Amélioration du suivi des patients atteints de maladies neuro-dégénératives à l'aide d'objets connectés

Pierre Drouin

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE NANTES

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs interactions*

Par

Pierre DROUIN

Amélioration du suivi des patients atteints de maladies neuro-dégénératives à l'aide d'objets connectés

Thèse présentée et soutenue à Nantes Université - UFR Sciences et Techniques, le 27 Septembre 2022
Unité de recherche : Laboratoire de Mathématiques Jean Leray Équie ALEA, - UmanIT, Pôle R&D

Rapporteurs avant soutenance :

Rosanna VERDE — Full Professor of statistics, University of Campania Luigi Vanvitelli, Caserta, Italy
Basile CHAIX — Directeur de recherche, Sorbonne Université, Paris, France

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Julien CHIQUET	Senior reasearcher, Université Paris-Saclay, France
Examinateurs :	Rosanna VERDE	Full Professor of statistics, University of Campania Luigi Vanvitelli, Caserta, Italy
	Basile CHAIX	Directeur de recherche, Sorbonne Université, Paris, France
Dir. de thèse :	Lise BELLANGER	Maître de conférence, Laboratoire de Mathématiques Jean Leray, Nantes Université, France
Co-dir. de thèse :	Laurent CHEVREUIL	Directeur associé de l'entreprise UmanIT, France
Co-encadrant. de thèse :	Aymeric STAMM	Ingénieur de recherche CNRS, Laboratoire de Mathématiques Jean Leray, Nantes Université, France

Invité(s) :

David-Axel LAPLAUD — Professeur des Universités - Praticien Hospitalier, Centre Hospitalier Universitaire de Nantes, France
Simone VANTINI — Associate professor, MOX Politecnico di Milano, Italy

Remerciements

Cette thèse a été réalisée dans le cadre du dispositif "Conventions Industrielles de Formation par la Recherche" (CIFRE) proposée par l'Association Nationale Recherche et Technologie (ANRT), qui a co-financé les travaux de recherche avec l'entreprise UmanIT. Des financements supplémentaires ont été obtenus par le biais de réponses à des appels à projet proposés par l'Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société (AMIES) et la fondation pour l'Aide à la Recherche dans la Sclérose En Plaques (ARSEP).

Au cours de ces quatre années, j'ai bénéficié d'un encadrement soutenu de la part de ma directrice de thèse Lise Bellanger et de mon co-encadrant Aymeric Stamm. j'ai beaucoup appris à leur côté, et leur forte implication dans ce projet, leurs conseils et leurs expériences dans le domaine de la recherche en statistique appliquée m'ont apporté une aide précieuse dans l'accomplissement de ces travaux. Ayant réalisé ma thèse dans le cadre d'une convention CIFRE, j'ai également découvert le monde de la recherche et du développement dans le secteur privé en tant que doctorant-salarié au sein de l'entreprise UmanIT. Sous la supervision de mon co-directeur Laurent Chevreuil, j'ai disposé d'une grande autonomie et de toutes les ressources nécessaires pour mener à bien les tâches qui m'ont été confiées. L'organisation de mon temps de travail par l'entreprise a notamment su prendre en compte les contraintes inhérentes à la rédaction d'une thèse et l'obtention d'un doctorat.

En tant que membres du Comité de Suivi Individuel, les professeurs el-Mostafa Qannari, Pierre-Antoine Gourraud et David Causeur ont apporté leur regard extérieur et veillé à la bonne progression de mon projet de recherche.

Ma thèse a été examinée consciencieusement par les professeurs Basile Chaix et Rosanna Verde, qui ont accepté de jouer le rôle de rapporteur. Les remarques pertinentes qu'ils ont formulées à la suite de leur relecture m'ont permis de clarifier certains passages de ce manuscrit et de préparer ma soutenance. Le jury de cette dernière est constitué de mes encadrants, des rapporteurs, du professeur Julien Chiquet en tant que président, ainsi que du professeur David-Axel Laplaud et du Docteur Simone Vantini en tant que membres invités.

Les professeurs David-Axel Laplaud et Pierre-Antoine Gourraud ont cru au potentiel du projet e-Gait à un stade précoce de sa conception. Ils nous ont ainsi permis, l'équipe de développement et moi-même, d'intégrer le dispositif à l'étude clinique qu'ils menaient au Centre Hospitalier Universitaire de Nantes. Ainsi, nous avons pu obtenir très tôt des données de marche mesurées auprès de patients atteints de Sclérose En Plaques par l'équipe Neurologie du Centre d'Investigation Clinique. Je tiens à souligner ici la réactivité de Fabienne Le Frère, Leslie Airiau, Melinda Moyon et Marine Thomas. Nous avons aussi pu bénéficier de l'aide du docteur Laetitia Barbin dans la rédaction de protocoles de recherche clinique et de réponses aux appels d'offre. Les analyses présentées dans ce document n'auraient pas pus être appliquées à un cas concret sans leur concours.

J'ai bénéficié d'un accueil chaleureux par mes collègues d'UmanIT, qui ont volontiers accepté de jouer le rôle de cobayes. Cela m'a permis de me familiariser avec le matériel et de développer les algorithmes et méthodes d'analyse de la marche. Vincent Graillot m'a apporté ses connaissances en électronique et programmation, et s'est chargé du développement de l'application permettant de piloter le système de capteurs et l'enregistrement des données. Le travail de Fanny Doistau a permis d'explorer les différentes options de financement et de marché pour le projet eGait. Vincent "Beardy" Cloâtre, Julie Cottu, Hélène Gaillard, Solène Garda-Krebs et Maxime Potiron se sont également prêtés à l'exercice difficile de la relecture du manuscrit pour y débusquer les coquilles et fautes d'orthographe.

Ces longues années d'étude n'auraient pas été possibles sans la confiance indéfectible et le soutien sans faille de mon père, ma mère, ma soeur, qui a su donner un certain cachet à la fin de mon doctorat, Nicolas et Jeanne. J'ai également une pensée pour ma grand-mère, Olivier, Frédérique, Jérôme, Étienne, Juliette, Thierry, Mehdi, Audrey et Sacha. Je me souviens de ceux qui sont partis.

J'ai pu compter sur mes amis et amies pour me soutenir et me changer les idées tout au long de ces années, en particulier Beardy, JCO, Pierre Bis et les montois (PDA, Guéno, Jojo, Quentin, Ronan, Pedro, Raph, Jules et LK4) qui savent se montrer quand il le faut.

Lore, j'ai eu la chance que nos routes se croisent il y a deux ans. Je n'oublierai jamais ce que tu as fait pour moi et je n'imagine pas ces instants passés et ceux qui suivront sans ton amour.

En conclusion, j'adresse mes remerciements à toutes les personnes et organisations citées précédemment pour avoir, chacune à leur échelle, permis la rédaction de cette thèse.

SOMMAIRE

INTRODUCTION

Contexte

Sclérose en plaques et troubles de la marche

Les pathologies neurogénératives se caractérisent par un vieillissement accéléré des cellules du systèmes nerveux, pouvant gravement nuire à la qualité de vie des patients. Environ 1,5 millions de personnes sont touchées en France, et une augmentation de cette prévalence est attendue selon les prévisions données par le ministère des solidarités et de la santé. [1]. La Sclérose En Plaques (SEP) est la troisième pathologie neuro-dégénérative la plus fréquente (derrière les maladies d'Alzheimer et de Parkinson) avec une estimation de 100 000 patients en France, et elle est celle qui se déclare chez les adultes les plus jeunes (diagnostic autour 25 et 35 ans), avec une prédominance féminine (3 femmes pour un homme) [2]. S'il n'y a pas d'évolution type de la pathologie, elle peut généralement prendre deux formes, la première étant caractérisée par des phases de poussées de symptômes suivies de récupération partielles ou complètes des fonctions neurologiques (SEP récurrente-rémittente) et la seconde par une dégradation progressive des fonctions neurologiques du patient sans récupération (SEP progressive) [123]. Elle réduit peu l'espérance de vie du patient dans la plupart des cas [3], mais les poussées de symptômes et la diminution des fonctions neurologiques qu'elles entraînent ont un fort impact sur la qualité de vie des patients [14]. Plusieurs traitements existent pour ralentir sa progression [123], mais aucun ne permet actuellement de guérir de cette pathologie. Ces facteurs nécessitent donc l'émergence de solutions permettant de quantifier l'état de santé des patients, permettant ainsi leur suivi médical et l'évaluation de nouvelles stratégies thérapeutiques [79].

Les déficits de la marche sont parmi les troubles les plus fréquents et considérés comme les plus handicapant par les patients atteints de SEP [67, 100]. L'évaluation de la marche tient donc une place de première importance dans leur suivi médical et dans la recherche

1. https://solidarites-sante.gouv.fr/IMG/pdf/plan_pmnd_version_longue.pdf
2. https://solidarites-sante.gouv.fr/soins-et-maladies/maladies/maladies-neurodegeneratives/article/la-sclerose-en-plaques
3. https://urlz.fr/hMZ8

clinique [114]. Plusieurs tests et échelles sont couramment utilisés en pratique. Les capacités de marche observées par le clinicien peuvent être intégrées avec l'évaluation d'autres fonctions neurologiques pour établir un score relatif à la sévérité générale de la pathologie, tel que l'*Expended Disability Status Scale* (EDSS) [97]. D'autres tests consistent à évaluer la vitesse des patients en les chronométrant sur une distance fixe [45] ou en mesurant la distance qu'ils peuvent parcourir pendant un temps limité [52]. Enfin, des questionnaires d'auto-évaluation sont utilisés pour obtenir une mesure subjective des patients sur leurs capacités de marche [69].

Si les échelles et tests chronométrés ont démontré leur utilité pour évaluer les troubles de la marche dans un contexte clinique [91], ils présentent un certain nombre de limites. Tout d'abord, bien qu'étant l'une des mesures les plus utilisées en pratique [111], le score EDSS est critiqué pour son manque de fiabilité inter-examinateur [31]. Il consiste en une mesure de l'état de santé général des patients et n'est pas spécifique des troubles de la marche. Les tests de marche chronométrés ne fournissent pas une description précise de la démarche du patient [125] et sont peu sensibles aux altérations pouvant survenir sur la période d'un essai clinique chez les patients atteints de forme progressive [170]. Ces observations traduisent le besoin de proposer des solutions permettant une mesure quantitative et objective des différents aspects de la marche des patients, en tirant parti des dispositifs numériques dédiés à l'analyse du mouvement humain [125].

Analyse de la marche par dispositif numérique

De nombreux systèmes numériques dédiés à l'analyse de la marche humaine sont décrits dans la littérature. Leur objectif commun est de mesurer des informations quantitatives relatives à la *démarche* d'un individu [119]. La *démarche* est un terme se rapportant à la façon de marcher d'un individu et consiste en une succession de *cycles de marche* [142]. Un cycle de marche est constitué de la succession d'évènements survenant entre deux poses successives d'un même pied au sol [8]. La démarche peut donc être décrite sous la forme de *paramètres spatio-temporels*, *cinétiques* et/ou *cinématiques*. Les *paramètres spatio-temporels* décrivent certains aspects d'un cycle de marche (*e.g.* durée, longueur ou vitesse). Les *paramètres cinématiques* décrivent les angles des articulations (i) entre les différents segments du corps humain ou (ii) entre l'orientation d'un segment observé à un temps donné et son orientation dans la position anatomique (*i.e.* position debout avec les pieds à plat sur une même ligne). Les *paramètres cinétiques* consistent en une description des forces et moments causant le mouvement [152, 142, 81].

Il existe une relativement grande variété de dispositifs et méthodes permettant la description quantitative de la marche. Ils peuvent être répartis en trois grandes catégories : les systèmes de traitement d'image, les capteurs au sol et les dispositifs portatifs [119]. Les systèmes de traitement d'image permettent une représentation détaillée de la démarche par une reconstruction en trois dimensions de l'individu et sont considérés comme les *Gold Standard* pour l'analyse de la marche [81], y compris pour la SEP [17, 102]. Les systèmes de capteurs au sol se présentent sous la forme de plate-formes ou de tapis équipés de capteurs de force ou de pression [119]. Ils permettent d'estimer certains paramètres spatio-temporels de la marche (longueur/durée du cycle et de ses différentes phases [35]) ainsi que la force transmise au sol par le pied [119]. S'ils permettent une représentation de la marche précise, les systèmes externes (par traitement vidéo et capteurs au sol) sont coûteux, et leur utilisation nécessite des laboratoires dédiés et du personnel spécialisé, avec des phases d'acquisition de données qui peuvent être longues [25]. Enfin, les dispositifs portatifs regroupent les systèmes de capteurs pouvant être fixés sur une partie du corps de l'individu. Ils présentent l'avantage d'être moins coûteux que les systèmes externes, et sont utilisables dans des contextes plus représentatifs des conditions de vie réelle du patient [125]. Plusieurs types de dispositifs sont utilisés. Les capteurs de pression placés au niveau de la semelle permettent de quantifier la force transmise au sol par le pied, et donc d'identifier le début et la fin des phases d'appui [71]. Un autre groupe de dispositifs portables rassemble les capteurs *inertiels*. Ces capteurs mesurent leur propre mouvement selon un ou plusieurs axes sous la forme (i) de l'accélération due à une force rectiligne pour les accéléromètres ou (ii) de la vitesse angulaire due au moment d'une force pour les gyroscopes [81]. Ils peuvent être utilisés seuls ou en combinaison, notamment dans un type de système de capteurs appelé *centrale inertielle*. Les centrales inertielles assemblent dans un même boîtier des accéléromètres et gyroscopes alignés sur 3 axes orthogonaux, parfois avec des magnétomètres [143], et permettent d'estimer l'orientation du dispositif dans l'espace à 3 dimensions.

Différents aspects du mouvement du porteur peuvent être représentés (paramètres spatio-temporels, cinématique ou cinétique) selon le type de système utilisé. La détermination de ces paramètres nécessite l'élaboration d'algorithmes adaptés au signal mesuré pour identifier les évènements particuliers du cycle de marche (*e.g.* pose du pied, décollement des orteils). Ces évènements correspondent la plupart du temps à des phénomènes remarquables dans le signal (*e.g.* extremums locaux, phases stationnaires ou inversions de signe dans les données mesurées) [81]. Les algorithmes correspondent à un ensemble de

règles permettant d'identifier ces phénomènes (*e.g.* algorithme de détection d'extremums [85]) et de déterminer à quels évènements du cycle de marche ils correspondent en fonction du type de signal et de l'aspect du mouvement qui est mesuré [142].

La détection des différents évènements du cycle permet d'une part le calcul de paramètres spatio-temporels de la démarche. Elle permet également la formation de segments décrivant un ou plusieurs aspects du mouvement sous la forme d'une séquence d'états sur une période correspondant à un cycle de marche [47]. Ces segments peuvent par exemple représenter la cinématique de la partie basse du corps, correspondant aux angles entre les différents segments des membres inférieurs au cours du cycle [142]. Cette orientation peut être représentée sous la forme de *quaternions unitaires*, nombres hyper-complexes de dimension 4 représentant des rotations en 3 dimensions [54]. Les segments décrivant les cycles de marche peuvent permettre de déterminer plusieurs paramètres tels que les angles des articulations observés à des instants particuliers du cycle ou l'amplitude entre deux orientations différentes d'un même segment [152].

Nombreuses sont les études mettant en évidence l'existence de relations significatives entre des paramètres de la marche mesurés par dispositifs numériques et la sévérité de la SEP. Elles démontrent ainsi l'intérêt de l'utilisation de ces technologies dans le suivi de la pathologie [103, 117, 125, 46, 7]. Une autre approche consiste à analyser les données de marche par méthodes de classification pour former des groupes de patients présentant des déficits similaires.

Classification de données de marche

Plusieurs types de méthodes d'*apprentissage automatique* sont utilisées dans l'analyse de la marche. Des méthodes de *classification supervisée* peuvent être appliquées pour déterminer la présence ou non d'une pathologie chez un individu en fonction de ses données de marche [72, 104]. Ces méthodes s'appliquent dans le cas où les groupes dans lesquels classer les individus sont connus *a priori*. Une autre approche consiste à explorer les données pour identifier l'existence de groupes de patients partageant des déficits de marche communs. Cette approche se base sur la *classification non supervisée*, dans laquelle les méthodes d'apprentissage consistent à répartir les observations d'un jeu de données dans des groupes en fonction de leur similarité [2]. La *classification non supervisée* peut être réalisée à partir d'un ensemble de *paramètres spatio-temporels* et *cinématiques* déterminés à partir des données de marche [41, 118]. Une autre approche consiste à former des groupes de patients en fonction de la forme des segments représentant leur cycles de marche (ou

pattern de marche) par *Classification Ascendante Hiérarchique* [139]. Pour ce faire, une mesure de la dissimilarité entre observations appropriée à leur type de données est nécessaire. Les *patterns* de marche se présentant sous la forme d'une succession d'états observés au cours du temps [47], ils peuvent être considérés comme des *séries chronologiques* (ou séries temporelles) ou des *données fonctionnelles* (ou *courbes*).

Une problématique commune à ces deux types de données est le mauvais *alignement temporel des données*. Il correspond aux situations pour lesquelles un évènement équivalent est présent dans deux observations, mais à des instants différents. Dans le cas de l'analyse de la marche par exemple, le décalage entre la survenue du décollement du pied observé entre deux cycles de marche peut être dû à une imprécision de la méthode ayant permis la segmentation des données. Pour s'affranchir de ce mauvais alignement, les méthodes de classification de *séries chronologiques* peuvent recourir à des mesures de dissimilarité dites *élastiques* [4], permettant le réalignement temporel des données. Le *Dynamic Time Warping* est réputé comme la dissimilarité élastique la plus utilisée en classification de séries chronologiques [2]. Il est également employé dans l'analyse de la marche [11, 134, 163]. Dans le cas des données fonctionnelles, ce phénomène est appelé *variation de phase*, en opposition à la *variation d'amplitude* qui se rapporte à la forme des courbes [109]. Plusieurs méthodes permettant l'alignement temporel de courbes par suppression de la variation de phase existent, telle que la distorsion de l'axe des temps par fonction de *warping* [171, 185].

Durant leur suivi, les patients atteints de maladies neurodégénératives telles que la Sclérose En Plaques sont régulièrement examinés par un clinicien au cours de consultations médicales. Des informations supplémentaires concernant leur état de santé global sont donc la plupart du temps disponibles. Contrairement aux méthodes de classification non supervisées, qui ne permettent pas de tirer profit de telles connaissances *a priori*, les méthodes de classification *semi-supervisée* permettent l'intégration d'informations supplémentaires dans le processus de construction des groupes. Il a été démontré que l'ajout, même minime, d'informations supplémentaires augmentait de manière significative les performances des algorithmes de classification [39]. Les méthodes de classification semi-supervisée se répartissent en trois grandes catégories : (i) les méthodes de classification avec *contraintes*, (ii) les *ensembles de classification* et (iii) les méthodes de classification par *compromis*. Les premières tirent leur nom de l'utilisation de *contraintes* qui définissent des règles que la classification finale doit respecter [39]. Il peut s'agir de paires d'observations devant appartenir au même groupe ou à des groupes différents dans la classification

finale, ou de contraintes définissant la structure autorisée pour les groupes finaux (*e.g.* cardinal ou taille maximale du groupe, séparation minimale entre groupes, *etc...*). Les méthodes par *ensemble de classification* consistent à compiler le résultat de plusieurs méthodes de classification appliquées sur le même ensemble d'observations dans une classification dite *consensus* [33]. Les méthodes de classification par *compromis* consistent à calculer la dissimilarité globale entre observations comme la somme pondérée des dissimilarités observées dans les espaces des différentes sources de données disponibles, de telle sorte que la classification finale soit le meilleur compromis entre les informations qu'elles apportent [16]. Il existe peu de méthodes de classification semi-supervisées adaptées aux séries chronologiques ou aux données fonctionnelles, et seules quelques méthodes de classification avec *contraintes* [98] et *d'ensemble de classification* [183, 172, 147] généralisées aux séries chronologiques sont décrites dans la littérature.

Ce manuscrit de thèse présente les travaux réalisés pour analyser la marche à partir des rotations de la hanche. Les approches proposées diffèrent de celles décrites dans la littérature par le fait qu'elles se basent sur l'analyse de données mesurées par *un unique capteur inertiel placé à la ceinture*, et que les données représentent *l'orientation en 3 dimensions de la hanche* au cours de la marche sous la forme d'une *séquence de quaternions unitaires*. Tout d'abord, un algorithme appelé `STRIde PAttern GEneration` (`STRIPAGE`) est développé pour identifier les cycles de marche dans les données mesurées par le dispositif. Cet algorithme repose sur un ensemble de règles permettant d'identifier les instants correspondant à la pose du pied au sol et à son décollement à partir des changements d'orientation de la hanche. L'identification de ces évènements permet de segmenter les données en un ensemble de séquences correspondant aux cycles de marche. Deux approches sont abordées pour analyser la marche à partir de ces cycles. La première approche repose sur la détermination de *paramètres spatio-temporels* représentant plusieurs aspects des cycles de marche. La seconde approche consiste à former des groupes rassemblant des individus en fonction de la similarité de leur démarche. Pour ce faire, des méthodes de *classification non supervisée* et *semi-supervisée* sont adaptées aux séquences de quaternions unitaires. La classification de données fonctionnelles de quaternions permet notamment le calcul du centre de l'ensemble des cycles de marche détectés chez un individu pour définir le biomarqueur appelé *Signature de Marche*. Ces travaux ont été réalisés dans le cadre du projet *eGait* du pôle *Recherche & Développement* de l'entreprise UmanIT, en collaboration avec le Laboratoire de Mathématiques Jean Leray UMR CNRS 6629 de Nantes.

Présentation du projet *eGait*

Collaborateurs

L'entreprise UmanIT [4] est une société nantaise dont le secteur principal d'activité est le développement de solutions web et mobile pour entreprise. Par le biais de son pôle Recherche et Développement, elle souhaite apporter des solutions numériques innovantes aux différents acteurs de la santé en tirant partie de son expertise dans le domaine de l'informatique. En 2017, elle s'associe par un contrat de collaboration de recherche avec le Laboratoire de Mathématiques Jean Leray de Nantes Université (LMJL) [5] pour l'assister dans le développement de méthodes d'analyse statistique des données de santé. Le premier projet porte sur la numérisation du score *Multiple Sclerosis Fonctional Composite* (MSFC) sous la forme d'une application pour *smartphone* à écran tactile, appelée *e-MSFC*. Le score MSFC représente l'état de santé global d'un patient atteint de SEP et est calculé à partir d'un test d'adresse, d'un test cognitif et d'un test de marche [36]. Les versions numérisées des tests d'adresse et cognitifs sont développées en 2017. Je rejoins le projet *e-MSFC* dans le cadre d'un stage de Master 2 *Modélisation en Pharmacologie Clinique et Épidémiologie* en avril 2017. Le sujet de mon stage porte sur la validation des versions numérisées du test d'adresse et du test cognitif de l'échelle *e-MSFC*, développées en collaboration avec le Centre Hospitalier Universitaire de Toulouse, en comparant les résultats obtenus par des patients atteints de SEP avec ceux des tests de l'échelle MSFC classique. La poursuite du projet *e-MSFC* consiste à coupler les deux tests numérisés de l'échelle *e-MSFC* à une solution permettant l'analyse de la marche appelée *eGait*. La solution *eGait* repose sur l'utilisation d'un système de capteurs de mouvements portatif. Le développement de la solution *eGait* fait l'objet de mes travaux de recherche en qualité de doctorant dans le cadre d'une Conventions Industrielles de Formation par la REcherche (CIFRE), qui débute en 2018.

Caractéristiques de la solution *eGait*

Après la revue de l'état de la littérature sur l'analyse de la marche dans la SEP et des systèmes numériques, il a été convenu de diriger le développement de la solution *eGait* vers l'utilisation d'un système de capteurs de mouvement portatif. La solution doit respecter

4. https://www.umanit.fr/
5. https://www.math.sciences.univ-nantes.fr/fr

les contraintes suivantes :

— Permettre la mesure quantitative d'un ou plusieurs aspects de la marche.

— Être sensible aux déficits de la marche chez le porteur et à leur sévérité.

— Être compatible avec une utilisation en contexte clinique par son prix et par son ergonomie.

— Pouvoir être utilisable pour une mesure en vie quotidienne du patient.

Pour respecter ces contraintes, il a été choisi d'utiliser un unique système de capteurs placé à la ceinture en position latérale droite. Le dispositif *MetaMotionR*[6] développé par *Mbientlab* est sélectionné pour le développement du premier prototype. Ce système de capteurs embarque un accéléromètre, un gyroscope et un magnétomètre sur 3 axes, et une unité centrale de calcul. Cette dernière fusionne les données des 3 types de capteurs pour déterminer l'orientation du dispositif en 3 dimensions au cours du temps sous la forme de *quaternions unitaires*. Le système est piloté par une application Androïd installée sur un *smartphone* qui stocke les données mesurées. Par sa position, le dispositif mesure les rotations en 3 dimensions de la hanche du porteur durant la marche. Des méthodes de traitement de ce signal sont nécessaires pour l'analyse quantitative de la démarche du porteur.

Missions de la thèse

— Acquisition de données de marche d'individus sains.

— Co-organisation et co-rédaction de protocole d'essais cliniques pour l'acquisition de données de marche de patients atteints de Sclérose En Plaques.

— Data management des données de marche des essais cliniques.

— Co-rédaction de réponses à appels à projet et de demandes de financement pour soutenir le projet.

— Développement de méthodes d'analyse de la marche adaptées aux quaternions unitaires.

— Validation de ces méthodes à partir de données réelles mesurées chez des individus sains et chez des patients atteints de Sclérose En Plaques.

— Encadrement de stagiaires de niveau Master 2.

6. https://mbientlab.com/metamotionr/

Contributions

Financements

— Décembre 2018 : Co-rédaction d'un dossier Projet Exploratoire, Premier Soutien (PEPS) adressé à l'Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société (AMIES) pour le financement d'une étude clinique préparée en collaboration avec le CHU de Nantes. La somme de 15 000€ a pu être levée.

— Septembre 2019 : Co-rédaction d'une candidature à l'appel à projet « Approche personnalisée, éthique, sociologique et économique de la SEP par la recherche » de la fondation pour l'Aide à la Recherche dans la Sclérose En Plaques (rédacteur principal : Aymeric Stamm). La somme de 150 000€ a pu être levée.

Études cliniques.

— Une première étude exploratoire a pu être organisée en partenariat avec l'équipe Neurologie du Centre d'Investigation Clinique de Nantes. Nous avons contacté les Pr. David Laplaud, (Neurologue et PU-PH au CHU de Nantes) et le Pr. Pierre-Antoine Gourraud (PU-PH au CHU de Nantes) pour leur présenter le projet *eGait* en 2018, alors qu'ils menaient une étude pour étudier le signal nerveux de patients atteints de SEP mesuré par le bracelet électronique MYO. La rédaction d'un amendement a permis d'ajouter à cette étude principale une étude ancillaire pour mesurer la marche de 30 patients. La période d'inclusion de ces patients s'est étendue de Septembre 2019 à Mai 2020. Les analyses des données de marche ont permis une première validation de la relation entre les paramètres spatio-temporels de la démarche mesurés par la solution *eGait* et la gravité de la pathologie des patients. Elles constituent également une première preuve de la possibilité de former des groupes de patients en fonction de leur démarche et de la sévérité de leur pathologie par méthode de classification semi-supervisée.

— Une seconde étude clinique a été préparée avec l'équipe du CIC de neurologie du CHU de Nantes. L'objectif de cette étude est de confirmer les conclusions de l'étude ancillaire quant à la relation entre les données mesurées par la solution *eGait* et la sévérité de la Sclérose En plaques. Cette étude incluant 44 patients a pu être financée grâce au soutien du PEPS de l'AMIES. La période d'inclusion de ces patients est toujours en cours et s'étend de Aout 2021 à Aout 2022.

Contributions scientifiques.

- Algorithme de détection des cycles de marche `Stride PAttern GEneration` (`STRIPAGE`, langage de programmation R) et génération d'un biomarqueur nommé *Signature de Marche* (SdM).
- Méthodes de détermination de paramètres spatio-temporels du cycle de marche à partir de séquences de quaternions unitaires.
- Généralisation des méthodes de classifications ascendantes hiérachiques, *K-means* et *K-medoids* de séries chronologiques et données fonctionnelles aux quaternions unitaires.
- Généralisation des méthodes de classification semi-supervisée `hcluscompro` et `mergeTrees` aux séries chronologiques de quaternions unitaires.

Encadrement de stages.

- Étudiant : Benjamin Martineau
  Diplôme préparé : Master 2 Ingénierie Statistiques, Nantes Université.
  Sujet : Utilisation d'objets connectés dans le développement de dispositif médical : Détection de cycles de marche
  Période : Second semestre 2020
- Étudiant : Raphael Brard
  Diplôme préparé : Master 2 Ingénierie Statistiques, Nantes Université
  Sujet : Détection de la marche à l'aide de système de capteurs de mouvement en vie quotidienne
  Période : Second semestre 2021
- Étudiante : Léonie Veuille
  Diplôme préparé : Master 2 Ingénierie Statistiques, Nantes Université
  Sujet : Analyse des données d'un système de capteurs de mouvement
  Période : Second semestre 2022

Valorisations

Brevet. La méthode de détection des cycles de marche et du calcul du biomarqueur Signature De Marche à partir d'une séquence de quaternions unitaires fait l'objet de la demande de brevet n° 21 00309 : « Méthode et dispositif de détermination d'un cycle de

marche » (Co-auteurs : Pierre Drouin, Lise Bellanger, Aymeric Stamm, Laurent Chevreuil, Vincent Graillot). Elle a été déposée à l'Institut National de la Propriété Industrielle le 13 janvier 2021 et est actuellement en cours de validation.

Publications scientifiques.

- En tant que premier auteur :
  - "Semi-supervised clustering of quaternion time series : application to gait analysis in multiple sclerosis using motion sensor data". Pierre Drouin, Aymeric Stamm, Laurent Chevreuil, Vincent Graillot, Laetitia Barbin, Pierre-Antoine Gourraud, David-Axel Laplaud, Lise Bellanger, Statistics in Medicine, Wiley Online Library [7] : Accepté avec révision mineure en Mai 2022.
  - "Gait impairment monitoring in multiple sclerosis using a wearable motion sensor". Pierre Drouin, Aymeric Stamm, Laurent Chevreuil, Vincent Graillot, Laetitia Barbin, Philippe Nicolas, Pierre-Antoine Gourraud, David-Axel Laplaud, Lise Bellanger. (Medical Case Reports and Reviews, Open Access Text [8]). Publié en Février 2022.
- En tant que co-auteur (hors travaux de thèse) :
  - "Smoothing Method for Unit Quaternion Time Series : An application to motion data". Elena Ballante, Lise Bellanger, Pierre Drouin, Silvia Figini, Aymeric Stamm, Journal of Statistical Planning and Inference [9]. Soumission mars 2022.
  - Brard, Raphaël, Lise Bellanger, Laurent Chevreuil, Fanny Doistau, Pierre Drouin, and Aymeric Stamm . 2022. "A Novel Walking Activity Recognition Model for Rotation Time Series Collected by a Wearable Sensor in a Free-Living Environment" Sensors 22, no. 9 : 3555. https://doi.org/10.3390/s22093555

Conférences et communications orales.

- 14-16 décembre 2019 : Présentation orale "Clustering time-varying quaternion data : Application to the detection of gait abnormalities". 12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019 [10]), Londres, Angleterre.

7. https://onlinelibrary.wiley.com/journal/10970258
8. https://rb.gy/gnsawu
9. https://www.journals.elsevier.com/journal-of-statistical-planning-and-inference
10. http://www.cmstatistics.org/CMStatistics2019/

— 15-16 septembre 2020 : Poster "Comparaison de méthodes de clustering pour détecter des troubles de la marche à partir de données issues d'un capteur de mouvement". 14ème Conférence Francophone d'EPIdémiologie CLINique [11].

— 7-9 avril 2021 : Poster "Compromise-based clustering for quaternion time series and its application to gait analysis in Multiple Sclerosis". 8th Channel Network Conference – Société Française de Biométrie [12].

— 29 juin - 01 juillet 2022 : Présentation orale "Classification semi-supervisée de séries temporelles de quaternions pour l'analyse des troubles de la marche dans la sclérose en plaques" - Conférence Intelligence Artificielle et santé : approches interdisciplinaires, Nantes [13].

— 10-15 Juillet 2022 : Présentation orale "Semi-supervised clustering of quaternion time series for the analysis of gait impairment in multiple sclerosis". 31st International Biometric Conference – International Biometric Society, Riga, Lettonie [14].

Vulgarisation.

— Communications orales

  — Journées Scientifiques de l'Université de Nantes – Colloque IA et interdisciplinarité : application en santé et industrie, 7 juin 2021.

  — Communication orale : Journée Portes Ouvertes de la fondation pour l'Aide à la Recherche sur la Sclérose En Plaques (ARSEP), CHU de Nantes, 26 novembre 2021.

— Article : « Détecter des troubles de la marche », Tangente Hors-série n°73. « Les mathématiques au cœur de l'emplois ». Parution : 1 février 2021.

— *AMIES Success Story* : "Daily-life gait impairment detection", UmanIT company and the Dept of Mathematics Jean Leray (Nantes) (2022) [15].

Organisation du mémoire

Le chapitre 1 présente l'état de l'art de l'analyse de la marche dans la Sclérose En Plaques dans la section 1.1, le dispositif *MetaMotionR* utilisé pour la solution *eGait*,

11. http://epiclin2020.congres-scientifique.com/
12. https://cnc21.sciencesconf.org/
13. https://www.lebesgue.fr/fr/conf_IA_sante2022
14. https://www.ibc2022.org/home
15. https://eu-maths-in.eu/portfolio/umanit-daily-life-gait-impairment-detection/

l'algèbre des données mesurées, *i.e.* les quaternions unitaires, ainsi que les bases de données utilisées dans la suite des analyses dans la section 1.2. La section 1.3 présente l'état de l'art des méthodes de classification non supervisées de séries chronologiques et fonctionnelles, ainsi que des méthodes de classification semi-supervisées.

Le chapitre 2 décrit l'algorithme `StriPaGe` qui permet l'identification des cycles de marche dans les données mesurées par le dispositif *MetaMotionR* ainsi que sa validation sur une base de données d'apprentissage. Les méthodes développées pour calculer les paramètres spatio-temporels de la marche à partir de ces cycles sont détaillées et évaluées à partir de données de marche de patients atteints de SEP dans la section 2.2.

Le chapitre 3 présente la généralisation des méthodes de classification ascendante hiérarchique, *K-means*, *K-medoids* aux séries chronologiques et de données fonctionnelles aux quaternions unitaires dans la section 3.1. La méthode du *K-means* sur données fonctionnelles permet de calculer le biomarqueur *Signature de la Marche* (SdM), représentatif de la démarche de l'individu. Les méthodes de classification non-supervisées de séries chronologiques et de données fonctionnelles sont ensuite comparées quant à leur capacité à regrouper les SdM d'individus sains en fonction des conditions de marche dans lesquelles elles ont été mesurées. Enfin, la section 3.2 présente la généralisation des deux méthodes de classification semi-supervisées par compromis `hclustcompro` et par consensus `mergeTrees` aux séries chronologiques de quaternions. Elles sont comparées en les appliquant aux SdM de patients atteints de SEP en intégrant leur score EDSS comme source d'information supplémentaire.

Chapitre 1

Prérequis

Ce chapitre présente les concepts qui seront abordés dans le reste du manuscrit. L'état de l'art de l'analyse de la marche dans la Sclérose En Plaques et par dispositifs numériques est passé en revue dans la section 1.1. Le dispositif *MetaMotionR* utilisé pour le projet *eGait*, le format mathématique des données mesurées et les bases de données qui sont analysées dans les chapitres suivants sont décrits dans la section 1.2. La possibilité de former des groupes d'individus en fonction de leurs données de marche par méthode de classification est également explorée dans ce manuscrit. Les données mesurées se présentent sous la forme d'une séquence de mesure au cours du temps, elles peuvent donc être appréhendées comme des séries chronologiques ou des données fonctionnelles. Les méthodes de classification non-supervisées et semi-supervisées adaptées à ces types de données sont présentées dans la section 1.3.

1.1 La marche et son évaluation chez les patients atteints de Sclérose En Plaques

1.1.1 Marche et dispositifs numériques existants

1.1.1.1 Définition générale de la marche

Si la marche est exécutée quotidiennement par les personnes valides, et ce de manière quasiment inconsciente, elle n'en reste pas moins un phénomène complexe faisant intervenir un grand nombre de fonctions physiologiques. Largement étudiée en biomécanique, elle y est définie comme un déplacement consistant en une translation de l'ensemble du corps, consécutive à des mouvements de rotations [20]. Son exécution implique la répétition de séquences de mouvements des segments corporels afin de déplacer le corps vers l'avant tout en le gardant en équilibre [127]. Ces séquences répétées sont généralement appelées *cycles de marche*.

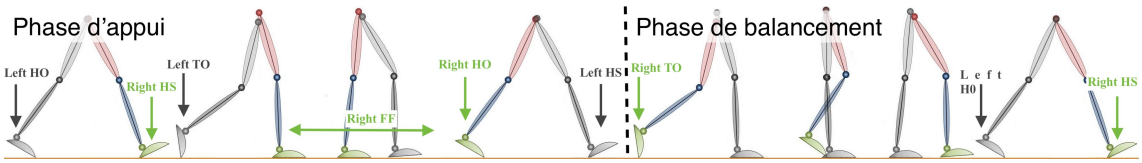


Fig. modifiée à partir de : https://commons.wikimedia.org/wiki/File:GaitCycle_by_JaquelinPerry.jpg

FIGURE 1.1 – Cyle de marche et ses phases

Un cycle de marche est défini comme l'ensemble des mouvements réalisés entre deux contacts consécutifs du talon d'un même pied avec le sol [8]. Un cycle de marche peut donc être décrit en considérant les évènements du pied droit ou du pied gauche. Un exemple de cycle de marche est schématisé sur la figure1.1, en considérant le pied droit (représenté en vert sur la figure). Plusieurs évènements surviennent durant un cycle de marche :

- Le contact du talon au sol marque le début d'un cycle (noté HS pour *Heel Strike*).
- Le contact des orteils avec le sol, qui correspond à l'instant où le pied est à plat sur le sol (notée FF pour *Flat Foot*).
- Le décollement du talon du sol (noté HO pour *Heel Off*).
- Le décollement des orteils du sol (noté TO pour *Toe Off*).

Un *cycle de marche* peut être décomposé en plusieurs phases à partir de ces évènements [7] :

Un pas est défini comme la période comprise entre le contact d'un pied avec le sol et celui du pied opposé. Un *pas pied droit* est délimité par le contact du talon droit (*right* HS) et le contact du talon gauche qui lui succède (*left* HS).

La phase d'appui d'un pied est la période durant laquelle il est en contact avec le sol. Pour un *cycle de marche pied droit*, elle est délimitée par le contact du talon droit (*right) HS* et le décollement des orteils du pied droit (*right* TO).

La phase de balancement d'un pied est la période durant laquelle il n'est pas en contact avec le sol. Pour un *cycle de marche pied droit*, elle est délimitée par le décollement des orteils du pied droit (*right* TO) et le contact du talon droit qui lui succède (*right* HS).

La phase de simple appui est la période durant laquelle un seul pied est en contact avec le sol. Pour un *cycle de marche pied droit*, elle est délimitée par le

décollement du pied gauche du sol (*left* TO) et le contact du pied gauche qui lui succède (*left* HS).

La phase de double appui est la période durant laquelle les deux pieds sont en contact avec le sol. Pour un *cycle de marche pied droit*, elle est délimitée par le contact du talon gauche au sol (*left* HS) et le décollement des orteils du pied droit (*right* TO).

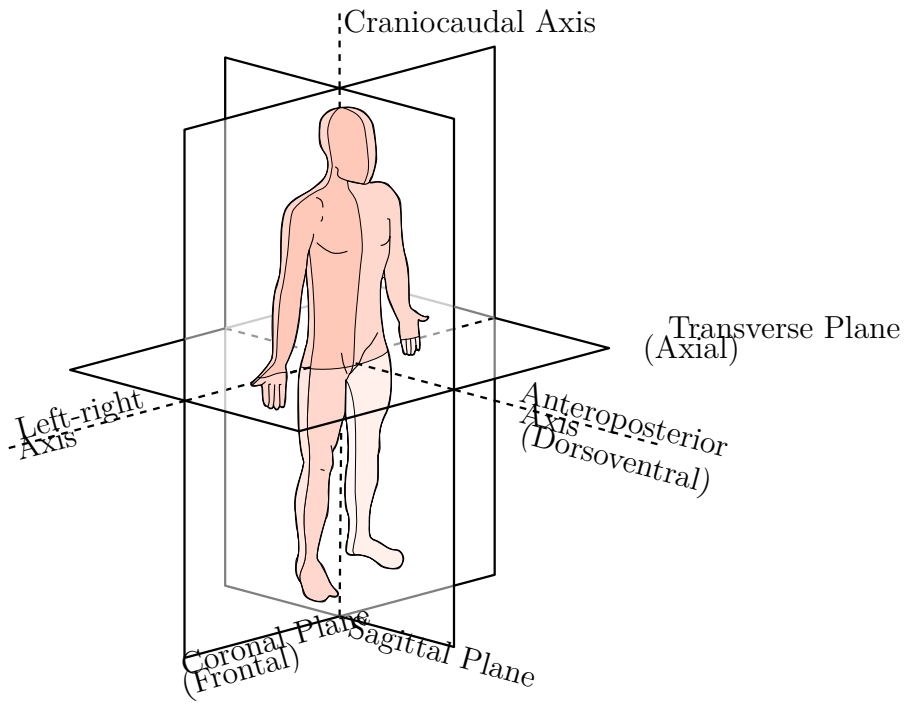
La *démarche* désigne la manière de marcher d'un individu, et son analyse consiste à décrire quantitativement plusieurs aspects du *cycle de marche* et de ses différentes phases [26] :

Les paramètres spatio-temporels sont généralement des valeurs scalaires décrivant différentes caractéristiques du cycle de marche. Les *paramètres spatiaux* représentent des distances, telles que la longueur parcourue durant un pas [103, 125, 117] ou la largeur d'appui (écartement des pieds lorsqu'ils entrent en contact avec le sol) [103, 46] Les *paramètres temporaux* sont relatifs au rythme de la marche (leur écart type caractérisant la variabilité temporelle du cycle) [7]. Ils peuvent correspondre à la durée du cycle, du pas, des phases d'appui, de balancement et/ou des phases de simple et double appui [7, 46, 117, 125, 103]. Ces durées peuvent être exprimées en unité de temps, en pourcentage de la durée du cycle, ou être converties en fréquence (*e.g.* nombre de pas ou cycles par unité de temps), Les paramètres *spatio-temporels* sont relatifs aux changements de position du corps au cours du temps. L'estimation de la vitesse de marche de l'individu est par exemple largement décrite [103, 46, 125, 117, 115, 164].

La cinématique est l'étude de l'orientation des segments du corps humain au cours de la marche [64]. Cette orientation est typiquement calculée comme l'angle d'une articulation entre deux segments du corps [92]. Ils sont exprimés dans le plan sagittal, défini par les axes antéro-postérieur et cranio-caudal, le plan frontal , défini par les axes gauche-droite et cranio-caudal, et le plan transverse, défini par les axes antéro-postérieur et gauche-droite (voir figure 1.2).

La cinétique de la marche est l'étude des forces permettant le mouvement. Il peut s'agir de la pression transmise au sol par le pied, du moment (en Newton-mètre par radian) ou de la puissance (en Watts par kilogramme) des articulations [64].

La démarche est connue pour présenter une variabilité *inter* et *intra-individuelle* [68, 47]. La *durée* d'un cycle et de ses différentes phases peut donc être variable entre deux



Source : https://commons.wikimedia.org/wiki/File:Anatomical_Planes-en.svg

FIGURE 1.2 – Position anatomique et plans de coupe

cycles mesurés chez un même individu ou des individus différents. Cette variabilité est désignée sous le terme de *quasi-* ou *semi- périodicité de la marche* [68]. La variabilité n'impacte pas seulement l'aspect temporel des cycles de marche, mais également leur aspect *spatial*. Les mouvements du corps varient ainsi également entre des cycles observés chez un même individu ou chez des individus différents. La variabilité *inter-individuelle* des cycles étant supérieure à la variabilité *intra-individuelle*, chaque individu présente sa propre *démarche* (permettant par exemple d'identifier un individu à partir de sa *démarche* [161]). La *démarche* individuelle dépend des fonctions musculo-squelettiques et nerveuses de l'individu [180], le déficit d'une ou plusieurs de ces fonctions peut donc avoir un impact sur plusieurs aspects du cycle de marche [86]. L'analyse quantitative de la marche a donc un intérêt en clinique pour identifier les aspects de la démarche impactés par la pathologie [181]. Les différents systèmes numériques et méthodes utilisés pour quantifier la marche sont présentés dans la section suivante.

1.1.1.2 Analyse de la marche par systèmes numériques

Les *systèmes numériques* pour l'analyse de la marche rassemblent les dispositifs permettant de quantifier plusieurs aspects de la démarche d'un individu. Ils se répartissent

principalement en trois catégories [119] :

Les systèmes d'analyse d'image. Ces systèmes se basent sur l'utilisation de caméras filmant l'individu en train de marcher. Les méthodes basées sur l'analyse d'image se répartissent en deux catégories [94] :

Capture de mouvement utilisant des marqueurs. Ces méthodes recourent à des marqueurs placés à plusieurs endroits du corps de l'individu. Des systèmes de reconnaissance sont utilisés pour identifier la position et l'orientation de ces marqueurs dans l'espace en 3 dimensions à partir d'images acquises par des caméras vidéos. La combinaison de la position de ces capteurs avec des informations anthropométriques permet de représenter le corps humain par une chaîne de segments articulés. La *cinématique*, la *cinétique* des articulations du corps humain, ainsi que plusieurs paramètres de la marche peuvent être déterminés à partir de ces données [102]. Le système *Viper*[1] développé par *Vicon Motion Systems*, ou ceux développés par *OptiTrack*[2] sont des exemples de systèmes de capture de mouvement utilisant des marqueurs. *La capture de mouvement avec marqueurs* est considérée comme le *Gold Standard* pour l'analyse de la marche [102]. Elle présent néanmoins plusieurs limitations :

— Elle est sensible à la position des marqueurs sur le corps [94].

— Son utilisation nécessite un laboratoire dédié, du personnel qualifié, et l'acquisition des données peut être longue et coûteuse [25, 125].

Capture de mouvement sans marqueurs. Ces méthodes se répartissent en deux groupes en fonction du type de système vidéo et du pré-traitement des images recueillies.

(i) *Les systèmes actifs avec reconnaissance de motifs.* Les systèmes *actifs* émettent de la lumière qui sera réfléchie par les objets et l'individu présents dans la scène. La distance entre les surfaces réfléchissant la lumière et le système émetteur est estimée. Des algorithmes de reconnaissance de posture peuvent alors être utilisés pour estimer la position des segments du corps en 3D [156]. Comme pour la capture de mouvement avec marqueurs, la démarche peut être représentée par sa cinématique, sa cinétique ou sous forme de paramètres spatio-temporels. Si ces systèmes sont moins coûteux

1. https://www.vicon.com/hardware/cameras/viper/
2. https://optitrack.com/applications/movement-sciences/

que la capture de mouvement avec marqueurs, leur utilisation n'en reste pas moins limitée à la mesure de la marche dans un espace clos. Ils sont également sensibles à la lumière ambiante [94].

(ii) *Détermination de la silhouette par systèmes passifs.* Les données sont mesurées par des caméras *classiques* (*i.e. qui reçoivent la lumière sans en émettre*). Les images sont pré-traitées pour identifier la zone correspondant à l'individu, appelée *silhouette*. Une image pré-traitée se présente sous la forme d'une matrice de pixels blancs s'ils sont inclus dans la *silhouette* de l'individu, noirs sinon. Une image pré-traitée représente donc la *posture* de l'individu à un instant donné par un ensemble de pixels blancs dont la forme correspond à sa silhouette. Un ensemble d'images de mêmes dimensions représentant les différentes postures de l'individu au cours de la marche est ainsi obtenu. La démarche est alors représentée sous la forme de la matrice de pixels dont le niveau de gris est déterminé comme la moyenne des pixels observés sur l'ensemble des images prétraitées, appelée la *Gait Energy Image* [63]. Cette méthode présente l'avantage de pouvoir être réalisée avec du matériel moins spécifique que les systèmes actifs. Elle est cependant sensible à la corpulence de l'individu, aux vêtements qu'il porte et à la présence de plusieurs individus dans le champ, bien que des méthodes de traitements spécifiques soient décrites pour s'affranchir de ces limitations [94].

Les plateformes et tapis de capteurs de pression. Ils sont constitués d'une série de capteurs mesurant la force de réaction du sol au moment de la pose du pied. Ils permettent ainsi d'identifier facilement les évènements du cycle de marche et d'étudier la force transmise au sol durant l'appui du pied [81, 94]. Ils ne permettent cependant pas de représenter la cinématique de la marche [94]. Le tapis *GaitRite*[3] est un exemple de ce type de système.

Les systèmes portatifs. Ils correspondent aux dispositifs qu'il est possible de placer directement sur l'individu. Les types de systèmes portatifs se distinguent par le type de données qu'ils mesurent.

Les capteurs de pression ou de force mesurent la force à laquelle ils sont soumis. Ils sont placés sous le pied pour mesurer la force transmise par le pied au

3. https://www.gaitrite.com/

sol [119]. En fonction de la surface du pied recouverte, ils peuvent détecter la pose et/ou le décollement du talon et/ou des orteils [157], et ainsi déterminer plusieurs paramètres spatio-temporels du cycle de marche (durée des phases d'appui, de balancement, du pied à plat,*etc...*). Ils disposent d'une excellente précision pour la détection de ces évènements chez les sujets sains, mais elle peut être diminuée dans le cas de patients traînant leurs pieds au sol [107]. Leur aspect cosmétique et leur durabilité constituent également une limitation à leur utilisation quotidienne [154].

Les capteurs inertiels peuvent mesurer leur propre mouvement. Ils sont principalement représentés par les accéléromètres et les gyroscopes [81]. Ils mesurent l'accélération linéaire (resp. la vitesse angulaire) résultant d'une force extérieure (resp. d'un torque) selon un axe. Ils peuvent être utilisés seuls ou assemblés dans un même dispositif. Un assemblage courant consiste notamment à aligner 3 accéléromètres orthogonaux avec 3 gyroscopes orthogonaux dans un même dispositif, appelé *Centrale Inertielle* (*Inertial Measurement Unit*, ou IMU). L'utilisation de capteurs inertiels pour la mesure de la marche présente un certain nombre d'*avantages* : leur faible coût en comparaison des autres systèmes (analyse vidéo, tapis de capteurs,*etc...*), leurs faibles taille et poids, et le fait que leur usage ne soit pas restreint à un laboratoire dédié [88]. Ces caractéristiques permettent également une mesure de la marche dans la vie quotidienne de l'individu. Ils ont donc le potentiel de fournir des informations quantitatives et objectives reflétant l'impact de la pathologie du patient sur son quotidien.

Du fait de leurs avantages énoncés précédemment, les travaux de recherche de la thèse se sont portés sur l'analyse de la marche par systèmes de capteurs inertiels, qui est détaillée de manière plus approfondie dans la suite de cette section.

Analyse de la marche et systèmes de capteurs inertiels portatifs. La miniaturisation des capteurs inertiels et l'augmentation de leur puissance de calcul ont conduit à dynamiser la recherche dans leur utilisation pour analyser la marche [142, 81]. Les données renvoyées par ces capteurs se présentent typiquement sous la forme d'une *séquence* d'*éléments* $S = (s_1, \ldots, s_N)$ associés à leur *temps de mesure* $T = (t_1, \ldots, t_N)$, où $i \in \{1, \ldots, N\}$ correspond à l'*indice* d'un élément et N est le nombre d'éléments de la séquence S. Dans cette notation générique, s_i peut être un scalaire, un vecteur ou une ma-

trice en fonction du type de dispositif utilisé. Par exemple, dans le cas d'un *accéléromètre* (resp. gyroscope) à un axe, s_i est de dimension 1 et sa valeur correspond à l'*accélération* en m/s^2 (resp. *vitesse angulaire* en °/s) mesurée au temps t_i dans le sens de cet axe. s_i peut également être multi-dimensionnel, dans le cas par exemple des centrales inertielles composées de plusieurs capteurs.

Les capteurs inertiels peuvent être placés à différents endroits du corps, par exemple au niveau du pied [46, 117], des chevilles/mollets [40, 34, 7] et/ou des lombaires [125, 34, 115, 164, 7]. Les aspects de la marche qu'il est possible de mesurer dépendent de leur position, mais également du type de système utilisé, de leur nombre et des méthodes et algorithmes utilisés pour traiter leur signal.

Si les accéléromètres (resp. gyroscopes) mesurent l'accélération (resp. la vitesse angulaire), d'autres aspects du mouvement peuvent en être déduits. L'intégration de la vitesse angulaire permet de déterminer le déplacement angulaire, et donc l'orientation du segment sur lequel un gyroscope est fixé. Les intégrations simple et double de l'accélération mesurée par un accéléromètre permettent d'estimer respectivement la vitesse linéaire et le déplacement du segment sur lequel il est fixé [81]. Cependant, ces intégrations sont sujettes à des dérives (ou *drifts*), dues aux erreurs de mesures des capteurs qui s'accumulent avec le temps [182]. Ce phénomène se traduit par une augmentation avec le temps de l'écart entre le paramètre estimé et sa valeur réelle. De plus, l'accélération de la gravité terrestre (accélération verticale d'environ $9.8 m/s^2$) doit être retirée des données de l'accéléromètre [173]. Pour ce faire, l'orientation du dispositif dans le référentiel terrestre doit être estimée. C'est la raisons pour laquelle les gyroscopes et accéléromètres peuvent utilisés conjointement, parfois avec un magnétomètre (*e.g.* dans les centrales inertielles). Des algorithmes de pré-traitement des données, dits de *fusion de capteurs*, peuvent alors être utilisés pour fusionner les données de ces différents capteurs afin de corriger les *drifts*[81].

Après cette potentielle étape de pré-traitement du signal, des méthodes d'analyses des données peuvent être appliquées pour extraire des informations relatives à la démarche. Il peut s'agir de décrire les phases du cycles de marche sous la forme de paramètres spatio-temporels décrivant différents aspects des cycles de marche (*c.f.* description du cycle de marche et de ses évènements section 1.1.1.1). Les paramètres les plus fréquemment étudiés sont listés ci dessous :

- **Paramètres temporels :**
  - Durée des pas
  - Durée des cycles

  - Durée des phases du cycle (appui, balancement, simple appui, double appui)
  - Cadence du pas
- **Paramètres spatiaux :**
  - Longueur et largeur du pas
  - Longueur et largeur du cycle
  - Hauteurs maximale et minimale du pied durant la phase de balancement
  - Amplitude de rotation maximale de l'articulation de la hanche, de la cuisse, du tibia et/ou de la cheville au cours du cycle
- **Paramètres spatio-temporels :**
  - Vitesse de marche

Le calcul de ces paramètres nécessite d'identifier les évènements des cycles de marche (*e.g.* la pose du pied au sol et/ou le décollement des orteils) dans le signal renvoyé par les dispositifs numériques. Les méthodes d'identification de ces évènements peuvent être réparties en 3 catégories [50] :

(i) Les méthodes basées sur l'identification de *points caractéristiques*. Elles sont basées sur des méthodes d'identification de points remarquables dans le signal, qui sont supposés correspondre à des évènements particuliers du cycle de marche [142, 81, 85]. Les points remarquables peuvent correspondre à des changements de signes, des *extremums* locaux ou des phases stationnaires dans le signal. Leur identification est souvent basée sur des *valeurs-seuils*, et elle dépend du type de données mesurées ainsi que de la position du capteur, par exemple :
   - L'accélération mesurée au niveau du pied est *quasi*-nulle lorsque ce dernier est en contact avec le sol [93], permettant d'estimer les instants correspondant à la pose du talon et au décollement des orteils.
   - La vitesse angulaire mesurée dans le plan sagittal au niveau de la cheville change de signe au moment de la pose du pied et de son décollement [40, 76].
   - L'impact du pied au sol provoque un pic dans l'accélération verticale mesurée à la partie basse du tronc [187].

(ii) Les méthodes basées sur *un modèle* et/ou des algorithmes *adaptatifs* (*i.e.* d'apprentissage supervisé [24]). Ces deux types d'approches sont regroupés dans la même catégorie car ils nécessitent une phase d'apprentissage, soit pour générer un modèle

de cycle de marche [13] soit pour entraîner un algorithme (par exemple un modèle caché de Markov [61]). Elles requièrent donc une base d'apprentissage conséquente contenant des données dans lesquelles les différentes phases de marche sont annotées.

(iii) Les méthodes d'*analyse des propriétés dynamiques du signal*. Elles se basent sur l'auto-similarité locale du signal pour identifier les cycles de marche. Cette composante est déterminée par transformation du signal par *ondelettes* [110, 90], ou évaluée par des méthodes basées sur la corrélation ou sur la similarité de forme [162].

D'autres méthodes permettent l'étude de la *cinématique* des articulations au cours de la marche (*c.f.* section 1.1.1.1). Pour ce faire, l'angle formé entre les deux segments de l'articulation est estimé à partir des données renvoyées par deux capteurs, chacun étant placé sur l'un des deux segments [120]. Étudier l'ensemble des articulations des membres inférieurs nécessitent donc potentiellement un grand nombre de capteurs. Par exemple, les méthodes décrites par Tadano *et al.* (2013) et Narváez *et al.* (2018) [120] sont basées sur l'utilisation d'un réseau de 7 capteurs (un placé au niveau des lombaires, 2 au niveau des cuisses, deux au niveau des tibias et 2 au niveau des pieds). Elles permettent de représenter l'orientation des différents segments des membres inférieurs sous forme de quaternions unitaires.

La fiabilité et la validité de toute méthode de quantification de la marche nécessitent d'être évaluées avant son utilisation en pratique. Cette validation est réalisée en confrontant ses résultats avec la *vérité terrain*, qui peut être représentée à l'aide de différentes méthodes. Pour évaluer une méthode de détection des évènements du cycle de marche, le nombre d'évènements identifiés par la méthode (*e.g.* le nombre de cycles de marche) est comparé avec le nombre d'évènements réellement réalisés par le porteur. Ce nombre peut être connu si la méthode est appliquée sur une base de données existante [50, 162] ou si ce nombre est identifié par un dispositif non portatif (*e.g.* le tapis GaitRite [76]) ou déterminé à partir d'une vidéo prise durant l'acquisition des données [89, 113, 76]. Les critères utilisés pour cette évaluation sont basés sur le nombre de points identifiés à raison comme des évènements, appelés *vrais positifs* (TP pour *True Positives*), le nombre d'évènements non identifiés par la méthode, appelés *faux négatifs* (FN pour *False Negatives*), et le nombre de points identifiés à tort comme des évènements, appelés *faux positifs* (FP pour *False Positives*). Ces indicateurs servent à calculer deux critères :

(i) La *sensibilité* (ou *True Positive Rate*) : $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. La *sensibilité* est égale à 1 si tous les évènements sont détectés par l'algorithme.

(ii) La précision (ou *Positive Predictive Value*) : PPV = $\frac{\text{TP}}{\text{TP+FP}}$ La précision est égale à 1 si tous les points considérés comme des évènements par l'algorithme sont *True Positives*.

Une méthode avec une très grande sensibilité peut considérer à tort un grand nombre de points comme des évènements du cycle, les faux positifs n'étant pas pris en compte dans son calcul. À l'inverse, les faux négatifs n'étant pas pris en compte dans le calcul de la précision, une méthode avec une grande précision peut ne pas identifier un grand nombre d'évènements du cycle [13]. L'exactitude, représentée par la moyenne harmonique de ces deux indicateurs, dit score F_1, est donc souvent utilisée pour évaluer la performance de ces algorithmes [13, 50, 162] : $F_1 = 2 \times \frac{\text{PPV}\times\text{TPR}}{\text{PPV+TPR}}$. L'*erreur relative* est un autre critère d'évaluation. Elle correspond au rapport entre le nombre d'erreurs de l'algorithme FP+FN et le nombre d'évènements de la vérité terrain [113, 89].

Pour évaluer la précision d'une méthode d'estimation de paramètres spatio-temporels ou de la cinématique de la marche, la valeur du paramètre estimé est comparée avec celle mesurée par un dispositif considéré comme *Gold Standard* (*e.g.* tapis de pression GaitRite [40, 168] ou système d'analyse vidéo [22, 169, 120, 166]). La différence entre le paramètre estimé $\hat{p}$ et sa véritable valeur p peut être évaluée par différents critères :

— *Critère graphique* :
  — Graphique de Bland-Altman [5] : il représente la différence entre $\hat{p}$ et p en fonction de leur valeur moyenne.
— *Critères quantitatifs* :
  — *Racine de l'erreur quadratique moyenne* RMSE (pour *Root Mean Square Error*) [9, 6, 149] : RMSE = $\sqrt{\frac{\sum_{i=1}^{n}(p-\hat{p}_n)^2}{n}}$, avec n le nombre d'observations du paramètre p.
  — *Pourcentage d'erreur* [50, 168] : Err = $\frac{|\hat{p}-p|}{p} \times 100$
  — *Coefficient de corrélation linéaire* entre p et $\hat{p}$ [115, 164, 162]

Dans une revue de l'état de l'art de l'analyse du mouvement par Iosa *et al.* (2016), les auteurs indiquent que la détection des évènements du cycle de marche est plus précise en utilisant des dispositifs placés à proximité des pieds [81]. Par exemple, Hundza *et al.* (2014) décrivent une méthode basée sur l'analyse de la vitesse angulaire mesurée au niveau du tibia et permettant une détection parfaite des cycles de marche (aucun faux positif ni faux négatif) [76]. Cependant, cette tendance est à nuancer au regard des développements

récents qui permettent d'obtenir de très bonnes performances avec des données mesurées au niveau de la ceinture et/ou de la taille. Ainsi, Ghersi *et al.* (2020) présentent une méthode basée sur *l'identification de points caractéristiques* dans l'accélération mesurée au niveau de la taille avec un score F_1 proche de 0.999. Cette méthode présente de meilleures performances que deux méthodes appliquées à des données mesurées au niveau du pied : Sprager *et al.* (2018)[162] (F_1 entre 0.94 et 0.99), basée sur l'*analyse des propriétés dynamiques*, et Barth *et al.* (2015) [13] (F_1 entre 0.94 et 0.98). Il convient de relever que les données servant à l'évaluation des performance de la méthode de Ghersi *et al.* ont été mesurées auprès de sujet sains, tandis que les moins bons résultats des méthodes de Sprager *et al.* et Barth *et al.* ont été obtenus chez des sujets âgés ou atteints de maladie de Parkinson. Il est donc difficile de dégager un consensus net quant à la position du capteur et quant au type de méthode de détection des évènements du cycle de marche qui permettent d'obtenir les meilleures performances.

Les dispositifs numériques portatifs permettent donc la mesure quantitative de plusieurs aspects de la marche. Ils présentent l'avantage d'être légers, peu coûteux, sont faciles d'utilisation et peuvent potentiellement être employés pour mesurer la marche dans la vie quotidienne. Ils sont donc de bons candidats pour obtenir des mesures quantitatives de la démarche associées aux troubles de la marche de patients, par exemple atteints de la sclérose en plaques.

1.1.2 Sclérose en Plaques et mesures des troubles de la marche

La **Sclérose En Plaques** (SEP) est une maladie *neuro-dégénérative auto-immune*. Elle est causée par la dégradation de la *gaine de myéline* par les cellules du système immunitaire du patient. La gaine de myéline est une structure du système nerveux composée de cellules spécialisées (les *cellules gliales myélinisantes*) s'enroulant autour des *axones* des *neurones*. Elle joue un rôle d'isolant électrique et permet d'améliorer la conduction du signal nerveux. Sa dégradation entraîne donc une perte de la transmission de l'information nerveuse. Plusieurs fonctions neurologiques peuvent ainsi être atteintes chez le patient. Cette pathologie est dite *dégénérative* car elle persiste durant toute la vie du patient et tend à s'aggraver au cours du temps. Plusieurs formes de progression de la pathologie existent :

(i) La forme *récurrente-rémittente* caractérisée par des phases de poussée de symptômes suivies de phases de récupération partielle ou totale des fonctions neurologiques.

(ii) La forme *progressive* caractérisée par une dégradation sans récupération des fonctions neurologiques du patient.

Les fonctions neurologiques pouvant être affectées par la SEP sont très variées. Les déficits de la marche comptent parmi les handicaps les plus fréquemment observés et sont considérés par les patients atteints comme ceux ayant le plus d'impact sur la vie quotidienne [100]. C'est pourquoi l'évaluation de la marche tient une place très importante dans leur suivi médical [114]. Pour ce faire, les cliniciens utilisent traditionnellement des scores ou échelles chronométrés ou basés sur leurs observations.

1.1.2.1 Échelles et tests de marche classiques dans la SEP.

L'***Expanded Disability Status Scale*** ou EDSS [97] est l'échelle la plus largement utilisée pour évaluer l'incapacité globale des patients diagnostiqués avec la SEP [80]. L'évaluation du score EDSS est basée sur la détermination par le neurologue du degré de sévérité des déficits observés sur plusieurs fonctions neurologiques [4] :

1. *Fonction pyramidale.* Sa principale fonction concerne la contraction volontaire des muscles. Son atteinte peut provoquer une *paralysie* ou *parésie i.e.* paralysie partielle ou transitoire) des muscles, réduisant la mobilité des membres supérieurs et/ou inférieurs. Elle est notée de 0 pour une fonction normale à 4 dans le cas d'un patient tétraplégique.
2. *Fonction cérébelleuse.* Elle est responsable de la coordination des mouvements. Son atteinte se traduit par une *ataxie* (*i.e.* trouble de l'équilibre et de la coordination des mouvements), pouvant conduire à des troubles de la marche. Elle est notée de 1 pour une fonction normale à 5 pour une ataxie très sévère.
3. *Fonction du tronc cérébral.* Le tronc cérébral est responsable d'un grand nombre de fonctions différentes (cardiaque, respiratoire, motricité du visage et des yeux, élocution, déglutition *etc...*). Dans la SEP, les principales atteintes du tronc cérébral peuvent se traduire entre autre par un *nystagmus* (mouvement saccadé de l'oeil), une faiblesse ou une paralysie des muscles moteurs de l'oeil, une *dysarthrie* (difficulté à parler) ou des troubles de la déglutition. Cette fonction est évaluée sur une échelle de 0 (absence de trouble) à 5 (impossibilité de déglutir ou de parler).

4. https://www.edmus.org/fr/proj/ms_fs.html

4. *Fonction sensitive*. Elle concerne les sens comme le *toucher*, la *proprioception* (perception de la position des segments du corps dans l'espace), la *douleur*, la thermoception (perception de la chaleur), *etc...* Le déficit de ces fonctions peut apparaître chez les patients atteint de SEP, et est noté de 0 (absence de déficit) à 6 (perte de sensation pour l'ensemble du corps sous la tête).

