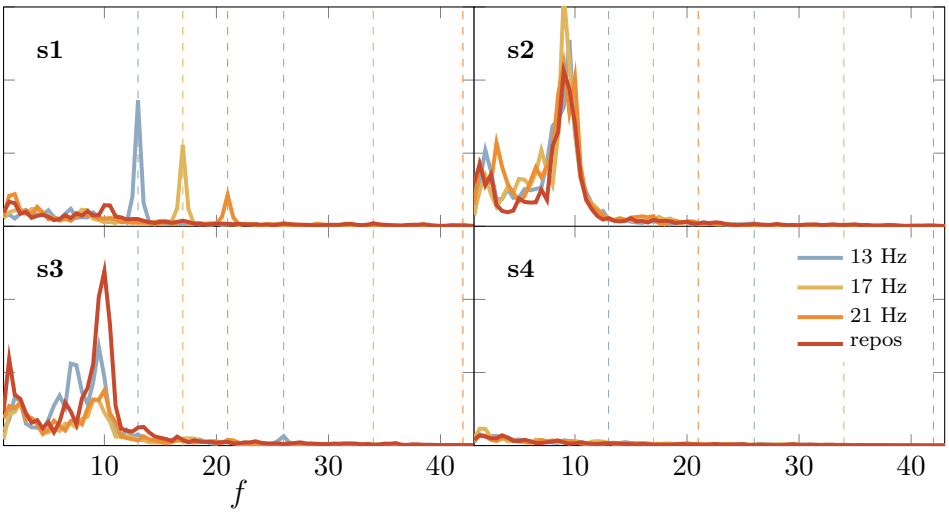

FIGURE *4.11 – Moyenne pour chaque classe des densités spectrales de puissance (unité arbitraire) calculées sur chaque essai avec la méthode de Welch des quatre sources les plus puissantes (en moyenne) obtenues avec la méthode rKL-$\mathcal{P}$-i-a pour la session 5 du sujet 12. Les lignes pointillées verticales marquent les fréquences d'intérêt :* 13, 17, 21, 26, 34 *et* 42 *Hz.*

$I_{\text{SSVEP}}(f)$	F-$\mathcal{P}$-i-a	ℓKL-$\mathcal{P}$-i-a	rKL-$\mathcal{P}$-i-a	sKL-$\mathcal{P}$-i-a	αLD-$\mathcal{P}$-i-a	R-$\mathcal{P}$-i-a	LE-$\mathcal{P}$-i-a	W-$\mathcal{P}$-i-a
13 Hz	$0,95$	$\mathbf{0,96}$	$\mathbf{0,96}$	$\mathbf{0,96}$	$\mathbf{0,96}$	$\mathbf{0,96}$	$\mathbf{0,96}$	$0,95$
17 Hz	$0,87$	$\mathbf{0,91}$	$\mathbf{0,91}$	$\mathbf{0,91}$	$\mathbf{0,91}$	$\mathbf{0,91}$	$\mathbf{0,91}$	$0,89$
21 Hz	$0,50$	$\mathbf{0,60}$	$\mathbf{0,60}$	$\mathbf{0,60}$	$\mathbf{0,60}$	$\mathbf{0,60}$	$\mathbf{0,60}$	$0,54$

TABLE *4.2 – Scores obtenus avec la mesure de performance* (4.3) *pour les huit méthodes testées sur les classes* 13*,* 17 *et* 21 *Hz pour la session 1 du sujet 12.*

$I_{\text{SSVEP}}(f)$	F-$\mathcal{P}$-i-a	ℓKL-$\mathcal{P}$-i-a	rKL-$\mathcal{P}$-i-a	sKL-$\mathcal{P}$-i-a	αLD-$\mathcal{P}$-i-a	R-$\mathcal{P}$-i-a	LE-$\mathcal{P}$-i-a	W-$\mathcal{P}$-i-a
13 Hz	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$	$\mathbf{0,83}$
17 Hz	$0,61$	$\mathbf{0,62}$	$\mathbf{0,62}$	$\mathbf{0,62}$	$\mathbf{0,62}$	$\mathbf{0,62}$	$\mathbf{0,62}$	$0,61$
21 Hz	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$	$\mathbf{0,30}$

TABLE *4.3 – Scores obtenus avec la mesure de performance* (4.3) *pour les huit méthodes testées sur les classes* 13*,* 17 *et* 21 *Hz pour la session 5 du sujet 12.*

De plus, F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a sont moins efficaces que les autres algorithmes dans de plus nombreux cas. Pour la session 5 du sujet 12, il y a très peu de différences. Pour la session 2 du sujet 3, on n'observe pas de différences à 13 Hz, les sources de SSVEP estimées par F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a capturent mieux le pic à 17 Hz, mais moins bien celui à 21 Hz. Enfin, pour la session 1 du sujet 12, les sources de SSVEP estimées par F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a sont moins bonnes pour les trois classes 13, 17 et 21 Hz. Les différences de performances augmentent avec la fréquence de la classe et atteignent 10% pour F-$\mathcal{P}$-i-a et 6% pour W-$\mathcal{P}$-i-a pour la classe 21 Hz. Cela est à mettre en relation avec le fait qu'en général, le rapport signal sur bruit des sources de SSVEP diminue avec la fréquence. Les sources de SSVEP estimées par F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a capturent donc moins bien les pics lorsque le rapport signal sur bruit diminue pour ces données. La tendance observée est donc en concordance avec les résultats des données simulées pour lesquelles F-$\mathcal{P}$-i-a et W-$\mathcal{P}$-i-a sont moins robustes au bruit que les autres méthodes.

Conclusions Pour conclure notre étude de la séparation aveugle de sources de ces trois électroencéphalogrammes, nous observons que les critères de Frobenius et de Wasserstein donnent des résultats moins consistants que les autres méthodes sur ces données électroencéphalographiques. Pour les trois enregistrements, on identifie au moins deux sources physiologiquement problables : la source de SSVEP et une source qui contient le rythme occipital dominant. Pour confirmer ces résultats et étendre notre étude, il faut considérer un plus grand nombre d'enregistrements. De plus, l'analyse de données acquises avec plus d'électrodes permettrait de retrouver plus de sources et de les localiser précisémment avec des méthodes inverses.

Conclusions et perspectives

Ce travail de thèse s'intéresse à plusieurs questions de modélisation et de résolution sur la diagonalisation conjointe approximée de matrices symétriques positives définies pour des applications en traitement de signaux électroencéphalographiques. Le problème de la diagonalisation conjointe est caractérisé par (i) le choix du critère à minimiser, (ii) la contrainte de non-dégénérescence imposée à la solution et (iii) l'algorithme de résolution. Les méthodes que nous proposons sont modulaires et on peut faire varier ces trois composantes indépendamment. Plusieurs conclusions sur les résultats théoriques et sur les expériences numériques ressortent de ce travail, notamment par rapport à ces trois composantes. Les résultats théoriques obtenus ainsi que les expériences numériques amènent des conclusions et ouvrent des perspectives sur les trois points (i), (ii) et (iii).

En ce qui concerne (i), notre étude sur la possibilité d'adopter une approche géométrique du problème (chapitre 2) est fructueuse. Elle permet en effet de modéliser les critères en mesurant la position relative des matrices à diagonaliser par rapport à l'ensemble des matrices diagonales positives définies. Cette formulation inclut les critères traditionnels : le critère des moindres carrés via la distance de Frobenius et le critère log-vraissemblance via la divergence de Kullback-Leibler. Elle permet également de créer de nouveaux critères, notamment en lien avec la théorie de l'information, pour lesquels aucune solution numérique n'existait. Nous considérons ainsi la divergence log-det α, la distance riemannienne naturelle, la distance log-euclidienne et la distance de Wasserstein. Ces critéres possèdent des propriétés différentes et nous déterminons plusieurs d'entre elles, qui sont souhaitables dans le cadre de la diagonalisation conjointe. Les illustrations numériques sur des données simulées et sur des enregistrements électroencéphalographiques (chapitre 4) montrent que les critères basés sur les distances de Frobenius et de Wasserstein sont plus impactés par le bruit que les autres méthodes. On constate que ce sont les deux critères qui vérifient le moins de propriétés souhaitables. Le critère obtenu à partir la mesure de Kullback-Leibler droite est particulièrement efficace dans l'ensemble des situations considérées. Les critères basés sur la divergence log-det α pour $\alpha = 0$, la distance riemannienne naturelle et la mesure de Kullback-Leibler symétrisée donnent aussi de bons résultats. Ces trois critères sont les seuls à posséder toutes les propriétés souhaitables, ce qui les rend très attractifs d'un point de vue théorique. Bien que les critères exploitant la divergence log-det α pour $\alpha = 0$ et la distance riemannienne naturelle sont plus performants que celui basé sur la mesure de Kullback-Leibler symétrisée, leur coût numérique est significativement plus important dans l'état actuel de nos développements car on a besoin d'une méthode itérative pour estimer les matrices diagonales cibles. Cela peut être modifié en choisissant d'autres matrices diagonales cibles appropriées (permettant de conserver l'ensemble des propriétés), comme par exemple celles de la mesure de Kullback-Leibler symétrisée.

Pour traiter (ii), dans le chapitre 3, nous construisons un cadre d'optimisation riemannienne adapté à la diagonalisation conjointe incluant naturellement les contraintes de non-dégénérescence. Nous considérons trois géométries différentes pour la variété des matrices inversibles, à laquelle la solution appartient. Pour les deux premières, nous équipons la variété

des matrices inversibles avec les métriques invariantes à gauche et à droite définies dans (1.47). Bien qu'elle soit adaptée à la structure des solutions du problème, l'utilisation de la métrique invariante à gauche est inédite pour la diagonalisation conjointe. La métrique invariante à droite a un lien naturel avec le modèle de la diagonalisation conjointe [4, 11, 12]. Pour la dernière, nous proposons d'exploiter la décomposition polaire et nous définissons la variété polaire, issue du produit de la variété des matrices symétriques positives définies et de celle des matrices orthogonales. Additionnellement, nous considérons trois contraintes pour éviter les solutions dégénérées : la contrainte oblique (1.7), la contrainte intrinsèque (1.8) et la contrainte non-holonomique, qui exploite directement la géométrie des classes d'équivalence (1.6). Les contraintes oblique et intrinsèque sont inclues dans des sous-variétés riemanniennes de la variété polaire et de celle des matrices inversibles équipée des métriques invariantes à gauche et à droite. La contrainte non-holonomique engendre une variété quotient de la variété des matrices inversibles, qui définit une variété quotient riemannienne pour la métrique invariante à gauche, mais pas pour celle invariante à droite. Nous proposons tout de même des outils pour définir des algorithmes d'optimisation "pseudo-riemanniens" sur la variété quotient non-riemannienne dans ce cas et également dans celui où le critère à minimiser n'est pas bien défini sur le quotient. Les simulations montrent que, pour le critère de Kullback-Leibler gauche, tous les outils d'optimisation développés donnent des solutions équivalentes. On remarque notamment que les outils pseudo-riemannien proposés pour associer la contrainte non-holonomique avec la métrique invariante à droite fonctionnent correctement. Pour le critère de Frobenius, les résultats sont variables selon le choix des outils d'optimisation. Les meilleures performances sont obtenues avec la variété polaire associée à la contrainte intrinsèque. La contrainte oblique donne les résultats les plus consistants pour les trois géométries considérées. La contrainte non-holonomique est globalement moins performante, ce qui est probablement dû au fait que le critère de Frobenius ne définit pas de fonction objectif sur le quotient. Le comportement du critère de Frobenius varie selon le choix de l'échelle diagonale du diagonalisateur conjoint alors que le critère de Kullback-Leibler gauche est invariant, ce qui peut expliquer que l'on a des différences seulement pour le critère de Frobenius.

Enfin, pour (*iii*), notre approche géométrique du chapitre 2 permet de définir trois stratégies d'optimisation pour résoudre le problème de diagonalisation conjointe en pratique : une stratégie directe classique, une stratégie indirecte inédite et une stratégie indirecte inverse qui est généralisation de l'approche proposée dans [107], réinterprétée de façon géométrique. Les éléments manquants pour définir un algorithme de résolution sont fournis par l'optimisation riemannienne. En effet, de nombreuses solutions algorithmiques génériques existent et permettent d'optimiser n'importe quel critère sur n'importe quelle variété riemannienne [2]. De plus, la recherche dans ce domaine est active et de nouveaux algorithmes apparaissent, comme par exemple [61]. Les expériences sur les données simulées indiquent que pour le critère de Kullback-Leibler gauche, les trois stratégies d'optimisation sont équivalentes. Pour le critère de Frobenius, les stratégies directe et indirecte donnent des résultats semblables et la stratégie indirecte inverse est meilleure. Pour cette stratégie, à chaque itération, on considère un sous-problème d'optimisation en partant de l'identité et on réalise une inversion, ce qui semble inhiber l'influence de l'échelle diagonale du diagonalisateur conjoint.

Nous voyons plusieurs perspectives à ce travail de thèse. Pour le modèle géométrique de

la diagonalisation conjointe, nous avons l'intention d'étudier la possibilité d'utiliser la notion d'angle sur la variété des matrices symétriques positives définies pour mesurer la position relative des matrices à diagonaliser par rapport à l'ensemble des matrices diagonales dans le modèle (2.2), comme avancé dans le chapitre 2 (voir la figure 2.2). Une autre piste à investiguer concerne l'étude pratique du modèle (2.27), pour lequel le diagonalisateur vient rapprocher les matrices diagonales des matrices à diagonaliser, qui restent fixes (voir la figure 2.6). De plus, il serait intéressant de déterminer les liens éventuels entre les diagonalisateurs des différents critères, comme il en existe pour les moyennes induites par différentes divergences [20]. Une meilleure compréhension des performances des critères pourrait permettre de trouver un moyen de les combiner pour maximiser la qualité des résultats. Une étude plus fine des propriétés des solutions optimales reste à faire, notamment concernant la convexité géodésique des critères, qui dépend aussi des outils d'optimisation. En ce qui concerne l'optimisation riemannienne sur la variété des matrices inversibles, la variété polaire, qui exploite la décomposition polaire, est efficace. Elle a aussi l'avantage de se généraliser aux matrices rectangulaires de rang plein en remplaçant la variété des matrices orthogonales par celle de Stiefel, présentée par exemple dans [2]. De plus, nous équipons la variété des matrices symétriques positives définies avec la métrique riemannienne classique pour l'optimisation sur la variété polaire. Il est également possible de considérer d'autres structures riemanniennes pour cette variété comme celle qui engendre la distance de Wasserstein décrite dans [21]. En outre, il reste à étudier la possibilité d'exploiter d'autres décompositions telles que la décomposition LU, la décomposition QR ou la décomposition de Schur. En effet, toutes les décompositions n'ont pas le même comportement par rapport aux perturbations comme montré dans [19] et l'une d'entre elle peut se révéler plus avantageuse. Une autre perspective est d'élargir ce travail à d'autres modèles de séparation aveugle de sources. Notre approche géométrique et les outils d'optimisation développés peuvent être adaptés à la séparation aveugle de sources conjointe [40, 72], où plusieurs jeux de donnés sont considérés simultanéments, à l'analyse en sous-espaces indépendants conjointe [74], ou encore à la diagonalisation conjointe bilinéaire [43], où on cherche à diagonaliser des matrices non-symétriques.

Annexe A

Matrice diagonale la plus proche

Dans le chapitre 2, nous introduisons l'approche géométrique au problème de la diagonalisation conjointe approximée par le biais du modèle (2.2), qui dépend du choix de $h(\cdot,\cdot)$ dans (2.2) mesurant la position relative d'une matrice par rapport à une autre et de celui des matrices diagonales cibles $\boldsymbol{B} \mapsto \boldsymbol{\Lambda}_k(\boldsymbol{B})$. Étant donné $h(\cdot,\cdot)$, le choix naturel pour les $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ est donné dans (2.3). Dans un contexte général, cette équation se traduit par : étant donné $\boldsymbol{M} \in \mathcal{S}_n^{++}$, on cherche $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$ telle que

$$\boldsymbol{\Lambda} = \underset{\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}}{\operatorname{argmin}} \quad h(\boldsymbol{M}, \boldsymbol{\Lambda}). \tag{A.1}$$

La matrice $\boldsymbol{\Lambda}$ est alors appelée la matrice diagonale la plus proche de $\boldsymbol{M}$ selon $h(\cdot,\cdot)$. Pour résoudre ce problème, il faut considérer la fonction $\bar{h} : \boldsymbol{\Lambda} \mapsto h(\boldsymbol{M}, \boldsymbol{\Lambda})$ et trouver la matrice $\boldsymbol{\Lambda}$ qui annule son gradient. Comme évoqué dans la section 2.3, nous ne disposons pas de formule analytique en forme fermée pour certaines mesures que nous considérons dans ce travail, à savoir la divergence log-det α, la distance riemannienne naturelle et la distance de Wasserstein. Dans de tels cas, nous pouvons malgré tout utiliser ces matrices en les estimant avec un processus itératif d'optimisation riemannienne sur $\mathcal{D}_n^{++}$. Dans cette annexe, nous présentons d'abord les ingrédients de l'optimisation riemannienne sur $\mathcal{D}_n^{++}$ puis nous traitons les trois cas de la divergence log-det α, la distance riemannienne naturelle et la distance de Wasserstein.

