Nantes Université

Laboratoire de Mathématiques Jean Leray

cnrs

Rapport de stage de fin d'étude

Information commune et spécifique aux données d'imagerie médicale et de données issues d'un système de capteurs de mouvements

---

Master 2 Ingénierie Statistique

Auteur **:**
Margot Bornet
Professeure référente **:**
Lise Bellanger

Encadrants **:**
Lise Bellanger
Aymeric Stamm

28 Août 2024

Sommaire

Remerciements

Je souhaite tout d'abord exprimer ma profonde gratitude à Lise Bellanger et Aymeric Stamm pour m'avoir offert l'opportunité de réaliser ce stage au sein du laboratoire de mathématiques et d'avoir ainsi pu découvrir ce projet. Je les remercie également pour leur soutien continu tout au long de ce stage et pour l'aide précieuse qu'ils m'ont apportée.

Je tiens également à les remercier de m'avoir permis de participer aux rencontres SEP, un événement organisé par l'Arsep consacré à la recherche sur la sclérose en plaques. Cette expérience enrichissante, qui m'a permis de réaliser un poster pour l'occasion, fut une première pour moi et s'est avérée particulièrement instructive.

Je tiens à exprimer ma reconnaissance à toutes les personnes qui m'ont aidé durant ce stage. Je remercie tout particulièrement Madame Véronique Cariou pour ses précieux conseils concernant l'analyse multibloc.

Je souhaite enfin remercier Klervi Le Gall, Nadia Negab et Manon Simonot, qui travaillent toutes les trois sur le projet eGait et qui, grâce à leur expérience et leurs connaissances sur le projet, ont su répondre à mes questions et m'ont ainsi accompagner dans la découverte de ce projet.

Introduction

L'étude réalisée durant ce stage s'inscrit dans le cadre du projet *eGait* porté par Lise Bellanger et Aymeric Stamm au sein du Laboratoire de Mathématiques Jean Leray (LMJL). Ce projet s'incrit dans le domaine de la Recherche et du Développement en biostatistique et a pour but l'analyse de la marche à l'aide d'un système de capteurs de mouvement. Un dispositif eGait a été développé au cours du projet par l'entreprise UmanIT et le LMJL et est composé d'une application mobile permettant de piloter une centrale inertielle, c'est-à-dire l'assemblage de capteurs de mouvement. Cette centrale inertielle permet la mesure des informations quantitatives relatives à la démarche d'un individu et le traitement des données récoltées permet de générer un biomarqueur appelé signature de marche (SDM). Ce biomarqueur caractérise la rotation de la hanche d'un individu au cours d'un cycle de marche moyen (voir section 1.4.1) et permet ainsi de suivre l'évolution des troubles de la marche d'un individu.

Les troubles de la marche peuvent engendrer des modifications significatives dans la vitesse de la marche ou encore dans l'amplitude des mouvements. Ils peuvent être provoqués par divers facteurs, comme le vieillissement et la maladie. Le projet est ainsi tourné vers l'étude de la marche chez les patients atteints de Sclérose en Plaques (SEP). Cette maladie auto-immune affecte le système nerveux central, entraînant ainsi des problèmes de coordination et de mobilité chez les malades.
Le déficit ambulatoire est habituellement évalué lors d'un examen clinique en mesurant le temps nécessaire pour parcourir huit mètres, à l'aide de scores de sévérité de la maladie tel que l'EDSS et avec une IRM en mesurant le volume lésionnel cérébral. Seulement, ces indicateurs ne semblent pas être suffisants pour l'analyse de la marche puisqu'ils ne permettent pas de dissocier les différents troubles de la marche, comme les troubles moteur, de spasticité et d'équilibre. C'est dans ce cadre que l'utilisation du biomarqueur SDM constitue une opportunité unique de collecter des informations quantitatives sur la santé du patient, précises, sans contraintes et peu coûteuses.

Nous nous concentrons sur l'évaluation de l'atteinte de la marche chez les patients atteints de SEP, le projet ayant pour but l'étude de l'association entre la SDM individuelle du patient obtenue lors de son examen clinique et la charge lésionnelle mesurée par IRM. L'objectif de ce stage est ainsi de : (i) évaluer l'association entre la SDM, l'EDSS et charge lésionnelle mesurée par IRM, (ii) établir des profils type de SDM en fonction de la charge lésionnelle observée par circuit et de la sévérité de la maladie mesurée via les scores EDSS, (iii) expliquer et prévoir le volume lésionnel par circuit, ou l'appartenance aux groupes établis en (ii), à partir des profils type de SDM.
L'établissement de profils type pourrait faciliter le diagnostic du handicap ambulatoire d'un patient en identifiant le profil type dont il se rapproche le plus. Cela pourrait permettre une évaluation spécifique des troubles de la marche dès l'examen clinique et ainsi permettre aux neurologues de proposer un suivi adapté au patient le plus tôt possible, limitant les risques d'aggravation.