5. *Fonction visuelle*. Elle correspond au sens de la *vue*. Une atteinte de la vue dans la SEP peut se traduire par une baisse de l'acuité visuelle et/ou un *scotome* (perte de la vue dans une partie du champ visuel). Elle est notée sur une échelle de 1 (vue normale) à 6 (acuité visuelle < 0.3 du meilleur oeil, < 0.1 de l'oeil le plus atteint, en notation de Monoyer [5]).

6. *Fonction cognitive*. Elle correspond aux capacités cognitives comme la mémoire, la concentration, l'humeur, *etc...* . Son atteinte dans la SEP est mesurée sur une échelle de 0 (normale) à 5 (démence).

7. *Fonctions sphinctériennes*. Elles correspondent à la capacité de rétention des selles et de l'urine. Leur altération peut conduire à des fuites pouvant aller jusqu'à une incontinence. Leur déficit dans la SEP est noté de 0 (fonction normale) à 6 (perte des fonctions intestinale et urinaire).

8. *Autres fonctions*. Le neurologue peut attribuer un score de 1 si une ou plusieurs fonctions neurologiques sont atteintes par la SEP.

Le score EDSS est attribué sur la base du nombre de fonctions neurologiques présentant un déficit, la sévérité de ce dernier, et/ou de la distance sur laquelle un patient peut marcher sans aide. Ainsi, un score EDSS de 0 correspond à un patient ayant toutes ses fonctions neurologiques cotées à 0 (ou 1 pour la fonction cérébelleuse). À partir de 1, le score EDSS augmente de 0.5 point jusqu'à 10. Entre 1 et 3.5, le score est attribué en fonction du nombre de fonctions neurologiques touchées et du degré de sévérité (par exemple, un score de 2.5 est attribué à un patient ayant deux fonctions neurologiques cotées à 2, et un score de 3 est attribué à un patient ayant une fonction cotée à 3, ou trois ou quatre fonctions neurologiques notés à 2). Les scores entre 4 et 6.5 sont attribués en fonctions des déficits des fonctions neurologiques du patient et de ses capacités de marche. Ainsi, les patients ayant un score de 4 ont une fonction neurologique cotée à 4, ou toute combinaison de déficits de fonctions neurologiques supérieure à celles correspondant au score 3.5, et sont capables de marcher 500 mètres sans assistance à la marche. Les patients

5. http://campus.cerimes.fr/semiologie/enseignement/esemio7/site/html/2_4.html

ayant un score de 6.5 nécessitent une assistance à la marche pour parcourir une distance de 20 mètres sans pause (et ont généralement une combinaison d'au moins deux fonctions neurologiques avec un score de sévérité 3 ou plus). À partir de 7, les patients ne peuvent plus marcher plus de 5 mètres sans assistance. L'augmentation du score correspond à la perte de fonctions permettant de se déplacer en chaise roulante, de se tenir assis, de s'alimenter et de communiquer. Le score 10 correspond au décès du patient dû à la SEP.

Malgré la forte utilisation du score EDSS en pratique, plusieurs caractéristiques lui sont reprochées. Tout d'abord, l'échelle manque de linéarité. En effet, une augmentation de 1 point ne traduit pas la même évolution de la sévérité selon le score considéré [179]. L'EDSS est également critiqué pour son manque d'objectivité, qui se traduit par une faible fiabilité inter examinateur [31]. Enfin, l'altération de la marche n'est qu'une des informations prises en compte lors de l'attribution d'un score EDSS, qui vise à donner une appréciation générale du handicap global du patient. Ainsi, deux patients ayant un score EDSS identique peuvent néanmoins présenter des troubles de la marche individuels différents. En raison de ces propriétés, l'EDSS ne peut être considéré comme une mesure quantitative de la capacité de marche.

Compte tenu des limites de l'EDSS, un groupe de travail a été créé par la National MS Society afin d'identifier des résultats quantitatifs dotés de bonnes propriétés psychométriques pour évaluer la gravité de la SEP [36]. Ce dernier propose le score ***Multiple Sclerosis Functional Composite*** comme une nouvelle échelle de mesure pour représenter l'état de santé global du patient. Il se destine principalement à être utilisé dans les essais cliniques [114]. Il se présente sous la forme d'un *z-score* calculé à partir des résultats du patient à 3 tests : le *Nine Hole Peg Test* (NHPT) testant la dextérité des membres supérieurs, le *Paced Auditory Serial Addition Test* (PASAT) testant les capacités cognitives, et le *Timed 25 Foot Walk* (T25FW) testant la vitesse de marche [45]. Si une utilisation croissante du score MSFC dans les essais cliniques est observée, il n'en reste pas moins critiqué pour l'interprétation peu intuitive du *z-score* et pour la dépendance de son calcul à la population choisie comme référence [53].

Si le score MSFC n'a donc pas été accepté par les autorités réglementaires comme critère d'évaluation principal à utiliser dans les essais cliniques, le **T25FW** utilisé seul pour évaluer la marche connaît un intérêt croissant [114]. Les instructions données au patient au cours de ce test consistent à marcher une distance de $7m60$ (25 foot) aussi vite que possible sans courir. Le patient effectue deux fois cette distance de marche en suivant ces instructions [45]. Le temps moyen de ces 2 essais est ensuite calculé. Une

pratique commune consiste ensuite à le convertir en vitesse de marche. Cette dernière a tendance à diminuer avec l'augmentation de la sévérité de la maladie du patient. La forte relation entre la vitesse de marche et l'état de santé de patient combinée à la bonne fiabilité du test et sa facilité d'exécution au cours d'une consultation en font *un critère d'évaluation de l'ambulation idéal pour la recherche et la pratique clinique dans la SEP* [114]. Ce test ne donne cependant que des informations relatives à la vitesse de marche du patient pour parcourir une distance fixe. Intuitivement, on peut supposer que des patients présentant différents types d'incapacités dues à la SEP (par exemple un défaut d'équilibre, une boiterie induite par une parésie, une fatigue chronique,*etc...*) peuvent avoir la même vitesse de marche mais différer dans leur démarche. Il doit également être réalisé à l'hôpital sous la supervision d'un clinicien.

D'autres tests de marche chronométrés sont utilisés en pratique pour l'évaluation de la SEP. Une première catégorie consiste à faire varier les consignes et conditions de test du T25FW, telles que l'allure de la marche (rapide ou naturelle), le fait de partir à l'arrêt ou non, ou la distance (10 mètres ou 30 mètres par exemple) [56]. Une autre catégorie consiste à évaluer la distance que le patient peut marcher en un temps défini (*périmètre de marche*), par exemple pendant six minutes (*Timed* 6 *Minutes Walking*) [52], ou une version de 2 minutes, qui est jugée plus faisable en pratique pour les patients avec des déficits sévères [17].

D'autres tests intègrent des activités supplémentaires à la marche pour évaluer la motricité des membres inférieurs. Dans le test chronométré du *Timed Up and Go* (TUG), le patient part en position assise, et doit se lever, marcher trois mètres, faire demi-tour puis s'asseoir à nouveau [133]. Les résultats de ce test dépendent de la mobilité générale du patient, ce qui peut être considéré comme un avantage dans certains contextes, mais aussi comme un inconvénient lorsque l'objectif est d'étudier spécifiquement la marche [17]. Le test chronométré *Six Spot Step Test* (SSST) consiste pour le patient à marcher en suivant un parcours marqué par six plots cylindriques qu'il doit déplacer hors de leur marque au sol avec un pied prédéfini [122]. Le test est répété quatre fois avec deux essais pour chaque pied. Le SSST présente une forte corrélation avec l'EDSS et le T25FW, dont il est une alternative incluant l'équilibre et la coordination. Cependant ses propriétés psychométriques, particulièrement sa sensibilité aux changements de sévérité de la pathologie, ne sont pas aussi bien établies que pour le T25FW, et il est plus long et difficile à mettre en place en condition clinique [17].

Enfin, des questionnaires peuvent être employés pour que le patient évalue sa percep-

tion de la sévérité de sa pathologie. Le *Multiple Sclerosis Walking Scales 12* (MSWS-12) est un questionnaire spécifiquement développé pour évaluer l'impact de la SEP sur les capacités de marche [69]. Il est constitué de douze questions dans lesquelles il est demandé au patient de noter sur une échelle de 1 à 5 les limitations dues à la SEP resenties pour certains aspects de la marche. Le MSWS-12 a montré une très bonne sensibilité aux changements d'état de santé du patient [69].

Ces tests et questionnaires sont tous couramment utilisés en pratique pour le suivi des patients et/ou dans un contexte de recherche clinique. Cependant, ils sont généralement critiqués pour leur manque de sensibilité aux différences entre les patients avec un stade peu avancé de la pathologie [153] et aux changements ténus des déficits de marche qui peuvent survenir durant le temps d'un essai clinique [170] Ils ne fournissent également qu'une information partielle concernant la marche, et ne permettent pas de quantifier certains aspects de la démarche du patient. La mesure de la marche peut donc encore être améliorée par des solutions fournissant des informations quantitatives, objectives et plus précises afin de servir de bio-marqueurs de la progression de la pathologie [170, 123]. Les technologies numériques permettant la mesure du mouvement offrent de telles solutions.

1.1.2.2 Analyse de la marche par dispositifs numériques dans la SEP.

De nombreuses méthodes recourant aux dispositifs numériques (tels que présentés section 1.1.1.2) sont utilisées ou en cours de développement pour l'analyse de la marche dans la SEP.

Comme pour de nombreuses autres pathologies, la *capture de mouvements par système vidéo avec marqueurs* est considérée comme *Gold Standard* [17, 102] pour l'analyse quantitative de la marche, notamment de la cinématique et cinétique. Certains systèmes vidéo sans marqueurs ont également été testés, tels que la caméra Kinect pour estimer la vitesse de marche et le déplacement du centre de gravité au cours du T25FW [15]. Parmi les tapis de capteurs de pression, le *Gaitrite* est peut être le plus fréquemment utilisé pour identifier des paramètres spatio-temporels du cycle de marche associés à la gravité de la pathologie [103, 51, 159, 160].

Plusieurs relations entre paramètres spatio-temporels et/ou cinétiques de la marche et les déficits de la marche de patients atteints de SEP et/ou la gravité globale de leur pathologie sont décrites dans la littérature. L'analyse quantitative de la marche par système à reconnaissance vidéo menée par Severini *et al.* (2017) [152] compare plusieurs paramètres spatio-temporels et cinétique de la marche entre sujets sains et 3 groupes de patients

formés en fonction de la sévérité globale de la pathologie (EDSS < 5 : forme légère, EDSS $\in [5, 5.5]$: forme modérée, EDSS ≤ 6 : forme sévère). Leurs résultats démontrent que la durée des cycles (en s) et de la phases d'appui (en pourcentage de durée du cycle) tend à augmenter avec la sévérité de la pathologie, tandis que la cadence de pas (en pas par minute), la vitesse de marche (en mètre par seconde) et la longueur des pas (en % de la hauteur du patient) tendent à diminuer. Ils démontrent également un lien entre l'amplitude des mouvements des articulations de la hanche, des genoux et des chevilles et la vitesse de marche mesurée au cours du test *Timed 6 Minutes Walking*. Les relations entre paramètres spatio-temporels et score EDSS ont également été évaluées par Lizrova Preiningerova *et al.* (2015) à l'aide du tapis GaitRite. Dans cette étude, les patients sont regroupés par niveau d'EDSS ($0-1.5$, $2-2.5$, $3-3.5$,*etc...*) [103]. Les résultats montrent que seule la vitesse de marche diffère significativement entre tous les niveaux d'EDSS. Avec la sévérité, la longueur du pas (en mètre) diminue (significativement entre 0 et 6), la durée du pas augmente (en seconde, significativement à partir de 5), ainsi que la durée de la phase de double appui (en pourcentage de la durée du cyle, significativement à partir de 3). Aucune relation entre la sévérité de la pathologie et la variabilité des paramètres spatio-temporels n'a pu être démontrée.

Les capteurs inertiels sont considérés comme une potentielle alternative aux dispositifs non portatifs pour l'analyse quantitative de la marche. Du fait de leurs caractéristiques évoquées dans la section 1.1.1.2, ils sont en effet moins coûteux, plus faciles d'utilisation et peuvent être employés hors du contexte hospitalier, pour potentiellement mesurer la marche en vie quotidienne [125, 7]. Plusieurs configurations ont été explorées, et varient par les dispositifs utilisés, leur position et le type d'information sur la marche mesurée. Leur utilisation en clinique nécessite préalablement de confronter les informations qu'ils permettent de mesurer avec les indicateurs et les tests utilisés par les neurologues en pratique. L'étude des données mesurées par des capteurs inertiels fixés sur les chaussures au cours du test T25FW démontre une différence significative dans les valeurs de paramètres spatio-temporels du cycle entre un groupe de patients sans troubles de marche (EDSS ≤ 3.5) et un groupe présentant des troubles de marche (EDSS > 3.5). Plus précisément, la longueur (en mètre) et la vitesse de la marche (en m/s) est significativement plus faible chez le groupe de patients atteints de troubles, alors que la durée du cycle (en seconde) et la durée relative de la phase d'appui (en % de la durée du cycle) tend à augmenter [46]. Ces différences sont aussi observées entre les patients avec un score EDSS > 3.5 et des individus sains. La réduction de la vitesse de marche et de la longueur des pas, ainsi que

l'augmentation de la durée proportionnelle de la phase d'appui sont également observées chez les patients sans restriction de distance maximale de marche (EDSS ≤ 3.5) et les individus sains. Les capteurs inertiels peuvent donc mesurer des informations quantitatives de la marche différenciant les patients des individus sains, y compris pour des stades peu avancés. Ces résultats sont en accord avec ceux présentés par Müller *et al.* (2021), qui identifient une différence significative entre sujets sains et patients avec une faible sévérité (EDSS entre 1.5 et 2) de la vitesse de marche, la longueur des cycles (en mètre) et la durée de la phase d'appui (en seconde) mesurée par un capteur inertiel fixé sur le dessus du pied au cours du test *Timed 6 Minutes Walking*[117]. Des résultats similaires ont été observés par Angelini *et al.* (2020) entre sujets sains, patients atteints de forme progressive modérée (EDSS compris entre 3 et 5) et forme sévère (EDSS > 5) avec un réseau de capteurs inertiels (2 au pied et un au niveau des lombaires) [7]. Enfin, Pau *et al.* (2016) démontrent une diminution significative de la vitesse de marche évaluée avec un accéléromètre placé au niveau des lombaires, chez des patients atteints de formes modérée (EDSS $2.0 - 4.0$) et sévère (EDSS $4.5 - 6.5$) par rapport à des sujet sains, et une diminution de la cadence des pas, une augmentation de la phase d'appui et de la phase de double appui chez les patients atteints d'une forme sévère par rapport aux sujets sains et aux patients atteints de forme moins avancée (EDSS $0 - 1.5$ et $2.0 - 4.0$). Toutes ces expériences ont été réalisées à partir de données mesurées dans un contexte clinique, et sont une première preuve de l'intérêt de l'utilisation des capteurs inertiels dans ce contexte. D'autres études proposent et évaluent des méthodes permettant d'estimer la vitesse de marche en contexte clinique et en vie quotidienne à partir de l'accélération mesurée au niveau des lombaires. Supratak *et al.* (2018) et Atrsaei *et al.* (2021) estiment la vitesse des patients atteints de SEP à partir de paramètres estimés à partir de segments du signal mesurés au cours de la marche (correspondant aux cycles de marche dans le premier cas et à une fenêtre de 2 secondes dans le second), en y appliquant respectivement un modèle de *Support Vector Regression* et de *régression linéaire Gaussien* [164, 9]. Les performances de la méthode proposée par Supratak *et al.* sont meilleures (Erreur d'environ 0.01 m/s entre la vitesse estimée et réelle contre 0.1 m/s), car elle construit un modèle personnalisé pour chaque patient. Cependant, seuls Atrsaei *et al.* évaluent les performances de leur méthode en la confrontant à la vitesse mesurée dans la vie quotidienne des patients par un dispositif fixé au pied.

Les systèmes de capteurs inertiels présentent donc un intérêt pour l'analyse des troubles de la marche dans la SEP en contexte clinique, et La littérature apporte les premières

preuves de la faisabilité de leur utilisation dans la vie quotidienne. En effet, de nombreuses études démontrent leur capacité à apporter des informations quantitatives associées à la gravité de la pathologie. En dehors de ces considération techniques, l'aspect cosmétique du dispositif doit aussi être pris en compte dans le développement d'une telle solution pour une bonne acceptation par le patient. Certains dispositifs ont été considérés par les patients comme trop encombrants et inconfortables [113]. Rueterbories *et al.* (2010) suggèrent notamment d'utiliser des dispositifs aussi petits et légers que possible, et ne nécessitant pas de système d'attache trop contraignant pour le patient [142].

1.2 Présentation de la solution *eGait*

L'entreprise UmanIT et le LMJL souhaitent développer un dispositif et une méthode d'analyse de la marche tirant profit de leurs expertises respectives. En s'appuyant sur les observations précédemment présentées, il a été convenu que la solution développée devait être un équilibre entre :

— Obtenir une mesure quantitative de la marche suffisamment précise pour être sensible aux déficits du porteur du dispositif.

— Utiliser un dispositif aussi ergonomique, peu invasif et peu coûteux que possible.

Ces caractéristiques sont supposées permettre de faciliter l'utilisation de la solution dans un contexte clinique pour apporter des informations quantitatives supplémentaires sur la démarche du patient. La solution doit également être conçue de telle sorte que des développements supplémentaires puissent permettre une utilisation en vie quotidienne. Le choix du type de dispositif utilisé pour la solution `eGait` repose sur les points suivants, qui ont été déduits de la revue de la littérature présentée dans section 1.1.1.2.

Tout d'abord, seuls les systèmes portatifs sont compatibles avec une utilisation en vie quotidienne. Le choix s'est donc porté vers les systèmes de capteurs inertiels portatifs. Les méthodes impliquant l'utilisation de plusieurs capteurs fixés sur plusieurs parties du corps humain permettent les représentations de la marche les plus complètes. L'orientation des différents segments du corps humains au cours du temps peut être déterminée, permettant ainsi de calculer la cinématique de la démarche et plusieurs paramètres spatio-temporels. Cependant, ces méthodes nécessitent de placer des dispositifs sur plusieurs segments du corps, ce qui complique la mise en place. Il a donc été décidé d'utiliser un seul dispositif, ce qui nécessite de définir sa localisation sur le corps du porteur. Le choix de la position

du capteur est basé sur des critères ergonomiques/esthétiques et sur la caractéristique de la marche à quantifier.

Le positionnement proche des pieds augmente la précision de détection des évènements du cycle de marche [81]. Cependant, il nécessite des systèmes d'attache tels que des élastiques ou des sangles qui peuvent être inconfortables pour le porteur, et l'aspect esthétique des capteurs placés au niveau des chaussures/pieds a été décrit comme un facteur limitant de l'observance des patients [142]. En prenant ces considérations en compte, le choix s'est porté sur un unique capteur inertiel placé à la ceinture. Cette position est supposée facile d'accès, et peu invasive. La placement du capteur ne nécessite rien d'autre qu'une ceinture, qui est un accessoire fréquemment utilisé. Le dispositif est donc discret et peut même être caché par un vêtement. L'intégration d'un dispositif numérique à une ceinture est une approche qui a déjà été adoptée, par exemple avec le dispositif Actibelt [38].

Les méthodes d'analyse de la marche à partir d'un dispositif placé au niveau de la taille ou de la hanche se basent le plus souvent sur l'étude de l'accélération, avec un système placé en position centrale et proche du centre de gravité (*cf* section 1.1.1.2). Il n'est cependant pas possible d'obtenir des informations relatives aux rotations de la partie basse du corps durant la marche. Pour ces raisons, il a été choisi d'étudier les rotations de la hanche en plaçant le dispositif en position latérale. En effet, la hanche étant mobilisée pendant la marche, on peut supposer que l'étude de ses mouvements permet de quantifier plusieurs aspects de la démarche de l'individu. De plus, il a été démontré que la cinématique de l'articulation de la hanche est affectée par la Sclérose En Plaques [152]. Pour permettre la comparaison des données mesurées chez différents individus, il a été convenu de positionner le capteur au niveau de la *hanche droite*. Ce choix est arbitraire, et aurait pu tout aussi bien être la hanche gauche.

Pour étudier cet aspect du mouvement, l'orientation du système de capteurs est considérée comme représentative de celle de la hanche droite, et doit être estimée au cours du temps. Pour limiter les problèmes liés à la *dérive* due à l'intégration des seules données du gyroscope (*cf* section 1.1.1.2), l'orientation est déterminée par algorithme de *fusion de capteurs* permettant d'obtenir le *vecteur d'orientation absolue* du dispositif. Cette information peut être obtenue sous plusieurs formes : (i) *matrice de rotation*, (ii) *angles d'Euler* et (iii) *quaternions unitaires*. Les quaternions unitaires présentent l'avantage d'être moins coûteux en espace de stockage et en puissance de calcul que les deux autres types de représentation, et ne sont pas sujets au problème du blocage de cardan [176]. La section

1.2.1 présente le dispositif utilisé pour mesurer les données de marche qui seront analysées dans la suite de ce manuscrit. Le format mathématique des données mesurées est présenté dans la section 1.2.2

1.2.1 Description du dispositif utilisé : *MetaMotionR* (MMR)

Le système de capteurs *MetaMotionR* (MMR) développé par Mbientlab [6] est un dispositif de type *centrale inertielle*. Il rassemble dans un même boîtier en plastique de dimension 27mm × 27mm × 4mm les capteurs suivants :

— BMI160 3-axis Accéléromètre

— BMI160 3-axis Gyroscope

— BMM150 3-axis Magnétomètre

— BMP280 Baromètre/Pression/Altimètre

— BMP280 Température

— LTR-329ALS Luminosité/lumière ambiante

Seuls les accéléromètres, gyromètres et magnétomètres seront utilisés pour mesurer les données dans la suite de ce manuscrit. Ces derniers sont alignés mutuellement, formant le référentiel du dispositif défini par les axes (s_1, s_2, s_2). L'assemblage des capteurs permet d'aligner le référentiel du dispositif selon la géométrie du boîtier qui les contient. Sur la figure 1.3, le boîtier est représenté en haut, et la puce du système de capteur en bas. Le référentiel du dispositif est schématisé par les flèches bleues. L'axe s_3 étant dirigé vers l'observateur, il est matérialisé par un point entouré d'un cercle.

Le système de capteurs embarque également l'algorithme BOSCH 9-axis IMU Sensor Fusion [7]". Cet algorithme permet de fusionner les données mesurées par les accéléromètres, gyroscopes et/ou magnétomètres. Quatre modes sont disponibles : (i) `NDoF` calcule l'orientation absolue à partir des données de l'accéléromètre, du gyroscope et du magnétomètre, (ii) `IMUPlus` calcule l'orientation relative dans l'espace à partir des données de l'accéléromètre et du gyroscope, (iii) `Compass` détermine la direction géographique à partir du champ magnétique terrestre et (iv) `M4G` qui est similaire à `IMUPlus`, mais remplace le gyroscope par le magnétomètre. Le mode `NDoF` est choisi pour obtenir le vecteur d'orientation absolu du dispositif sous la forme de *quaternion unitaire*. La méthode de calcul de

6. https://mbientlab.com/metamotionr/

7. https://www.bosch-sensortec.com/bst/products/motion/fusionlibsoftware/overview_fusionlibsoftware

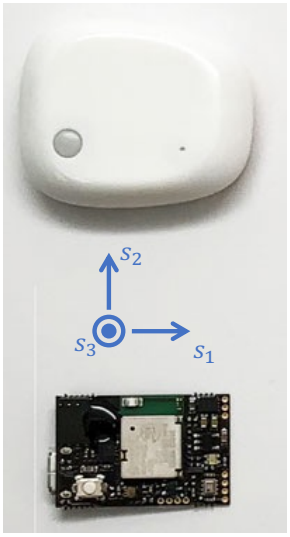


FIGURE 1.3 – Dispositif MMR et son référentiel

l'orientation par l'algorithme de *fusion de capteurs* étant exécutée par le dispositif, elle n'est pas détaillée dans ce manuscrit.

Le dispositif MMR est contrôlé via *Bluetooth* avec un *smartphone* par une application développée par UmanIT pour système d'exploitation *Android*. Le taux d'échantillonnage est fixé à 100 Hz. Le système MMR transfère toutes les 10 ms son orientation absolue sous forme de *quaternion unitaire* au *smartphone* qui les enregistre dans un tableau au format *.csv*. L'algèbre des quaternions unitaires et leur relation avec les rotations en trois dimensions sont présentées dans la section suivante.

1.2.2 Format mathématique

1.2.2.1 Algèbre des quaternions unitaires

Les quaternions sont des nombres hypercomplexes décrits pour la première fois par Hamilton en 1844 [43]. La présentation détaillée des quaternions et de leur algèbre est donnée dans de nombreux ouvrages [32, 176]. Dans cette section, seuls les concepts utiles à la compréhension de la suite du manuscrit sont abordés.

Formule générale des quaternions. Les quaternions sont des vecteurs à 4 dimensions de l'espace noté $\mathbb{H}$:

$$\mathbf{q} = (w, x, y, z)^\top , \tag{1.1}$$

La spécificité des quaternions réside dans les propriétés de leur algèbre. En effet, un

quaternion $\mathbf{q}$ peut également être considéré comme un nombre hypercomplexe de rang 4 de formule :

$$\mathbf{q} = w + ix + jy + kz, \tag{1.2}$$

où i, j et k généralisent le nombre imaginaire i selon la règle

$$i^2 = j^2 = k^2 = ijk = -1. \tag{1.3}$$

Enfin, une autre notation d'un quaternion peut être donnée en séparant sa *partie scalaire* (ou *réelle*) w de sa *partie vectorielle* (ou *imaginaire*) $\mathbf{v} = (x, y, z)^\top$:

$$\mathbf{q} = \left(w, \mathbf{v}^\top\right)^\top. \tag{1.4}$$

La fonction sélective $\mathrm{Re}(\mathbf{q}) = w$ (resp. $\mathrm{Im}(\mathbf{q}) = \mathbf{v}$) renvoie la partie réelle de $\mathbf{q}$ (resp. sa partie imaginaire).

Produit de Hamilton. La règle (1.3) implique que le produit de deux quaternions, appelé *produit de Hamilton*, n'est pas commutatif. Ainsi, soient deux quaternions $\mathbf{q}_1 = w_1 + ix_1 + jy_1 + kz_1$ et $\mathbf{q}_2 = w_2 + ix_2 + jy_2 + kz_2$, le *produit de Hamilton* de $\mathbf{q}_1$ et $\mathbf{q}_2$ vaut :

$$\begin{aligned}\mathbf{q}_1\mathbf{q}_2 =& w_1w_2 - x_1x_2 - y_1y_2 - z_1z_2 \\ &+(w_1x_2 + x_1w_2 + y_1z_2 - z_1y_2)i \\ &+(w_1y_2 - x_1z_2 + y_1w_2 + z_1x_2)j \\ &+(w_1z_2 + x_1y_2 - y_1x_2 + z_1w_2)k\end{aligned} \tag{1.5}$$

Quaternion conjugué. Comme pour les nombres complexes, le *conjugué* d'un quaternion $\mathbf{q} = (w, x, y, z)^\top$, noté $\mathbf{q}^t$, désigne le quaternion dont la partie imaginaire est opposée à $\mathbf{q}$:

$$\mathbf{q}^t = (w, -\mathbf{v}^\top)^\top. \tag{1.6}$$

Norme. D'après le produit de Hamilton (équation (1.5)), le produit d'un quaternion par son conjugué vaut

$$\mathbf{q}\mathbf{q}^t = w^2 + x^2 + y^2 + z^2,$$

Permettant de définir la *norme* d'un quaternion :

$$\|\mathbf{q}\| = \sqrt{\mathbf{q}\mathbf{q}^t} = \sqrt{w^2 + x^2 + y^2 + z^2}. \tag{1.7}$$

$\|\mathbf{q}\|$ correspond à la norme euclidienne associée au produit scalaire sur $\mathbb{R}^4$.

Quaternion inverse. Tout quaternion non nul admet un *quaternion inverse* :

$$\mathbf{q}^{-1} = \frac{\mathbf{q}^t}{\|\mathbf{q}\|^2} \tag{1.8}$$

Ainsi, le produit d'un quaternion par son inverse vaut :

$$\mathbf{q}\mathbf{q}^{-1} = \frac{\mathbf{q}\mathbf{q}^t}{\|\mathbf{q}\|^2} = 1 \tag{1.9}$$

Exponentiel et logarithme d'un quaternion. L'exponentiel d'un quaternion $\mathbf{q}$ vaut :

$$\exp(\mathbf{q}) = \exp(w)\left(\cos(\|\mathbf{v}\|), \frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\sin(\|\mathbf{v}\|)\right)^\top \tag{1.10}$$

Le logarithme d'un quaternion $\mathbf{q}$ vaut :

$$\ln(\mathbf{q}) = \left(\ln(\|\mathbf{q}\|), \frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\arccos\frac{w}{\|\mathbf{q}\|}\right)^\top \tag{1.11}$$

Puissance d'un quaternion Un quaternion unitaire élevé à la puissance $p \in \mathbb{R}$ vaut :

$$\mathbf{q}^p = \exp\left(\ln(\mathbf{q}) \times p\right) \tag{1.12}$$

Quaternion unitaire. Le groupe des quaternions unitaires rassemble les quaternions de norme 1. On le notera $\mathbb{H}_u = \{\mathbf{q} \in \mathbb{H}; \|\mathbf{q}\| = 1\}$. La formule générale d'un quaternion unitaire est :

$$\mathbf{q} = \cos\frac{\theta}{2} + \mathbf{u}\sin\frac{\theta}{2} = \left(\cos\frac{\theta}{2}, u_1\sin\frac{\theta}{2}, u_2\sin\frac{\theta}{2}, u_3\sin\frac{\theta}{2}\right)^\top, \tag{1.13}$$

avec

$$\theta = 2\arctan2(\|\mathbf{v}\|, w), \tag{1.14}$$

un scalaire et

$$\mathbf{u} = u_1 + u_2 + u_3 = \frac{\mathbf{v}}{\|\mathbf{v}\|}, \tag{1.15}$$

un vecteur unitaire.

Forme polaire et logarithme des quaternions unitaires. On remarque une simila-

rité de l'équation (1.13) avec l'équation d'Euler pour les nombres complexes unitaires $\exp(i\theta/2) = \cos(\theta/2) + i\sin(\theta/2)$. En effet, il est possible de généraliser cette équation aux quaternions unitaires, qui peuvent être écrits sous leur forme polaire :

$$\mathbf{q} = \exp\left(\mathbf{u}\frac{\theta}{2}\right) = \cos\frac{\theta}{2} + \mathbf{u}\sin\frac{\theta}{2} \tag{1.16}$$

Le logarithme d'un quaternion unitaire se déduit simplement de la forme polaire :

$$\ln(\mathbf{q}) = \ln\left(\exp\left(\mathbf{u}\frac{\theta}{2}\right)\right) = \mathbf{u}\frac{\theta}{2} = \left(0, \frac{\theta}{2}u_1, \frac{\theta}{2}u_2, \frac{\theta}{2}u_3\right)^\top \tag{1.17}$$

La transformation logarithmique est une application entre l'espace des quaternions unitaires $\mathbb{H}_u$ et l'espace tangent $\mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u) \subseteq \mathbb{R}^3$ au point $\mathbf{q}_0 = (1,0,0,0)$[131]. Du fait de la non commutativité du produit de Hamilton, certaines identités habituellement associées aux fonctions exponentielle et logarithmique ne sont cependant pas respectées. Ainsi, $\ln(\mathbf{q}_1\mathbf{q}_2)$ n'est pas toujours égal à $\ln(\mathbf{q}_1) + \ln(\mathbf{q}_2)$, de même que le produit $\exp(\mathbf{q}_1)\exp(\mathbf{q}_2)$ n'est pas toujours égal à $\exp(\mathbf{q}_1 + \mathbf{q}_2)$.

Quaternions unitaires et rotations en 3D. Une représentation naturelle d'une rotation dans un espace à 3 dimensions est donnée par son **angle de rotation** $\theta \in [0, 2\pi]$ et son **axe de rotation** $\mathbf{u} \in \mathcal{S}^2$, où $\mathcal{S}^2$ est la 2-sphère. Il existe une relation entre l'axe et l'angle d'une rotation et sa représentation *quaternionique*. En effet, le quaternion représentant la rotation directe d'angle θ autour d'un vecteur unitaire $\mathbf{u}$ est obtenu par l'équation (1.13).

L'ensemble des quaternions unitaires $\mathbb{H}_u$ forme un groupe multiplicatif, sous-groupe de $\mathbb{H}^\times = \mathbb{H} \setminus \{0\}$. Il forme un groupe de Lie isomorphe au groupe unitaire spécial SU(2), lequel étant exactement deux fois plus grand que le groupe spécial orthogonal SO(3) des matrices de rotation à 3 dimensions. En effet, les quaternions $\mathbf{q}$ et $-\mathbf{q}$ représentent la même rotation. L'algèbre des quaternions unitaires peut donc être considérée comme une structure de groupe particulière sur la $3-$sphère $\mathcal{S}^3 \subset \mathbb{R}^4$ [82].

Le groupe $\mathbb{H}_u$ est doté de l'élément neutre $\mathbf{q}^{(0)} = (1,0,0,0)^\top$ tel que $\mathbf{q}\mathbf{q}^{(0)} = \mathbf{q}^{(0)}\mathbf{q} = \mathbf{q}$. $\mathbf{q}^0$ représente "la" rotation identité, cette dernière désignant toute rotation d'angle 0 (ou 2π), quel que soit son axe.

Étant de norme 1, le conjugué d'un quaternion unitaire est égal à son inverse. En recourant à la représentation axe-angle, on prouve facilement que $\mathbf{q}^{-1}$ est associé à

la rotation de même axe que $\mathbf{q}$ mais avec un angle opposé de $-\theta$, ou indistinctement la rotation de même angle que $\mathbf{q}$ autour de l'axe opposé $-\mathbf{u}$.

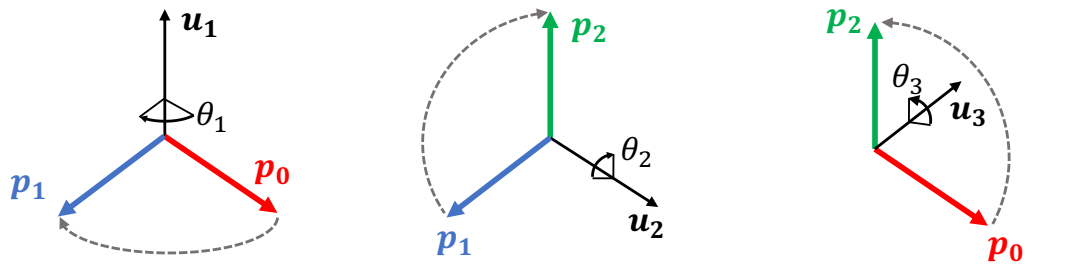
Composer des rotations avec des quaternions unitaires. Soit $\mathbf{p_0} \in \mathbb{R}^3$ un vecteur de dimension 3. La rotation codée par $\mathbf{q}_1$ transforme $\mathbf{p_0}$ en un nouveau vecteur $\mathbf{p_1}$ donné par :

$$\begin{pmatrix} 0 \\ \mathbf{p}_1 \end{pmatrix} = \mathbf{q}_1 \begin{pmatrix} 0 \\ \mathbf{p_0} \end{pmatrix} \mathbf{q}_1^{-1}. \tag{1.18}$$

La rotation codée par $\mathbf{q}_1$ est illustrée sur la figure 1.4a en utilisant sa représentation axe-angle. De même, la figure 1.4b illustre l'effet de l'application d'une autre rotation $\mathbf{q}_2$ à $\mathbf{p}_1$ pour obtenir $\mathbf{p}_2$. Selon l'équation (1.18), on peut écrire :

$$\begin{pmatrix} 0 \\ \mathbf{p}_2 \end{pmatrix} = \mathbf{q}_2 \begin{pmatrix} 0 \\ \mathbf{p_1} \end{pmatrix} \mathbf{q}_2^{-1} = \mathbf{q}_2 \left(\mathbf{q}_1 \begin{pmatrix} 0 \\ \mathbf{p_0} \end{pmatrix} \mathbf{q}_1^{-1} \right) \mathbf{q}_2^{-1} = (\mathbf{q}_2\mathbf{q}_1) \begin{pmatrix} 0 \\ \mathbf{p_0} \end{pmatrix} (\mathbf{q}_2\mathbf{q}_1)^{-1}. \tag{1.19}$$

L'équation (1.19) montre que le *produit de Hamilton* de deux quaternions unitaires $\mathbf{q}_3 = \mathbf{q}_2\mathbf{q}_1$ est équivalent à l'application de la rotation $\mathbf{q}_1$ suivie de la rotation $\mathbf{q}_2$, tel que présenté figure 1.4c.



(a) Rotation de $\mathbf{p}_0$ par $\mathbf{q}_1$ (b) Rotation de $\mathbf{p}_1$ par $\mathbf{q}_2$ (c) Rotation de $\mathbf{p}_0$ par $\mathbf{q}_3 = \mathbf{q}_2\mathbf{q}_1$

FIGURE 1.4 – Composition de rotations par produit de quaternions unitaires.

Définition d'une distance propre aux quaternions unitaires . La définition d'une métrique permettant de représenter la distance entre 2 quaternions unitaires $d(\mathbf{q}_1, \mathbf{q}_2)$ doit respecter les propriétés suivantes [77] :

- *non-négativité* $d : \mathbb{H}_u \times \mathbb{H}_u \to \mathbb{R}^+$
- *identité* : $\forall (\mathbf{q}_1, \mathbf{q}_2) \in \mathbb{H}_u^2, d(\mathbf{q}_1, \mathbf{q}_2) = 0 \Leftrightarrow \mathbf{q}_1 = \mathbf{q}_2$
- *symétrie* : $\forall (\mathbf{q}_1, \mathbf{q}_2) \in \mathbb{H}_u^2, d(\mathbf{q}_1, \mathbf{q}_2) = d(\mathbf{q}_2, \mathbf{q}_1)$

— *inégalité triangulaire* : $\forall(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3) \in \mathbb{H}_u^3, d(\mathbf{q}_1, \mathbf{q}_3) \leq d(\mathbf{q}_1, \mathbf{q}_2) + d(\mathbf{q}_2, \mathbf{q}_3)$

— *Respect de la topologie de* $SO(3)$ (voir Huyngh 2009 [77] pour plus d'informations sur cette propriété)

— *bi-invariance* : $\forall(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}) \in \mathbb{H}_u^3, d(\mathbf{q}\mathbf{q}_1, \mathbf{q}\mathbf{q}_2) = d(\mathbf{q}_1, \mathbf{q}_2)$ et $d(\mathbf{q}_1\mathbf{q}, \mathbf{q}_2\mathbf{q}) = d(\mathbf{q}_1, \mathbf{q}_2)$

Huynh (2009) [77] et Piorek *et al.* (2020) [132] présentent plusieurs fonctions candidates pour définir la distance entre deux rotations de l'espace à 3 dimensions :

— distance basée sur la norme euclidienne :

$$d_1(\mathbf{q}_1, \mathbf{q}_2) = \min(\|\mathbf{q}_1 - \mathbf{q}_2\|, \|\mathbf{q}_1 + \mathbf{q}_2\|) \tag{1.20}$$

— distance basée sur le produit scalaire :

$$d_2(\mathbf{q}_1, \mathbf{q}_2) = \arccos\left(|\mathbf{q}_1 \cdot \mathbf{q}_2|\right), \tag{1.21}$$

où $\cdot\cdot$ représente le produit scalaire de deux vecteurs (il ne s'agit pas du *produit de Hamilton* de deux quaternions qui produit un autre quaternion).

— distance basée sur le produit scalaire après élimination de la fonction cosinus inverse :

$$d_3(\mathbf{q}_1, \mathbf{q}_2) = 1 - (|\mathbf{q}_1 \cdot \mathbf{q}_2|) \tag{1.22}$$

— distance géodésique (telle que proposée par Piorek *et al.* (2020) [132] :

$$d_4(\mathbf{q}_1, \mathbf{q}_2) = 2 \arccos \mathrm{Re}\left(\mathbf{q}_1^{-1}\mathbf{q}_2\right), \tag{1.23}$$

Elle correspond à la longueur minimale d'une ligne géodésique reliant les deux quaternions sur la 3−sphère [82].

La distance géodésique d_4 présente l'avantage d'avoir une interprétation "physique" relativement simple. En effet, le quaternion $\mathbf{q}_1^{-1}\mathbf{q}_2$ représente la rotation nécessaire pour obtenir $\mathbf{q}_2$ à partir de $\mathbf{q}_1$ [165]. Sa partie réelle correspond donc à l'angle associé à cette rotation, ou en traduisant Piorek dans Analysis of Chaotic Behavior in Non-linear Dynamical Systems : *"[La distance géodésique] peut également être considérée comme la quantité d'énergie ou de rotation nécessaire pour faire tourner le quaternion* $\mathbf{q}_1$ *jusqu'à la rotation définie par le quaternion* $\mathbf{q}_2$*"* [131]. La distance géodésique a également été utilisée par Jablonksi [82] pour généraliser la dissimilarité

Dynamic Time Warping aux séries chronologiques de quaternions unitaires, et pour l'appliquer à la classification non supervisée de données de mouvements [82].

Dans la suite de ce manuscrit, la *distance géodésique* définie par l'équation 1.23 est donc utilisée pour calculer la distance entre 2 quaternions unitaires.

Interpolation entre deux quaternion unitaires L'interpolation sur la ligne géodésique connectant deux quaternions unitaires est permise par l'algorithme *slerp* (pour *Spherical Linear intERPolation*) [155] :

$$slerp(\mathbf{q}_1, \mathbf{q}_2, p) = \mathbf{q}_1(\mathbf{q}_1^{-1}\mathbf{q}_2)^p = \mathbf{q}_1 \exp(p \ln(\mathbf{q}_1^{-1}\mathbf{q}_2)), \tag{1.24}$$

où $p \in [0, 1]$ est le paramètre déterminant le point d'interpolation entre $\mathbf{q}_1$ et $\mathbf{q}_2$. Il a une interprétation physique relativement simple. En effet, $\mathbf{q}_1^{-1}\mathbf{q}_2$ est la rotation complète pour obtenir $\mathbf{q}_2$ à partir de $\mathbf{q}_1$. Si l'on considère l'exemple dans lequel $p = 1/3$, $\mathbf{q}_1(\mathbf{q}_1^{-1}\mathbf{q}_2)^p$ correspond donc au quaternion obtenu après avoir effectué un tiers de la rotation complète entre $\mathbf{q}_1$ et $\mathbf{q}_2$.

Quaternion unitaire moyen. Le quaternion unitaire moyen $\bar{\mathbf{q}}$ d'un ensemble de N quaternions unitaires $\{\mathbf{q}_i\}_{i=1,2,\ldots,N}$ ne peut pas être calculé par la moyenne algébrique classique $\left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbf{q}_i$. En effet, le quaternion ainsi obtenu n'est pas unitaire et le fait que $\mathbf{q}_i$ et $-\mathbf{q}_i$ représentent la même rotation n'est pas pris en compte. Markley *et al.* proposent en 2007 une méthode permettant d'estimer $\bar{q}$ à partir de l'équation suivante [108] :

$$\text{avg}_{\text{Markley}}(\mathbf{q}_1, \ldots, \mathbf{q}_N) = \underset{\mathbf{q}\in\mathbb{H}_u}{\arg\max}\, \mathbf{q}^\top M \mathbf{q}, \tag{1.25}$$

avec M la matrice de dimension 4×4 :

$$M = \sum_{i=1}^{N} \mathbf{q}_i \mathbf{q}_i^T, \tag{1.26}$$

Il est à noter que les quaternions $\mathbf{q}_i$ sont considérés comme des vecteurs de $\mathbb{R}^4$ dans cette méthode, et les règles du produit matriciel sont respectées dans les équations (1.25) et (1.26). La solution de l'équation (1.25) est donnée par le *vecteur propre* de M associé à la plus *forte valeur propre*.

Dans la suite de ce manuscrit, par souci d'optimisation de certains algorithmes utilisés pour l'analyse des données, on supposera que l'exponentiel de la moyenne

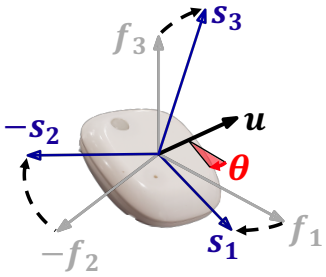
arithmétique des log-quaternions unitaires est une approximation suffisante de $\bar{\mathbf{q}}$:

$$\text{avg}(\mathbf{q}_1, \ldots, \mathbf{q}_N) = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \ln(\mathbf{q}_i)\right). \tag{1.27}$$

Remarque : L'équation (1.27) est supposée une bonne approximation de $\bar{\mathbf{q}}$ *dans la mesure où les quaternions* $\mathbf{q}_1, \ldots, \mathbf{q}_N$ *restent proches du point* $(1, 0, 0, 0)$.

1.2.2.2 Mesure de l'orientation de la hanche avec le dispositif MMR

L'algorithme *SensorFusion* détermine l'orientation du dispositif MMR à un moment donné comme la rotation entre un référentiel fixe (aligné avec le système de coordonnées de la Terre) $\mathfrak{R}_f = (\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$ vers le référentiel propre du système de capteurs $\mathfrak{R}_s = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3)$. Cette rotation est décrite par deux paramètres : son axe de rotation $\mathbf{u} = (u_1, u_2, u_3)^\top$ et son angle de rotation θ. La figure 1.5a représente un exemple de rotation entre le référentiel $\mathfrak{R}_s$ et $\mathfrak{R}_f$ (par souci de lisibilité, les axes $-s_2$ et $-f_2$ sont matérialisés sur la figure. Ils sont opposés aux axes s_2 et f_2 des référentiels $\mathfrak{R}_s$ et $\mathfrak{R}_f$, qui seraient en partie confondus avec $\mathbf{u}$ et θ sur la droite de la figure).



(a) Orientation de $\mathfrak{R}_s$ dans $\mathfrak{R}_f$

(b) Localisation du dispositif *eGait*

FIGURE 1.5 – Dispositif *eGait* et orientation de la hanche

L'orientation du système de capteurs est enregistrée à un instant t sous la forme d'un *quaternion unitaire* calculé à partir de $\mathbf{u}$ et θ selon l'équation (1.13). Le dispositif enregistre son orientation à une fréquence $F = 100$ Hertz. Les données se présentent donc comme une séquence de N quaternion unitaire $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$ associés à leur temps de mesure $T = (t_1, \ldots, t_N)$. Par la position du système de capteurs au niveau de la ceinture (voir figure 1.5b), ces données représentent le changement d'orientation de la hanche au cours du temps. Par conséquent, pour une acquisition durant laquelle le porteur du dispositif marche, les données constituent une mesure des mouvements de rotation de la hanche

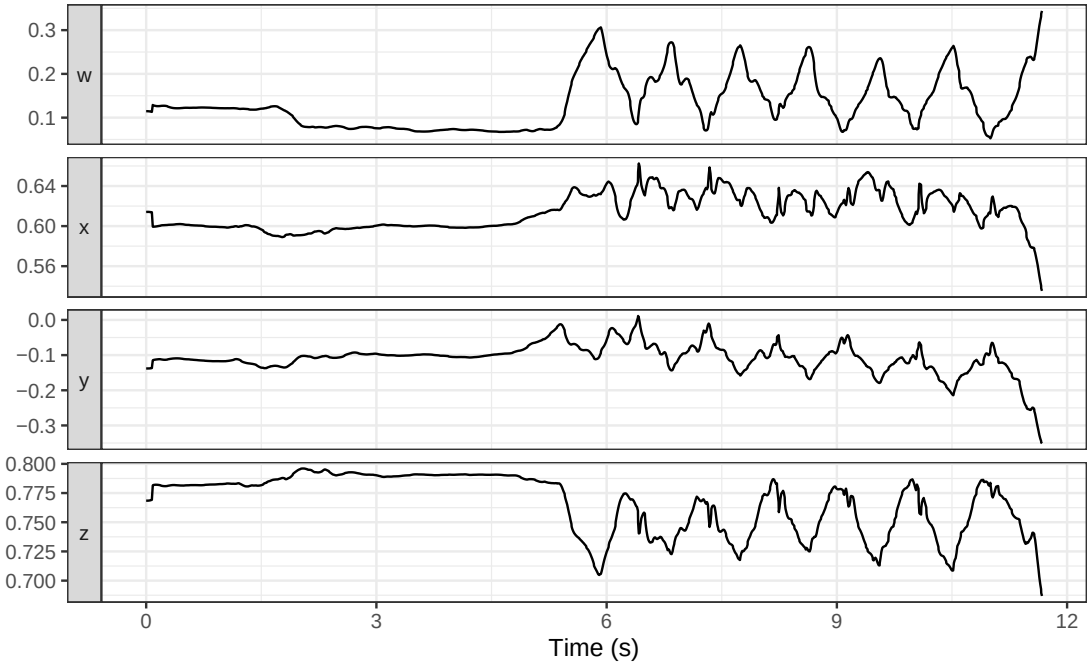


FIGURE 1.6 – Exemple de données de marche

durant l'ambulation.

Un exemple de données est présenté figure 1.6. Durant leur acquisition, le volontaire est resté en position statique pendant une période d'environ 6 secondes avant de marcher une distance d'environ 7m60. Les courbes représentent l'évolution des valeurs prises par les composantes w_i, x_i, y_i et z_i du *quaternion unitaire* $\mathbf{q}_i$ représentant l'orientation de sa hanche à un instant t_i. Durant la phase statique, les valeurs de w_i, x_i, y_i et z_i varient peu, ce qui est attendu, le volontaire n'étant pas en mouvement. La phase de marche débute peu avant la 6^{eme} seconde du jeu de données. Les valeurs de w_i, x_i, y_i et z_i varient alors de manière quasi-cyclique comme escompté. 5 ou 6 cycles peuvent être observés entre le début et la fin du jeu de données. Le volontaire entame un demi-tour juste avant la fin du jeu de données, ce qui se traduit notamment par la dérive des courbes observés sur les composantes w, x et y.

La section suivante décrit les bases de données de marche acquises auprès de volontaires sains et de patients atteints de SEP avec le dispositif MMR fixé à leur ceinture. Les protocoles utilisés pour leur acquisition sont également présentés. Ces bases ont permis le développement de méthodes d'analyse permettant d'obtenir des informations quantitatives sur la démarche du porteur. Elles permettent également d'étudier la relation entre ces informations et les caractéristiques cliniques décrivant les déficits de la marche des

patients atteints de SEP. Ces méthodes et analyses sont présentées dans les chapitres 2 et 3 de ce manuscrit.

1.2.3 Présentation des bases de données de travail

Les deux bases de données qui permettent d'obtenir les résultats présentés dans la suite du manuscrit sont constituées des données mesurées par le dispositif MMR. Elles sont présentées dans la suite de cette section, ainsi que le protocole qui a permis leur mesure.

1.2.3.1 BDDtest

Cette première base de données a été constituée pour répondre à deux objectifs :

— Permettre le développement et l'évaluation de l'algorithme de traitement des données de marche, *i.e.* permettant la segmentation du jeu de données renvoyé par le dispositif MMR en cycle de marche. Cet algorithme nommé `STRIPAGE` est présenté dans la section 2.1.

— Évaluer la relation entre les données de marche et la présence d'un trouble de la marche simulé.

Elle est constituée des données de marche mesurées par le dispositif *MMR* (*i.e.* des séquences de quaternions unitaires) auprès de 27 volontaires sains de sexe, âge, taille et corpulence variés. Aucun ne présente de pathologie connue affectant la marche à la date de leur acquisition. La période d'acquisition s'étend d'Août à Octobre 2020. (*Remarque : d'autres données de marche ont été acquises avant la constitution de la* `BDDtest`*. Elles ont permis d'expérimenter plusieurs méthodes pour la prise en main des données, et le développement de versions antérieures de l'algorithme* `STRIPAGE`*. C'est cependant pour la constitution de la* `BDDtest` *que le protocole d'acquisition a été fixé de façon rigoureuse. De plus, la version finale de l'algorithme* `STRIPAGE` *décrite dans la suite de ce manuscrit a été finalisée et évaluée à partir de ces données. C'est pourquoi elle est la seule base de données de marche mesurées chez des individus sains présentée dans ce manuscrit.*)

Chacun des volontaires a réalisé un test durant lequel il marche une distance d'environ $7,60$ mètres le plus vite possible et sans courir. Les volontaires ont parcouru cette distance deux fois dans des conditions identiques au cours d'un même test. Ces consignes correspondent au test neurologique *Time 25 Foot Walk* (T25FW) utilisé en pratique pour évaluer la marche des patients atteints de Sclérose en plaques [114] (*cf* section 1.1.2.1).

Ce test a été réalisé dans deux conditions, l'une en marche libre et l'autre en marche contrainte. Pour la seconde, les volontaires portent une orthèse bloquant complètement l'articulation du genou, afin de simuler un déficit mécanique de la marche. Simuler un déficit de la marche par le port d'une attelle est une méthode déjà rencontrée dans la littérature [19]. La base de données complète, constituée des données de marche mesurées pour chaque volontaire en condition de marche libre et contrainte, est appelée `BDDtest`.

11 volontaires ont été filmés à l'aide de la caméra d'un smartphone durant les acquisitions de données, afin de contrôler *a posteriori* le nombre de cycles de marche qu'ils ont réalisés. Les données de marche mesurées par le dispositif MMR chez ces volontaires ainsi que le nombre de cycles de marche observés par vidéo constituent une sous-partie de la base de données appelée `BDDapp`. Elle permet de comparer les évènements détectés par l'algorithme de détection des cycles avec la réalité terrain afin d'améliorer leur performance.

1.2.3.2 BDDsep

Cette base est constituée de données de marche mesurées avec le dispositif MMR auprès de patients atteints de Sclérose En Plaques (SEP) ainsi que de données cliniques décrivant la sévérité de leur pathologie. Ces patients ont donné leur accord pour rejoindre la cohorte OFSEP-HD [8] de l' *Observatoire Français de la Sclerose En Plaques* (OFSEP [9]) et pour participer à une étude clinique organisée par les Professeurs Pierre-Antoine Gourraud et David-Axel Laplaud, responsable de l'équipe Neurologie du Centre d'Investigation Clinique (CIC) du Centre Hospitalier Universitaire de Nantes [10].

Cette étude principale, qui sera désignée par *étude MYO*, avait pour but d'évaluer les déficits neurologiques des patients par l'utilisation du bracelet électronique MYO [11] permettant d'enregistrer l'activité électrique des muscles. L'étude MYO intégrait la détermination du score EDSS des patients inclus, ainsi que l'évaluation de leur capacité de marche par le test T25FW (*cf* section 1.1.2.1). La mesure de leur données de marche par le dispositif MMR a pu s'intégrer facilement dans ce protocole sous la forme d'une *étude ancillaire*. Une étude dite *ancillaire* permet d'explorer une problématique annexe au projet de recherche principal d'une étude clinique. Elle est généralement réalisée sur un sous-effectif de la cohorte recrutée dans l'étude principale. Elle ne doit pas impliquer de

8. http://www.ofsep.org/en/hd-cohort
9. https://www.ofsep.org/en/
10. http://cicnantes.fr/fr/equipe-neurologie.html
11. https://fr.aliexpress.com/item/32860368130.html

modification majeure du protocole. Une demande d'amendement au protocole de l'étude MYO a donc pu être déposée auprès du CHU de Nantes pour intégrer une étude ancillaire. Son protocole implique la mesure des données de marche de 30 patients par le dispositif MMR durant le test T25FW. Les données de marche de ces patients (*i.e.* séquence de quaternions unitaires), ainsi que le temps de marche du T25FW, leur score EDSS et les sous-scores des fonctions neurologiques pyramidale, cérébelleuse et sensitive constituent la `BDDsep`.

Cette étude ancillaire a permis d'obtenir des données mesurées auprès de patients dans des délais très courts comparativement au temps nécessaire à la mise en place d'une étude clinique, en particulier en période de pandémie du COVID-19. En effet, à la suite de cette première expérience, une étude a été organisée avec le CHU de Nantes à l'aide d'un financement obtenu par un dossier *Projet Exploratoire, Premier Soutien* proposé par l'*Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société* (AMIES). Le financement a été obtenu en Décembre 2018, et la rédaction du protocole a démarré immédiatement. Les premières inclusions des 44 patients prévus dans cette étude ont débuté en Août 2021 et sont encore en cours à la date de la rédaction de ce manuscrit. Une autre étude multi-centrique co-organisée avec les CHU de Rennes et de Nantes prévoyant l'inclusion d'une centaine de patients atteints de SEP a pu être financée grace à un appel à projet de la *Fondation pour l'Aide à la Recherche dans la Sclérose En Plaques* obtenu en septembre 2019. Le début des inclusions des patients est prévu pour fin 2022.

Les données du dispositif MMR constituent donc une mesure quantitative de la marche d'un individu. Il est nécessaire de développer des méthodes d'analyses appropriées afin d'en extraire des informations relatives à la démarche de l'individu et aux potentiels déficits l'impactant. Pour ce faire, un algorithme permettant la détection des cycles de marche doit être développé. Plusieurs paramètres spatio-temporels associés aux cycles de marche ainsi identifiés peuvent ensuite être calculés. Cet aspect est abordé dans le chapitre 2. Une autre approche consiste à comparer la forme des données représentant les cycles de marche des individus par méthode de classification. L'objectif est d'identifier des groupes de patients présentant des troubles de la marche similaires à partir de leurs données de marche. Pour ce faire, des méthodes de classifications adaptées aux données du dispositif MMR sont nécessaires. La section 1.3 présente la problématique de la classification non supervisée de façon générale et les spécificités liées aux données se présentant sous la

forme de mesures d'un paramètre répétées au cours du temps. Ces notions sont adaptées et appliquées aux séquences de quaternions unitaires représentant des cycles de marche dans le chapitre 3.

1.3 Méthodes de classification pour données de marche

1.3.1 Classification non supervisée : présentation générale

La classification non supervisée est une méthode d'apprentissage qui consiste à répartir un ensemble d'objets en groupes en fonction de leur similarité. Ce regroupement est réalisé directement à partir des caractéristiques des données, sans autres informations a priori concernant la structure des groupes à former ou leur nombre. Elle s'applique donc aux cas où il n'y a pas (ou peu) d'informations disponibles autres que le jeu de données en lui même. La classification est un vaste sujet de recherche, et de très nombreux aspects et méthodes sont décrits dans la littérature. Parmi les ouvrages de référence, on peut citer les livres *Finding Groups in Data : An Introduction to Cluster Analysis* de Leonard Kaufman et Peter J. Rousseeuw [87], et *Data Clustering Algorithms and Applications* de Charu C. Aggarwal et Chandan K. Reddy [2]. Dans la suite de cette section, seuls les aspects et concepts nécessaires à la compréhension de la suite du manuscrit sont présentés. Notamment, s'il existe des méthodes de classification permettant d'identifier des groupes de *variables* fortement corrélés [174], seule la la classification non supervisée *d'observations* est abordée dans ce manuscrit.

La problématique générale de la *classification non supervisée* peut être formulée intuitivement comme *le processus de distribution d'un ensemble de n observations* $Q = \{Q_1, Q_2, \ldots, Q_n\}$ *dans un ensemble de groupes* $C = \{C_1, \ldots, C_K\}$*, tels que* $\bigcup_{k=1}^{K} C_k = Q$ *et que les groupes* C_k *soient aussi homogènes et séparés que possible* [33]. De nombreuses catégories de méthodes de classification non-supervisée existent :

— Les *modèles génératifs et probabilistes* : Ces approches consistent à optimiser l'adéquation entre les données observées et un modèle mathématique par une approche probabiliste. Par exemple, on peut supposer que les données suivent un modèle génératif, *e.g.* qu'elles suivent une loi de mélange. Un groupe est alors défini comme l'ensemble des données observées qui suivent une même distribution.

— Les méthodes *basées sur la densité* : Dans cette approche, aucune hypothèse concernant le nombre ou la distribution des groupes n'est formulée, mais ils sont supposées être des régions de l'espaces des données avec une forte densité d'observations (*e.g* la méthode *DBSCAN*).

— Les méthodes *basées sur une grille* : Les algorithmes de cette catégorie divisent l'espace des données observées en un nombre fini de *cellule* formant une grille. Les

groupes sont alors les régions de la grille contenant une forte densité d'observations par rapport à leur entourage.

— Les méthodes *basées sur la distance*.

La suite de cette section se concentre sur la catégorie des *algorithmes basés sur la distance*, *i.e.* ceux pour lesquels les notions d'*homogénéité* (ou *cohésion*) et de *séparation* dépendent d'une fonction quantifiant la proximité entre paire d'observations du jeu de données $d : Q \times Q \to \mathbb{R}^+$. Si le terme "basée sur la distance" est utilisé pour définir cette catégorie de méthodes, la fonction d peut être :

— une *dissimilarité*, *i.e.* ayant les propriétés de *non-négativité*, *symétrie* et *identité*,

— une distance, *i.e.* respectant la propriété de l'*inégalité triangulaire* en plus des trois propriétés précédentes [105] (*c.f.* description de ces propriétés section 1.2.2.1, §**Définition d'une distance propre aux quaternions unitaires**).

Les méthodes de classification *basées sur la distance* sont décrites par Aggarwal *et al.* (2014) comme les plus utilisées en pratique [2], car elles peuvent être utilisées sur tout type de données pour lequel une fonction de *distance*/*dissimilarité* appropriée peut être définie. Le résultat de ces algorithmes se présente généralement sous la forme d'une partition de Q, dont les groupes sont disjoints : $C_k \bigcap C_{k'} = \emptyset$ pour $k \neq k'$. Les résultats de ce type de méthode dépendent donc fortement de la fonction de *distance*/*dissimilarité* utilisée pour quantifier la similarité entre les observations. Elle doit donc être choisie avec une grande attention, en tenant compte du type de données analysées et de la caractéristique qui doit être comparée entre deux observations.

Les méthodes de classification *basée sur la distance* se répartissent en deux grandes catégories en fonction de la stratégie utilisée pour former les groupes d'observation [146] :

1. La classification par partitionnement.
2. La classification hiérarchique.

Ces deux types d'approche sont présentés dans la section suivante.

1.3.1.1 Type d'approche de classification basée sur la distance

Classification par partitionnement Les méthodes de classification par partitionnement rassemblent les méthodes de recherche itérative de la partition optimale des observations en un nombre prédéfini de groupes K. Dans le cas des algorithmes basés sur la distance, elles sont toutes basées sur le principe général de la méthode *K-means* [2].

Dans ce type d'algorithmes, chaque groupe est représenté par un *prototype*, et la partition dite optimale est déterminée de manière itérative. Pour ce faire, K prototypes sont initialement choisis parmi les observations du jeu de données. Les observations sont alors attribuées au groupe représenté par le prototype dont elles sont le plus proches. Le nouveau prototype de chaque groupe est calculé à partir des observations qu'il contient. Les itérations d'attribution des observations et de calcul des nouveaux prototypes sont répétées jusqu'à ce qu'une itération supplémentaire d'attribution des observations ne change pas les prototypes, ou qu'un critère de convergence soit atteint. Ce critère de convergence est basé sur la minimisation d'une fonction objective permettant d'évaluer la qualité de la partition (tels qu'ils seront présentés plus bas dans cette section).

Le pseudo-code des méthodes de partitionnement est présenté dans l'algorithme 1.

Algorithm 1 Classification par partitionnement

Entrées :
$Q = \{Q_i\}, i \in \{1 \dots n\}$: Un ensemble d'observations.
K : Le nombre de groupes à former.
Fonctions :
prototype(C) : Calcule le prototype du groupe C.
$d(Q_i, Q_j)$: Calcule la dissimilarité entre deux observations.
Début :
Sélection aléatoire des prototypes des groupes $\bar{C}_k$ parmi Q
$C_k \leftarrow \emptyset,\ \forall k \in \{1 \dots K\}$ // Initialisation des groupes
for $i = 1$ **to** n **do**
  $C_{k^*} \leftarrow \{\{C_{k^*}\}, \{Q_i\}\}$, with $k^* = \arg\min_k d(\bar{C}_k, Q_i)$ // Attribution aux groupes
end for
$\bar{C}_k \leftarrow \text{center}\,(C_k)\,,\ \forall k \in \{1 \dots K\}$ // Calcul des prototypes des groupes
repeat
  $C_k^{-1} \leftarrow C_k,\ \forall k \in \{1 \dots K\}$
  $C_k \leftarrow \emptyset,\ \forall k \in \{1 \dots K\}$
  for $i = 1$ **to** n **do**
    $C_{k^*} \leftarrow \{\{C_{k^*}\}, \{Q_i\}\}$, with $k^* = \arg\min_k d(\bar{C}_k, x_i)$
  end for
  $\bar{C}_k \leftarrow \text{center}\,(C_k)\,,\ \forall k \in \{1 \dots K\}$
until $C_k^{-1} \neq C_k,\ \forall k \in \{1 \dots K\}$
Renvoyer C
Fin

S'il est prouvé que cette méthode converge en théorie vers la partition optimale [150], elle reste sensible à la sélection initiale des prototypes. Ainsi, la partition finale peut

correspondre à un minimum local sans être la partition optimale [126]. Une stratégie possible pour fiabiliser cette méthode est d'appliquer l'algorithme 1 de manière répétée, en sélectionnant des prototypes initiaux aléatoirement parmi les observations (tel que proposé par MacQueen (1967) [106]), puis de sélectionner la partition présentant la meilleure qualité.

Plusieurs variantes de l'algorithme de classification par partitionnement sont décrites [2]. La première, et peut être la plus *populaire*, est la méthode du *K-means*, dans laquelle le prototype d'un groupe est représenté par sa moyenne :

$$\bar{C}_k \leftarrow \text{moy}\,(C_k)\,. \tag{1.28}$$

avec moy une fonction permettant de calculer la moyenne des observations (*e.g* la moyenne arithmétique dans le cas de données euclidienne). La fonction objective permettant d'évaluer la qualité de la partition C est la *somme du carré des erreurs intra groupe* WSS (pour *Within Sum of Squared error*) :

$$\text{WSS}(C) = \sum_{k=1}^{K} \sum_{Q_i \in C_k} d(Q_i, \bar{C}_k)^2, \tag{1.29}$$

avec $d(.,.)$ la distance entre deux observations (*e.g.* la distance euclidienne).

Une variante populaire de l'algorithme du *K-means* est le *K-medoid* (décrit comme *Partition Around Medoïd* par Kaufman et Rousseeuw [87]), pour laquelle le prototype d'un groupe est défini comme son médoïde, i.e. l'observation la plus centrale du groupe :

$$\bar{C}_k \leftarrow \underset{Q \in C_k}{\arg\min} \sum_{Q_i \in C_k} d(Q, Q_i), \tag{1.30}$$

Par cette méthode de détermination du prototype, le *K-medoid* est supposé plus robuste à la présence de potentielles valeurs aberrantes, ou *outliers*, que le *K-means* [112].