Optimisation sur $\mathcal{D}_n^{++}$

Commençons par rappeler que la variété des matrices diagonales positives définies $\mathcal{D}_n^{++}$ est une sous-variété riemannienne fermée de la variété des matrices symétriques positives définies $\mathcal{S}_n^{++}$. Notons que la variété $\mathcal{D}_n^{++}$ est d'autant plus facile à manipuler étant donné que tous ses éléments commutent. L'espace tangent à tout point de $\mathcal{D}_n^{++}$ est identifié à l'ensemble des matrices diagonales $\mathcal{D}_n$ et hérite de la métrique (1.30) qui équipe $\mathcal{S}_n^{++}$. Pour obtenir un nouveau point sur la variété $\mathcal{D}_n^{++}$ à partir d'un vecteur tangent $\boldsymbol{\xi} \in \mathcal{D}_n$ de $\boldsymbol{\Lambda} \in \mathcal{D}_n^{++}$, nous utilisons l'exponentielle riemannienne de $\mathcal{D}_n^{++}$ qui est donc identique à celle de $\mathcal{S}_n^{++}$, *i.e.*,

$$\exp_{\boldsymbol{\Lambda}}^{\mathcal{D}_n^{++}}(\boldsymbol{\xi}) = \boldsymbol{\Lambda} \exp(\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}). \tag{A.2}$$

Étant donné la fonction objectif $\bar{h}$ définie sur $\mathcal{D}_n^{++}$, dont le gradient euclidien à $\boldsymbol{\Lambda}$ est $\operatorname{grad}_{\mathcal{E}} \bar{h}(\boldsymbol{\Lambda})$ et dont la hessienne euclidienne à $\boldsymbol{\Lambda}, \boldsymbol{\xi}$ est $\operatorname{hess}_{\mathcal{E}} \bar{h}(\boldsymbol{\Lambda})[\boldsymbol{\xi}]$, les versions riemanniennes

du gradient et de la hessienne sont

$$\operatorname{grad}_{\mathcal{D}_n^{++}} \bar{h}(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^2 \operatorname{ddiag}(\operatorname{grad}_{\mathcal{E}} \bar{h}(\boldsymbol{\Lambda})), \tag{A.3}$$

et

$$\operatorname{hess}_{\mathcal{D}_n^{++}} \bar{h}(\boldsymbol{\Lambda})[\boldsymbol{\xi}] = \boldsymbol{\Lambda}^2 \operatorname{ddiag}(\operatorname{hess}_{\mathcal{E}} \bar{h}(\boldsymbol{\Lambda})[\boldsymbol{\xi}]) + \boldsymbol{\Lambda}\boldsymbol{\xi} \operatorname{ddiag}(\operatorname{grad}_{\mathcal{E}} \bar{h}(\boldsymbol{\Lambda})). \tag{A.4}$$

Divergence log-det α

Les versions riemanniennes du gradient et de la hessienne de $\bar{h}_{\alpha\mathrm{LD}} : \boldsymbol{\Lambda} \mapsto d_{\alpha\mathrm{LD}}(\boldsymbol{M}, \boldsymbol{\Lambda})$ sont données par

$$\operatorname{grad}_{\mathcal{D}_n^{++}} \bar{h}_{\alpha\mathrm{LD}}(\boldsymbol{\Lambda}) = \frac{2}{1-\alpha}\boldsymbol{\Lambda}(\boldsymbol{\Delta}\boldsymbol{\Lambda} - \boldsymbol{I}_n), \tag{A.5}$$

et

$$\operatorname{hess}_{\mathcal{D}_n^{++}} \bar{h}_{\alpha\mathrm{LD}}(\boldsymbol{\Lambda})[\boldsymbol{\xi}] = \frac{2}{1-\alpha}\boldsymbol{\Lambda}(\boldsymbol{\Delta}\boldsymbol{\xi} + \dot{\boldsymbol{\Delta}}\boldsymbol{\Lambda}), \tag{A.6}$$

où

$$\boldsymbol{\Delta} = \operatorname{ddiag}\left(\left(\frac{1-\alpha}{2}\boldsymbol{M} + \frac{1+\alpha}{2}\boldsymbol{\Lambda}\right)^{-1}\right),$$

et

$$\dot{\boldsymbol{\Delta}} = -\frac{1+\alpha}{2}\operatorname{ddiag}\left(\left(\frac{1-\alpha}{2}\boldsymbol{M} + \frac{1+\alpha}{2}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\xi}\left(\frac{1-\alpha}{2}\boldsymbol{M} + \frac{1+\alpha}{2}\boldsymbol{\Lambda}\right)^{-1}\right).$$

Distance riemannienne naturelle

Les versions riemanniennes du gradient et de la hessienne de $\bar{h}_{\mathrm{R}} : \boldsymbol{\Lambda} \mapsto \delta_{\mathrm{R}}^2(\boldsymbol{M}, \boldsymbol{\Lambda})$ sont données par

$$\operatorname{grad}_{\mathcal{D}_n^{++}} \bar{h}_{\mathrm{R}}(\boldsymbol{\Lambda}) = 2\boldsymbol{\Lambda} \operatorname{ddiag}(\log(\boldsymbol{M}^{-1}\boldsymbol{\Lambda})), \tag{A.7}$$

et

$$\operatorname{hess}_{\mathcal{D}_n^{++}} \bar{h}_{\mathrm{R}}(\boldsymbol{\Lambda})[\boldsymbol{\xi}] = 2\boldsymbol{\Lambda} \operatorname{ddiag}(\mathrm{D}\log(\boldsymbol{M}^{-1}\boldsymbol{\Lambda})[\boldsymbol{\xi}]), \tag{A.8}$$

où la dérivée du logarithme matriciel (voir par exemple [47]) est dans ce cas

$$\mathrm{D}\log(\boldsymbol{M}^{-1}\boldsymbol{\Lambda})[\boldsymbol{\xi}] = \int_0^1 [(\boldsymbol{M}^{-1}\boldsymbol{\Lambda} - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}\boldsymbol{M}^{-1}\boldsymbol{\xi}[(\boldsymbol{M}^{-1}\boldsymbol{\Lambda} - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1} ds.$$

Notons que la dérivée du logarithme matriciel $\mathrm{D}\log(\boldsymbol{M}^{-1}\boldsymbol{\Lambda})[\boldsymbol{\xi}]$ peut numériquement être évaluée efficacement avec l'algorithme proposé dans [8].

Distance de Wasserstein

Les versions riemanniennes du gradient et de la hessienne de $\bar{h}_{\mathrm{W}} : \boldsymbol{\Lambda} \mapsto \delta^2_{\mathrm{W}}(\boldsymbol{M}, \boldsymbol{\Lambda})$ sont données par

$$\mathrm{grad}_{\mathcal{D}_n^{++}} \bar{h}_{\mathrm{W}}(\boldsymbol{\Lambda}) = \frac{1}{2}\boldsymbol{\Lambda}(\boldsymbol{\Lambda} - \mathrm{ddiag}(\boldsymbol{Q})), \tag{A.9}$$

et

$$\mathrm{hess}_{\mathcal{D}_n^{++}} \bar{h}_{\mathrm{W}}(\boldsymbol{\Lambda})[\boldsymbol{\xi}] = \frac{1}{2}\boldsymbol{\Lambda}(\boldsymbol{\xi} - \mathrm{ddiag}(\dot{\boldsymbol{Q}})), \tag{A.10}$$

où $\boldsymbol{Q} = (\boldsymbol{\Lambda}^{1/2}\boldsymbol{M}\boldsymbol{\Lambda}^{1/2})^{1/2}$ et $\dot{\boldsymbol{Q}}$ est la solution de l'équation de Sylvester

$$\boldsymbol{Q}\dot{\boldsymbol{Q}} + \dot{\boldsymbol{Q}}\boldsymbol{Q} = \mathrm{sym}(\boldsymbol{\xi}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}\boldsymbol{\Lambda}^{1/2}).$$

Notons qu'on est assuré de l'existence et de l'unicité de la solution de cette équation de Sylvester comme $\boldsymbol{Q}$ est symétrique positive définie.

Annexe B

Gradients et hessiennes des critères étudiés

Dans cette annexe, nous dérivons la version euclidienne des gradients et hessiennes de l'ensemble des critères de diagonalisation conjointe approximée présentés dans la section 2.3. Les versions riemanniennes sur les différentes variétés que nous considérons dans ce manuscript sont ensuite obtenues en utilisant les relations fournies dans la section 1.5 du chapitre 1 et dans le chapitre 3. Par souci de simplicité, nous enlevons ici les sommes pondérées qu'il faut donc restituer pour obtenir les formules complètes. De plus, nous utilisons abondamment les notations suivantes : $\boldsymbol{M}_k = \boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^T$, $\dot{\boldsymbol{M}}_k = \boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T + \boldsymbol{\xi}\boldsymbol{C}_k\boldsymbol{B}^T$, $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_k(\boldsymbol{B})$, $\dot{\boldsymbol{\Lambda}}_k = \mathrm{D}\,\boldsymbol{\Lambda}_k(\boldsymbol{B})[\boldsymbol{\xi}]$, $\boldsymbol{N}_k = \boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{A}$ et $\dot{\boldsymbol{N}}_k = \boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{\xi}^T + \boldsymbol{\xi}\boldsymbol{\Lambda}_k\boldsymbol{A}^T$. Pour chacune des divergences que l'on considère, on détaille les calculs pour le critère (2.2) en guise d'exemple de la façon de procéder et on donne uniquement les résultats pour les critères (2.4) et (2.5) comme les calculs sont similaires. Rappelons que la proposition 2.1 nous permet de traiter $\boldsymbol{\Lambda}_k(\boldsymbol{B})$ comme une constante lorsqu'on dérive le gradient.

Distance de Frobenius

On peut réécrire le critère f_{F} comme

$$f_{\mathrm{F}}(\boldsymbol{B}) = \operatorname{tr}\left((\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)^2\right),$$

où $\boldsymbol{\Lambda}_k = \operatorname{ddiag}(\boldsymbol{M}_k)$, donc

$$\mathrm{D}\, f_{\mathrm{F}}(\boldsymbol{B})[\boldsymbol{\xi}] = 2\operatorname{tr}((\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\dot{\boldsymbol{M}}_k) = 4\operatorname{tr}((\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T).$$

Par identification, on a donc

$$\operatorname{grad}_{\mathcal{E}} f_{\mathrm{F}}(\boldsymbol{B}) = 4(\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\boldsymbol{B}\boldsymbol{C}_k. \tag{B.1}$$

Dériver ce gradient dans la direction $\boldsymbol{\xi}$ donne

$$\operatorname{hess}_{\mathcal{E}} f_{\mathrm{F}}(\boldsymbol{B})[\boldsymbol{\xi}] = 4(\dot{\boldsymbol{M}}_k - \dot{\boldsymbol{\Lambda}}_k)\boldsymbol{B}\boldsymbol{C}_k + 4(\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\boldsymbol{\xi}\boldsymbol{C}_k, \tag{B.2}$$

où $\dot{\boldsymbol{\Lambda}}_k = \operatorname{ddiag}(\dot{\boldsymbol{M}}_k)$.

Pour $\widehat{f}_{\mathrm{F}}$ et $\widetilde{f}_{\mathrm{F}}$, on obtient

$$\mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{\mathrm{F}}(\boldsymbol{B}) = 4(\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\boldsymbol{B}\boldsymbol{C}_k, \qquad \mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{\mathrm{F}}(\boldsymbol{B})[\boldsymbol{\xi}] = 4(\dot{\boldsymbol{M}}_k\boldsymbol{B} + (\boldsymbol{M}_k - \boldsymbol{\Lambda}_k)\boldsymbol{\xi})\boldsymbol{C}_k, \quad \text{(B.3)}$$

et

$$\mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{F}}(\boldsymbol{A}) = 4(\boldsymbol{N}_k - \boldsymbol{M}_k)\boldsymbol{A}\boldsymbol{\Lambda}_k, \qquad \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{F}}(\boldsymbol{A})[\boldsymbol{\xi}] = 4(\dot{\boldsymbol{N}}_k\boldsymbol{A} + (\boldsymbol{N}_k - \boldsymbol{M}_k)\boldsymbol{\xi})\boldsymbol{\Lambda}_k. \quad \text{(B.4)}$$

Divergence de Kullback-Leibler

Rappelons que pour toutes matrices $\boldsymbol{X}$ et $\boldsymbol{\xi}$, on a

$$\mathrm{D}\log\det(\boldsymbol{X})[\boldsymbol{\xi}] = \mathrm{tr}(\boldsymbol{X}^{-1}\,\mathrm{D}\,\boldsymbol{X}[\boldsymbol{\xi}]), \quad \text{(B.5)}$$

et

$$\mathrm{D}(\boldsymbol{X}^{-1})[\boldsymbol{\xi}] = -\boldsymbol{X}^{-1}\boldsymbol{\xi}\boldsymbol{X}^{-1}. \quad \text{(B.6)}$$

Notons que du fait que cette divergence n'est pas symétrique, les calculs de la différenciation de $\widetilde{f}_{\ell\mathrm{KL}}$ sont similaires à ceux de $f_{r\mathrm{KL}}$ et les calculs pour $\widetilde{f}_{r\mathrm{KL}}$ ressemblent à ceux de $f_{\ell\mathrm{KL}}$.

Mesure gauche

En utilisant (B.5), on obtient

$$\mathrm{D}\, f_{\ell\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = \mathrm{tr}(\dot{\boldsymbol{M}}_k(\boldsymbol{\Lambda}_k^{-1} - \boldsymbol{M}_k^{-1})) = 2\,\mathrm{tr}((\boldsymbol{\Lambda}_k^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T),$$

où $\boldsymbol{\Lambda}_k = \mathrm{ddiag}(\boldsymbol{M}_k)$. Par identification, on a donc

$$\mathrm{grad}_{\mathcal{E}}\, f_{\ell\mathrm{KL}}(\boldsymbol{B}) = 2(\boldsymbol{\Lambda}_k^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k. \quad \text{(B.7)}$$

En utilisant (B.6), on dérive ce gradient dans la direction $\boldsymbol{\xi}$, ce qui donne

$$\mathrm{hess}_{\mathcal{E}}\, f_{\ell\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = 2\boldsymbol{\Lambda}_k^{-1}(\boldsymbol{\xi} - \boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{B})\boldsymbol{C}_k + 2\boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T}, \quad \text{(B.8)}$$

où $\dot{\boldsymbol{\Lambda}}_k = \mathrm{ddiag}(\dot{\boldsymbol{M}}_k)$.

Pour $\widehat{f}_{\ell\mathrm{KL}}$ et $\widetilde{f}_{\ell\mathrm{KL}}$, on obtient

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{\ell\mathrm{KL}}(\boldsymbol{B}) & = & 2(\boldsymbol{\Lambda}_k^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k, \\ \mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{\ell\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] & = & 2\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\xi}\boldsymbol{C}_k + 2\boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T}, \end{array} \quad \text{(B.9)}$$

et

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{\ell\mathrm{KL}}(\boldsymbol{A}) & = & 2(\boldsymbol{I}_n - \boldsymbol{N}_k^{-1}\boldsymbol{M}_k)\boldsymbol{A}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{\ell\mathrm{KL}}(\boldsymbol{A})[\boldsymbol{\xi}] & = & 2\boldsymbol{N}_k^{-1}((\boldsymbol{M}_k - \boldsymbol{N}_k)\boldsymbol{A}^{-T}\boldsymbol{\xi}^T + \dot{\boldsymbol{N}}_k\boldsymbol{N}_k^{-1}\boldsymbol{M}_k)\boldsymbol{A}^{-T}. \end{array} \quad \text{(B.10)}$$

Mesure droite

En utilisant (B.5) et (B.6), on obtient cette fois-ci

$$\mathrm{D}\, f_{r\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = \mathrm{tr}(\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k - \boldsymbol{\Lambda}_k\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}) = 2\,\mathrm{tr}((\boldsymbol{I}_n - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T),$$

où $\boldsymbol{\Lambda}_k = \mathrm{ddiag}(\boldsymbol{M}_k^{-1})^{-1}$. Donc

$$\mathrm{grad}_{\mathcal{E}}\, f_{r\mathrm{KL}}(\boldsymbol{B}) = 2(\boldsymbol{I}_n - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \tag{B.11}$$

et

$$\mathrm{hess}_{\mathcal{E}}\, f_{r\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = 2\boldsymbol{M}_k^{-1}((\boldsymbol{\Lambda}_k - \boldsymbol{M}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T + (\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1} - \dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1})\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \tag{B.12}$$

où $\dot{\boldsymbol{\Lambda}}_k = \boldsymbol{\Lambda}_k\,\mathrm{ddiag}(\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1})\boldsymbol{\Lambda}_k$.

Pour $\widehat{f}_{r\mathrm{KL}}$ et $\widetilde{f}_{r\mathrm{KL}}$, on obtient

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{r\mathrm{KL}}(\boldsymbol{B}) & = & 2(\boldsymbol{I}_n - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{r\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] & = & 2\boldsymbol{M}_k^{-1}((\boldsymbol{\Lambda}_k - \boldsymbol{M}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T + \dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \end{array} \tag{B.13}$$

et

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{r\mathrm{KL}}(\boldsymbol{A}) & = & 2(\boldsymbol{M}_k^{-1} - \boldsymbol{N}_k^{-1})\boldsymbol{A}\boldsymbol{\Lambda}_k, \\ \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{r\mathrm{KL}}(\boldsymbol{A})[\boldsymbol{\xi}] & = & 2\boldsymbol{M}_k^{-1}\boldsymbol{\xi}\boldsymbol{\Lambda}_k + 2\boldsymbol{A}^{-T}\boldsymbol{\xi}^T\boldsymbol{A}^{-T}. \end{array} \tag{B.14}$$

Mesure symétrisée

Dans ce cas, les termes en $\log\det(\cdot)$ s'annulent et on a

$$\mathrm{D}\, f_{s\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = \frac{1}{2}\,\mathrm{tr}(\dot{\boldsymbol{M}}_k\boldsymbol{\Lambda}_k^{-1} - \boldsymbol{\Lambda}_k\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}) = \mathrm{tr}((\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T),$$

où $\boldsymbol{\Lambda}_k = \mathrm{ddiag}(\boldsymbol{M}_k)^{1/2}\,\mathrm{ddiag}(\boldsymbol{M}_k^{-1})^{-1/2}$. Donc

$$\mathrm{grad}_{\mathcal{E}}\, f_{s\mathrm{KL}}(\boldsymbol{B}) = (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \tag{B.15}$$

et

$$\begin{aligned} \mathrm{hess}_{\mathcal{E}}\, f_{s\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] = {} & \boldsymbol{\Lambda}_k^{-1}(\boldsymbol{\xi} - \dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\boldsymbol{B})\boldsymbol{C}_k \\ & + \boldsymbol{M}_k^{-1}(\boldsymbol{\Lambda}_k\boldsymbol{B}^{-T}\boldsymbol{\xi}^T + \dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k - \dot{\boldsymbol{\Lambda}}_k)\boldsymbol{B}^{-T}, \end{aligned} \tag{B.16}$$

où $\dot{\boldsymbol{\Lambda}}_k = \frac{1}{2}\boldsymbol{\Lambda}_k(\mathrm{ddiag}(\boldsymbol{M}_k)^{-1}\,\mathrm{ddiag}(\dot{\boldsymbol{M}}_k) + \mathrm{ddiag}(\boldsymbol{M}_k^{-1})^{-1}\,\mathrm{ddiag}(\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}))$.