1 Prérequis

1.1 La sclérose en plaques

La **sclérose en plaques** (SEP) est une maladie auto-immune affectant le système nerveux central; le cerveau et la moelle épinière. Le système immunitaire est touché et attaque la myéline, gaine protectrice des fibres nerveuses qui permet la propagation de l'influx nerveux du cerveau aux différentes parties du corps, entraînant ainsi différentes lésions.

De nombreux signes de la maladie sont observables : on retrouve des troubles moteurs, des fourmillements, des troubles de l'équilibre ou encore des troubles visuels ou urinaires. Les symptômes sont variés et évoluent chez un même patient, l'évolution et l'expression de la maladie sont donc imprévisibles.

Il existe deux modes évolutifs de la maladie : la forme rémittente et la forme progressive. La SEP rémittente, plus fréquente, évolue sous forme de poussées tandis que la SEP progressive correspond à une aggravation continue des symptômes, sans poussées. Une poussée correspond à l'apparition de nouveaux symptômes ou l'aggravation de symptômes déjà existants pendant plus de 24 heures et en dehors d'une période de fièvre.

La SEP touche 120 000 personnes en France et aucun traitement ne la guérit, mais il en existe permettant d'améliorer le quotidien des malades face aux nombreux symptômes.
Le suivi de la maladie se fait à l'aide d'un examen clinique durant lequel plusieurs tests sont réalisés. Parmi eux, le score EDSS est déterminé par le neurologue, et la vitesse de marche des malades est mesurée lors du *Timed 25-Foot Walk* (T25FW). Le T25FW est un test durant lequel le patient marche environ 7.60 mètres en ligne droite, avec ou sans aide, et encadré par les praticiens. Les troubles de la marche font en effet partis des symptômes les plus fréquents et entraînent une diminution de la qualité de vie chez les personnes atteintes de SEP, son étude est donc devenu un moyen important pour la compréhension et le suivi de la maladie.
Des IRM sont également réalisées afin de suivre l'évolution des lésions du cerveau et de la moelle épinière.

1.1.1 Echelle EDSS

L'**échelle EDSS** (Expanded Disability Status Scale) est une échelle discrète et non linéaire utlisée pour évaluer le handicap d'un patient. Ses valeurs varient de 0 à 10. Très souvent utilisée, elle permet le suivi des patients et l'évaluation de l'évolution du handicap, et est déterminée par le neurologue lors de l'examen clinique.
L'EDSS est un score global de sévérité de la maladie. Lors de l'examen clinique, le neurologue évalue les différents symptômes liés à la SEP, il évalue alors les fonctions neurologiques suivantes :

- la fonction visuelle : liée à la vue
- la fonction du tronc cérébral : gère le rythme cardiaque, la respiration, la motricité du visage et des yeux, l'élocution, la déglutition
- la fonction pyramidale : liée à la contraction volontaire des muscles
- la fonction cérébelleuse : gère la coordination des mouvements, l'équilibre
- la fonction sensitive : liée à la proprioception, la douleur, la thermoception
- la fonction sphinctérienne : liée aux troubles urinaires et intestinaux
- la fonction cérébrale (ou cognitive) : la mémoire, la concentration, l'humeur
- les autres fonctions neurologiques touchées

Le neurologue attribue un score à chacune de ces fonctions, c'est ce qu'on appelle les sous-scores, et il obtient le score final en combinant ces sous-scores.

1.1.2 Volume lésionnel cérébral mesuré par IRM

La sévérité de la maladie peut également être évaluée à l'aide de **mesures IRM** (Imagerie par Résonance Magnétique). Les données IRM permettent d'évaluer les lésions faites au cerveau et à la moelle épinière, et ainsi déterminer l'évolution de la maladie et son impact sur la démarche du patient.
On s'intéresse pour cela plusieurs régions relatives à l'atteinte de la marche : le cerveau, la moelle cervicale et la moelle thoracique. Grâce à l'IRM, il est possible de récupérer pour chaque région d'intérêt le volume des lésions et le volume total de la région en question, comprenant ainsi la substance blanche et les tractus corticospinaux gauche et droit. Le tractus corticospinal permet le contrôle des mouvements des membres et du tronc en facilitant la transmission des signaux moteurs du cortex cérébral aux muscles des membres et du tronc. Il constitue la voie motrice de la matière blanche, et commence au niveau du cortex cérébral, jusque dans la moelle épinière.

Ici, les lésions sont localisées sur les images IRM à l'aide d'une segmentation faite par un algorithme d'apprentissage profond, puis vérifiées par un médecin. La charge lésionnelle par circuit est alors comptabilisée pour le cerveau, la moelle cervicale et la moelle thoracique.