Remarque : D'autres variantes du K-means existent, telles que le K-medians dans laquelle le prototype du groupe est calculée comme la médiane des observations, le K-modes adapté aux données non numériques [73]*, ou le Fuzzy K-means (ou Fuzzy C-means) où l'appartenance d'une observation à un groupe est évaluée par un score de* 0 *à* 1 [42, 96]*. Ces méthodes ne seront pas détaillées dans ce manuscrit, car elles n'ont pas été adaptées aux données de marche mesurées par le système MMR (i.e. séquence de quaternions unitaires).*

Classification hiérarchique La classification hiérarchique présente l'avantage d'être directement applicable sur la matrice D de dissimilarités entre paires d'observations, soit $d_{ij} = d(Q_i, Q_j)$ pour tout $i, j \in \{1 \dots n\}$, où $d(.,.)$ renvoie la dissimilarité entre deux observations du jeu de données. Deux approches principales peuvent être adoptées en classification hiérarchique. L'approche *descendante* consiste à considérer que toutes les observations appartiennent à un seul groupe à l'état initial. À chaque itération, un groupe est choisi pour être ensuite divisé en deux. Ce processus itératif s'arrête lorsque chaque observation se retrouve dans son propre groupe. L'approche *ascendante* consiste à considérer que chaque observation se trouve dans son propre groupe à l'état initial. À chaque itération, les deux groupes les plus proches sont fusionnés. Ce processus itératif s'arrête lorsque toutes les observations se retrouvent dans le même groupe. La classification descendante hiérarchique ne sera pas étudiée plus en détail dans ce manuscrit, cette dernière pouvant être considérée comme une répétition d'algorithmes de partitionnement ([105]). Dans la suite, nous allons donc nous concentrer sur la classification ascendante hiérarchique (CAH).

Soit l'état initial dans lequel chaque observation est son propre groupe, chacun étant donc constitué d'une seule observation : $C_i = \{Q_i\}, \forall i \in \{1 \dots n\}$. La première étape est de déterminer les deux groupes les plus proches C_i et C_j qui seront fusionnés en un seul groupe noté C_{ij}, avec $C_{ij} = C_i \cup C_j$ et $(C_i, C_j) = \arg\min_{C_k, C_l} D(C_k, C_l)$. L'étape suivante consiste à mettre à jour la matrice D avec la nouvelle valeur de dissimilarité entre le nouveau cluster C_{ij} et chaque autre cluster C_k, $\forall k \in \{1, \dots, n\} \setminus \{i, j\}$. Plusieurs critères de liaison ont été décrits dans la littérature pour déterminer comment mettre à jour la dissimilarité entre les groupes. Ils ont été unifiés par la formule de récurrence de Lance et Williams[99] :

$$D\left(C_{ij}, C_k\right) = \alpha_1 D\left(C_i, C_k\right) + \alpha_2 D\left(C_j, C_k\right) + \beta D\left(C_i, C_j\right) + \gamma \left| D\left(C_i, C_k\right) - D\left(C_j, C_k\right) \right|, \tag{1.31}$$

où α_1, α_2, β et γ sont des nombres scalaires réels. Les critères de liaison peuvent être divisés en deux catégories. Les critères *géométriques* (par exemple, Ward, centroïde, médiane) supposent que les observations appartiennent à un espace euclidien et la distance euclidienne est donc implicitement supposée pour mesurer la dissimilarité entre observations. Les critères de *graphe* calculent la dissimilarité entre deux groupes uniquement à partir des dissimilarités entre les observations qu'ils contiennent [130]. Cette propriété en fait les seuls critères adaptés aux matrices de dissimilarité non euclidiennes. Les trois critères

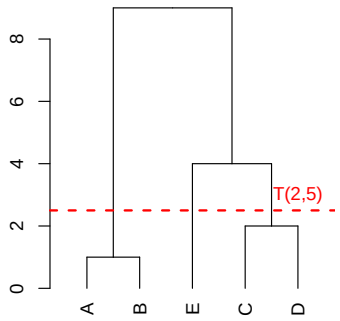


FIGURE 1.7 – Exemple de dendrogram construit par CAH

de liaison de graphe les plus courants sont : *liaison simple*, *liaison complète* et *liaison moyenne*. Les coefficients de la formule de Lance et Williams (1.31) associés à chacun d'eux sont donnés dans le tableau 1.1, où $|C_i|$ désigne le nombre d'observations dans le groupe C_i.

TABLE 1.1 – Coefficients de Lance et Williams associés aux critères de liaison de graph

Critère	α_1	α_2	β	γ
Liaison simple	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Liaison complète	$\frac{1}{2}$	$\frac{1}{2}$	0	$+\frac{1}{2}$
Liaison moyenne	$\frac{\lvert C_i\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$\frac{\lvert C_j\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	0	0

Les étape de fusion de groupe et mise à jour de la matrice de dissimilarité sont ainsi répétées jusqu'à ce que toutes les observations appartiennent au même groupe. Le résultat final d'une CAH se présente sous la forme d'un dendrogramme noté $\mathcal{T}$, dont une représentation commune est visible sur la figure 1.7. Le dendrogramme se présente sous la forme d'un arbre dont les feuilles représentent les observations du jeux de données et la hauteur des branches correspond à la dissimiliarité à partir de laquelle 2 groupes sont fusionnés.

On note $\mathcal{T}(h)$ l'état du dendrogramme à une hauteur fixée t. Ainsi sur l'exemple figure 1.7, à la hauteur $t = 2,5$, la structure de $\mathcal{T}$ correspond au regroupement des observations

suivant : $\mathcal{T}(2,5) = \{C_1 = \{A,B\}, C_2 = \{E\}, C_3 = \{C,D\}\}$. On peut observer que les groupes $\{C,D\}$ et $\{E\}$ fusionneront à l'étape suivante, pour $\mathcal{T}(4)$. Le dendrogramme présente ainsi l'état de la classification à chaque étape du regroupement et peut être parcouru pour étudier les relations entre les observations et les groupes. La structure de l'arbre constitue aussi un outil pour la détermination du nombre de groupe à former. Ainsi, la CAH peut être stoppée à n'importe quelle étape de la construction de l'arbre, pour obtenir une partition en un nombre K de groupes.

1.3.1.2 Sélection du nombre de groupes et validation de la classification

Le choix du nombre de groupes à former est un aspect fondamental de la classification non supervisée. En effet, il est une des entrées des algorithmes par partitionnement (*e.g. K-means*), et constitue une étape nécessaire pour obtenir une partition à partir d'un dendrogramme formé par CAH. Dans certains cas d'application, le nombre de groupes désirés peut être connu à l'avance, mais l'objectif d'une classification non supervisée peut également consister à identifier l'existence de groupes dans les données. Dans ce cas, le choix du nombre de groupe à former se base sur un ou des critères permettant l'évaluation de la qualité de la classification, *i.e.* sa validation. Outre le choix du nombre de groupes, l'étape de validation peut permettre d'identifier les valeurs d'hyper-paramètres d'un algorithme permettant de maximiser la qualité d'une classification, ou de comparer les classifications obtenues à partir d'algorithmes différents. Elle consiste donc à calculer le critère d'évaluation choisi en fonction des différentes classifications obtenues avec un ensemble d'algorithmes, de paramètres et de nombres de groupes différents, pour sélectionner la configuration aboutissant à la partition ayant la meilleure qualité [2].

Si cette étape revêt une grande importance, elle reste un problème encore ouvert [62]. De nombreux critères d'évaluation sont décrits, et se répartissent en deux catégories [2, 4] :

1. Les critères d'évaluation interne.
2. Les critères d'évaluation externe.

Critères internes Les critères internes permettent d'évaluer la classification uniquement à partir du jeu de données et des groupes formés, sans autres informations externes. Ils se basent en général sur l'évaluation de la *cohésion* et/ou de la *séparation* des groupes, qui sont les deux propriétés attendues d'une classification non supervisée. Parmi les indices les plus utilisés [4], on peut citer :

— Le *carré des dissimilarités intra-groupe* (WSS pour *Within Sum of Squares error*), autrement appelée *Inertie intra-groupe* (I_W) [128] : Cet indice représente la *cohésion* des groupes comme la somme des carrés des dissimilarités observées entre les observations constituant un groupe et son prototype. Il peut être calculé par groupe pour comparer leur *cohésion*, ou à l'échelle de la classification pour évaluer sa qualité globale. Sa valeur minimale est 0, qui correspond au cas théorique où toutes les observations sont égales au groupe auquel elles appartiennent. Les partitions de meilleure qualité sont associées à une faible valeur de I_W.

— L' *indice de silhouette* [140] : Cet indice est calculé pour chaque observation à partir de ses coefficients de *cohésion* et de *séparation*. La *cohésion* associée à une observation correspond à la moyenne des distances entre cette observation et les autres observations appartenant à son groupe. La *séparation* d'une observation correspond à la moyenne des distances entre cette observation et celles appartenant au groupe le plus proche. La silhouette d'une observation est d'autant plus grande qu'elle est proche des autres observations de son groupe et éloignée des observations des autres groupes. À l'inverse, la silhouette peut être négative dans le cas où une observation est en moyenne plus proche des observations d'un groupe différent du sien. L'étude de la silhouette de chaque observation permet d'évaluer si les groupes sont compacts et éloignés. La qualité globale de la partition peut être évaluée par la valeur moyenne des silhouettes.

— L'*indice de Dunn (*DI*)* : cet indice est calculé à l'échelle de la partition, comme le rapport entre la *séparation* et la *cohésion*. La *séparation* correspond à la plus petite dissimilarité entre deux observations appartenant à des groupes différents. La *cohésion* correspond à la plus haute dissimilarité entre deux observations appartenant au même groupe. La valeur de DI est d'autant plus grande que les groupes d'une partition sont compacts et séparés [62].

Ces critères sont adaptés à la validation d'une partition en groupe, et peuvent être utilisés pour choisir le nombre de groupes à former à partir d'un algorithme de partitionnement ou d'un dendrogramme obtenu par CAH.

D'autres critères sont spécifiques de l'évaluation de la qualité d'un dendrogramme. Par exemple, la *corrélation cophénétique* correspond à la corrélation entre la dissimilarité entre deux observations et la hauteur de la branche qui les relie dans un dendrogramme [158]. Ainsi, plus la *corrélation cophénétique* est élevée, plus la structure d'un dendrogramme représente fidèlement les distances entre paires d'observations du jeu de données initial.

La validation de la classification non supervisée est très étudiée dans la littérature, cependant elle reste un problème ouvert et il n'existe pas de consensus quant au meilleur critère interne à utiliser [2]. Une solution consiste à comparer la qualité d'une classification évaluée avec plusieurs critères afin d'identifier un consensus (*e.g.* identifier un nombre de groupes associé qui maximise/minimise plusieurs critères).

Critères externes Ils consistent à comparer la partition obtenue après classification avec une répartition en classe des observations connue *a priori*. Les classes initiales sont souvent définies à partir d'une vérité de terrain, l'avis d'un expert ou une classification obtenue à partir d'un autre algorithme [4].

Ces critères sont calculés à partir d'une table de contingence dont chaque cellule i, j correspond au nombre d'observations appartenant à la classe i et au groupe j. Ils représentent le degré de correspondance entre les groupes obtenus par la classification et les classes connues *a-priori*, *i.e.* si l'algorithme de classification tend à rassembler les observations de même classe dans les mêmes groupes et/ou s'il tend à séparer les observations appartenant à des classes différentes dans des groupes distincts.

Parmi les indices externes les plus utilisés, on peut citer l'indice de *Rand* (RI), qui correspond à la somme des paires d'observations correctement associées (*i.e.* les observations de même classe attribuées au même groupe et les paires d'observations de classe différentes attribuées à des groupes différents) divisée par le nombre total de paire d'observation [136], et sa version ajustée (ARI) [74]. L'indice de Rand prend valeur entre 0 et 1, et est d'autant plus proche de 1 que les répartitions des observations en groupes et en classe sont proches. L'ARI correspond au RI *ajusté sur la chance*, *i.e.* le RI auquel est soustrait la proportion de paires d'observations correctement associées qu'on peut attendre dans le cas où la répartition des observations en groupe est aléatoire. l'ARI peut donc prendre des valeurs négatives, dans le cas où la répartition en groupe obtenue par l'algorithme de classification est moins bonne que le hasard. On pourra citer également les indices de *Jaccard*, [2], ou encore le *Kappa de Cohen* [30].

La validation par critère externe permet d'évaluer un algorithme de classification sur un jeu de données dont la structure en groupes est déjà connue. Elle permet donc d'estimer le niveau de confiance que l'on peut accorder aux résultats qu'il produit sur ce type de jeu de données [2]. Dans les cas d'application en pratique, le nombre et la structure des groupes d'un nouveau jeu données est très rarement connu. Seuls les critères d'évaluation interne peuvent donc être utilisés pour estimer quantitativement la qualité d'une classification.

L'interprétation des résultats par des experts du domaine dans lequel l'algorithme de classification sera appliqué est également de première importance dans le processus de validation [33].

1.3.2 Méthodes de classification adaptées aux données de marche

Les données du dispositif MMR se présentent sous la forme d'une séquence de quaternions unitaires (*c.f* 1.2.2) représentant l'orientation de la hanche au cours du temps. Elles peuvent donc être appréhendées comme des *séries chronologiques* ou des *données fonctionnelles*. Leur analyse par méthodes de *classification basées sur la distance* nécessite donc de définir une fonction de dissimilarité et une méthode permettant de calculer le prototype d'un ensemble d'observations pour ces deux types de données. Ces aspects sont présentés pour ces deux types de données dans la suite de cette section.

1.3.2.1 Classification de séries chronologiques

Une *série chronologique* se présente sous la forme de valeurs discrètes ordonnées prises par un ou plusieurs paramètres au cours du temps. Sous sa forme brute, une série chronologique correspond à une série de N valeurs prises par un paramètre $Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N})$, chacune de ces mesures étant associée au temps auquel elle a été observée $T = (t_1, \ldots, t_N)$, avec $\mathbf{q}_{i,j}$ un élément de la série (pouvant être numérique ou non, uni- ou multidimensionnel) observé au temps $t_{i,j}$.

La suite de cette section se concentre sur la présentation des aspects liés à la *classification de séries complètes* (*Whole time-series clustering*). Deux autres catégories de méthode de classification de *séries chronologiques* peuvent être citées [4] :

— La classification de segments (*Subsequence clustering*) extraits d'une même série chronologique via une fenêtre glissante.

— La Classification d'éléments d'une même série temporelle, basée sur le temps de leur mesure et leur valeur. Cette approche est proche de la classification sur segments , bien que tous les points ne sont pas nécessairement attribués à une classe et peuvent être considérés comme du bruit

Si certaines méthodes de classification sur séries complètes peuvent être appliquées directement sur les données brutes, une étape de pré-traitement est souvent utilisée pour en donner une autre représentation. Le choix de pré-traiter les données vise le plus souvent à réduire leur dimension et est motivé par la réduction du temps de calcul des algorithmes,

permettant ainsi de traiter efficacement les jeux de données de très grande taille [4, 178, 137]. La transformation peut également permettre de réduire l'impact du bruit dans les données.

Les travaux de recherche présentés dans ce manuscrit consistent à appliquer des méthodes de classification à des données correspondant à des cycles de marche. Elles correspondent donc à des données de dimension relativement faible, constituées d'une centaine d'éléments mesurés sur des périodes de l'ordre de la seconde. Les problématiques liées aux données de grande dimension ne sont donc pas rencontrées. La suite du manuscrit présente les méthodes adaptées à la classification de séries chronologiques complètes sans pré-traitement. Le terme *observation* se rapporte donc à une série chronologique d'indice $i \in \mathbb{N}$ notée $Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N})$ et la grille de ses temps de mesure $T = (t_{i,1}, \ldots, t_{i,N})$. Comme évoqué précédemment, l'application de méthode de classification basée sur la distance consiste en premier lieu à identifier une fonction mesurant la dissimilarité entre paires de séries chronologiques.

Mesure de dissimilarité Plusieurs aspects des séries chronologiques font de la détermination de la dissimilarité une problématique complexe [4, 2]. En effet, les séries à comparer peuvent différer par leur nombre d'éléments ou par l'étendue des intervalles de temps séparant les éléments. Deux évènements (ou patterns) similaires peuvent être présents dans plusieurs séries chronologiques mais observés à des temps différents, par exemple en raison d'un mauvais alignement temporel des données. À cela s'ajoute le fait que les données peuvent être uni ou multivariées, et sujettes à la présence de bruit.

Plusieurs mesures de dissimilarité entre 2 séries chronologiques Q_1 et Q_2, respectivement de dimensions N_1 et N_2, sont décrites dans la littérature, et se distinguent par la stratégie adoptée pour s'adapter aux contraintes énoncées plus haut. Elles sont classées en fonction des aspects des données qu'elles permettent de comparer [4, 2]. La suite du manuscrit se focalise sur les mesures évaluant la dissimilarité de **forme** entre séries chronologiques.

La distance *point par point* est la dissimilarité la plus "intuitive". En effet, elle correspond à la somme des distances entre les éléments de Q_1 et Q_2 observés aux mêmes instants :

$$d_{pw}(Q_1, Q_2) = \sqrt[p]{\sum_{j=1}^{N} d(\mathbf{q}_{1,j}, \mathbf{q}_{2,j})^p}, \tag{1.32}$$

avec $d(.,.)$ une fonction mesurant la distance entre deux éléments. Pour l'utiliser, les deux

séries doivent être de même dimension. Dans le cas de données euclidiennes, $d(\mathbf{q}_{1,i}, \mathbf{q}_{2,i})$ peut être définie comme $|\mathbf{q}_{1,i} - \mathbf{q}_{2,i}|$. La distance *point par point* correspond alors à la distance de Minkowsi d'ordre p :

$$L_p - \text{norm}(Q_1, Q_2) = \sqrt[p]{\sum_{i=1}^{N} |\mathbf{q}_{1,i} - \mathbf{q}_{2,i}|^p}. \tag{1.33}$$

Cette dissimilarité présente l'avantage d'avoir les propriétés d'une distance, c'est à dire qu'elle satisfait les conditions de *non-négativité*, d'*identité*, de *symétrie* et d'*inégalité triangulaire*. La distance *point par point* est sensible à la présence de bruit et de mauvais alignement temporel entre deux séries chronologiques [29]. D'autres dissimilarités dites *élastiques* [4] sont proposées pour pallier ces limitations. Le terme *élastique* renvoie au fait que ces méthodes déforment les séries chronologiques initiales en modifiant les temps de mesure associés à leurs éléments ou en supprimant/remplaçant certains éléments des séries. Elles sont pour la plupart des généralisations aux séries numériques de l'*Edit Distance*, qui représente la différence entre 2 chaînes de caractère par le nombre d'opérations (délétion, insertion, modification de caractères) nécessaires pour transformer une chaîne en une autre [29]. Un nombre conséquent de mesures de dissimilarité entre séries chronologiques est décrit dans la littérature, dont certaines sont présentées ci-dessous.

La méthode *Edit Distance on Real sequence* (EDR) définit la dissimiarité entre deux séries Q_1 et Q_2 comme le nombre d'opérations d'insertion, de suppression ou de remplacement nécessaires pour transformer Q_1 en Q_2 [29]. La dissimilarité EDR s'écrit formellement comme :

$$\text{EDR}(Q_1, Q_2) = \begin{cases} N_1 \text{ si } N_2 = 0, \\ N_2 \text{ si } N_1 = 0, \\ \min \begin{cases} \text{EDR}(Q_{1,2:N_1}, Q_{2,2:N_2}) + \text{I}(c(\mathbf{q}_{1,1}, \mathbf{q}_{2,1}) > \epsilon), \\ \text{EDR}(Q_{1,2:N_1}, Q_2) + 1, \\ \text{EDR}(Q_1, Q_{2,2:N_2}) + 1 \end{cases} \end{cases}, \tag{1.34}$$

avec c(.,.) une fonction de coût mesurant la dissimilarité entre 2 éléments des séries Q_1 et Q_2, et $Q_{1,1:i}$ la sous-série $(\mathbf{q}_{1,1}, \ldots, \mathbf{q}_{1,i})$ et $Q_{2,1:j}$ la sous série $(\mathbf{q}_{2,1}, \ldots, \mathbf{q}_{2,j})$, I la fonction indicatrice, et ϵ une valeur seuil, en dessous laquelle les deux éléments $\mathbf{q}_{1,i}$ et $\mathbf{q}_{2,j}$ sont suffisamment proches pour être considérés comme similaires. La dissimilarité EDR est robuste à la présence de bruit et de décalage entre les séries temporelles [29]. Elle ne

respecte cependant pas la condition d'inégalité triangulaire, du fait de l'utilisation du seuil ϵ. De plus, les auteurs ne détaillent pas de méthode standardisée pour la définition de la valeur de ce paramètre, qui semble donc être un problème non résolu.

La dissmilarité *Edit Distance with Real Penalty* (ERP) repose sur le concept de *gap*, consistant en l'ajout d'un élément noté g de valeur constante à l'une ou l'autre des séries Q_1 et Q_2 et à des positions permettant de maximiser leur alignement. La fonction Q_1 alignée par l'ajout d'éléments g est notée $\tilde{Q_1}$. La dissimilarité entre Q_1 et Q_2 se calcule selon la formule suivante :

$$\mathrm{ERP}(Q_1, Q_2) = \begin{cases} \sum_{i=1}^{N_1} c(\mathbf{q}_{1,i}, g) \text{ si } N_2 = 0 \\ \sum_{i=1}^{N_2} c(\mathbf{q}_{2,i}, g) \text{ si } N_1 = 0 \\ \min \begin{cases} \mathrm{ERP}(Q_{1,2:N_1}, Q_{2,2:N_2}) + c(\mathbf{q}_{1,1}, \mathbf{q}_{2,1}), \\ \mathrm{ERP}(Q_{1,2:N_1}, Q_2) + c(\mathbf{q}_{1,1}, g), \\ \mathrm{ERP}(Q_1, Q_{2,2:N_2}) + c(g, \mathbf{q}_{2,1}) \end{cases} \end{cases} \tag{1.35}$$

La détermination de la valeur de l'élément g est une problématique naturellement soulevée par la formulation de la méthode ERP. Dans [28], les auteurs se placent dans le cas de séries de réels dans un espace euclidien, lesquelles ayant été normalisées $\mathrm{Norm}(Q_1) = \left\{\frac{x_1-\mu}{\sigma}, \ldots, \frac{x_n-\mu}{\sigma}\right\}$, avec μ et σ respectivement la valeur moyenne et l'écart type des valeurs $\{x_1, \ldots, x_n\}$. La fonction de distance $c(x_i, y_j)$ étant défini comme la $L_1 - norm : |x_i - y_j|$. Dans ce cas, si n'importe quelle valeur pour g permet à la dissimilarité ERP de satisfaire la condition d'inégalité triangulaire, la valeur $g = 0$ permet de définir $ERP(Q_1, Q_2)$ comme l'aire entre les deux séries, de plus les aires sous les trajectoires formées par Q_1 et $\tilde{Q_1}$ sont égales. Par ces propriétés, la dissimilarité ERP satisfait donc les conditions pour être considérées comme une distance [2, 28, 29], et la rend robuste à la présence de problème d'alignement entre les séries. Elle reste cependant sensible à la présence de bruit [29].

La méthode *Longest Common Sub-Sequence* (LCSS) a été développée spécialement pour déterminer la dissimilarité entre deux trajectoires mesurées par capteurs de mouvements, une trajectoire étant définie comme le changement de position d'un point dans un espace à 2 dimensions au cours du temps [175]. Une attention particulière est portée par les auteurs à rendre cette dissimilarité robuste à la présence de bruit dans les données. La

LCSS entre Q_1 et Q_2 est mesurée par la fonction suivante :

$$\mathrm{LCSS}_{\delta,\epsilon}(Q_1, Q_2) = \begin{cases} 0 \text{ si } N_1 = N_2 = 0 & \\ 1 + LCSS_{\delta,\epsilon}\left(Q_{1,1:N_1-1}, Q_{2,1:N_2-1}\right) \text{ si} & c(\mathbf{q}_{1,N_1}, \mathbf{q}_{2,N_2}) < \epsilon \\ & \text{et } |N_1 - N_2| \leq \delta \\ \max \begin{cases} LCSS_{\delta,\epsilon}\left(Q_{1,1:N_1-1}, Q_2\right), \\ LCSS_{\delta,\epsilon}\left(Q_1, Q_{2,1:N_2-1}\right) \end{cases} & \end{cases} \tag{1.36}$$

avec ϵ et δ deux valeurs seuils. Si la distance entre $c(\mathbf{q}_{1,i}, \mathbf{q}_{2,j})$ est inférieure à δ, les deux éléments sont jugés suffisamment proche pour être considérés comme similaires. La constante δ définit un seuil de tolérance pour les temps auxquels sont mesurés les éléments $\mathbf{q}_{1,i}$ et $\mathbf{q}_{2,j}$: si l'écart des temps de ces deux mesures est inférieur à δ, ces deux points sont jugés suffisamment proches dans le temps pour pouvoir être similaires. La similarité calculée par LCSS correspond donc à longueur maximale des sous-séquences de Q_1 et de Q_2 composés d'éléments suffisamment proches en valeur et en temps de mesure pour être similaires. Cette longueur est ensuite normalisée en la divisant par la taille de la plus petite des deux séries Q_1 et Q_2 pour calculer la similarité suivante :

$$\mathrm{S1}(\delta, \epsilon, Q_1, Q_2) = \frac{\mathrm{LCSS}_{\delta,\epsilon}(Q_1, Q_2)}{\min(N_1, N_2)} \tag{1.37}$$

L'équation suivante permet ensuite la conversion de S1 en une mesure de dissimilarité

$$\mathrm{D1}(\delta, \epsilon, Q_1, Q_2) = 1 - \mathrm{S1}(\delta, \epsilon, Q_1, Q_2) \tag{1.38}$$

L'un des avantages de la dissimilarité LCSS est sa robustesse à la présence de bruit dans les données. Elle ne respecte cependant pas la condition d'inégalité triangulaire [28, 175]. Elle nécessite également de fixer à priori les seuils δ et ϵ, et les auteurs ne donnent pas de méthodes standardisées pour définir leur valeur.

Le *Dynamic Time Warping* (DTW) est une des mesures de dissimilarité de séries

temporelles les plus populaires [3].

$$\mathrm{DTW}\left(Q_1, Q_2\right) = \begin{cases} 0 \text{ si } N_1 = N_2 = 0, \\ \infty \text{ si } N_1 = 0 \text{ ou } N_2 = 0, \\ c\left(\mathbf{q}_{1,N_1}, \mathbf{q}_{2,N_2}\right) + \min \begin{cases} \mathrm{DTW}\left(Q_{1,1:(N_1-1)}, Q_{2,1:(N_2-1)}\right) \\ \mathrm{DTW}\left(Q_{1,1:N_1}, Q_{2,1:(N_2-1)}\right) \\ \mathrm{DTW}\left(Q_{1,1:(N_1-1)}, Q_{2,1:N_2}\right) \end{cases} \end{cases}, \tag{1.39}$$

La dissimilarité DTW permet de s'affranchir de mauvais alignements locaux entre les séries chronologiques. Elle est cependant sensible à la présence de bruit dans les données et ne respecte pas les propriétés d'inégalité triangulaire et d'identité, et n'est donc pas une *distance* [29, 116]. Elle est la seule à avoir été généralisée aux séries chronologiques de quaternions unitaires par Jablonski *et al.* (2012) [82], dans la méthode *Quaternion Dynamique Time Warping* (QDTW). Dans la suite de ce manuscrit, seule cette dissimilarité sera donc détaillée et utilisée pour l'analyse des données mesurées par le dispositif MMR (*c.f.* section 3.1.2.1).

Détermination du prototype des groupes Plusieurs méthodes, en particulier les algorithmes de partitionnement, ainsi que certains critères de validation, nécessitent la détermination du meilleur représentant d'un groupe, appelé *prototype*. De façon similaire au cas des données réelles, le prototype d'un groupe correspond à la série chronologique qui minimise la somme des distances avec les observations de ce groupe. Une telle série est appelée *séquence de Steiner* [59]. Pour un ensemble de séries chronologiques $Q = \{Q_1, \ldots, Q_n\}$, elle est la série chronologique C telle que [128]

$$\sum_{i=1}^{n} d(Q_i, C)^2 \leq \sum_{i=1}^{n} d(Q_i, C')^2, \forall C' \neq C \tag{1.40}$$

Si $d(.,.)$ est la distance euclidienne ($L_p$norm avec $p = 2$), la solution C de l'équation 1.40 peut simplement être calculée comme la moyenne arithmétique des éléments observés aux même temps de mesure. L'utilisation d'une mesure de dissimilarité *élastique* pour $d(.,.)$ (*e.g.* DTW) rend la détermination exacte de C impossible à calculer en pratique, du fait de la puissance de calcul nécessaire [128]. Ainsi, plusieurs solutions pour en définir une approximation suffisante sont proposées.

Le prototype d'un groupe de séries chronologiques peut être défini par sa médoïde

(*c.f.* équation 1.30) [4]. Il s'agit de l'approche la plus simple à mettre en place, puisque le médoide est la série du groupe qui minimise la somme des distances avec les autres séries qu'il contient.

D'autres méthodes permettent de calculer une série moyenne en tenant compte des alignements réalisés entre les séries par une mesure de dissimilarité élastique. Plusieurs méthodes de calcul du prototype sont adaptées au DTW. La première calcule la moyenne de séries alignées séquentiellement par paires [58]. Le prototype final dépend de l'ordre des paires choisies [121]. Une autre consiste à calculer la moyenne d'un groupe de séries alignées sur le médoïde [1]. Le prototype du groupe sera donc de même taille que le médoïde. La dernière méthode identifiée dans la littérature, appelée *Dynamic time warping Barycenter Averaging* (DBA) permet de calculer itérativement la moyenne globale d'un ensemble de séries alignées par DTW sur une série définie comme le prototype initial [128]. Si sa convergence est prouvée par ses auteurs, le prototype final est dépendant de l'*initialisation*, *i.e.* de la série définie comme prototype initial. DBA peut également converger vers un minimum local. Cette méthode a été généralisée par Tomasz *et. al* (2017) [60] pour calculer la série moyenne d'un groupe de séries chronologiques de quaternions unitaires. Seule cette méthode sera donc détaillée dans la suite de ce manuscrit (*e.g.* cf section 3.1.2.2).

Dans les deux paragraphes précédents, des méthodes appropriées aux séries chronologiques ont été identifiées pour mesurer la dissimilarité entre observations et calculer le prototype d'un groupe d'observations. Dans le paragraphe suivant, quelques exemples d'application de classification de séries chronologiques pour l'analyse de la marche identifiés dans la littérature sont présentés.

Applications à l'analyse de la marche dans la littérature Une approche souvent décrite dans la littérature consiste à appliquer des méthodes de classification sur un ensemble de paramètres spatio-temporels des cycles de marche (voir 1.1.1.1 déterminés à partir des données de marche des individus [151, 41, 118]. D'autres approches transforment les données de marche par décomposition en valeurs singulières [141] ou par ondelettes discrètes [18].

Plusieurs études appliquent une CAH sur les séries brutes en utilisant DTW comme mesure de dissimilarité. Baghdadi et al. (2019) utilisent cette approche pour décrire leurs

données dans l'étude de la fatigue des travailleurs par des dispositifs portables fixés à la cheville [11]. Pullido-Valdeolivas et al. (2018) forment des groupes de patients atteints de paraplégie spastique pédiatrique héréditaire pour découvrir des phénotypes de la marche par CAH sur séries temporelles représentant la cinématiques de la démarche [134]. Steinmetzer et al. (2018) regroupent des patients diagnostiqués avec une maladie de Parkinson en comparant leurs données de marche mesurées par des semelles équipées d'un accéléromètre [163]. Les groupes formés tendent à rassembler des patients présentant des troubles de la marche similaires. Aucune étude équivalente utilisant un *algorithme de partitionnement* n'a été identifiée dans la littérature. Un seul exemple de classification de QTS brutes appliquée à l'analyse de données de mouvement a été décrit dans la littérature par Jablonski (2011) [82]. Cette approche passe par la généralisation du DTW aux séries temporelles que quaternions unitaires (*unit Quaternion Time Series*, QTS), appelée Quaterinon Dynamic Time Warping (QDTW). Dans cette étude, l'auteur montre que l'utilisation de QDTW comme mesure de dissimilarité dans une CAH (avec critère de liaison moyenne) permet de former des groupes de données de mouvement correspondant aux individus chez qui elles ont été mesurées. L'une des contributions des travaux de thèse présentés dans ce manuscrit consiste à adapter l'algorithme *K-means* au QTS, et à l'appliquer avec la CAH, sur un ensemble de données de marche (*c.f.* chapitre 3).

1.3.2.2 Classification de données fonctionnelles

Cette section présente brièvement certaines notions relatives à la classification de données fonctionnelles. Une description détaillée de l'analyse de ce type de données est donnée dans le livre *Functional Data Analysis* (2005) de J. O. Ramsay et B. W. Silverman [135].

Données fonctionnelles L'approche par données fonctionnelles consiste à considérer que les éléments d'une séquence $\mathbf{q}_1, \ldots, \mathbf{q}_N$ observés aux temps $t_1, \ldots, t_N$ sont les valeurs d'une fonction continue f sur un maillage fini du domaine de définition T de cette fonction. Toute fonction continue peut être décomposée dans une base de fonctions $(\phi_\ell)_{1 \leq \ell \leq L}$.

$$f(t) = \sum_{\ell=1}^{L} \alpha_\ell \phi_\ell(t),$$

La décomposition exacte requiert une somme infinie ($L = \infty$). Dans la pratique, une approximation est donc nécessaire pour reconstruire le caractère fonctionnel des données observées. La base de fonctions $(\phi_\ell)_{1 \leq \ell \leq L}$ est donc de dimension L finie, et la valeur de L

doit être déterminée. Les coefficients α_ℓ peuvent être estimés à partir des éléments de Q selon deux cas de figure :

— Les observations sont supposées sans erreur de mesure :

$$\mathbf{q}_i = f(t_i),\ i \in \{1, \ldots, N\}.$$

Une procédure d'interpolation entre les valeurs successives peut alors être utilisée.

— Les observations sont des prédictions ponctuelles de la fonction avec erreur :

$$\mathbf{q}_i = f(t_i) + \epsilon_i,\ i \in \{1, \ldots, N\}.$$

Une procédure de lissage par moindres carrés est utilisée en choisissant une base de fonctions appropriée (fonctions trigonométriques, B-splines ou ondelettes) [135].

De façon similaire au cas des séries chronologiques abordé dans la section 1.3.2.1, l'application de méthodes de classification *basée sur la distance* sur un ensemble de données fonctionnelles nécessite d'identifier une mesure de dissimilarité et une méthode de calcul du prototype adaptée à ce type de données.

Dissimilarité et alignement des données fonctionnelles Une mesure couramment utilisée pour déterminer la dissimilarité entre fonctions dans la classification est la métrique associée à l'espace $L_2(\mathbb{R}^p)$ (qui sera désigné dans la suite par L_2 afin de ne pas alourdir les notations) :

$$d^2_{L^2}(f_1, f_2) = \int_{\mathbb{R}} ||f_1(t) - f_2(t)||^2_{\mathbb{R}^p} dt = \sum_{i=1}^{p} \int_{\mathbb{R}} \left(f_{1,i}(t) - f_{2,i}(t)\right)^2 dt,$$

avec $f_{1,i}(t)$ la valeur de la $i^{ème}$ composante du vecteur $f_1(t)$. Elle peut être utilisée comme mesure de dissimilarité entre données fonctionnelles dans un algorithme de classification hiérarchique [48], ou dans un algorithme *K-means* [78, 167].

Cependant, les données fonctionnelles représentent le même type d'observations que les séries chronologiques (*i.e.* mesures répétées au cours du temps). Elles sont donc sujettes à des problématiques similaires, notamment la possible présence de bruit, d'erreur de mesure ou de mauvais alignement entre les fonctions. Si les deux premières peuvent être en partie résolues lors de la reconstruction des données fonctionnelles, remédier au mauvais alignement, ou variation de phase, nécessite des méthodes spécifiques. L'alignement d'une

fonction consiste à modifier son axe des temps par une fonction de transformation h [109]. Si f est la fonction à aligner par la transformation h et g est la fonction f après alignement, alors $g = f \circ h$, c'est à dire $\forall t \in T$, $g(t) = f(h(t))$.

L'une des possibilités consiste à identifier les instants remarquables tels que des extremas locaux partagés entre les fonctions et le prototype, puis de transformer les données fonctionnelles pour que les temps associés à ces extremas soient concomitants [109]. Une autre approche consiste à rechercher parmi une classe de fonctions de transformation W celles qui minimisent la dissimilarité entre les fonctions réalignées et le prototype. Dans la suite de cette section, le cadre théorique de l'alignement de données fonctionnelles est présenté pour l'ensemble des fonctions $F = \{f : \Omega \subseteq \mathbb{R} \to E \subseteq \mathbb{R}^p\}$ à aligner par un ensemble W de fonctions de *warping* de *classe affine strictement croissante*. Les notations et la présentation du cadre théorique sont tirées de Vantini (2012) [171].

L'espace F auquel appartiennent les fonctions et la classe des fonctions de *warping* W doivent respecter les propriétés suivantes :

(a) $F = \{f : \Omega \subseteq \mathbb{R} \to E \subseteq \mathbb{R}^p\}$ est un espace métrique doté d'une métrique $d : F \times F \to \mathbb{R}_0^+$.

(b) W est un sous-groupe du groupe des automorphismes continues : $W = \{h : \Omega \subseteq \mathbb{R} \to \Omega \subseteq \mathbb{R}\}$.

(c) $\forall f \in F$ et $\forall h \in W$, alors :$f \circ h \in F$.

(d) Soient deux fonctions $f_1, f_2 \in F$ et $h \in W$, la distance d entre f_1 et f_2 est invariante à la composition de f_1 et f_2 par h (elle est dite invariante à la classe W) :

$$d(f_1, f_2) = d(f_1 \circ h, f_2 \circ h). \tag{1.41}$$

Les propriétés (a)-(d) permettent de définir une semi-métrique $d_W : F \times F \to \mathbb{R}_0^+$ qui est déterminée conjointement par la métrique d et la classe W :

$$d_W(f_1, f_2) = \min_{h1,h2 \in W} d(f_1 \circ h_1, f_2 \circ h_2) \tag{1.42}$$

Le choix de la classe de fonctions de *warping* W dépend de la nature des données à aligner et de la source de potentielle variabilité de phase. Dans l'application aux données de marche, cette variabilité peut être due à des erreurs dans l'identification du début ou de la fin d'un cycle de marche par un algorithme ou à une différence de leur durée. La première source de variabilité se traduit par un décalage temporel uniforme, *i.e.* le même

évènement sera observé à des instants décalés entre deux fonctions. La seconde se traduit par une dilatation ou contraction de la forme d'une fonction par rapport à une autre. La classe de fonction des fonctions de *warping* affine strictement croissante permet de limiter l'effet conjugué de ces deux type de variabilité [109] :

$$W = \{h : h(t) = \alpha t + \gamma, \text{avec } \alpha \in \mathbb{R}^+, \gamma \in \mathbb{R}\}. \tag{1.43}$$

Il reste donc à définir une distance d appropriée permettant de respecter la propriété (d).

L'espace métrique $L_2(\mathbb{R}^p)$ est couramment utilisé dans les méthodes de classification non supervisée de données fonctionnelles [83]. On prouve cependant aisément qu'il ne respecte pas la propriété (d) pour le calcul de la distance entre fonctions alignées par fonctions de *warping* affines strictement croissantes.

En effet, l'espace $L_2(\mathbb{R}^p)$ est défini par la métrique

$$d^2_{L^2}(f_1, f_2) = \int_{\mathbb{R}} ||f_1(t) - f_2(t)||^2_{\mathbb{R}^p} dt,$$

On devrait observer $d^2_{L_2}(f_1 \circ h, f_2 \circ h) = d^2_{L_2}(f_1, f_2)$. Or, en développant :

$$\begin{aligned} d^2_{L_2}(f_1 \circ h, f_2 \circ h) &= \int_{\mathbb{R}} ||f_1(h(t)) - f_2(h(t))||^2_{\mathbb{R}^p} dt \\ &= \int_{\mathbb{R}} ||f_1(\alpha t + \gamma) - f_2(\alpha t + \gamma)||^2_{\mathbb{R}^p} dt \end{aligned}$$

En posant $S = \alpha t + \gamma \leftrightarrow dS = \alpha \times dt$, on obtient

$$\begin{aligned} d^2_{L_2}(f_1 \circ h, f_2 \circ h) &= \int_{\mathbb{R}} ||f_1(S) - f_2(S)||^2_{\mathbb{R}^p} \frac{dS}{\alpha} \\ &= \frac{d^2_{L_2}(f_1, f_2)}{\alpha}. \end{aligned}$$

La métrique $d^2_{L_2}$ n'est donc pas invariante à la classe des fonctions affines strictement croissantes. Elle peut être normalisée de la façon suivante :

$$d^2_{L_2 m}(f_1, f_2) = \frac{d^2_{L_2}(f_1, f_2)}{||f_1||^2_{L^2} + ||f_2||^2_{L^2}} \tag{1.44}$$

avec :

$$||f_1||^2_{L^2} = d^2_{L^2}(f_1, 0_{L^2}) = \int_{\mathbb{R}} ||f_1(t)||^2_{\mathbb{R}^p} dt. \tag{1.45}$$

Ainsi,

$$||f_1 \circ h||^2_{L^2} = \frac{||f_1||^2_{L^2}}{\alpha}$$

ce qui donne :

$$d^2_{L_2 m}(f_1 \circ h, f_2 \circ h) = \frac{\frac{d^2_{L_2}(f_1,f_2)}{\alpha}}{\frac{||f_1||^2_{L^2}}{\alpha} + \frac{||f_2||^2_{L^2}}{\alpha}} = \frac{d^2_{L_2}(f_1, f_2)}{||f_1||^2_{L^2} + ||f_2||^2_{L^2}} = d^2_{L_2 m}(f_1, f_2)$$

La distance $d^2_{L_2 m}$ est donc invariante à la classe des fonctions affines strictement croissante, et permet de respecter la propriété (d). On peut donc définir la métrique permettant le calcul de dissimilarité entre paires de fonctions alignées par fonction de *warping* affines strictement croissante comme :

$$d_W(f_1, f_2) = \min_{h1,h2 \in W} d^2_{L_2 m}(f_1 \circ h_1, f_2 \circ h_2), \tag{1.46}$$

La distance (1.46) peut ainsi être utilisée dans un algorithme de classification non supervisée pour mesurer la dissimilarité entre deux observations d'un ensemble de données fonctionnelles (une *observation* désignant ici une fonction). Cette distance permet également de s'affranchir de la variabilité de phase entre les données. Le prochain paragraphe présente une méthode permettant de calculer la fonction moyenne d'un ensemble d'observations alignées par fonction de *warping*. Cette méthode permet ainsi de calculer les prototypes des groupes et de généraliser les méthodes de classification par *partitionnement* aux données fonctionnelles.

Alignement d'un ensemble de données fonctionnelles et calcul du prototype. Si l'alignement de paires de fonctions a été précédemment décrit, l'objectif est ici d'aligner un ensemble de n fonctions $\{f_i\}_{i=1,2,\ldots,n}$ sur son prototype noté f_0 par un ensemble de fonction $\{h_i\}_{i=1,2,\ldots,n}$. f_0 représente la forme moyenne des n fonctions alignées. La problématique consiste à minimiser l'équation suivante :

$$\sum_{i=1}^{n} d^2_{L_2 m}\left(f_i \circ h_i, f_0\right). \tag{1.47}$$

Il s'agit donc de déterminer conjointement l'ensemble $\{h_i\}_{i=1,2,\ldots,n}$ et le prototype f_0 qui minimisent la somme (1.47). La résolution de ce problème est permise par la succession de deux étapes réalisées itérativement. En notant $f_0^{[\ell-1]}$ le prototype déterminé à l'itération

$\ell - 1$:

(1) La première étape consiste à aligner les fonctions $\{f_i\}_{i=1,2,\ldots,n}$ sur $f_0^{[\ell-1]}$ par l'ensemble des fonctions de warping $\{h_i^{[\ell]}\}_{i=1,2,\ldots,n}$ tel que :

$$\left\{h_i^{[\ell]}\right\}_{i=1,2,\ldots,n} = \left\{\underset{h\in W}{\arg\min}\, d_{L_2m}^2\left(f_i \circ h, f_0^{[\ell-1]}\right)\right\}_{i=1,2,\ldots,n} \tag{1.48}$$

(2) La seconde étape consiste à calculer le prototype $f_0^{[\ell]}$.

Deux méthodes peuvent être utilisées pour déterminer le prototype à l'étape (2) :

(i) Le prototype est calculé par la moyenne de Fréchet des fonctions alignées :

$$f_0^{[\ell]} = \underset{f\in F}{\arg\min} \sum_{i=1}^{n} d_{L_2m}^2\left(f_i \circ h_i^{[\ell]}, f\right). \tag{1.49}$$

(ii) Le prototype est calculé comme le médoïde de l'ensemble des fonctions alignées :

$$f_0^{[\ell]} = \underset{f\circ h, f\in\{f_i\}, h\in\{h_i^{[\ell]}\}}{\arg\min} \sum_{i=1}^{n} d_{L_2m}^2\left(f_i \circ h, f \circ h\right). \tag{1.50}$$

Dans le cas (i) où la méthode de détermination du prototype par moyenne de Fréchet est utilisée, l'algorithme peut être initialisé en calculant le premier prototype $f_0^{[0]}$ comme la moyenne de Fréchet des fonctions initiales $\{f_i\}_{i=1,2,\ldots,n}$. Dans le cas (ii), le prototype initial $f_0^{[0]}$ peut être déterminé comme le médoïde de $\{f_i\}_{i=1,2,\ldots,n}$.

Cette méthode d'alignement d'un ensemble de fonctions sur son prototype permet de généraliser les méthodes de classification par *partitionnement* aux données fonctionnelles, en utilisant (1.49) pour le *K-means*, ou (1.50) pour le *K-medoid*. Ces méthodes sont implémentées sous R dans la fonction `fda` (option `center_method = "mean"` pour le *K-means alignment* et `center_method = "medoid"` pour le *K-medoids alignment*) du *package* `fdacluster` [12] disponible sur CRAN.

Si plusieurs applications de l'analyse de données fonctionnelles à des données représentant la marche sont décrites dans la littérature [37], aucune ne fait intervenir de *méthode de classification*. D'autres part, il n'existe pas de méthode d'analyse de données fonctionnelles de quaternions unitaires. L'une des contributions des travaux de thèse présentés dans ce manuscrit consiste donc à adapter les méthodes de classification *K-means*, *K-medoids* et *CAH* à ce type de données (*c.f.* chapitre 3).

12. https://cran.r-project.org/web/packages/fdacluster/index.html

1.3.3 Analyse de données de marche avec données supplémentaires par classification semi-supervisée

Les méthodes de classification non supervisées abordées dans la section 1.3.1 consistent à regrouper les individus en fonction de leur similarité, cette dernière étant déterminée à partir d'une seule source d'information (par exemple une série temporelle représentant leur démarche). Dans la pratique il est pourtant fréquent de disposer de plusieurs sources d'information associées à un même individu. C'est par exemple le cas en médecine où de nombreuses informations sont disponibles concernant les patients atteints de pathologies chroniques telles que la SEP. Leur suivi médical implique en effet d'évaluer la sévérité de l'atteinte de plusieurs fonctions neurologiques au cours des consultations hospitalières. La prise en considération de ces informations supplémentaires dans le processus de classification pourrait donc permettre d'améliorer la pertinence clinique des résultats.

Les méthodes permettant de tenir compte d'informations supplémentaires dans la classification sont dites semi-supervisées. Le développement de ces méthodes est motivé par la difficulté de converger vers la *meilleure solution* (c'est à dire le partitionnement maximisant la variabilité inter-groupe et minimisant la variabilité intra-groupe) par le biais d'un seul algorithme appliqué sur un ensemble d'observations représentées par une source de données [33]. Les méthodes semi-supervisées se répartissent en plusieurs classes en fonction de l'approche par laquelle les informations supplémentaires sont intégrées dans le processus de classification.

1.3.3.1 Classification avec contraintes

L'approche de *classification avec contraintes* s'applique dans les cas pour lesquels des informations concernant la structure de la classification attendue sont connues *a priori* [57]. Ces connaissances *a priori* peuvent se présenter sous la forme de contraintes ou de labels déjà disponibles pour une partie des observations [39]. Les contraintes sont elles-même catégorisées en deux niveaux : (i) les contraintes relatives aux instances et (ii) les contraintes relatives aux groupes. Les premières se présentent sous la forme de paires d'observations du jeu de données devant appartenir au même groupe dans la partition finale (notées contraintes "ML" pour *Must Link*) ou devant appartenir à des groupes différents (notées contraintes "CL" pour *Cannot Link*). Les contraintes relatives aux groupes définissent certaines propriétés structurelles que doivent respecter les groupes de la partition finale. Elles peuvent définir le nombre de groupes de la partition finale, le nombre

minimal ou maximal d'observation qu'ils doivent contenir, la distance minimale entre deux groupes ou la distance maximale entre deux observations du même groupe (*i.e.* le diamètre maximal d'un groupe) [98].

Que les informations supplémentaires soient sous la forme de labels ou de contraintes, l'algorithme peut être forcé à converger vers une partition qui les respecte toutes ou bien des violations peuvent être autorisées dans une certaine mesure. Les contraintes seront alors considérées comme fortes dans le premier cas et comme douces dans le second [39].

Enfin, il existe trois approches d'intégration des contraintes dans le processus de classification, répartissant les méthodes en autant de catégories : méthodes *basées sur la recherche*, méthodes *basées sur la distance* et méthodes *hybrides*. Dans les méthodes *basées sur la recherche*, ou *basées sur les contraintes*, les algorithmes de classification sont modifiés pour ajuster l'espace des solutions aux partitions satisfaisants les contraintes. Dans les méthodes *basées sur la distance*, les mesures de dissimilarité des algorithmes sont modifiées pour rapprocher les observations devant appartenir au même groupe et éloigner celles devant être dans des groupes différents. Les méthodes *hybrides* intègrent les deux autres types d'approches.

Les méthodes avec contraintes sont très largement développées à partir des algorithmes de classification par *partitionnement*, l'utilisation de contraintes étant peu adaptée à la *classification hiérarchique* [105]. Ceci peut s'expliquer partiellement par deux observations. Tout d'abord les approches par contraintes relatives aux instances sont les plus fréquemment décrites et utilisées dans la littérature [98]. Ensuite, la structure d'une classification hiérarchique implique que les observations soient intégrées à un même groupe à différentes étapes de la construction du dendrogramme, rendant les contraintes "ML" et "CL" inappropriées [10].

Un exemple d'application de classification avec contraintes pour l'analyse des troubles de la marche a été récemment publié par Yang et al. (2021) [184]. Les auteurs y décrivent une méthode de reconnaissance de troubles de la marche (appelés "anomalies" dans l'article) à partir de données vidéo. Elle se base sur une version modifiée de la méthode `COP K means`, une approche *basées sur la recherche* à partir de contraintes relatives aux instances [177]. La présence de troubles de la marche est simulée par le port de bandage ou de matériel limitant les mouvements. La démarche y est représentée par la *Gait Energy Image*, une approche spatio-temporelle dans laquelle la démarche est représentée par une image.

Aucune application de *classification avec contraintes* pour l'analyse de la marche dans

laquelle la démarche est représentée par série chronologique n'a été identifiée dans la littérature. En effet, si on observe une croissance dans le développement des approches de classification par contrainte, peu de méthodes adaptées aux séries chronologiques sont décrites dans la littérature. Ceci peut s'expliquer par la difficulté à adapter les méthodes existantes à ce type de données, comme en témoignent les travaux présentés par Lampert et al. (2018)[98]. Dans cet article, les auteurs présentent en effet la généralisation d'un ensemble de méthodes pour la classification avec contraintes ML et CL de séries temporelles. L'approche adoptée consiste à utiliser le DTW pour la mesure de dissimilarité entre observations et la détermination des prototypes de groupes par DBA. Parmi un large éventail de méthodes considérées par les auteurs, cinq se sont avérées utilisables, les autres étant inadaptées soit par construction, soit par leur trop grand temps de calcul ou par leurs difficultés à converger vers une solution.

1.3.3.2 Approches par ensemble de classifications

L'approche par *ensemble de classifications* consiste à combiner plusieurs classifications relatives aux mêmes observations afin de produire une unique classification globale [33]. Les classifications combinées peuvent être obtenues à partir de différentes sources de données représentant les mêmes observations et/ou d'algorithmes de classification différents. L'objectif est de compenser les erreurs pouvant survenir dans l'une ou l'autre des classifications afin de générer une "meilleure" solution globale.

La combinaison des méthodes de classification s'articule autour de trois caractéristiques. Les méthodes peuvent être appliquées séquentiellement ou en parallèle. Dans le premier cas, les informations générées par une méthode de classification sont transmises à la suivante.

Les algorithmes de classification peuvent être appliqués de manière isolée ou interagir ensemble. Dans la *classification coopérative*, les méthodes sont appliquées indépendamment, et la classification consensus est générée par une étape de post-traitement. Dans la *classification collaborative*, le résultat final est défini par itérations durant lesquelles (i) plusieurs méthodes de classification sont appliquées (ii) une classification consensus est définie et (iii) des informations relatives aux différentes classifications sont utilisées afin de générer de nouveaux paramètres utilisés par les méthodes de classification lors de la nouvelle itération (retour à l'étape (i)). Ces itérations sont répétées jusqu'à stabilisation de la classification consensus.

Les différentes méthodes de classification utilisées peuvent être appliquées sur le jeu de

données complet ou sur une sous partie. Dans le second cas, deux cas de figure peuvent être rencontrés. Les méthodes peuvent (i) être appliquées sur un sous-ensemble contenant les mêmes observations représentées par des sources d'informations différentes, appelée *classification verticale* ou (ii) être appliquées sur des sous-ensembles contenant des observations différentes représentées par les mêmes sources d'informations, appelée *classification horizontale*. Des cas de figures combinant des classifications horizontales et verticales peuvent également être rencontrés, à condition que certaines informations soient partagées par les différents sous-ensembles de données [33].

Certaines approches par *ensemble de classifications* sont développées pour intégrer spécifiquement des classifications par partitionnement [12] ou des classifications hiérarchiques [75], et d'autres sont construites pour pouvoir intégrer les deux types de classifications [186].

Des approches par *ensemble de classifications* adaptées à l'analyse de séries temporelles sont décrites dans littérature. Certaines permettent notamment de combiner les classifications obtenues par plusieurs types de représentation des données et/ou plusieurs mesures de dissimilarité [183, 172, 147].

1.3.3.3 Approche par compromis

L'approche par compromis permet une classification d'un ensemble d'observations représentées par deux sources d'information, l'une étant considérées comme principale et l'autre comme supplémentaire. Cette approche est relativement récente et deux méthodes sont décrites dans la littérature. La première proposée par Ma et al.. (2018)[105] permet d'intégrer une source d'information supplémentaire sous la forme d'un dendrogramme dit *ontologique*, représentant une classification hiérarchique des observations. La dissimilarité entre les observations utilisée dans la construction est pondérée par une dissimilarité ultramétrique dérivée du dendrogramme ontologique. La pondération est déterminée soit en maximisant une mesure interne de qualité des groupes (comme l'indice de Dunn), soit par validation croisée si certains labels sont connus *a priori*. La pénalité est donc optimisée pour un nombre donné de groupes. Par conséquent, une optimisation sur deux paramètres doit être effectuée : la pondération provenant de l'information supplémentaire et le nombre de groupes.

La seconde méthode de classification par compromis est présentée par Bellanger et al. (2020) [16] sous le nom de `Perioclust`, et sera nommée `hclustcompro` dans la suite de ce manuscrit. Elle est adaptée aux situations dans lesquelles deux matrices peuvent

être calculées, la première représentant les dissimilarités entre deux observations dans l'espace de l'information principale et la seconde représentant les dissimilarités dans l'espace de l'information supplémentaire. Une matrice de dissimilarité globale entre paire d'observations est alors calculée comme la moyenne pondérée des dissimilarités observées dans les deux sources d'information. Son calcul repose sur un coefficient qui représente la proportion de chacune des deux sources d'informations dans la dissimilarité globale. La valeur du coefficient de pondération est déterminée à partir d'un critère inspiré de la corrélation cophénétique [158]. La minimisation de ce critère permet la détermination d'une matrice de dissimilarité globale qui aboutit à une classification hiérarchique dont la structure représente le meilleur compromis entre la source d'information principale et la source d'information supplémentaire.

Ces deux méthodes ont été développées pour le traitement de données statiques, et les auteurs présentent une application sur des données représentant des habitudes de consommateurs pour la première et sur des données archéologiques pour la seconde. Cependant, étant applicables sur des matrices de dissimilarité, leur généralisation est possible par l'utilisation d'une mesure adaptée aux séries temporelles de quaternions unitaires telle que le QDTW. Cet aspect est présenté dans le chapitre 3.

Chapitre 2

Étude de la marche par séquence de quaternions unitaires

Ce chapitre présente la méthode développée pour identifier les cycles de marche dans les données brutes mesurées par le système de capteurs MetaMotionR (MMR) (*c.f.* section 1.2), et les différents paramètres spatio-temporels qui peuvent en être déduits pour représenter la marche de l'individu. Les performances de l'algorithme de détection des cycles sont ensuite évaluées sur la base de données `BDDapp` (*c.f.* section 1.2.3.1). La relation entre les paramètres spatio-temporels et les troubles de la marche de patients atteints de Sclérose En Plaques est ensuite évaluée à partir des données de la base `BDDsep` (*c.f.* section 1.2.3.2).

2.1 Algorithme de détection des cycles de marche `STRIde PAttern GEneration`

2.1.1 Présentation de l'algorithme

La première étape pour analyser les données mesurées par le système de capteurs MMR consiste à identifier les instants correspondant à des évènements des cycles de marche. Cette détection permet d'extraire du signal des segments correspondant à des cycles de marche. Comme décrit dans la section 1.1.1.2, il s'agit en effet d'une approche couramment utilisée dans les méthodes d'analyse de la marche par dispositif numérique. Dans la suite de ce manuscrit, un élément d'une séquence est appelé un *point*. Les algorithmes développés pour y parvenir reposent le plus souvent sur un ensemble de règles qui permettent de déterminer si un point d'une séquence correspond à un évènement particulier du cycle de marche [142]. Ces règles sont définies en fonction du type de dispositif utilisé, du type de données qu'il mesure et, dans le cas des dispositifs portatifs, de la position du dispositif sur le corps du porteur. L'aspect du mouvement représenté est en effet directement dépendant

de ces facteurs. Si de nombreux algorithmes de détection des cycles de marche sont décrits dans la littérature, aucun n'est adapté au cas où un unique système de capteurs porté à la ceinture mesure l'orientation de la hanche sous forme de quaternions unitaires.

Par conséquent, l'algorithme `STRIde PAttern GEneration` (`STRIPAGE`) est proposé. Cette section présente les différentes étapes de cet algorithme. Leur conception se base sur un ensemble de règles qui reposent sur les caractéristiques suivantes, dues à la semi-périodicité de la marche [47, 68] (*c.f.* section 1.1.1.1) :

— Les mouvements de la partie inférieure du corps survenant durant un cycle de marche peuvent présenter une variabilité inter-individu, *i.e.* différence observée entre les cycles de marche de plusieurs individus, et intra-individu *i.e.* différence observée entre les cycles de marche d'un même individu.

— Cependant, les variabilités restent suffisamment faibles pour que ces mouvements soient proches entre des cycles de marche mesurés chez un même individu et chez plusieurs individus.

Les règles sur lesquelles se basent l'algorithme `STRide PAttern GEneration (STRIPAGE)` sont aussi définies en fonction de l'aspect de la marche qui est mesuré par les données. Elles représentent l'orientation en 3 dimensions de la hanche droite au cours du temps (*c.f.* 1.2.2.2). Si l'on considère la représentation schématique d'un cycle de marche sur la figure 1.1 et la description qui en est donnée section 1.1.1.1, plusieurs hypothèses sont formulées concernant la variation de l'orientation de la hanche au cours du cycle de marche :

Hypothèse 1 : L'orientation de la hanche d'un individu est relativement similaire entre deux évènements équivalents survenant au cours de cycles de marche différents (par exemple, au moment de la pose du pied droit au sol).

Hypothèse 2 : Les deux orientations de la hanche les plus éloignées au cours d'un cycle de marche sont observées au *début de la phase d'appui* et au *début de la phase de balancement*.

Hypothèse 3 : On considère un cycle de marche dont le premier évènement est la pose du pied droit au sol. L'orientation de la hanche observée à cet instant est appelée *orientation initiale*. Durant le cycle, l'orientation de la hanche s'éloigne de son orientation initiale jusqu'à la fin de la phase d'appui. L'orientation observée à cet instant est appelée *orientation intermédiaire*. Le pied droit quitte le sol et la *phase de balancement* commence. Durant la phase de balancement, l'orientation de la hanche s'éloigne de l'*orientation intermédiaire* pour se rapprocher de *l'orientation initiale*.

Le pied droit entre en contact avec le sol, et un nouveau cycle de marche commence.

Hypothèse 4 : L'*orientation moyenne* de la hanche observée au cours d'un cycle de marche est proche de l'orientation de la hanche observée à l'instant où les deux jambes sont alignées. Ce phénomène survient deux fois au cours du cycle, une au cours de la phase d'appui et une au cours de la phase de balancement.

En tenant compte de ces hypothèses et l'aspect semi-périodique de la marche, l'algorithme `STRIPAGE` est conçu de telle sorte que ces règles soient respectées :

Règle ♯1 : La périodicité du signal correspond à la durée des cycles de marche.

Règle ♯2 : Les instants correspondant au *début de la phase d'appui* et au *début de la phase de balancement* correspondent aux deux orientations les plus éloignées observées dans un intervalle de temps correspondant à la durée d'un cycle de marche (*c.f.* **Hypothèse 2**). Ces instants sont appelés *points de segmentation*.

Règle ♯3 : Les cycles de marche d'un même individu ont une durée similaire.

Règle ♯4 : Les orientations de la hanche observées au *début de la phase d'appui* (resp. *au début de la phase de balancement*) sont similaires entre plusieurs cycles de marche détectés chez un même individu.

Règle ♯5 : Les orientations observées au *début de la phase d'appui* (resp. *au début de la phase de balancement*) sont relativement similaires entre les cycles de marche détectés chez des individus différents.

Règle ♯6 : Les changements d'orientation de la hanche observés au cours d'un cycle de marche sont similaires entre plusieurs cycles d'un même individu.

La suite de cette section présente la manière dont les différentes étapes de l'algorithme `STRIPAGE` se basent sur ces règles pour identifier les cycles de marche dans les données renvoyées par le dispositif MMR porté à la ceinture, au niveau de la hanche droite.

2.1.1.1 Données mesurées

Les données renvoyées par le système de capteurs MMR sont celles renseignant l'orientation en 3 dimensions du système de capteurs par rapport à un référentiel fixe (*c.f.* section 1.2). On rappelle que cette orientation se présente sous la forme d'un quaternion unitaire $\mathbf{q}$ représentant la rotation en 3 dimensions d'angle θ autour de l'axe $\mathbf{u}$ entre le référentiel fixe noté $\mathfrak{R}_f$ le référentiel propre au dispositif, noté $\mathfrak{R}_{capt}$ (*c.f.* section 1.2.2.1 et section 1.2.2.2). L'orientation est mesurée de manière répétée au cours du temps, chacune de ces

valeurs est donc associée à un temps de mesure t. Les données du système de capteurs représentant l'évolution de son orientation au cours du temps, elles peuvent être considérées comme une série temporelle de quaternions unitaires notée :

$$Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)^\top = \begin{pmatrix} \mathbf{w} = (w_1, \ldots, w_N) \\ \mathbf{x} = (x_1, \ldots, x_N) \\ \mathbf{y} = (y_1, \ldots, y_N) \\ \mathbf{z} = (z_1, \ldots, z_N) \end{pmatrix}^\top \tag{2.1}$$

mesurée sur la grille de temps $T = (t_1, \ldots, t_N)$.

2.1.1.2 Étape 1 : Estimation de la durée des cycles de marche par périodogramme

La première étape pour identifier les cycles de marche consiste à en estimer la durée. Pour ce faire, on se base sur la **Règle ♯1** et on suppose que la semi-periodicité de la marche transparaît dans au moins l'une des composantes de Q. Ainsi, on considérera de manière indépendante les séries univariées w, x, y et z, et on supposera qu'au moins l'une de ces séries est de forme sinusoïdale avec une période correspondant à celle d'un cycle de marche. La méthode suivante est appliquée aux 4 séries $\mathbf{w}$, $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$, l'exemple étant donné pour la série $\mathbf{x}$.

La fréquence $\omega_{\mathbf{x}}$, correspondant à la fréquence dominante des oscillations de $\mathbf{x}$, est déterminée par *periodogramme* [70]. $\omega_{\mathbf{x}}$ correspond à la fréquence associée à la plus grande valeur de la densité spectrale de X. L'estimation de la durée des cycles de marche par analyse spectrale des données est une méthode déjà utilisée dans la littérature (par exemple Ghersi *et. al* (2020) dans l'accélération mesurée au niveau du tronc [50]).

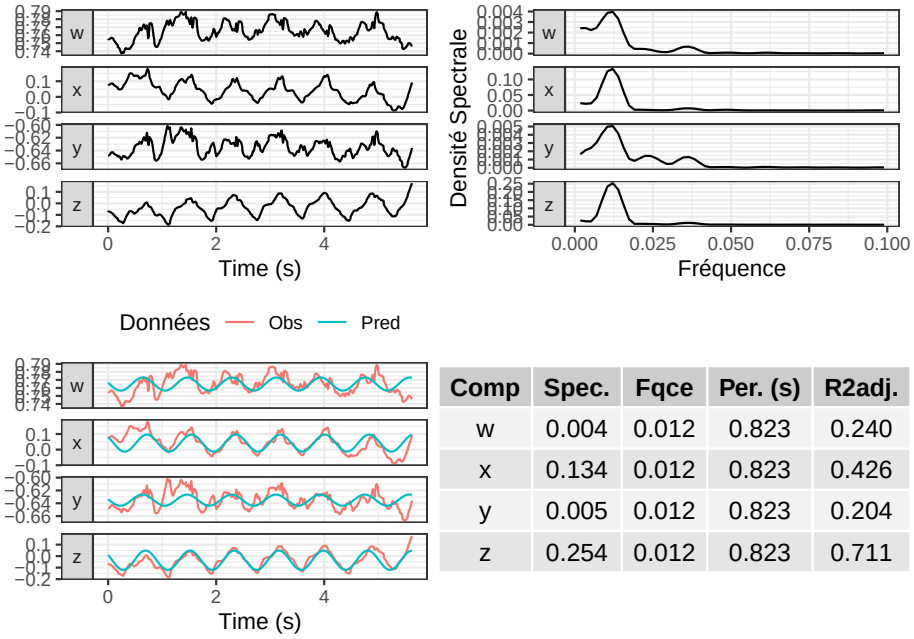
Un modèle linéaire est ajusté sur les valeurs de X :

$$x(t) = \beta_0 + \beta_1 \cos 2\pi\omega_{\mathbf{x}} t + \beta_2 \sin 2\pi\omega_{\mathbf{x}} t + \epsilon, \tag{2.2}$$

L'adéquation du modèle est évaluée par le coefficient de détermination ajusté R^2.