Pour $\widehat{f}_{s\mathrm{KL}}$ et $\widetilde{f}_{s\mathrm{KL}}$, on obtient

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{s\mathrm{KL}}(\boldsymbol{B}) & = & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k - \boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{s\mathrm{KL}}(\boldsymbol{B})[\boldsymbol{\xi}] & = & \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\xi}\boldsymbol{C}_k + \boldsymbol{M}_k^{-1}(\boldsymbol{\Lambda}_k\boldsymbol{B}^{-T}\boldsymbol{\xi}^T + \dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)\boldsymbol{B}^{-T} \end{array} \tag{B.17}$$

et

$$\begin{array}{rcl} \mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{s\mathrm{KL}}(\boldsymbol{A}) & = & (\boldsymbol{M}_k^{-1}\boldsymbol{N}_k - \boldsymbol{N}_k^{-1}\boldsymbol{M}_k)\boldsymbol{A}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{s\mathrm{KL}}(\boldsymbol{A})[\boldsymbol{\xi}] & = & \boldsymbol{M}_k^{-1}\boldsymbol{\xi}\boldsymbol{\Lambda}_k + \boldsymbol{N}_k^{-1}(\boldsymbol{M}_k\boldsymbol{A}^{-T}\boldsymbol{\xi}^T + \dot{\boldsymbol{N}}_k\boldsymbol{N}_k^{-1}\boldsymbol{M}_k)\boldsymbol{A}^{-T} \end{array} \tag{B.18}$$

Divergence log-det α

On peut réécrire $f_{\alpha\mathrm{LD}}$ comme

$$f_{\alpha\mathrm{LD}}(\boldsymbol{B}) = \frac{4}{1-\alpha^2}\log\det(\boldsymbol{Q}_{k,\alpha}) - \frac{2}{1+\alpha}\log\det(\boldsymbol{M}_k) - \frac{2}{1-\alpha}\log\det(\boldsymbol{\Lambda}_k),$$

où $\boldsymbol{\Lambda}_k$ est la solution de l'équation (2.17) pour $\boldsymbol{M} = \boldsymbol{M}_k$ et $\boldsymbol{Q}_{k,\alpha} = \frac{1-\alpha}{2}\boldsymbol{M}_k + \frac{1+\alpha}{2}\boldsymbol{\Lambda}_k$. En utilisant (B.5), on obtient alors

$$\mathrm{D}\, f_{\alpha\mathrm{LD}}(\boldsymbol{B})[\boldsymbol{\xi}] = \frac{2}{1+\alpha}\operatorname{tr}(\boldsymbol{Q}_{k,\alpha}^{-1}\dot{\boldsymbol{M}}_k - \boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k) = \frac{4}{1+\alpha}\operatorname{tr}((\boldsymbol{Q}_{k,\alpha}^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T).$$

Donc,

$$\operatorname{grad}_{\mathcal{E}} f_{\alpha\mathrm{LD}}(\boldsymbol{B}) = \frac{4}{1+\alpha}(\boldsymbol{Q}_{k,\alpha}^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k. \tag{B.19}$$

La hessienne est donnée par

$$\operatorname{hess}_{\mathcal{E}} f_{\alpha\mathrm{LD}}(\boldsymbol{B})[\boldsymbol{\xi}] = \frac{4}{1+\alpha}\left(\boldsymbol{Q}_{k,\alpha}^{-1}(\boldsymbol{\xi} - \dot{\boldsymbol{Q}}_{k,\alpha}\boldsymbol{Q}_{k,\alpha}^{-1}\boldsymbol{B})\boldsymbol{C}_k + \boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T}\right), \tag{B.20}$$

où $\dot{\boldsymbol{Q}}_{k,\alpha} = \frac{1-\alpha}{2}\dot{\boldsymbol{M}}_k + \frac{1+\alpha}{2}\dot{\boldsymbol{\Lambda}}_k$. La matrice $\dot{\boldsymbol{\Lambda}}_k$, obtenue en dérivant l'équation (2.17), est la solution de l'équation

$$\dot{\boldsymbol{\Lambda}}_k \operatorname{ddiag}(\boldsymbol{Q}_{k,\alpha}^{-1}) - \frac{1+\alpha}{2}\boldsymbol{\Lambda}_k \operatorname{ddiag}(\boldsymbol{Q}_{k,\alpha}^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{Q}_{k,\alpha}^{-1}) = \frac{1-\alpha}{2}\boldsymbol{\Lambda}_k \operatorname{ddiag}(\boldsymbol{Q}_{k,\alpha}^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{Q}_{k,\alpha}^{-1}).$$

Nous ne proposons cependant pas de moyen pour résoudre cette équation dans ce manuscript et nous nous limitons donc à des méthodes de gradient pour optimiser $f_{\alpha\mathrm{LD}}$.

Pour $\widehat{f}_{\alpha\mathrm{LD}}$, on obtient

$$\begin{aligned} \operatorname{grad}_{\mathcal{E}} \widehat{f}_{\alpha\mathrm{LD}}(\boldsymbol{B}) &= \tfrac{4}{1+\alpha}(\boldsymbol{Q}_{k,\alpha}^{-1} - \boldsymbol{M}_k^{-1})\boldsymbol{B}\boldsymbol{C}_k, \\ \operatorname{hess}_{\mathcal{E}} \widehat{f}_{\alpha\mathrm{LD}}(\boldsymbol{B})[\boldsymbol{\xi}] &= \tfrac{4}{1+\alpha}\left(\boldsymbol{Q}_{k,\alpha}^{-1}(\boldsymbol{\xi} - \dot{\boldsymbol{Q}}_{k,\alpha}\boldsymbol{Q}_{k,\alpha}^{-1}\boldsymbol{B})\boldsymbol{C}_k + \boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T}\right), \end{aligned} \tag{B.21}$$

où $\boldsymbol{Q}_{k,\alpha} = \frac{1-\alpha}{2}\boldsymbol{M}_k + \frac{1+\alpha}{2}\boldsymbol{\Lambda}_k$ et $\dot{\boldsymbol{Q}}_{k,\alpha} = \frac{1-\alpha}{2}\dot{\boldsymbol{M}}_k$. Enfin, on a pour $\widetilde{f}_{\alpha\mathrm{LD}}$

$$\begin{aligned} \operatorname{grad}_{\mathcal{E}} \widetilde{f}_{\alpha\mathrm{LD}}(\boldsymbol{A}) &= \tfrac{4}{1-\alpha}(\boldsymbol{Q}_{k,\alpha}^{-1} - \boldsymbol{N}_k^{-1})\boldsymbol{A}\boldsymbol{\Lambda}_k, \\ \operatorname{hess}_{\mathcal{E}} \widetilde{f}_{\alpha\mathrm{LD}}(\boldsymbol{H})[\boldsymbol{\xi}] &= \tfrac{4}{1-\alpha}\left(\boldsymbol{Q}_{k,\alpha}^{-1}(\boldsymbol{\xi} - \dot{\boldsymbol{Q}}_{k,\alpha}\boldsymbol{Q}_{k,\alpha}^{-1}\boldsymbol{A})\boldsymbol{\Lambda}_k + \boldsymbol{A}^{-T}\boldsymbol{\xi}^T\boldsymbol{A}^{-T}\right), \end{aligned} \tag{B.22}$$

où $\boldsymbol{Q}_{k,\alpha} = \frac{1-\alpha}{2}\boldsymbol{M}_k + \frac{1+\alpha}{2}\boldsymbol{N}_k$ et $\dot{\boldsymbol{Q}}_{k,\alpha} = \frac{1+\alpha}{2}\dot{\boldsymbol{N}}_k$.

Distance riemannienne naturelle

On trouve dans [81] que

$$\mathrm{D}\|\log(\boldsymbol{X})\|_{\mathrm{F}}^2[\boldsymbol{\xi}] = 2\operatorname{tr}(\log(\boldsymbol{X})\boldsymbol{X}^{-1}\,\mathrm{D}\,\boldsymbol{X}[\xi]). \tag{B.23}$$

De ce fait,

$$\mathrm{D}\, f_{\mathrm{R}}(\boldsymbol{B})[\boldsymbol{\xi}] = 2\,\mathrm{tr}(\log(\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{-1/2})\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{\Lambda}_k^{-1/2}),$$

où $\boldsymbol{\Lambda}_k$ est la solution de (2.19) pour $\boldsymbol{M} = \boldsymbol{M}_k$. Comme $\boldsymbol{X}^{-1}\log(\boldsymbol{Y})\boldsymbol{X} = \log(\boldsymbol{X}^{-1}\boldsymbol{Y}\boldsymbol{X})$,

$$\mathrm{D}\, f_{\mathrm{R}}(\boldsymbol{B})[\boldsymbol{\xi}] = 2\,\mathrm{tr}(\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k) = 4\,\mathrm{tr}(\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T).$$

Donc,

$$\mathrm{grad}_{\mathcal{E}}\, f_{\mathrm{R}}(\boldsymbol{B}) = 4\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{B}^{-T}, \tag{B.24}$$

et

$$\begin{aligned}\mathrm{hess}_{\mathcal{E}}\, f_{\mathrm{R}}(\boldsymbol{B})[\boldsymbol{\xi}] = 4\,\mathrm{D}\log\left(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k\right)\left[\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{M}}_k - \boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k\right]\boldsymbol{B}^{-T} \\ - 4\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T},\end{aligned} \tag{B.25}$$

où $\mathrm{D}\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)[\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{M}}_k - \boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k]$ est la dérivée directionnelle du logarithme matriciel à $\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k$ dans la direction $\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{M}}_k - \boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k$. Dans ce cas, la matrice $\dot{\boldsymbol{\Lambda}}_k$, obtenue en dérivant l'équation (2.19), est la solution de

$$\mathrm{ddiag}(\mathrm{D}\log(\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k)[\boldsymbol{M}_k^{-1}\dot{\boldsymbol{\Lambda}}_k - \boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k]) = 0.$$

Nous ne proposons cependant pas de moyen pour résoudre cette équation dans ce manuscrit et nous nous limitons donc à des méthodes de gradient pour optimiser f_{R}.

Pour $\widehat{f}_{\mathrm{R}}$, on obtient

$$\begin{aligned}\mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{\mathrm{R}}(\boldsymbol{B}) &= 4\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{B}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{\mathrm{R}}(\boldsymbol{B})[\boldsymbol{\xi}] &= 4\,\mathrm{D}\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)[\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{M}}_k]\boldsymbol{B}^{-T} - 4\log(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{M}_k)\boldsymbol{B}^{-T}\boldsymbol{\xi}^T\boldsymbol{B}^{-T}.\end{aligned} \tag{B.26}$$

Enfin, on a pour $\widetilde{f}_{\mathrm{R}}$

$$\begin{aligned}\mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{R}}(\boldsymbol{A}) &= 4\log(\boldsymbol{M}_k^{-1}\boldsymbol{N}_k)\boldsymbol{A}^{-T}, \\ \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{R}}(\boldsymbol{A})[\boldsymbol{\xi}] &= 4\,\mathrm{D}\log(\boldsymbol{M}_k^{-1}\boldsymbol{N}_k)[\boldsymbol{M}_k^{-1}\dot{\boldsymbol{N}}_k]\boldsymbol{A}^{-T} - 4\log(\boldsymbol{M}_k^{-1}\boldsymbol{N}_k)\boldsymbol{A}^{-T}\boldsymbol{\xi}^T\boldsymbol{A}^{-T}.\end{aligned} \tag{B.27}$$

Pour estimer le logarithme matricielle et sa dérivée directionnelle première, nous utilisons les méthodes proposées dans [8] et présentées dans l'annexe C.

Distance log-euclidienne

Le critère f_{LE} peut se réécrire

$$f_{\mathrm{LE}}(\boldsymbol{B}) = \mathrm{tr}\left((\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k))^2\right),$$

avec $\boldsymbol{\Lambda}_k = \exp(\mathrm{ddiag}(\log(\boldsymbol{M}_k)))$. Dans [47], il est montré que

$$\mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{M}}_k] = \int_0^1 [\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}\dot{\boldsymbol{M}}_k[\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}ds.$$

Il suit que

$$\begin{aligned}
\mathrm{D}\, f_{\mathrm{LE}}(\boldsymbol{B})[\boldsymbol{\xi}] &= 2\,\mathrm{tr}\left((\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k))\,\mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{M}}_k]\right) \\
&= 2\,\mathrm{tr}\left((\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k))\int_0^1 [\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}\dot{\boldsymbol{M}}_k[\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}ds\right) \\
&= 2\int_0^1 \mathrm{tr}\left([\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}(\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k))[\boldsymbol{M}_k - \boldsymbol{I}_n)s + \boldsymbol{I}_n]^{-1}\dot{\boldsymbol{M}}_k\right) ds \\
&= 2\,\mathrm{tr}\left(\mathrm{D}\log(\boldsymbol{M}_k)[\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k)]\dot{\boldsymbol{M}}_k\right) \\
&= 4\,\mathrm{tr}\left(\mathrm{D}\log(\boldsymbol{M}_k)[\log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k)]\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T\right).
\end{aligned}$$

On note $\boldsymbol{O}_k = \log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k) = \log(\boldsymbol{M}_k) - \mathrm{ddiag}(\log(\boldsymbol{M}_k))$. On obtient donc

$$\mathrm{grad}_{\mathcal{E}}\, f_{\mathrm{LE}}(\boldsymbol{B}) = 4\,\mathrm{D}\log(\boldsymbol{M}_k)[\boldsymbol{O}_k]\boldsymbol{B}\boldsymbol{C}_k, \tag{B.28}$$

et

$$\begin{aligned}
\mathrm{hess}_{\mathcal{E}}\, f_{\mathrm{LE}}(\boldsymbol{B})[\boldsymbol{\xi}] = 4\,\mathrm{D}\log(\boldsymbol{M}_k)[\boldsymbol{O}_k]\boldsymbol{\xi}\boldsymbol{C}_k \\
+ 4\left(\mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{O}}_k] + \mathrm{D}^2\log(\boldsymbol{M}_k)[\boldsymbol{O}_k, \dot{\boldsymbol{M}}_k]\right)\boldsymbol{B}\boldsymbol{C}_k,
\end{aligned} \tag{B.29}$$

où $\dot{\boldsymbol{O}}_k = \mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{M}}_k] - \mathrm{ddiag}(\mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{M}}_k])$.

Pour $\widehat{f}_{\mathrm{LE}}$, on obtient

$$\begin{aligned}
\mathrm{grad}_{\mathcal{E}}\, \widehat{f}_{\mathrm{LE}}(\boldsymbol{B}) &= 4\,\mathrm{D}\log(\boldsymbol{M}_k)[\boldsymbol{O}_k]\boldsymbol{B}\boldsymbol{C}_k, \\
\mathrm{hess}_{\mathcal{E}}\, \widehat{f}_{\mathrm{LE}}(\boldsymbol{B})[\boldsymbol{\xi}] &= 4\,\mathrm{D}\log(\boldsymbol{M}_k)[\boldsymbol{O}_k]\boldsymbol{\xi}\boldsymbol{C}_k \\
&\quad +4\left(\mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{O}}_k] + \mathrm{D}^2\log(\boldsymbol{M}_k)[\boldsymbol{O}_k, \dot{\boldsymbol{M}}_k]\right)\boldsymbol{B}\boldsymbol{C}_k,
\end{aligned} \tag{B.30}$$

avec $\boldsymbol{O}_k = \log(\boldsymbol{M}_k) - \log(\boldsymbol{\Lambda}_k)$ et $\dot{\boldsymbol{O}}_k = \mathrm{D}\log(\boldsymbol{M}_k)[\dot{\boldsymbol{M}}_k]$. Enfin, on a pour $\widetilde{f}_{\mathrm{LE}}$

$$\begin{aligned}
\mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{LE}}(\boldsymbol{A}) &= 4\,\mathrm{D}\log(\boldsymbol{N}_k)[\boldsymbol{O}_k]\boldsymbol{A}\boldsymbol{\Lambda}_k, \\
\mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{LE}}(\boldsymbol{A})[\boldsymbol{\xi}] &= 4\,\mathrm{D}\log(\boldsymbol{N}_k)[\boldsymbol{O}_k]\boldsymbol{\xi}\boldsymbol{\Lambda}_k \\
&\quad +4\left(\mathrm{D}\log(\boldsymbol{N}_k)[\dot{\boldsymbol{O}}_k] + \mathrm{D}^2\log(\boldsymbol{M}_k)[\boldsymbol{O}_k, \dot{\boldsymbol{N}}_k]\right)\boldsymbol{A}\boldsymbol{\Lambda}_k,
\end{aligned} \tag{B.31}$$

avec $\boldsymbol{O}_k = \log(\boldsymbol{N}_k) - \log(\boldsymbol{M}_k)$ et $\dot{\boldsymbol{O}}_k = \mathrm{D}\log(\boldsymbol{N}_k)[\dot{\boldsymbol{N}}_k]$. Dans annexe C, nous présentons des méthodes pour estimer le logarithme matriciel et sa dérivée directionnelle première, et nous proposons une nouvelle méthode pour évaluer numériquement la dérivée directionnelle seconde.