La fréquence estimée sur la série permettant d'ajuster le modèle avec la meilleure adéquation est notée ω_{wc} et permet d'estimer la durée d'un cycle de marche :

$$\tau_{wc} = \omega_{wc}^{-1} \tag{2.3}$$



Comp	Spec.	Fqce	Per. (s)	R2adj.
w	0.004	0.012	0.823	0.240
x	0.134	0.012	0.823	0.426
y	0.005	0.012	0.823	0.204
z	0.254	0.012	0.823	0.711

FIGURE 2.1 – Estimation de la période de cycle de marche

L'application de cette méthode sur un exemple de jeu de données est présentée sur la figure 2.1. Les 4 séries **w**, **x**, **y** et **z** sont représentées en haut à gauche, leur *périodogramme* en haut à droite. Les modèles ajustés à partir de leur fréquence dominante estimée à partir de leur *périodogramme* sont représentés en bas à gauche (données observées en rouge, courbe ajustée en bleu). Les valeurs de fréquence dominante pour chacune des composantes sont renseignées dans le tableau en bas à droite (colonne "Fqce"), avec la valeur de densité spectrale correspondante (colonne "Spec"), la période ("Per." en seconde) et la qualité d'ajustement du modèle ("R2adj."). Dans cet exemple, le modèle ajusté sur la composante **z** est celui présentant la meilleure adéquation aux données observées, avec une période de cycle de marche de 0.823 secondes.

En résumé, la méthode précédemment présentée permet d'estimer la durée d'un cycle de marche τ_{wc} par *périodogramme* en tirant partie de la semi-périodicité de la marche.

2.1.1.3 Étape 2 : Identification des points de segmentation

L'objectif est ici d'identifier les instants particuliers du signal correspondant à des évènements permettant de segmenter les données en cycles de marche, appelés *points de segmentation*. La méthode développée, schématisée sur la figure 2.2, repose sur la **Règle** ♯2 : "Les instants correspondant au *début de la phase d'appui* et au *début de la phase*

de balancement correspondent aux deux orientations les plus éloignées observées dans un intervalle de temps correspondant à la durée d'un cycle de marche".

Pour les identifier, on définit la fenêtre de recherche glissante

$$sw(i) = (i, \ldots, \min\{i + \delta - 1, N\}), i \in \{1, \ldots, N\}. \tag{2.4}$$

$sw(i)$ correspond à une sous séquence des données dans laquelle un *point de segmentation* est recherché. Son étendu est définie par un paramètre δ. La valeur de ce paramètre doit être telle que l'intervalle de temps $[t_{sw_1}, t_{sw_\delta}]$ soit d'une durée suffisante pour contenir une phase d'appui ou de balancement d'un cycle de marche. L'estimation de la durée des cycles de marche τ_{wc} décrite au paragraphe 2.1.1.2 permet d'adapter la valeur de δ à la cadence de marche de l'individu. En effet, le paramètre δ est déterminé de telle sorte que la fenêtre de recherche recouvre une période représentant 80% de τ_{wc} par l'arrondi à l'entier supérieur de :

$$\delta = 0.8 \times \tau_{wc} \times \omega_{IMU}, \tag{2.5}$$

avec ω_{IMU} le nombre de points mesurés par seconde par le système de capteurs (dans le cas du système de capteur MMR, $\omega_{IMU} = 100$ Hz).

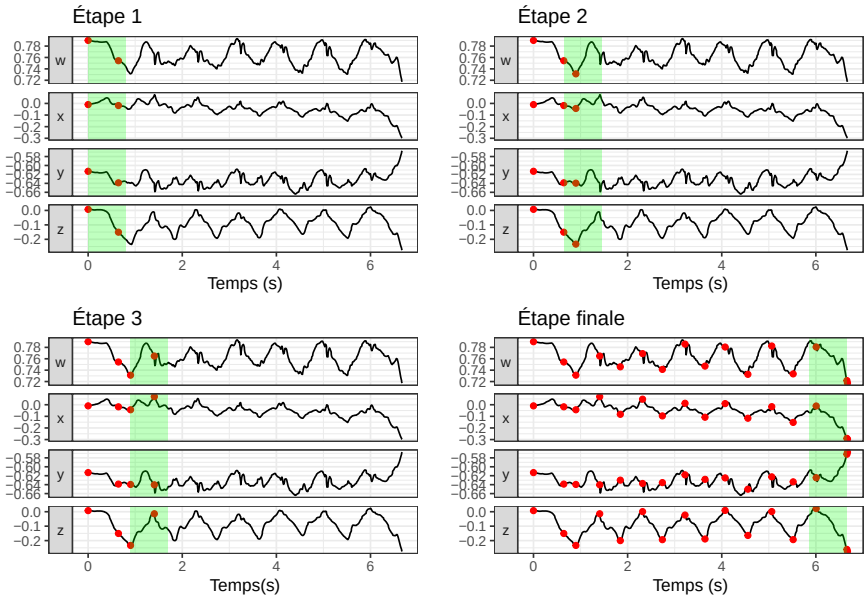


FIGURE 2.2 – Schématisation de l'algorithme `WCDSW` (avec $\delta = 66$)

Le premier *point de segmentation* noté r_1 est défini par défaut comme le premier point de la série (premier point rouge sur la figure 2.2, Étape 1 en haut à gauche). Le second

point de segmentation r_2 est identifié comme le point appartenant à la fenêtre de recherche $sw(r_1)$ (matérialisée par la zone verte sur la figure 2.2) correspondant au quaternion le plus éloigné de r_1 :

$$r_2 = \arg\max_j d(\mathbf{q}_{r_1}, \mathbf{q}_j), j \in sw(r_1),$$

avec $d(.,.)$ la distance géodésique entre deux quaternions unitaires (équation (1.23)). Le point r_2 est représenté par le second point rouge de la figure 2.2, Étape 1 en haut à gauche.

La fenêtre de recherche prend les nouvelles valeurs $sw(r_2) = (r_2, \ldots, r_2 + \delta - 1)$. Le *point de segmentation* suivant est identifié dans la nouvelle fenêtre $sw(r_2)$ par la même méthode qu'à l'étape précédente (voir Étape 2 en haut à droite sur la figure 2.2).

Cette phase de recherche se répète de façon itérative jusqu'à ce que la fenêtre de recherche atteigne la fin du jeu de données (figure 2.2, Étape 3 en bas à gauche et Étape finale en bas à droite). Cette méthode est appelée `Walking Cycle Detection Swiping Window` (`WCDSW`) et est décrite dans l'algorithme 2.

Le vecteur obtenu de taille R noté $\mathbf{r} = \{r_1, \ldots, r_R\}$ contient l'ensemble des points permettant la segmentation du jeu de données.

Algorithm 2 `Walking Cycle Detection Sweeping Window (WCDSW)`

Paramètres
$Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$: une série temporelle de quaternions unitaires
τ_{wc} : Estimation de la durée d'un cycle de marche
ω_{IMU} : La fréquence du dispositif (ici, $\omega_{IMU} = 100$ Hz)
Fonctions
$d(\mathbf{q}_1, \mathbf{q}_2)$: Calcule la distance géodésique entre les quaternions unitaires $\mathbf{q}_1$ et $\mathbf{q}_2$
Initialisation
$\delta \leftarrow 0.8 \times \tau_{wc} \times \omega_{IMU}$
$sw_1 \leftarrow 1$; $sw_2 \leftarrow \delta$; $\mathbf{r} \leftarrow r \leftarrow 1$
Début
while $sw_2 < N$ **do**
  $r \leftarrow \arg\max_j \left(d(\mathbf{q}_r, \mathbf{q}_j)\right), j \in \{sw_1, \ldots, sw_2\}$
  $\mathbf{r} \leftarrow \{\mathbf{r}, r\}$
  $sw_1 \leftarrow r$; $sw_2 \leftarrow \min\{r + \delta - 1, N\}$
end while
$\mathbf{r} \leftarrow \{\mathbf{r}, N\}$
Renvoyer r
Fin

À ce stade, 2 points de segmentation successifs $(r_i, r_{i+1})\forall i \in \{1, \ldots, R-1\}$ peuvent délimiter 3 types de périodes :

— une phase d'appui d'un cycle de marche,

— une phase de balancement d'un cycle de marche,

— une période ne correspondant pas à une phase de cycle de marche

On souhaite identifier les périodes correspondant à des cycles de marche complets, composés d'une phase d'appui suivie d'une phase de balancement. Pour identifier ces cycles de marche, on forme un ensemble de segments dits *candidats*, qui sont délimités par les *points de segmentation* r_i et r_{i+2}, $i \in \{1, \ldots, R-2\}$. La période correspondant à chaque segment *candidat* comprend donc 3 *points de segmentation*. Les premier et troisième *points de segmentation* de chaque segment correspondent respectivement à son point initial et à son point final, le deuxième étant son point intermédiaire. L'ensemble des triplets de *points de segmentation* permettant de former les segments *candidats* est noté :

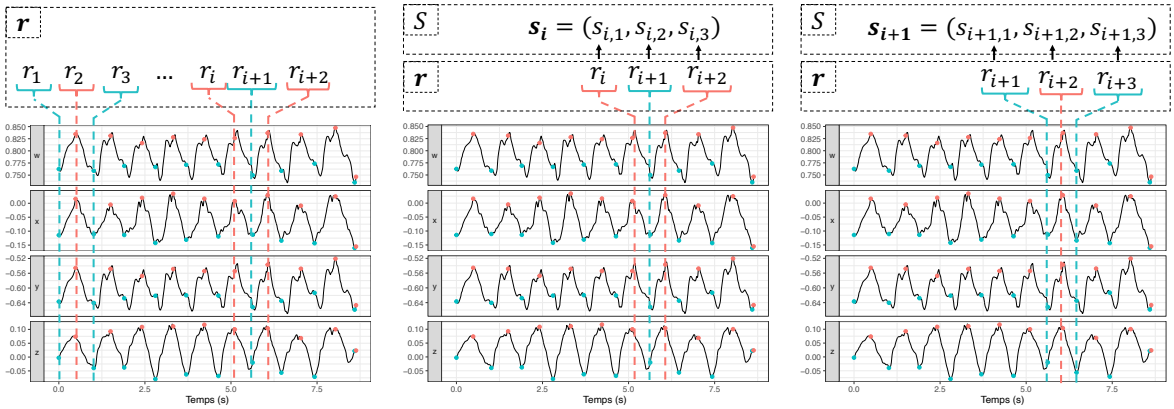
$$S = \{\mathbf{s}_1, \ldots, \mathbf{s}_{R-2}\}, \tag{2.6}$$

avec

$$\begin{aligned} \forall i \in \{1, \ldots, R-2\}, \mathbf{s}_i &= (s_{i,1}, s_{i,2}, s_{i,3}), \\ s_{i,1} &= r_i, \\ s_{i,2} &= r_{i+1}, \\ s_{i,3} &= r_{i+2} \end{aligned} \tag{2.7}$$

On note $M = R - 2$ le nombre de triplets contenus dans l'ensemble S. La figure 2.3a représente un exemple de jeu de données de marche dans lequel sont matérialisés les *points de segmentation* $\mathbf{r}$ identifiés par l'algorithme 2 `WCDSW`. Les *segments candidats* sont délimités par deux points successifs de la même couleur (deux points successifs rouges ou deux points successifs bleus). Deux exemples de segments candidats sont mis en évidence sur les figures 2.3b et 2.3c, représentés respectivement par les triplets de points de segmentation $\mathbf{s_i}$ et $\mathbf{s_{i+1}}$ les constituant.

Les prochaines étapes de l'algorithme `STRIPAGE` consistent à supprimer de l'ensemble des *segments candidats* S ceux qui ne sont pas constitués d'une phase d'appui suivie d'une phase de balancement. Ces segments sont appelés *outliers*. À ce stade, une première partie des *outliers* peut être identifiée à partir de leur durée.



(a) Segments *candidats* (b) Exemple du triplet $\mathbf{s_i}$ (c) Exemple du triplet $\mathbf{s_{i+1}}$

FIGURE 2.3 – Segments candidats et triplets

2.1.1.4 Étape 3 : Suppression des *outliers* de durée

Selon la **Règle** ♯2, la marche est supposée suffisamment régulière pour que plusieurs cycles de marche mesurés chez un même individu soient similaires de durée. Parmi les segments *candidats* délimités par l'ensemble des *points de segmentation*, on souhaite donc supprimer ceux dont la durée est jugée *trop différente*. Ces segments sont appelés *outliers de durée*.

Pour ce faire, on calcule la durée de tous les segments *candidats* :

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M), \tau_i = t_{s_{i,3}} - t_{s_{i,1}}, i \in \{1, \ldots, M\}. \tag{2.8}$$

La durée médiane des durées des segments *candidats* est ensuite calculée, ainsi que sa *Median Absolute Deviation* (*MAD*) :

$$\tilde{\tau} = \text{median}(\boldsymbol{\tau}) \tag{2.9}$$

$$\text{MAD} = \text{median}(\{|\tau_i - \tilde{\tau}|\}_{i=1,2,\ldots,M}) \tag{2.10}$$

Un Intervalle de Durée (ID) attendue pour les cycles de marche est déterminé à partir de ces 2 paramètres :

$$\text{ID} = \tilde{\tau} \pm \delta_{\text{ID}} \times \text{MAD} \tag{2.11}$$

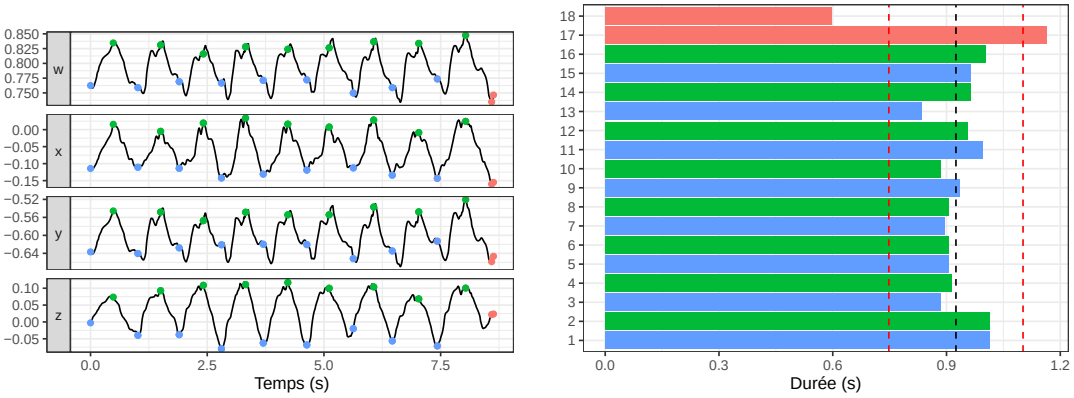
avec δ_{ID} un hyper-paramètre définissant l'étendue de l'intervalle ID. L'ensemble des triplets de *points de segmentation* formant des segments dont la durée n'est pas comprise

dans l'intervalle ID est noté T_{out} et retiré de l'ensemble S :

$$S \leftarrow S \setminus \{T_{out}\} \tag{2.12}$$

On note $M \leftarrow |S|$ le nombre de segments restant (avec $|S|$ le nombre de segments restant dans l'ensemble S)

La figure 2.4 présente un exemple d'identification d'*outliers de durée*, avec $\delta_{\text{ID}} = 3$. Les segments *candidats* sont délimités par deux points successifs de la même couleur sur la 2.4a. La durée correspondant à ces segments est présentée par les barres horizontales sur la figure 2.4b. La couleur d'une barre horizontale correspond à la couleur des points délimitant le segment dont elle représente la durée. La première barre bleue en bas de la 2.4b représente donc la durée du premier segment délimité par des points bleus sur la figure 2.4a. La durée médiane $\tilde{\tau}$ est matérialisée par la ligne verticale en pointillés noirs et l'intervalle ID est matérialisé par les 2 lignes en pointillés rouge. Les *outliers* dont la durée est située en dehors de l'intervalle ID sont représentés par des barres rouges. Sur la figure 2.4a, seuls les points correspondant à la fin de ces segments sont représentés. En effet le dernier point d'un segment correspond au premier point du segment suivant. Le premier point des deux *outliers de durée* de la figure 2.4 correspond donc au dernier point d'un segment *candidat*.



(a) Segments *candidats* et *outliers de durée*

(b) Durées des segments *candidats*

FIGURE 2.4 – Identification du groupe des cycles de marche

2.1.1.5 Étape 4 : Identification des phases d'appui et des phases de balancement

Par construction de l'algorithme 2 `WCDSW`, deux *points de segmentation* successifs dans un triplet correspondent aux deux orientations de la hanche les plus éloignées observées dans un intervalle de temps. Certains de ces points délimitent des phases de cycle de marche (phase de balancement ou phase d'appui), et d'autres peuvent ne pas correspondre à ce type d'évènement.

Il est nécessaire d'identifier le type de phase que délimitent deux *points de segmentation* successifs. Cette identification permet par la suite de sélectionner les segments débutant par une phase d'appui, suivie d'une phase de balancement. La méthode développée repose sur la **Règle** ♯4 "Les orientations observées au *début de la phase d'appui* (resp. *au début de la phase de balancement*) sont similaires entre plusieurs cycles de marche détectés chez un même individu" et la **Règle** ♯5 "Les orientations observées au *début de la phase d'appui* (resp. *au début de la phase de balancement*) sont relativement similaires entre les cycles de marche détectés chez des individus différents."

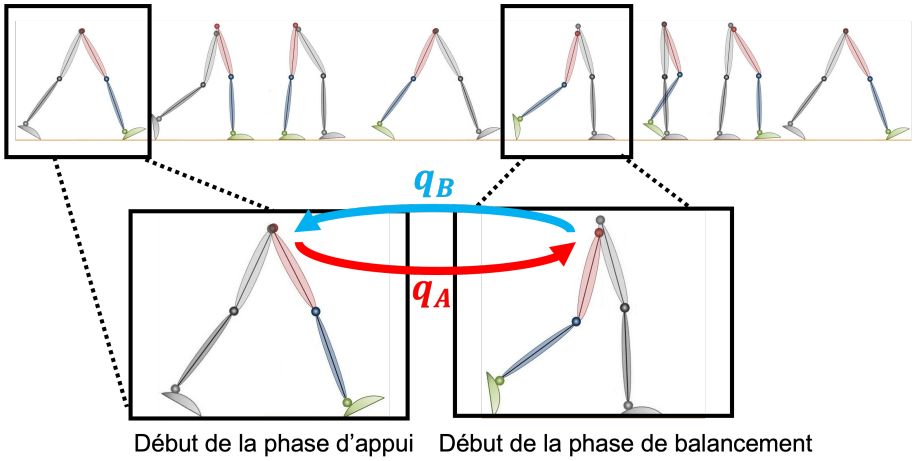


FIGURE 2.5 – Rotation de référence $\mathbf{q}_A$ et $\mathbf{q}_B$

Prenant ces observations en compte, on considère un cycle de marche *typique* représenté en haut de la figure 2.5, et plus particulièrement l'orientation de la hanche observée au début de la *phase d'appui* et celle observée au début de la *phase balancement*. On définit alors deux rotations dites *de référence* :

— La première est la rotation entre les orientations de la hanche observées au début et à la fin de la phase d'appui (*i.e.* au début de la phase de balancement), appelée *rotation d'appui de référence* (représentée par la flèche rouge sur la figure 2.5).

— La seconde est la rotation entre les orientations de la hanche observées au début et à la fin de la phase de balancement (*i.e.* au début de la phase de d'appui), appelée *rotation de balancement de référence* (représentée par la flèche bleue sur la figure 2.5).

Les orientations de la hanche étant exprimées dans un espace à 3 dimensions, ces deux rotations de référence peuvent être exprimées sous forme de quaternions unitaires. On note $\mathbf{q}_A = (w_A, x_A, y_A, z_A)$ et $\mathbf{q}_B = (w_B, x_B, y_B, z_B)$ les deux quaternions unitaires représentant respectivement la *rotation d'appui de référence* et la *rotation de balancement de référence*. On rappelle qu'une rotation en 3 dimensions dépend de deux paramètres : son axe $\mathbf{u}$ et son angle de rotation θ (*c.f.* section 1.2.2.2). On suppose que la variabilité inter-individuelle de la démarche peut se traduire par une différence dans l'*amplitude du mouvement* de la hanche, *i.e.* une différence de la valeur de l'angle θ de la rotation d'appui (resp. de balancement) entre individus. On suppose cependant que l'axe de la rotation d'appui (resp. de balancement) est similaire entre les individus (*c.f.* **Règle** ♯5). Ces axes sont notés $\mathbf{u}_A$ pour la rotation d'appui de référence et $\mathbf{u}_B$ pour la rotation de balancement de référence.

On souhaite donc estimer les *axes des rotations de référence* $\mathbf{u}_A$ et $\mathbf{u}_B$ pour les comparer avec les axes des rotations entre deux *points de segmentation* successifs identifiés par l'algorithme 2 `WCDSW` dans les données de marche d'un individu. Cette comparaison permettrait d'identifier à quels évènements correspondent les *points de segmentation* : début de phase d'appui, début de phase de balancement ou aucun de ces deux évènements.

Estimation des axes de rotation de référence $\mathbf{u}_A$ et $\mathbf{u}_B$ Les données de la base `BDDapp` (*c.f.* description de cette base de données section 1.2.3.1) sont utilisées pour estimer les coordonnées de $\mathbf{u}_A$ et $\mathbf{u}_B$. Pour ce faire, les *points de segmentation* correspondant au début des phases d'appui et de balancement sont identifiés en comparant les données aux vidéos acquises durant les tests de marche. Pour chaque volontaire, on note $\mathbf{q}_{s_{i,1}}$ et $\mathbf{q}_{s_{i,2}}$ deux quaternions correspondant à l'orientation de la hanche observée à deux *points de segmentation* successifs dans les données de marche. La rotation entre ces deux orientations est obtenue par :

$$\mathbf{q}_i^* = \mathbf{q}_{s_{i,1}}^{-1} \mathbf{q}_{s_{i,2}} \tag{2.13}$$

L'axe $\mathbf{u}_i^*$ de la rotation $\mathbf{q}_i^*$ est calculé par l'équation (1.15). Deux cas de figure peuvent être alors observés :

— Le quaternion $\mathbf{q}_{s_{i,1}}$ est observé au début d'une phase d'appui, et le quaternion $\mathbf{q}_{s_{i,2}}$

est observé au début de la phase de balancement qui lui succède. L'axe $\mathbf{u}_i^*$ de la rotation $\mathbf{q}_i^*$ est donc un axe de *rotation d'appui*.

— Le quaternion $\mathbf{q}_{s_{i,1}}$ est observé au début d'une phase de balancement, et le quaternion $\mathbf{q}_{s_{i,2}}$ est observé au début de la phase d'appui qui lui succède. L'axe $\mathbf{u}_i^*$ de la rotation $\mathbf{q}_i^*$ est donc un axe de rotation *de balancement*.

L'ensemble des axes de rotations d'appui et de balancement est ainsi identifié dans la base `BDDapp` en comparant les données mesurées avec les vidéos prises lors de l'acquisition des données. L'*axe de la rotation d'appui de référence* $\mathbf{u}_A$ (resp. *axe de la rotation de balancement de référence* $\mathbf{u}_B$) est estimé par le calcul de l'axe médian de l'ensemble des axes de *rotation d'appui* (resp. axes de *rotation de balancement*) par l'alogrithme adapté aux données circulaires décrit par J.Cabrera et G.S. Watson (1990) [23].

La figure 2.6 représente la projection des coordonnées des axes de références $\mathbf{u}_A$ (triangle rouge) et $\mathbf{u}_B$ (triangle bleu) sur les plan XY et XZ. Les points rouges (resp. bleus) représentent l'ensemble des *axes de rotation d'appui* (resp. *de balancement*) identifiés dans la base `BDDapp`. Il est intéressant de noter que les coordonnées des axes $\mathbf{u}_A$ et $\mathbf{u}_B$ sont opposées. Cette observation était attendue, compte tenu du fait que les *rotation d'appui de référence* et *rotation de balancement de référence* sont supposées représenter des rotations "opposées". Si les axes de rotation d'appui et de balancement mesurés sur la `BDDapp` forment deux groupes distincts, on remarque qu'il existe malgré tout une certaine variabilité entre les cycles de marche mesurés chez plusieurs individus.

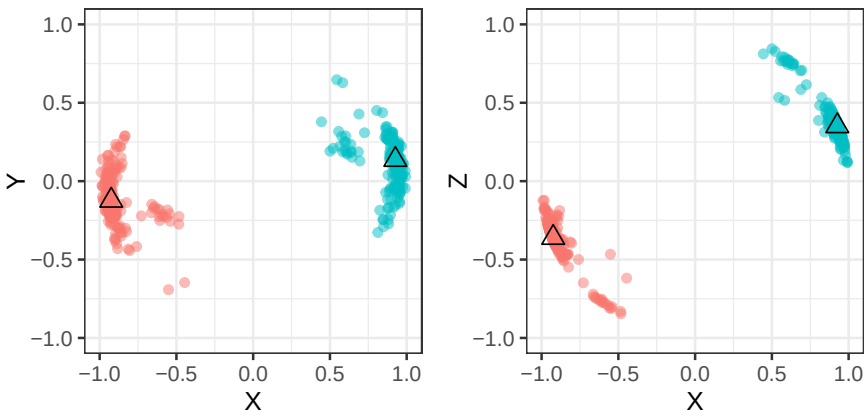


FIGURE 2.6 – Estimation des axes de référence $\mathbf{u}_A$ et $\mathbf{u}_B$ à partir de la `BDDapp`

Identification des phases d'appui et de balancement dans un nouveau jeu de données. On note :

- $Q = (\mathbf{q}_i, \ldots, \mathbf{q}_N)$: les données mesurées par le dispositif MMR sous la forme d'une séquence de quaternions unitaires mesurés aux temps $T = (t_1, \ldots, t_N)$.
- $S = \{\mathbf{s}_1 = (s_{1,1}, s_{1,2}, s_{1,3}), \ldots, \mathbf{s}_M = (s_{M,1}, s_{M,2}, s_{M,3})\}$: l'ensemble des triplets de *points de segmentation* identifiés dans Q par l'algorithme 2 `WCDSW`, auquel a été retiré l'ensemble des *outliers de durée*.
- $Q^* = \{\mathbf{q}_1^* = \mathbf{q}_{s_{1,1}}^{-1}\mathbf{q}_{s_{1,2}}, \ldots, \mathbf{q}_M^* = \mathbf{q}_{s_{M,1}}^{-1}\mathbf{q}_{s_{M,2}}\}$: l'ensemble des rotations entre deux orientations observées à deux *points de segmentation* successifs.
- $U^* = \{\mathbf{u}_1^*, \ldots, \mathbf{u}_M^*\}$: l'ensemble des axes des rotations Q^* calculés par l'équation (1.15).

$\mathbf{u}_i, i \in \{1, \ldots, M\}$ représente donc l'axe de la rotation entre les orientations $\mathbf{q}_{s_{i,1}}$ et $\mathbf{q}_{s_{i,2}}$ observées respectivement au premier et second point du triplet $\mathbf{s}_i$.

Étape 4.1 : Répartition des axes U^* en 2 groupes. Il a été constaté que les axes de *rotation d'appui* et *de balancement* mesurés sur la `BDDsep` se répartissent en 2 groupes distincts (*c.f.* figure 2.5). Le même phénomène est attendu pour les données représentant la démarche de tout individu. Les axes U^* sont donc répartis en 2 groupes U_1^* et U_2^* par *K-means*, avec $K = 2$.

Étape 4.2 : Calcul de l'axe médian des groupes U_1^* et U_2^* On souhaite déterminer un axe représentatif pour chacun des groupes U_1^* et U_1^*. À ce stade, certains des segments *candidats* formés à partir de l'ensemble de triplets de *points de segmentations* S peuvent encore être des *outliers*. Leur axe de rotation $\mathbf{u}_i^*$ peut donc être éloigné du reste des axes formant le groupe auquel ils ont été attribués par le $K - means$. Pour s'affranchir de la présence de potentiels *outliers*, l'axe représentatif du groupe U_1^* (resp. U_2^*) est défini comme son axe median $\tilde{\mathbf{u}}_1^*$ (resp. $\tilde{\mathbf{u}}_2^*$), obtenu par l'alogrithme adapté aux données circulaires décrit par J.Cabrera et G.S. Watson (1990) [23]. La figure 2.7 représente les deux groupes d'axe U_1^* et U_2^* identifiés à partir d'un jeu de données exemple. Les deux points en forme de triangle représentent leur axe median $\tilde{\mathbf{u}}_1^*$ et $\tilde{\mathbf{u}}_2^*$.

Étape 4.3 : Identification du type de rotation des axes U_1^* et U_2^*. Pour déterminer le type de rotation représenté par chacun des deux groupes U_1^* et U_2^*, deux cas de figure sont envisagés :

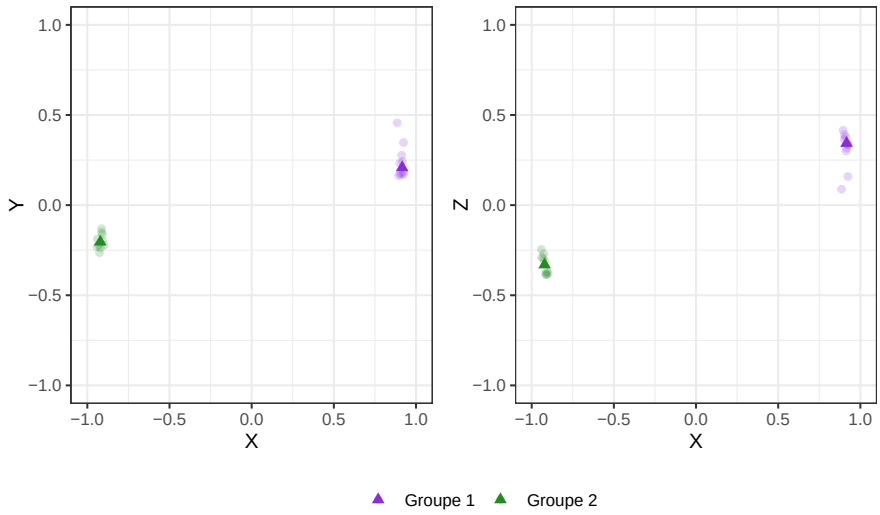


FIGURE 2.7 – Groupes U_1^* et U_2^* et leurs axes medians $\tilde{\mathbf{u}}_1^*$ et $\tilde{\mathbf{u}}_2^*$

— Cas 1 : U_1^* regroupe les axes de rotation de balancement et U_2^* les axes de rotation d'appui.

— Cas 2 : U_1^* regroupe les axes de rotation d'appui et U_2^* les axes de rotation de balancement.

Les cas 1 et 2 respectivement aux coûts c_1 et c_2 définis par (2.14) et (2.15)

$$c_1 = \frac{1 - (\tilde{\mathbf{u}}_1^* \cdot \mathbf{u}_B)}{2} + \frac{1 - (\tilde{\mathbf{u}}_2^* \cdot \mathbf{u}_A)}{2} \tag{2.14}$$

$$c_2 = \frac{1 - (\tilde{\mathbf{u}}_1^* \cdot \mathbf{u}_A)}{2} + \frac{1 - (\tilde{\mathbf{u}}_2^* \cdot \mathbf{u}_B)}{2}, \tag{2.15}$$

$(\tilde{\mathbf{u}}_1^* \cdot \mathbf{u}_B)$ étant le produit scalaire entre ces 2 vecteurs.

Les vecteurs $\tilde{\mathbf{u}}_1^*$, $\tilde{\mathbf{u}}_2^*$, $\mathbf{u}_A$ et $\mathbf{u}_B$ sont de norme 1. Le produit scalaire maximal entre deux de ces vecteurs vaut donc 1 dans le cas où ils sont co-linéaires, -1 s'ils sont opposés, et 0 s'ils sont orthogonaux. On suppose que le Cas 1 est vrai. Il est attendu que $(\tilde{\mathbf{u}}_1^* \cdot \mathbf{u}_B)$ soit proche de 1 et grand par rapport à $(\tilde{\mathbf{u}}_1^* \cdot \mathbf{u}_A)$. Par conséquent, le premier terme de la somme (2.14) est d'autant plus proche de 0 que les vecteurs ($\tilde{\mathbf{u}}_1^*$ et $\mathbf{u}_B$) sont proches, et le premier terme de la somme (2.15) est d'autant plus grand que les vecteurs ($\tilde{\mathbf{u}}_1^*$ et $\mathbf{u}_A$) sont éloignés. Il est aussi attendu que $(\tilde{\mathbf{u}}_2^* \cdot \mathbf{u}_B)$ soit petit par rapport à $(\tilde{\mathbf{u}}_2^* \cdot \mathbf{u}_A)$. De la même façon, le second terme de la somme (2.14) sera proche de 0 et le second terme de la somme (2.15) sera grand. Les coûts c_1 et c_2 sont donc bornés entre 0 et 2, et sont

d'autant plus faibles que le cas auquel ils sont associés est probable. Ainsi, le cas associé au coût le plus faible est considéré comme vrai.

La figure 2.8 présente les deux groupes d'axes précédemment identifiés à l'étape 4.2, ainsi que les coordonnées de l'axe type d'une rotation d'appui $\mathbf{u}_A$ (triangle rouge) et celui d'une rotation de balancement $\mathbf{u}_B$ (triangle bleu). Dans cet exemple l'axe $\tilde{\mathbf{u}}_1^*$ est proche de $\mathbf{u}_B$, et $\tilde{\mathbf{u}}_2^*$ est proche de $\mathbf{u}_A$. De ce fait, le coût c_1 est plus faible que c_2. Le premier groupe rassemble donc les axes de rotation de *balancement*, alors que le second correspond aux axes de rotations d'*appui*.

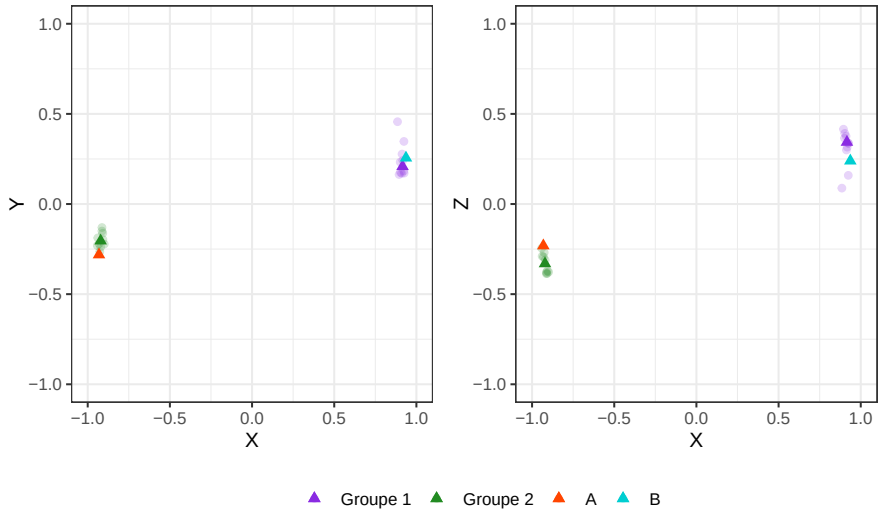


FIGURE 2.8 – Axes medians et axes types $\mathbf{u}_A$ et $\mathbf{u}_B$

Étape 4.4 : Identification des *outliers d'axe*. Comme énoncé dans l'étape 4.2, certains des *segments candidats* formés par les triplets de points sont des *outliers*. Une partie de ces *outliers* peut être identifiée en comparant les coordonnées de leurs axes de rotation avec celles de l'axe médian de leur groupe $\tilde{\mathbf{u}}_g^*$. Les axes ayant des coordonnées *éloignées* de l'axe médian du groupe auquel ils sont attribués sont considérés comme *outliers d'axe*. Un hyperparamètre $\theta_{\mathbf{u}}$ est défini pour identifier ces *outliers*. Il représente une *valeur-seuil* associée au produit scalaire $\mathbf{u}_i^* \cdot \tilde{\mathbf{u}}_g^*$. Si $\mathbf{u}_i^* \cdot \tilde{\mathbf{u}}_g^*$ est inférieur à $\theta_{\mathbf{u}}$, $\mathbf{u}_i^*$ est trop éloignés de $\tilde{\mathbf{u}}_g^*$ pour considérer que le *point de segmentation* $s_{i,1}$ corresponde au début d'une phase d'appui ou de balancement. Du fait de la variabilité de la démarche intra et inter individuelle, une dispersion des coordonnées des axes de rotation autour de leur axe médian est attendue dans les données mesurées en vie réelle. La valeur de $\theta_{\mathbf{u}}$

doit donc être définie de telle sorte que ce critère ne soit pas trop restrictif, *i.e.* qu'un nombre trop important de segments correspondant à des cycles de marche ne soit considéré comme *outlier*. Les triplets de *points de segmentation* $\mathbf{s}_i =$ dont les axes de la rotation $\mathbf{q}_1^* = \mathbf{q}_{s_{i,1}}^{-1} \_ s_{i,2}$ sont trop éloignés de l'axe médian de leur groupe sont considérés comme *outlier d'axe* et sont retirés de la liste des *segments candidats* S.

La figure 2.9a représente un exemple d'axes identifiés dans un jeu de données. Les points de segmentation correspondant à ces axes sont matérialisés sur figure 2.9b. Les points rouges correspondent au groupe des axes de rotation d'appui. Les points bleus correspondent au groupe des axes de rotation de balancement. Les axes médians de ces deux groupes sont représentés respectivement par les triangles rouges et bleus. Les zones rouge et bleue correspondent aux portions de la sphère définies par le paramètre $\theta_{\mathbf{u}}$. Si les coordonnées d'un axe $\mathbf{u}_i^*$ appartenant à un groupe g (par exemple le groupe A) ne sont pas comprises dans la zone définie autour de son axe médian (par exemple la zone rouge), alors $\mathbf{u}_i^* \cdot \tilde{\mathbf{u}}_g^* < \theta_{\mathbf{u}}$. Le triplet correspondant est considéré comme un `outlier d'axe`.

L'ensemble des triplets de points de segmentation $\mathbf{S}$ est désormais réparti en 3 sous ensemble, en fonction de la nature de la période délimitée par leur premier et deuxième *point de segmentation*. On note :

— S_A : les triplets pour lesquels cette période correspond à une phase d'appui.

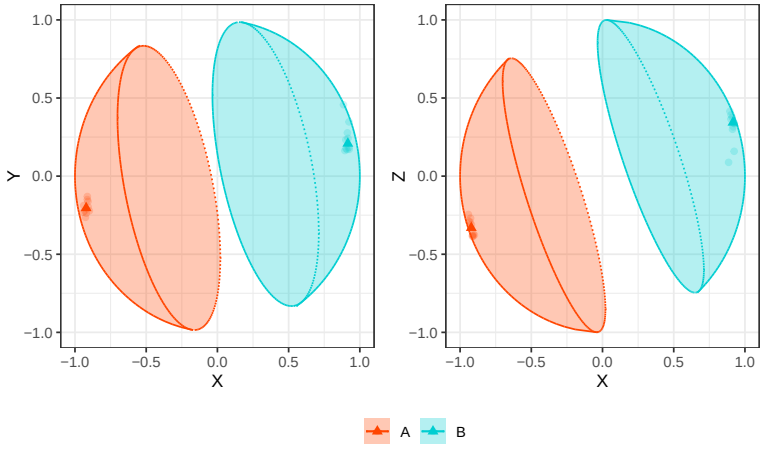
— S_B : les triplets pour lesquels cette période correspond au début d'une phase de balancement.

— S_{out} : les triplets pour lesquels cette période ne correspond pas à une phase d'un cycle de marche.

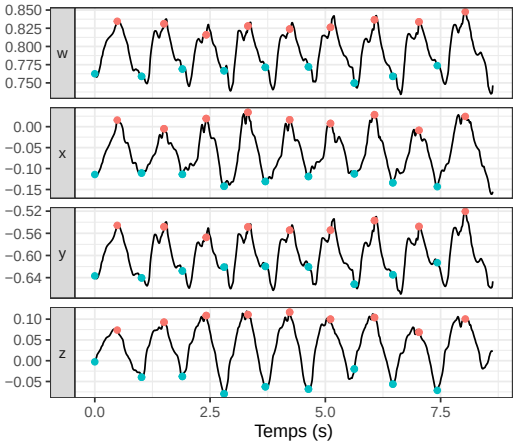
Les triplets de l'ensemble S_A sont utilisés dans la suite de l'algorithme pour former des *segments candidats* pouvant être des cycles de marche débutant par une phase d'appui suivie d'une phase de balancement. On note $M \leftarrow |S_A|$ le nombre de segments candidats restant, correspondant au cardinal de l'ensemble S_A.

2.1.1.6 Étape 5 : Segmentation des cycles de marche

Étape 5.1 : Segmentation du signal. Cette étape a pour but de former les segments correspondant aux cycles de marche commençant par une phase d'appui à partir du jeu de données de marche Q. Ces segments sont délimités par les premiers et derniers points de segmentation des triplets S_A identifiés par la méthode décrite dans la section 2.1.1.5. Pour ce faire, l'algorithme 3 `Get Segments` est appliqué sur Q.



(a) Axe de rotation et seuil $\theta_{\mathbf{u}}$



(b) Points de segmentation et type de rotation

FIGURE 2.9 – Points de segmentation et type de rotation

L'algorithme 3 `Get Segments` renvoie un ensemble de M segments. Un segment est noté $Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N_i})^\top$, avec $i \in \{1 \ldots M\}$ et $N_i = s_{i,1} - s_{i,3} + 1$ définissant le nombre de quaternions unitaires contenus dans le segment. La grille des temps de mesure de ce segment est notée $T_i = (t_{i,1} \ldots t_{i,N_i})$, avec $t_{i,1} = 0$ et t_{i,N_i} correspondant à la durée du segment.

La figure 2.10 présente un exemple de segmentation d'un jeu de données par l'algorithme 3 `Get Segments` à partir des points de segmentation correspondant au début d'une phase d'appui. Les points de segmentation correspondant au début d'une phase d'appui sont représentés sur les données de marche par les points rouges sur la figure 2.10a. Les segments formés par l'algorithme 3 `Get Segments` sont visibles sur la figure 2.10b.

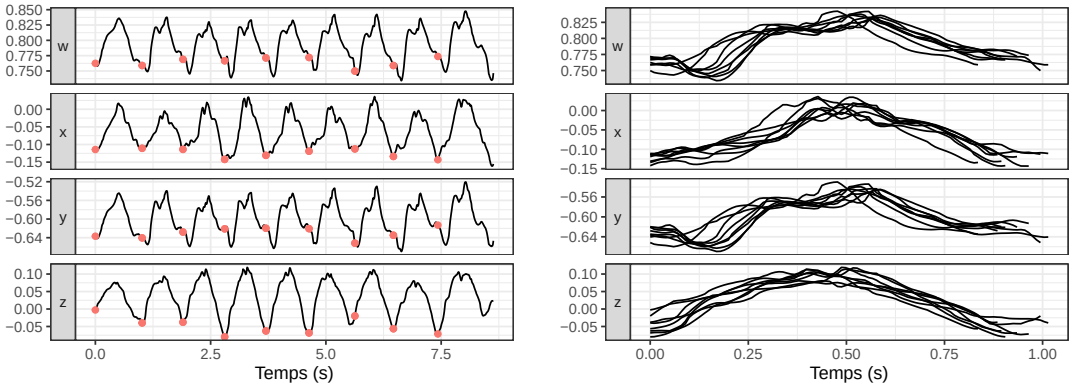
Algorithm 3 `Get Segments`

Paramètres
$Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$: les données mesurées par le dispositif MMR sous la forme d'une séquence de quaternions unitaires mesurés aux temps $T = (t_1, \ldots, t_N)$.
$S_A = \{\mathbf{s}_1 = (s_{1,1}, s_{1,2}, s_{1,3}), \ldots, \mathbf{s}_M = (s_{M,1}, s_{M,2}, s_{M,3})\}$: l'ensemble des triplets de *points de segmentation* permettant de former les segments *candidats* pouvant correspondre à des cycles de marche débutant par une phase d'appui.
Initialisation
$i \leftarrow 1$
Début
while $i \leq M$ **do**
  $Q_i \leftarrow \left(\mathbf{q}_{s_{i,1}}, \mathbf{q}_{(s_{i,1}+1)}, \ldots, \mathbf{q}_{s_{i,3}}\right)$
  $T_i \leftarrow \left(t_{s_{i,1}}, t_{(s_{i,1}+1)}, \ldots, t_{s_{i,3}}\right) - t_{s_{i,1}}$
  $i \leftarrow i + 1$
end while
Fin

Étape 5.2 : Ré-orientation des segments. Pour permettre la comparabilité des segments identifiés dans des jeux de données mesurés chez un même individu au cours de différentes sessions d'acquisition, l'orientation de la hanche doit être exprimée à partir d'un référentiel fixe équivalent entre les segments. Pour ce faire, le quaternion unitaire moyen de chaque segment i est calculé par l'équation 3.9 (*c.f.* section 1.2.2.1) : $\bar{\mathbf{q}}_i = \text{avg}(Q_i)$. Chaque segment est ensuite *réorienté* sur son quaternion unitaire moyen, *i.e.* l'orientation du système de capteurs dans le segment $Q_i, \forall i \in \{1, \ldots, M\}$ est calculée à partir de l'orientation moyenne du segment :

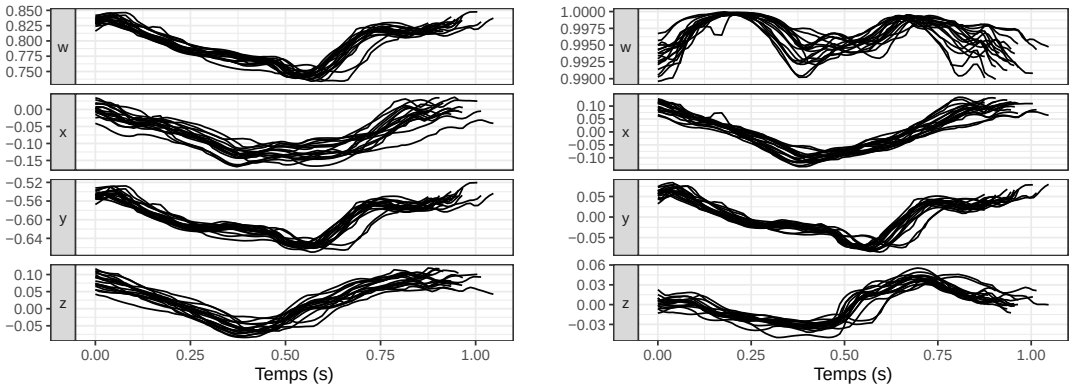
$$\mathbf{q}_{i,j} \leftarrow \bar{\mathbf{q}}_i^{-1} \mathbf{q}_{i,j}, \; j \in \{1, \ldots, N_m\}, \; i \in \{1, \ldots, M\} \tag{2.16}$$

La figure 2.11 présente un exemple de réorientation de segments. On peut remarquer que les segments après réorientation se rapprochent du quaternion unitaire représentant la rotation identité $\mathbf{q}_0 = (1, 0, 0, 0)$. Ceci confirme l'**Hypothèse 4** formulée en préambule de cette section. En effet, l'orientation de la hanche au cours d'un cycle de marche réorienté est calculée en considérant son orientation moyenne comme référence. La hanche se rapproche de cette orientation moyenne par deux fois au cours du cycle, la première durant la phase d'appui, la seconde durant la phase de balancement, aux instants où les deux jambes se croisent.



(a) Données de marche et *points de segmentation* (b) Segments formés par l'algorithme 3 `Get Segments`

FIGURE 2.10 – Application de l'algorithme 3 `Get Segments`



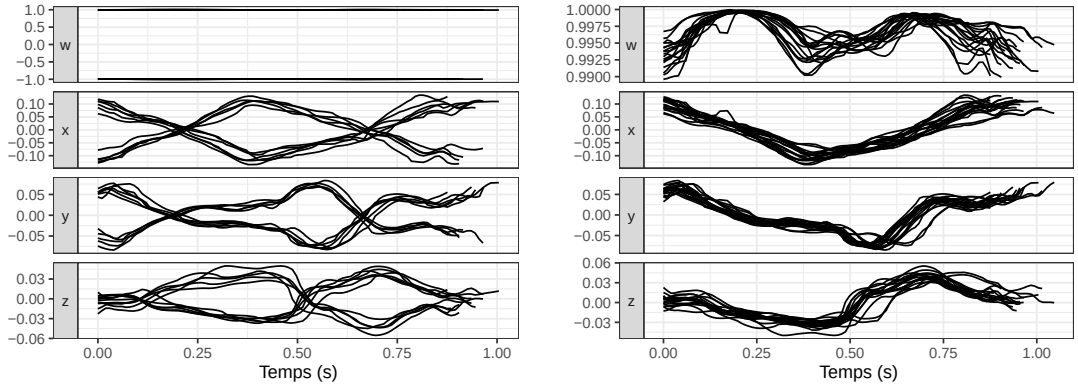
(a) Cycles de marche avant réorientation (b) Cycles de marche après réorientation

FIGURE 2.11 – Réorientation des cycles de marche

Étape 5.3 : Hémisphérisation des segments. Dans certains cas, les segments formés à partir de jeux de données mesurés chez un même individu au cours de sessions d'acquisitions différentes sont constitués de quaternions unitaires situés sur 2 hémisphères différents de la 3-sphère $\mathcal{S}^3$ (voir l'exemple présenté dans la figure 2.12a).

On rappelle que l'ensemble des quaternions unitaires forme une algèbre de Lie isomorphe au groupe unitaire spécial $SU(2)$, lequel étant exactement deux fois plus grand que le groupe des rotations en trois dimensions $SO(3)$ (*c.f.* 1.2.2.2). Par conséquent un quaternion unitaire $\mathbf{q}$ et son opposé $-\mathbf{q}$ encodent la même rotation 3D (voir 1.2.2.1). Les segments mesurés chez un même individu sont transformés pour s'assurer qu'ils soient constitués de quaternions unitaires situés dans le même hémisphère de $\mathcal{S}^3$. L'hémisphère a été choisi arbitrairement comme celui constitué de quaternions unitaires dont la com-

posante w est positive. Les quaternions unitaires des segments présentant des valeurs négatives de leur composante w sont donc remplacés par leur quaternion unitaire opposé. Un exemple de cette transformation est présenté sur la figure 2.12b.



(a) Cycles de marche avant hémisphérisation (b) Cycles de marche après hémisphérisation

FIGURE 2.12 – Hémisphérisation des cycles de marche

2.1.1.7 Étape 6 : Suppression des outliers de forme

À ce stade, il peut rester des *outliers* parmi les M segments *candidats* restants. La méthode développée pour les identifier s'appuie sur la **Règle** ♯6 : "Les changements d'orientation de la hanche observés au cours d'un cycle de marche sont similaires entre plusieurs cycles d'un même individu". On admet donc que la forme des segments *outliers* diffère de celle des segments correspondant à des cycles de marche, supposés présenter des formes similaires. La méthode d'identification des *outliers* restants est composée des étapes présentées ci-dessous.

Étape 6.1 : Calcul de la matrice de dissimilarité des segments. La première étape consiste à définir la matrice des dissimilarités entre paire de segments notée D de dimension $M \times M$. Ici, on souhaite comparer la *forme* des segments *candidats*. Un mauvais alignement temporel peut être observé entre ces derniers. Celui ci peut être dû à une imprécision dans la détection des *points de segmentation* par l'algorithme 2 `WCDSW`, ou à la variabilité naturelle de la démarche observée chez un individu. C'est pourquoi la dissimilarité entre un segment i et un segment j est déterminée par la méthode *Quaternion Dynamic Time Warping* (QDTW) décrite par Jablonski (2011)[82]. La méthode QDTW généralise le DTW aux séries temporelles de quaternions unitaires et est présentée plus

en détail dans la section 3.1.2.1. En résumé, elle permet de s'affranchir des potentiels décalages entre deux séries par alignement temporel de leurs éléments. On définit donc :

$$D[i, j] = \text{QDTW}(Q_i, Q_j)$$

pour tout $i, j' \in \{1, \ldots, M\}$.

Étape 6.2 : Regroupement hiérarchique des segments. Par la suite, un algorithme de Classification Ascendante Hiérarchique (CAH) est appliqué à la matrice D avec le critère de *liaison simple* pour former un dendrogramme $\mathcal{T}$. Il est attendu que les cycles de marche forment un groupe compact lors des premières étapes de construction de $\mathcal{T}$. À l'inverse, il est attendu que les segments *outliers* soient inclus dans le groupe des cycles de marche à des étapes tardives de la construction de $\mathcal{T}$.

Étape 6.3 : Choix du nombre de groupes. Le choix du nombre de groupes à former à partir d'un dendrogramme d'une CAH est un problème encore non résolu dans le domaine de la classification non supervisée. La plupart des critères, tels que ceux présentés dans la section 1.3.1.2, cherchent à minimiser la dissimilarité entre les observations d'un même groupe et maximiser la dissimilarité entre des observations de groupes distincts [124]. Cependant, dans le cas considéré ici, le seul groupe d'intérêt est celui formé par les cycles de marche. Il importe donc peu que les autres groupes soient des *singletons* (i.e. constitués d'une unique observation) ou constitués de segments présentant une forte dissimilarité entre eux. Les critères d'évaluation de la qualité d'une partition traditionnellement utilisés ne sont donc pas appropriés. Ainsi, le critère décrit ci-après a été développé pour déterminer à quelle hauteur stopper la construction de $\mathcal{T}$. Soit $\mathbf{h}$ le vecteur des hauteurs des branches du dengrogramme $\mathcal{T}$. La hauteur à laquelle couper $\mathcal{T}$ est déterminée par :

$$\hbar = \tilde{h} + (\text{Q3}(\mathbf{h}) - \text{Q1}(\mathbf{h})) \times \delta_{\text{IQR}} \tag{2.17}$$

avec $\tilde{h}$ la médiane de $\mathbf{h}$, $\text{Q1}(\mathbf{h})$ et $\text{Q3}(\mathbf{h})$ respectivement les premier et troisième quartiles de $\mathbf{h}$ et δ_{IQR} un hyperparamètre permettant de régler la permissivité du critère (IQR pour *Inter Quartile Range*). Plus la valeur de δ_{IQR} est grande, plus le dernier segment intégré au groupe des cycles de marche pourra présenter une dissimilarité élevée avec ledit groupe. Le choix du critère de liaison simple permet de tirer partie de sa tendance à produire des groupes étendus avec l'ajout d'observation une à une à chaque étape de la

construction de $\mathcal{T}$, aussi décrit comme "effet de chaînage" [65].

Étape 6.4 : Identification du groupe des cycles de marche. Dans l'hypothèse selon laquelle les données sont mesurées pendant une période durant laquelle l'activité de l'individu correspond majoritairement à de la marche, les segments formés à ce stade de l'analyse correspondent en majorité à des cycles de marche. Ces derniers étant similaires, l'identification des segments correspondant à des cycles de marche revient à sélectionner le groupe de la partition obtenu en coupant $\mathcal{T}$ à la hauteur $\hbar$ qui contient le plus grand nombre d'observations.

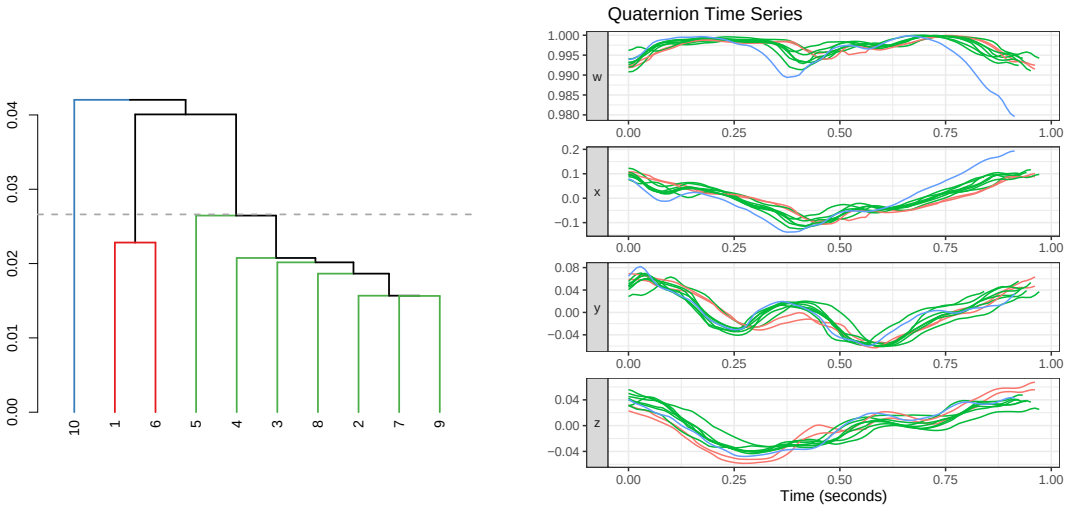
La figure 2.13 présente l'application de cette méthode pour différencier les cycles de marche des *outliers* d'un jeu de données exemple. 10 segments pouvant potentiellement être des cycles de marche sont identifiés à ce stade de l'analyse des données. Le dendrogramme de la figure 2.13a a été obtenu par CAH appliquée à ces segments avec le critère de liaison simple. Le numéro identifiant les segments est représenté sur les feuilles de l'arbre $\mathcal{T}$. La limite en pointillés symbolise la valeur de $\hbar$ calculée par (2.17), en fixant $\delta_{\text{IQR}} = 0.75$. 3 groupes sont formés, le premier étant un singleton constitué du segment 10, le second rassemble les segments 1 et 6 et le reste des segments forme le dernier groupe. La distribution des segments en groupe est représentée sur la figure 2.13b. La comparaison des segments avec les vidéos prises durant l'acquisition de ces données permet de déterminer que le segment 10 correspond à l'arrêt de la marche du volontaire, durant lequel il effectue un demi tour pour se retrouver face au parcours. Les segments 1 et 6 correspondent à la mise en marche du volontaire, au début de laquelle il se tient en position statique, les 2 pieds alignés. Enfin, le groupe composé du plus de segments (en vert) correspond aux cycles de marche réalisé durant le test.

Les différentes étapes précédemment présentées permettent l'identification des cycles de marche dans les jeux de données mesurées par le dispositif MMR. L'algorithme qu'elles constituent, appelé `STRIde PAttern GEneration` (`STRIPAGE`), prend en entrée :

— Une série temporelle de N quaternions unitaires $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$ représentant l'orientation en 3 dimensions de la hanche au cours de la marche.

— Un hyper-paramètre δ_{ID} définissant la zone de rejet des outliers de durée.

— Un hyper-paramètre $\theta_{\mathbf{u}}$ définissant la zone de rejet des outliers d'axe.



(a) CAH (b) Segments par groupe

FIGURE 2.13 – Identification du groupe des cycles de marche

— Un hyper-paramètre δ_{IQR} définissant la hauteur de rejet des outliers de forme.

À partir de ces entrées, `STRIPAGE` détermine un ensemble de M sous-séquences $Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N_i})$, chacune étant constituée de N_i quaternions unitaires, associés avec leurs temps de mesure $T_i = (t_{i,1} \ldots, t_{i,N_i})$. Ces séries correspondent aux cycles de marche du porteur du dispositif MMR, la première phase étant une phase d'appui du pied droit, et la seconde étant une phase de balancement.

Les étapes de l'algorithme `STRIPAGE` sont résumées dans la figure 2.14.

2.1.2 Évaluation de l'algorithme `STRIPAGE`

Cette section présente l'évaluation des performances de l'algorithme `Stride Pattern Generation` pour détecter les cycles de marche pied droit dans un jeu de données mesurées par le dispositif MMR placé à la ceinture du porteur en position latérale droite.

2.1.2.1 Plan d'expérience

Présentation des données. Les données de marche mesurées de la base de données `BDDapp` sont utilisées pour évaluer l'algorithme `STRIPAGE`. On rappelle que cette base est constituée des données de marche mesurées par le dispositif MMR chez 11 volontaires sains. Ces volontaires ont réalisé deux tests dans deux conditions différentes, l'une en condition de marche libre et l'autre en condition de marche contrainte. La condition de

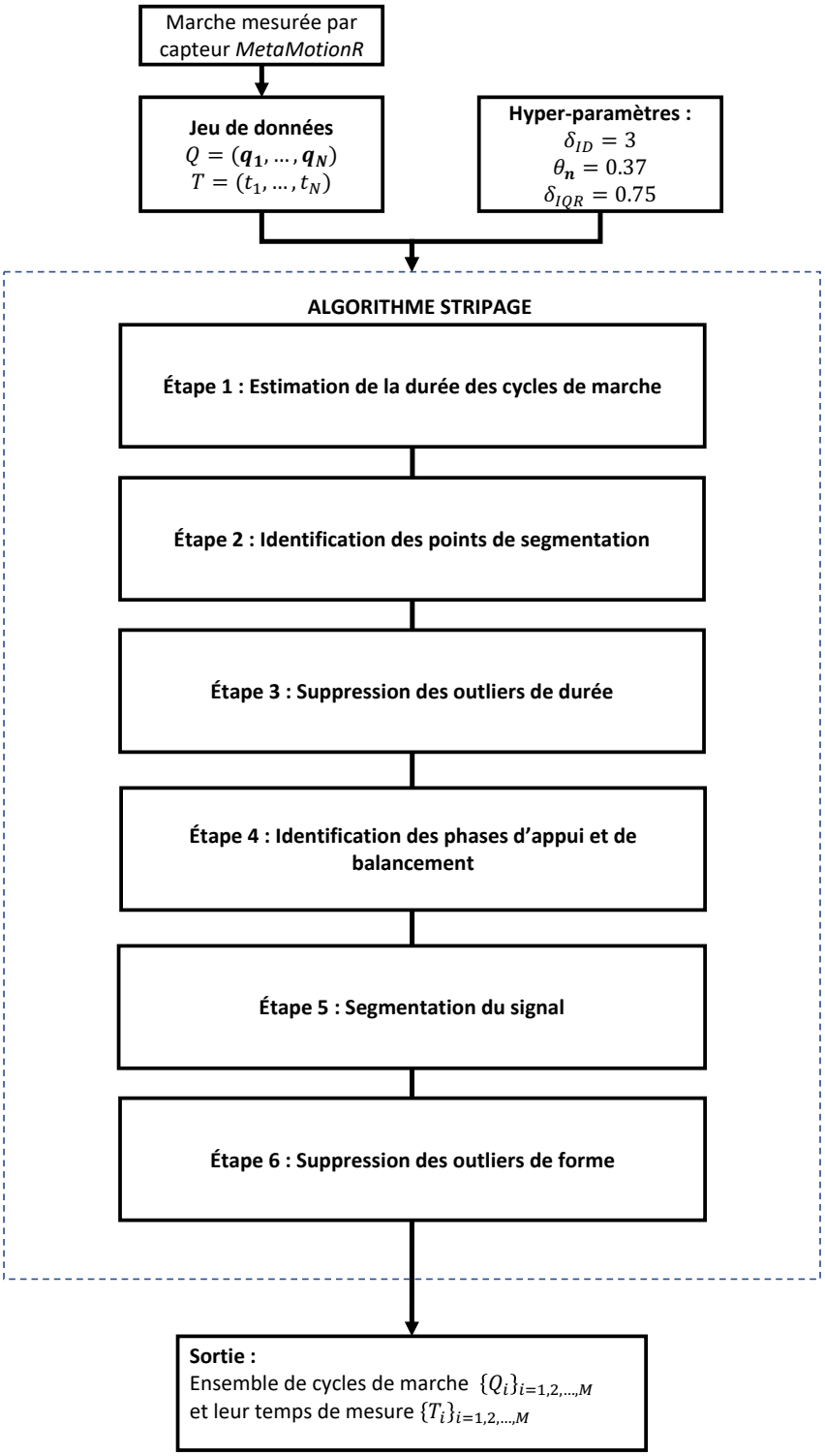


FIGURE 2.14 – Résumé de l'algorithme STRIPAGE

marche contrainte consiste à porter une orthèse bloquant d'articulation du genou droit. Ainsi, les performances de l'algorithme `STRIPAGE` sont évaluées à partir de données de marche mesurées chez des individus sains et chez des individus présentant un déficit de la marche simulé.

Traitement des données. Les données de marche sont traitées par l'algorithme `STRIPAGE` avec les hyperparamètres prenant les valeurs suivantes :

— $\delta_{\text{ID}} = 3$

— $\theta_{\mathbf{u}} = 0.37$

— $\delta_{\text{IQR}} = 0.75$

Ces valeurs ont été définies empiriquement lors du développement de l'algorithme `STRIPAGE`, à partir du jeu de données de la base de données `BDDapp`.

Les vidéos prises durant l'acquisition des données de la base de données `BDDapp` permettent d'identifier le nombre de cycles de marche pied droits réalisés par les volontaires durant les acquisitions de données de marche. Les performances de l'algorithme `STRIPAGE` sont donc évaluées par rapport à la *vérité de terrain*.

Critères d'évaluation. La comparaison entre les cycles de marche identifiés par l'algorithme `STRIPAGE` et ceux identifiés dans les vidéos permet le calcul des indicateurs suivants :

— Le nombre de cycles de marche comptés sur la vidéo : N_c

— Le nombre de cycles de marche correctement identifiés (*True Positive*) : TP.

— Le nombre de cycles de marche non identifiés (*False Negative*) : FN.

— Le nombre de segments considérés à tort comme des cycles de marche (*False Positive*) : FP.

À partir de ces indicateurs, l'évaluation des performances de l'algorithme se base sur le calcul des critères suivants :

— La sensibilité (ou *True Positive Rate*) : $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$

— La précision (ou *Positive Predictive Value*) : $\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}$

— L'exactitude, déterminée comme le score $\text{F}_1 = 2 \times \frac{\text{PPV} \times \text{TPR}}{\text{PPV+TPR}}$

— Erreur relative (en %) : $\text{Err} = \frac{\text{FP+FN}}{N_c} \times 100$

Tous ces critères prennent une valeur comprise dans l'intervalle $[0, 1]$, la valeur 1 étant associée à la meilleure performance possible. Ces critères sont calculés pour chacun des jeux de données de la base `BDDapp`, et les performances globales de l'algorithme `STRIPAGE` sont évaluées par la moyennes des indicateurs.

2.1.2.2 Résultats

Le tableau 2.1 présente les résultats obtenus avec l'algorithme de détection des cycles de marche sur les données de la base `BDDapp`. Le nombre de cycles de marche identifiés à partir des vidéos est renseigné dans la colonne "Cycle Vidéo". La colonne "Cycles `STRIPAGE`" renseigne le nombre de cycle de marche détectés par l'algorithme `STRIPAGE`. Les valeurs des indicateurs TP, FP et FN, ainsi que celles des critères TPR, PPV et F_1 sont renseignées pour les tests de chaque volontaire dans chaque condition. Leur valeur est aussi donnée pour l'ensemble des tests par condition, et sur l'ensemble des données de la base `BDDapp`.

TABLE 2.1 – Performances de l'algorithme `STRIPAGE` sur la `BDDapp`

Volontaire	**Condition**	**Cycles vidéo**	**Cycles `STRIPAGE`**	**TP**	**FP**	**FN**	**TPR**	**PPV**	**F1**	**Err**
V1	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	10	9	9	0	1	0,9	1	0,95	10
V2	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	11	9	9	0	2	0,818	1	0,9	18,182
V3	Libre	8	7	7	0	1	0,875	1	0,93	12,5
	Contrainte	8	7	7	0	1	0,875	1	0,93	12,5
V4	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	8	8	8	0	0	1	1	1	0
V5	Libre	8	7	7	0	1	0,875	1	0,93	12,5
	Contrainte	8	9	8	1	0	1	0,889	0,94	0
V6	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	8	6	6	0	2	0,75	1	0,86	25
V7	Libre	10	9	9	0	1	0,9	1	0,95	10
	Contrainte	10	10	10	0	0	1	1	1	0
V8	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	9	7	7	0	2	0,778	1	0,88	22,222
V9	Libre	8	5	5	0	3	0,625	1	0,77	37,5
	Contrainte	8	6	6	0	2	0,75	1	0,86	25
V10	Libre	8	8	8	0	0	1	1	1	0
	Contrainte	8	7	7	0	1	0,875	1	0,93	12,5
V11	Libre	8	5	5	0	3	0,625	1	0,77	37,5
	Contrainte	8	8	8	0	0	1	1	1	0
Total	Libre	90	81	81	0	9	0,9	1	0,95	10
	Contrainte	96	86	85	1	11	0,885	0,988	0,93	11,458
	Global	186	167	166	1	20	0,892	0,994	0,94	10,753

La précision de l'algorithme `STRIPAGE` est parfaite en condition de marche libre (PPV = 1), ce qui signifie que tous les segments qu'il considère comme des cycles de marche correspondent effectivement à des cycles réalisés par le volontaire. La précision de l'algorithme est très bonne pour la condition de marche contrainte (PPV = 0.988), un seul segment identifié comme cycle de marche étant un faux positif (voir Volontaire **V5** dans le tableau 2.1). La sensibilité de l'algorithme est plus variable que la précision. Elle est parfaite pour 6 volontaires en condition de marche normale et pour 3 volontaires en condition de marche contrainte. La plus faible précision (TPR = 0.625) est observée chez les volontaires $V9$ et $V11$ en condition de marche libre, avec 5 cycles de marche identifiés sur 8 dans les deux cas. Cependant, la sensibilité globale de `STRIPAGE` reste légèrement meilleure en condition de marche libre (TPR = 0.9) qu'en marche contrainte (TPR = 0.885). De manière générale, l'algorithme `STRIPAGE` favorise la précision (PPV = 0.994) au détriment de la sensibilité (TPR = 0.892), conduisant à une exactitude plutôt élevée (F_1 = 0.94). L'exactitude de l'algorithme est satisfaisante pour les deux conditions de marche. Elle est légèrement supérieure en condition de marche libre (F_1 = 0.95) qu'en condition de marche contrainte (F_1 = 0.93).

2.1.2.3 Discussion

L'algorithme `STRIPAGE` permet la détection des cycles de marche dans les données mesurées par le dispositif MMR sous forme de quaternions unitaires. Il permet une détection des cycles de marche de manière automatisée, et ne nécessite l'intervention de l'utilisateur que pour déterminer les valeurs des hyperparamètre δ_{ID}, $\theta_{\mathbf{u}}$ et δ_{IQ}, pour lesquels des valeurs par défaut sont proposées. L'identification des segments représentant les cycles de marche reposent sur des hypothèses a-priori qui dérivent de propriétés de la marche humaine, telles que les mouvements de rotation des hanches ou la régularité de la démarche des individus.

Beaucoup de méthodes sont décrites pour la détection d'évènements du cycle de marche, et ce domaine de recherche est très actif [129]. Cependant, aucun algorithme de détection des cycles de marche a partir de données représentant l'orientation en 3D de la hanche mesurée par un unique système de capteurs portatif n'a été identifié dans la littérature. Il n'est donc pas possible de comparer directement les performances de `STRIPAGE` avec celles d'un autre algorithme appliqué sur les données de la `BDDsep`. En revanche, les critères de validation présentés dans la section 2.1.2.2 sont aussi utilisés pour évaluer plusieurs méthodes de détection de cycles de marche à partir de données de

capteurs de mouvements. [162] Bien que l'algorithme STRIPAGE présente de très bonnes performances avec une sensibilité très satisfaisante et une précision très élevée, elles restent moins élevées que celles de certains algorithmes présentés dans la littérature. Par exemple, la méthode décrite par Ghersi *et al.* (2020) [50], également basée sur l'identification de points caractéristiques du signal, présente un score F_1 proche de 0.999. Ces résultats ont cependant été obtenus à partir de données d'un accéléromètre placé à la ceinture. L'algorithme intègre entre outre une étape de correction des cycles de marche détectés, fusionnant les segments trop courts et divisant les segments trop longs. L'adaptation de ces étapes aux séquences de quaternions unitaires pourraient donc améliorer les performances de l'algorithme STRIPAGE. De plus, la base de données servant à l'évaluation de la méthode proposée par Ghersi *et al.* contient plus de 3600 cycles mesurés chez 11 sujets différents, là où la BDDapp contient entre 8 et 10 cycles de marche mesurés chez 11 sujets dans deux conditions différentes, pour un total de 186 cycles. Ce plus faible nombre de cycles entraîne donc une diminution plus importante de la sensibilité, spécificité et du score F_1 pour les jeux de données dans lesquels 1 ou 2 erreurs de mesure (*i.e.* faux positifs ou faux negatifs) sont observées.

Plusieurs perspectives de recherche peuvent être identifiées. Tout d'abord, l'estimation de la durée des cycles de marche nécessite de calculer le périodogramme sur 4 séries temporelles correspondant aux valeurs prises par les composantes w, x, y et z. La généralisation du périodogramme à l'analyse spectrale d'une série temporelle de quaternions unitaires permettrait de simplifier cette étape du pipeline et d'estimer la durée des cycles de marche sur la seule série Q. La fiabilité avec laquelle l'algorithme STRIPAGE identifie les instants correspondant aux débuts et à la fin des phases d'appui et de balancement n'a pas pu être évaluée compte tenu du matériel à disposition pour son évaluation. Cette évaluation nécessite en effet de comparer les résultats de l'algorithme STRIPAGE avec ceux d'un dispositif suffisamment précis pour être considérés comme *Gold Standard* tel que les systèmes à reconnaissance optique ou le tapis de capteurs de pression GaitRite. De plus, les données ont été mesurées sur des individus ne présentant pas de troubles de la marche. Bien que le port de l'orthèse bloquante du genou est supposé simuler un déficit, la fiabilité de l'algorithme STRIPAGE telle qu'évaluée ici est difficilement transposable aux cas où le porteur du dispositif présente un fort déficit de la marche. C'est pourquoi, à la date de rédaction de ce manuscrit, une étude de fiabilité est en cours de développement en partenariat avec le CHU de Nantes, pour comparer les données de marche mesurées chez des personnes âgées avec la solution *eGait* avec celles du tapis *GaitRite*. Cette étude

permettra également de déterminer avec plus de précision quelles valeurs pour les hyperparamètre δ_{ID}, $\theta_{\mathbf{u}}$ et δ_{IQR} permettent de maximiser les performances de l'algorithme `STRIPAGE`. Enfin, la solution *eGait* nécessite un matériel peu coûteux et encombrant. Elle est donc en l'état facilement utilisable pour la mesure de la marche dans un contexte clinique.

2.1.3 Valorisation

L'algorithme `STRIPAGE` fait l'objet d'une partie de la demande de brevet n° 21 00309 « Méthode et dispositif de détermination d'un cycle de marche », déposée le 13 janvier 2021, et dont la validation est toujours en cours d'évaluation par l'Institut National de la Propriété Industrielle à date de rédaction de ce manuscrit.

L'algorithme `STRIPAGE` fait partie de la librairie interne à l'entreprise UmanIT `Stride Pattern Analysis` (`SPA`), développée en langage R et $C++$. Les méthodes plus générales pour l'analyse statistique des séries chronologiques de quaternions unitaires sont incluses dans la librairie `Statistics for QUAternion Temporal data` (`SQUAT`), développée en langage R et $C++$. Elle est disponible sur GitHub [1].

Les données mesurées sont de nature quantitative, et peuvent ainsi être utilisées pour déterminer plusieurs paramètres spatio-temporels afin de caractériser la démarche du porteur. Cet aspect est abordé dans la section 2.2.

2.2 Analyse de la démarche individuelle par paramètres spatio-temporels

2.2.1 Détermination des paramètres spatio-temporels (PST)

Cette section présente les méthodes de détermination de différents paramètres de la marche à partir de l'ensemble des M cycles Q_m détectés par l'algorithme `StriPagGe` dans les données d'un individu. Ces paramètres permettent de décrire plusieurs aspects de la démarche du porteur du système de capteurs.

1. https://github.com/astamm/squat

Afin d'illustrer les méthodes de détermination des différents paramètres spatiaux et temporels de la marche, l'angle entre l'orientation du dispositif observée au cours d'un cycle de marche Q_m et son orientation initiale est calculé :

$$\boldsymbol{\theta}_m = (\theta_{m,1}, \ldots, \theta_{m,N_m}) \,, \tag{2.18}$$

avec $\theta_{m,j} = d(\mathbf{q}_{m,1}, \mathbf{q}_{m,j}),\ \forall j \in \{1, \ldots, N_m\}$, la distance géodésique entre 2 quaternions unitaires (voir équation 1.23)).

On rappelle qu'un cycle de marche est constitué d'une phase d'appui suivie d'une phase de balancement du pied droit (voir définition du cycle de marche dans la section 1.1.1.1). On rappelle également l'hypothèse selon laquelle la transition de la phase d'appui vers la phase de balancement correspond à l'instant où l'orientation du dispositif est la plus éloignée de son orientation initiale. Ainsi, l'indice correspondant à cet instant est noté :

$$j_m^* = \arg\max_j \theta_{m,j},\ \forall j \in \{1, \ldots, N_m\} \tag{2.19}$$

Un cycle de marche peut donc être décrit par plusieurs paramètres temporels décrivant ses différentes phases :

— La durée du cycle :

$$\tau_m^{(wc)} = t_{m,N_m} \tag{2.20}$$

— La durée de la phase d'appui :

  — En seconde :

$$\tau_m^{(st)} = t_{m,j_m^*} \tag{2.21}$$

  — En pourcentage de durée du cycle :

$$p_m^{(st)} = \frac{\tau_m^{(st)}}{\tau_m^{(wc)}} \times 100 \tag{2.22}$$

— La durée de la phase de balancement :

  — En seconde :

$$\tau_m^{(sw)} = \tau_m^{(wc)} - t_{m,j_m^*} \tag{2.23}$$

  — En pourcentage de durée du cycle :

$$p_m^{(sw)} = \frac{\tau_m^{(sw)}}{\tau_m^{(wc)}} \times 100 \tag{2.24}$$

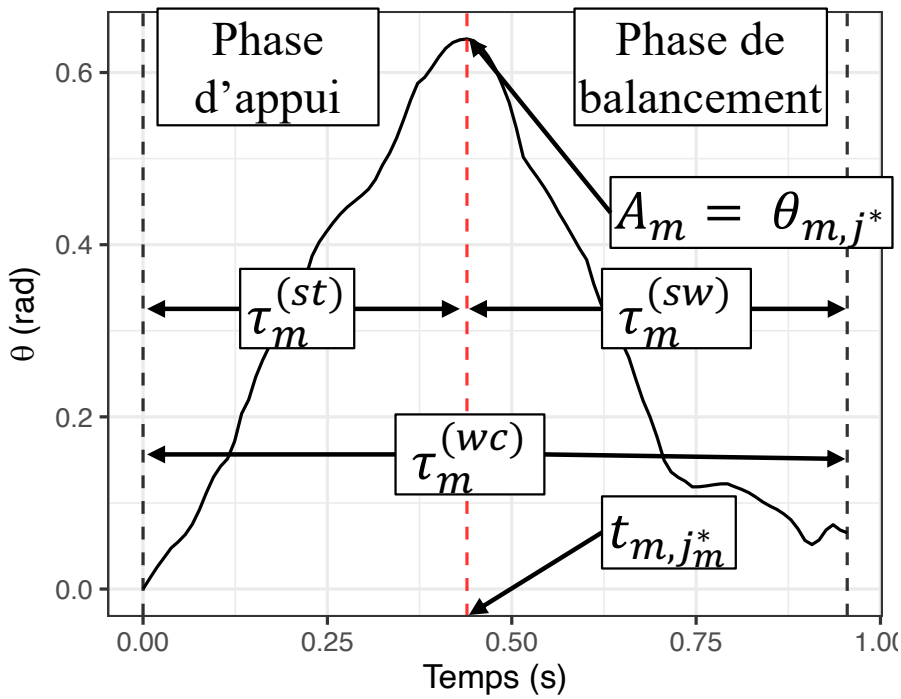


FIGURE 2.15 – Paramètres spatiaux et temporels du cycle de marche

Un cycle de marche peut également être décrit par son amplitude, *i.e.* l'angle entre l'orientation initiale, correspondant au début de la phase d'appui, et l'orientation observée au début de la phase de balancement :

$$A_m = \theta_{m,j_m^*} \tag{2.25}$$

La figure 2.15 schématise un cycle de marche et ses différents paramètres spatiaux et temporels. La courbe noire représente la série $\boldsymbol{\theta}_m$ (équation (2.18)). La durée du cycle est représentée par l'intervalle délimité par les pointillés noirs. Les pointillés rouges représentent l'instant correspondant à la transition entre la phase d'appui et la phase de balancement.

Le dernier paramètre pouvant être déterminé à partir d'un cycle de marche est l'angle moyen observé entre deux orientations successives. Pour le déterminer, on note la série des angles observés entre 2 orientations successives :

$$\boldsymbol{\Delta\theta}_m = (\Delta\theta_{m,1}, \ldots, \Delta\theta_{m,N_m-1}) , \tag{2.26}$$

avec $\Delta\theta_{m,j} = d(\mathbf{q}_{m,j}, \mathbf{q}_{m,j+1}),\ \forall j \in \{1, \ldots, N_m - 1\}$. L'angle moyen entre deux orientations successives est alors calculé comme la moyenne circulaire [84] de $\boldsymbol{\Delta\theta}_m$:

$$\overline{\Delta\theta}_m = \arctan 2 \left(\frac{1}{N_m - 1} \sum_{j=1}^{N_m - 1} \sin A_m, \frac{1}{N_m - 1} \sum_{j=1}^{N_m - 1} \cos A_m \right), \tag{2.27}$$

La marche d'un individu est ensuite représentée par la moyenne des paramètres calculés sur l'ensemble des cycles de marche Q_m détectés par l'algorithme `STRIPAGE` :

— Durée moyenne du cycle (en seconde) :

$$\bar{\tau}^{(wc)} = \frac{1}{M} \sum_{m=1}^{M} \tau_m^{(wc)} \tag{2.28}$$

— Durée moyenne de la phase d'appui (en pourcentage) :

$$\bar{p}^{(st)} = \frac{1}{M} \sum_{m=1}^{M} p_m^{(st)} \tag{2.29}$$

— Durée moyenne de la phase de balancement (en pourcentage) :

$$\bar{p}^{(sw)} = \frac{1}{M} \sum_{m=1}^{M} p_m^{(sw)} \tag{2.30}$$

— Amplitude moyenne du cycle (en radian) :

$$\bar{A} = \arctan 2 \left(\frac{1}{M} \sum_{m=1}^{M} \sin A_m, \frac{1}{M} \sum_{m=1}^{M} \cos A_m \right) \tag{2.31}$$

— Vitesse angulaire moyenne du cycle (en radian par seconde) :

$$\overline{\omega} = \frac{1}{\Delta t} \arctan 2 \left(\frac{1}{M} \sum_{m=1}^{M} \sin \overline{\Delta\theta}_m, \frac{1}{M} \sum_{m=1}^{M} \cos \overline{\Delta\theta}_m \right), \tag{2.32}$$

avec $\Delta t = F^{-1}$ la durée séparant deux mesures successives du système de capteurs. Ainsi, la vitesse angulaire moyenne $\overline{\omega}$ représente la vitesse de rotation moyenne de la hanche au cours d'un cycle de marche.

Ces paramètres quantifient plusieurs aspects de la démarche d'un individu. La mise en évidence d'un lien entre ces paramètres et le déficit de la marche est nécessaire pour

prouver l'intérêt de l'intégration de la solution *eGait* dans l'évaluation de la sévérité de la Sclérose En Plaques.

2.2.2 PST et troubles de la marche chez les patients SEP

Cette section présente la comparaison des paramètres mesurés par la solution *eGait* chez des patients atteints de Sclérose En Plaques avec leur déficit de la marche.

2.2.2.1 Matériel et méthode

L'analyse est menée à partir des données de la base `BDDsep`. Les données mesurées avec le système de capteurs MMR sont traitées par l'algorithme `STRIPAGE` pour identifier les cycles de marche. La marche de chaque patient est ensuite quantifiée par les paramètres tel que décrits section 2.2.1 : durée moyenne d'un cycle, durée moyenne de la phase d'appui, durée moyenne de la phase de balancement, amplitude moyenne d'un cycle et vitesse angulaire moyenne d'un cycle. La relation entre ces paramètres et la sévérité de la pathologie est évaluée d'après plusieurs critères.

Ces paramètres sont comparés entre groupes de patients formés selon la sévérité de leur pathologie estimée d'après leur score EDSS. Trois groupes de patients sont formés : Sévérité Faible (SF-EDSS : EDSS < 2), Sévérité Modérée (SM-EDSS : EDSS ≥ 2 et EDSS < 4) et Sévérité Élevée (SE-EDSS : EDSS ≥ 4) [125].

Les paramètres sont également comparés entre groupes de patients formés selon leur score aux systèmes fonctionnels Pyramidal, Cérébelleux et Sensitif (*cf* site web EDMUS [2]) :

— Pyramidal− (sous score Pyramidal ≤ 1 : Normal ou absence de handicap) vs Pyramidal+ (sous score Pyramidal > 1 : présence de handicap ou parésie)

— Cérébelleux− (sous score Cérébelleux ≤ 1 : Normal ou absence de handicap) vs Cérébelleux+ (sous score Cérébelleux > 1 : présence d'ataxie)

— Sensitif− (sous score Sensitif ≤ 1 : Normal ou faibles troubles sensitifs) vs Sensitif+ (sous score Sensitif > 1 : présence de troubles sensitif)

La différence des paramètres de la marche entre les groupes formés selon l'EDSS est évaluée par analyse de la variance [148]. Lorsqu'une différence significative entre les groupes EDSS est identifiée par l'analyse de variance, une analyse *post hoc* par test de

2. https://www.edmus.org/fr/proj/ms_fs.html

Différence Significative Honnête de Tukey-Kramer (Tukey DSH) est réalisée pour identifier les différences significatives entre sous groupes [95].

Les paramètres sont comparés entre groupes formés à partir des scores aux fonctions neurologiques par un modèle de régression linéaire multivarié. Chaque modèle est formé à partir d'un paramètre de marche comme variable réponse, les groupes Pyramidal+/−, Cérébelleux+/− et Sensitif+/− étant les trois variables explicatives catégorielles à deux modalités chacune.

Enfin, ces paramètres sont comparés avec la vitesse de marche des patients, sans stratification en groupe. Un modèle de régression linéaire est ajusté pour chaque paramètre avec la vitesse de marche mesurée lors du T25FW comme variable réponse. La significativité de la relation est évaluée par test de corrélation de Pearson. La relation entre un paramètre et la vitesse de marche est évaluée par le coefficient de détermination R^2 ajusté. La significativité de toutes les analyses statistiques est acceptée au seuil $\alpha = 0.05$.

2.2.2.2 Résultats

Description de l'échantillon. La distribution des scores EDSS et sous scores neurologiques entre les groupes de patients est présentée dans le tableau 2.2. Parmi les 30 patients de la base `BDDsep`, 11 présentent un déficit faible (SF-EDSS), 16 un déficit modéré (SM-EDSS) et 3 un déficit élevé (SE-EDSS). Aucun patient du groupe SF-EDSS ne présente de handicap ou trouble du système Pyramidal, Cérébelleux ou Sensitif. Si tous les patients du groupe SE-EDSS présentent des handicaps dûs à un déficit du système Pyramidal, ils n'ont pas tous une altération des systèmes Cérébelleux et Sensitif. Les patients du groupe SM-EDSS présentent également des atteintes variables de leurs systèmes fonctionnels. Cette hétérogénéité des profils des patients présentant une sévérité globale de la pathologie similaire justifie donc l'analyse des paramètres de la marche en fonction de leurs atteintes des fonctions neurologiques.

Relation entre paramètres de la marche et sévérité globale. Le tableau 2.3 présente la vitesse de marche des patients en fonction de leur sévérité globale, ainsi que les valeurs des paramètres de la marche et le nombre de cycle de marche détectés par l'algorithme `STRIPAGE`. Comme attendu, la vitesse de marche des patients tend à décroître avec la sévérité de la pathologie (1.788 m/s pour le groupe SF-EDSS, 1.356 m/s pour le groupe SM-EDSS et 0.952 pour le groupe SE-EDSS). Les paramètres par groupe de sévérité globale sont également représentés sur la figure 2.16. L'hypothèse d'homogénéité

TABLE 2.2 – Scores EDSS et sous scores fonctionnels par groupe

		SF-EDSS			**SM-EDSS**				**SE-EDSS**	
	EDSS	**0**	**1**	**1.5**	**2**	**2.5**	**3**	**4**	**5.5**	**6**
	Nb. patients	7	3	1	5	5	2	4	1	2
Pyramidal	Median	0	0	1	1	2	2	2	3	3
	Min-Max	0-0	0-1		0-2	1-3	2-2	1-2		3-4
Cérébelleux	Median	0	0	1	1	2	2	2	2	0
	Min-Max	0-0	0-0		0-2	1-3	2-2	1-3		0-2
Sensitif	Median	0	0	0	1	1	1	1	1	1
	Min-Max	0-0	0-1		0-2	0-2	1-2	0-2		1-2

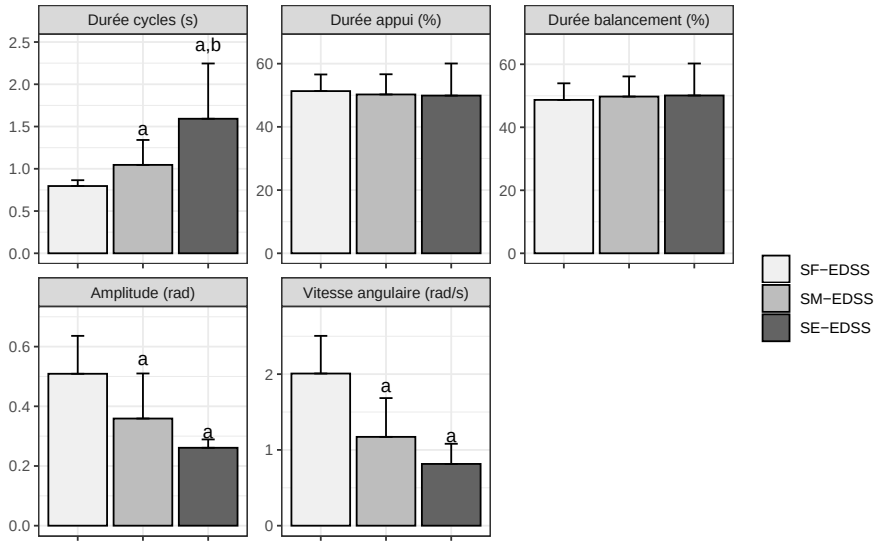
des variances entre groupe a été considérée comme respectée pour l'ensemble des paramètres spatio-temporels ($p - value > 0.05$ pour le test de Fligner-Killeen). L'analyse par variance montre une différence significative entre les groupes pour la durée, l'amplitude et la vitesse angulaire des cycles de marche. Les analyses *post hoc* permettent d'identifier une différence significative de durée des cycles pour les groupes SF-EDSS vs SM-EDSS, SF-EDSS vs SE-EDSS et SM-EDSS vs SE-EDSS. La durée moyenne des cycles tend donc à augmenter avec la sévérité globale de la pathologie. Aucune différence significative n'est identifiée pour la durée des phases d'appui et de balancement (en pourcentage de durée totale du cycle). L'amplitude et la vitesse angulaire du cycle sont signficativement différentes entre les groupes SF-EDSS vs SM-EDSS et SF-EDSS vs SE-EDSS. La rotation de la hanche au cours de la marche des patients présentant une sévérité globale moyenne et élevée a donc tendance à être moins ample et rapide que ceux présentant une faible sévérité de la pathologie. Bien que l'amplitude et la vitesse angulaire soient plus faibles pour les patients avec une forte sévérité globale que ceux avec une sévérité modérée, les données de cette étude n'ont pas permis de les considérer comme statistiquement significatives.

Relation entre paramètres de la marche et sous scores neurologiques. Le tableau 2.4 présente les paramètres de marche en fonction des groupes formés selon les sous scores des systèmes fonctionnels Pyramidal, Cérébelleux et Sensitif. Ces derniers sont également présentés dans la figure 2.17. Parmi les 30 patients de la base `BDDsep`, 12 présentent des troubles du système pyramidal se traduisant par un handicap minime, une paraparésie ou hémiparésie légère à marquée, une tétraparésie modérée, ou une monoplé-

TABLE 2.3 – Paramètres spatio-temporels de la marche - Stratification sur score EDSS

		SF-EDSS (N=11)	**SM-EDSS (N=16)**	**SE-EDSS (N=3)**
Vitesse de marche (m/s)	Moy. (E.t.)	1.788 (0.193)	1.356 (0.399)	0.952 (0.359)
	Min-Max	1.506-2.16	0.59-1.812	0.364-0.78
Nb. Cycles	Médian	7	7	10
	Min-Max	5-9	5-13	6-15
Durée du cycle (s)	Moy. (E.t.)	0.796 (0.069)	1.046 (0.295)[a]	1.592 (0.654)[a,b]
	Min-Max	0.699-0.935	0.758-2.062	1.141-2.342
Appui (%)	Moy. (E.t.)	51.313 (5.266)	50.25 (6.391)	49.902 (10.15)
	Min-Max	46.31-61.454	39.321-63.327	41.952-61.335
Balancement (%)	Moy. (E.t.)	48.687 (5.266)	49.75 (6.391)	50.098 (10.15)
	Min-Max	38.546-53.69	36.673-60.679	38.665-58.048
Amplitude (rad)	Moy. (E.t.)	0.509 (0.127)	0.359 (0.151)[a]	0.261 (0.028)[a]
	Min-Max	0.289-0.814	0.061-0.608	0.235-0.3
Vitesse angulaire (rad/s)	Moy. (E.t.)	2.008 (0.497)	1.172 (0.512)[a]	0.815 (0.266)[a]
	Min-Max	1.337-3.167	0.171-2.131	0.448-1.066

Moy. : Moyenne, E.t. : Écart type
[a] Différence significative du groupe SF-EDSS (Tukey HDS : $p.adj < 0.05$)
[b] Différence significative du groupe SM-EDSS (Tukey HDS : $p.adj < 0.05$)



a : Différence significative du groupe SF-EDSS (Tukey HDS : $p.adj < 0.05$)
b : Différence significative du groupe SM-EDSS (Tukey HDS : $p.adj < 0.05$)

FIGURE 2.16 – Paramètres de la marche et sévérité de la pathologie

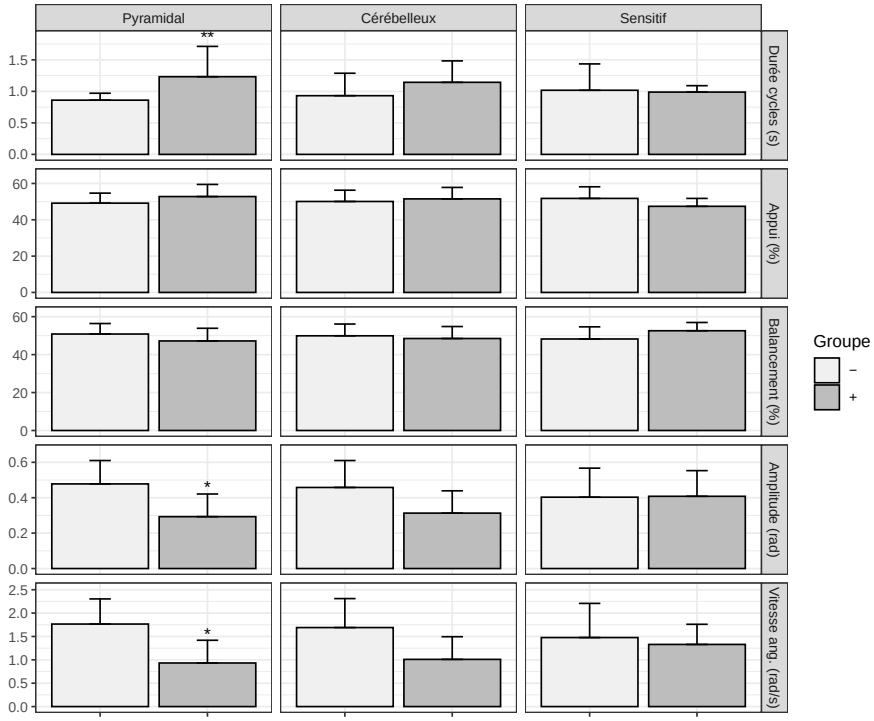
TABLE 2.4 – Paramètres spatio-temporels de la marche - Stratification sur scores neurologiques

		Pyramidal		**Cérébelleux**		**Sensitif**	
		- (N=18)	**+ (N=12)**	**- (N=19)**	**+ (N=11)**	**- (N=22)**	**+ (N=8)**
Durée du cycle (s)	Moy. (E.t.)	0.86 (0.11)	1.232 (0.481)**	0.931 (0.356)	1.143 (0.34)	1.017 (0.418)	0.988 (0.102)
	Min-Max	0.699-1.064	0.758-2.342	0.699-2.342	0.758-2.062	0.699-2.342	0.82-1.141
Appui (%)	Moy. (E.t.)	49.148 (5.53)	52.791 (6.672)	50.082 (6.205)	51.508 (6.308)	51.765 (6.406)	47.416 (4.35)
	Min-Max	39.321-61.454	41.952-63.327	39.321-61.454	41.952-63.327	41.61-63.327	39.321-53.102
Balancement (%)	Moy. (E.t.)	50.852 (5.53)	47.209 (6.672)	49.918 (6.205)	48.492 (6.308)	48.235 (6.406)	52.584 (4.35)
	Min-Max	38.546-60.679	36.673-58.048	38.546-60.679	36.673-58.048	36.673-58.39	46.898-60.679
Amplitude (rad)	Moy. (E.t.)	0.478 (0.132)	0.293 (0.128)*	0.458 (0.152)	0.313 (0.126)	0.403 (0.164)	0.408 (0.145)
	Min-Max	0.221-0.814	0.061-0.532	0.17-0.814	0.061-0.532	0.061-0.814	0.221-0.608
Vitesse angulaire (rad/s)	Moy. (E.t.)	1.766 (0.538)	0.933 (0.486)*	1.691 (0.62)	1.01 (0.485)	1.477 (0.731)	1.33 (0.431)
	Min-Max	1.099-3.167	0.171-2.131	0.448-3.167	0.171-2.131	0.171-3.167	0.74-2.083

Moy. : Moyenne, E.t. : Écart type
* : $p < 0.05$, ** : $p < 0.01$

gie (Pyramidal+). 16 patients présentent une ataxie légère à modérée (Cérébelleux+). 8 présentent des troubles sensitifs (Sensitif+). Après ajustement des modèles sur les scores des trois systèmes fonctionnels, seule l'atteinte du système pyramidal se traduit par une différence significative de la durée du cycle, de l'amplitude et de la vitesse angulaire. Les patients présentant un déficit de la fonction pyramidale tendent à avoir des cycles de marche plus longs, et une rotation de la hanche moins ample et moins rapide durant la marche. Bien que ces différences soient également observées entre les groupes Cérébelleux− et Cérébelleux+, les données de l'étude ne permettent pas de conclure qu'elles soient statistiquement significatives. Il est à noter que 10 patients présentent conjointement des troubles Pyramidaux et Cérébelleux, dont 4 patients avec des troubles Sensitifs. L'effet de l'atteinte d'une seule fonction neurologique sur les paramètres de la marche est donc difficile à évaluer sur la base de données de l'étude.

Relation entre paramètres et vitesse de marche La figure 2.18 présente la relation de la vitesse de la marche des patients au T25FW avec les paramètres de la marche, ainsi que la droite de régression linéaire associée. La relation entre vitesse de marche et durée des cycles est représentée et évaluée après transformation par logarithme de ces deux variables. Les corrélations correspondantes et leur évaluation sont présentées dans le tableau 2.5. La durée du cycle, l'amplitude et la vitesse angulaire du cycle sont corrélées significativement avec la vitesse de marche. Ainsi, plus la vitesse du patient diminue, plus ses cycles de marche tendent à être longs avec une rotation de hanche moins ample et moins rapide. La corrélation est particulièrement forte entre la vitesse de marche et (i)



* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

FIGURE 2.17 – Paramètres de la marche et sous scores neurologiques

TABLE 2.5 – Relation entre paramètres et vitesse de marche

Paramètre	**Coef. Pearson**	**p – value**	**R^2 aj.**
Durée du cycle (log)	-0.936	<0.0001	0.871
Appui (%)	-0.089	0.639	-0.027
Balancement (%)	0.089	0.639	-0.027
Amplitude (rad)	0.684	<0.0001	0.448
Vitesse angulaire (rad/s)	0.841	<0.0001	0.697

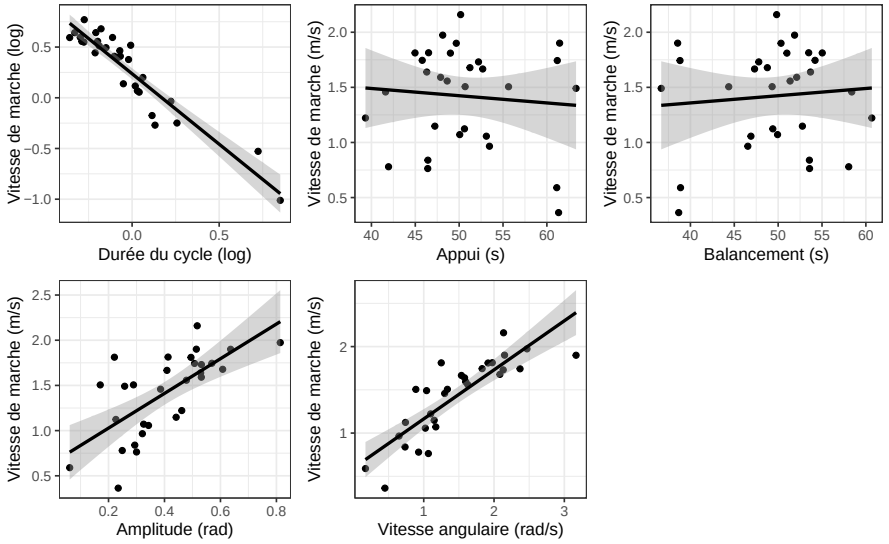


FIGURE 2.18 – Paramètres de la marche et vitesse de marche

la durée du cycle après transformation logarithmique (R^2 **aj.** = 0.871), et (ii) la vitesse angulaire (R^2 **aj.** = 0.697). Ni la durée de la phase d'appui ni celle de balancement ne sont corrélées avec la vitesse de marche.

2.2.2.3 Discussion

L'analyse statistique exploratoire des données de la base `BDDsep` montre une relation entre les paramètres de marche déterminés par la solution *eGait* et plusieurs marqueurs de la sévérité globale et des déficits de la marche des patients atteints de SEP. La différence entre plusieurs de ces paramètres (vitesse de rotation et amplitude de la hanche, durée des cycles) est observée y compris entre les patients ayant une sévérité faible et modérée. La sensibilité de ces paramètres aux effets d'une sévérité modérée est importante pour détecter les déficits apparaissant au début de la pathologie. La différence non significative pour l'amplitude et la vitesse de rotation de la hanche entre les sévérités modérée et élevée

est potentiellement due au faible nombre de patients du dernier groupe. Ces résultats peuvent également être commentés au regard d'autres études décrites dans la littérature.

Tout d'abord, le fait que la durée des cycles de marche tend à diminuer avec l'état de santé général du patient et la vitesse de marche a été observé dans de nombreuses études [125, 152, 159, 103, 7]. Cela constitue une première preuve de la validité des méthodes développées pour évaluer la durée des cycles de marche pour solution *eGait*. Cependant, l'absence de relation entre le pourcentage de durée de la phase d'appui/de balancement avec la sévérité de la pathologie est en contradiction avec ces mêmes études. En effet, Severini *et al.* (2017) [152] ont établi une relation entre ce paramètre et la sévérité de la pathologie par analyse de la marche avec un système vidéo, ce qui a aussi été observé par Pau *et al.* (2016) [125] avec un système d'analyse de la marche par accéléromètre. Les valeurs observées sur la `BDDsep` pour la phase d'appui et de balancement sont également en contradiction avec celles de ces deux études. En effet, la phase d'appui représente généralement 60% de la durée du cycle, et augmente avec la sévérité (jusqu'à plus de 70% chez les patients les plus sévèrement atteints [152]). Cela peut potentiellement s'expliquer par le fait que l'algorithme `STRIPAGE` ne détecte pas précisément les transitions phases d'appui/phase de balancement, soit parce qu'il est sensible au bruit, soit parce que le changement de sens de la rotation de la hanche n'est pas concomitant avec la pose du talon au sol ou le décollement des orteils. Si les relations entre l'amplitude de la rotation de la hanche, la sévérité de la pathologie et la vitesse de marche sont en accord avec les observations de Severini *et al.* (2017)[152], ces mêmes relations avec la vitesse de rotation de la hanche n'ont, à ma connaissance, pas été décrites dans la littérature.

Pour conclure, ces résultats ouvrent de nombreuses perspectives de recherche. En premier lieu, la taille de l'échantillon étant relativement faible, ces résultats méritent d'être confirmés sur une cohorte de patients plus grande. Un plus grand nombre de patients pourrait également permettre d'évaluer plus précisément l'effet des troubles des systèmes neurologiques pyramidaux, cérébelleux et sensitif. Pour ce faire, une nouvelle étude a été organisée avec le CHU de Nantes incluant 40 patients atteints de SEP, dont une dizaine présentant des troubles de la marche sévère. L'inclusion des patients a commencé en Août 2021 et est toujours en cours à la date de rédaction de ce manuscrit. Cette étude a été financée grâce au dispositif Projet Exploratoire, Premier Soutien (PEPS) proposé par l'Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société (AMIES) [3]. Les très bonnes corrélations entre la durée des cycles, la vitesse de rotation de la hanche

3. https://www.agence-maths-entreprises.fr/public/pages/activities/contrats.html

et la vitesse de la marche présentent un intérêt tout particulier. En effet, compte tenu du faible encombrement du dispositif de la marche, ces paramètres pourraient permettre d'évaluer la vitesse de la marche des patients dans leur quotidien. Pour ce faire, des travaux supplémentaires sont nécessaires, en premier lieu pour évaluer et améliorer si besoin la précision avec laquelle la vitesse de marche peut être estimée avec la solution *eGait*. Cette estimation pourrait recourir à des méthodes d'apprentissage supervisé personnalisé, comme celles utilisées dans les travaux de Supratak *et al.* (2018) [164]. D'autre part, l'algorithme `STRIPAGE` a été développé pour traiter des données représentant en majorité une activité de marche. Il est donc nécessaire de développer une méthode permettant d'identifier les périodes durant lesquelles le porteur du dispositif marche dans un jeu de données mesuré sur une journée complète. Cet aspect a été abordé par Raphaël Brard lors de son stage de Master 2 que j'ai co-encadré entre avril et septembre 2021, puis poursuivi lors d'un contrat à durée déterminé au sein d'UmanIT de septembre 2021 à février 2022. Les résultats de ces travaux sont présentés dans l'article "*A Novel Walking Activity Recognition Model for Rotation Time Series Collected by a Wearable Sensor in a Free-Living Environment*" [21].

2.2.3 Valorisation

La présentation des résultats de l'analyse de la marche par paramètres spatio-temporels sur la `BDDsep` a fait l'objet d'un rapport de recherche adressé le 3 février 2021 au Département promotion de la direction de la Recherche et de l'Innovation du CHU de Nantes. La comparaison de la durée et de la vitesse angulaire moyenne des cycles de marche avec la vitesse de marche des patients atteints de Sclérose En Plaques est présentée dans l'article "Gait impairment monitoring in multiple sclerosis using a wearable motion sensor", publié dans la revue Medical Case Reports and Reviews [4], édité par Open Access Text en février 2022.

4. https://www.oatext.com/pdf/MCRR-5-175.pdf

Chapitre 3

Méthodes de classification pour données de marche

L'algorithme `StriPaGe` présenté dans la section 2.1 permet d'identifier les cycles de marche dans les données mesurées par le dispositif *MetaMotionR*, représentant la rotation de la hanche au cours du temps sous forme de quaternions unitaires. Une première approche décrite dans la section 2.2 consiste à représenter quantitativement plusieurs aspects de la marche sous forme de paramètres spatio-temporels. L'analyse des données de patients atteints de Sclérose En Plaques présentée dans la section 2.2.2 apporte les premières preuves de la relation entre ces paramètres spatio-temporels et les déficits de la marche. Cette section aborde une approche dans laquelle la forme des données de marche est comparée entre individus par méthode de classification. Les données de marche se présentent sous la forme de mesures successives de l'orientation en 3 dimensions de la hanche au cours du temps. Elles peuvent donc être considérées comme des *séries chronologiques* ou des données fonctionnelles de *quaternions unitaires*.

La première partie de ce chapitre décrit l'adaptation aux quaternions unitaires des méthodes de classification non supervisée par regroupement hiérarchique (CAH) et par partitionnement (*K-means* et *K-medoids*) de *séries chronologiques* et de *données fonctionnelles*. Ces méthodes sont employées pour l'analyse de la marche au travers de deux applications.

(i) Tout d'abord, l'utilisation du *K-means* de données fonctionnelles adapté aux quaternions unitaires permet le calcul du prototype des cycles de marche identifiés chez un même individu par l'algorithme `StriPaGe`. Ce prototype permet de représenter la démarche d'un individu sous la forme d'une unique séquence de quaternions unitaires. Il constitue le biomarqueur appelé *Signature de Marche* (SdM).

(ii) Les méthodes de classification CAH, *K-means* et *K-medoids* sont ensuite appliquées pour former des groupes d'individus présentant des troubles de la marche simulés à partir de leur SdM.

La seconde partie de ce chapitre présente la généralisation de méthodes de *classification semi-supervisée* aux séries chronologiques de quaternions unitaires. L'intérêt des méthodes de classification semi-supervisée réside dans la possibilité d'intégrer plusieurs sources d'informations concernant les observations dans la construction des groupes. Ainsi, dans le contexte de l'analyse de la marche, elles permettent de classifier les patients en fonction de leurs données de marche en tenant compte d'informations connues *a priori* concernant le niveau de sévérité global de leur pathologie. Les deux méthodes adaptées sont (i) `mergeTrees` [75], une méthode *d'ensemble de classification* basée sur la construction d'un arbre consensus à partir d'un ensemble de dendrogrammes obtenus par classification ascendante hiérarchique, et (ii) `hclustcompro` [16], une méthode de classification ascendante hiérarchique *par compromis*, basée sur le calcul d'une dissimilarité globale entre observations à partir des dissimilarités observées sur les espaces d'une source d'information principale et d'une source d'information supplémentaire. Ces deux méthodes sont comparées quant à leurs capacités à former des groupes de patients atteints de SEP présentant une démarche et un niveau de sévérité global de la pathologie similaire.

3.1 Classification non supervisée et données de marche

Cette section présente l'adaptation des méthodes de classification ascendante hiérarchique (CAH), *K-means* et *K-medoids* aux séries chronologiques de quaternions unitaires et aux données fonctionnelles de quaternions unitaires. Cette adaptation repose sur la modification des algorithmes classiques par l'utilisation de fonctions appropriées au type de données pour (i) mesurer la dissimilarité entre observations et (ii) déterminer le prototype des groupes formés pour les méthodes *K-means*, *K-medoids*. Les méthodes proposées permettent également de s'affranchir du potentiel mauvais alignement temporel entre les séquences.

3.1.1 Classification non supervisée de données fonctionnelles quaternioniques

Dans cette section, on considère qu'une séquence de quaternions unitaires $\mathbf{q}_1, \ldots, \mathbf{q}_N$ observés aux temps $t_1, \ldots, t_N$ constitue un ensemble d'observations discrètes d'une fonction quaternionique appartenant à l'espace des fonctions à valeurs dans l'espace des quaternions unitaires $\mathcal{Q} = \{Q : \Omega \subseteq \mathbb{R} \to \mathbb{H}_u\}$. La première étape pour appliquer une mé-

thode de classification sur un sous-ensemble de fonctions quaternioniques consiste à doter l'espace $\mathcal{Q}$ d'une métrique représentant la dissimilarité entre deux observations, i.e. deux séquences de quaternions unitaires. De façon similaire aux séries chronologiques, un mauvais alignement peut être observé entre les fonctions. Ce phénomène est appelé *variabilité de phase* dans le domaine de l'analyse des données fonctionnelles [109].

La méthode d'alignement de données fonctionnelles par fonctions de *warpings* affines strictement croissantes est décrite dans la section 1.3.2.2. Elle permet le calcul (i) de la *dissimilarité* entre données fonctionnelles alignées et (ii) du *centre*, ou *prototype*, d'un ensemble de données fonctionnelles alignées. Cette section détaille comment tirer partie de cette solution pour adapter les méthodes de Classification Ascendante Hiérarchique, de *K-means alignment* et de *K-medoids alignment* aux données fonctionnelles de quaternions unitaires.

3.1.1.1 Transformation logarithmique des fonctions quaternioniques

La solution décrite section 1.3.2.2 pour l'alignement des fonctions et le calcul du prototype est définie pour des fonctions qui prennent valeur dans $\mathbb{R}^p$. Une adaptation est donc nécessaire pour l'utiliser afin d'aligner des fonctions de quaternions unitaires. Les fonctions Q sont transformées par fonction logarithmique (voir équation (1.17)) :

$$\ln(\mathbf{q}) = \ln\left(\cos\frac{\theta}{2} + \mathbf{u}\sin\frac{\theta}{2}\right) = \left(0, \mathbf{u}\frac{\theta}{2}\right) \tag{3.1}$$

On définit la transformation logarithmique $\ln : \mathcal{Q} \to \mathcal{Q}^{(ln)}$, avec $\mathcal{Q}^{(ln)}$ l'ensemble des fonctions de *log-quaternions* :

$$\mathcal{Q}^{(ln)} = \left\{Q^{(ln)} : \Omega \subseteq \mathbb{R} \to \mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u) \subseteq \mathbb{R}^3\right\}, \tag{3.2}$$

avec $\mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u)$ l'espace tangent à l'espace $\mathbb{H}_u$ au point $\mathbf{q}_0 = (1, 0, 0, 0)$ (*c.f.* section 1.2.2.1, paragraphe Forme polaire et logarithme des quaternions unitaires.). Les fonctions de *log-quaternions* appartiennent au sous-ensemble $\mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u)$ inclus dans l'espace $\mathbb{R}^3$. La transformation logarithmique permet donc d'appliquer les méthodes de classification de données fonctionnelles décrites dans la suite de ce paragraphe à un ensemble de n fonctions de log-quaternions $\{Q_i^{(ln)} \in \mathcal{Q}^{(ln)}\}_{i=1,2,\dots,n}$. La transformation inverse est l'exponentielle $\exp : \mathcal{Q}^{(ln)} \to \mathcal{Q}$, qui permet d'obtenir une fonction de quaternions unitaires à partir d'une fonction de log-quaternions.

3.1.1.2 Classification Ascendante Hiérarchique.

Soit un ensemble de n fonctions de log-quaternions $\{Q_i^{(ln)}\}_{i=1,2,\dots,n}$, la matrice de dissimilarité D de dimension $n \times n$ est calculée telle que :

$$\begin{aligned} d(i,j) &= d_W(Q_i^{(ln)}, Q_j^{(ln)}) \\ &= \min_{h_1,h_2 \in W} d^2_{L^2_m}(Q_i^{(ln)} \circ h_1, Q_j^{(ln)} \circ h_2), \end{aligned} \tag{3.3}$$

$d^2_{L^2_m}$ étant la métrique définie dans l'équation (1.44), adaptée pour des fonctions appartenant à $\mathbb{R}^3$:

$$d^2_{L^2_m}\left(Q_i^{(ln)}, Q_j^{(ln)}\right) = \frac{d^2_{L_2}(Q_i^{(ln)}, Q_j^{(ln)})}{||Q_i^{(ln)}||^2_{L^2} + ||Q_j^{(ln)}||^2_{L^2}}, \tag{3.4}$$

avec

$$||Q_i^{(ln)}||^2_{L^2} = d^2_{L^2}(Q_i^{(ln)}, 0_{L^2}) = \int_{\mathbb{R}} ||Q_i^{(ln)}(t)||^2_{\mathbb{R}^3} dt = \sum_{j=1}^{3} \int_{\mathbb{R}} \left[Q_{i,j}^{(ln)}(t)\right]^2 dt. \tag{3.5}$$

Une Classification Ascendante Hiérarchique peut alors être appliquée sur la matrice D. Les observations $\{Q_1^{(ln)}, \dots, Q_n^{(ln)}\}$ n'appartenant pas à un espace euclidien, seuls les critères de liaison simple, complète et moyenne peuvent être utilisés dans les étapes de fusion et de mise à jour de la matrice D (*c.f.* présentation de la formule de récurrence de Lance et Williams 1.31 au paragraphe 1.3.1.1).

3.1.1.3 *K-means alignment* et *K-medoids alignment*

Les algorithmes de *K-means* et *K-medoids* (*c.f.* section 1.3.1.1) permettent de répartir un ensemble n observations en K groupes, ce nombre de groupes K étant déterminé *a priori* par l'utilisateur. Ils sont construits autour de deux étapes répétées itérativement :

1. L'affectation des observations au groupe dont elles sont les plus proches.
2. Le calcul du prototype des groupes en fonction des observations qu'ils rassemblent.

Les méthodes de *K-means alignment* et *K-medoids alignment* sont des généralisations de ces algorithmes aux données fonctionnelles, permettant la classification d'un ensemble de fonctions alignées par fonctions de *warping*. Elles sont décrites par Sangalli *et al.* (2010) [145, 144]. Pour l'adaptation de ces algorithmes à la classification d'un ensemble de n données fonctionnelles de log-quaternions $\{Q_i^{(ln)}\}_{i=1,2,\dots,n}$, on définit :

— $\mathbf{\Phi} = \{\Phi_k \in \mathcal{Q}^{(ln)}\}_{k=1,2,\dots,K}$: l'ensemble des prototypes représentant les K groupes.

— $d_W\left(Q_i^{(ln)}, \Phi_k\right) = \min\limits_{h_i \in W} d^2_{L^2_m}\left(Q_i^{(ln)} \circ h_i, \Phi_k\right)$: la mesure de dissimilarité entre l'observation i et le groupe k, $d^2_{L^2_m}$ étant la métrique définie dans l'équation (3.4).

— $\lambda\left(\mathbf{\Phi}, Q_i^{(ln)}\right) = \underset{r\in\{1,\dots,K\}}{\arg\min}\, d_W\left(Q_i^{(ln)}, \Phi_r\right)$: la fonction d'attribution de l'observation i. $\lambda\left(\mathbf{\Phi}, Q_i^{(ln)}\right) = k$ signifie que le meilleur alignement de la fonction $Q_i^{(ln)}$ est observé avec le prototype Φ_k, l'observation i devant donc être affectée au groupe k.

— $V_k = \{i \in [\![1, n]\!] : \lambda\left(Q_i^{(ln)}, \mathbf{\Phi}\right) = k\}$: définit l'ensemble des observations appartenant au groupe k.

— $n_k = |V_k|$: Le nombre d'observations affectées au groupe k.

— $\left\{h_i = \arg\min_{h\in W} d_W\left(Q_i^{(ln)}, \Phi_{\lambda\left(\mathbf{\Phi}, Q_i^{(ln)}\right)}\right)\right\}_{i=1,2,\dots,n}$: l'ensemble des fonctions de *warping* alignant les observations sur le prototype du groupe auquel elles sont affectées.

Après l'itération $\ell - 1$ (pour $\ell \in \mathbb{N}^+$), on définit :

— $\mathbf{\Phi}^{[\ell-1]} = \left\{\Phi_k^{[\ell-1]}\right\}_{k=1,2,\dots,K}$ l'ensemble des prototypes des groupes,

— $\left\{Q_i^{(ln)[\ell-1]} = Q_i^{(ln)} \circ h_i^{[\ell-1]}\right\}_{i=1,2,\dots,n}$ les n fonctions de log-quaternions alignées au prototype du groupe auquel elles ont été affectées

L'itération ℓ consiste en la succession des étapes suivantes :

1. *Calcul des prototypes :*
   — Algorithme *K-means alignment* : Pour $k \in \{1, \dots, K\}$, le prototype $\Phi_k^{[\ell]}$ est calculé par la moyenne de Fréchet de l'ensemble des fonctions $Q_i^{(ln)[\ell-1]}$ pour lesquelles $\lambda\left(\mathbf{\Phi}^{[\ell-1]}, Q_i^{(ln)[\ell-1]}\right) = k$:

$$\Phi_k^{[\ell]} = \underset{\Phi\in\mathcal{Q}^{(ln)}}{\arg\min} \sum_{i\in V_k} d_W\left(Q_i^{(ln)[\ell-1]} \circ h_i^{[\ell-1]}, \Phi\right).$$

   — Algorithme *K-Medoids alignment* : Pour $k \in \{1, \dots, K\}$, le prototype $\Phi_k^{[\ell]}$ est déterminé comme le médoïde du groupe :

$$\Phi_k^{[\ell]} = \underset{Q_i^{(ln)[\ell-1]}, i\in V_k}{\arg\min} \sum_{j\in V_k} d_W\left(Q_i^{(ln)[\ell-1]}, Q_j^{(ln)[\ell-1]}\right)$$

2. *Affectation et alignement des fonctions* - Pour $i \in \{1, \dots n\}$, la fonction $Q_i^{(ln)[\ell-1]}$ est alignée au prototype $\Phi_{\lambda\left(\mathbf{\Phi}^{[\ell]}, Q_i^{(ln)[\ell-1]}\right)}$ et la fonction alignée $\hat{Q}_i^{(ln)[\ell]} = Q_i^{(ln)[\ell-1]} \circ h_i^{[\ell]}$, est affectée au groupe $\lambda\left(\mathbf{\Phi}^{[\ell]}, \hat{Q}_i^{(ln)[\ell]}\right)$.

3. *Normalisation* - Pour $k \in \{1, \ldots, K\}$, les $n_k^{[\ell]}$ fonctions alignées au groupe k sont transformées par la fonction $\left(\bar{h}_k^{[\ell]}\right)^{-1}$, telle que

$$\bar{h}_k^{[\ell]} = \frac{1}{n_k^{[\ell]}} \sum_{i \in V_k} h_i^{[\ell]},$$

pour obtenir les fonctions $Q_i^{(ln)[\ell-1]} = \hat{Q}_i^{(ln)[\ell]} \circ \left(\bar{h}_k^{[\ell]}\right)^{-1}$.
L'étape de normalisation permet d'empêcher les groupes de fonctions de s'éloigner les uns des autres sur l'axe des temps, ou d'empêcher l'ensemble des fonctions de "dériver" temporellement par rapport à l'état initial au cours des itérations.

Les algorithmes de *K-means alignment* et *K-medoids alignment* sont initialisés en choisissant aléatoirement l'ensemble des prototypes initiaux $\mathbf{\Phi}^{[0]}$ parmi l'ensemble des fonctions initiales $\{Q_1^{(ln)[0]}, \ldots, Q_n^{(ln)[0]}\} = \{Q_1^{(ln)}, \ldots, Q_n^{(ln)}\}$. Ils sont stoppés quand une itération supplémentaire des étapes d'affectation et d'alignement entraîne une diminution de la dissimilarité globale $\sum_{k=1}^{K} \sum_{i \in V_k} d_W \left(\Phi_k^{[\ell]}, Q_i^{(ln)[\ell-1]}\right)$ inférieure à un seuil définit par l'utilisateur.

Soit Ω la dernière itération de l'algorithme *K-means alignment* ou *K-medoids alignment*, l'application exponentielle $\exp : Q^{(ln)} \subseteq \mathcal{Q}^{(ln)} \to Q \subseteq \mathcal{Q}$ permet de transformer les fonctions de log-quaternions $\{Q_1^{(ln)[\Omega]}, \ldots, Q_n^{(ln)[\Omega]}\}$ et les prototypes $\left\{\Phi_k^{[\Omega]}\right\}_{k=1,2,\ldots,K}$ dans l'espace des fonctions de quaternions unitaires.

3.1.2 Classification non supervisée de séries temporelles de quaternions unitaires

Dans cette section, on considère une séquence de quaternions unitaires $\mathbf{q}_1, \ldots, \mathbf{q}_N$ observés aux temps $t_1, \ldots, t_N$ comme une série chronologique de quaternions unitaires $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$ associés aux temps de mesure $T = (t_1, \ldots, t_N)$. Les méthodes de calcul de la dissimilarité entre deux séries temporelles (TS) par *Quaternion Dynamic Time Warping* (DTW) et de calcul du prototype d'un ensemble de TS par *Dtw Barycenter Averaging* (DBA) sont présentées, ainsi que leur adaptation aux séries chronologiques de quaternions unitaires (QTS). Elles sont utilisées pour généraliser les méthodes de Classification Ascendante Hiérarchique, *K-means* et *K-medoids* aux QTS.

3.1.2.1 Mesure de dissimilarité et alignement temporel

Le *Dynamic Time Warping* (DTW) est l'une des mesures de dissimilarité de séries temporelles les plus populaires [3]. La dissimilarité est calculée à partir des éléments alignés entre les deux séries. L'algorithme DTW consiste à rechercher l'alignement des éléments de deux séries temporelles tel que la dissimilarité entre ces deux séries alignées soit la plus faible. La généralisation de cet algorithme aux QTS, appelée *Quaternion Dynamic Time Warping* (QDTW), est décrite par Jablonski (2012) [82]. Le QTDW est présenté dans la suite de cette section. Certaines notations et notions utilisées dans la présentation de l'algorithme DTW classique par Petitjean *et al.* (2011) [128] et Müller (2007) [116] sont également reprises.

Soient deux QTS $Q_1 = (\mathbf{q}_{1,1}, \ldots, \mathbf{q}_{1,N_1})$ et $Q_2 = (\mathbf{q}_{2,1}, \ldots, \mathbf{q}_{2,N_2})$ respectivement de dimension N_1 et N_2, et $d(.,.)$ la distance géodésique entre 2 quaternions unitaires (voir équation (1.23)). La problématique du QDTW peut être présentée comme la recherche de la fonction de *warping* respectant les propriétés correspondant à la solution de l'équation suivante [82] :

$$\mathrm{QDTW}(Q_1, Q_2) = \min_{W} \sum_{i=1}^{K} d(\mathbf{q}_{1,w_{i,1}}, \mathbf{q}_{2,w_{i,2}}), \tag{3.6}$$

où W est appelée fonction de *warping*. Elle aligne les éléments de Q_1 et Q_2, en respectant les propriétés suivantes [116] :

(1) $W = (w_1, \ldots, w_K)$, $\max(N_1, N_2) \leq K \leq (N_1 + N_2 - 1)$, avec K correspondant au nombre d'alignements entre Q_1 et Q_2, et :

$$w_k = (w_{k,1}, w_{k,2}), \text{ avec } k \in \{1, \ldots, K\}, \; w_{k,1} \in \{1, \ldots, N_1\}, \text{et } w_{k,2} \in \{1, \ldots, N_2\}$$

(2) Les premiers et derniers éléments de Q_1 et Q_2 sont alignés : $w_{1,1} = w_{1,2} = 1$, $w_{K,1} = N_1$, $w_{K,2} = N_2$ (cette propriété est appelée *contrainte de liaison* dans la littérature)

(3) Chaque élément de Q_1 est aligné avec au moins un élément de Q_2, et inversement (*contrainte de continuité*)

(4) L'ordre des éléments est conservé : $w_{k,1} \leq w_{k+1,1}$ et $w_{k,2} \leq w_{k+1,2}$, $\forall k \in \{1, \ldots, K-1\}$ (*contrainte de monotonicité*)

(5) Les possibilités d'alignement à l'étape k dépendent de l'alignement de l'étape $k-1$: $w_{k,1} - w_{k-1,1} \leq 1$ et $w_{k,2} - w_{k-1,2} \leq 1$ (*contrainte d'étape*)

La fonction de *warping* correspondant à la solution de l'équation (3.6) est déterminée récursivement grâce à l'équation suivante :

$$\mathrm{QDTW}\left(Q_1, Q_2\right) = \begin{cases} 0 \text{ si } N_1 = N_2 = 0, \\ \infty \text{ si } N_1 = 0 \text{ ou } N_2 = 0, \\ d\left(\mathbf{q}_{1,N_1}, \mathbf{q}_{2,N_2}\right) + \min \begin{cases} \mathrm{QDTW}\left(Q_{1,1:(N_1-1)}, Q_{2,1:(N_2-1)}\right) \\ \mathrm{QDTW}\left(Q_{1,N_1:i}, Q_{2,1:(N_2-1)}\right) \\ \mathrm{QDTW}\left(Q_{1,1:(N_1-1)}, Q_{2,1:N_2}\right) \end{cases} \end{cases}, \tag{3.7}$$

où $Q_{1,1:i}$ est la sous-série $(\mathbf{q}_{1,1}, \ldots, \mathbf{q}_{1,i})$ et $Q_{2,1:j}$ la sous série $(\mathbf{q}_{2,1}, \ldots, \mathbf{q}_{2,j})$. L'algorithme 4 `QDTW` permet de déterminer l'alignement par QDTW entre 2 séries chronologiques et de calculer la dissimilarité correspondante.

Algorithm 4 `QDTW`

Entrées
Q_1 : une série chronologique de dimension N_1
Q_2 : une série chronologique de dimension N_2
Fonctions
$d(.,.)$: La distance géodésique entre deux quaternions unitaires (équation 1.23))
Initialisation
C ;P : deux matrices de dimensions $N_1 \times N_2$
Début
$C[1,1] \leftarrow d(\mathbf{q}_{1,1}, \mathbf{q}_{2,1})$; $P[1,1] \leftarrow (0,0)$
for all $i \in \{2, \ldots, N_1\}$ **do**
  $C[i,1] \leftarrow d(\mathbf{q}_{1,i}, \mathbf{q}_{2,1}) + C[i-1,1]$
  $P[i,1] \leftarrow (i-1,1)$
end for
for all $j \in \{2, \ldots, N_2\}$ **do**
  $C[1,j] \leftarrow d(\mathbf{q}_{1,1}, \mathbf{q}_{2,j}) + C[1,j-1]$
  $P[1,j] \leftarrow (1,j-1)$
end for
for all $i \in \{2, \ldots, N\}$ **do**
  for all $j \in \{2, \ldots, M\}$ **do**
    $C[i,j] \leftarrow d(\mathbf{q}_{1,i}, \mathbf{q}_{2,j}) + \min\limits_{\{i^*,j^*\} \in \{\{i-1,j-i\},\{i-1,j\},\{i,j-i\}\}} C[i^*, j^*]$
    $P[i,j] \leftarrow \arg\min\limits_{\{i^*,j^*\} \in \{\{i-1,j-i\},\{i-1,j\},\{i,j-i\}\}} C[i^*, j^*]$
  end for
end for
Fin

L'élément de la *Cost Matrix* $C[i,j]$ correspond à la solution de $\text{QDTW}(Q_{1,1:i}, Q_{2,1:j})$, et l'élément de la *Path matrix* $P[i,j]$ contient les indices $\{i^*, j^*\} \in \{\{i-1, j-i\}, \{i-1, j\}, \{i, j-i\}\}$ tels que la somme $d(\mathbf{q}_{1,i}, \mathbf{q}_{2,j}) + \text{QDTW}\left(Q_{1,1:i^*}, Q_{2,1:j^*}\right)$ soit minimale. $C[N_1, N_2]$ correspond donc à la dissimilarité entre Q_1 et Q_2 après leur alignement par fonction de *warping* (solution de l'équation (3.6)). La fonction de *warping* W entre 2 séries Q_1 et Q_2 qui minimise la somme (3.6) peut être déterminée à partir de la matrice P grâce à l'algorithme 5 `GetWarpingFunction`.

Algorithm 5 `GetWarpingFunction`

Entrées
P : *Path matrix* de dimension $N_1 \times N_2$ entre 2 séries chronologiques Q_1 et Q_2, respectivement de dimension N_1 et N_2, déterminée par l'algorithme 4 `QDTW`.
Début
Initialisation
$W \leftarrow \emptyset$
$i \leftarrow N_1$
$j \leftarrow N_2$
Début
while $i \geq 1$ and $j \geq 1$ **do**
  $W \leftarrow P[i,j] \cup W$
  $(i,j) \leftarrow P[i,j]$
end while
Fin

L'algorithme 4 permet de trouver la solution de l'équation (3.6) et de calculer la dissimilarité entre deux séries chronologiques de quaternions unitaires (QTS). Il peut être utilisé comme mesure de dissimilarité pour généraliser les méthodes de Classification Ascendante Hiérachique, *K-means* et *K-medoids*. Les méthodes de *K-means* et *K-medoids* nécessitent également de déterminer le prototype d'un groupe de QTS. La détermination du médoïde peut être réalisée directement à partir des dissimilarités QDTW entre les QTS (tel que présenté dans la section 3.1.2.3). En revanche, une méthode est nécessaire pour calculer la QTS moyenne d'un ensemble de QTS en tenant compte des alignements réalisés par QDTW.

3.1.2.2 Détermination du prototype d'un groupe de QTS

Soit un groupe de n QTS $\{Q_i\}_{i=1,2,\ldots,n}$, de dimension $\{N_i \in \mathbb{N}\}_{i=1,2,\ldots,n}$. On définit le *prototype* de ce groupe comme la série *consensus* $\Phi = (\phi_1, \ldots, \phi_{N_\Phi})$ telle que, pour toute

autre QTS $\Psi = (\psi_1, \ldots, \psi_{N_\Psi})$:

$$\sum_{i=1}^{n} \mathrm{QDTW}^2 (Q_i, \Phi) \leq \sum_{i=1}^{n} \mathrm{QDTW}^2 (Q_i, \Psi) . \tag{3.8}$$

La définition de Φ donnée dans l'équation (3.8) correspond à la *séquence de Steiner* [59]. La somme de l'équation (1.40) correspond à la *Somme des carrés intra-groupes* (ou *within cluster sum of square*, WSS), également appelée *inertie intra-groupe*. Trouver la solution exacte de l'équation (3.8) n'est pas réalisable à cause de la puissance de calcul nécessaire pour explorer l'ensemble des QTS "candidates" pour Φ (voir Petitjean *et al.* [128] (2011) pour une description plus précise de cette problématique).