Distance de Wasserstein

Bien que la distance de Wasserstein soit définie sur l'ensemble des matrices symétriques positives semi-définies, nous considérons que les matrices $\boldsymbol{C}_k$ sont positives définies pour dériver les gradients et les hessiennes. On a

$$\mathrm{D}\, f_{\mathrm{W}}(\boldsymbol{B})[\boldsymbol{\xi}] = \mathrm{tr}\left(\frac{1}{2}\dot{\boldsymbol{M}}_k - \mathrm{D}(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2})^{1/2}[\boldsymbol{\xi}]\right),$$

où $\boldsymbol{\Lambda}_k$ est la solution de (2.23) pour $\boldsymbol{M} = \boldsymbol{M}_k$. La matrice $\boldsymbol{\eta}_k = \mathrm{D}(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2})^{1/2}[\boldsymbol{\xi}]$ est la solution de l'équation de Sylvester

$$\left(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2}\right)^{1/2}\boldsymbol{\eta}_k + \boldsymbol{\eta}_k\left(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2}\right)^{1/2} = \boldsymbol{\Lambda}_k^{1/2}\dot{\boldsymbol{M}}_k\boldsymbol{\Lambda}_k^{1/2}.$$

De plus, en notant $\boldsymbol{Q}_k = (\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2})^{-1/2}$,

$$\mathrm{tr}(\boldsymbol{\eta}_k) = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\eta}_k\boldsymbol{Q}_k^{-1}\boldsymbol{Q}_k + \boldsymbol{Q}_k\boldsymbol{Q}_k^{-1}\boldsymbol{\eta}_k\right) = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\dot{\boldsymbol{M}}_k\boldsymbol{\Lambda}_k^{1/2}\right).$$

Donc

$$\mathrm{D}\,f_{\mathrm{W}}(\boldsymbol{B})[\boldsymbol{\xi}] = \frac{1}{2}\,\mathrm{tr}\left(\dot{\boldsymbol{M}}_k - \boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\dot{\boldsymbol{M}}_k\boldsymbol{\Lambda}_k^{1/2}\right) = \mathrm{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{\Lambda}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\right)\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{\xi}^T\right).$$

On obtient finalement

$$\mathrm{grad}_{\mathcal{E}}\,f_{\mathrm{W}}(\boldsymbol{B}) = \left(\boldsymbol{I}_n - \boldsymbol{\Lambda}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\right)\boldsymbol{B}\boldsymbol{C}_k. \tag{B.32}$$

Comme $\mathcal{D}_n$ est commutatif, $\mathrm{D}\,\boldsymbol{\Lambda}_k^{1/2}[\boldsymbol{\xi}] = \frac{1}{2}\boldsymbol{\Lambda}_k^{-1/2}\dot{\boldsymbol{\Lambda}}_k$. La hessienne est donc

$$\begin{aligned}\mathrm{hess}_{\mathcal{E}}\,f_{\mathrm{W}}(\boldsymbol{B})[\boldsymbol{\xi}] = &\left(\boldsymbol{I}_n - \boldsymbol{\Lambda}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\right)\boldsymbol{\xi}\boldsymbol{C}_k \\ &- \boldsymbol{\Lambda}_k^{1/2}\left(\dot{\boldsymbol{Q}}_k + \frac{1}{2}\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{Q}_k + \frac{1}{2}\boldsymbol{Q}_k\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\right)\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{B}\boldsymbol{C}_k,\end{aligned} \tag{B.33}$$

où $\dot{\boldsymbol{Q}}_k$ est la solution de l'équation de Sylvester

$$\boldsymbol{Q}_k\dot{\boldsymbol{Q}}_k + \dot{\boldsymbol{Q}}_k\boldsymbol{Q}_k = -\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{M}_k^{-1}\left(\dot{\boldsymbol{M}}_k + \frac{1}{2}\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{M}_k + \frac{1}{2}\boldsymbol{M}_k\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\right)\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k^{-1/2}.$$

Dans ce cas, la matrice $\dot{\boldsymbol{\Lambda}}_k$, obtenue en dérivant l'équation (2.23), est la solution de

$$\mathrm{ddiag}\left(\boldsymbol{\nu}_k\left(\dot{\boldsymbol{\Lambda}}_k\right)\right) = \dot{\boldsymbol{\Lambda}}_k,$$

où $\boldsymbol{\nu}_k(\dot{\boldsymbol{\Lambda}}_k)$ est la solution de l'équation de Sylvester

$$\begin{aligned}&\left(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2}\right)^{1/2}\boldsymbol{\nu}_k\left(\dot{\boldsymbol{\Lambda}}_k\right) + \boldsymbol{\nu}_k\left(\dot{\boldsymbol{\Lambda}}_k\right)\left(\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2}\right)^{1/2} = \\ &\qquad\qquad \boldsymbol{\Lambda}_k^{1/2}\left(\dot{\boldsymbol{M}}_k + \frac{1}{2}\boldsymbol{\Lambda}_k^{-1}\dot{\boldsymbol{\Lambda}}_k\boldsymbol{M}_k + \frac{1}{2}\boldsymbol{M}_k\dot{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_k^{-1}\right)\boldsymbol{\Lambda}_k^{1/2}.\end{aligned}$$

Ici encore, nous ne proposons pas de méthode pour trouver $\dot{\boldsymbol{\Lambda}}_k$ et nous nous limitons donc à des méthodes de gradient pour optimiser f_{W}.

Pour $\widehat{f}_{\mathrm{W}}$, on obtient

$$\begin{aligned}\mathrm{grad}_{\mathcal{E}}\,\widehat{f}_{\mathrm{W}}(\boldsymbol{B}) &= \left(\boldsymbol{I}_n - \boldsymbol{\Lambda}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\right)\boldsymbol{B}\boldsymbol{C}_k, \\ \mathrm{hess}_{\mathcal{E}}\,\widehat{f}_{\mathrm{W}}(\boldsymbol{B})[\boldsymbol{\xi}] &= \left(\boldsymbol{I}_n - \boldsymbol{\Lambda}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{\Lambda}_k^{1/2}\right)\boldsymbol{\xi}\boldsymbol{C}_k - \boldsymbol{\Lambda}_k^{1/2}\dot{\boldsymbol{Q}}_k\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{B}\boldsymbol{C}_k,\end{aligned} \tag{B.34}$$

où $\boldsymbol{Q}_k = (\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{M}_k\boldsymbol{\Lambda}_k^{1/2})^{-1/2}$ et $\dot{\boldsymbol{Q}}_k$ est la solution de l'équation de Sylvester

$$\boldsymbol{Q}_k\dot{\boldsymbol{Q}}_k + \dot{\boldsymbol{Q}}_k\boldsymbol{Q}_k = -\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{M}_k^{-1}\dot{\boldsymbol{M}}_k\boldsymbol{M}_k^{-1}\boldsymbol{\Lambda}_k^{-1/2}.$$

Enfin, pour $\widetilde{f}_{\mathrm{W}}$

$$\begin{aligned} \mathrm{grad}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{W}}(\boldsymbol{A}) &= \left(\boldsymbol{I}_n - \boldsymbol{M}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{M}_k^{1/2}\right)\boldsymbol{A}\boldsymbol{\Lambda}_k, \\ \mathrm{hess}_{\mathcal{E}}\, \widetilde{f}_{\mathrm{W}}(\boldsymbol{A})[\boldsymbol{\xi}] &= \left(\boldsymbol{I}_n - \boldsymbol{M}_k^{1/2}\boldsymbol{Q}_k\boldsymbol{M}_k^{1/2}\right)\boldsymbol{\xi}\boldsymbol{\Lambda}_k - \boldsymbol{M}_k^{1/2}\dot{\boldsymbol{Q}}_k\boldsymbol{M}_k^{1/2}\boldsymbol{A}\boldsymbol{\Lambda}_k, \end{aligned} \tag{B.35}$$

où $\boldsymbol{Q}_k = (\boldsymbol{M}_k^{1/2}\boldsymbol{N}_k\boldsymbol{M}_k^{1/2})^{-1/2}$ et $\dot{\boldsymbol{Q}}_k$ est la solution de l'équation de Sylvester

$$\boldsymbol{Q}_k\dot{\boldsymbol{Q}}_k + \dot{\boldsymbol{Q}}_k\boldsymbol{Q}_k = -\boldsymbol{M}_k^{-1/2}\boldsymbol{N}_k^{-1}\dot{\boldsymbol{N}}_k\boldsymbol{N}_k^{-1}\boldsymbol{M}_k^{-1/2}.$$

Les trois équations de Sylvester ci-dessus possèdent des solutions uniques car $\boldsymbol{Q}_k$ est symétrique positive définie dans les trois cas.

Annexe C

Logarithme matriciel et ses dérivées

Sommaire

Dans cette annexe, nous présentons des méthodes pour évaluer numériquement le logarithme matriciel et ses dérivées de Fréchet première et seconde. Dans cette thèse, nous utilisons ces outils pour estimer les critères de diagonalisation conjointe approximée du chapitre 2 basés sur la distance riemannienne naturelle et sur la distance log-euclidienne, ainsi que leurs gradients et hessiennes, qui sont donnés dans l'annexe B. Soient $\mathbb{M}_n = \mathbb{C}^{n\times n}$, $\mathcal{M}$ l'ensemble des matrices de $\mathbb{M}_n$ qui n'ont pas de valeurs propres sur la ligne fermée $\mathbb{R}^-$ et $\mathcal{N}$ l'ensemble des matrices de $\mathbb{M}_n$ dont les valeurs propres sont dans $\{\lambda \in \mathbb{C} : -\pi < \Im(\lambda) < \pi\}$, où $\Im(\cdot)$ retourne la partie imaginaire de son argument. Une matrice $\boldsymbol{Y} \in \mathbb{M}_n$ est un logarithme de $\boldsymbol{X}$ dans $\mathcal{M}$ si $\exp(\boldsymbol{Y}) = \boldsymbol{X}$, où $\exp(\cdot)$ est l'exponentielle matricielle définie par la série

$$\exp(\boldsymbol{Y}) = \sum_{k=0}^{\infty} \frac{\boldsymbol{Y}^k}{k!}. \tag{C.1}$$

Il y a un unique logarithme $\boldsymbol{Y}$ de $\boldsymbol{X}$ dans $\mathcal{N}$, qu'on appelle le logarithme principal de $\boldsymbol{X}$ noté $\log(\boldsymbol{X})$ [58, théorème 1.31]. L'estimation numérique du logarithme matriciel est un sujet étudié dans de nombreux travaux, voir par exemple [6-8, 35, 37, 47, 57, 58, 70, 71]. Pour estimer la dérivée première, on trouve une méthode générale valable pour toute fonction différentiable dans [58, chapitre 3] et la méthode qui constitue l'état de l'art pour le logarithme matriciel est l'algorithme spécifique proposé dans [8]. Pour la dérivée seconde, on trouve une méthode générale valable pour toute fonction différentiable deux fois dans [60]. Cependant, aucun algorithme spécifique au logarithme matriciel n'existe.

Notre contribution est d'étendre les méthodes de [8], qui permettent d'estimer le logarithme matriciel et sa dérivée première, afin de développer une méthode spécifique pour évaluer la

dérivée seconde du logarithme matriciel. Dans la section C.1, nous rappelons les définitions et les propriétés des dérivées (de Fréchet) première et seconde puis nous définissons un nouvel opérateur lié à la dérivée seconde et qui possède les propriétés essentielles pour étendre les travaux de [8]. Dans la section C.2, nous expliquons les méthodes qui permettent d'estimer le logarithme et sa dérivée première puis nous développons une nouvelle méthode qui permet d'évaluer la dérivée seconde du logarithme matriciel. Enfin, dans la section C.3, nous donnons les algorithmes détaillés dans le cas complexe et dans le cas réel.

C.1 Dérivées de Fréchet

On s'intéresse aux dérivées de Fréchet première et seconde de fonctions f définies de $\mathbb{M}_n$ dans $\mathbb{M}_n$ et différentiables deux fois. Dans la section C.1.1, nous présentons les dérivées première et seconde et nous rappelons leurs propriétés par rapport à la somme, au produit, à la composition et à l'inversion de fonctions. La dérivée de Fréchet seconde ne possède pas la propriété par rapport à l'inversion de fonctions qui est centrale pour étendre les travaux de [8]. Dans la section C.1.2, nous proposons un nouvel opérateur pour remédier à ce problème. Nous montrons que cet opérateur possède toutes les propriétés souhaitées et qu'il permet d'obtenir la dérivée de Fréchet seconde.

C.1.1 Dérivées première et seconde

L'ensemble des résultats présentés dans cette section est issu de [18, 58, 60]. La dérivée de Fréchet première d'une fonction $f : \mathbb{M}_n \to \mathbb{M}_n$ au point $\boldsymbol{X} \in \mathbb{M}_n$, notée $\mathrm{D}\, f(\boldsymbol{X})$, est l'application linéaire de $\mathbb{M}_n$ vers $\mathbb{M}_n$ telle que, pour tout $\boldsymbol{\xi} \in \mathbb{M}_n$,

$$\mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] = f(\boldsymbol{X} + \boldsymbol{\xi}) - f(\boldsymbol{X}) + o(\|\boldsymbol{\xi}\|). \tag{C.2}$$

La dérivée de Fréchet est proche de la dérivée de Gâteaux ou dérivée directionnelle, qui est donnée par

$$\lim_{t \to 0} \quad \frac{f(\boldsymbol{X} + t\boldsymbol{\xi}) - f(\boldsymbol{X})}{t}. \tag{C.3}$$

En effet, si la dérivée de Fréchet existe, alors elle est égale à la dérivée de Gâteaux et si la dérivée de Gâteaux existe, est linéaire par rapport à $\boldsymbol{\xi}$ et est continue par rapport à $\boldsymbol{X}$, alors c'est aussi la dérivée de Fréchet. Soient deux fonctions différentiables $g : \mathbb{M}_n \to \mathbb{M}_n$ et $h : \mathbb{M}_n \to \mathbb{M}_n$. La fonction $f = \alpha g + \beta h$ est différentiable et

$$\mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] = \alpha\, \mathrm{D}\, g(\boldsymbol{X})[\boldsymbol{\xi}] + \beta\, \mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\xi}]. \tag{C.4}$$

La fonction $f = gh$ est différentiable et

$$\mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] = \mathrm{D}\, g(\boldsymbol{X})[\boldsymbol{\xi}] h(\boldsymbol{X}) + g(\boldsymbol{X})\, \mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\xi}]. \tag{C.5}$$

La fonction $f = g \circ h$ est différentiable et

$$\mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] = \mathrm{D}\, g(h(\boldsymbol{X}))\, [\mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\xi}]]\,. \tag{C.6}$$

Enfin, si f et f^{-1} sont toutes deux définies et différentiables, alors en prenant $\boldsymbol{X} \in \mathbb{M}_n$, $\boldsymbol{Y} = f(\boldsymbol{X})$ et $\boldsymbol{\xi} \in \mathbb{M}_n$, on a la relation

$$\mathrm{D}\, f(\boldsymbol{X}) \left[\mathrm{D}\, f^{-1}(\boldsymbol{Y})[\boldsymbol{\xi}]\right] = \boldsymbol{\xi}. \tag{C.7}$$

La dérivée de Fréchet seconde d'une fonction $f : \mathbb{M}_n \to \mathbb{M}_n$ au point $\boldsymbol{X} \in \mathbb{M}_n$, notée $\mathrm{D}^2 f(\boldsymbol{X})$, est l'application bilinéaire de $(\mathbb{M}_n)^2$ vers $\mathbb{M}_n$ telle que, pour tout $(\boldsymbol{\xi}, \boldsymbol{\eta}) \in (\mathbb{M}_n)^2$,

$$\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \mathrm{D}\, f(\boldsymbol{X} + \boldsymbol{\eta})[\boldsymbol{\xi}] - \mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] + o(\|\boldsymbol{\eta}\|). \tag{C.8}$$

Elle correspond à la dérivée de Fréchet de l'application $\boldsymbol{X} \mapsto \mathrm{D}\, f(\boldsymbol{X})$. Cette application est symétrique par rapport à ses deux variables, *i.e.*,

$$\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\eta}, \boldsymbol{\xi}]. \tag{C.9}$$

Soient deux fonctions deux fois différentiables $g : \mathbb{M}_n \to \mathbb{M}_n$ et $h : \mathbb{M}_n \to \mathbb{M}_n$. La fonction $f = \alpha g + \beta h$ est deux fois différentiable et

$$\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \alpha\, \mathrm{D}^2 g(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] + \beta\, \mathrm{D}^2 h(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}]. \tag{C.10}$$

La fonction $f = gh$ est deux fois différentiable et

$$\begin{aligned}\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \mathrm{D}\, g(\boldsymbol{X})[\boldsymbol{\xi}]\, \mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\eta}] + \mathrm{D}\, g(\boldsymbol{X})[\boldsymbol{\eta}]\, \mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\xi}] \\ + \mathrm{D}^2 g(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] h(\boldsymbol{X}) + g(\boldsymbol{X})\, \mathrm{D}^2 h(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}].\end{aligned} \tag{C.11}$$

La fonction $f = g \circ h$ est deux fois différentiable et

$$\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \mathrm{D}^2 g(h(\boldsymbol{X}))\left[\mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\xi}], \mathrm{D}\, h(\boldsymbol{X})[\boldsymbol{\eta}]\right] + \mathrm{D}\, g(h(\boldsymbol{X}))\left[\mathrm{D}^2 h(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}]\right]. \tag{C.12}$$

Pour illustrer, prenons $f(\boldsymbol{X}) = \boldsymbol{X}^2$. La dérivée première est $\mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}] = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\xi}\boldsymbol{X}$. Pour prouver ce résultat, on peut soit employer la définition (C.2) en remarquant que $f(\boldsymbol{X} + \boldsymbol{\xi}) - f(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\xi}\boldsymbol{X} + \boldsymbol{\xi}^2$, soit utiliser (C.5) avec $g(\boldsymbol{X}) = h(\boldsymbol{X}) = \boldsymbol{X}$. La dérivée seconde est quant à elle $\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \boldsymbol{\eta}\boldsymbol{\xi} + \boldsymbol{\xi}\boldsymbol{\eta}$. Dans ce cas également, on trouve le résultat en employant la définition (C.8) ou la propriété (C.11) avec $g(\boldsymbol{X}) = h(\boldsymbol{X}) = \boldsymbol{X}$. De plus, on a $f^{-1}(\boldsymbol{X}) = \boldsymbol{X}^{1/2}$ et, en utilisant (C.7), on obtient que $\boldsymbol{\xi}_{1/2} = \mathrm{D}\, f^{-1}(\boldsymbol{X})[\boldsymbol{\xi}]$ est la solution de l'équation de Sylvester $\boldsymbol{X}^{1/2}\boldsymbol{\xi}_{1/2} + \boldsymbol{\xi}_{1/2}\boldsymbol{X}^{1/2} = \boldsymbol{\xi}$. Nous revenons sur cet exemple pour la dérivée seconde à la fin de la section C.1.2.

C.1.2 Un nouvel opérateur lié à la dérivée de Fréchet seconde

Pour définir notre nouvel opérateur, la clé est d'adopter un formalisme différent de celui utilisé dans la section C.1.1. Dans la suite, $f : \mathbb{M}_n \to \mathbb{M}_n$ est une fonction deux fois différentiable. Commençons par définir l'opérateur $\mathfrak{D}f$ comme l'application de $(\mathbb{M}_n)^2$ vers $\mathbb{M}_n$ telle que, pour tout $(\boldsymbol{X}, \boldsymbol{\xi}) \in (\mathbb{M}_n)^2$,

$$\mathfrak{D}f(\boldsymbol{X}, \boldsymbol{\xi}) = \mathrm{D}\, f(\boldsymbol{X})[\boldsymbol{\xi}]. \tag{C.13}$$

Étant donné $(\boldsymbol{X}, \boldsymbol{\xi}) \in (\mathbb{M}_n)^2$, on définit l'opérateur $\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})$ comme la dérivée première de $\mathfrak{D} f$ à $(\boldsymbol{X}, \boldsymbol{\xi})$, *i.e.*, $\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi}) = \mathrm{D}(\mathfrak{D} f)(\boldsymbol{X}, \boldsymbol{\xi})$. Autrement dit, $\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})$ est l'application bilinéaire de $(\mathbb{M}_n)^2$ vers $\mathbb{M}_n$ telle que, pour tout $(\boldsymbol{\eta}, \boldsymbol{\nu}) \in (\mathbb{M}_n)^2$,

$$\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \mathfrak{D} f(\boldsymbol{X} + \boldsymbol{\eta}, \boldsymbol{\xi} + \boldsymbol{\nu}) - \mathfrak{D} f(\boldsymbol{X}, \boldsymbol{\xi}) + o(\|(\boldsymbol{\eta}, \boldsymbol{\nu})\|). \tag{C.14}$$

L'opérateur $\mathfrak{D}^2$ est lié à la dérivée de Fréchet seconde. En effet, on peut montrer que

$$\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] + \mathrm{D} f(\boldsymbol{X})[\boldsymbol{\nu}]. \tag{C.15}$$

En particulier, on obtient la dérivée de Fréchet seconde avec

$$\mathrm{D}^2 f(\boldsymbol{X})[\boldsymbol{\xi}, \boldsymbol{\eta}] = \mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{0}_n]. \tag{C.16}$$

L'opérateur $\mathfrak{D}^2$ s'avère également très utile lorsqu'on manipule des champs de vecteurs, *i.e.*, lorsque $\boldsymbol{\xi}$ est fonction de $\boldsymbol{X}$ et qu'on souhaite différencier $\boldsymbol{X} \mapsto \mathrm{D} f(\boldsymbol{X})[\boldsymbol{\xi}(\boldsymbol{X})]$ [1]. On a alors

$$\mathrm{D}\left(\mathrm{D} f(\boldsymbol{X})[\boldsymbol{\xi}(\boldsymbol{X})]\right)[\boldsymbol{\eta}] = \mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi}(\boldsymbol{X}))\left[\boldsymbol{\eta}, \mathrm{D}\,\boldsymbol{\xi}(\boldsymbol{X})[\boldsymbol{\eta}]\right]. \tag{C.17}$$

Proposition C.1
Soient deux fonctions différentiables deux fois $g : \mathbb{M}_n \to \mathbb{M}_n$ et $h : \mathbb{M}_n \to \mathbb{M}_n$. L'opérateur $\mathfrak{D}^2$ a les propriétés suivantes par rapport à la somme, au produit, à la composition et à l'inversion. La fonction $f = \alpha g + \beta h$ est différentiable deux fois et

$$\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \alpha \mathfrak{D}^2 g(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] + \beta \mathfrak{D}^2 h(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}].$$

La fonction $f = gh$ l'est aussi et

$$\begin{aligned}\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \mathfrak{D}^2 g(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] h(\boldsymbol{X}) + g(\boldsymbol{X}) \mathfrak{D}^2 h(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] \\ + \mathfrak{D} g(\boldsymbol{X}, \boldsymbol{\xi}) \mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\eta}) + \mathfrak{D} g(\boldsymbol{X}, \boldsymbol{\eta}) \mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\xi}).\end{aligned}$$

La fonction $f = g \circ h$ l'est également et

$$\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \mathfrak{D}^2 g(h(\boldsymbol{X}), \mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\xi}))\left[\mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\eta}), \mathfrak{D}^2 h(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]\right].$$

Enfin, si $f : \mathbb{M}_n \to \mathbb{M}_n$ et $f^{-1} : \mathbb{M}_n \to \mathbb{M}_n$ sont toutes deux définies et différentiables deux fois, alors en prenant $\boldsymbol{X} \in \mathbb{M}_n$, $\boldsymbol{Y} = f(\boldsymbol{X})$ et $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\nu} \in \mathbb{M}_n$, on a la relation

$$\mathfrak{D}^2 f(\boldsymbol{X}, \mathfrak{D} f^{-1}(\boldsymbol{Y}, \boldsymbol{\xi}))\left[\mathfrak{D} f^{-1}(\boldsymbol{Y}, \boldsymbol{\eta}), \mathfrak{D}^2 f^{-1}(\boldsymbol{Y}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]\right] = \boldsymbol{\nu}.$$

Démonstration. Pour la somme $f = \alpha g + \beta h$, la preuve est immédiate en utilisant la définition de l'opérateur $\mathfrak{D}^2$. Pour le produit $f = gh$, on a par définition, $\mathfrak{D} f(\boldsymbol{X}, \boldsymbol{\xi}) = \mathfrak{D} g(\boldsymbol{X}, \boldsymbol{\xi}) h(\boldsymbol{X}) + g(\boldsymbol{X}) \mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\xi})$. Il suit que

$$\mathfrak{D}^2 f(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = \mathrm{D}\left(\mathfrak{D} g(\boldsymbol{X}, \boldsymbol{\xi}) h(\boldsymbol{X}) + g(\boldsymbol{X}) \mathfrak{D} h(\boldsymbol{X}, \boldsymbol{\xi})\right)[\boldsymbol{\eta}, \boldsymbol{\nu}].$$

1. C'est notamment le cas pour le calcul de la hessienne de la fonction de coût de diagonalisation conjointe approximée basée sur la distance log-euclidienne (voir annexe B).

Le résultat est obtenu en utilisant la propriété (C.5) et en remarquant que $\mathrm{D}(g(\boldsymbol{X}))[\boldsymbol{\eta},\boldsymbol{\nu}] = \mathrm{D}\,g(\boldsymbol{X})[\boldsymbol{\eta}] = \mathfrak{D}g(\boldsymbol{X},\boldsymbol{\eta})$ et $\mathrm{D}(\mathfrak{D}g)(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = \mathfrak{D}^2 g(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]$. Pour la composition $f = g \circ h$, on a par définition, $\mathfrak{D}f(\boldsymbol{X},\boldsymbol{\xi}) = \mathfrak{D}g(h(\boldsymbol{X}),\mathfrak{D}h(\boldsymbol{X},\boldsymbol{\xi}))$. Soit $\bar{h} : (\mathbb{M}_n)^2 \to (\mathbb{M}_n)^2$ telle que $\bar{h}(\boldsymbol{X},\boldsymbol{\xi}) = (h(\boldsymbol{X}),\mathfrak{D}h(\boldsymbol{X},\boldsymbol{\xi}))$. Il suit que $\mathfrak{D}f = (\mathfrak{D}g) \circ \bar{h}$ et (C.6) entraîne

$$\mathfrak{D}^2 f(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = \mathrm{D}(\mathfrak{D}g)(\bar{h}(\boldsymbol{X},\boldsymbol{\xi}))\left[\mathrm{D}\,\bar{h}(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]\right].$$

Le résultat est obtenu en remarquant que $\mathrm{D}\,\bar{h}(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = (\mathfrak{D}h(\boldsymbol{X},\boldsymbol{\eta}),\mathfrak{D}^2 h(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}])$. Pour l'inversion, le membre de gauche de l'équation est obtenu en utilisant la propriété pour la composition avec $g = f$ et $h = f^{-1}$ et le membre de droite est obtenu en considérant $\varphi(\boldsymbol{Y}) = \boldsymbol{Y}$ pour laquelle on a $\mathfrak{D}\varphi(\boldsymbol{Y},\boldsymbol{\xi}) = \boldsymbol{\xi}$ et $\mathfrak{D}^2\varphi(\boldsymbol{Y},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = \boldsymbol{\nu}$. □

Pour illustrer, reprenons la fonction $f(\boldsymbol{X}) = \boldsymbol{X}^2$ pour laquelle on a $\mathfrak{D}f(\boldsymbol{X},\boldsymbol{\xi}) = \boldsymbol{X}\boldsymbol{\xi}+\boldsymbol{\xi}\boldsymbol{X}$. Dans ce cas, on obtient $\mathfrak{D}^2 f(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = \boldsymbol{\eta}\boldsymbol{\xi}+\boldsymbol{\xi}\boldsymbol{\eta}+\boldsymbol{X}\boldsymbol{\nu}+\boldsymbol{\nu}\boldsymbol{X}$ en utilisant la définition (C.14) ou la propriété par rapport au produit de la propostion C.1 avec $g(\boldsymbol{X}) = h(\boldsymbol{X}) = \boldsymbol{X}$. Pour la fonction inverse $f^{-1}(\boldsymbol{X}) = \boldsymbol{X}^{1/2}$, la propriété par rapport à l'inversion de la proposition C.1 entraîne que $\boldsymbol{\nu}_{1/2} = \mathfrak{D}^2 f^{-1}(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]$ est la solution de l'équation de Sylvester

$$\boldsymbol{X}^{1/2}\boldsymbol{\nu}_{1/2} + \boldsymbol{\nu}_{1/2}\boldsymbol{X}^{1/2} = \boldsymbol{\nu} - \boldsymbol{\xi}_{1/2}\boldsymbol{\eta}_{1/2} - \boldsymbol{\eta}_{1/2}\boldsymbol{\xi}_{1/2},$$

où $\boldsymbol{\xi}_{1/2} = \mathrm{D}\,f^{-1}(\boldsymbol{X})[\boldsymbol{\xi}]$ est la solution de l'équation de Sylvester $\boldsymbol{X}^{1/2}\boldsymbol{\xi}_{1/2} + \boldsymbol{\xi}_{1/2}\boldsymbol{X}^{1/2} = \boldsymbol{\xi}$ et $\boldsymbol{\eta}_{1/2} = \mathrm{D}\,f^{-1}(\boldsymbol{X})[\boldsymbol{\eta}]$ est la solution de l'équation de Sylvester $\boldsymbol{X}^{1/2}\boldsymbol{\eta}_{1/2} + \boldsymbol{\eta}_{1/2}\boldsymbol{X}^{1/2} = \boldsymbol{\eta}$.

C.2 Évaluation du logarithme et de ses dérivées

Étant donné $\boldsymbol{X} \in \mathcal{M}$ et $\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{\nu} \in \mathbb{M}_n$, nous présentons dans cette section les méthodes proposées dans [8] pour évaluer le logarithme matriciel $\log(\boldsymbol{X})$ et sa dérivée $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$. Nous dérivons ensuite une nouvelle méthode pour estimer l'opérateur $\mathfrak{D}^2\log(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]$, qui permet d'obtenir la dérivée de Fréchet seconde du logarithme matriciel grâce à (C.16). Quand $\boldsymbol{X}$ est suffisamment proche de l'identité, on peut évaluer ces fonctions en utilisant des approximants de Padé, voir par exemple [8, 58]. Lorsque ce n'est pas le cas, on rapproche $\boldsymbol{X}$ de l'identité et on utilise une propriété du logarithme matriciel, de sa dérivée de Fréchet première et de l'opérateur $\mathfrak{D}^2$.

C.2.1 Logarithme et dérivée première

Nous nous plaçons d'abord dans le cas où $\boldsymbol{X}$ est suffisamment proche de l'identité $\boldsymbol{I}_n$. Le logarithme matriciel $\log(\boldsymbol{X})$ peut alors être évalué par un approximant de Padé $r_m(\boldsymbol{X})$, pour un entier m. On peut choisir de définir la fonction r_m de différentes façons. Dans [57], il est montré qu'on obtient le meilleur équilibre entre efficacité et stabitilité numérique avec

$$r_m(\boldsymbol{X}) = \sum_{i=1}^{m} \alpha_i^{(m)} \left[\left(1-\beta_i^{(m)}\right)\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{X}\right]^{-1}(\boldsymbol{X}-\boldsymbol{I}_n), \tag{C.18}$$

Algorithme C.1 : Logarithme matriciel et dérivée première : algorithme de base

Input : $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{\xi} \in \mathbb{M}_n$
Output : $\log(\boldsymbol{X})$, $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$
Définir $\boldsymbol{X}_0 = \boldsymbol{X}$ et $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$
for $j = 1$ *à* s **do**
  $\boldsymbol{X}_j = \boldsymbol{X}_{j-1}^{1/2}$
  Résoudre l'équation de Sylvester $\boldsymbol{X}_j\boldsymbol{\xi}_j + \boldsymbol{\xi}_j\boldsymbol{X}_j = \boldsymbol{\xi}_{j-1}$
$\log(\boldsymbol{X}) \approx 2^s r_m(\boldsymbol{X}_s)$
$\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}] \approx 2^s \mathrm{D}\, r_m(\boldsymbol{X}_s)[\boldsymbol{\xi}_s]$

où les coefficients $\alpha_i^{(m)} \in [0,1]$ et $\beta_i^{(m)} \in [0,1]$ sont les poids et les noeuds de la méthode de quadrature de Gauss-Legendre à m points. Dans ce cas, la dérivée directionnelle première $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$ peut être estimée par $\mathrm{D}\, r_m(\boldsymbol{X})[\boldsymbol{\xi}]$ [8], où

$$\mathrm{D}\, r_m(\boldsymbol{X})[\boldsymbol{\xi}] = \sum_{i=1}^{m} \alpha_i^{(m)} \left[\left(1-\beta_i^{(m)}\right)\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{X}\right]^{-1} \boldsymbol{\xi} \left[\left(1-\beta_i^{(m)}\right)\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{X}\right]^{-1}. \quad \text{(C.19)}$$

Lorsque $\boldsymbol{X}$ n'est pas suffisamment proche de l'identité, la solution est d'utiliser le fait que la suite $\boldsymbol{X}^{1/2^s}$ converge vers $\boldsymbol{I}_n$ quand s tend vers l'infini et que $\log(\boldsymbol{X}) = 2^s \log(\boldsymbol{X}^{1/2^s})$ et $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}] = 2^s \mathrm{D}\log(\boldsymbol{X}^{1/2^s})[\boldsymbol{\xi}_s]$, où $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$ et $\boldsymbol{X}^{1/2^j}\boldsymbol{\xi}_j + \boldsymbol{\xi}_j\boldsymbol{X}^{1/2^j} = \boldsymbol{\xi}_{j-1}$, $j \in \{1,\ldots,s\}$. On obtient alors l'algorithme C.1, qui correspond à [8, algorithme 2.1]. En ce qui concerne le choix des entiers s et m, nous renvoyons vers [7, algorithme 4.1]. L'algorithme C.1 peut être amélioré en termes de vitesse et de précision en utilisant la décomposition de Schur [8], ce que nous faisons dans la section C.3.

C.2.2 Extension à la dérivée seconde

En utilisant le même raisonnement, lorsque $\boldsymbol{X} \in \mathcal{M}$ est suffisamment proche de l'identité, la dérivée seconde $\mathfrak{D}^2\log(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]$ peut être approximée par $\mathfrak{D}^2 r_m(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}]$, où

$$\mathfrak{D}^2 r_m(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = \mathrm{D}\, r_m(\boldsymbol{X})[\boldsymbol{\nu}] - \sum_{i=1}^{m} \alpha_i^{(m)}\beta_i^{(m)} \boldsymbol{M}_i^{(m)} \left(\boldsymbol{\eta}\boldsymbol{M}_i^{(m)}\boldsymbol{\xi} + \boldsymbol{\xi}\boldsymbol{M}_i^{(m)}\boldsymbol{\eta}\right)\boldsymbol{M}_i^{(m)}, \quad \text{(C.20)}$$

avec $\boldsymbol{M}_i^{(m)} = [(1-\beta_i^{(m)})\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{X}]^{-1}$. Lorsque $\boldsymbol{X}$ n'est pas suffisamment proche de l'identité, on peut utiliser la relation

$$\mathfrak{D}^2\log(\boldsymbol{X},\boldsymbol{\xi})[\boldsymbol{\eta},\boldsymbol{\nu}] = 2^s\mathfrak{D}^2\log(\boldsymbol{X}_s,\boldsymbol{\xi}_s)[\boldsymbol{\eta}_s,\boldsymbol{\nu}_s], \quad \text{(C.21)}$$

où, pour $j \in \{1,\ldots,s\}$, $\boldsymbol{X}_j = \boldsymbol{X}^{1/2^j}$, $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$ et $\boldsymbol{X}_j\boldsymbol{\xi}_j + \boldsymbol{\xi}_j\boldsymbol{X}_j = \boldsymbol{\xi}_{j-1}$, $\boldsymbol{\eta}_0 = \boldsymbol{\eta}$ et $\boldsymbol{X}_j\boldsymbol{\eta}_j + \boldsymbol{\eta}_j\boldsymbol{X}_j = \boldsymbol{\eta}_{j-1}$, $\boldsymbol{\nu}_0 = \boldsymbol{\nu}$ et $\boldsymbol{X}_j\boldsymbol{\nu}_j + \boldsymbol{\nu}_j\boldsymbol{X}_j = \boldsymbol{\nu}_{j-1} - \boldsymbol{\xi}_j\boldsymbol{\eta}_j - \boldsymbol{\eta}_j\boldsymbol{\xi}_j$. On obtient enfin l'algorithme C.2.

Algorithme C.2 : Logarithme matriciel et ses dérivées : algorithme de base

Input : $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\nu} \in \mathbb{M}_n$
Output : $\log(\boldsymbol{X})$, $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$, $\mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]$
1 Définir $\boldsymbol{X}_0 = \boldsymbol{X}$, $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$, $\boldsymbol{\eta}_0 = \boldsymbol{\eta}$ et $\boldsymbol{\nu}_0 = \boldsymbol{\nu}$
2 **for** $j = 1$ *à* s **do**
3 $\quad \boldsymbol{X}_j = \boldsymbol{X}_{j-1}^{1/2}$
4 $\quad$ Résoudre les équations de Sylvester :
5 $\quad \boldsymbol{X}_j\boldsymbol{\xi}_j + \boldsymbol{\xi}_j\boldsymbol{X}_j = \boldsymbol{\xi}_{j-1}$
6 $\quad \boldsymbol{X}_j\boldsymbol{\eta}_j + \boldsymbol{\eta}_j\boldsymbol{X}_j = \boldsymbol{\eta}_{j-1}$
7 $\quad \boldsymbol{X}_j\boldsymbol{\nu}_j + \boldsymbol{\nu}_j\boldsymbol{X}_j = \boldsymbol{\nu}_{j-1} - \boldsymbol{\xi}_j\boldsymbol{\eta}_j - \boldsymbol{\eta}_j\boldsymbol{\xi}_j$
8 $\log(\boldsymbol{X}) \approx 2^s r_m(\boldsymbol{X}_s)$
9 $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}] \approx 2^s\,\mathrm{D}\,r_m(\boldsymbol{X}_s)[\boldsymbol{\xi}_s]$
10 $\mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] \approx 2^s\mathfrak{D}^2 r_m(\boldsymbol{X}_s, \boldsymbol{\xi}_s)[\boldsymbol{\eta}_s, \boldsymbol{\nu}_s]$

C.3 Algorithmes

Dans cette section, nous détaillons l'algorithme C.2. Nous faisons la distinction entre le cas où les matrices manipulées sont complexes et celui où elles sont réelles. Pour construire des algorithmes efficaces, nous utilisons en effet la décomposition de Schur qui n'est pas la même dans le cas complexe que dans le cas réel. Bien qu'il soit possible d'utiliser l'algorithme complexe lorsqu'on manipule des matrices réelles, l'algorithme réel est avantageux en termes de vitesse, de mémoire et de précision [8]. Si on enlève les lignes qui correspondent à l'évaluation de la dérivée seconde du logarithme matriciel, les algorithmes donnés se réduisent aux algorithmes 5.1 et 6.1 de [8]. Notre contribution revient ainsi à ajouter les lignes 17, 18, 22, 23, 25, 29, 31 et 33 de l'algorithme C.3, ce qui impacte la ligne 17 de l'algorithme C.4. Ces modifications permettent ainsi de calculer $\mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]$ pour, essentiellement, le même coût de calcul.