Pour obtenir une approximation du prototype d'un ensemble de QTS, l'algorithme `Dtw Barycenter Averaging` (`DBA`) décrit initialement par Petitjean *et al.* (2011) [128] est adapté aux quaternions unitaires. L'algorithme `DBA` est conçu à l'origine pour calculer itérativement le prototype Φ d'un ensemble de séries temporelles $\{S_i = (s_{i,1}, \ldots, s_{i,N_i})\}_{i=1,2,\ldots,n}$ dont les éléments appartiennent à un espace euclidien par la répétition de deux étapes :

(1) Étape d'alignement : Aligner chacune des séries temporelles par `DTW` avec le prototype $\Phi^{[\ell-1]} = \left(\phi_1^{[\ell-1]}, \ldots, \phi_{N_\Phi}^{[\ell-1]}\right)$, défini à l'itération $\ell - 1$.

(2) Étape de calcul du prototype : Calculer le nouveau prototype initial $\Phi^{[\ell]}$, dont chaque élément $\phi_j^{[\ell]}$ est le barycentre des éléments des séries temporelles $\{S_i\}_{i=1,2,\ldots,n}$ qui ont été alignés avec $\phi_j^{[\ell-1]}$.

L'adaptation de l'algorithme `DBA` aux QTS nécessite donc d'étendre le concept de barycentre à un ensemble de m quaternions unitaires $\{\mathbf{q}_i\}_{i=1,2,\ldots,m}$. Tomasz *et. al* (2017) [60] proposent une adaptation de l'algorithme `DBA` aux QTS en calculant la moyenne d'un quaternion par l'algorithme de Markley *et. al* (2007) [108]. Par souci d'optimisation, on suppose ici que l'exponentiel de la moyenne arithmétique des log-quaternions unitaires en est une approximation suffisante (*c.f.* section 1.2.2) :

$$\bar{\mathbf{q}} = \exp\left(\sum_{i=1}^{m} \frac{\ln(q)}{m}\right) \tag{3.9}$$

L'utilisation de l'algorithme 4 `QDTW` et de l'équation (3.9) permet d'adapter l'algorithme `DBA` aux séries de quaternions unitaires. Cette adaptation, appelée `QDBA` est présentée dans l'algorithme 6.

L'algorithme 6 `QDBA` peut nécessiter plusieurs itérations pour converger vers une solution stable, *i.e.* que le prototype obtenu reste le même après une itération supplémentaire

Algorithm 6 `QDBA`

Entrées
$\{Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N_i})\}_{i=1,2,\ldots,n}$: Un ensemble de n QTS.
$\Phi^{[\ell-1]} = \left(\phi_1^{[\ell-1]}, \ldots, \phi_{N_\Phi}^{[\ell-1]}\right)$: le prototype de l'ensemble de QTS $\{Q_i\}_{i=1,2,\ldots,n}$ défini à l'itération $\ell - 1$.
Fonctions
`QDTW.P`(Q_1, Q_2) : Aligne les séries Q_1 et Q_2 par l'algorithme 4 `QDTW` et renvoie la *Path matrix* P
Initialisation
for all $j \in \{1, \ldots, N_\Phi\}$ **do**
  $A_j \leftarrow \{\emptyset\}$
end for
Début
for all $i \in \{1, \ldots, n\}$ **do**
  $P \leftarrow$ DTW.P$(\Phi^{[\ell-1]}, Q_i)$ // Alignement par QDTW entre $\Phi^{[\ell-1]}$ et Q_i
  $j \leftarrow N_\Phi$
  $k \leftarrow N_i$
  while $j \geq 1$ and $k \geq 1$ **do**
    $A_j \leftarrow \{A_j\} \cup \{\mathbf{q}_{i,k}\}$
    $(j, k) \leftarrow P[j, k]$
  end while
end for
// A_j contient tous les quaternions des QTS $\{Q_i\}_{i=1,2,\ldots,n}$ alignés avec $\phi_j^{[\ell-1]}$
for all $j \in \{1, \ldots, N_\phi\}$ **do**
  $\phi_j^{[\ell]} \leftarrow \exp \sum_{\mathbf{q} \in A_j} \frac{\ln(\mathbf{q})}{|A_j|}$ // Calcul de la moyenne des quaternions alignés avec $\phi_j^{[\ell-1]}$
end for
Renvoyer $\Phi^{[\ell]}$
Fin

des étapes d'alignement et de calcul du prototype. Le choix du prototype initial $\Phi^{[0]}$ lors de l'initialisation de `QDBA` influence grandement le prototype final obtenu. Cet aspect est décrit et étudié par Petitjean *et al.* (2011) [128], et les auteurs proposent le choix d'une des séries du groupe comme prototype initial.

L'algorithme `QDBA` permet donc de calculer le prototype d'un groupe de QTS en tenant compte des alignements temporels entre séries par l'algorithme `QDTW`. La section suivante présente comment l'utilisation de `DTW` et `QDBA` permet de généraliser les méthodes de Classification Ascendante Hiérachique, du *K-means* et du *K-medoids*.

3.1.2.3 Algorithmes de classification de QTS

Classification Ascendante Hiérarchique. Soit un ensemble de n QTS $\{Q_i\}_{i=1,2,\dots,n}$, la matrice D de dissimilarité de dimension $n \times n$ est calculée telle que :

$$d(i,j) = \text{QDTW}(Q_i, Q_j), \tag{3.10}$$

$\text{QDTW}(Q_i, Q_j)$ étant défini comme l'élément $C[N_i, N_j]$ de la *Cost Matrix* renvoyée par l'algorithme 4 `QDTW`. Une Classification Ascendante Hiérarchique peut alors être appliquée sur la matrice D. Comme pour la classification de données fonctionnelles de quaternions unitaires, les observations $\{Q_1, \dots, Q_n\}$ n'appartenant pas à un espace euclidien, seuls les critères de liaison simple, moyenne et complète sont adaptés pour les étapes de fusion et de mise à jour de la matrice D (*c.f.* présentation de la formule de récurrence de Lance and Williams 1.31 au paragraphe 1.3.1.1).

K-means* et *K-medoids Pour l'adaptation du *K-means* et du *K-medoids* à la classification d'un ensemble de n QTS $\{Q_i\}_{i=1,2,\dots,n}$, on définit :

- $\boldsymbol{\Phi} = \{\Phi_k\}_{k=1,2,\dots,K}$: l'ensemble des prototypes (ou prototypes) représentant les K groupes.
- $d(Q_i, \Phi_k) = \text{QDTW}(Q_i, \Phi_k)$: la mesure de dissimilarité entre l'observation i et le groupe k, correspondant à l'élément $C[N_i, N_{\Phi_k}]$ de la *Cost Matrix* calculée par l'algorithme 4 `QDTW`.
- $\lambda(\boldsymbol{\Phi}, Q_i) = \underset{r \in \{1,\dots,K\}}{\arg\min}\ \text{QDTW}(Q_i, \Phi_r)$: la fonction d'attribution de l'observation i. $\lambda(\boldsymbol{\Phi}, Q_i) = k$ signifie que le meilleur alignement de la QTS Q_i est observé avec le prototype Φ_k, l'observation i devant donc être affectée au groupe k.

— $V_k = \{i \in [\![1, n]\!] : \lambda\left(Q_i^{(ln)}, \mathbf{\Phi}\right) = k\}$: définit l'ensemble des observations appartenant au groupe k.

— $n_k = |V_k|$: Le nombre d'observations affectées au groupe k.

Soient $\mathbf{\Phi}^{[\ell-1]} = \left\{\Phi_k^{[\ell-1]}\right\}_{k=1,2,\ldots,K}$ l'ensemble des prototypes des groupes après l'itération $\ell - 1$. L'itération ℓ consiste en la succession d'étapes suivante :

1. *Calcul des prototypes :*

   — Algorithme *K-means* : Pour $k \in \{1, \ldots, K\}$, le prototype $\Phi_k^{[\ell]}$ est calculé par itérations de l'algorithme 6 `QDBA` jusqu'à convergence à partir de l'ensemble des fonctions $Q_1^{(ln)[\ell-1]}$ pour lesquelles $\lambda\left(\mathbf{\Phi}^{[\ell-1]}, Q_1^{(ln)[\ell-1]}\right) = k$:

   $$\Phi_k^{[\ell]} = \text{QDBA}\left(\Phi_k^{[\ell-1]}, \{Q_i\}_{i \in V_k}\right)$$

   — Algorithme *K-Medoids* : Pour $k \in \{1, \ldots, K\}$, le prototype $\Phi_k^{[\ell]}$ est déterminé comme le médoïde du groupe k :

   $$\Phi_k^{[\ell]} = \underset{Q_i, i \in V_k}{\arg\min} \sum_{j \in V_k} \text{QDTW}\left(Q_i, Q_j\right)$$

2. *Affectation des QTS* - Pour $i \in \{1, \ldots n\}$, la fonction Q_i est affectée au groupe $\lambda\left(\mathbf{\Phi}^{[\ell]}, Q_i\right)$.

Les algorithmes de *K-means* et *K-medoids* sont initialisés en choisissant choisissant aléatoirement l'ensemble des prototypes initiaux $\mathbf{\Phi}^{[0]}$ parmi l'ensemble des fonctions initiales $\{Q_1, \ldots, Q_n\}$. Ils sont stoppés quand une itération supplémentaire entraîne une diminution de l'inertie globale $\sum_{k=1}^{K} \sum_{i \in V_k} \text{QDTW}^2\left(\Phi_k^{[\ell]}, Q_i\right)$ inférieure à un seuil.

Les sections 3.1.1 et 3.1.2 présentent les modifications apportées aux méthodes de CAH, *K-means* et *K-medoids* de *données fonctionnelles* et de *séries chronologiques* pour la classification de séquences de quaternions unitaires. Ces modifications reposent sur l'adaptation du calcul de dissimilarité entre observations et du calcul du prototype des groupes à ce type de données. Les méthodes choisies permettent de s'affranchir du mauvais alignement temporel entre les données.

Les généralisations aux séries chronologiques de quaternions unitaires des algorithmes `DTW` pour le calcul de la dissimilarité entre QTS et `DBA` pour le calcul du prototype d'un groupe de QTS sont déjà décrites respectivement par Jablonski (2011) [82] et Tomasz *et.*

al (2017) [60]. Elles permettent donc d'adapter aisément les méthodes de classification CAH, *K-means* et *K-medoids* aux QTS.

Aucune méthode de classification adaptée aux données fonctionnelles de quaternions unitaires n'a été identifiée dans la littérature. La transformation des données dans l'espace des log-quaternions est donc proposée pour adapter les méthodes de CAH, *K-means alignment* et *K-medoids alignment* à ce type de données. Toutes ces méthodes reposent sur l'alignement des données par fonctions de *warping* affines strictement croissantes et d'une métrique adaptée pour respecter le cadre théorique défini par Vantini (2011) [171].

Des applications de ces méthodes de classification aux séquences de quaternions unitaires sont présentées. La section 3.1.3 présente le calcul du biomarqueur *Signature de Marche* d'un individu par l'application du *K-means alignment* sur les cycles de marche détectés par l'algorithme `STRIPAGE`. Dans la section 3.1.4, les méthodes CAH, *K-means* et *K-medoids* de *données fonctionnelles* et de *séries chronologiques* sont appliquées pour classifier les SdMs de volontaires sains en fonction de la présence ou non d'un déficit de marche simulé.

3.1.3 Méthode *K-means alignment* pour le calcul du biomarqueur *Signature de Marche*

L'algorithme `STRIPAGE` présenté dans la section 2.1.1 permet d'identifier un ensemble de séquences de quaternions unitaires représentant l'orientation de la hanche au cours des cycles de marche d'un individus portant le dispositif *MetaMotionR*. L'objectif est ici de synthétiser la démarche d'un individu par une unique séquence de quaternions unitaires dont la forme est représentative de ses cycles de marche.

3.1.3.1 Calcul du prototype des cycles de marche par *K-means alignment*

On considère un ensemble de m_i séquences de quaternions unitaires $\{Q_{i,j}\}_{j=1,2,\dots,m_i}$ représentant les cycles de marche détectés par l'algorithme `STRIPAGE` (présenté dans la section 2.1) dans les données mesurées par le dispositif *MetaMotionR* porté par un individu i au cours d'une phase d'acquisition. Le prototype de ces cycles, appelé *Signature de Marche* (SdM), est obtenu par le procédé suivant :

(1) Les séquences $\{Q_{i,j}\}_{j=1,2,\dots,m_i}$ sont transformées par fonction logarithmique (voir équation (1.17)) pour obtenir les séquences $\left\{Q_{i,j}^{(ln)}\right\}_{j=1,2,\dots,m_i}$.

(2) Le prototype des séquences $\left\{Q_{i,j}^{(ln)}\right\}_{j=1,2,\ldots,m_i}$ est calculé après alignement par fonctions de *warping* affines par la méthode du *K-means alignment*, avec $K = 1$ (la problématique d'alignement de données fonctionnelles et de l'adaptation du *K-means alignment* aux fonctions de log-quaternions est présentée section 3.1.1). Le prototype obtenu par cette méthode est noté $Q_i^{(ln)}$.

(3) Le prototype $Q_i^{(ln)}$ est transformé par fonction exponentielle (voir équation 1.10) pour obtenir le prototype Q_i sous la forme d'une fonction quaternionique).

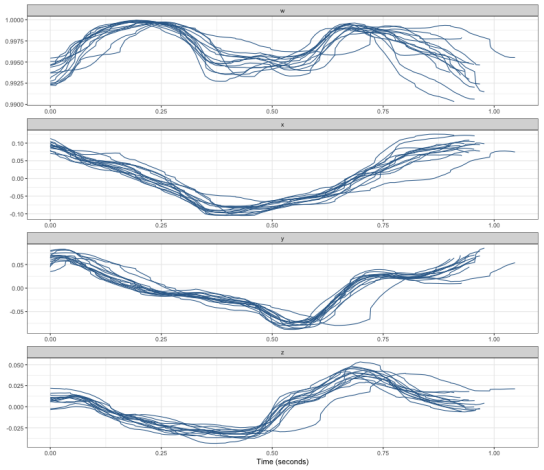
(4) Le prototype Q_i est échantillonné en une grille de 101 quaternions unitaires par interpolation linéaire sphérique (*c.f.* description de la méthode *slerp*, équation (1.24), section 1.2.2.1), et les temps associés à ces mesures sont exprimés en pourcentage de la durée totale du prototype Q_i, sur une grille T_i allant de 0% à 100% (les mesures étant séparées d'un temps équivalent à 1% de la durée totale du prototype).

Le prototype obtenu est une séquence de quaternions unitaires $Q_i = (\mathbf{q}_1, \ldots, \mathbf{q}_{101})$ associés aux temps de mesure $T_i = (t_{i,1}, \ldots, t_{i,101}) = (0\%, \ldots, 100\%)$. Il est supposé représenter la démarche de l'individu par les rotations mesurées au niveau de sa hanche au cours d'un cycle de marche, et est appelé *Signature de Marche* (SdM). La figure 3.1 présente un exemple de calcul de SdM à partir d'un ensemble de cycle de marche (figure 3.1a). Les cycles de marche exprimés dans l'espace des log-quaternions sont représentés (i) avant alignement en haut de la figure 3.1b, et (ii) après alignement avec leur prototype (courbe rouge graissée) en bas de la figure 3.1b. Enfin, la *Signature de Marche* (en bleu) et les cycles de marche à partir desquels elle a été calculée (en gris) sont présentés en pourcentage de durée du cycle sur la figure 3.1c.

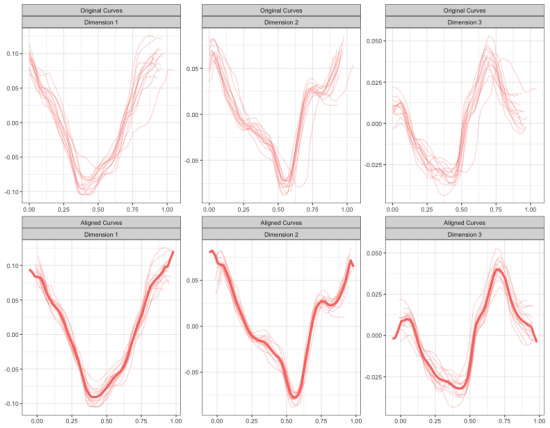
3.1.3.2 Signature de marche et orientation de référence

Il est nécessaire d'exprimer les rotations de la hanche en considérant une orientation de référence commune entre les SdMs de différents individus pour pouvoir en comparer la forme. Cette *orientation de référence* $\mathbb{R}_f$ correspond au référentiel dans lequel est calculée l'orientation du dispositif MMR, *i.e.* le quaternion unitaire de la rotation entre le référentiel fixe $\mathbb{R}_f$ et le référentiel propre au système de capteurs $\mathbb{R}_s$ (*c.f.* section 1.2.2.2). On définit une *orientation commune* entre les SdMs de différents individus comme une orientation observée à des instants du cycles de marche équivalents entre les SdMs. Deux *orientations de références* sont aisément identifiables au cours d'un cycle de marche :

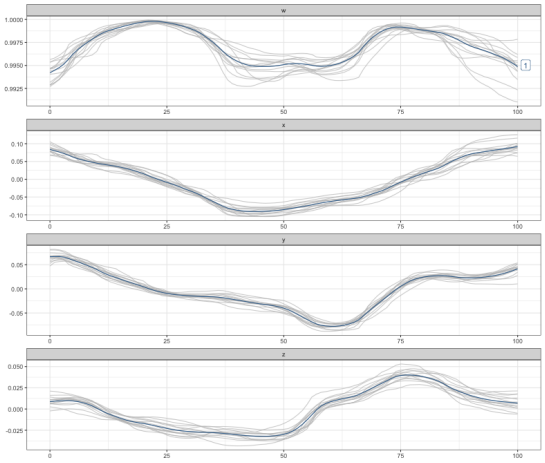
— L'*orientation moyenne* correspond à l'instant où les deux jambes sont alignées, cet



(a) Ensemble de cycles de marche $\{Q_{i,j}\}_{j=1,2,\ldots,m_i}$



(b) Alignement des cycles de marche $\{Q_{i,j}^{(ln)}\}_{j=1,2,\ldots,m_i}$ et calcul du prototype $Q_i^{(ln)}$



(c) Cycles de marche $\{Q_{i,j}\}_{j=1,2,\ldots,m_i}$ et *Signature de Marche* (en pourcentage de durée du cycle)

FIGURE 3.1 – Calcul de la Signature de Marche

évènement survenant à deux reprises durant un cycles de marche : la première durant la phase d'appui et la seconde durant la phase de balancement (*c.f.* section 2.1.1.6, **Étape 5.2. : Réorientation des segments**.

— L'*orientation initiale* correspond à la première et la dernière orientation observée au cours du cycle, *i.e.* par définition la pose du talon droit au sol.

Il est donc possible de définir deux représentations différentes de la SdM, qui diffèrent par l'orientation de référence utilisée pour calculer l'orientation de la hanche et donc par les instants auxquels la trajectoire de la SdM se rapprochent du quaternion unitaire $\mathbf{q}_0 = (1, 0, 0, 0)$. Il n'y a *a priori* pas de raison de privilégier l'une de ces représentations par rapport à l'autre pour comparer les SdMs de plusieurs individus. Par conséquent, les deux approches seront explorées.

La première représentation, appelée *Signature de marche brute*, est calculée directement par la méthode présentée dans la section 3.1.3.1 à partir des cycles de marche identifiés par l'algorithme `STRIPAGE`, qui ont été traités pour que l'orientation de la hanche au cours d'un cycle soit calculée à partir de l'orientation moyenne (*c.f.* section 2.1.1.6, **Étape 5.2. : Réorientation des segments**). Par construction, les quaternions $\mathbf{q}_{i,j}$ qui composent la SdM représentent donc la rotation entre l'*orientation moyenne* de la hanche observée au cours de la SdM et l'orientation de la hanche observée au $(j-1)^{eme}$ pourcent de la SdM.

La seconde représentation de la SdM est obtenue à partir de la *SdM brute*, en déterminant l'orientation observée au temps initial $t_1 = 0\%$ comme l'orientation de référence. Les orientations initiales et finales de la hanche au cours d'un cycle de marche étant théoriquement identiques, la méthode de réorientation est construite de telle sorte que les premier et dernier éléments de la SdM correspondent à la rotation identité, *i.e.* le quaternion $\mathbf{q} = (1, 0, 0, 0)$. Pour ce faire, la SdM est transformée par fonction ln (voir équation 1.11) pour obtenir la log-SdM $Q_i^{(ln)} = \left(\mathbf{q}_{i,1}^{(ln)}, \ldots, \mathbf{q}_{i,101}^{(ln)}\right)$. Les composantes $\mathbf{x}_i^{(ln)}$, $\mathbf{y}_i^{(ln)}$, et $\mathbf{z}_i^{(ln)}$ de $Q_i^{(ln)}$ sont considérées comme des séries uni-variées : $\mathbf{x}^{(ln)} = \left(x_{i,1}^{(ln)}, \ldots, x_{i,101}^{(ln)}\right)$, $\mathbf{y}^{(ln)} = \left(y_{i,1}^{(ln)}, \ldots, y_{i,101}^{(ln)}\right)$ et $\mathbf{z}^{(ln)} = \left(z_{i,1}^{(ln)}, \ldots, z_{i,101}^{(ln)}\right)$. Chacune de ces séries est soustraite par la droite passant par son premier et dernier élément. La SdM *réalignée* sur ses premier et dernier éléments est ensuite obtenue par transformation exponentielle des quaternions formés par les composantes ainsi obtenues (on rappelle que la composante w d'un log-quaternion unitaire est égale à 0, il n'est donc pas nécessaire d'y appliquer cette transformation). Cette transformation appelée `StraightenQTS` est présentée dans l'algorithme 7.

Algorithm 7 `StraigthenQTS`

Paramètres
$Q_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,101})$: une série temporelle de quaternions unitaires
$T_i = (t_{i,1}, \ldots, t_{i,101}) = (0, \ldots, 100)$: la grille des temps de mesure de la série Q_i
Début
$Q_i^{(ln)} \leftarrow (\ln \mathbf{q}_{i,1}, \ldots, \ln \mathbf{q}_{i,101})$, avec $\ln \mathbf{q}_{i,j} = \left(0, x_{i,j}^{(ln)}, y_{i,j}^{(ln)}, z_{i,j}^{(ln)}\right) \forall j \in \{1, \ldots, 101\}$
$\mathbf{x}_i^{(ln)} \leftarrow \left(x_{i,1}^{(ln)}, \ldots, x_{i,101}^{(ln)}\right)$; $\mathbf{y}_i^{(ln)} \leftarrow \left(y_{i,1}^{(ln)}, \ldots, y_{i,101}^{(ln)}\right)$; $\mathbf{z}_i^{(ln)} \leftarrow \left(z_{i,1}^{(ln)}, \ldots, z_{i,101}^{(ln)}\right)$
for all $j \in \{1, \ldots, 101\}$ **do**

$$x_{i,j}^{s,(ln)} \leftarrow x_{i,j}^{(ln)} - \left(x_{i,1}^{(ln)} + \left(\frac{x_{i,101}^{(ln)} - x_{i,1}^{(ln)}}{100}\right) \times (j-1)\right)$$

$$y_{i,j}^{s,(ln)} \leftarrow y_{i,j}^{(ln)} - \left(y_{i,1}^{(ln)} + \left(\frac{y_{i,101}^{(ln)} - y_{i,1}^{(ln)}}{100}\right) \times (j-1)\right)$$

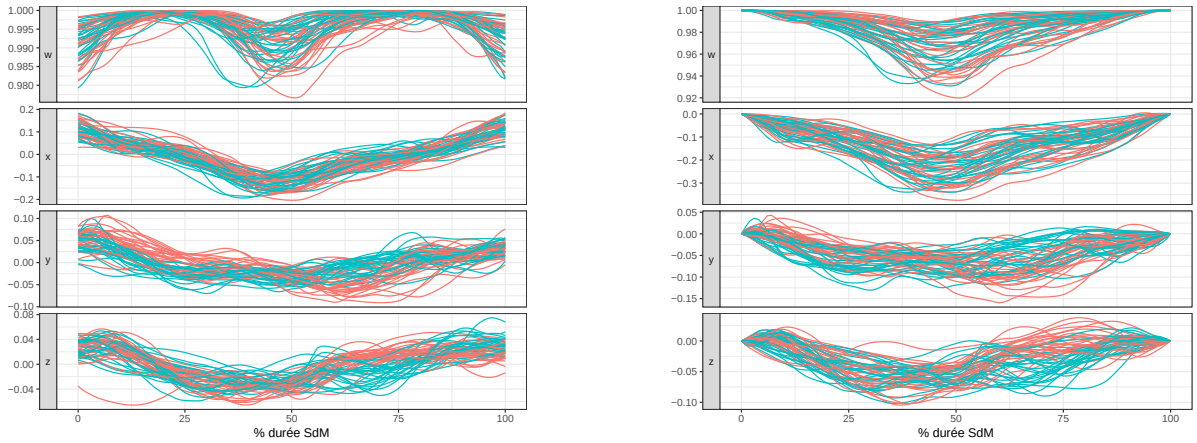
$$z_{i,j}^{s,(ln)} \leftarrow z_{i,j}^{(ln)} - \left(z_{i,1}^{(ln)} + \left(\frac{z_{i,101}^{(ln)} - z_{i,1}^{(ln)}}{100}\right) \times (j-1)\right)$$

end for
$Q_i^{s,(ln)} \leftarrow \left(\mathbf{q}_{i,1}^{s,(ln)}, \ldots, \mathbf{q}_{i,101}^{s,(ln)}\right)$, avec $\mathbf{q}_{i,j}^{s,(ln)} = \left(0, x_{i,j}^{s,(ln)}, y_{i,j}^{s,(ln)}, z_{i,j}^{s,(ln)}\right) \forall j \in \{1, \ldots, 101\}$
$Q_i^s = \left(\exp \mathbf{q}_{i,1}^{s,(ln)}, \ldots, \exp \mathbf{q}_{i,101}^{s,(ln)}\right)$
Renvoyer Q_i^s
Fin

Pour résumer, deux représentations de la *Signature de Marche* peuvent être calculées, en fonction de l'orientation choisie comme référence :

— *Signature de Marche "brute"* : L'orientation de référence correspond à l'orientation *moyenne* observée au cours de la SdM. Chaque quaternion unitaire correspond à la rotation entre cette orientation de référence et l'orientation observée au cours de la SdM. La trajectoire de la SdM passe proche du quaternion unitaire représentant la rotation identité $\mathbf{q}_0 = (1, 0, 0, 0)$ à deux reprises. La première est observée durant la phase d'appui, la seconde durant la phase de balancement, et elles correspondent aux instants où les deux jambes se croisent (*c.f.* section 2.1.1.6, paragraphe **Étape 5.2 : Réorientation des segments**).

— *Signature de Marche "réalignée"* : Elle est calculée à partir de la SdM *brute* avec l'algorithme 7 `StraightenQTS`. L'orientation de référence correspond à l'orientation initiale de la SdM. Chaque quaternion unitaire correspond à la rotation entre cette orientation de référence et l'orientation observée au cours de la SdM. Le premier et le dernier élément de la SdM dite *réalignée* correspondent au quaternion $\mathbf{q}_0 = (1, 0, 0, 0)$.



(a) SdM *brutes* (b) SdM réalignées

FIGURE 3.2 – *Signature de Marche* de 27 volontaires sains

3.1.3.3 Exemple de calcul de la *Signature de Marche* de volontaires sains

Un exemple de calcul de *Signature de Marche brute* et *réalignées* à partir des données de la `BDDtest` est présenté ici (*c.f.* description de cette base de données section 1.2.3.1). Pour chaque volontaire, un ensemble de cycles de marche est détecté par l'algorithme `STRIPAGE` (*c.f.* description de l'algorithme section 2.1.1). Deux SdMs *brutes* sont calculées pour chaque volontaire d'après la méthode présentée section 3.1.3.1, la première à partir de l'ensemble de ses cycles de marche mesurés en condition de marche libre et la seconde à partir de l'ensemble de ses cycles de marche mesurés en condition de marche contrainte. Les SdMs *réalignées* sont aussi calculées en appliquant l'algorithme 7 `StraightenQTS` sur les SdMs *brutes*.

Les résultats sont présentés sur la figure 3.2. Les SdMs mesurées en condition de marche libre sont représentées par les courbes rouges et celles mesurées en condition de marche contrainte sont représentées par les courbes bleues. Les SdMs *brutes* sont visibles sur la figure 3.2a. Sur la figure 3.2b, on peut voir que chaque SdM *réalignée* par l'algorithme 7 `StraightenQTS` débute et termine par le quaternion unitaire $\mathbf{q}_0 = (1, 0, 0, 0)$.

3.1.4 Application : classification non supervisée de données de marche avec et sans déficit de marche simulé

Cette section présente l'application des algorithmes *K-means*, *K-medoids* et CAH sur la base `BDDtest` (*c.f.* description de cette base de données section 1.2.3.1). On rappelle

que cette base est composée des données de marche de volontaires sains mesurées par le systèmes de capteurs *MetaMotionR* en condition de marche réelle et marche contrainte. La condition de marche contrainte a pour but de simuler un déficit mécanique de la marche, l'articulation du genou étant bloquée par le port d'une orthèse. L'objectif est de comparer la capacité des méthodes de classification à construire des groupes d'individus en fonction des conditions dans lesquelles les données de marche ont été mesurées.

Pour ce faire, la démarche d'un individu dans une condition de marche est représentée par sa *Signature de Marche* (SdM) calculée par la méthode présentée dans la section 3.1.3 à partir des segments détectés par l'algorithme *STRIPAGE*. Chaque individu est représenté par deux SdMs, la première représentant sa démarche en condition de marche normale et la seconde représentant sa démarche en condition de marche contrainte. Des échantillons sont générés à partir de l'ensemble des SdMs selon deux règles : (i) un volontaire n'est représenté que par une SdM (sa SdM mesurée en condition de marche libre ou contrainte), (ii) les nombres de SdMs mesurées en condition de marche libre et contrainte sont équilibrés.

Les méthodes de CAH, *K-means* et *K-medoid* pour données fonctionnelles de quaternions unitaires et séries chronologiques de quaternions unitaires sont appliquées sur ces échantillons pour former 2 groupes. Les résultats sont évalués par le calcul de l'indice de Rand ajusté [74] entre les groupes formés par les méthodes de classification et les conditions de marche dans lesquelles les données ont été mesurées. La robustesse des méthodes de classification est évaluée en calculant le coefficient de variation de l'indice de Rand.

3.1.4.1 Plan d'expérience

La méthodologie pour évaluer les méthodes de classification est présentée dans cette section. Elle décrit la méthode d'échantillonage des SdMs, leur pré-traitement et les critères d'évaluation des partitions obtenues.

Formation d'échantillon de SdMs Deux SdMs sont calculées pour chaque volontaire, la première à partir de l'ensemble de ses cycles de marche mesurés en *condition de marche libre* et la seconde à partir de l'ensemble de ses cycles de marche mesurés en *condition de marche contrainte*. Les SdMs ensuite donc échantillonnées de telle sorte que :

(i) Un échantillon soit constitué d'un nombre équivalent de SdMs mesurées en condition de marche contrainte et en marche libre, *i.e.* 13 SdMs en marche libre et 14 en marche contrainte.

(ii) Un volontaire ne soit représenté dans un échantillon que par l'une de ses SdMs (mesurée en marche libre ou contrainte).

Deux versions de chaque échantillon sont formées. Un volontaire est représenté par la SdM mesurée dans la même condition de marche dans chacune des deux versions de l'échantillon, la première version contenant les SdMs *brutes* et la seconde version contenant les SdMs *réalignées* par l'algorithme 7 `StraightenQTS`. La figure 3.3 schématise la méthode d'échantillonnage des SdMs.

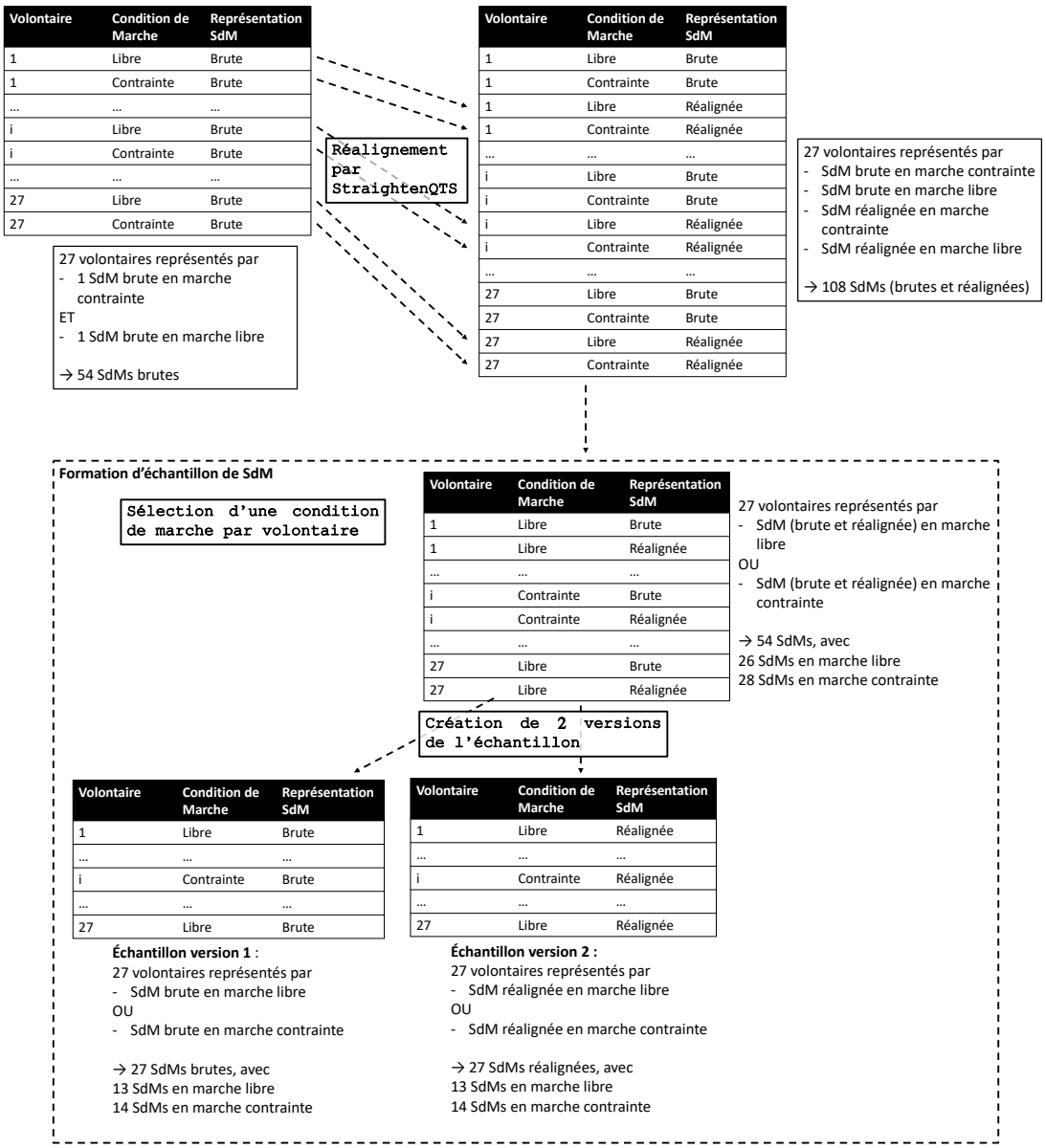


FIGURE 3.3 – Méthode d'échantillonnage

Transformation des SdMs par égalisation des angles de rotation Un premier essai d'application des méthodes de classification sur les SdMs a montré que, quel que soit leur type (brutes ou réalignées), les groupes obtenus ne correspondaient pas aux conditions de marche dans lesquelles elles avaient été mesurées. Les méthodes de classification semblent former des groupes en fonction de l'amplitude de la SdM, et cette amplitude n'est pas liée au port ou non de l'orthèse bloquante du genou. Pour s'affranchir de la variabilité d'amplitude entre les SdMs, la méthode de transformation suivante est appliquée pour chaque version de chaque échantillon. Elle permet d'exprimer un ensemble de SdMs de telle sorte que l'angle de rotation observé à un pourcentage de durée donné soit égal entre toutes les SdMs.

On note $\{Q_i\}_{i=1,2,\dots,27}$ l'ensemble des 27 SdMs d'une version d'un échantillon, avec $Q_i = \{\mathbf{q}_{i,1}, \dots, \mathbf{q}_{i,101}\} \forall i \in \{1, \dots, 27\}$. La moyenne *point par point* des SdMs de l'échantillon est calculée :

$$\bar{Q} = (\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_{101}) = \left(\exp \sum_{i=1}^{27} \frac{\ln \mathbf{q}_{i,1}}{27}, \dots, \exp \sum_{i=1}^{27} \frac{\ln \mathbf{q}_{i,101}}{27} \right). \tag{3.11}$$

De façon similaire à l'adaptation de l'algorithme `DBA` aux séries chronologiques de quaternions unitaires (voir section 3.1.2.2), on suppose qu'une bonne approximation de la moyenne d'un ensemble de quaternions unitaires peut être obtenue par l'exponentiel de leur moyenne arithmétique dans l'espace des log-quaternions (*c.f.* équation (1.17) et équation (1.10) pour la présentation des transformations logarithmique et exponentielle des quaternions). La série des angles des rotations représentées par $\bar{Q}$ est calculée (*c.f.* équation (1.14) présentant le calcul d'un angle de rotation à partir d'un quaternion unitaire) :

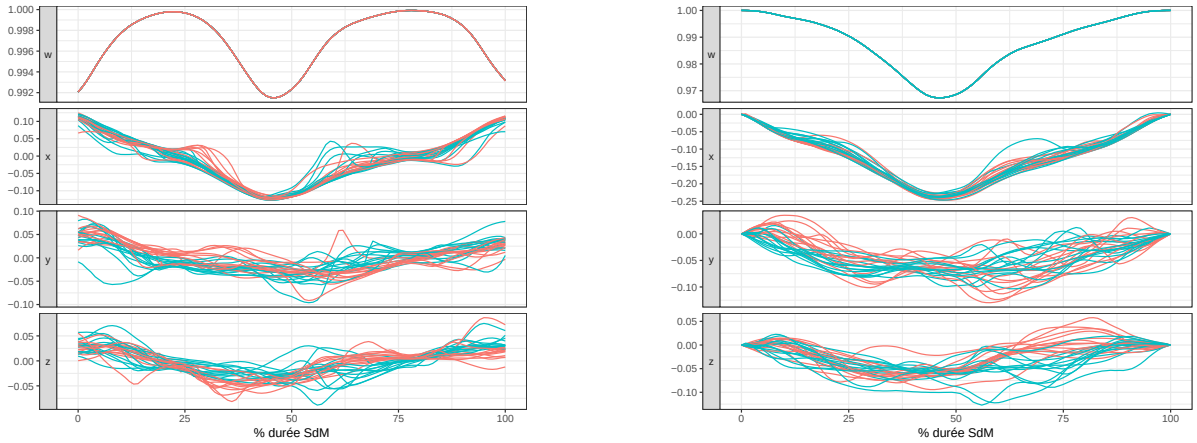
$$\bar{\boldsymbol{\theta}} = \left(2 \arccos \mathrm{Re}(\bar{\mathbf{q}}_1), \dots, 2 \arccos \mathrm{Re}(\bar{\mathbf{q}}_{101}) = (\bar{\theta}_1, \dots, \bar{\theta}_{101} \right). \tag{3.12}$$

La série des axes des rotations de chacune des SdMs est calculée (*c.f.* équation (1.15) présentant le calcul d'un axe de rotation à parti d'un quaternion unitaire) :

$$\left\{ U_i = (\mathbf{u}_{i,1}, \dots, \mathbf{u}_{i,101}) = \left(\frac{\mathrm{Im}(\mathbf{q}_{i,1})}{\|\mathrm{Im}(\mathbf{q}_{i,1})\|}, \dots, \frac{\mathrm{Im}(\mathbf{q}_{i,101})}{\|\mathrm{Im}(\mathbf{q}_{i,101})\|} \right) \right\}_{i=1,2,\dots,n} \tag{3.13}$$

Enfin, les SdMs sont centrées sur la série $\bar{\boldsymbol{\theta}}$:

$$\left\{ Q_i^{(\bar{\boldsymbol{\theta}})} = \left(\mathbf{q}_{i,1}^{(\bar{\theta}_1)}, \dots, \mathbf{q}_{i,101}^{(\bar{\theta}_{101})} \right) \right\}_{i=1,2\dots,n}, \tag{3.14}$$



(a) SdMs *brutes θ-égalisées* (b) SdMs *réaligneés θ-égalisées*

FIGURE 3.4 – *Signature de Marche θ-égalisées* de 27 volontaires sains

avec

$$\mathbf{q}_{i,j}^{(\bar{\theta})} = \left(\cos\frac{\bar{\theta}_j}{2}, \sin\frac{\bar{\theta}_j}{2}\mathbf{u}_{i,j}\right), j \in \{1,\ldots,101\}, i \in \{1,\ldots,n\}. \tag{3.15}$$

Les SdMs $Q_i^{(\bar{\boldsymbol{\theta}})}$ sont dites *θ-égalisées*. Les figures 3.4a et 3.4b présentent respectivement les SdMs *brutes* et *réalignées θ-égalisées* d'un échantillon généré selon les règles (i) et (ii) énoncées au paragraphe précédent.

Classification et évaluation des partitions obtenues Pour résumer, une SdM *brute* et une SdM *réalignée* est calculée pour chaque volontaire et pour chaque condition de marche (libre et contrainte). Plusieurs échantillons sont générés, chaque échantillon (i) étant constitué d'un nombre équilibré de SdMs mesurées en condition de marche libre et en condition de marche contrainte, et (ii) ne contenant qu'une des deux SdMs d'un volontaire (la SdM mesurée en condition de marche libre ou contrainte). Chaque échantillon est décliné en deux versions, l'une contenant des SdM *brutes* et l'autre contenant des SdMs *réalignées*. Pour chacune des versions de chacun des échantillons, les SdMs sont *θ-égalisées* à partir de la série des angles de rotation calculée à partir de leur moyenne *point par point*. Une version d'un échantillon contient donc un ensemble de SdMs *θ-égalisées*, *brutes* ou *réalignées* et mesurées en condition de marche libre ou contrainte.

Les méthodes de Classification Ascendante Hiérarchique (CAH), de *K-means* et de *K-medoid* adaptées aux séries chronologiques de quaternions unitaires (*c.f.* section 3.1.2.3) et les méthodes de CAH, de *K-means alignment* et de *K-medoid alignment* adaptées aux

données fonctionnelles de quaternions unitaires (*c.f.* section 3.1.1) sont appliquées sur chacune des versions de chacun des échantillons. Le nombre de groupes à former pour les méthodes dérivées du *K-means* et du *K-medoid* est fixé à $K = 2$. Les dendrogrammes obtenus par les adaptations de la CAH sont coupés à la hauteur permettant de former 2 groupes.

Dans le contexte de cette expérience, les classes auxquelles appartiennent les observations sont connues. Les méthodes de classification peuvent donc être évaluées au moyen d'un critère de validation externe. Pour ce faire, l'*indice de Rand ajusté* [74] est calculé à partir du tableau de contingence ci-dessous :

$C \diagdown G$	G_1	G_2	
C_1	n_{11}	n_{12}	$n_{1.}$
C_2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

avec G_1 et G_2 les groupes formés par la méthode de classification, C_1 et C_2 les classes connues *a priori* pour les observations (ici : conditions de marche contrainte et libre), et $n_{i,j}$ le nombre d'observations appartenant à la fois aux classes C_i et G_j. L'*indice de Rand Ajusté* (ARI) est alors obtenu d'après l'équation suivante :

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}} \tag{3.16}$$

L'ARI vaut 1 quand la répartition en deux groupes correspond parfaitement à la répartition en condition de marche, et diminue à mesure que les deux partitions diffèrent l'une de l'autre. Il peut prendre une valeur négative quand la correspondance entre les groupes et les classes connues *a priori* est moins bonne que celle qui pourrait être attendue dans le cas d'une attribution aléatoire des observations aux groupes.

L'ARI moyen est calculé pour chaque méthode de classification sur l'ensemble des échantillons et par version d'échantillon, *i.e.* ARI moyen des échantillons contenant des signatures *brutes θ-égalisées* et ARI moyen des échantillons contenant des signatures *réalignées θ-égalisées*). La robustesse des algorithmes est évaluée à l'aide du Coefficient de Variation de l'ARI (écart-type de l'ARI divisé par sa moyenne). Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande. La méthode avec le plus faible coefficient de variation est donc considérée comme la plus

TABLE 3.1 – Indice de Rand Ajusté (ARI)

Méthode	**SdM**	**Moy. (E.T.)**	**CV**	**Min-Max**
CAH (QTS)	Brute	0.017 (0.104)	6.000	-0.038 - 0.715
	Réalignée	0.059 (0.109)	1.846	-0.036 - 0.475
CAH (fonc.)	Brute	0.012 (0.096)	8.062	-0.038 - 0.715
	Réalignée	0.075 (0.134)	1.779	-0.024 - 0.589
K-means (QTS)	Brute	0.042 (0.157)	3.703	-0.038 - 0.852
	Réalignée	0.264 (0.238)	0.903	-0.034 - 0.852
KMA (fonc.)	Brute	0.076 (0.202)	2.636	-0.038 - 0.715
	Réalignée	0.407 (0.295)	0.725	-0.038 - 1
K-medoids (QTS)	Brute	0.054 (0.183)	3.372	-0.038 - 0.852
	Réalignée	0.248 (0.241)	0.972	-0.036 - 0.715
KMedA (fonc.)	Brute	0.062 (0.184)	2.946	-0.038 - 0.715
	Réalignée	0.333 (0.285)	0.857	-0.024 - 1

robuste.

3.1.4.2 Résultats

Les SdMs *brutes* et *réalignées* sont calculées à partir des données de marche mesurées chez 27 volontaires en conditions de marche libre et contrainte. 60 échantillons sont générés d'après les règles définies dans la section 3.1.4.1.

Le tableau 3.1 présente la moyenne, l'écart type, le coefficient de variation (CV) et les valeurs minimales et maximales de l'ARI calculé à partir des résultats des méthodes de classification en fonction du type de SdM constituant les échantillons sur lesquels elles ont été appliquées. Ces résultats sont également présentés sur la figure 3.5. La mention (QTS) indique les méthodes de classification adaptées aux séries temporelles de quaternions unitaires. La mention (fonc.) indique les méthodes de classification adaptées aux données fonctionnelles de quaternions unitaires. "KMA" et "KMedA" correspondent respectivement aux abréviations de *K-means alignement* et *K-medoids alignement*.

Les valeurs d'ARI sont plus grandes pour les SdMs *alignées* que pour les SdMs *brutes*, et ce pour toutes les méthodes de classification. Cela tend à montrer que la représentation des SdMs en prenant l'orientation initiale comme référence est plus adaptée pour former des groupes d'individus en fonction de la présence d'un déficit de la marche mécanique. Les résultats montrent la supériorité des méthodes par partitionnement sur la classification hiérarchique. Cette observation est cependant à nuancer par le contexte de

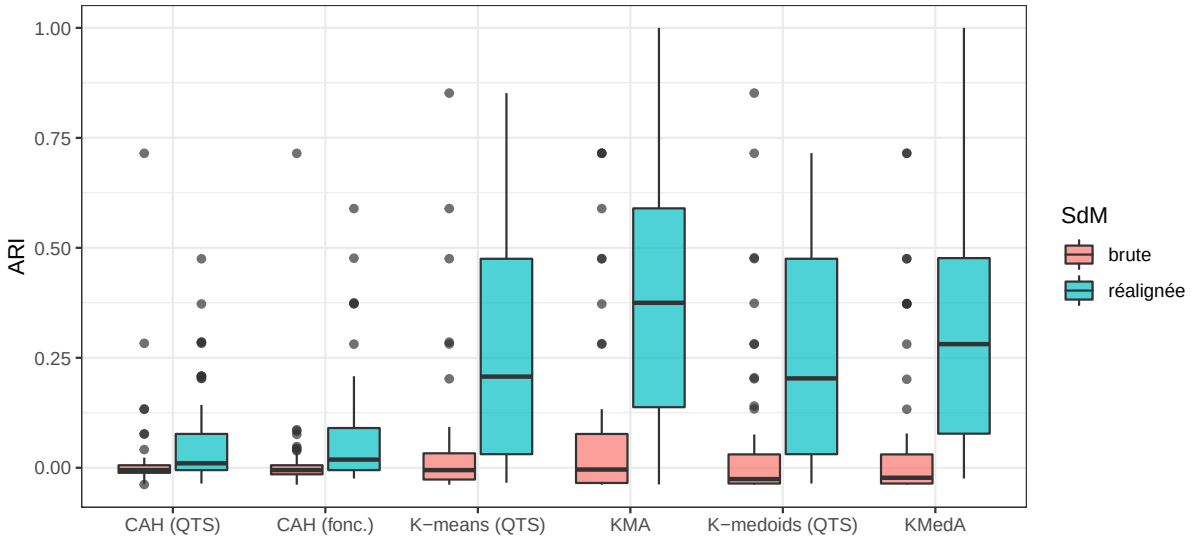


FIGURE 3.5 – ARI des méthodes de classification

l'étude. Aucune méthode pour déterminer le nombre de groupes à former n'a été utilisée, ce dernier étant connu à l'avance. Les méthodes de CAH sont plus appropriées dans un contexte exploratoire, et le nombre de groupes à former est souvent défini *a posteriori* à partir de la structure du dendrogramme obtenu. Dans le contexte de cette expérience, ces méthodes sont donc plus sensibles à la présence de données aberrantes, un des 2 groupes formés pouvant être constitué d'une seule SdM *outlier*. De plus, les échantillons ont été construits de telle sorte qu'ils soient constitués d'un nombre équilibré de SdMs mesurées dans l'une et l'autre des conditions de marche. Ceci constitue donc un autre paramètre favorisant les méthodes de *K-means* et *K-medoid* qui sont connues pour leur tendance à former des groupes de taille équilibrés. Pour finir, les méthodes présentant les meilleures performances selon l'ARI sont les méthodes de partitionnement sur données fonctionnelles, la meilleure étant en moyenne le *K-means alignment*. Elle est également la plus robuste et présente le coefficient de variation le plus faible. Les deux méthodes de partitionnement sur données fonctionnelles donnent des répartitions en 2 groupes parfaites sur certains échantillons, mais peuvent également produire des partitions aussi mauvaises que celles qu'il serait possible d'observer après une répartition au hasard des observations entre les groupes. Ceci peut s'expliquer par la présence éventuelle de SdMs de forme atypique, qui pourraient être considérées comme ayant des formes aberrantes.

3.1.4.3 Discussion

Dans cette section, plusieurs méthodes de classification couramment utilisées pour former des groupes de séries chronologiques et de données fonctionnelles ont été adaptées aux quaternions unitaires. Ces méthodes reposent sur l'utilisation de dissimilarités de forme et de calculs des prototypes de groupes permettant de tenir compte des problèmes de décalages entre les données. La comparaison des approches sur séries chronologiques et sur données fonctionnelles a permis d'identifier de meilleures performances pour les algorithmes de partitionnement sur données fonctionnelles dans le contexte de cette expérience. Cette étude, consistant à former des groupes d'individus à partir d'un biomarqueur représentant leur démarche par la variation de l'orientation de la hanche au cours du temps, permet d'identifier plusieurs points d'intérêt pour l'analyse de la marche par le dispositif MMR. Le premier concerne l'importance du choix de l'orientation de référence pour l'expression de la SdM. Les résultats montrent que définir comme référence l'orientation initiale du système de capteurs améliore fortement la capacité à former des groupes en fonction de la condition de marche par rapport à l'orientation moyenne. Le second est que la variabilité de l'amplitude des cycles de marche entre individus constitue une information parasite de laquelle il semble nécessaire de s'affranchir dans l'analyse des données de marche. Une transformation de centrage des SdMs sur l'angle de la moyenne point par point est proposée et permet d'améliorer les performances des méthodes de classification.

Il est important de noter que les résultats observés sont fortement liés au contexte de l'étude. Le fait que le nombre de groupes soit connu à l'avance et que les 2 groupes à former soient constitués d'un nombre d'observations équilibré joue en faveur des méthodes de partitionnement. Les algorithmes avec les meilleures performances présentent cependant des valeurs d'ARI relativement faibles, avec une moyenne de 0.407 pour le meilleur d'entre eux. Ceci peut s'expliquer par la taille de la population d'étude relativement restreinte. Une autre explication possible est la présence d'*outliers* dans les données. En effet, on constate que la qualité des partitions varie fortement en fonction des échantillons, les meilleurs méthodes de classification parvenant à former des partitions parfaites dans certains cas, mais génèrent des partitions avec un grand nombre d'erreurs dans d'autres. Enfin, le déficit de marche a été simulé par l'utilisation d'une orthèse, et est donc supposé déformer la marche de manière similaire entre les individus. Lors d'une utilisation dans la vie réelle, les atteintes de la marche peuvent avoir différentes causes d'origine neurologiques et/ou mécaniques, et un même patient peut en présenter plusieurs, ce qui complexifie nécessairement l'identification de groupes à former.

Plusieurs pistes d'améliorations peuvent être envisagées pour améliorer ces résultats. En premier lieu, l'augmentation de la taille de l'échantillon pourrait permettre de diminuer la sensibilité des méthodes à la présence d'*outliers*. Les méthodes de classification utilisées dans cette étude sont basées sur des mesures de dissimilarité de la forme globale des données, or on peut remarquer sur les figures que les `SdMs` semblent se regrouper par conditions de marche à certains endroits de la courbe. Ceci est visible sur la figure 3.2 pour la composante y des SdMs *brutes* et sur la composante z sur les SdMs réalignées. Aussi, une mesure de dissimilarité détectant les sous-domaines dans lesquels les données sont les plus similaires ou différentes pourrait permettre l'identification de patterns propres à un type de déficit de marche, telle que la méthode *Longest Common Sub Section* [175] pour les séries chronologiques. Enfin, d'autres types de transformation que celle proposée dans cette étude pourraient être développées pour s'affranchir de la variabilité inter-individuelle de l'amplitude des cycles, telles que la réduction de dimension par décomposition en valeurs singulières, l'Analyse par Composante Principale ou l'utilisation de bases de fonctions en analyse fonctionnelle.

3.1.5 Valorisation

Le calcul du biomarqueur *Signature De Marche* par *K-means alignment* à partir des cycles de marche identifiés par l'algorithme `STRIPAGE` fait l'objet d'une partie de la demande de brevet n° 21 00309 « Méthode et dispositif de détermination d'un cycle de marche », déposée le 13 janvier 2021, et dont la validation est toujours en cours d'évaluation par l'Institut National de la Propriété Industrielle à date de rédaction de ce manuscrit.

Des résultats intermédiaires de l'étude de comparaison des méthodes de classification de séquences de quaternions unitaires ont fait le sujet d'une présentation orale à la 12^{th} *International Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methological Statistics (CMCStatistics 2019)*, le 15 décembre 2019 à la *Senate House University of London*, Angleterre, et d'une présentation sous forme de poster en ligne à la $14^{ème}$ Conférence Francophone d'EPIdémiologie CLINique et $27^{ème}$ journées des Statisticiens des Centre de Lutte Contre le Cancer, les 15 et 16 septembre 2020.

3.2 Classification semi-supervisée et analyse de la marche dans la SEP

Cette section présente l'adaptation de méthodes semi-supervisées pour la classification de séries temporelles de quaternions unitaires et leur application pour l'analyse de la marche de patients atteints de Sclérose En Plaques. Nous avons vu dans la section 2.2.2 que les patients étaient très fréquemment sujets à l'apparition de troubles de la marche et que ces symptômes étaient considérés comme les plus handicapants par ces derniers [100], donnant une place centrale à l'évaluation de la marche dans le suivi de cette pathologie [114].

L'objectif est ici de former des groupes de patients présentant des signatures de marche similaires. Au cours de son suivi, un patient atteint de SEP est régulièrement examiné à l'hôpital par un clinicien qui évalue ses fonctions neurologiques pour déterminer la sévérité de la pathologie. L'échelle la plus utilisée dans le cas de SEP est l'*Expendend Disability Status Scale* (EDSS) [111].

L'Expanded Disability Status Scale ou EDSS [97] est l'échelle la plus largement utilisée pour évaluer l'étendue du handicap global des patients diagnostiqués avec la SEP [111]. L'EDSS est un système d'évaluation allant de 0 (état neurologique normal) à 10 (décès dû à la SEP) [111]. Les scores EDSS inférieurs à 4 sont attribués aux patients présentant des troubles neurologiques légers, les scores EDSS modérés (de 4 à $6,5$) sont attribués aux patients présentant des troubles sévères de la marche et les scores EDSS élevés (supérieurs à 7) sont attribués aux patients qui ne peuvent plus marcher sans aide bilatérale (cannes, béquilles ou appareils orthopédiques). L'EDSS est critiquée pour son manque de linéarité, une augmentation du score de 1 point étant associée à une augmentation de la sévérité inégale entre les différents degrés de l'échelle [179]. De plus, l'altération de la marche n'est qu'une des informations prises en compte lors de l'attribution du score EDSS, qui vise à donner une appréciation générale du handicap global du patient. Ainsi, deux patients ayant un score EDSS identique peuvent néanmoins présenter des troubles de marche différents.

Malgré ses imperfections, l'échelle EDSS constitue une source d'informations supplémentaire dont la prise en compte dans une classification par approche semi-supervisée (section 1.3.3) pourrait permettre de contrebalancer la variabilité des données de marche et ainsi améliorer l'interprétabilité des résultats. Aucune des méthodes existantes dans la littérature n'est cependant adaptée à la classification semi-supervisée de séries chro-

nologiques de quaternions unitaires. La section 3.2.1 présente donc l'adaptation de deux méthodes, l'une étant une approche par *compromis* et la seconde par *ensemble de classifications*. Ces deux méthodes sont ensuite comparées en les appliquant à la base de données `BDDsep` dans la section 3.2.2.

3.2.1 Méthodes

La sélection des méthodes à généraliser aux séries chronologiques de quaternions unitaires est basée sur plus plusieurs caractéristiques liées au contexte de l'application visée. Tout d'abord, le nombre de groupes à former n'est pas connu *a priori*. Il n'y a en effet pas d'étude antérieure décrivant l'analyse de la marche par représentation de la rotation de la hanche au cours du temps sous forme de quaternions unitaires. Les approches de classifications hiérarchiques semblent plus appropriées dans ce cadre, le dendrogramme qui en résulte permettant d'explorer les regroupements à différents niveaux de sa structure.

De plus, les données supplémentaires disponibles se prêtent peu à la détermination de contraintes telles que celles utilisées dans les méthodes de *classification avec contraintes* [39] (*c.f.* présentation des types de contraintes section 1.3.3.1). En effet, deux patients ayant le même score EDSS peuvent présenter des troubles de la marche différents. Définir des paires de patients devant appartenir au même groupe en fonction de leur score EDSS serait donc hasardeux. D'autre part, définir des paires de patients devant appartenir à des groupes différents nécessiterait de déterminer une règle arbitraire telle qu'un écart maximal toléré entre scores EDSS au sein d'un même groupe.

D'après ces considérations, les méthodes de *classification avec contraintes* sont considérées comme peu adaptées au cadre d'application. Les deux prochaines sections présentent donc une méthode de *classification par compromis* et une méthode par *ensemble de classifications* ainsi que leur adaptation aux séries temporelles de quaternions unitaires par l'utilisation d'une mesure de dissimilarité adaptée.

3.2.1.1 Méthode de classification par compromis : `hclustcompro`

La méthode `hclustcompro` est décrite par Bellanger et al. en 2021 [16] dans le cadre d'une application à des données archéologiques et géographiques.

Soit un jeu de données constitué de n observations représentées par deux types d'information différents, appelés respectivement information principale et information supplémentaire. La première étape consiste à déterminer D_1 et D_0, deux matrices de dimension

$n \times n$ représentant les dissimilarités normalisées entre les paires d'observation du jeu de données, respectivement calculées à partir de l'information principale et l'information supplémentaire. Le principe de la méthode `hclustcompro` consiste à appliquer un algorithme de CAH à partir de la combinaison convexe suivante [16] :

$$D_\alpha = \alpha D_1 + (1 - \alpha) D_0 \tag{3.17}$$

où $\alpha \in [0; 1]$ est un paramètre de pondération fixé, dont la valeur doit être déterminé. La méthode résulte en un dendrogramme $\mathcal{T}_\alpha$ dont la structure est dépendante de la valeur du paramètre α et de la stratégie d'agrégation définie par le critère de liaison utilisé dans la CAH.

La détermination de la valeur du paramètre α dépend d'un critère objectif qui se base sur le principe de la corrélation cophénétique proposée par Sokal *et al.*[158]. Tout d'abord, la matrice cophénétique associée au dendrogramme $\mathcal{T}_\alpha$ est calculée. Chaque élément de cette matrice correspond à la hauteur de $\mathcal{T}_\alpha$ à partir de laquelle deux observations appartiennent au même groupe. Selon le principe décrit par Sokal et al., la corrélation cophénétique est calculée comme la corrélation entre la matrice de dissimilarité initiale et la matrice cophénétique calculée à partir de $\mathcal{T}_\alpha$. Sa valeur quantifie avec quelle fidélité les dissimilarités entre paires d'objets observées dans le jeu de données initial sont représentées dans le dendrogramme.

En se basant sur ce principe, le critère de sélection de α est défini par l'équation suivante :

$$CorCrit_\alpha = |Cor(D_\alpha^{Coph}, D_1) - Cor(D_\alpha^{Coph}, D_0)| \tag{3.18}$$

où D_α^{Coph} est la matrice cophénétique de $\mathcal{T}_\alpha$, le dendrogramme obtenu par CAH appliquée sur D_α, pour une valeur de α fixée in 3.17. Le critère $CorCrit_\alpha$ correspond à une différence absolue entre deux termes, chacun correspondant à la corrélation quantifiant à quel point la structure de $\mathcal{T}_\alpha$ représente fidèlement les informations apportées par les matrices D_1 et D_0. Pour pondérer l'importance de D_1 et D_0 dans la classification, la valeur du coefficient α est déterminée par l'optimisation de la fonction suivante :

$$\hat{\alpha} = \arg\min_\alpha CorCrit_\alpha, \tag{3.19}$$

$\hat{\alpha}$ correspond à la valeur de α permettant de construire un dendrogramme dont la structure hiérarchique correspond au meilleur compromis entre les dissimilarités observées dans D_1 and D_0.

3.2.1.2 Méthode par ensemble de classifications : `mergeTrees`

La méthode `mergeTrees` est une méthode de *classification coopérative* décrite par Hulot *et al.* en 2020 [75] dans le cadre de l'analyse de données omics.

Soient m jeux de données observés sur n individus, et $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m\}$ l'ensemble des m dendrogrammes obtenus à partir de ces jeux de données par CAH. La méthode `mergeTrees` consiste en la construction d'un dendrogramme dit "consensus", noté $C(\boldsymbol{\mathcal{T}})$ et respectant la propriété suivante : *Pour toute paire d'individus i et j, $i \neq j$, si i et j appartiennent à des groupes différents dans au moins l'un des arbres de l'ensemble $\boldsymbol{\mathcal{T}}$ à une hauteur h, alors ils appartiennent à des groupes différents dans l'arbre consensus $C(\boldsymbol{\mathcal{T}})$ à la hauteur h.* Formulé de manière différente, la hauteur d'une branche reliant deux observations dans l'arbre consensus $C(\boldsymbol{\mathcal{T}})$ sera égale à la hauteur maximale de la branche reliant ces deux observations dans l'ensemble des dendrogrammes. La figure 3.6 présente un exemple de l'agrégation de 2 dendrogrammes par l'algorithme `mergeTrees`. Les traits pointillés bleus matérialisent la hauteur maximale de la branche reliant deux observations dans les deux arbres $\mathcal{T}_1$ et $\mathcal{T}_2$. Dans l'arbre $\mathcal{T}_1$, la hauteur de la branche reliant l'observation 4 et l'observation 3 vaut 7, et la hauteur de cette branche vaut 2 dans l'arbre $\mathcal{T}_2$. La branche reliant ces deux observations dans $C(\mathcal{T}_1, \mathcal{T}_2)$ vaut donc 7.

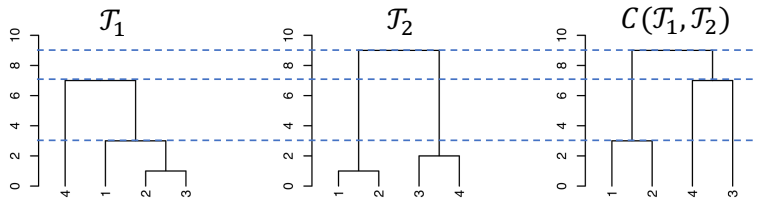


FIGURE 3.6 – Principe de l'agrégation de 2 dendrogrammes par `mergeTrees`. Design de la figure inspiré de Hulot *et al.* (2020) [75], à partir des données l'exemple transmis dans la documentation de la fonction `mergeTrees` du package R `mergeTrees`.

L'algorithme est construit de telle sorte que $C(\boldsymbol{\mathcal{T}})$ respecte les règles d'*Anonymat*, de *Neutralité* et d'*Unanimité*. L'*Anonymat* implique que le résultat ne dépend pas de l'ordre des dendrogrammes dans l'ensemble $\boldsymbol{\mathcal{T}}$. La *Neutralité* implique qu'un changement des labels associés aux observations des arbres de l'ensemble $\boldsymbol{\mathcal{T}}$ se traduit par un simple changement de ces mêmes labels dans $C(\boldsymbol{\mathcal{T}})$. Pour finir, l'*Unanimité* implique que la méthode appliquée à un ensemble constitué d'une répétition d'un même arbre $\mathcal{T}$ résulte à la construction de l'arbre consensus $C(\boldsymbol{\mathcal{T}}) = \mathcal{T}$.

La hauteur des branches doit être comparable entre les arbres de l'ensemble. En effet,

si toutes les branches d'un arbre $\mathcal{T}_\alpha$ présentent une hauteur supérieure à toute autre branche de n'importe quel arbre de l'ensemble, l'arbre consensus sera alors $C(\boldsymbol{\mathcal{T}}) = \mathcal{T}_\alpha$. Une étape de pré-traitement pour normaliser les dissimilarités mesurées sur les différents jeux de données peut donc être nécessaire.

Enfin, la méthode `mergeTrees` peut résulter en un arbre non-binaire, *i.e.* un arbre dans lequel plus de deux branches sont créées à une même hauteur, par exemple dans les cas où plusieurs branches de dendrogrammes différents sont de même hauteur. Cela peut par exemple se traduire par la présence de trois branches reliées au même noeud dans l'arbre consensus $C(\boldsymbol{\mathcal{T}})$.

3.2.1.3 Adaptation aux séries temporelles de quaternions

Les méthodes `hclustcompro` et `mergeTrees` sont basées sur l'algorithme de la CAH, qui peut être appliqué sur une matrice de dissimilarité. On considère un cadre d'application dans lequel les données sont représentées par deux sources d'information, la principale étant constituée de séries temporelles de quaternions unitaires, la source d'information supplémentaire pouvant correspondre à tout type de données pour lequel une mesure de dissimilarité existe. L'application des méthodes `hclustcompro` et `mergeTrees` passe par le calcul de la matrice de dissimilarité D_1 entre séries chronologiques de quaternions par QDTW, et par la matrice de dissimilarité D_0 entre les données de la source d'information supplémentaire.