C.3.1 Cas complexe

Dans le cas complexe, la décomposition de Schur d'une matrice $\boldsymbol{X}$ est

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^H, \tag{C.22}$$

où $\boldsymbol{Q}$ est une matrice unitaire et $\boldsymbol{T}$ est une matrice triangulaire supérieure appelée forme de Schur de $\boldsymbol{X}$ (voir par exemple [54] pour plus de détails). L'opérateur $\cdot^H$ correspond à la transposée conjuguée. On a alors

$$\begin{aligned} \log(\boldsymbol{X}) &= \boldsymbol{Q}\log(\boldsymbol{T})\boldsymbol{Q}^H, \\ \mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}] &= \boldsymbol{Q}\,\mathrm{D}\log(\boldsymbol{T})[\boldsymbol{Q}^H\boldsymbol{\xi}\boldsymbol{Q}]\boldsymbol{Q}^H, \\ \mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] &= \boldsymbol{Q}\mathfrak{D}^2\log(\boldsymbol{T}, \boldsymbol{Q}^H\boldsymbol{\xi}\boldsymbol{Q})[\boldsymbol{Q}^H\boldsymbol{\eta}\boldsymbol{Q}, \boldsymbol{Q}^H\boldsymbol{\nu}\boldsymbol{Q}]\boldsymbol{Q}^H. \end{aligned} \tag{C.23}$$

Travailler avec une matrice triangulaire est moins coûteux et plus précis [8]. On obtient enfin l'algorithme C.3.

Algorithme C.3 : Logarithme matriciel et ses dérivées : algorithme complexe

Input : $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\nu} \in \mathbb{M}_n$

Output : $\log(\boldsymbol{X})$, $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$, $\mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]$

Calculer la décomposition de Schur complexe $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^H$

$\boldsymbol{T}_0 = \boldsymbol{T}$

Déterminer les entiers s et m comme dans [7, algorithme 4.1]

for $j = 1$ *à* s **do**

  Calculer $\boldsymbol{T}_j = \boldsymbol{T}_{j-1}^{1/2}$ avec la récurrence donnée dans [22] et [58, algorithme 6.1]

$\boldsymbol{R} = \boldsymbol{T}_s - \boldsymbol{I}_n$

Remplacer la diagonale et la première super-diagonale de $\boldsymbol{R}$ par celles de $\boldsymbol{T}_0^{1/2^s} - \boldsymbol{I}_n$ évaluées avec [6, algorithme 2] et [59, équation (5.6)]

$\boldsymbol{Y} = \boldsymbol{0}_n$

for $i = 1$ *à* m **do**

  Résoudre $(\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{R})\boldsymbol{U} = \alpha_i^{(m)}\boldsymbol{R}$ d'inconnue $\boldsymbol{U}$ par substitution

  $\boldsymbol{Y} \leftarrow \boldsymbol{Y} + \boldsymbol{U}$

$\boldsymbol{Y} \leftarrow 2^s\boldsymbol{Y}$

Remplacer la diagonale de $\boldsymbol{Y}$ par le logarithme de la diagonale de $\boldsymbol{T}_0$

Remplacer la première super-diagonale de $\boldsymbol{Y}$ par [58, équation (11.28)]

$\log(\boldsymbol{X}) = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^H$

$\boldsymbol{\xi}_0 = \boldsymbol{Q}^H\boldsymbol{\xi}\boldsymbol{Q}$

$\boldsymbol{\eta}_0 = \boldsymbol{Q}^H\boldsymbol{\eta}\boldsymbol{Q}$

$\boldsymbol{\nu}_0 = \boldsymbol{Q}^H\boldsymbol{\nu}\boldsymbol{Q}$

for $j = 1$ *à* s **do**

  Résoudre les équations de Sylvester (par substitution) :

  $\boldsymbol{T}_j\boldsymbol{\xi}_j + \boldsymbol{\xi}_j\boldsymbol{T}_j = \boldsymbol{\xi}_{j-1}$ d'inconnue $\boldsymbol{\xi}_j$

  $\boldsymbol{T}_j\boldsymbol{\eta}_j + \boldsymbol{\eta}_j\boldsymbol{T}_j = \boldsymbol{\eta}_{j-1}$ d'inconnue $\boldsymbol{\eta}_j$

  $\boldsymbol{T}_j\boldsymbol{\nu}_j + \boldsymbol{\nu}_j\boldsymbol{T}_j = \boldsymbol{\nu}_{j-1} - \boldsymbol{\xi}_j\boldsymbol{\eta}_j - \boldsymbol{\eta}_j\boldsymbol{\xi}_j$ d'inconnue $\boldsymbol{\nu}_j$

$\boldsymbol{L} = \boldsymbol{0}_n$

$\boldsymbol{\mathfrak{L}} = \boldsymbol{0}_n$

for $i = 1$ *à* m **do**

  $\boldsymbol{M}_i^{(m)} = [\boldsymbol{I}_n + \beta_i^{(m)}\boldsymbol{R}]^{-1}$

  $\boldsymbol{V} = \alpha_i^{(m)}\boldsymbol{M}_i^{(m)}\boldsymbol{\xi}_s\boldsymbol{M}_i^{(m)}$

  $\boldsymbol{W} = \alpha_i^{(m)}\boldsymbol{M}_i^{(m)}(\boldsymbol{\nu}_s - \beta_i^{(m)}(\boldsymbol{\xi}_s\boldsymbol{M}_i^{(m)}\boldsymbol{\eta}_s + \boldsymbol{\eta}_s\boldsymbol{M}_i^{(m)}\boldsymbol{\xi}_s))\boldsymbol{M}_i^{(m)}$

  $\boldsymbol{L} \leftarrow \boldsymbol{L} + \boldsymbol{V}$

  $\boldsymbol{\mathfrak{L}} \leftarrow \boldsymbol{\mathfrak{L}} + \boldsymbol{W}$

$\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}] = 2^s\boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^H$

$\mathfrak{D}^2\log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}] = 2^s\boldsymbol{Q}\boldsymbol{\mathfrak{L}}\boldsymbol{Q}^H$

C.3.2 Cas réel

Dans le cas réel, la décomposition de Schur d'une matrice $\boldsymbol{X}$ est

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^T, \tag{C.24}$$

où $\boldsymbol{Q}$ est une matrice orthogonale et $\boldsymbol{T}$ est une matrice quasi-triangulaire supérieure, *i.e.*, une matrice triangulaire supérieure par bloc dont les blocs diagonaux sont de dimension 1×1 ou 2×2. L'équation (C.23) reste valide dans le cas réel et on obtient enfin l'algorithme C.4.

Algorithme C.4 : Logarithme matriciel et ses dérivées : algorithme réel

Input : $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\nu} \in \mathbb{R}^{n \times n}$
Output : $\log(\boldsymbol{X})$, $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$, $\mathfrak{D}^2 \log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]$
Calculer la décomposition de Schur réelle $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^T$
$\boldsymbol{T}_0 = \boldsymbol{T}$
Déterminer les entiers s et m comme dans [7, algorithme 4.1]
for $j = 1$ *à* s **do**
    Calculer $\boldsymbol{T}_j = \boldsymbol{T}_{j-1}^{1/2}$ avec la récurrence donnée dans [56] et [58, algorithme 6.7]
$\boldsymbol{R} = \boldsymbol{T}_s - \boldsymbol{I}_n$
Remplacer les blocs diagonaux de $\boldsymbol{R}$ par ceux de $\boldsymbol{T}_0^{1/2^s} - \boldsymbol{I}_n$ évalués avec [6, algorithme 2] pour les blocs 1×1 et avec [7, algorithme 5.1] pour les blocs 2×2 (où les racines carrées sont estimées avec [58, équation (6.9)])
for $i = 1$ *à* $n - 1$ **do**
    if $\boldsymbol{R}_{i,i}$ *et* $\boldsymbol{R}_{i+1,i+1}$ *sont dans des blocs* 1×1 **then**
        Recalculer $\boldsymbol{R}_{i,i+1}$ avec [59, équation (5.6)]
Évaluer $\boldsymbol{Y}$ comme dans les lignes 8 à 12 de l'algorithme C.3
Remplacer les blocs diagonaux de $\boldsymbol{Y}$ par $\log((\boldsymbol{T}_0)_{i,i})$ pour les blocs 1×1 et en utilisant [8, équation (6.1)] pour les blocs 2×2
for $i = 1$ *à* $n - 1$ **do**
    if $\boldsymbol{Y}_{i,i}$ *et* $\boldsymbol{Y}_{i+1,i+1}$ *sont dans des blocs* 1×1 **then**
        Recalculer $\boldsymbol{Y}_{i,i+1}$ avec [58, équation (11.28)]
$\log(\boldsymbol{X}) = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^T$
Exécuter les lignes 16 à 33 de l'algorithme C.3 pour obtenir $\mathrm{D}\log(\boldsymbol{X})[\boldsymbol{\xi}]$ et $\mathfrak{D}^2 \log(\boldsymbol{X}, \boldsymbol{\xi})[\boldsymbol{\eta}, \boldsymbol{\nu}]$

Annexe D

Sous-variétés de GL_η : Preuves

Cette annexe contient les preuves des propositions 3.3, 3.5, 3.6 et 3.8 qui se trouvent dans la section 3.2 du chapitre 3. Ces propositions concernent l'intégration de la contrainte oblique (1.7) et de la contrainte intrinsèque (1.8) dans des sous-variétés de GL_n. Dans chaque cas, nous considérons trois structures riemanniennes qui proviennent de (i) GL_n équipé de la métrique invariante à gauche définie dans (1.47), (ii) GL_n équipé de la métrique invariante à droite également définie dans (1.47) et (iii) la variété polaire $\mathcal{P}_n$. Les preuves données ici sont celles des projecteurs orthogonaux sur les espaces tangents et des hessiennes riemanniennes.

Contrainte oblique

Preuve de la proposition 3.3

Étant donné $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{o}}$, l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$ lorsque GL_n est équipé de la métrique invariante à gauche, défini par $\{\boldsymbol{\eta} \in \mathbb{R}^{n\times n} : \langle \boldsymbol{\xi}, \boldsymbol{\eta}\rangle_{\boldsymbol{B}}^{\ell} = 0, \forall \boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}\}$, est $\{\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{B} : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. Cet espace possède en effet la bonne dimension, qui est n, et pour tout $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$

$$\langle \boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{B}\rangle_{\boldsymbol{B}}^{\ell} = \mathrm{tr}(\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{\Lambda}) = \mathrm{tr}(\mathrm{ddiag}(\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}) = 0.$$

De ce fait, $P_{\boldsymbol{B}}^{\mathrm{o},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B}$ et $\boldsymbol{\Lambda}_{\mathrm{o},\ell}$ est la matrice diagonale solution de l'équation $\mathrm{ddiag}(P_{\boldsymbol{B}}^{\mathrm{o},\ell}(\boldsymbol{Z})\boldsymbol{B}^T) = \boldsymbol{0}_n$, qui peut être vectorisée par

$$\left(\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{B}\boldsymbol{B}^T\right)\mathrm{diag}(\boldsymbol{\Lambda}_{\mathrm{o},\ell}) = \mathrm{diag}(\boldsymbol{Z}\boldsymbol{B}^T).$$

Comme $\boldsymbol{B} \in \mathrm{GL}_n$, $\boldsymbol{B}\boldsymbol{B}^T$ est positive définie et le théorème du produit de Schur assure que $\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{B}\boldsymbol{B}^T$ est inversible, ce qui complète la preuve pour le projecteur $P^{\mathrm{o},\ell}$.

De la même façon, l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$ quand GL_n est équipé de la métrique invariante à droite est $\{\boldsymbol{\Lambda}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B} : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. Il suit que $P_{\boldsymbol{B}}^{\mathrm{o},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B}$. Enfin, la résolution de $\mathrm{ddiag}(P_{\boldsymbol{B}}^{\mathrm{o},r}(\boldsymbol{Z})\boldsymbol{B}^T) = \boldsymbol{0}_n$ amène $\boldsymbol{\Lambda}_{\mathrm{o},r} = \mathrm{ddiag}(\boldsymbol{Z}\boldsymbol{B}^T)\,\mathrm{ddiag}\left((\boldsymbol{B}\boldsymbol{B}^T)^2\right)^{-1}$.

Pour terminer, étant donné $\mathcal{B} \in \mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$, l'espace normal à $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ dans $T_{\mathcal{B}}\mathcal{P}_n$ est donné par $\{(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{\Lambda}\boldsymbol{S})\boldsymbol{S}, \boldsymbol{0}_n) : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. En effet, on peut réécrire $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n} = \{(\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{P}_n : \mathrm{ddiag}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{S}) = \boldsymbol{0}_n\}$ et pour tout $(\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$, on a

$$\langle(\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}), (\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{\Lambda}\boldsymbol{S})\boldsymbol{S}, \boldsymbol{0}_n)\rangle_{\mathcal{B}}^{\mathcal{P}_n} = \mathrm{tr}(\boldsymbol{\xi}_{\boldsymbol{S}}\,\mathrm{sym}(\boldsymbol{\Lambda}\boldsymbol{S})) = \mathrm{tr}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{S}\boldsymbol{\Lambda}) = \mathrm{tr}(\mathrm{ddiag}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{S})\boldsymbol{\Lambda}) = 0.$$

Donc $P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}(\mathcal{Z}) = (\boldsymbol{Z}_{\boldsymbol{S}} - \boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{Z}_{\boldsymbol{U}})$ et pour que $P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}(\mathcal{Z})$ soit dans $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$, $\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}$ doit être solution de l'équation $\operatorname{ddiag}((\boldsymbol{Z}_{\boldsymbol{S}} - \boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S})\boldsymbol{S}) = \boldsymbol{0}_n$, qui peut être vectorisée par

$$\left(\boldsymbol{S}^3 \odot \boldsymbol{S} + \boldsymbol{S}^2 \odot \boldsymbol{S}^2\right) \operatorname{diag}(\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}) = 2\operatorname{diag}(\boldsymbol{Z}_{\boldsymbol{S}}\boldsymbol{S}).$$

Comme $\boldsymbol{S} \in \mathcal{S}_n^{++}$, toute puissance de $\boldsymbol{S}$ est positive définie et le théorème du produit de Schur assure que $\boldsymbol{S}^3 \odot \boldsymbol{S}$ et $\boldsymbol{S}^2 \odot \boldsymbol{S}^2$ sont positives définies. De plus, la somme de matrices positives définies est positive définie donc $(\boldsymbol{S}^3 \odot \boldsymbol{S} + \boldsymbol{S}^2 \odot \boldsymbol{S}^2)$ est inversible, ce qui complète la preuve.

Preuve de la proposition 3.5

Par définition, on a

$$\operatorname{hess}_{\mathrm{o},\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] = P_{\boldsymbol{B}}^{\mathrm{o},\ell}\left(\nabla_{\boldsymbol{\xi}}^{\ell} P_{\boldsymbol{B}}^{\mathrm{o},\ell}(\operatorname{grad}_{\ell} f(\boldsymbol{B}))\right) = P_{\boldsymbol{B}}^{\mathrm{o},\ell}\left(\operatorname{hess}_{\ell} f(\boldsymbol{B})[\boldsymbol{\xi}] - \nabla_{\boldsymbol{\xi}}^{\ell}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B}\right),$$

où $\boldsymbol{\Lambda}_{\mathrm{o},\ell}$ est définie dans la proposition 3.3 pour $\boldsymbol{Z} = \operatorname{grad}_{\ell} f(\boldsymbol{B})$. On peut montrer que

$$\nabla_{\boldsymbol{\xi}}^{\ell}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}^T \mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},\ell}[\boldsymbol{\xi}]\boldsymbol{B} + \operatorname{sym}(\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B} + \boldsymbol{B}\operatorname{sym}(\boldsymbol{B}^{-1}\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},\ell}\boldsymbol{B}).$$

Comme $\boldsymbol{B}\boldsymbol{B}^T \mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},\ell}[\boldsymbol{\xi}]\boldsymbol{B}$ est dans l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$, $P_{\boldsymbol{B}}^{\mathrm{o},\ell}(\boldsymbol{B}\boldsymbol{B}^T \mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},\ell}[\boldsymbol{\xi}]\boldsymbol{B}) = \boldsymbol{0}_n$. La formule de $\operatorname{hess}_{\mathrm{o},\ell} f(\boldsymbol{B})[\boldsymbol{\xi}]$ est obtenue en fusionnant les deux équations ci-dessus.

De la même façon, on obtient

$$\operatorname{hess}_{\mathrm{o},r} f(\boldsymbol{B})[\boldsymbol{\xi}] = P_{\boldsymbol{B}}^{\mathrm{o},r}\left(\operatorname{hess}_{r} f(\boldsymbol{B})[\boldsymbol{\xi}] - \nabla_{\boldsymbol{\xi}}^{r}\boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B}\right),$$

où $\boldsymbol{\Lambda}_{\mathrm{o},r}$ est définie dans la proposition 3.3 pour $\boldsymbol{Z} = \operatorname{grad}_{r} f(\boldsymbol{B})$. On peut montrer que

$$\begin{aligned}\nabla_{\boldsymbol{\xi}}^{r}\boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B} = \mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},r}[\boldsymbol{\xi}]\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Lambda}_{\mathrm{o},r}\left(\boldsymbol{\xi}\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{B}\operatorname{sym}(\boldsymbol{\xi}^T\boldsymbol{B})\right)\\ + \operatorname{sym}(\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},r}\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{B} - \boldsymbol{\xi}\boldsymbol{B}^{-1}\operatorname{sym}(\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{o},r})\boldsymbol{B}.\end{aligned}$$

La formule de $\operatorname{hess}_{\mathrm{o},r} f(\boldsymbol{B})[\boldsymbol{\xi}]$ est ici aussi obtenue en remarquant que $\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},r}[\boldsymbol{\xi}]\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{B}$ est dans l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{o}}$ et en fusionnant les deux équations ci-dessus.

Sur $\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$, on obtient

$$\operatorname{hess}_{\mathrm{o},\mathcal{P}_n} f(\mathcal{B})[\Xi] = P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}\left(\operatorname{hess}_{\mathcal{P}_n} f(\mathcal{B})[\Xi] - \nabla_{\Xi}^{\mathcal{P}_n}(\boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{0}_n)\right),$$

où $\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}$ est définie dans la proposition 3.3 pour $\mathcal{Z} = \operatorname{grad}_{\mathcal{P}_n} f(\mathcal{B})$. De plus, on a

$$\begin{aligned}\nabla_{\Xi}^{\mathcal{P}_n}(\boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{0}_n) = (\operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\operatorname{sym}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n})\boldsymbol{S}), \boldsymbol{0}_n)\\ + (\boldsymbol{S}\operatorname{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n} + \boldsymbol{S}\,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}[\Xi])\boldsymbol{S}, \boldsymbol{0}_n).\end{aligned}$$

$(\boldsymbol{S}\operatorname{sym}(\boldsymbol{S}\,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{o},\mathcal{P}_n}[\Xi])\boldsymbol{S}, \boldsymbol{0}_n)$ est dans l'espace normal de $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{o},\mathcal{P}_n}$ et sa projection par $P_{\mathcal{B}}^{\mathrm{o},\mathcal{P}_n}$ est donc l'élément nul. La formule de la hessienne est enfin obtenue en fusionnant les deux équations ci-dessus.

Contrainte intrinsèque

Preuve de la proposition 3.6

Étant donné $\boldsymbol{B} \in \mathcal{M}_n^{\mathrm{i}}$, l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{i}}$ lorsque GL_n est équipé de la métrique invariante à gauche est $\{\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{Q}^T : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. En effet, cet espace est de dimension n et pour tout $\boldsymbol{\xi} \in T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{i}}$

$$\langle \boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{Q}^T\rangle_{\boldsymbol{B}}^{\ell} = \mathrm{tr}(\boldsymbol{B}^{-1}\boldsymbol{\xi}(\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{Q}^T)^T) = \mathrm{tr}(\boldsymbol{\xi}\boldsymbol{Q}\boldsymbol{\Lambda}) = \mathrm{tr}(\mathrm{ddiag}(\boldsymbol{\xi}\boldsymbol{Q})\boldsymbol{\Lambda}) = 0.$$

De ce fait, $P_{\boldsymbol{B}}^{\mathrm{i},\ell}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{Q}^T$ et $\boldsymbol{\Lambda}_{\mathrm{i},\ell}$ est la matrice diagonale solution de l'équation $\mathrm{ddiag}(P_{\boldsymbol{B}}^{\mathrm{i},\ell}(\boldsymbol{Z})\boldsymbol{B}^T) = \boldsymbol{0}_n$, qui peut être vectorisée par

$$\left(\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{Q}^T\boldsymbol{Q}\right)\mathrm{diag}(\boldsymbol{\Lambda}_{\mathrm{i},\ell}) = \mathrm{diag}(\boldsymbol{Z}\boldsymbol{Q}).$$

Comme $\boldsymbol{B} \in \mathrm{GL}_n$ et $\boldsymbol{Q} \in \mathrm{GL}_n$, $\boldsymbol{B}\boldsymbol{B}^T$ et $\boldsymbol{Q}^T\boldsymbol{Q}$ sont positives définies et le théorème du produit de Schur assure que $\boldsymbol{B}\boldsymbol{B}^T \odot \boldsymbol{Q}^T\boldsymbol{Q}$ est inversible, ce qui complète la preuve pour $P^{\mathrm{i},\ell}$.

De la même façon, on peut montrer que l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}_n^{\mathrm{i}}$ quand GL_n est équipé de la métrique invariante à droite est $\{\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B} : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. Le projecteur orthogonal est donc de la forme $P_{\boldsymbol{B}}^{\mathrm{i},r}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}$. Résoudre $\mathrm{ddiag}(P_{\boldsymbol{B}}^{\mathrm{i},r}(\boldsymbol{Z})\boldsymbol{Q}) = \boldsymbol{0}_n$ donne $\boldsymbol{\Lambda}_{\mathrm{i},r} = \mathrm{ddiag}(\boldsymbol{Z}\boldsymbol{Q})\,\mathrm{ddiag}(\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Q})^{-1}$.

Étant donné $\mathcal{B} \in \mathcal{M}_n^{\mathrm{i},\mathcal{P}_n}$, l'espace normal à $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{i},\mathcal{P}_n}$ dans $T_{\mathcal{B}}\mathcal{P}_n$ est donné par $\{(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda})\boldsymbol{S}, P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T)) : \boldsymbol{\Lambda} \in \mathcal{D}_n\}$. En effet, pour tout $\Xi = (\boldsymbol{\xi}_{\boldsymbol{S}}, \boldsymbol{\xi}_{\boldsymbol{U}}) \in T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{i},\mathcal{P}_n}$,

$$\begin{aligned}\left\langle \Xi, \left(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda})\boldsymbol{S}, P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T)\right)\right\rangle_{\mathcal{B}}^{\mathcal{P}_n} &= \mathrm{tr}(\boldsymbol{\xi}_{\boldsymbol{S}}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda})) \\ &\quad + \mathrm{tr}\left(\boldsymbol{\xi}_{\boldsymbol{U}}^T(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T - \boldsymbol{U}\,\mathrm{sym}(\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T))\right).\end{aligned}$$

En remarquant que $\boldsymbol{U}\,\mathrm{sym}(\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T)$ est dans l'espace normal à $T_{\boldsymbol{U}}\mathcal{O}_n$, cette expression devient

$$\begin{aligned}\left\langle \Xi, \left(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda})\boldsymbol{S}, P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T)\right)\right\rangle_{\mathcal{B}}^{\mathcal{P}_n} &= \mathrm{tr}(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}) + \mathrm{tr}(\boldsymbol{\xi}_{\boldsymbol{U}}^T\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{Q}^T) \\ &= \mathrm{tr}(\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]\boldsymbol{Q}\Lambda) \\ &= \mathrm{tr}(\mathrm{ddiag}(\mathrm{D}\,\Gamma(\mathcal{B})[\Xi]\boldsymbol{Q})\Lambda) = 0.\end{aligned}$$

Il suit que $P_{\mathcal{B}}^{\mathrm{i},\mathcal{P}_n}(\mathcal{Z}) = (\boldsymbol{Z}_{\boldsymbol{S}} - \boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}, \boldsymbol{Z}_{\boldsymbol{U}} - P_{\boldsymbol{U}}^{\mathcal{O}_n}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T))$ et $\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}$ est obtenu en vectorisant l'équation $\mathrm{ddiag}(\mathrm{D}\,\Gamma(\mathcal{B})[P_{\mathcal{B}}^{\mathrm{i},\mathcal{P}_n}(\mathcal{Z})]\boldsymbol{Q}) = 0$.

Preuve de la proposition 3.8

Lorsque GL_n est équipée de la métrique invariante à gauche, on obtient

$$\mathrm{hess}_{\mathrm{i},\ell}\, f(\boldsymbol{B})[\boldsymbol{\xi}] = P_{\boldsymbol{B}}^{\mathrm{i},\ell}\left(\mathrm{hess}_{\ell}\, f(\boldsymbol{B})[\boldsymbol{\xi}] - \nabla_{\boldsymbol{\xi}}^{\ell}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{Q}^T\right),$$

et on peut montrer que

$$\nabla^{\ell}_{\boldsymbol{\xi}} \boldsymbol{B}\boldsymbol{B}^T \boldsymbol{\Lambda}_{\mathrm{i},\ell} \boldsymbol{Q}^T = \boldsymbol{B}\boldsymbol{B}^T \,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},\ell}[\boldsymbol{\xi}]\boldsymbol{Q}^T + \mathrm{sym}(\boldsymbol{\xi}\boldsymbol{B}^T)\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{Q}^T + \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\Lambda}_{\mathrm{i},\ell}\dot{\boldsymbol{Q}}^T \\ + \boldsymbol{B}\,\mathrm{sym}(\boldsymbol{B}^{-1}\boldsymbol{\xi}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{B}) - \boldsymbol{B}\,\mathrm{sym}(\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\ell}\boldsymbol{B})\boldsymbol{B}^{-1}\boldsymbol{\xi}.$$

Fusionner les deux équations ci-dessus et remarquer que $P^{\mathrm{i},\ell}_{\boldsymbol{B}}(\boldsymbol{B}\boldsymbol{B}^T \,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},\ell}[\boldsymbol{\xi}]\boldsymbol{Q}^T) = \boldsymbol{0}_n$, car $\boldsymbol{B}\boldsymbol{B}^T \,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},\ell}[\boldsymbol{\xi}]\boldsymbol{Q}^T$ est dans l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}^{\mathrm{i}}_n$, donne le résultat.

Quand GL_n est équipée de la métrique invariante à droite, on a

$$\mathrm{hess}_{\mathrm{i},r}\, f(\boldsymbol{B})[\boldsymbol{\xi}] = P^{\mathrm{i},r}_{\boldsymbol{B}}\left(\mathrm{hess}_r\, f(\boldsymbol{B})[\boldsymbol{\xi}] - \nabla^r_{\boldsymbol{\xi}}\boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}\right).$$

Dans ce cas, on peut montrer que

$$\nabla^r_{\boldsymbol{\xi}}\boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{Q}\boldsymbol{B}^T\boldsymbol{B} = \mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},r}[\boldsymbol{\xi}]\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Lambda}_{\mathrm{i},r}(\dot{\boldsymbol{Q}}^T\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{Q}^T\,\mathrm{sym}(\boldsymbol{\xi}^T\boldsymbol{B})) \\ + \mathrm{sym}(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},r}\boldsymbol{\xi}\boldsymbol{B}^{-1})\boldsymbol{B} - \boldsymbol{\xi}\boldsymbol{B}^{-1}\,\mathrm{sym}(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},r})\boldsymbol{B}.$$

Le résultat est obtenu en fusionnant les deux équations ci-dessus et en remarquant que $\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},r}[\boldsymbol{\xi}]\boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{B}$ est dans l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}^{\mathrm{i}}_n$.

Sur $\mathcal{M}^{\mathrm{i},\mathcal{P}_n}_n$, on obtient

$$\mathrm{hess}_{\mathrm{i},\mathcal{P}_n}\, f(\mathcal{B})[\Xi] = P^{\mathrm{i},\mathcal{P}_n}_{\mathcal{B}}\left(\mathrm{hess}_{\mathcal{P}_n}\, f(\mathcal{B})[\Xi] - \nabla^{\mathcal{P}_n}_{\Xi}(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}, P^{\mathcal{O}_n}_{\boldsymbol{U}}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T))\right),$$

et on peut montrer que

$$\nabla^{\mathcal{P}_n}_{\Xi}(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}, P^{\mathcal{O}_n}_{\boldsymbol{U}}(\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T)) = \\ \left(\boldsymbol{S}\,\mathrm{sym}\left((\boldsymbol{\xi}_{\boldsymbol{U}}\boldsymbol{Q} + \boldsymbol{U}\dot{\boldsymbol{Q}})\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\right)\boldsymbol{S} + \mathrm{sym}(\boldsymbol{\xi}_{\boldsymbol{S}}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n})\boldsymbol{S}), \boldsymbol{0}_n\right) \\ + \left(\boldsymbol{0}_n, P^{\mathcal{O}_n}_{\boldsymbol{U}}\left(\boldsymbol{\xi}_{\boldsymbol{S}}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T + \boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\dot{\boldsymbol{Q}}^T - \boldsymbol{\xi}_{\boldsymbol{U}}\,\mathrm{sym}(\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}\boldsymbol{Q}^T)\right)\right) \\ + \left(\boldsymbol{S}\,\mathrm{sym}(\boldsymbol{U}\boldsymbol{Q}\,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}[\Xi])\boldsymbol{S}, P^{\mathcal{O}_n}_{\boldsymbol{U}}\left(\boldsymbol{S}\,\mathrm{D}\,\boldsymbol{\Lambda}_{\mathrm{i},\mathcal{P}_n}[\Xi]\boldsymbol{Q}^T\right)\right).$$

La formule de $\mathrm{hess}_{\mathrm{i},\mathcal{P}_n}\, f(\mathcal{B})[\Xi]$ est obtenue en remarquant que le dernier terme est dans l'espace normal à $T_{\boldsymbol{B}}\mathcal{M}^{\mathrm{i}}_n$.

Bibliographie

[1] P.-A. Absil et K. A. Gallivan. "Joint Diagonalization on the Oblique Manifold for Independent Component Analysis". In : *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. T. 5. 2006, p. 945-948 (cf. p. 12, 13).

[2] P.-A. Absil, R. Mahony et R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA : Princeton University Press, 2008 (cf. p. 2, 13-15, 24, 47, 57, 62, 80, 81).

[3] P.-A. Absil et J. Malick. "Projection-like retractions on matrix manifolds". In : *SIAM Journal on Optimization* 22.1 (2012), p. 135-158 (cf. p. 52, 54).

[4] B. Afsari. "Sensitivity analysis for the problem of matrix joint diagonalization". In : *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), p. 1148-1171 (cf. p. 11, 13, 40, 45, 80).

[5] B. Afsari et P. S. Krishnaprasad. "Some gradient based joint diagonalization methods for ICA". In : *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, p. 437-444 (cf. p. 13).

[6] A. H. Al-Mohy. "A more accurate Briggs method for the logarithm". In : *Numerical Algorithms* 59.3 (2012), p. 393-402 (cf. p. 95, 102, 103).

[7] A. H. Al-Mohy et N. J. Higham. "Improved inverse scaling and squaring algorithms for the matrix logarithm". In : *SIAM Journal on Scientific Computing* 34.4 (2012), p. 153-169 (cf. p. 95, 100, 102, 103).

[8] A. H. Al-Mohy, N. J. Higham et S. D. Relton. "Computing the Fréchet derivative of the matrix logarithm and estimating the condition number". In : *SIAM Journal on Scientific Computing* 35.4 (2013), p. 394-410 (cf. p. 84, 91, 95, 96, 99-101, 103).

[9] K. Alyani, M. Congedo et M. Moakher. "Diagonality Measures of Hermitian Positive-Definite Matrices with Application to the Approximate Joint Diagonalization Problem". In : *Linear Algebra and its Applications* (2016) (cf. p. 14, 27, 35-37, 40).

[10] S.-I. Amari. *Differential-geometrical methods in statistics*. Springer, Heidelberg, 1985 (cf. p. 37).

[11] S.-I. Amari. "Natural gradient works efficiently in learning". In : *Neural computation* 10.2 (1998), p. 251-276 (cf. p. 13, 45, 80).

[12] S.-I. Amari, T. Chen et A. Cichocki. "Nonholonomic orthogonal learning algorithms for blind source separation". In : *Neural computation* 12.6 (2000), p. 1463-1484 (cf. p. 13, 45, 80).

[13] E. Andruchow et al. "The left invariant metric in the general linear group". In : *Journal of Geometry and Physics* 86 (2014), p. 241-257 (cf. p. 25).

[14] M. Arns et al. "Neurofeedback en psychiatrie : une technique du présent ?" In : *L'Encéphale* 43.2 (2017), p. 135-145 (cf. p. 6).

[15] V. Arsigny et al. “Geometric means in a novel vector space structure on symmetric positive-definite matrices”. In : *SIAM journal on matrix analysis and applications* 29.1 (2007), p. 328-347 (cf. p. 14, 37).

[16] A. Belouchrani et al. “A blind source separation technique using second-order statistics”. In : *IEEE Transactions on signal processing* 45.2 (1997), p. 434-444 (cf. p. 10).

[17] H. Berger. “Über das elektrenkephalogramm des menschen”. In : *Archiv für psychiatrie und nervenkrankheiten* 87.1 (1929), p. 527-570 (cf. p. 1, 6).

[18] R. Bhatia. *Matrix analysis*. T. 169. Springer Science & Business Media, 2013 (cf. p. 96).

[19] R. Bhatia. “Matrix factorizations and their perturbations”. In : *Linear Algebra and its applications* 197 (1994), p. 245-276 (cf. p. 81).

[20] R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009 (cf. p. 14, 22, 23, 36, 81).

[21] R. Bhatia, T. Jain et Y. Lim. “On the Bures-Wasserstein distance between positive definite matrices”. In : *preprint* (2017) (cf. p. 14, 38, 81).

[22] Å. Björck et S. Hammarling. “A Schur method for the square root of a matrix”. In : *Linear algebra and its applications* 52 (1983), p. 127-140 (cf. p. 102).

[23] F. Bouchard, J. Malick et M. Congedo. “Approximate joint diagonalization according to the natural Riemannian distance”. In : *International Conference on Latent Variable Analysis and Signal Separation*. Springer. 2017, p. 290-299 (cf. p. 3).

[24] F. Bouchard, J. Malick et M. Congedo. “Riemannian Optimization and Approximate Joint Diagonalization for Blind Source Separation”. In : *IEEE Transactions on Signal Processing* 66.8 (2018), p. 2041-2054 (cf. p. 3).

[25] F. Bouchard et al. “Approximate Joint Diagonalization with Riemannian Optimization on the General Linear Group”. In : *à soumettre à SIAM Journal of Matrix Analysis and Applications* () (cf. p. 3).

[26] F. Bouchard et al. “Approximate joint diagonalization within the Riemannian geometry framework”. In : *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE. 2016, p. 210-214 (cf. p. 3).

[27] F. Bouchard et al. “Réduction de dimension pour la Séparation Aveugle de Sources”. In : *GRETSI 2017*. 2017 (cf. p. 3).

[28] N. Boumal et al. “Manopt, a Matlab Toolbox for Optimization on Manifolds”. In : *Journal of Machine Learning Research* 15 (2014), p. 1455-1459 (cf. p. 62).

[29] A. Breloy et al. “Borne de Cramér-Rao intrinsèque pour la matrice de covariance des distributions elliptiques complexes”. In : *GRETSI 2017*. 2017 (cf. p. 3).

[30] A. Breloy et al. “Intrinsic Cramér-Rao bounds for scatter and shape matrices estimation in CES distributions”. In : *IEEE Signal Processing Letters* 26.2 (2019), p. 262-266 (cf. p. 3).

[31] G. Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006 (cf. p. 7, 10, 76).

[32] J.-F. Cardoso. “Blind signal separation : statistical principles”. In : *Proceedings of the IEEE* 86.10 (1998), p. 2009-2025 (cf. p. 10).

[33] J.-F. Cardoso et A. Souloumiac. “Blind beamforming for non Gaussian signals”. In : *IEEE Proceedings-F* 140.6 (1993), p. 362-370 (cf. p. 1, 2, 10, 12, 13).

[34] J.-F. Cardoso et A. Souloumiac. “Jacobi angles for simultaneous diagonalization”. In : *SIAM journal on matrix analysis and applications* 17.1 (1996), p. 161-164 (cf. p. 2, 12, 13).