Ces deux matrices sont introduites dans l'équation 3.17 pour appliquer la méthode `hclustcompro`. D'autre part, l'ensemble des dendrogrammes $(\mathcal{T}_1, \mathcal{T}_0)$ sur lequel appliquer `mergeTrees` est généré par CAH appliquée respectivement sur D_1 et D_0.

Se pose alors la question du choix du critère de liaison déterminant la stratégie d'agrégation de la CAH. La formule de récurrence de Lance et Williams implique que les critères de liaison *géométriques* ne sont adaptés qu'aux cas pour lesquels les données sont exprimées dans un espace euclidien et la distance euclidienne est utilisée comme mesure de dissimilarité. Cette propriété n'étant pas respectée pour les séries chronologiques de quaternions unitaires, les seuls critères de liaison utilisables sont ceux appartenant à la classe des graphes, soient liaisons *simple*, *moyenne* et *complète*.

En conclusion, les méthodes `hclustcompro` et `mergeTrees`, représentant une extension de la CAH à la classification semi-supervisée, sont adaptables au contexte de l'analyse de la base de données `BDDsep`. Leur adaptation se base sur l'utilisation de l'algorithme QDTW adapté aux séries temporelles de quaternions unitaires pour générer une matrice de

dissimilarité. Du fait des propriétés des séries chronologiques de quaternions unitaires et de l'algorithme QDTW, les critères de liaisons simple, moyenne et complète représentent les stratégies d'agrégation possibles pour la classification par `hclustcompro` et `mergeTrees`.

3.2.2 Application à la base SEP

Cette section présente l'application des méthodes `CAH`, `hclustcompro` et `mergeTrees` aux données de la base SEP. Les objectifs de cette étude sont (i) d'évaluer l'intérêt de l'ajout d'informations supplémentaires dans la classification, (ii) de comparer les résultats obtenus par les deux méthodes et les différentes stratégies d'agrégation possibles afin d'identifier l'approche la plus adaptée. L'évaluation des méthodes se base sur la pertinence clinique des classifications obtenues. La section 3.2.2.1 décrit le jeu de données de cette étude, la section 3.2.2.2 détaille la méthodologie employée et la section présente enfin les résultats 3.2.2.3.

3.2.2.1 Présentation des données

Cette étude est réalisée à partir des données de la base `BDDsep` (décrite dans la section 1.2.3.2). Chaque patient est représenté par deux sources d'information :

— Son score EDSS (*c.f.* description de l'échelle EDSS section 1.1.1.1).

— Les cycles de marche détectés par l'algorithme `STRIPAGE` (présenté dans la section 2.1.1) à partir des données mesurées par le système de capteurs MMR durant le test de marche T25FW.

Prétraitement des données et calcul de SdM des patients Durant la préparation de l'étude clinique, l'impact de la position du système de capteurs au niveau de la ceinture du porteur sur la forme des données recueillies n'était pas connu, ou sous-estimé. Les consignes transmises à l'équipe Neurologie du Centre d'Investigation Clinique (CIC) du CHU de Nantes en charge de recueillir les données de marche indiquaient donc de placer le dispositif MMR sur *la ceinture du patient, en position latérale droite*. Lors d'une réunion intermédiaire avec l'équipe du CIC, il s'est avéré que cette consigne n'était pas assez précise. Du fait de ce manque de précision, le dispositif a pu être placé à une position variable entre différents patients (*e.g.* légèrement vers l'avant ou vers l'arrière de la position latérale droite). Cette différence de placement du dispositif s'est traduite par une source de variabilité supplémentaire dans les données mesurées. Cette variabilité n'étant pas due

aux potentiels troubles de la marche des patients, on souhaite s'en affranchir autant que possible.

En se basant sur la règle ♯5 énoncée dans la section 2.1.1 *"Les orientations observées au début de la phase d'appui (resp. au début de la phase de balancement) sont relativement similaires entre les cycles de marche détectés chez des individus différents"* et sur le principe énoncé dans la section 2.1.1.5, on suppose que les axes de la rotation entre les orientations observées au début de la phase d'appui et au début de la phase de balancement sont similaires entre les individus. Pour limiter l'impact de la variabilité de positionnement du système de capteur, les SdMs des patients sont transformées de telle sorte que l'axe de la rotation entre l'orientation initiale (*i.e.* observée au début de la phase d'appui) et l'orientation observée à fin de la phase d'appui soient similaire entre toutes les SdMs. Pour ce faire, avant le calcule de la SdM, les cycles de marche de chaque patients ont été transformés par la méthode suivante.

On note :

— $\{Q_m = (\mathbf{q}_{m,1}, \ldots, \mathbf{q}_{m,N_m})\}_{m=1,2,\ldots,M}$: les M cycles de marche d'un patient détectés par l'algorithme `STRIPAGE`.

— $\{Q_m^{(1)} = (\mathbf{q}_{m,1}^{(1)}, \ldots, \mathbf{q}_{m,N_m}^{(1)})\}_{m=1,2,\ldots,M}$: les M cycles transformés de telle sorte que l'orientation initiale soit considérée comme l'orientation de référence, avec $\mathbf{q}_{m,j}^{(1)} = \mathbf{q}_{m,1}^{-1}\mathbf{q}_{m,j}, \forall j \in \{1, \ldots, N_m\}$.

— $U_m = (\mathbf{u}_{m,1}, \ldots, \mathbf{u}_{m,N_m})$: L'ensemble des axes des rotations de $Q^{(1)}$, où $\forall j \in \{1, \ldots, N_m\}$, $\mathbf{u}_{m,j}$ est déterminé à partir de $\mathbf{q}_{m,j}^{(1)}$ par l'équation (1.15).

— $\theta_m = (\theta_{m,1}, \ldots, \theta_{m,N_m})$: L'ensemble des angles des rotations de $Q^{(1)}$, où $\forall j \in \{1, \ldots, N_m\}$, $\theta_{m,j}$ est déterminé à partir de $\mathbf{q}_{m,j}^{(1)}$ par l'équation (1.14).

— $\mathbf{j}^* = (j_1^*, \ldots, j_m^*)$, où j_m^* est l'indice correspondant à la transition entre la phase de balancement et la phase d'appui dans le cycle m (*c.f.* 2.2.1).

— $U^* = \{\mathbf{u}_1^*, \ldots, \mathbf{u}_M^*\}$, où $\mathbf{u}_m^* = \mathbf{u}_{m,j^*}$ est l'axe de la rotation entre le début de la phase d'appui et le début de la phase de balancement du segment m.

L'objectif est de transformer les axes de rotation $\mathbf{u}_m^* \in U^*$ de telle sorte qu'ils aient les coordonnées d'un axe de *rotation d'appui de référence*. Pour ce faire, on considère un axe de rotation $\mathbf{u}_m^* \in U^*$ et l'axe de référence $\mathbf{u}_A$, estimé à partir de la `BDDapp` par la méthode présentée dans la section 2.1.1.5. L'objectif est de calculer la rotation la plus directe permettant de tourner $\mathbf{u}_m^*$ afin d'obtenir $\mathbf{u}_A$, *i.e.* le quaternion unitaire $\mathbf{q}_m^{(r)}$ telle

que :

$$\mathbf{q}_m^{(r)} \begin{pmatrix} 0 \\ \mathbf{u}_\mathbf{m}^* \end{pmatrix} \left(\mathbf{q}_m^{(r)}\right)^{-1} = \begin{pmatrix} 0 \\ \mathbf{u}_A \end{pmatrix}. \tag{3.20}$$

L'axe de la rotation $\mathbf{q}_m^{(r)}$, noté $\mathbf{v}_m^{(r)}$ est simplement perpendiculaire aux deux axes $\mathbf{u}_m^*$ et $\mathbf{u}_A$, il est donc estimé par :

$$\mathbf{v}_m^{(r)} = \frac{\mathbf{u}_m^* \times \mathbf{u}_A}{\|\mathbf{u}_m^* \times \mathbf{u}_A\|}, \tag{3.21}$$

où $\times$ est le *produit vectoriel* entre deux vecteurs. L'angle de la rotation $\mathbf{q}_m^{(r)}$, noté $\theta_m^{(r)}$, est alors calculé

$$\theta_m^{(r)} = \arctan\left(\frac{\|\mathbf{u}_m^* \times \mathbf{u}_A\|}{\mathbf{u}_m^* \cdot \mathbf{u}_A}\right), \tag{3.22}$$

où $\cdot$ est le *produit scalaire* entre deux vecteurs. Le quaternion $\mathbf{q}_m^{(r)}$ est calculé à partir de $\mathbf{v}_m^{(r)}$ et $\theta_m^{(r)}$ par l'équation (1.13).

La rotation $\mathbf{q}_m^{(r)}$ est appliquée à l'ensemble des axes U_m pour obtenir l'ensemble des axes réorientés $U_m^{(r)} = (\mathbf{u}_{m,1}^{(r)}, \ldots, \mathbf{u}_{m,N_m}^{(r)})$, où :

$$\begin{pmatrix} 0 \\ \mathbf{u}_{m,j}^{(r)} \end{pmatrix} = \mathbf{q}_m^{(r)} \begin{pmatrix} 0 \\ \mathbf{u}_{m,j} \end{pmatrix} \left(\mathbf{q}_m^{(r)}\right)^{-1} \tag{3.23}$$

Les rotations du cycle de marche $Q_m^{(1)}$ sont recalculées à partir de $U_m^{(r)}$ et de θ_m, de telle sorte que

$$\mathbf{q}_{m,j}^{(1)} \leftarrow \cos\frac{\theta_{m,j}}{2} + \mathbf{u}_{m,j}^{(r)} \sin\frac{\theta_{m,j}}{2}, \forall j \in \{1, \ldots, N_m\} \text{ et } \forall m \in \{1, \ldots, M\} \tag{3.24}$$

Les cycles $Q_m^{(1)}$ ainsi obtenus sont ensuite réorientés sur leur quaternions moyens par la méthode décrite dans la section 2.1.1.6, **Étape** 5.2, pour obtenir les *cycles de marche transformés* Q_m. L'ensemble de cette transformation est appelée *égalisation sur* $\mathbf{u}_A$.

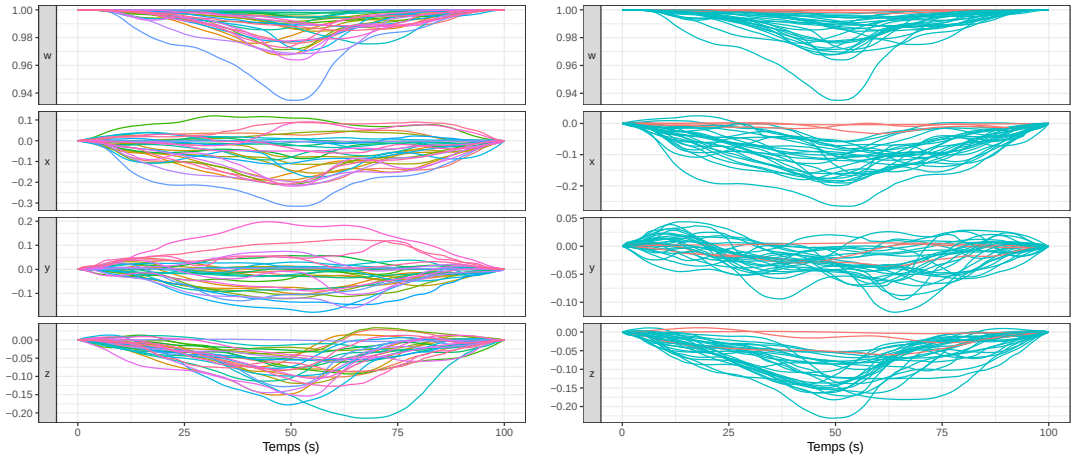
La *Signature de Marche* (SdM) de chaque patient est calculée à partir des *cycles de marche* $\mathbf{u}_A$*-égalisés*, puis *réalignée* par l'algorithme 7 `StraightenQTS` selon la méthode décrite dans la section 3.1.4.1. La SdM *réalignée* d'un patient i est notée

$$Q_i = \left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,101}^\top\right),$$

et chaque élément $\mathbf{q}_{i,j}, \forall j \in \{0, \ldots, 100\}$ représente l'orientation de la hanche au $(j-1)^{ème}$ pourcentage de la durée totale d'un cycle. On rappelle également que $\mathbf{q}_{i,j} = (1, 0, 0, 0), \forall i$

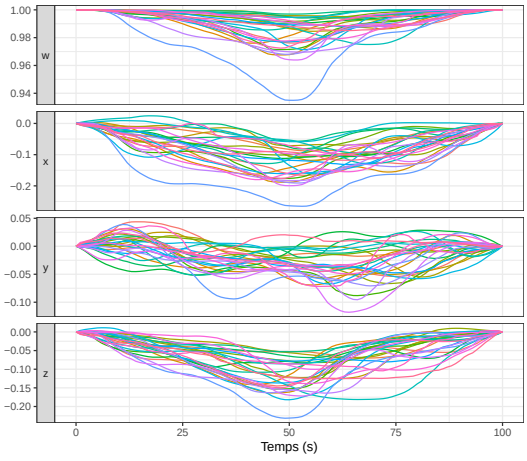
et $\forall j \in \{1, 100\}$.

Les figures 3.7a et 3.7b présentent les SdMs des 30 patients calculées respectivement à partir des cycles de marche *non centrés* et *centrés sur* $\mathbf{u}_A$. Malgré cette transformation, 3 présentent des SdMs dont la forme est jugée aberrante. Ils sont représentés en rouge sur la figure 3.7b. Ils sont donc exclus du reste de l'analyse. La figure 3.7c présente les SdMs *centrées sur* $\mathbf{u}_A$ des 27 patients sur lesquels les méthodes de classification seront appliquées.



(a) SdMs sans transformation (30 patients)

(b) SdMs transformées avec *outliers*



(c) SdMs transformées sans outliers (27 patients)

FIGURE 3.7 – Transformation des SdMs et outliers

La section suivante présente la méthodologie employée pour appliquer les différentes méthodes de classification sur ces données et évaluer leurs résultats.

3.2.2.2 Design d'expérience

Calcul des dendrogrammes Tout d'abord, deux matrices de dissimilarité sont calculées à partir des deux sources d'information relatives aux patients :

— La matrice D_1 contient les dissimilarités normalisées entre les SdMs des patients calculées par QDTW.

— La matrice D_0 contient les dissimilarités normalisées entre les scores EDSS. Cette échelle étant de nature quantitative ordinale, la dissimilarité entre paires d'EDSS est déterminée par la méthode de Gower [55].

Une `CAH` non supervisée est appliquée sur D_1 et D_0 pour obtenir respectivement les dendrogrammes $\mathcal{T}_1$ et $\mathcal{T}_0$. `hclustcompro` est appliqué sur D_1 et D_0. En suivant les notations présentées dans 3.2.1.1, la matrice de dissimilarité pondérée D_α est calculée selon 3.17 à partir de D_1 et D_0. Le dendrogramme résultant est noté $\mathcal{T}_\alpha$. `mergeTrees` est appliquée sur les deux dendrogrammes $\mathcal{T}_1$ et $\mathcal{T}_0$ et aboutit à l'arbre consensus $\mathcal{C}(\mathcal{T}_1, \mathcal{T}_0)$. Toutes ces méthodes sont appliquées avec les critères de liaisons simple, moyenne et complète.

Sélection du nombre de groupes Les dendrogrammes résultant des méthodes employées fournissent une hiérarchie de regroupements possibles pour les données. La détermination du nombre de groupes à former à partir de cette structure est une problématique complexe et encore ouverte. Il existe de nombreux critères dans la littérature sur lesquels s'appuyer pour réaliser cette tâche [27]. Ici le **critère du coude** déterminé en fonction de l'erreur quadratique intra-groupe notée WSS (pour *Within cluster Sum of Square error*) et la **largeur moyenne des silhouettes** sont généralisés aux QTS.

La méthode du coude consiste à tracer la WSS en fonction du nombre de clusters. Il s'agit d'une fonction décroissante qui atteint 0 lorsque le nombre de clusters correspond à la taille de l'échantillon. Ce tracé permet d'identifier le nombre optimal de groupes pour la valeur correspondant à une rupture de la pente. En d'autres termes, la WSS décroît généralement plus rapidement au début et ralentit à partir d'une valeur donnée. Définir le nombre optimal de groupes consiste à trouver ce point de transition. La WSS est une mesure de cohésion correspondant à la somme des carrés de dissimilarité entre les observations et le prototype de leur cluster. Le prototype d'un groupe TS est connu sous le nom de séquence de Steiner [59]. Il s'agit des TS théoriques qui minimisent la dissimilarité avec les TS qui forment un cluster. La méthode pour déterminer le prototype dépend donc de la mesure de dissimilarité choisie dans la méthode de classification. Dans le cas de la

mesure élastique, sa détermination doit tenir compte de l'alignement dans le temps, ce qui n'est pas une tâche triviale. Une approche possible est de donner une approximation du prototype par la médoïde du groupe. Cette méthode de représentation est fréquemment utilisée dans la pratique [4]. La médoïde est la TS la plus centrale du groupe, c'est-à-dire la TS observée qui a la plus petite somme de distance au carré avec les autres observations du cluster. Soit une partition d'un ensemble de QTS en K groupes $C_k, k \in \{1, \ldots, K\}$, et I_k définissant l'ensemble des QTS du groupe C_k : $I_k = \{i \in \{1, \ldots, n\}/Q_i \in C_k\}$, la WSS est calculée d'après l'équation suivante :

$$\text{WSS} = \sum_{k=1}^{K} \sum_{i \in I_k} \text{QDTW}(Q_i, \tilde{Q}_k)^2, \tag{3.25}$$

avec $\tilde{Q}_k$ le médoïde du groupe C_k, tel que décrit :

$$\tilde{Q}_k = \underset{Q_i, i \in I_k}{\arg\min} \sum_{j \in I_k, j \neq i} \text{QDTW}\left(Q_i, Q_j\right)^2 \tag{3.26}$$

La largeur de la silhouette est une mesure individuelle allant de -1 à $+1$ qui dépend du coefficient de séparation B et du coefficient de cohésion A. Le coefficient de cohésion d'une observation i est noté $A(i)$ et est défini comme la moyenne des dissimilarités de Q_i avec les autres observations de son groupe :

$$A(i) = \frac{1}{\mid I_k \mid -1} \sum_{j \in I_k, j \neq i} \text{QDTW}(Q_i, Q_j) \tag{3.27}$$

Le coefficient de séparation associé à l'observation i est noté $B(i)$ et représente la plus petite dissimilarité moyenne entre Q_i et les observations d'un autre groupe :

$$B(i) = \min_{k' \neq k} \frac{1}{\mid I_{k'} \mid} \sum_{i' \in I_{k'}} \text{QDTW}(Q_i, Q_{i'}) \tag{3.28}$$

La largeur de silhouette $s(i)$ associée à l'observation Q_i est alors calculée :

$$s(i) = \frac{B(i) - A(i)}{\max\left(A(i), B(i)\right)} \tag{3.29}$$

Une silhouette proche de 1 indique une bonne assignation de l'observation considérée. Enfin, la largeur de silhouette moyenne notée ASW (pour *Average Silhouette Width*) d'une partition est calculée comme la moyenne des largeurs de silhouette $s(i)$ associée à

chaque observation :

$$\text{ASW} = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{3.30}$$

Le choix du nombre de groupes à former peut être guidé par la recherche de la valeur de K la plus petite possible et maximisant l'ASW.

Des partitions sont formées à partir des dendrogrammes obtenus par `CAH` (appliquée sur D_1), `hclustcompro` et `mergeTrees` en choisissant un nombre de groupes K compris entre 1 et 10. L'ASW et la WSS sont calculées pour chaque méthode en fonction de K selon les équations 3.30 et 3.25 respectivement. Le choix du nombre de clusters se base sur ces deux critères et sur l'expertise d'un neurologue qui examine la pertinence clinique des partitions obtenues. Une attention est également portée sur le fait que le nombre d'observations par groupes ne soit pas trop faible et ainsi éviter au maximum les *singleton*.

Critères d'évaluation. Les partitions obtenues sont évaluées sur la base de critères internes et externes.

Critères internes. L'*inertie intra-groupe* $I_W^{(k)}$ de chaque groupe C_k est définie par :

$$I_W^{(k)} = \frac{1}{|V_k|}\sum_{i\in V_k} \text{QDTW}^2(Q_i, \widetilde{Q}_k).$$

Sa valeur doit être faible en comparaison de l' *inertie inter-groupe* $I_B^{(k)}$ définie comme :

$$I_B^{(k)} = \text{QDTW}^2\left(\widetilde{Q}_k, \widetilde{Q}\right),$$

où $\widetilde{Q}$ est le médoïde du jeu de données complet déterminé par :

$$\widetilde{Q} = Q_i, \quad \text{where} \quad i = \underset{i\in[\![1,n]\!]}{\arg\min} \sum_{j\in[\![1,n]\!], j\neq i} \text{QDTW}^2\left(Q_i, Q_j\right).$$

On peut déterminer cette condition par la *proportion d'inertie intra-groupe* définie par :

$$p_W^{(k)} = \frac{I_W^{(k)}}{I_W^{(k)} + I_B^{(k)}}. \tag{3.31}$$

Cette proportion peut être déterminée pour chaque groupe et représente l'éloignement entre les observations et le médoïde de leur groupe par rapport à l'éloignement des groupes par rapport au médoïde du jeu de données complet. Ce critère doit être

aussi petit que possible, ce qui est observé quand les observations d'un même groupe sont très proches et que les groupes sont fortement séparés les uns des autres.

Cette même proportion peut être calculée pour l'ensemble de la partition en définissant :

$$I_W = \frac{\sum_{k=1}^{K} |V_k| I_W^{(k)}}{\sum_{k=1}^{K} |V_k|}, \quad I_B = \frac{\sum_{k=1}^{K} |V_k| I_B^{(k)}}{\sum_{k=1}^{K} |V_k|} \quad \text{and} \quad p_W = \frac{I_W}{I_W + I_B}. \tag{3.32}$$

L' *indice de Dunn* est également utilisé pour évaluer à quel point les groupes sont compacts et séparés les uns des autres. Il est définit par :

$$\mathrm{DI} = \frac{\min_{k,\ell \in [\![1,K]\!]^2, \ell \neq k} \delta\left(C_k, C_\ell\right)}{\max_{k \in [\![1,K]\!]} \Delta(C_k)}, \tag{3.33}$$

avec

$$\delta\left(C_k, C_\ell\right) = \min_{i \in V_k, j \in V_\ell} \mathrm{QDTW}\left(Q_i, Q_j\right) \text{ (séparation).}$$

et

$$\Delta\left(C_k\right) = \max_{i,j \in V_k} \mathrm{QDTW}\left(Q_i, Q_j\right) \text{ (cohésion).}$$

Cet indice prend une valeur élevée quand les groupes sont fortement compacts et séparés les uns des autres [62].

Les deux critères présentés précédemment évaluent la qualité de la partition uniquement à partir des SdMs. Un critère d'évaluation interne est également utilisé pour évaluer la qualité de la partition à partir des scores EDSS. Cette évaluation est réalisée par l'observation de la distribution des scores EDSS par groupe. Elle représente à quel point les groupes rassemblent des patients présentant une sévérité globale de leur pathologie comparable.

Critères externes. Les deux critères externes suivants sont utilisés pour compléter l'évaluation de la qualité des partitions :

- La distribution intra-groupe des temps observés au test T25FW. Cette durée dépend de la vitesse des patients et est donc impactée par leur déficit de la marche.
- L'évaluation clinique de la structure des dendrogrammes par 5 neurologues experts de la Sclérose En Plaques. Une présentation du contexte de l'étude et une description brève des méthodes de classification utilisées leur ont été proposées. Il leur a été ensuite demandé d'évaluer en aveugle les dendrogrammes

et partitions obtenus par chacune des méthodes, à partir de leurs connaissances de la pathologie et des informations cliniques des patients.

La méthodologie complète du calcul et de la comparaison des dendrogrammes résultants des trois méthodes de classification `CAH`, `mergeTrees` et *hclustcompro* est présentée dans la figure 3.8.

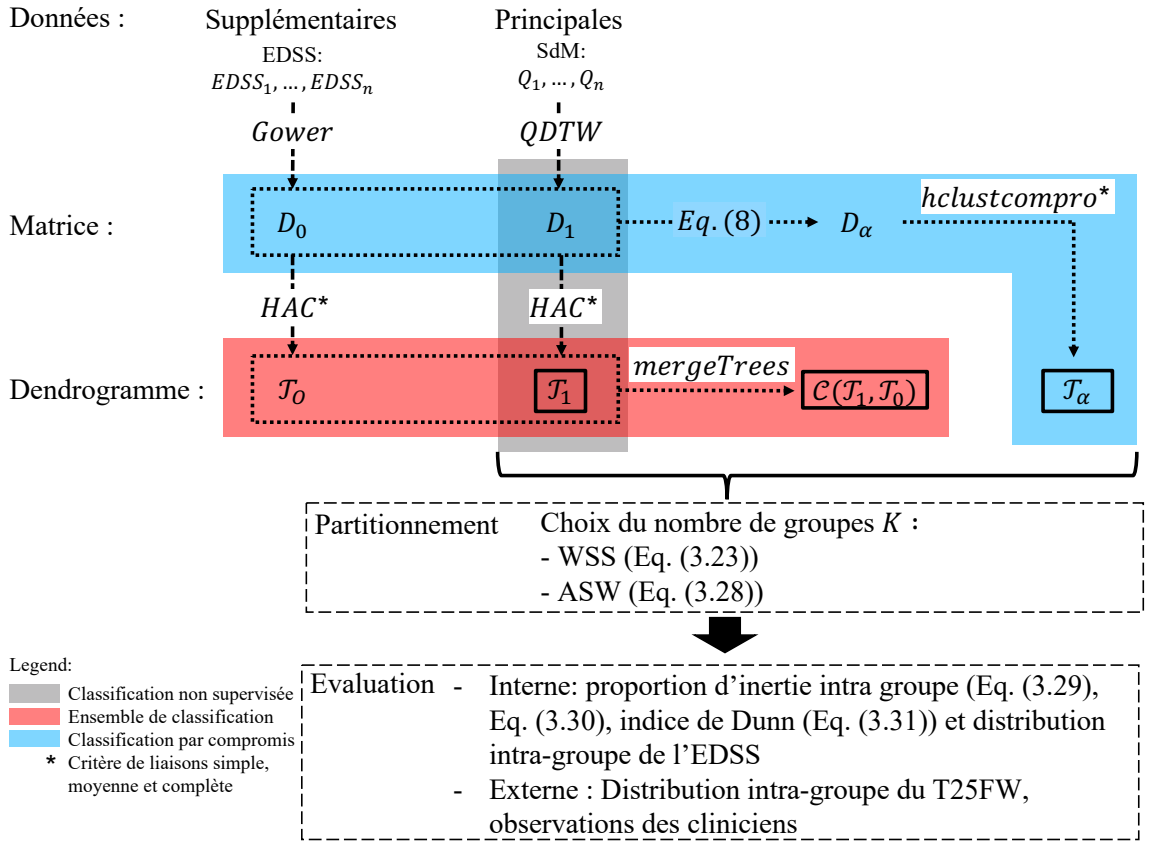
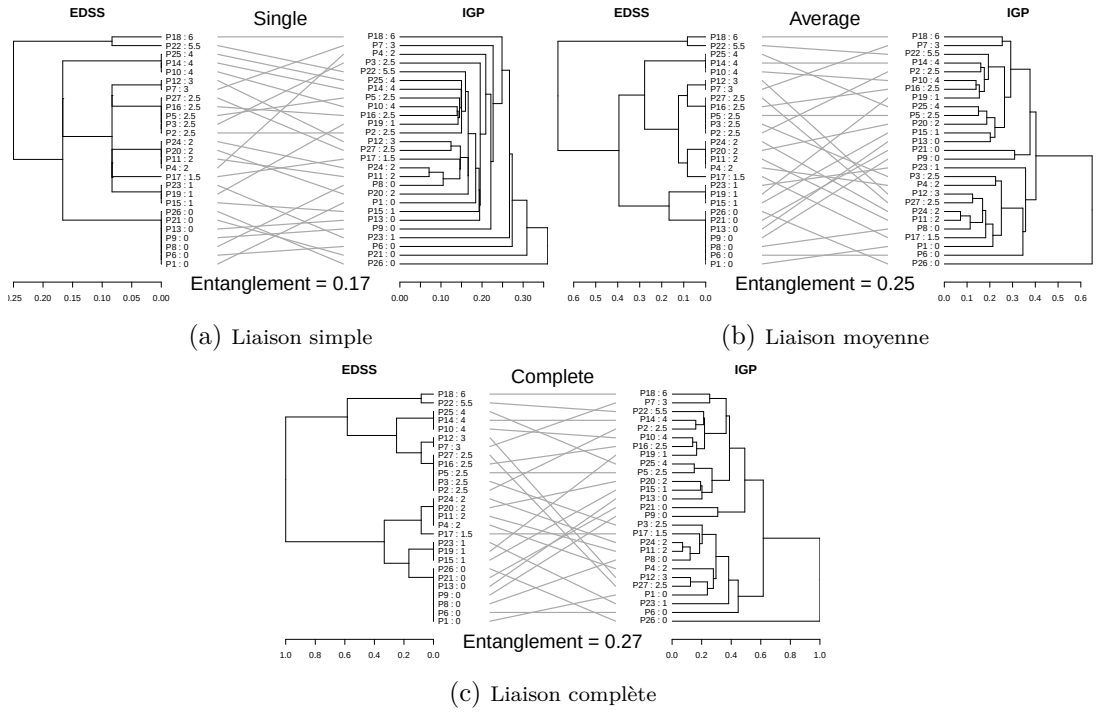


FIGURE 3.8 – Méthodologie de l'étude

3.2.2.3 Résultats

Enchevêtrement de $\mathcal{T}_0$ et $\mathcal{T}_1$. La figure 3.9 présente l'enchevêtrement (*entanglement*) entre les dendrogrammes $\mathcal{T}_0$ et $\mathcal{T}_1$ obtenus par CAH appliquée respectivement sur D_0 et D_1 en utilisant les critères de liaisons simple, moyenne et complète. Les résultats sont anonymisés en remplaçant le nom des patients par un identifiant allant de P1 à P27. Les feuilles des dendrogrammes correspondent à l'identifiant des patients et leur score EDSS.



(a) Liaison simple

(b) Liaison moyenne

(c) Liaison complète

FIGURE 3.9 – Enchevêtrement entre EDSS (à gauche) et SdM (à droite) par critère de liaison

Le score d'enchevêtrement mesure la différence des positions relatives des observations entre deux dendrogrammes, et reflète donc leur différence de structures. Il prend une valeur comprise entre 0 et 1. Plus l'enchevêtrement est élevé, plus les dendrogrammes diffèrent [49]. Les valeurs 0.17, 0.25 et 0.27 observées respectivement pour les critères de liaisons simple, moyenne et complète suggèrent qu'il existe une part d'informations communes entre l'EDSS et la SdM, ce qui n'est pas surprenant étant donné que l'EDSS comprend une évaluation du handicap de la marche. Néanmoins, l'enchevêtrement n'est pas proche de 0, en particulier avec les critères de liaisons moyenne ou complète, ce qui suggère que les deux sources de données contiennent des informations différentes relatives au handicap du patient.

Estimation du paramètre de pondération de `hclustcompro`. La matrice D_α utilisée par la méthode `hclustcompro` pour construire le dendrogramme final dépend de la valeur du paramètre de pondération α d'après l'équation 3.17. Une estimation ponctuelle de ce paramètre est obtenue par l'équation 3.19. Ces calculs aboutissent pour les trois critères de liaison utilisés aux estimations de $\widehat{\alpha}$ décrites dans le tableau 3.2.

TABLE 3.2 – Estimation du paramètre de pondération $\widehat{\alpha}$ utilisé dans la méthode `hclustcompro`.

Paramètre de pondération	**Critère de liaison**		
	simple	moyenne	complète
$\widehat{\alpha}$	0.59	0.67	0.69

Par conséquent, la dissimilarité entre patients est évaluée comme une combinaison linéaire de la dissimilarité entre leur SdM (associée à une pondération de 59%, 67% ou 69% en fonction du critère de liaison considéré) et de la dissimilarité entre leur score EDSS (associée à la pondération restante de 41%, 33% and 31%).

On peut noter que `hclustcompro` intègre une plus large part d'informations provenant des SdMs dans la matrice de dissimimarité finale pour les critères de liaison qui aboutissent à un plus grand enchevêtrement entre les dendrogrammes obtenus à partir des SdMs et du score EDSS. Cela semble indiquer que `hclustcompro` tend naturellement à intégrer plus d'informations provenant de la SdM quand les deux sources d'informations différentes produisent des dendrogrammes différents.

Choix du nombre de groupes. La figure 3.10 représente la variation de la WSS (équation 3.25) et de l'ASW (équation 3.29) pour un nombre de groupes allant de 1 à 10. Ils sont calculés à partir des partitions obtenues en coupant les arbres représentés dans la figure 3.11 à différentes hauteurs pour former le nombre désiré de groupes. Les figures 3.10 et 3.11 sont analysées ensemble afin de déterminer le nombre optimal de groupes commun à toutes les méthodes. Les résultats obtenus à partir du critère de liaison simple sont d'abord commentés.

Les dendrogrammes correspondants, visibles dans la première ligne de 3.11, sont symptomatiques de la tendance bien connue de ce critère à produire des groupes chaînés [66]. Ce phénomère rend difficile la recherche de la hauteur appropriée à laquelle couper le dendrogramme et conduit généralement à former soit un très grand nombre de groupes, soit un seul. De plus, le critère de liaison simple est celui qui conduit à former les dendrogrammes $\mathcal{T}_1$ (à partir de la SdM) et $\mathcal{T}_0$ (à partir de l'EDSS) les plus similaires. Il est écarté dans la suite de cette section. La suite de l'analyse se concentre sur les résultats produits par les critères de liaisons **moyenne et complète**, qui sont décrits par méthode de classification.

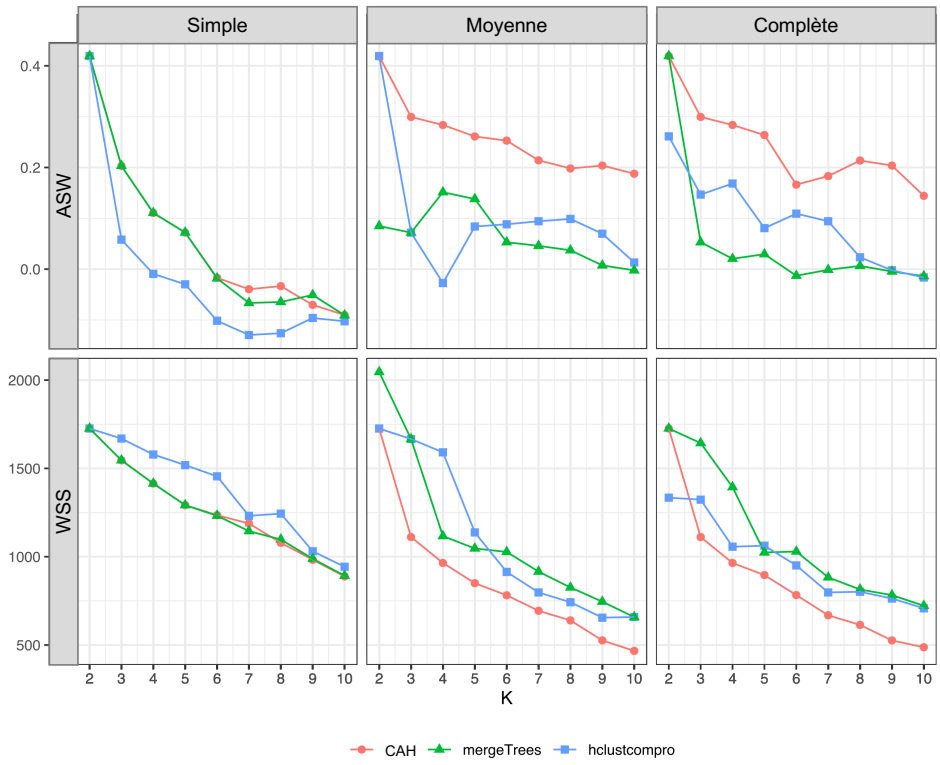


FIGURE 3.10 – WSS et ASW en fonction du nombre de groupes. Les couleurs représentent les différentes méthodes de classification, et les colonnes correspondent aux différents critères de liaison.

`CAH` sur les SdMs. La figure 3.10 suggère que le nombre optimal de clusters devrait être de trois, quel que soit le critère de liaison. En effet, un coude est visible sur les courbes WSS et l'ASW est la deuxième valeur la plus élevée pour $K = 3$. Les deux dendrogrammes formés avec les critères de liaisons complète et moyenne, représentés dans la première colonne de la figure 3.11, présentent une structure relativement similaire. Tous deux semblent indiquer une répartition des patients en seulement trois groupes de taille non équilibrée. Au sein des groupes, la distribution des scores EDSS est très variable, ce qui signifie que les partitions obtenues ne sont pas représentatives du handicap global des patients.

`mergeTrees`. La figure 3.10 indique que quatre ou cinq groupes peuvent être formés pour le critère de liaison moyenne, et cinq groupes lorsque pour le critère de liaison complète. La figure 3.11 montre que la méthode produit deux dendrogrammes très différents selon le critère de liaison (colonne du milieu, deux dernières lignes). Au sein des groupes, la distribution des scores EDSS est moins hétérogène que celle obtenue avec la méthode `CAH`. Cependant, certains groupes rassemblent des patients présentant de grandes différences de handicap global et, à l'inverse, certains patients présentant un handicap global similaire sont attribués à des groupes différents.

`hclustcompro`. La figure 3.10 suggère que cinq clusters devraient être formés avec le critère de liaison moyenne. En effet, on constate que la pente de la courbe WSS devient moins forte à partir de cette valeur et que l'ASW se stabilise pour $K = 5$. Si l'on considère les résultats obtenus à partir du critère de liason complet, les graphiques suggèrent que le nombre optimal de groupes est compris dans l'intervalle $K \in \{4, 5, 6\}$. Dans la figure 3.11, la structure des dendrogrammes semble naturellement suggérer de diviser les patients en cinq groupes. Le dendrogramme produit avec le critère de liaison moyenne présente un singleton. Le dendrogramme produit par le critère de liaison complète suggère de former cinq groupes de tailles relativement similaires et de distributions homogènes des scores EDSS.

En tenant compte de ces observations, et parce que cela permet une comparaison groupe par groupe des classifications obtenues par toutes les méthodes et tous les critères de liaison, 5 groupes sont générés pour toutes les méthodes et les deux critères de liaison. Les groupes de chaque partition ainsi formés sont identifiés par la couleur de leur branche dans la figure 3.11.

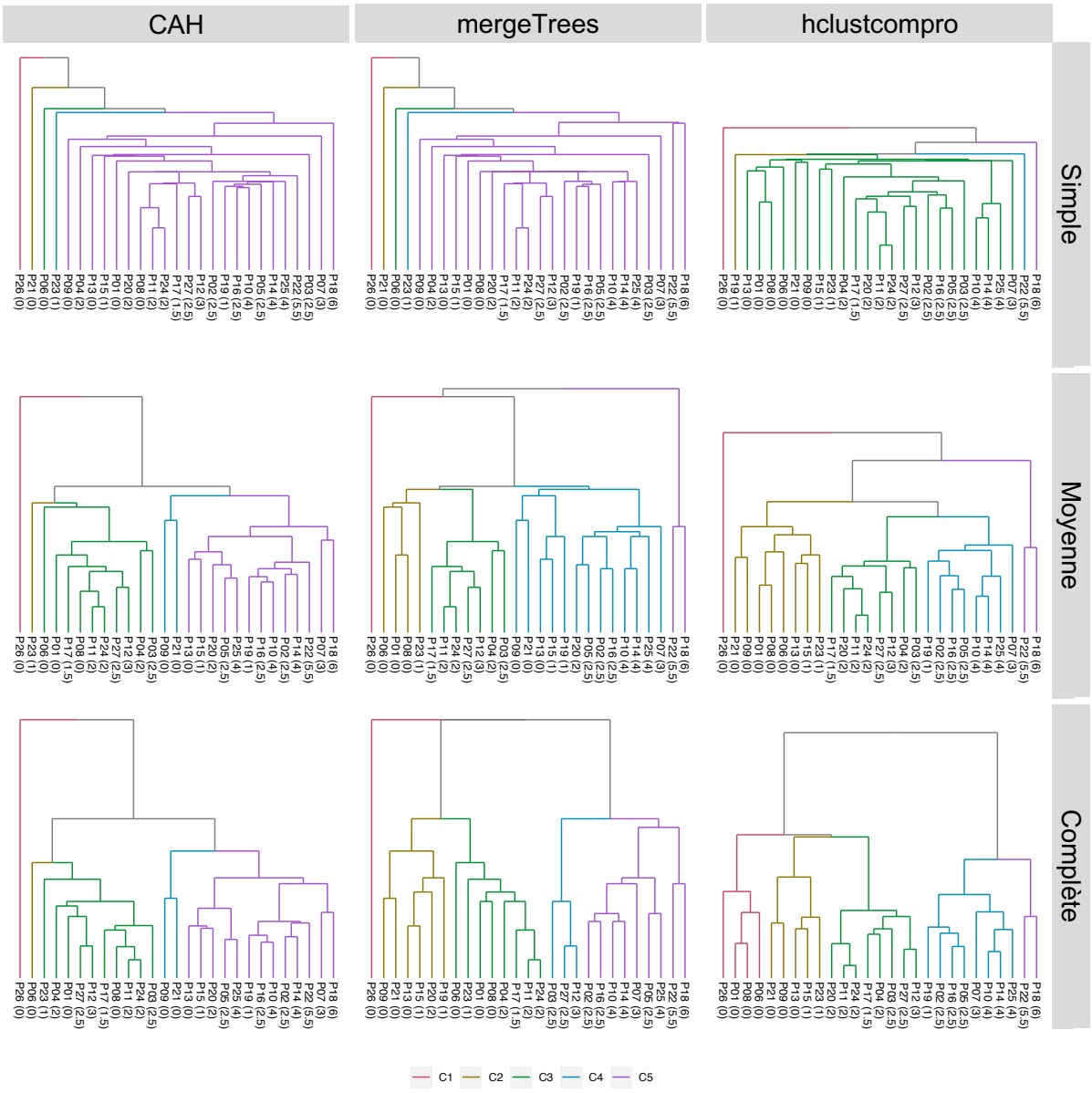


FIGURE 3.11 – Dendrogrammes par méthodes et critère de liaison

Comparaison des partitions obtenues par `CAH`, `hclustcompro` et `mergeTrees` Le tableau 3.3 fournit une description des critères d'évaluation internes associés aux groupes, c'est-à-dire des critères calculés à partir des données ayant été utilisées pour générer les groupes. Plus précisément, chaque groupe est résumé par :

— Sa *taille* n : le nombre de patients dans le groupe k ;

— Sa *proportion d'inertie intra-groupe* $\left(p_W^{(k)}\right)$: cette valeur est calculée selon l'équation (3.31) et correspond au rapport entre l'inertie au sein du groupe et la dissimilarité entre le médoïde du groupe et le médoide global des données ; une valeur faible indique un groupe bien isolé des autres et compact ;

— Son *score EDSS médian* (EDSS) : il renseigne le degré de sévérité médian des patients appartenant à ce groupe.

Dans le tableau 3.3, pour chaque méthode de classification et chaque critère de liaison, les groupes sont ordonnés autant que possible du plus petit au plus grand score EDSS médian. Il n'a pas toujours été possible d'obtenir un classement monotone des groupes selon l'EDSS médian, car parfois les groupes ayant des scores EDSS médians similaires sont trop éloignés les uns des autres dans la structure du dendrogramme. Par exemple, les groupes C1 et C4 formés avec la méthode `CAH` et le critère de liaison moyenne présentent le même EDSS médian, mais la structure du dendrogramme obtenu est telle qu'aucune permutation des branches ne permet de les approcher

TABLE 3.3 – **Résumé des groupes en fonction de leur taille (n), de leur proportion d'inertie intra-groupe $\left(p_W^{(k)}\right)$ et de l'EDSS (EDSS).** Pour un groupe donné, la proportion d'inertie intra-groupe calculée à partir des SdMs selon l'équation (3.31).

Classification	Critère de liaison	C1			C2			C3			C4			C5		
		n	$p_W^{(1)}$	EDSS	n	$p_W^{(2)}$	EDSS	n	$p_W^{(3)}$	EDSS	n	$p_W^{(4)}$	EDSS	n	$p_W^{(5)}$	EDSS
`CAH`	moyenne	1	0.0	0	1	0.0	1	10	47.0	2	2	28.6	0	13	43.8	2.5
`mergeTrees`	moyenne	1	0.0	0	11	52.6	2	5	65.5	0	8	37.0	3	2	43.5	5.5
`hclustcompro`	moyenne	1	0.0	0	8	69.8	0	8	52.0	2	8	33.9	3	2	43.5	5.5
`CAH`	complète	1	0.0	0	1	0.0	0	10	50.5	2	2	28.6	0	13	43.8	2.5
`mergeTrees`	complète	1	0.0	0	6	63.6	1	8	55.2	1.5	3	22.3	2.5	9	38.0	4
`hclustcompro`	complète	4	64.7	0	5	63.6	0	8	52.0	2	8	33.9	3	2	43.5	5.5

En examinant la taille des groupes (colonne n du tableau 3.3), on remarque que les méthodes `CAH` et `mergeTrees` produisent des singletons, c'est-à-dire des groupes avec un seul patient. Il ne s'agit pas d'un artefact dû au choix du nombre de groupes. En effet, quelle que soit la valeur de K, la figure 3.11 montre que les deux méthodes produisent un

singleton à la première séparation de leur dendrogramme. Les *singletons* relèvent d'avantage du domaine de la détection de valeurs aberrantes que de celui de la classification. En effet, les méthodes retirent d'un groupe une observation qui est trop éloignée des autres pour créer un autre groupe uniquement pour cette observation aberrante. Par conséquent, les deux groupes bénéficient d'une réduction de l'inertie comme on peut l'apprécier à partir des colonnes $p_W^{(k)}$ dans le tableau 3.3. Cependant, du point de vue de la classification des individus, c'est-à-dire de la recherche de groupes homogènes d'individus, il n'est pas logique de créer un groupe pour un seul individu. Cela est en accord avec la méthode du coude qui est habituellement utilisée pour avoir un aperçu du nombre optimal de groupes à partir de la courbe WSS. La WSS minimale est trivialement atteinte pour $K = n$, c'est-à-dire lorsque tous les individus sont dans leur propre groupe, aboutissant à une partition inutile. Enfin, nous pouvons voir dans la colonne EDSS que les méthodes `CAH` et `mergeTrees` ne parviennent pas à produire une séquence de groupes avec un EDSS médian croissant. En revanche, la méthode `hclustcompro` produit toujours une séquence de groupes avec un EDSS médian croissant. En outre, lorsqu'elle est combinée avec le critère de liaison complète, cette méthode ne génère pas de singletons, quel que soit le nombre de groupes choisi (voir figure 3.11). Cela se traduit par des proportions d'inertie intra-groupe non nulles pour tous les groupes dans le tableau 3.3.

L'évaluation des performances des méthodes de classification est également décrite pour leur partition globale plutôt que par groupe. Le tableau 3.4 présente les proportions globales d'inertie au sein des groupes pour l'ensemble de la partition (p_W). L'indice de Dunn est également indiqué. On rappelle que cet indice est élevé lorsque les groupes sont bien séparés les uns des autres et composés d'individus ayant des SdMs similaires. Ces mêmes critères sont également proposés en ne les calculant que pour les trois derniers groupes C3, C4 et C5, aucun de ces groupes n'étant des *singletons* pour aucune méthode. Deux points justifient ce choix :

(i) Les groupes sont appariés par les scores EDSS médians entre les méthodes

(ii) Les *singletons* dont indésirables dans le cadre de l'établissement d'une partition significative et améliorent à tort les critères de validation internes

Comme attendu, si l'on examine les mesures de performances calculées à partir des cinq groupes, la méthode `CAH` surpasse les méthodes semi-supervisées car elle ne prend en compte que les SdMs. Elle parvient donc mieux à former des groupes de patients présentant des SdMs similaires. La méthode `mergeTrees` arrive en deuxième position, ce qui peut largement être attribué à la présence d'un singleton. Cette interprétation

TABLE 3.4 – Validation interne des partitions.

Méthode	Critère	Tous les groupes		Groupes sans *singletons*	
		p_W	DI	p_W	DI
`CAH`	moyenne	34.0	0.377	42.5	0.377
`mergeTrees`	moyenne	42.7	0.288	49.2	0.288
`hclustcompro`	moyenne	45.9	0.192	42.1	0.467
`CAH`	complète	35.0	0.428	43.8	0.428
`mergeTrees`	complète	40.2	0.288	41.4	0.329
`hclustcompro`	complète	51.6	0.214	42.1	0.467

est confirmée lorsque l'on examine les mesures de performances calculées en utilisant uniquement les groupes qui ne sont pas des singletons. De ce point de vue, la méthode `hclustcompro` présente des performances similaires à la méthode `CAH` en termes d'inertie et surpasse nettement les méthodes `CAH` et `mergeTrees` lors de la comparaison des indices de Dunn.

La figure 3.12 affiche les distributions intra-groupes des scores EDSS et la durée du test T25FW pour les trois méthodes de classification (`CAH`, `hclustcompro`, `MergeTrees`) utilisées avec les critères de liaisons complète et moyenne. Elle fournit une description des groupes du point de vue des indicateurs cliniques standards de l'incapacité globale (EDSS) et des troubles de la marche (T25FW), qui complète la description fournie dans le tableau 3.3.

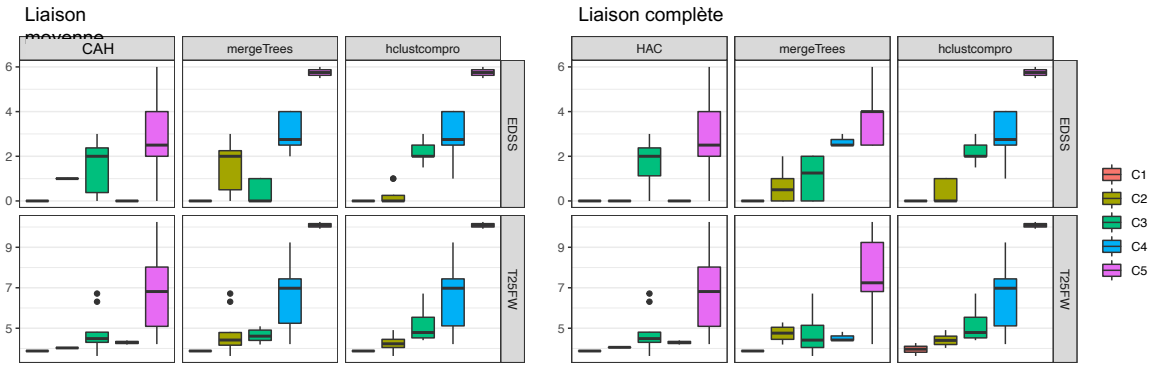


FIGURE 3.12 – Description de la distribution des scores EDSS et des temps de marche au T25FW pour chaque groupe.

Les distributions des scores EDSS et du temps de marche lors du test T25FW sont présentées par groupe dans la figure 3.12. Elles sont utilisées pour évaluer la pertinence

clinique des partitions générées. Nous avons également demandé à cinq neurologues des CHU de Nantes et de Rennes d'interpréter les résultats. Leur analyses a conduit aux observations suivantes :

`CAH`. Les groupes formés par `CAH` rassemblent des patients dont la gravité de la SEP est très hétérogène. Par exemple, le groupe C5 formé avec les critères de liaisons complète et moyenne contient des patients ayant des fonctions neurologiques normales (EDSS=0) et des patients qui ne peuvent pas marcher 100 mètres sans aide à la marche (EDSS=6). Les patients de ce groupe présentent également des temps de marche très différents lors du test T25FW, comme illustré dans la figure 3.12.

`mergeTrees`. Les partitions fournies par `mergeTrees` regroupent des patients dont la gravité de la SEP est plus similaire. Ils ont tendance à présenter des scores EDSS plus élevés dans les deux derniers groupes C4 et C5 que dans les trois premiers clusters C1, C2 et C3. Cependant, certains patients présentant des scores EDSS similaires sont répartis dans différents groupes, tandis que d'autres groupes rassemblent des patients présentant une gravité de la SEP différente. De plus, les groupes ne semblent pas rassembler les patients en fonction de leur temps de marche.

`hclustcompro`. Les partitions fournies par `hclustcompro` sont plus homogènes. Il est intéressant de noter que les groupes C3, C4 et C5 sont les mêmes pour les critères de liaison moyenne et complète. Avec le critère de liaisons moyenne, le groupe C1 est un singleton. Ce patient isolé (P27) se trouve également dans un singleton pour toutes les autres partitions sauf celle obtenue avec `hclustcompro` avec le critère de liaison complète. Le groupe C2 regroupe les patients ayant un score EDSS compris entre 0 et 1, qui présentent le temps de marche le plus faible lors du test T25FW. Avec le critère de liaison complet, le groupe C1 regroupe les patients ayant un score EDSS de 0 (c'est-à-dire identiques aux individus sains d'un point de vue neurologique). La figure 3.12 montre également qu'ils correspondent aux patients ayant le temps de marche le plus faible pendant le test T25FW. Le groupe C2 rassemble les patients ayant des scores EDSS de 0 et 1. Ils ont tendance à avoir des temps de marche légèrement plus élevés pendant le test T25FW que les patients du groupe C1. Si l'on compare avec le cluster C1, cela suggère que la SdM pourrait être capable de séparer les patients ayant une condition neurologique apparemment normale sur la base de leur déficience de la marche. Le groupe C5 rassemble les deux patients dont l'évaluation neurologique est la plus mauvaise. Ce sont également les deux patients avec les temps de marche les plus élevés au test T25FW. Les groupes C3 et C4 sont

moins homogènes en termes de caractéristiques cliniques, mais ils regroupent les patients présentant un degré de gravité intermédiaire de la pathologie et le groupe C4 regroupe les patients présentant des scores EDSS légèrement plus élevés que les patients du groupe C3, bien que cette différence ne semble pas significative.

D'après ces observations, la méthode `hclustcompro` avec critère de liaison complète conduit à la meilleure classification de la cohorte. Comme illustration finale de ces résultats, la variation de l'angle de rotation de la hanche au cours d'un cycle de marche est calculée à partir de la SdM des patients (voir équation 2.18, section 2.2.1). Elle est représentée dans la figure 3.13. Les courbes en gras représentent les médoïdes de chaque groupe (*c.f.* équation (3.26)). La figure 3.13 révèle que les patients des groupes C1, C2 et C3 ont tendance à présenter une plus grande amplitude de rotation de la hanche que les patients des groupes C4 et C5. Cette observation est à mettre en parallèle avec le fait que les groupes C1 à C3 rassemblent également des patients présentant des sévérités globales de la pathologie et des troubles de la marche plus légers que ceux des groupes C4 et C5.

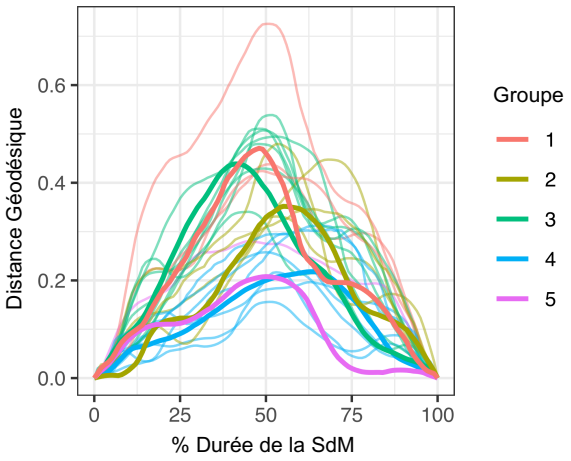


FIGURE 3.13 – Angle de rotation de la hanche associé aux SdMs par groupes (méthode `hclustcompro`)

3.2.2.4 Discussion

Dans cette section ont été présentées les adaptations de deux méthodes de classification hiérarchique semi-supervisée aux séries temporelles de quaternions unitaires. La première nommée `hclustcompro` est une méthode de classification par compromis. La seconde nommée `mergeTrees` est une méthode d'ensemble de classifications. L'adaptation de ces

deux méthodes repose sur l'utilisation du QDTW, une mesure de dissimilarité adaptée aux séries chronologiques de quaternions unitaires.

Si ces deux méthodes permettent de tenir compte de sources d'informations supplémentaires, elles se distinguent par la façon dont ces dernières sont intégrées dans la classification. Dans `hclustcompro`, la dissimilarité globale entre les observations est calculée comme la somme pondérée des dissimilarités observées dans les espaces des deux sources d'informations. La classification est alors générée à partir de cette dissimilarité globale. Dans `mergeTrees`, ce sont les classifications générées dans l'espace des informations principales et supplémentaires qui sont agrégées pour construire une classification consensus. L'ordre des étapes de classification et d'agrégation des différentes sources d'informations est donc en quelque sorte inversé entre les deux approches.

Les deux méthodes se différencient également par le nombre de sources d'informations qu'elles peuvent intégrer. Dans sa formulation actuelle, `hclustcompro` est limitée à deux sources d'information là où `mergeTrees` peut en intégrer un nombre virtuellement illimité. Cependant, le coefficient de pondération intervenant dans `hclustcompro` permet une interprétation claire et directe concernant la proportion avec laquelle chaque source d'information est utilisée dans la détermination de la dissimilarité globale. Il n'existe pas d'équivalent pour `mergeTrees`.

Leur application dans le contexte de l'analyse de la marche des patients atteints de SEP démontre à la fois l'importance et l'utilité d'injecter des informations supplémentaires quand elles sont disponibles, les deux méthodes semi-supervisées fournissant des partitions plus interprétables sur le plan clinique que la CAH non supervisée appliquée uniquement sur les données de marche. Les résultats tendent également à démontrer la supériorité de l'approche de la classification par compromis dans ce contexte. Ce résultat n'est pas surprenant car `mergeTrees` a été décrit par ses auteurs comme moins efficace lorsque les sources contiennent des informations différentes [75].

Il faut cependant rappeler que le jeu de données de cette étude est de taille relativement modeste. Par conséquent, ces résultats et conclusions doivent être confirmés sur une cohorte de patients plus importante. L'ajout de nouveaux patients à l'analyse peut conduire à la formation de nouveaux groupes et à l'identification de caractéristiques de marche plus précises partagées par les patients d'un même groupe. Pour ce faire, une seconde étude clinique a été préparée avec l'équipe du CIC de neurologie du CHU de Nantes. Cette étude incluant 44 patients a pu être financée grâce au soutien financier du PEPS de l'AMIES. La période d'inclusion de ces patients est toujours en cours et s'étend

de Aout 2021 à Aout 2022.

De plus, étudier la relation entre la SdM et d'autres paramètres spatio-temporels caractérisant la marche, tels que ceux mesurés par d'autres dispositifs (par exemple : `Gaitrite`[1] [159]) peut améliorer l'interprétation des résultats.

Une perspective de développement méthodologique pour `hclustcompro` consisterait naturellement à l'étendre aux cas pour lesquels plus de deux sources d'informations sont à prendre en compte. Cette extension pourrait s'appuyer sur les méthodes factorielles pour l'analyse conjointe de plusieurs tableaux de données, telles que `STATIS` [44, 101]. Dans le contexte de la marche, il pourrait être possible de prendre en compte des paramètres en plus de la forme de la SdM, par exemple la durée moyenne des cycles ou sa variabilité intra-individuelle qui a été décrite comme liée aux capacités ambulatoires (par exemple dans les risques de chute chez les personnes âgées [138]).

La CAH étant une méthode basée sur la distance, les résultats sont par construction influencés par le choix de la mesure de dissimilarité. Ici la dissimilarité QDTW a été utilisée car elle est à ce jour la seule adaptée aux séries temporelles de quaternions unitaires décrite dans la littérature. Il existe plusieurs mesures décrites pour les séries temporelles d'observations euclidiennes, lesquelles étant élastiques, basées sur des modèles ou sur des représentations spécifiques des données. La généralisation de ces méthodes aux QTS permettrait d'étendre l'éventail des choix possibles pour mieux s'adapter aux différents contextes d'étude.

3.2.3 Valorisation

Les travaux de généralisation de la méthode `hclustcompro` aux QTS et son application à l'analyse de la marche des patients atteints de SEP ont fait l'objet d'une communication scientifique orale sous forme de poster durant la 8th Channel Network Conference. Cet évènement s'est déroulé en ligne du 7 au 9 avril 2021, et a été organisée par la Société Française de Biométrie.

Les travaux de généralisation de la méthode `hclustcompro` aux QTS et sa comparaison avec la méthode `mergeTrees` dans le contexte de l'analyse de la marche de patients atteints de SEP sont présentés dans l'article "Semi-supervised clustering of quaternion time series : application to gait analysis in multiple sclerosis using motion sensor data" accepté après révisions mineures par la revue Statistics in Medicine (Wiley Library)[2]. Ils

1. http://www.biometrics.fr/web/fr/gaitrite-et-cirface/71-gaitrite.html
2. https://onlinelibrary.wiley.com/journal/10970258

feront également l'objet d'une présentation orale lors de la 31st International Biometric Conference, organisée à Riga, Lettonie, par l'International Biometric Society, du 10 au 15 Juillet 2022.

Conclusions générales

L'analyse quantitative de la marche par dispositifs numériques est un champ de recherche vaste. Ces dispositifs peuvent permettre de quantifier de nombreux aspects représentant la démarche de l'individu. Leur utilisation dans le domaine de la santé est particulièrement prometteuse, car ils fournissent une mesure quantitative et objective de certains aspects de la marche directement impactés par la pathologie du patient. Ces informations peuvent venir compléter les examens médicaux réalisés par les cliniciens. Parmi les différents types de technologies, les dispositifs portatifs présentent de nombreux avantages, tels que leur coût faible, la facilité avec laquelle ils peuvent être manipulés, et leur potentiel usage dans la vie courante. Cette dernière caractéristique est de première importance, car elle pourrait permettre une mesure non biaisée par le contexte hospitalier et régulière de l'impact de la pathologie sur le quotidien du patient. Leur usage aurait donc une application pour suivre l'état de santé des patients au cours du temps et pour évaluer les bénéfices de nouveaux traitements. C'est pourquoi le développement de méthodes d'analyse des données renvoyées par ce type de dispositif pour caractériser les déficits de la marche revêt un très grand intérêt.

Les premiers travaux de recherche présentés dans ce manuscrit pour le développement de la solution *eGait* ont été menés dans cette optique. Cette solution se base sur la mesure de l'orientation de la hanche au cours de la marche par un unique système de capteurs inertiel appelé *MetaMotionR*. Les données mesurées se présentent sous la forme d'une séquence de quaternions unitaires. L'algèbre de ces nombres hypercomplexes à 4 dimensions facilite la représentation des rotations en 3 dimensions, mais définit des règles de calcul spécifiques qui limitent l'application de méthodes adaptées aux données euclidiennes. Le développement de méthodes appropriées aux quaternions unitaires a donc été nécessaire pour analyser les données de marche. En effet, de nombreuses approches de mesure de la marche par capteurs portatifs sont décrites dans la littérature. Quelques unes se basent sur l'utilisation de quaternions unitaires, la plupart du temps pour représenter l'orientation de segments du corps humain, ou l'angle de certaines de ses articulations. Cependant, aucune méthode existante pour l'analyse de l'orientation de la hanche mesurée par un dispositif portatif sous forme de quaternions unitaires n'a été identifiée. Les

approches proposées sont donc inspirées des solutions existantes en les adaptant à ce type particulier de données.

La première étape, présentée dans la section 2.1, a consisté à développer un algorithme permettant d'identifier les instants délimitant les phases d'appui et de balancement d'un cycle de marche. Cet algorithme, appelé `STRIde PAttern GEneration`, se base sur un ensemble de règles formulées en considérant la position du dispositif sur le corps, le type de données mesurées et les mouvements de la hanche durant la marche. Il permet d'extraire les segments correspondant aux mouvements de la hanche durant les cycles de marche délimités par la pose du pied droit au sol d'un jeu de données brut mesuré par le dispositif *MetaMotionR*. Une première évaluation des performances de cet algorithme a pu être réalisée en comparant le nombre de cycles de marche détectés par l'algorithme avec le nombre de cycles de marche identifiés sur des données vidéos prises durant l'acquisition des données. Les données ont été mesurées pendant la marche de volontaires sains. Si la difficulté des algorithmes à détecter les cycles de marche d'individus présentant des troubles de la marche est un phénomène connu, les résultats permettent de conclure que l'algorithme `STRIPAGE` présente des performances tout à fait satisfaisantes. Sa spécificité et sensibilité sont en effet très élevées, y compris pour détecter des cycles de marche dans des données où un trouble de la marche a été simulé par le port d'une orthèse bloquante du genou. Quelques limites sont toutefois associées au contexte de l'expérience. Tout d'abord, le matériel à disposition n'a pas pu permettre d'évaluer la précision avec laquelle l'algorithme `STRIPAGE` détecte les évènements du cycle de marche à la milliseconde près, comme cela a pu être décrit pour d'autres dispositifs. Cette évaluation nécessite en effet de synchroniser les données mesurées par le dispositif testé avec un autre dispositif considéré comme *Gold Standard* (*e.g.* reconnaissance visuelle ou tapis de capteurs de pression) pour comparer leurs résultats. Le port de l'orthèse bloquante du genou n'est pas représentatif du large éventail de troubles de la marche pouvant être observés chez des patients atteints de pathologies diverses, telles que la Sclérose En Plaques. D'autres commentaires peuvent être formulés concernant la structure même de l'algorithme `STRIPAGE`. Il nécessite de définir *a priori* des valeurs pour plusieurs hyper-paramètres pour la détection des cycles. Si certaines valeurs par défaut sont proposées et permettent d'obtenir de bonnes performances, ces dernières ont été estimées de manière empirique. Une étude plus poussée pourrait permettre de définir des valeurs permettant d'optimiser les performances de l'algorithme. Enfin, si la durée des cycles de marche du porteur du dispositif est estimée de façon automatique par la méthode du *periodogram*, cette dernière n'est pas adaptée aux séquences

de quaternions unitaires. Il est donc nécessaire de l'appliquer de manière indépendante sur les 4 composantes des quaternions. La généralisation de cette méthode à ce type de données, en plus de présenter une perspective intéressante de recherche, pourrait permettre d'optimiser l'algorithme. Enfin, l'algorithme `STRIPAGE` est conçu pour identifier des cycles à partir de séquences mesurées sur une courte durée et durant laquelle le porteur du dispositif marche. Il est donc très improbable qu'il puisse permettre de détecter des cycles de marche dans des données plus complexes mesurées, par exemple mesurées sur une journée complète. Une phase de pré-traitement permettant d'identifier les phases de marche est donc nécessaire avant d'envisager son utilisation en vie quotidienne.