[35] J. R. Cardoso et F. Silva Leite. “Theoretical and numerical considerations about Padé approximants for the matrix logarithm”. In : *Linear Algebra and its Applications* 330.1 (2001), p. 31-42 (cf. p. 95).

[36] Z. Chebbi et M. Moakher. “Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function”. In : *Linear Algebra and its Applications* 436.7 (2012), p. 1872-1889 (cf. p. 14, 36).

[37] S. H. Cheng et al. “Approximating the logarithm of a matrix to specified accuracy”. In : *SIAM Journal on Matrix Analysis and Applications* 22.4 (2001), p. 1112-1125 (cf. p. 95).

[38] P. Comon. “Independent component analysis, a new concept ?” In : *Signal processing* 36.3 (1994), p. 287-314 (cf. p. 8, 10).

[39] P. Comon et C. Jutten. *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. 1st. Academic Press, 2010 (cf. p. 1, 8, 10, 35).

[40] M. Congedo. *EEG Source Analysis*. CNRS, University of Grenoble Alpes, Grenoble Institute of Technology, 2013 (cf. p. 1, 7, 8, 81).

[41] M. Congedo, C. Gouy-Pailler et C. Jutten. “On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics”. In : *Clinical Neurophysiology* 119.12 (2008), p. 2677-2686 (cf. p. 1, 8, 10, 11, 39, 71).

[42] M. Congedo et D.-T. Pham. “Least-squares joint diagonalization of a matrix set by a congruence transformation”. In : *SinFra’09 - 2nd Singaporean-French IPAL Symposium*. 2009, p. 96-106 (cf. p. 13, 61).

[43] M. Congedo, R. Phlypo et D. T. Pham. “Approximate Joint Singular Value Decomposition of an Asymmetric Rectangular Matrix Set”. In : *IEEE Transactions on Signal Processing* 59.1 (2011), p. 415-424 (cf. p. 81).

[44] M. Congedo et al. “A closed-form unsupervised geometry-aware dimensionality reduction method in the Riemannian manifold of SPD matrices”. In : *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE. 2017, p. 3198-3201 (cf. p. 3).

[45] M. Congedo et al. “Approximate Joint Diagonalization and Geometric Mean of Symmetric Positive Definite Matrices”. In : *PLoS ONE* 10.4 (avr. 2015), e0121423 (cf. p. 14, 28).

[46] M. Congedo et al. “"Brain Invaders" : a prototype of an open-source P300-based video game working with the OpenViBE platform”. In : *5th International Brain-Computer Interface Conference 2011 (BCI 2011)*. 2011, p. 280-283 (cf. p. 6).

[47] L. Dieci, B. Morini et A. Papini. “Computational techniques for real logarithms of matrices”. In : *SIAM Journal on Matrix Analysis and Applications* 17.3 (1996), p. 570-593 (cf. p. 84, 91, 95).

[48] D. C. Dowson et B. V. Landau. “The Fréchet distance between multivariate normal distributions”. In : *Journal of multivariate analysis* 12.3 (1982), p. 450-455 (cf. p. 38).

[49] I. Feinberg, R. L. Koresko et N. Heller. “EEG sleep patterns as a function of normal and pathological aging in man”. In : *Journal of psychiatric research* 5.2 (1967), p. 107-144 (cf. p. 6).

[50] P. Fillard et al. “A Riemannian framework for the processing of tensor-valued images”. In : *Deep Structure, Singularities, and Computer Vision*. Springer, 2005, p. 112-123 (cf. p. 14, 36, 37).

[51] B. N. Flury et W. Gautschi. “An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form”. In : *SIAM Journal on Scientific and Statistical Computing* 7.1 (1986), p. 169-184 (cf. p. 1, 12, 13).

[52] S. Gallot, D. Hulin et J. Lafontaine. *Riemannian geometry*. 3rd. Springer, 2004 (cf. p. 14, 57).

[53] A. Goh et R. Vidal. “Unsupervised Riemannian clustering of probability density functions”. In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2008, p. 377-392 (cf. p. 23, 36).

[54] G. H. Golub et C. F. Van Loan. *Matrix computations*. 3e éd. John Hopkins Unversity Press, 1996 (cf. p. 63, 101).

[55] X. Guo et al. “Approximate joint diagonalization by nonorthogonal nonparametric Jacobi transformations”. In : *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE. 2010, p. 3774-3777 (cf. p. 13).

[56] N. J. Higham. “Computing real square roots of a real matrix”. In : *Linear Algebra and its applications* 88 (1987), p. 405-430 (cf. p. 103).

[57] N. J. Higham. “Evaluating Padé approximants of the matrix logarithm”. In : *SIAM Journal on Matrix Analysis and Applications* 22.4 (2001), p. 1126-1135 (cf. p. 95, 99).

[58] N. J. Higham. *Functions of matrices : theory and computation*. SIAM, 2008 (cf. p. 95, 96, 99, 102, 103).

[59] N. J. Higham et L. Lin. “A Schur-Padé algorithm for fractional powers of a matrix”. In : *SIAM Journal on Matrix Analysis and Applications* 32.3 (2011), p. 1056-1078 (cf. p. 102, 103).

[60] N. J. Higham et S. D. Relton. “Higher order Fréchet derivatives of matrix functions and the level-2 condition number”. In : *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), p. 1019-1037 (cf. p. 95, 96).

[61] W. Huang, P.-A. Absil et K.A. Gallivan. "A Riemannian BFGS Method for Nonconvex Optimization Problems". In : *Numerical Mathematics and Advanced Applications ENUMATH 2015*. Springer International Publishing, 2016, p. 627-634 (cf. p. 13, 62, 80).

[62] H. H. Jasper. "Report of the Committee on Methods of Clinical Examination in Electroencephalography". In : *Electroencephalography and Clinical Neurophysiology* 10 (1958), p. 370-375 (cf. p. 6).

[63] B. Jeuris, R. Vandebril et B. Vandereycken. "A survey and comparison of contemporary algorithms for computing the matrix geometric mean". In : *Electronic Transactions on Numerical Analysis* 39 (2012), p. 379-402 (cf. p. 23).

[64] X. Jiang, Z.-Q. Luo et T. T. Georgiou. "Geometric methods for spectral analysis". In : *IEEE Transactions on Signal Processing* 60.3 (2012), p. 1064-1074 (cf. p. 38).

[65] M. Joho. "Newton method for joint approximate diagonalization of positive definite Hermitian matrices". In : *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), p. 1205-1218 (cf. p. 13).

[66] M. Joho et H. Mathis. "Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation". In : *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*. 2002, p. 273-277 (cf. p. 13).

[67] M. Joho et K. Rahbar. "Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation". In : *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*. 2002, p. 403-407 (cf. p. 13).

[68] C. Jutten et J. Herault. "Blind separation of sources, part I : An adaptive algorithm based on neuromimetic architecture". In : *Signal processing* 24.1 (1991), p. 1-10 (cf. p. 8).

[69] E. K. Kalunga et al. "Online SSVEP-based BCI using Riemannian geometry". In : *Neurocomputing* 191 (2016), p. 55-68 (cf. p. 70).

[70] C. S. Kenney et A. J. Laub. "A Schur-Fréchet algorithm for computing the logarithm and exponential of a matrix". In : *SIAM journal on matrix analysis and applications* 19.3 (1998), p. 640-663 (cf. p. 95).

[71] C. S. Kenney et A. J. Laub. "Condition estimates for matrix functions". In : *SIAM Journal on Matrix Analysis and Applications* 10.2 (1989), p. 191-209 (cf. p. 95).

[72] T. Kim, T. Eltoft et T. Lee. "Independent vector analysis : An extension of ICA to multivariate components". In : *International Conference on Independent Component Analysis and Signal Separation*. Springer. 2006, p. 165-172 (cf. p. 81).

[73] L. Korczowski et al. "Mining the bilinear structure of data with approximate joint diagonalization". In : *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE. 2016, p. 667-671 (cf. p. 3).

[74] D. Lahat et C. Jutten. "Joint independent subspace analysis using second-order statistics". In : *IEEE Transactions on Signal Processing* 64.18 (2016), p. 4891-4904 (cf. p. 81).

[75] J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003 (cf. p. 14, 55).

[76] S. Lee et al. "Geometric direct search algorithms for image registration". In : *IEEE Transactions on Image Processing* 16.9 (2007), p. 2215-2224 (cf. p. 25).

[77] V. Maurandi et E. Moreau. "A decoupled Jacobi-like algorithm for non-unitary joint diagonalization of complex-valued matrices". In : *IEEE Signal Processing Letters* 21.12 (2014), p. 1453-1456 (cf. p. 13).

[78] L. Mayaud et al. "Brain-computer interface for the communication of acute patients : a feasibility study and a randomized controlled trial comparing performance with healthy participants and a traditional assistive device". In : *Brain-Computer Interfaces* 3.4 (2016), p. 197-215 (cf. p. 6).

[79] G. Meyer. "Geometric optimization algorithms for linear regression on fixed-rank matrices". Thèse de doct. University of Liège, 2011 (cf. p. 23).

[80] M. I. Miller, A. Trouvé et L. Younes. "The metric spaces, Euler equations, and normal geodesic image motions of computational anatomy". In : *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. T. 2. IEEE. 2003, p. 635-638 (cf. p. 25).

[81] M. Moakher. "A differential geometric approach to the geometric mean of symmetric positive-definite matrices". In : *SIAM Journal on Matrix Analysis and Applications* 26.3 (2005), p. 735-747 (cf. p. 14, 36, 90).

[82] M. Moakher. "Divergence measures and means of symmetric positive-definite matrices". In : *New Developments in the Visualization and Processing of Tensor Fields*. Springer, 2012, p. 307-321 (cf. p. 35, 36).

[83] E. Moreau. "A generalization of joint-diagonalization criteria for source separation". In : *IEEE Transactions on Signal Processing* 49.3 (2001), p. 530-541 (cf. p. 10).

[84] E. Moreau et O. Macchi. "A one stage self-adaptive algorithm for source separation". In : *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*. T. 3. 1994, p. 49-52 (cf. p. 63).

[85] F. Mormann et al. "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients". In : *Physica D : Nonlinear Phenomena* 144.3-4 (2000), p. 358-369 (cf. p. 6).

[86] E. Niedermeyer et F. H. Lopes da Silva. *Electroencephalography : basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005 (cf. p. 6).

[87] L. Ning, X. Jiang et T. T. Georgiou. "Geometric methods for estimation of structured covariances". In : *arXiv preprint arXiv :1110.3695* (2011) (cf. p. 38).

[88] P. L. Nunez et R. Srinivasan. *Electric fields of the brain : the neurophysics of EEG*. Oxford University Press, USA, 2006 (cf. p. 6).

[89] I. Olkin et F. Pukelsheim. "The distance between two random vectors with given dispersion matrices". In : *Linear Algebra and its Applications* 48 (1982), p. 257-263 (cf. p. 38).

[90] R. D. Pascual-Marqui. “Review of methods for solving the EEG inverse problem”. In : *International journal of bioelectromagnetism* 1.1 (1999), p. 75-86 (cf. p. 7).

[91] D. T. Pham. “Blind separation of instantaneous mixture of sources via an independent component analysis”. In : *IEEE Transactions on Signal Processing* 44.11 (1996), p. 2768-2779 (cf. p. 10).

[92] D. T. Pham. “Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion”. In : *Signal Processing* 81.4 (2001), p. 855-870 (cf. p. 10).

[93] D.-T. Pham. “Joint Approximate Diagonalization of Positive Definite Hermitian Matrices”. In : *SIAM J. Matrix Anal. Appl.* 22.4 (2000), p. 1136-1152 (cf. p. 10, 12-14, 61).

[94] D. T. Pham. “Mutual information approach to blind separation of stationary sources”. In : *IEEE Transactions on Information Theory* 48.7 (2002), p. 1935-1946 (cf. p. 10).

[95] D. T. Pham et J.-F. Cardoso. “Blind separation of instantaneous mixtures of nonstationary sources”. In : *IEEE Transactions on Signal Processing* 49.9 (2001), p. 1837-1848 (cf. p. 2, 10, 35).

[96] D.-T. Pham et M. Congedo. “Least square joint diagonalization of matrices under an intrinsic scale constraint”. In : *Independent Component Analysis and Signal Separation*. Springer, 2009, p. 298-305 (cf. p. 13).

[97] D. T. Pham et P. Garat. “Blind separation of mixture of independent sources through a quasi-maximum likelihood approach”. In : *IEEE transactions on Signal Processing* 45.7 (1997), p. 1712-1725 (cf. p. 10).

[98] K. Rahbar et J. P. Reilly. “Geometric Optimization Methods for Blind Source Separation of Signals”. In : *in Proc. ICA*. 2000, p. 375-380 (cf. p. 13).

[99] C. R. Rao. “Information and accuracy attainable in the estimation of statistical parameters”. In : *Bull. Calcutta Math. Soc* 37.3 (1945), p. 81-91 (cf. p. 37).

[100] P. Rodrigues et al. “Dimensionality Reduction for BCI classification using Riemannian geometry”. In : *7th Graz Brain-Computer Interface Conference*. 2017 (cf. p. 3).

[101] P. Rodrigues et al. “Géométrie Riemannienne appliquée à la réduction de la dimension de signaux EEG pour les interfaces cerveau-machine”. In : *GRETSI*. 2017 (cf. p. 3).

[102] L. T. Skovgaard. “A Riemannian geometry of the multivariate normal model”. In : *Scandinavian Journal of Statistics* (1984), p. 211-223 (cf. p. 23, 36).

[103] A. Souloumiac. “Nonorthogonal joint diagonalization by combining Givens and hyperbolic rotations”. In : *Signal Processing, IEEE Transactions on* 57.6 (2009), p. 2222-2231 (cf. p. 13).

[104] S. Sra. “Positive definite matrices and the S-divergence”. In : *arXiv preprint arXiv :1110.1773* (2013) (cf. p. 14, 36).

[105] A. Takatsu. “On Wasserstein geometry of the space of Gaussian measures”. In : *arXiv preprint arXiv :0801.2250* (2009) (cf. p. 38).

[106] F. J. Theis, T. P. Cason et P.-A. Absil. "Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold". In : *Independent Component Analysis and Signal Separation*. Springer, 2009, p. 354-361 (cf. p. 13).

[107] P. Tichavskỳ et A. Yeredor. "Fast approximate joint diagonalization incorporating weight matrices". In : *Signal Processing, IEEE Transactions on* 57.3 (2009), p. 878-891 (cf. p. 12-14, 28, 32, 34, 61, 80).

[108] K. Todros et J. Tabrikian. "Fast approximate joint diagonalization of positive definite Hermitian matrices". In : *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. T. 3. IEEE. 2007, p. 1373-1376 (cf. p. 13).

[109] B. Vandereycken, P.-A. Absil et S. Vandewalle. "A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank". In : *IMA Journal of Numerical Analysis* 33.2 (2012), p. 481-514 (cf. p. 25).

[110] F.-B. Vialatte et al. "Steady-state visually evoked potentials : focus on essential paradigms and future perspectives". In : *Progress in Neurobiology* 90.4 (2010), p. 418-438 (cf. p. 70).

[111] C. Villani. *Optimal transport : old and new*. T. 338. Springer Science & Business Media, 2008 (cf. p. 38).

[112] P. H. Westfall et Young S.S. *Resampling-based multiple testing : examples and methods for p-value adjustment*. Wiley, New York, 1993 (cf. p. 63).

[113] A. Yeredor. "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation". In : *IEEE Transactions on Signal Processing* 50.7 (2002), p. 1545-1553 (cf. p. 12).

[114] A. Yeredor, A. Ziehe et K.-R. Müller. "Approximate joint diagonalization using a natural gradient approach". In : *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, p. 89-96 (cf. p. 13).

[115] E. Zacur, M. Bossa et S. Olmos. "Left-invariant Riemannian geodesics on spatial transformation groups". In : *SIAM Journal on Imaging Sciences* 7.3 (2014), p. 1503-1557 (cf. p. 25).

[116] A. Ziehe et al. "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation". In : *The Journal of Machine Learning Research* 5 (2004), p. 777-800 (cf. p. 13).

Résumé — La *diagonalisation conjointe approximée* d'un ensemble de matrices permet de résoudre le problème de *séparation aveugle de sources* et trouve de nombreuses applications, notamment pour l'*électroencéphalographie*, une technique de mesure de l'activité cérébrale. La diagonalisation conjointe se formule comme un problème d'optimisation avec trois composantes : le choix du critère à minimiser, la contrainte de non-dégénérescence de la solution et l'algorithme de résolution. Les approches existantes considèrent principalement deux critères, les moindres carrés et la log-vraissemblance. Elles sont spécifiques à une contrainte et se restreignent à un seul type d'algorithme de résolution. Dans ce travail de thèse, nous proposons de formuler le problème de diagonalisation conjointe selon un *modèle géométrique*, qui généralise les travaux précédents et permet de définir des critères inédits, notamment liés à la théorie de l'information. Nous proposons également d'exploiter l'*optimisation riemannienne* et nous définissons un ensemble d'outils qui permet de faire varier les trois composantes indépendamment, créant ainsi de nouvelles méthodes et révélant l'influence des choix de modélisation. Des expériences numériques sur des données simulées et sur des enregistrements électroencéphalographiques montrent que notre approche par optimisation riemannienne donne des résultats compétitifs par rapport aux méthodes existantes. Elles indiquent aussi que les deux critères traditionnels ne sont pas les meilleurs dans toutes les situations.

Mots clés : géométrie riemannienne, optimisation riemannienne, diagonalisation conjointe approximée, séparation aveugle de sources, électroencéphalographie.

Abstract — The *approximate joint diagonalisation* of a set of matrices allows the solution of the *blind source separation* problem and finds several applications, for instance in *electroencephalography*, a technique for measuring brain activity. The approximate joint diagonalisation is formulated as an optimization problem with three components : the choice of the criterion to be minimized, the non-degeneracy constraint on the solution and the solving algorithm. Existing approaches mainly consider two criteria, the least-squares and the log-likelihood. They are specific to a constraint and are limited to only one type of solving algorithms. In this thesis, we propose to formulate the approximate joint diagonalisation problem in a *geometrical fashion*, which generalizes previous works and allows the definition of new criteria, particularly those linked to information theory. We also propose to exploit *Riemannian optimisation* and we define tools that allow to have the three components varying independently, creating in this way new methods and revealing the influence of the choice of the model. Numerical experiments on simulated data as well as on electroencephalographic recordings show that our approach by means of Riemannian optimisation gives results that are competitive as compared to existing methods. They also indicate that the two traditional criteria do not perform best in all situations.

Keywords : Riemannian geometry, Riemannian optimization, approximate joint diagonalization, blind source separation, electroencephalography.