Plusieurs approches ont été explorées pour identifier des informations relatives aux déficits de marche du porteur du dispositif. La première, présentée dans la section 2.2, consiste à calculer un ensemble de paramètres spatio-temporels représentant la démarche, à partir des cycles de marche identifiés par l'algorithme `STRIPAGE`. Ces paramètres correspondent à la durées moyenne des cycles, des phases d'appui et de balancement, ansi que l'amplitude moyenne de la rotation de hanche et sa vitesse angulaire moyenne au cours des cycles. Ils ont été comparés à plusieurs scores et paramètres utilisés classiquement pour quantifier l'état de santé global et les déficits neurologiques et de marche de patients atteints de Sclérose En Plaques. Les résultats obtenus à partir des données mesurées chez 30 patients constituent une première preuve de la relation entre plusieurs de ces paramètres de marche et leur état de santé. La durée des cycles de marche tend à baisser chez les patients présentant une sévérité modérée et élevée par rapport aux patients avec un stade précoce de la pathologie, tandis que l'amplitude de la rotation de leur hanche et sa vitesse angulaire tend à être plus faible. Le même phénomène semble également être observé pour les patients présentant un trouble de la fonction neurologique pyramidale. Cette observation est cependant à nuancer au vu de la taille de l'échantillon. Enfin, la diminution de la vitesse de marche des patients, principal indicateur de la présence de troubles de marche, tend à diminuer avec l'augmentation de la durée des cycles et la diminution de l'amplitude et de la vitesse de rotation de la hanche. Ces observations sont en accord avec la littérature, cependant aucune étude faisant état de la relation directe entre la vitesse de rotation de la hanche et la vitesse de marche n'a été identifiée dans la littérature. Des modèles statistiques plus complexes pourraient permettre de modéliser la vitesse de la marche à partir de ces seuls paramètres, et ainsi permettre son estimation en vie quotidienne.

Une deuxième approche a été explorée pour regrouper les individus en fonction de leur

démarche par méthode de classification non supervisée et semi-supervisée. Pour ce faire, la démarche des individus est représentée par un biomarqueur appelé "Signature De Marche" (SdM). Elle se présente sous la forme d'une séquence de quaternions unitaires représentant l'orientation de la hanche de l'individu au cours d'un cycle. La forme de la SdM est supposée représentative de sa démarche. Pour l'obtenir, le centre des cycles de marche détecté chez un individu par l'algorithme `STRIPAGE` est calculé par la méthode du *K-means alignment*. Cette méthode permet d'aligner un ensemble de données fonctionnelles par des fonctions de *warping* pour s'affranchir de leur variabilité de phase, et d'en calculer le centre. Le *K-means alignment* a été initialement développé pour être appliqué sur des données fonctionnelles prenant valeur dans un espace euclidien, une adaptation aux quaternions unitaires a dû être proposée. De la même façon, les méthodes de Classification Ascendante Hiérarchique (CAH) et du *K-medoids alignment* de données fonctionnelles, ainsi que les méthodes de CAH, *K-means* et *K-medoids* de séries chronologiques ont été adaptées aux quaternions unitaires. Elles ont été comparées quant à leur capacité à former des groupes rassemblant des volontaires présentant ou non un déficit de marche simulé à partir de leur SdM. Les résultats montrent l'importance du choix de l'orientation de référence à partir de laquelle exprimer la rotation de la hanche. En effet, les résultats sont généralement meilleurs pour l'ensemble des méthodes lorsque cette dernière est exprimée en définissant l'orientation de la hanche observée au début de la SdM comme orientation de référence. L'importance de s'affranchir de la variabilité de l'amplitude de la rotation de la hanche a également été identifiée. Une méthode proposée pour calculer la signature de marche à partir d'un angle commun entre toutes les SdMs a en effet permis d'améliorer les performances des méthodes de classification. Dans le contexte de cette étude, les méthodes de *K-medoids alignement* et *K-means alignement* sont celles présentant les meilleures performances, parvenant dans certains cas à séparer parfaitement les volontaires en fonction de leur trouble de marche simulé. Les résultats restent cependant globalement mitigés, et les groupes formés par l'ensemble des méthodes de classification correspondent peu à la présence de déficit de marche simulé. Une explication possible est leur sensibilité à la présence de SdMs aux formes atypiques, voir aberrantes. Il est aussi envisageable que le port de l'orthèse bloquante du genou n'entraîne pas une modification de la rotation de la hanche suffisamment importante.

Enfin, deux méthodes de classification semi-supervisées ont été adaptées aux séries chronologiques de quaternions unitaires. Elles permettent d'introduire des sources d'informations supplémentaires dans la construction des groupes. La première, nommée `hclustcompro`,

est une méthode de classification ascendante hiérarchique (CAH) *par compromis*. Elle calcule la dissimilarité globale entre un ensemble d'observations comme la moyenne pondérée de la dissimilarité entre observations dans l'espace d'une source d'information principale et la dissimilarité dans l'espace d'une source d'informations supplémentaire. La seconde, nommée `mergeTrees`, est une méthode *d'ensemble de classification*. Elle construit un arbre dit "consensus" à partir d'un ensemble de dendrogrammes obtenus par méthodes de CAH appliquées sur plusieurs sources d'informations associées aux mêmes observations. Ces deux méthodes sont comparées sur leurs capacités à former des groupes rassemblant des patients atteints de SEP ayant une démarche et une sévérité globale de la pathologie similaire. Les résultats montrent que l'ajout d'informations supplémentaires améliore la pertinence clinique des partitions obtenus par rapport à la méthode de CAH non supervisée. Il est également constaté que la méthode `hclustcompro` utilisée avec le critère de liaison complète permet d'obtenir la partition la plus pertinente du point de vue clinique. Les groupes qu'elle forme sont en effet les plus homogènes du point de vue de la forme des SdMs et de la sévérité de la pathologie. Elle permet également l'identification de deux groupes séparant des patients présentant un stade peu avancé, ce qui peut laisser suggérer que cette méthode permet d'identifier des sous-populations chez les patients présentant une sévérité de la pathologie apparemment faible. Le statut indésirable des singletons dans le contexte de la classification a également été discuté.

L'ensemble de ces résultats constitue donc une première preuve de l'intérêt scientifique de l'usage de la solution *eGait* dans le suivi de patients atteints de maladies neurodégénératives. Plusieurs perspectives de recherche peuvent être identifiées. Tout d'abord, les performances de l'algorithme `STRIPAGE` doivent évaluées à l'aide de méthodes d'analyse de la marche considérées comme *Gold Standard* telles que le tapis de pression *GaitRite*. Pour ce faire, une étude nommée *eValGait* est en cours de préparation avec l'hôpital Bellier de Nantes. Elle prévoit l'inclusion de sujets sains et de personnes âgées à risque de chute qui réaliseront des tests de marche sur un tapis *GaitRite* en portant le dispositif MMR à la ceinture. Cette comparaison pourra également permettre d'optimiser le choix des hyper-paramètres afin d'améliorer les performances de la détection des cycles de marche par `STRIPAGE`. De plus, elle constitue un premier cas d'application de la solution *eGait* pour évaluer les troubles de la marche causés par une pathologie différente de la SEP. Les résultats de classification semi-supervisée des SdMs et EDSS ainsi que les relations entre les paramètres spatio-temporels des cycles de marche et les déficits de la marche observés chez les patients atteints de SEP nécessitent d'être confirmés sur une plus large

cohorte. Pour cela, une étude financée par un PEPS proposé par AMIES et incluant une quarantaine de patients atteints de SEP a été menée au CHU de Nantes. De plus, une étude multi-centrique financée par l'ARSEP est a été préparée avec les CHUs de Nantes et de Rennes. Ces deux études prévoient la mesure des données de marche des patients avec le dispositif MMR au cours du T25FW ainsi que la détermination des scores EDSS et sous scores neurologiques associés. L'étude multicentrique prévoie également d'intégrer les examens par Imagerie par Résonance Magnétique (IRM) pour mesurer la charge lésionelle neurologique des patients. Ces données pourront être comparées avec la SdM, et intégrées dans les analyses, par exemple en généralisant la méthode de classification semi-supervisée `hclustcompro` aux cas pour lesquels plus de deux sources d'informations sont à prendre en compte (*e.g.* variabilité de la marche). D'autres perspectives de recherche sont actuellement en cours d'exploration. Dans le cadre du stage de Master 2 de Raphaël Brard, les méthodes d'apprentissage supervisé *Classification And Regression Tree*, *Support Vector Machine*, *k-Nearest Neighbour* et régression logistique ont été adaptées aux séries de quaternions unitaires. Ces méthodes ont été comparées dans le contexte de la détection de périodes de marche dans des données mesurées par le dispositif MMR représentatives de l'activité quotidienne d'un individu. Ces travaux constituent une première étape pour rendre la solution *eGait* compatible avec la mesure de la marche en vie réelle. Enfin, la généralisation des méthodes de classification semi-supervisée et supervisée aux séquences de quaternions unitaires fait l'objet du sujet de la thèse de Klervi Le Gall "Apprentissage et reconnaissance des différents troubles de la marche à l'aide d'un capteur de mouvements : le cas des patients atteints de Sclérose en Plaques[3]", ayant débutée en octobre 2021. Dans le contexte de ces travaux, la généralisation de l'Analyse par Composante Principale aux séries chronologiques de quaternions unitaires est également abordée, dans le but de réaliser des analyses exploratoires sur un ensemble de SdMs, de synthétiser de nouvelles SdMs à l'aide d'une méthode basée sur l'ACP et les proches voisins et d'appliquer des méthodes de classification après réduction des dimensions des SdMs.

3. https://agence.lebesgue.fr/node/204

SIGNIFICATION DES SIGLES ET ACRONYMES

Acronyme/sigle	Signification
AMIES	Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société
ARI	Adjusted Rand Index
ARSEP	Aide à la Recherche dans la Sclérose En Plaques
ASW	*Average Silhouette Width coefficient*
BDDapp	Base de données apprentissage
BDDsep	Base de données SEP
BDDtest	Base de données test
CAH	Classification Ascendante Hiérarchique
CHU	Centre Hospitalier Universitaire
CIC	Centre d'investigation Clinique
CL	*Cannot Link constraint*
CV	Coefficient de Variation
DBA	Dynamic time warping Barycenter Averaging
DI	*Dunn Index*
DTW	*Dynamic Time Warping*
EDR	*Edit Distance on Real sequence*
EDSS	*Expended Disability Status Scale*
ERP	*Edit Distance with Real Penalty*
FF	*Flat Foot*
FN	*False Negative*
FP	*False Positive*
HO	*Heel Off*
HS	*Heel Strike*
ID	*Intervalle de Durée*
IMU	*Inertial Measurement Unit*

IQR	*Inter Quartile Range*
LCSS	*Longest Common Sub Sequence*
LMJL	Laboratoire de Mathématiques Jean Leray
MAD	*Median Absolute Deviation*
ML	*Must Link constraint*
MMR	*MetaMotionR*
MS	*Multiple Sclerosis*
MSFC	*Multiple Sclerosis Functional Composite*
OFSEP	Observatoire Français de la SEP
PEPS	Projet Exploratoire, Premier Soutien
PPV	*Positive Predictive Value*
PST	Paramètre spatio-temporel
PU-PH	Professeur des Université, Praticien Hospitalier
QDBA	Quaternion Dynamic time warping Barycenter Averaging
QDTW	*Quaternion Dynamic Time Warping*
QTS	*unit Quaternion Time Series*
RI	*Rand Index*
RMSE	*Root Mean Square Error*
SdM	Signature de marche
SE-EDSS	Sévérité élevée
SEP	Sclérose En Plaques
SF-EDSS	Sévérité Faible
slerp	*spherical linear interoplation*
SM-EDSS	Sévérité Moyenne
SPA	*Stride Pattern Analysis*
SQUAT	*Statistics for QUAternion Temporal data*
STRIPAGE	*STRIde PAttern GEneration*
T25FW	*Timed 25 Foot Walk*
TN	*True Negative*
TO	*Toe off*
TP	*True Positive*
TPR	*True Positive Rate*
TS	*Time series*
WCDSW	*Walking Cycle Detection Sweeping Window*

WSS	*Within Sum of Squared error*

BIBLIOGRAPHIE

[1] W.H. ABDULLA, D. CHOW et G. SIN. « Cross-words reference template for DTW-based speech recognition systems ». In : *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*. T. 4. 2003, 1576-1579 Vol.4. DOI : 10.1109/TENCON.2003.1273186[4].

[2] Charu AGGARWAL et Chandan REDDY. *Data Clustering Algorithms and Applications*. Août 2013.

[3] Rakesh AGGARWAL, Christos FALOUTSOS et Arun SWAMI. « Efficient similarity search in sequence databases ». In : *Foundations of Data Organization and Algorithms*. Sous la dir. de David B. LOMET. Berlin, Heidelberg : Springer Berlin Heidelberg, 1993, p. 69-84. ISBN : 978-3-540-48047-1.

[4] Saeed AGHABOZORGI, Ali SEYED SHIRKHORSHIDI et Teh YING WAH. « Time-series clustering – A decade review ». en. In : *Information Systems* 53 (oct. 2015), p. 16-38. ISSN : 0306-4379. DOI : 10.1016/j.is.2015.04.007[5].

[5] D. G. ALTMAN et J. M. BLAND. « Measurement in Medicine : The Analysis of Method Comparison Studies ». In : *Journal of the Royal Statistical Society. Series D (The Statistician)* 32.3 (1983), p. 307-317. (Visité le 07/04/2022).

[6] K. AMINIAN et al. « Spatio-temporal parameters of gait measured by an ambulatory system using miniature gyroscopes ». In : *Journal of Biomechanics* 35.5 (2002), p. 689-699. ISSN : 0021-9290. DOI : https://doi.org/10.1016/S0021-9290(02)00008-8[6].

[7] Lorenza ANGELINI et al. « Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting ». In : *Journal of Neurology* 267.10 (oct. 2020), p. 2897-2909. ISSN : 1432-1459. DOI : 10.1007/s00415-020-09928-8[7].

4. https://doi.org/10.1109/TENCON.2003.1273186
5. https://doi.org/10.1016/j.is.2015.04.007
6. https://doi.org/https://doi.org/10.1016/S0021-9290(02)00008-8
7. https://doi.org/10.1007/s00415-020-09928-8

[8] Stéphane Armand. « Analyse Quantifiée de la Marche : extraction de connaissances à partir de données pour l'aide à l'interprétation clinique de la marche digitigrade ». Thèse de doct. Université de Valenciennes et du Hainaut-Cambresis, 2005.

[9] Arash Atrsaei et al. « Toward a Remote Assessment of Walking Bout and Speed : Application in Patients With Multiple Sclerosis ». In : *IEEE Journal of Biomedical and Health Informatics* 25.11 (2021), p. 4217-4228. doi : 10.1109/JBHI.2021.3076707 [8].

[10] Korinna Bade et Andreas Nürnberger. « Creating a cluster hierarchy under constraints of a partially known hierarchy ». In : *Proceedings of the 2008 SIAM international conference on data mining*. SIAM. 2008, p. 13-24.

[11] Amir Baghdadi et al. « Monitoring worker fatigue using wearable devices : A case study to detect changes in gait parameters ». In : *Journal of Quality Technology* 53.1 (2019), p. 47-71. doi : 10.1080/00224065.2019.1640097 [9].

[12] Liang Bai, Jiye Liang et Fuyuan Cao. « A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters ». In : *Information Fusion* 61 (2020), p. 36-47. issn : 1566-2535. doi : https://doi.org/10.1016/j.inffus.2020.03.009 [10].

[13] Jens Barth et al. « Stride Segmentation during Free Walk Movements Using Multi-Dimensional Subsequence Dynamic Time Warping on Inertial Sensor Data ». In : *Sensors* 15.3 (2015), p. 6419-6440. issn : 1424-8220. doi : 10.3390/s150306419 [11].

[14] Ann D. Bass et al. « Effect of Multiple Sclerosis on Daily Activities, Emotional Well-being, and Relationships : The Global vsMS Survey ». In : *International Journal of MS Care* 22.4 (août 2019), p. 158-164. issn : 1537-2073. doi : 10.7224/1537-2073.2018-087 [12].

[15] Janina Behrens et al. « Using perceptive computing in multiple sclerosis - the Short Maximum Speed Walk test ». In : *Journal of NeuroEngineering and Reha-*

8. https://doi.org/10.1109/JBHI.2021.3076707
9. https://doi.org/10.1080/00224065.2019.1640097
10. https://doi.org/https://doi.org/10.1016/j.inffus.2020.03.009
11. https://doi.org/10.3390/s150306419
12. https://doi.org/10.7224/1537-2073.2018-087

bilitation 11.1 (mai 2014), p. 89. ISSN : 1743-0003. DOI : 10.1186/1743-0003-11-89 [13].

[16] Lise BELLANGER, Arthur COULON et Philippe HUSI. « PerioClust : A Simple Hierarchical Agglomerative Clustering Approach Including Constraints ». In : *Data Analysis and Rationality in a Complex World*. Sous la dir. de Theodore CHADJIPADELIS et al. Springer International Publishing. Cham, 2021, p. 1-8.

[17] Francois BETHOUX et Susan BENNETT. « Evaluating Walking in Patients with Multiple Sclerosis : Which Assessment Tools Are Useful in Clinical Practice ? » In : *International Journal of MS Care* 13.1 (mar. 2011), p. 4-14. DOI : 10.7224/1537-2073-13.1.4 [14].

[18] Akash Kumar BHOI. « Classification and clustering of Parkinson's and healthy control gait dynamics using LDA and K-means ». In : *International Journal Bioautomation* 21.1 (2017), p. 19.

[19] Michalina BŁAŻKIEWICZ, Karol LANN VEL LACE et Anna HADAMUS. « Gait Symmetry Analysis Based on Dynamic Time Warping ». In : *Symmetry* 13.5 (2021). ISSN : 2073-8994. DOI : 10.3390/sym13050836 [15].

[20] B BOUISSET et B MATON. *Muscles et mouvement : base et applications de méthode élécromyographique*. 1995.

[21] Raphaël BRARD et al. « A Novel Walking Activity Recognition Model for Rotation Time Series Collected by a Wearable Sensor in a Free-Living Environment ». In : *Sensors* 22.9 (2022). ISSN : 1424-8220. DOI : 10.3390/s22093555 [16].

[22] F. BUGANÉ et al. « Estimation of spatial-temporal gait parameters in level walking based on a single accelerometer : Validation on normal subjects by standard gait analysis ». In : *Computer Methods and Programs in Biomedicine* 108.1 (2012), p. 129-137. ISSN : 0169-2607. DOI : https://doi.org/10.1016/j.cmpb.2012.02.003 [17].

13. https://doi.org/10.1186/1743-0003-11-89
14. https://doi.org/10.7224/1537-2073-13.1.4
15. https://doi.org/10.3390/sym13050836
16. https://doi.org/10.3390/s22093555
17. https://doi.org/https://doi.org/10.1016/j.cmpb.2012.02.003

[23] J. Cabrera et G.S. Watson. « On a spherical median related distribution ». en. In : *Communications in Statistics - Theory and Methods* 19.6 (jan. 1990), p. 1973-1986. ISSN : 0361-0926, 1532-415X. DOI : 10.1080/03610929008830303 [18].

[24] Rafael Caldas et al. « A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms ». In : *Gait & Posture* 57 (2017), p. 204-210. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2017.06.019 [19].

[25] Michelle H. Cameron et Joanne M. Wagner. « Gait Abnormalities in Multiple Sclerosis : Pathogenesis, Evaluation, and Advances in Treatment ». In : *Current Neurology and Neuroscience Reports* 11.5 (juil. 2011), p. 507. ISSN : 1534-6293. DOI : 10.1007/s11910-011-0214-y [20].

[26] Aurelio Cappozzo. « Gait analysis methodology ». In : *Human Movement Science* 3.1 (1984), p. 27-50. ISSN : 0167-9457. DOI : https://doi.org/10.1016/0167-9457(84)90004-6 [21].

[27] Malika Charrad et al. « NbClust : An R Package for determining the relevant number of clusters in a data set ». In : *Journal of Statistical Software* 61.6 (2014), p. 1-36.

[28] Lei Chen et Raymond Ng. « On the Marriage of Lp-Norms and Edit Distance ». In : *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*. VLDB '04. Toronto, Canada : VLDB Endowment, 2004, p. 792-803. ISBN : 0120884690.

[29] Lei Chen, M. Tamer Özsu et Vincent Oria. « Robust and Fast Similarity Search for Moving Object Trajectories ». In : *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. SIGMOD '05. Baltimore, Maryland : Association for Computing Machinery, 2005, p. 491-502. ISBN : 1595930604. DOI : 10.1145/1066157.1066213 [22].

18. https://doi.org/10.1080/03610929008830303
19. https://doi.org/https://doi.org/10.1016/j.gaitpost.2017.06.019
20. https://doi.org/10.1007/s11910-011-0214-y
21. https://doi.org/https://doi.org/10.1016/0167-9457(84)90004-6
22. https://doi.org/10.1145/1066157.1066213

[30] Jacob Cohen. « A Coefficient of Agreement for Nominal Scales ». In : *Educational and Psychological Measurement* 20.1 (1960), p. 37-46. DOI : 10.1177/001316446002000104[23].

[31] Mikael Cohen et al. « Should we still only rely on EDSS to evaluate disability in multiple sclerosis patients ? A study of inter and intra rater reliability ». In : *Multiple Sclerosis and Related Disorders* 54 (2021), p. 103144. ISSN : 2211-0348. DOI : https://doi.org/10.1016/j.msard.2021.103144[24].

[32] John H Conway et Derek A Smith. *On quaternions and octonions : their geometry, arithmetic, and symmetry*. AK Peters/CRC Press, 2003.

[33] Antoine Cornuéjols et al. « Collaborative clustering : Why, when, what and how ». In : *Information Fusion* 39 (2018), p. 81-95. ISSN : 1566-2535. DOI : https://doi.org/10.1016/j.inffus.2017.04.008[25].

[34] Jordan J. Craig et al. « Instrumented balance and walking assessments in persons with multiple sclerosis show strong test-retest reliability ». In : *Journal of NeuroEngineering and Rehabilitation* 14.1 (mai 2017), p. 43. ISSN : 1743-0003. DOI : 10.1186/s12984-017-0251-0[26].

[35] Robert G Cutlip et al. « Evaluation of an instrumented walkway for measurement of the kinematic parameters of gait ». In : *Gait & Posture* 12.2 (2000), p. 134-138. ISSN : 0966-6362. DOI : https://doi.org/10.1016/S0966-6362(00)00062-X[27].

[36] Gary R. Cutter et al. « Development of a multiple sclerosis functional composite as a clinical trial outcome measure ». In : *Brain* 122.5 (mai 1999), p. 871-882. ISSN : 0006-8950. DOI : 10.1093/brain/122.5.871[28].

[37] Julia Dannenmaier et al. « Application of functional data analysis to explore movements : walking, running and jumping - A systematic review ». In : *Gait & Posture* 77 (2020), p. 182-189. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2020.02.002[29].

23. https://doi.org/10.1177/001316446002000104
24. https://doi.org/https://doi.org/10.1016/j.msard.2021.103144
25. https://doi.org/https://doi.org/10.1016/j.inffus.2017.04.008
26. https://doi.org/10.1186/s12984-017-0251-0
27. https://doi.org/https://doi.org/10.1016/S0966-6362(00)00062-X
28. https://doi.org/10.1093/brain/122.5.871
29. https://doi.org/https://doi.org/10.1016/j.gaitpost.2020.02.002

[38] Martin Daumer et al. « Steps towards a miniaturized, robust and autonomous measurement device for the long-term monitoring of patient activity : ActiBelt ». In : 52.1 (2007), p. 149-155. doi : doi:10.1515/BMT.2007.028 [30].

[39] Derya Dinler et Mustafa Kemal Tural. « A survey of constrained clustering ». In : *Unsupervised learning algorithms*. Springer, 2016, p. 207-235.

[40] Emer P. Doheny, Timothy G. Foran et Barry R. Greene. « A single gyroscope method for spatial gait analysis ». In : *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. 2010, p. 1300-1303. doi : 10.1109/IEMBS.2010.5626397 [31].

[41] Elham Dolatabadi et al. « Mixture-Model Clustering of Pathological Gait Patterns ». In : *IEEE Journal of Biomedical and Health Informatics* 21.5 (2017), p. 1297-1305. doi : 10.1109/JBHI.2016.2633000 [32].

[42] J. C. Dunn. « A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters ». In : *Journal of Cybernetics* 3.3 (1973), p. 32-57. doi : 10.1080/01969727308546046 [33].

[43] Sir William Rowan Hamilton LL.D. P.R.I.A. F.R.A.S. Hon. M.R.Soc. Ed. et Dub. « LXXVIII. On quaternions ; or on a new system of imaginaries in Algebra ». In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 25.169 (1844), p. 489-495. doi : 10.1080/14786444408645047 [34].

[44] Y Escoufier. « Operators related to a data matrix ». In : *Recent Developments in Statistics* (1977), p. 125-131.

[45] Jill S Fischer et al. « Multiple Sclerosis Functional Composite (MSFC) : administration and scoring manual ». In : *New York : National Multiple Sclerosis Society* (2001).

[46] Felix Flachenecker et al. « Objective sensor-based gait measures reflect motor impairment in multiple sclerosis patients : reliability and clinical validation of a wearable sensor device ». In : *Multiple Sclerosis and Related Disorders* 39 (2020), p. 101903.

30. https://doi.org/doi:10.1515/BMT.2007.028
31. https://doi.org/10.1109/IEMBS.2010.5626397
32. https://doi.org/10.1109/JBHI.2016.2633000
33. https://doi.org/10.1080/01969727308546046
34. https://doi.org/10.1080/14786444408645047

[47] A. Forner-Cordero, H.J.F.M. Koopman et F.C.T. van der Helm. « Describing gait as a sequence of states ». In : *Journal of Biomechanics* 39.5 (2006), p. 948-957. ISSN : 0021-9290. DOI : https://doi.org/10.1016/j.jbiomech.2005.01.019[35].

[48] « Functional Nonparametric Unsupervised Classification ». In : *Nonparametric Functional Data Analysis : Theory and Practice*. New York, NY : Springer New York, 2006, p. 125-147. ISBN : 978-0-387-36620-3. DOI : 10.1007/0-387-36620-2_9[36]. URL : https://doi.org/10.1007/0-387-36620-2_9.

[49] Tal Galili. « dendextend : an R package for visualizing, adjusting and comparing trees of hierarchical clustering ». In : *Bioinformatics* 31.22 (juil. 2015), p. 3718-3720. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btv428[37].

[50] Ignacio Ghersi et al. « Gait-cycle segmentation method based on lower-trunk acceleration signals and dynamic time warping ». In : *Medical Engineering & Physics* 82 (2020), p. 70-77. ISSN : 1350-4533. DOI : https://doi.org/10.1016/j.medengphy.2020.06.001[38].

[51] Uri Givon, Gabriel Zeilig et Anat Achiron. « Gait analysis in multiple sclerosis : Characterization of temporal–spatial parameters using GAITRite functional ambulation system ». In : *Gait & Posture* 29.1 (2009), p. 138-142. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2008.07.011[39].

[52] Myla D Goldman, Ruth Ann Marrie et Jeffrey A Cohen. « Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls ». In : *Multiple Sclerosis Journal* 14.3 (2008), p. 383-390. DOI : 10.1177/1352458507082607[40].

[53] Myla D. Goldman, Robert W. Motl et Richard A. Rudick. « Possible clinical outcome measures for clinical trials in patients with multiple sclerosis ». In : *Therapeutic Advances in Neurological Disorders* 3.4 (2010), p. 229-239. DOI : 10.1177/1756285610374117[41].

35. https://doi.org/https://doi.org/10.1016/j.jbiomech.2005.01.019
36. https://doi.org/10.1007/0-387-36620-2_9
37. https://doi.org/10.1093/bioinformatics/btv428
38. https://doi.org/https://doi.org/10.1016/j.medengphy.2020.06.001
39. https://doi.org/https://doi.org/10.1016/j.gaitpost.2008.07.011
40. https://doi.org/10.1177/1352458507082607
41. https://doi.org/10.1177/1756285610374117

[54] Juan Carlos González Islas, Omar Arturo Domínguez-Ramírez et Omar López Ortega. « Biped Gait Analysis based on Forward Kinematics Modeling using Quaternions Algebra ». In : *Mexican Journal of Biomedical Engineering* 41.3 (déc. 2020), p. 56-71.

[55] John C Gower. « A general coefficient of similarity and some of its properties ». In : *Biometrics* (1971), p. 857-871.

[56] James E. Graham et al. « Assessing walking speed in clinical research : a systematic review ». In : *Journal of Evaluation in Clinical Practice* 14.4 (2008), p. 552-562. doi : https://doi.org/10.1111/j.1365-2753.2007.00917.x [42]. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2753.2007.00917.x.

[57] Nizar Grira, Michel Crucianu et Nozha Boujemaa. « Unsupervised and Semi-supervised Clustering : a brief survey ». In : *A Review of Machine Learning Techniques for Processing Multimedia Content* (sept. 2005).

[58] L. Gupta et al. « Nonlinear alignment and averaging for estimating the evoked potential ». In : *IEEE Transactions on Biomedical Engineering* 43.4 (1996), p. 348-356. doi : 10.1109/10.486255 [43].

[59] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology*. Cambridge University Press, 1997. doi : 10.1017/CBO9780511574931.025 [44].

[60] Tomasz Hachaj et al. « Averaging Three-Dimensional Time-Varying Sequences of Rotations : Application to Preprocessing of Motion Capture Data ». In : *Image Analysis*. Sous la dir. de Puneet Sharma et Filippo Maria Bianchi. Cham : Springer International Publishing, 2017, p. 17-28. isbn : 978-3-319-59126-1.

[61] Nooshin Haji Ghassemi et al. « Segmentation of Gait Sequences in Sensor-Based Movement Analysis : A Comparison of Methods in Parkinson's Disease ». In : *Sensors* 18.1 (2018). issn : 1424-8220. doi : 10.3390/s18010145 [45].

[62] Maria Halkidi, Yannis Batistakis et Michalis Vazirgiannis. « On Clustering Validation Techniques ». In : *Journal of Intelligent Information Systems* 17.2 (déc. 2001), p. 107-145. issn : 1573-7675. doi : 10.1023/A:1012801612483 [46].

42. https://doi.org/https://doi.org/10.1111/j.1365-2753.2007.00917.x
43. https://doi.org/10.1109/10.486255
44. https://doi.org/10.1017/CBO9780511574931.025
45. https://doi.org/10.3390/s18010145
46. https://doi.org/10.1023/A:1012801612483

[63] J. Han et Bir Bhanu. « Individual recognition using gait energy image ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.2 (2006), p. 316-322. doi : 10.1109/TPAMI.2006.38 [47].

[64] Gerald F. Harris et Jacqueline J. Wertsch. « Procedures for gait analysis ». In : *Archives of Physical Medicine and Rehabilitation* 75.2 (1994), p. 216-225. issn : 0003-9993. doi : https://doi.org/10.1016/0003-9993(94)90399-9 [48].

[65] J. A. Hartigan. « Consistency of Single Linkage for High-Density Clusters ». In : *Journal of the American Statistical Association* 76.374 (1981), p. 388-394. doi : 10.1080/01621459.1981.10477658 [49].

[66] J. A. Hartigan. « Consistency of Single Linkage for High-Density Clusters ». In : *Journal of the American Statistical Association* 76.374 (1981), p. 388-394. doi : 10.1080/01621459.1981.10477658 [50].

[67] C Heesen et al. « Patient perception of bodily functions in multiple sclerosis : gait and visual function are the most valuable ». In : *Multiple Sclerosis Journal* 14.7 (2008), p. 988-991. doi : 10.1177/1352458508088916 [51].

[68] Nathaniel E. Helwig et al. « Methods to temporally align gait cycle data ». In : *Journal of Biomechanics* 44.3 (2011), p. 561-566. issn : 0021-9290.

[69] J. C. Hobart et al. « Measuring the impact of MS on walking ability ». In : *Neurology* 60.1 (2003), p. 31-36. doi : 10.1212/WNL.60.1.31 [52].

[70] James H. Horne et Sallie L. Baliunas. « A Prescription for Period Analysis of Unevenly Sampled Time Series ». In : *The Astrophysical Journal* 302 (mar. 1986), p. 757. doi : 10.1086/164037 [53].

[71] Adam M. Howell et al. « Kinetic Gait Analysis Using a Low-Cost Insole ». In : *IEEE Transactions on Biomedical Engineering* 60.12 (2013), p. 3284-3290. doi : 10.1109/TBME.2013.2250972 [54].

47. https://doi.org/10.1109/TPAMI.2006.38
48. https://doi.org/https://doi.org/10.1016/0003-9993(94)90399-9
49. https://doi.org/10.1080/01621459.1981.10477658
50. https://doi.org/10.1080/01621459.1981.10477658
51. https://doi.org/10.1177/1352458508088916
52. https://doi.org/10.1212/WNL.60.1.31
53. https://doi.org/10.1086/164037
54. https://doi.org/10.1109/TBME.2013.2250972

[72] Wei-Chun Hsu et al. « Multiple-Wearable-Sensor-Based Gait Classification and Analysis in Patients with Neurological Disorders ». In : *Sensors* 18.10 (2018). ISSN : 1424-8220. DOI : 10.3390/s18103397 [55].

[73] Zhexue Huang. « Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values ». In : *Data Mining and Knowledge Discovery* 2.3 (sept. 1998), p. 283-304. ISSN : 1573-756X. DOI : 10.1023/A:1009769707641 [56].

[74] Lawrence Hubert et Phipps Arabie. « Comparing partitions ». In : *Journal of classification* 2.1 (1985), p. 193-218.

[75] Audrey Hulot et al. « Fast tree aggregation for consensus hierarchical clustering ». In : *BMC Bioinformatics* 21.1 (mar. 2020), p. 120. ISSN : 1471-2105.

[76] Sandra R Hundza et al. « Accurate and reliable gait cycle detection in Parkinson's disease ». In : *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.1 (2013), p. 127-137.

[77] Du Q. Huynh. « Metrics for 3D Rotations : Comparison and Analysis ». In : *Journal of Mathematical Imaging and Vision* 35.2 (oct. 2009), p. 155-164. ISSN : 1573-7683. DOI : 10.1007/s10851-009-0161-2 [57].

[78] Francesca Ieva et al. « Multivariate functional clustering for the morphological analysis of electrocardiograph curves ». In : *Journal of the Royal Statistical Society : Series C (Applied Statistics)* 62.3 (2013), p. 401-418. DOI : https://doi.org/10.1111/j.1467-9876.2012.01062.x [58].

[79] Hernan Inojosa, Dirk Schriefer et Tjalf Ziemssen. « Clinical outcome measures in multiple sclerosis : A review ». In : *Autoimmunity Reviews* 19.5 (2020), p. 102512. ISSN : 1568-9972. DOI : https://doi.org/10.1016/j.autrev.2020.102512 [59].

[80] Hernan Inojosa, Dirk Schriefer et Tjalf Ziemssen. « Clinical outcome measures in multiple sclerosis : A review ». In : *Autoimmunity Reviews* 19.5 (2020), p. 102512. ISSN : 1568-9972. DOI : https://doi.org/10.1016/j.autrev.2020.102512 [60].

55. https://doi.org/10.3390/s18103397
56. https://doi.org/10.1023/A:1009769707641
57. https://doi.org/10.1007/s10851-009-0161-2
58. https://doi.org/https://doi.org/10.1111/j.1467-9876.2012.01062.x
59. https://doi.org/https://doi.org/10.1016/j.autrev.2020.102512
60. https://doi.org/https://doi.org/10.1016/j.autrev.2020.102512

[81] Marco Iosa et al. « Wearable inertial sensors for human movement analysis ». In : *Expert Review of Medical Devices* 13.7 (2016), p. 641-659. DOI : 10.1080/17434440.2016.1198694[61].

[82] Bartosz Jablonski. « Quaternion dynamic time warping ». In : *IEEE transactions on signal processing* 60.3 (2011), p. 1174-1183.

[83] Julien Jacques et Cristian Preda. « Functional data clustering : a survey ». In : *Advances in Data Analysis and Classification* 8 (sept. 2014), p. 231-255. ISSN : 1862-5355. DOI : 10.1007/s11634-013-0158-y[62].

[84] S Rao Jammalamadaka et Ashis SenGupta. *Topics in Circular Statistics*. WORLD SCIENTIFIC, 2001. DOI : 10.1142/4031[63]. eprint : https://www.worldscientific.com/doi/pdf/10.1142/4031.

[85] Shuo Jiang et al. « A robust algorithm for gait cycle segmentation ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017, p. 31-35. DOI : 10.23919/EUSIPCO.2017.8081163[64].

[86] Parinaz Kasebzadeh, Gustaf Hendeby et Fredrik Gustafsson. « Asynchronous Averaging of Gait Cycles for Classification of Gait and Device Modes ». In : *IEEE Sensors Journal* 21.1 (2021), p. 529-538. DOI : 10.1109/JSEN.2020.3014189[65].

[87] Leonard Kaufman et Peter J Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. T. 344. John Wiley & Sons, 1990.

[88] Justin J. Kavanagh et Hylton B. Menz. « Accelerometry : A technique for quantifying movement patterns during walking ». In : *Gait & Posture* 28.1 (juil. 2008), p. 1-15. ISSN : 0966-6362. DOI : 10.1016/j.gaitpost.2007.10.010[66].

[89] Alexander M. Keppler et al. « Validity of accelerometry in step detection and gait speed measurement in orthogeriatric patients ». In : *PLOS ONE* 14 (août 2019), p. 1-11. DOI : 10.1371/journal.pone.0221732[67].

61. https://doi.org/10.1080/17434440.2016.1198694
62. https://doi.org/10.1007/s11634-013-0158-y
63. https://doi.org/10.1142/4031
64. https://doi.org/10.23919/EUSIPCO.2017.8081163
65. https://doi.org/10.1109/JSEN.2020.3014189
66. https://doi.org/10.1016/j.gaitpost.2007.10.010
67. https://doi.org/10.1371/journal.pone.0221732

[90] Siddhartha KHANDELWAL et Nicholas WICKSTRÖM. « Novel methodology for estimating Initial Contact events from accelerometers positioned at different body locations ». In : *Gait & Posture* 59 (2018), p. 278-285. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2017.07.030 [68].

[91] Bernd C. KIESEIER et Carlo POZZILLI. « Assessing walking disability in multiple sclerosis ». In : *Multiple Sclerosis Journal* 18.7 (2012). PMID : 22740603, p. 914-924. DOI : 10.1177/1352458512444498 [69].

[92] C.Maria KIM et Janice J. ENG. « Magnitude and pattern of 3D kinematic and kinetic gait profiles in persons with stroke : relationship to walking speed ». In : *Gait & Posture* 20.2 (2004), p. 140-146. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2003.07.002 [70].

[93] Naoki KITAGAWA et Naomichi OGIHARA. « Estimation of foot trajectory during human walking by a wearable inertial measurement unit mounted to the foot ». In : *Gait & Posture* 45 (2016), p. 110-114. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2016.01.014 [71].

[94] Isabella KLÖPFER-KRÄMER et al. « Gait analysis – Available platforms for outcome assessment ». In : *Injury* 51 (2020). Optimizing Patient Function After Musculoskeletal Trauma, S90-S96. ISSN : 0020-1383. DOI : https://doi.org/10.1016/j.injury.2019.11.011 [72].

[95] Clyde Young KRAMER. « Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications ». In : *Biometrics* 12.3 (1956), p. 307-310.

[96] R. KRISHNAPURAM et J.M. KELLER. « The possibilistic C-means algorithm : insights and recommendations ». In : *IEEE Transactions on Fuzzy Systems* 4.3 (1996), p. 385-393. DOI : 10.1109/91.531779 [73].

[97] John F KURTZKE. « Rating neurologic impairment in multiple sclerosis : an expanded disability status scale (EDSS) ». In : *Neurology* 33.11 (1983), p. 1444-1444.

68. https://doi.org/https://doi.org/10.1016/j.gaitpost.2017.07.030
69. https://doi.org/10.1177/1352458512444498
70. https://doi.org/https://doi.org/10.1016/j.gaitpost.2003.07.002
71. https://doi.org/https://doi.org/10.1016/j.gaitpost.2016.01.014
72. https://doi.org/https://doi.org/10.1016/j.injury.2019.11.011
73. https://doi.org/10.1109/91.531779

[98] Thomas Lampert et al. « Constrained distance based clustering for time-series : a comparative and experimental study ». In : *Data Mining and Knowledge Discovery* 32.6 (2018), p. 1663-1707.

[99] G. N. Lance et W. T. Williams. « A General Theory of Classificatory Sorting Strategies ». In : *The Computer Journal* 9.4 (1967), p. 373-380. issn : 0010-4620.

[100] Nicholas G. LaRocca. « Impact of Walking Impairment in Multiple Sclerosis ». In : *The Patient : Patient-Centered Outcomes Research* 4.3 (sept. 2011), p. 189-201. issn : 1178-1661. doi : 10.2165/11591150-000000000-00000 [74].

[101] Christine Lavit et al. « The act (statis method) ». In : *Computational Statistics & Data Analysis* 18.1 (1994), p. 97-119.

[102] L Eduardo Cofré Lizama et al. « The use of laboratory gait analysis for understanding gait deterioration in people with multiple sclerosis ». In : *Multiple Sclerosis Journal* 22.14 (2016), p. 1768-1776. doi : 10.1177/1352458516658137 [75].

[103] Jana Lizrova Preiningerova et al. « Spatial and temporal characteristics of gait as outcome measures in multiple sclerosis (EDSS 0 to 6.5) ». In : *Journal of NeuroEngineering and Rehabilitation* 12.1 (fév. 2015), p. 14. issn : 1743-0003. doi : 10.1186/s12984-015-0001-0 [76].

[104] João Loureiro et Paulo Lobato Correia. « Using a Skeleton Gait Energy Image for Pathological Gait Classification ». In : *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 2020, p. 503-507. doi : 10.1109/FG47880.2020.00064 [77].

[105] Xiaofei Ma et Satya Dhavala. *Hierarchical Clustering with Prior Knowledge*. https://arxiv.org/abs/1806.03432. 2018. arXiv : 1806.03432 [stat.ML] [78].

[106] J. Macqueen. « Some methods for classification and analysis of multivariate observations ». In : *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. 1967, p. 281-297.

74. https://doi.org/10.2165/11591150-000000000-00000
75. https://doi.org/10.1177/1352458516658137
76. https://doi.org/10.1186/s12984-015-0001-0
77. https://doi.org/10.1109/FG47880.2020.00064
78. https://arxiv.org/abs/1806.03432

[107] Avril Mansfield et Gerard M Lyons. « The use of accelerometry to detect heel contact events for use as a sensor in FES assisted walking ». In : *Medical Engineering & Physics* 25.10 (2003), p. 879-885. ISSN : 1350-4533. DOI : https://doi.org/10.1016/S1350-4533(03)00116-4 [79].

[108] F. Landis Markley et al. « Averaging Quaternions ». In : *Journal of Guidance, Control, and Dynamics* 30.4 (2007), p. 1193-1197. DOI : 10.2514/1.28949 [80].

[109] J. S. Marron et al. « Functional Data Analysis of Amplitude and Phase Variation ». In : *Statistical Science* 30.4 (2015), p. 468-484. DOI : 10.1214/15-STS524 [81].

[110] John McCamley et al. « An enhanced estimate of initial contact and final contact instants of time using lower trunk inertial sensor data ». In : *Gait & Posture* 36.2 (2012), p. 316-318. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2012.02.019 [82].

[111] Sandra Meyer-Moock et al. « Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis ». In : *BMC neurology* 14.1 (2014), p. 58.

[112] Boris Mirkin. *Clustering for data mining : a data recovery approach*. Chapman et Hall/CRC, 2005.

[113] Yaejin Moon et al. « Monitoring gait in multiple sclerosis with novel wearable motion sensors ». In : *PLOS ONE* 12.2 (fév. 2017), p. 1-19. DOI : 10.1371/journal.pone.0171346 [83].

[114] Robert W Motl et al. « Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis ». In : *Multiple Sclerosis Journal* 23.5 (2017), p. 704-710.

[115] Robert W. Motl et al. « Accuracy of the actibelt accelerometer for measuring walking speed in a controlled environment among persons with multiple sclerosis ». In : *Gait & Posture* 35.2 (2012), p. 192-196. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2011.09.005 [84].

79. https://doi.org/https://doi.org/10.1016/S1350-4533(03)00116-4
80. https://doi.org/10.2514/1.28949
81. https://doi.org/10.1214/15-STS524
82. https://doi.org/https://doi.org/10.1016/j.gaitpost.2012.02.019
83. https://doi.org/10.1371/journal.pone.0171346
84. https://doi.org/https://doi.org/10.1016/j.gaitpost.2011.09.005

[116] Meinard Müller. « Dynamic Time Warping ». In : *Information Retrieval for Music and Motion*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 69-84. ISBN : 978-3-540-74048-3. DOI : 10.1007/978-3-540-74048-3_4 [85].

[117] Roy Müller et al. « Wearable inertial sensors are highly sensitive in the detection of gait disturbances and fatigue at early stages of multiple sclerosis ». In : *BMC Neurology* 21.1 (sept. 2021), p. 337. ISSN : 1471-2377. DOI : 10.1186/s12883-021-02361-y [86].

[118] Sara Mulroy et al. « Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke ». In : *Gait & Posture* 18.1 (2003), p. 114-125. ISSN : 0966-6362. DOI : https://doi.org/10.1016/S0966-6362(02)00165-0 [87].

[119] Alvaro Muro-De-La-Herran, Begonya Garcia-Zapirain et Amaia Mendez-Zorrilla. « Gait analysis methods : An overview of wearable and non-wearable systems, highlighting clinical applications ». In : *Sensors* 14.2 (2014), p. 3362-3394.

[120] Fabián Narváez, Fernando Árbito et Ricardo Proaño. « A Quaternion-Based Method to IMU-to-Body Alignment for Gait Analysis ». In : *Digital Human Modeling. Applications in Health, Safety, Ergonomics, and Risk Management*. Sous la dir. de Vincent G. Duffy. Cham : Springer International Publishing, 2018, p. 217-231.

[121] Vit Niennattrakul et Chotirat Ann Ratanamahatana. « On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping ». In : *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*. 2007, p. 733-738. DOI : 10.1109/MUE.2007.165 [88].

[122] M M Nieuwenhuis et al. « The Six Spot Step Test : a new measurement for walking ability in multiple sclerosis ». In : *Multiple Sclerosis Journal* 12.4 (2006). PMID : 16900764, p. 495-500. DOI : 10.1191/1352458506ms1293oa [89].

[123] Daniel Ontaneda et al. « Progressive multiple sclerosis : prospects for disease therapy, repair, and restoration of function ». In : *The Lancet* 389.10076 (2017), p. 1357-1366.

85. https://doi.org/10.1007/978-3-540-74048-3_4
86. https://doi.org/10.1186/s12883-021-02361-y
87. https://doi.org/https://doi.org/10.1016/S0966-6362(02)00165-0
88. https://doi.org/10.1109/MUE.2007.165
89. https://doi.org/10.1191/1352458506ms1293oa

[124] Julio-Omar Palacio-Niño et Fernando Berzal. *Evaluation Metrics for Unsupervised Learning Algorithms*. 2019. arXiv : 1905.05667 [cs.LG] [90].

[125] Massimiliano Pau et al. « Clinical assessment of gait in individuals with multiple sclerosis using wearable inertial sensors : Comparison with patient-based measure ». In : *Multiple sclerosis and related disorders* 10 (2016), p. 187-191.

[126] J. M Peña, J. A Lozano et P Larrañaga. « An empirical comparison of four initialization methods for the K-Means algorithm ». en. In : *Pattern Recognition Letters* 20.10 (oct. 1999), p. 1027-1040. issn : 0167-8655. doi : 10.1016/S0167-8655(99)00069-0 [91].

[127] Jacquelin Perry et Judith M Burnfield. « Gait analysis. Normal and pathological function 2nd ed ». In : *California : Slack* (2010).

[128] François Petitjean, Alain Ketterlin et Pierre Gançarski. « A global averaging method for dynamic time warping, with applications to clustering ». In : *Pattern Recognition* 44.3 (2011), p. 678-693.

[129] Pietro Picerno et al. « Wearable inertial sensors for human movement analysis : a five-year update ». In : *Expert Review of Medical Devices* 18.sup1 (2021). PMID : 34601995, p. 79-94. doi : 10.1080/17434440.2021.1988849 [92]. eprint : https://doi.org/10.1080/17434440.2021.1988849.

[130] Julien Ah-Pine et Xinyu Wang. « Similarity Based Hierarchical Clustering with an Application to Text Collections ». In : *Advances in Intelligent Data Analysis XV*. Sous la dir. d'Henrik Boström et al. Springer International Publishing. Cham, 2016, p. 320-331. isbn : 978-3-319-46349-0.

[131] Michał Piórek. « Analysis of Chaos for Quaternion Time Series ». In : *Analysis of Chaotic Behavior in Non-linear Dynamical Systems*. Springer, 2019, p. 73-88.

[132] Michał Piórek et Bartosz Jabłoński. « A quaternion clustering framework ». In : *International Journal of Applied Mathematics and Computer Science* 30.1 (2020), p. 133-147. issn : 1641-876X.

90. https://arxiv.org/abs/1905.05667
91. https://doi.org/10.1016/S0167-8655(99)00069-0
92. https://doi.org/10.1080/17434440.2021.1988849

[133] Diane Podsiadlo et Sandra Richardson. « The Timed "Up & Go" : A Test of Basic Functional Mobility for Frail Elderly Persons ». In : *Journal of the American Geriatrics Society* 39.2 (1991), p. 142-148. doi : https://doi.org/10.1111/j.1532-5415.1991.tb01616.x [93].

[134] Irene Pulido-Valdeolivas et al. « Gait phenotypes in paediatric hereditary spastic paraplegia revealed by dynamic time warping analysis and random forests ». In : *PLOS ONE* 13 (mar. 2018), p. 1-28. doi : 10.1371/journal.pone.0192345 [94].

[135] J. Ramsay et al. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005. isbn : 978-0-387-40080-8.

[136] William M. Rand. « Objective Criteria for the Evaluation of Clustering Methods ». In : *Journal of the American Statistical Association* 66.336 (1971), p. 846-850. doi : 10.1080/01621459.1971.10482356 [95].

[137] Chotirat Ratanamahatana et al. « A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering ». en. In : *Advances in Knowledge Discovery and Data Mining*. Sous la dir. de Tu Bao Ho, David Cheung et Huan Liu. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer, 2005, p. 771-777. isbn : 978-3-540-31935-1. doi : 10.1007/11430919_90 [96].

[138] Miriam F. Reelick et al. « Increased intra-individual variability in stride length and reaction time in recurrent older fallers ». In : *Aging Clinical and Experimental Research* 23.5 (oct. 2011), p. 393-399. issn : 1720-8319. doi : 10.1007/BF03337764 [97].

[139] Nicolas Roche et al. « Categorization of gait patterns in adults with cerebral palsy : A clustering approach ». In : *Gait & Posture* 39.1 (2014), p. 235-240. issn : 0966-6362. doi : https://doi.org/10.1016/j.gaitpost.2013.07.110 [98].

[140] Peter J. Rousseeuw. « Silhouettes : A graphical aid to the interpretation and validation of cluster analysis ». In : *Journal of Computational and Applied Mathematics*

93. https://doi.org/https://doi.org/10.1111/j.1532-5415.1991.tb01616.x
94. https://doi.org/10.1371/journal.pone.0192345
95. https://doi.org/10.1080/01621459.1971.10482356
96. https://doi.org/10.1007/11430919_90
97. https://doi.org/10.1007/BF03337764
98. https://doi.org/https://doi.org/10.1016/j.gaitpost.2013.07.110

20 (1987), p. 53-65. ISSN : 0377-0427. DOI : https://doi.org/10.1016/0377-0427(87)90125-7 [99].

[141] Adam ROZUMALSKI et Michael H. SCHWARTZ. « Crouch gait patterns defined using k-means cluster analysis are related to underlying clinical pathology ». In : *Gait & Posture* 30.2 (2009), p. 155-160. ISSN : 0966-6362. DOI : https://doi.org/10.1016/j.gaitpost.2009.05.010 [100].

[142] Jan RUETERBORIES et al. « Methods for gait event detection and analysis in ambulatory systems ». In : *Medical Engineering & Physics* 32.6 (2010), p. 545-552. ISSN : 1350-4533. DOI : https://doi.org/10.1016/j.medengphy.2010.03.007 [101].

[143] Angelo Maria SABATINI. « Estimating Three-Dimensional Orientation of Human Body Parts by Inertial/Magnetic Sensing ». In : *Sensors* 11.2 (2011), p. 1489-1525. ISSN : 1424-8220. DOI : 10.3390/s110201489 [102].

[144] Laura M SANGALLI et al. « K-mean alignment for curve clustering ». In : *Computational Statistics & Data Analysis* 54.5 (2010), p. 1219-1233.

[145] Laura M. SANGALLI et al. « Functional clustering and alignment methods with applications ». In : *Communications in Applied and Industrial Mathematics* 1 (2010), p. 205-224.

[146] Gilbert SAPORTA. *Probabilités, analyse des données et statistique*. Sous la dir. de TECHNIP. Juil. 2011.

[147] Sampasetty SARAVANAN et Gulam Mohideen Kadhar NAWAZ. « Ensemble-based time series data clustering for high dimensional data ». In : *International Journal of Innovative Computing, Information and Control* 10.4 (2014), p. 1457-1470.

[148] Henry SCHEFFE. *The analysis of variance*. T. 72. John Wiley & Sons, 1999.

[149] Thomas SEEL, Jörg RAISCH et Thomas SCHAUER. « IMU-Based Joint Angle Measurement for Gait Analysis ». In : *Sensors* 14.4 (2014), p. 6891-6909. ISSN : 1424-8220. DOI : 10.3390/s140406891 [103].

99. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7
100. https://doi.org/https://doi.org/10.1016/j.gaitpost.2009.05.010
101. https://doi.org/https://doi.org/10.1016/j.medengphy.2010.03.007
102. https://doi.org/10.3390/s110201489
103. https://doi.org/10.3390/s140406891

[150] Shokri Z. Selim et M. A. Ismail. « K-Means-Type Algorithms : A Generalized Convergence Theorem and Characterization of Local Optimality ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.1 (jan. 1984). Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 81-87. ISSN : 1939-3539. DOI : 10.1109/TPAMI.1984.4767478 [104].

[151] Mariano Serrao et al. « Identification of specific gait patterns in patients with cerebellar ataxia, spastic paraplegia, and Parkinson's disease : A non-hierarchical cluster analysis ». In : *Human Movement Science* 57 (2018), p. 267-279. ISSN : 0167-9457. DOI : https://doi.org/10.1016/j.humov.2017.09.005 [105].

[152] Giacomo Severini et al. « Evaluation of Clinical Gait Analysis parameters in patients affected by Multiple Sclerosis : Analysis of kinematics ». In : *Clinical Biomechanics* 45 (2017), p. 1-8.

[153] Camille J. Shanahan et al. « Technologies for Advanced Gait and Balance Assessments in People with Multiple Sclerosis ». In : *Frontiers in Neurology* 8 (2018). ISSN : 1664-2295. DOI : 10.3389/fneur.2017.00708 [106].

[154] Yoichi Shimada et al. « Clinical Application of Acceleration Sensor to Detect the Swing Phase of Stroke Gait in Functional Electrical Stimulation ». In : *The Tohoku Journal of Experimental Medicine* 207.3 (2005), p. 197-202. DOI : 10.1620/tjem.207.197 [107].

[155] Ken Shoemake. « Animating Rotation with Quaternion Curves ». In : *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '85. Association for Computing Machinery. New York, NY, USA, 1985, p. 245-254. ISBN : 0897911660.

[156] Jamie Shotton et al. « Real-time human pose recognition in parts from single depth images ». In : *CVPR 2011*. 2011, p. 1297-1304. DOI : 10.1109/CVPR.2011.5995316 [108].

104. https://doi.org/10.1109/TPAMI.1984.4767478
105. https://doi.org/https://doi.org/10.1016/j.humov.2017.09.005
106. https://doi.org/10.3389/fneur.2017.00708
107. https://doi.org/10.1620/tjem.207.197
108. https://doi.org/10.1109/CVPR.2011.5995316

[157] M.M. Skelly et H.J. Chizeck. « Real-time gait event detection for paraplegic FES walking ». In : *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 9.1 (2001), p. 59-68. doi : 10.1109/7333.918277 [109].

[158] Robert R Sokal et F James Rohlf. « The comparison of dendrograms by objective methods ». In : *Taxon* (1962), p. 33-40.

[159] Jacob J Sosnoff et al. « Quantifying gait impairment in multiple sclerosis using GAITRite™ technology ». In : *Gait & posture* 34.1 (2011), p. 145-147.

[160] Jacob J. Sosnoff, Brian M. Sandroff et Robert W. Motl. « Quantifying gait abnormalities in persons with multiple sclerosis with minimal disability ». In : *Gait & Posture* 36.1 (2012), p. 154-156. issn : 0966-6362. doi : https://doi.org/10.1016/j.gaitpost.2011.11.027 [110].

[161] Sebastijan Sprager et Matjaz B. Juric. « Inertial Sensor-Based Gait Recognition : A Review ». In : *Sensors* 15.9 (2015), p. 22089-22127. issn : 1424-8220. doi : 10.3390/s150922089 [111].

[162] Sebastijan Šprager et Matjaž B. Jurič. « Robust Stride Segmentation of Inertial Signals Based on Local Cyclicity Estimation ». In : *Sensors* 18.4 (2018). issn : 1424-8220. doi : 10.3390/s18041091 [112].

[163] Tobias Steinmetzer et al. « Clustering of Human Gait with Parkinson's Disease by Using Dynamic Time Warping ». In : *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE. 2018, p. 1-6. doi : 10.1109/IWOBI.2018.8464203 [113].

[164] Akara Supratak et al. « Remote Monitoring in the Home Validates Clinical Gait Measures for Multiple Sclerosis ». In : *Frontiers in Neurology* 9 (2018). issn : 1664-2295. doi : 10.3389/fneur.2018.00561 [114].

[165] Adam Switonski, Henryk Josinski et Konrad Wojciechowski. « Dynamic time warping in classification and selection of motion capture data ». In : *Multidimensional Systems and Signal Processing* 30.3 (2019), p. 1437-1468.

109. https://doi.org/10.1109/7333.918277
110. https://doi.org/https://doi.org/10.1016/j.gaitpost.2011.11.027
111. https://doi.org/10.3390/s150922089
112. https://doi.org/10.3390/s18041091
113. https://doi.org/10.1109/IWOBI.2018.8464203
114. https://doi.org/10.3389/fneur.2018.00561

[166] Shigeru TADANO, Ryo TAKEDA et Hiroaki MIYAGAWA. « Three Dimensional Gait Analysis Using Wearable Acceleration and Gyro Sensors Based on Quaternion Calculations ». In : *Sensors* 13.7 (2013), p. 9321-9343. ISSN : 1424-8220. DOI : 10.3390/s130709321 [115].

[167] Shuichi TOKUSHIGE, Hiroshi YADOHISA et Koichi INADA. « Crisp and fuzzy k-means clustering algorithms for multivariate functional data ». In : *Computational Statistics* 22.1 (avr. 2007), p. 1-16. ISSN : 1613-9658. DOI : 10.1007/s00180-006-0013-0 [116].

[168] Diana TROJANIELLO et al. « Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors : application to elderly, hemiparetic, parkinsonian and choreic gait ». In : *Journal of NeuroEngineering and Rehabilitation* 11.1 (nov. 2014), p. 152. ISSN : 1743-0003. DOI : 10.1186/1743-0003-11-152 [117].

[169] Can TUNCA et al. « Inertial Sensor-Based Robust Gait Analysis in Non-Hospital Settings for Neurological Disorders ». In : *Sensors* 17.4 (2017). ISSN : 1424-8220. DOI : 10.3390/s17040825 [118].

[170] Carmen TUR et Xavier MONTALBAN. « Progressive MS trials : Lessons learned ». In : *Multiple Sclerosis Journal* 23.12 (2017). PMID : 29041872, p. 1583-1592. DOI : 10.1177/1352458517729460 [119].

[171] Simone VANTINI. « On the definition of phase and amplitude variability in functional data analysis ». In : *TEST* 21.4 (déc. 2012), p. 676-696. ISSN : 1863-8260. DOI : 10.1007/s11749-011-0268-9 [120].

[172] Iago VÁZQUEZ et al. « An ensemble solution for multivariate time series clustering ». In : *Neurocomputing* 457 (2021), p. 182-192. ISSN : 0925-2312. DOI : https://doi.org/10.1016/j.neucom.2020.09.093 [121].

115. https://doi.org/10.3390/s130709321
116. https://doi.org/10.1007/s00180-006-0013-0
117. https://doi.org/10.1186/1743-0003-11-152
118. https://doi.org/10.3390/s17040825
119. https://doi.org/10.1177/1352458517729460
120. https://doi.org/10.1007/s11749-011-0268-9
121. https://doi.org/https://doi.org/10.1016/j.neucom.2020.09.093

[173] P.H Veltink et al. « Three dimensional inertial sensing of foot movements for automatic tuning of a two-channel implantable drop-foot stimulator ». In : *Medical Engineering & Physics* 25.1 (2003). Control Issues of Functional Electrical Stimulation : Current and Future Systems, p. 21-28. ISSN : 1350-4533. DOI : https://doi.org/10.1016/S1350-4533(02)00041-3 [122].

[174] E. Vigneau et E. M. Qannari. « Clustering of Variables Around Latent Components ». In : *Communications in Statistics - Simulation and Computation* 32.4 (2003), p. 1131-1150. DOI : 10.1081/SAC-120023882 [123].

[175] M. Vlachos, D. Gunopulos et G. Kollios. « Discovering similar multidimensional trajectories ». In : *Proceedings 18th International Conference on Data Engineering* (2002), p. 673-684.

[176] John Voight. *Quaternion algebras*. Springer Nature, 2021.

[177] Kiri Wagstaff et al. « Constrained k-means clustering with background knowledge ». In : *Icml*. T. 1. 2001, p. 577-584.

[178] Xiaoyue Wang et al. « Experimental comparison of representation methods and distance measures for time series data ». en. In : *Data Mining and Knowledge Discovery* 26.2 (mar. 2013), p. 275-309. ISSN : 1573-756X. DOI : 10.1007/s10618-012-0250-5 [124]. (Visité le 06/07/2021).

[179] JN Whitaker et al. « Outcomes assessment in multiple sclerosis clinical trials : a critical analysis ». In : *Multiple Sclerosis Journal* 1.1 (1995), p. 37-47.

[180] M.W. Whittle. « 10 - Gait analysis ». In : *The Soft Tissues*. Sous la dir. de G.R. McLatchie et C.M.E. Lennox. Butterworth-Heinemann, 1993, p. 187-199. ISBN : 978-0-7506-0170-2. DOI : https://doi.org/10.1016/B978-0-7506-0170-2.50017-0 [125].

[181] Michael W. Whittle. « Clinical gait analysis : A review ». In : *Human Movement Science* 15.3 (1996), p. 369-387. ISSN : 0167-9457. DOI : https://doi.org/10.1016/0167-9457(96)00006-1 [126].

122. https://doi.org/https://doi.org/10.1016/S1350-4533(02)00041-3
123. https://doi.org/10.1081/SAC-120023882
124. https://doi.org/10.1007/s10618-012-0250-5
125. https://doi.org/https://doi.org/10.1016/B978-0-7506-0170-2.50017-0
126. https://doi.org/https://doi.org/10.1016/0167-9457(96)00006-1

[182] Oliver J. Woodman. *An introduction to inertial navigation*. Research report 696. University of Cambridge, 2007. DOI : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.7402 [127].

[183] Yun Yang et Ke Chen. « Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations ». In : *IEEE Transactions on Knowledge and Data Engineering* 23.2 (2011), p. 307-320. DOI : 10.1109/TKDE.2010.112 [128].

[184] Zhenlun Yang. « An Efficient Automatic Gait Anomaly Detection Method Based on Semisupervised Clustering ». In : *Computational Intelligence and Neuroscience* 2021 (fév. 2021). Sous la dir. d'Amparo Alonso-Betanzos. Publisher : Hindawi, p. 8840156. ISSN : 1687-5265. DOI : 10.1155/2021/8840156 [129].

[185] Xiao Zang. « Clustering Functional Data Based on Amplitude-Phase Separation ». Thèse de doct. The Ohio State University, 2021.

[186] Li Zheng, Tao Li et Chris Ding. « A Framework for Hierarchical Ensemble Clustering ». In : *ACM Trans. Knowl. Discov. Data* 9.2 (sept. 2014). ISSN : 1556-4681. DOI : 10.1145/2611380 [130].

[187] Wiebren Zijlstra. « Assessment of spatio-temporal parameters during unconstrained walking ». In : *European Journal of Applied Physiology* 92.1 (juin 2004), p. 39-44. ISSN : 1439-6327. DOI : 10.1007/s00421-004-1041-5 [131].

127. https://doi.org/http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.7402
128. https://doi.org/10.1109/TKDE.2010.112
129. https://doi.org/10.1155/2021/8840156
130. https://doi.org/10.1145/2611380
131. https://doi.org/10.1007/s00421-004-1041-5

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Titre : Amélioration du suivi des patients atteints de maladies neuro-dégénératives à l'aide d'objets connectés

Mot clés : Biostatistique, Analyse de la marche, Séries temporelles, Données fonctionnelles, Classification, Quaternions unitaires, Sclérose en Plaques

Résumé :

Cette thèse s'inscrit dans le contexte du projet *e-Gait* dont l'objectif est de développer un nouvel outil de mesure basé sur l'utilisation de systèmes numériques pour quantifier les troubles de la démarche de patients atteints de maladie neurodégénative, et plus particulièrement la Sclérose En Plaques (SEP). La solution adoptée consiste à mesurer les rotations en trois dimensions de la hanche au cours de la marche à l'aide d'un système de capteurs inertiels placé à la ceinture. Ces rotations sont représentées sous la forme d'une séquence de quaternions unitaires. Des méthodes adaptées à ce type de données sont présentées pour en extraire des informations relatives à la démarche de l'individu. Un algorithme est proposé pour segmenter le signal en cycles de marche. Dans une première approche, la démarche individuelle est représentée sous forme de paramètres spatio-temporels. Dans une seconde, elle est représentée sous la forme d'une unique séquence de quaternions unitaire appelée "Signature de Marche" (SdM). Des méthodes de classification non supervisée et semi-supervisée sont adaptées pour permettre d'identifier des groupes de patients présentant des déficits de la marche similaires à partir de leur SdM.

Title: Improving the monitoring of patients with neurodegenerative diseases with digital technologies

Keywords: Biostatistics, Gait analysis, Time series, Functional data, Clustering, Unit quaternions, Multiple Sclerosis

Abstract: This thesis is part of the *e-Gait* project whose objective is to develop a new measurement tool based on the use of digital systems to quantify the gait impairment of patients with neurodegenerative diseases, and more particularly Multiple Sclerosis (MS). The solution adopted consists in measuring the three-dimensional rotations of the hip during walking using a system of inertial sensors placed on the belt. These rotations are represented as a sequence of unit quaternions. Methods adapted to this type of data are presented to extract information related to the individual's gait. An algorithm is proposed to segment the signal into gait cycles. Two approaches to compare gait cycles between several individuals are explored. The first one consists in representing the gait as spatio-temporal parameters. The second is to represent the individual gait as a single sequence of unit quaternions called "Individual Gait Pattern" (IGP). Unsupervised and semi-supervised clustering methods are adapted to identify groups of patients with similar gait impairment based on their IGP.