1.2 Dispositif eGait

Le suivi de la maladie et les méthodes d'évaluation des troubles de la marche d'un individu que nous avons évoqués rencontrent cependant certaines limites. En effet, les scores EDSS sont des données qualitatives évaluées par les neurologues. Cette évaluation ne donne pas la même importance à tous les symptômes. De plus, elle peut varier d'un neurologue à

l'autre, entraînant des scores différents pour un même patient, elle peut donc être biaisée. L'évaluation par mesures IRM est quant à elle assez contraignante car la réalisation d'une IRM est coûteuse et réalisée seulement une fois par an.

L'idée d'un capteur à porter au niveau de la hanche lors de la marche est apparue comme un nouveau moyen pour évaluer l'atteinte de la marche, donnant des valeurs quantitatives non biaisées et peu contraignantes. Le disposotif eGait est notamment constitué d'un système de capteurs, composé d'un accéléromètre, un gyroscope et un magnétomètre, d'un smartphone et d'une application mobile permettant l'acquisition et le stockage des données.

Le capteur se porte au niveau de la hanche et récupère des données à une fréquence de 100 Hz, c'est-à-dire toutes les 10 ms. Les données sont récupérées sous la forme de séquences de quaternions unitaires, définissant l'angle de rotation de la hanche au cours du temps. La rotation est ainsi décrite par 2 paramètres : son axe de rotation et son angle de rotation.

Figure 1: Capteur positionné à la hanche

Figure 2: Données récupérées par le capteur

Ce dispositif est le premier se plaçant au niveau de la hanche, les capteurs pré-existants se plaçant habituellement au niveau des pieds des patients. L'objectif de ce capteur est ainsi d'être porté quotidiennement par les patients afin de récupérer leurs données de marche et repérer d'éventuelles troubles de la marche, tout en assurance une utilisation simple et peu contraignante.

1.3 Les quaternions

Un **quaternion** est un vecteur à 4 dimensions, que l'on note $\mathbf{q} = (w, x, y, z)^t$, $\mathbf{q} \in \mathbb{R}^4$. Il peut aussi être considéré comme un nombre hypercomplexe de rang 4, on a alors la notation suivante :

$$\mathbf{q} = w + ix + jy + kz \text{ avec } i^2 = j^2 = k^2 = ijk = -1$$

avec w, x, y et z des nombres réels et i, j et k généralisant le nombre imaginaire i selon la règle :

$$i^2 = j^2 = k^2 = ijk = -1$$

Un **quaternion unitaire** est un quaternion normé; il vérifie $\|\mathbf{q}\| = \sqrt{w^2 + x^2 + y^2 + z^2} = 1$. Les quaternions unitaires permettent de décrire une rotation en 3 dimensions. Ils s'écrivent sous la forme :

$$\mathbf{q} = \cos\frac{\theta}{2} + \mathbf{u}\sin\frac{\theta}{2} = \cos\frac{\theta}{2} + u_x \sin\frac{\theta}{2} i + u_y \sin\frac{\theta}{2} j + u_z \sin\frac{\theta}{2} k$$

Le quaternion $\mathbf{q}$ représente la rotation d'angle $\theta \in \mathbb{R}$ autour de l'axe $\mathbf{u} = (u_x, u_y, u_z) \in \mathbb{S}^2$, où $\mathbb{S}^2$ est la sphère en trois dimensions. Cette rotation est illustrée sur la figure 3.

Les données d'intérêt sont les rotations de la hanche au cours du temps. On peut considérer N quaternions unitaires. Ces rotations se retrouvent sous la forme d'une série temporelle de quaternions (QTS) : $Q = (\mathbf{q}_1, ..., \mathbf{q}_N)^t = \begin{pmatrix} w = (w_1, ..., w_N)^t \\ x = (x_1, ..., x_N)^t \\ y = (y_1, ..., y_N)^t \\ z = (z_1, ..., z_N)^t \end{pmatrix}^t$ où la séquence des quaternions est associée aux temps $T = (t_1, ..., t_N)$.

Les rotations sont calculées par le dispositif, positionné à la hanche. Le système de capteurs est doté de son propre référentiel, noté $R_s = (s_1, s_2, s_3)$, et on considère le référentiel terrestre $R_f = (f_1, f_2, f_2)$. L'orientation du dispositif à un temps donné est alors déterminée comme la rotation entre un référentiel fixe, le référentiel terrestre, vers le référentiel du système de capteurs. On retrouve cela sur le schéma 4, où la rotation est décrite par son axe de rotation $\mathbf{u}$ et son angle de rotation θ.

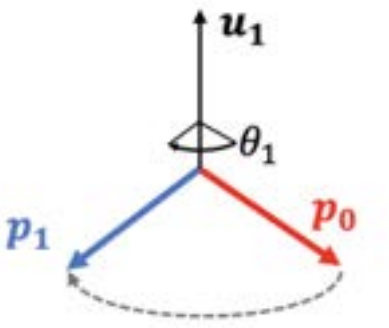


Figure 3: Rotation de p_0 d'axe de rotation $\mathbf{u}_1$ et d'angle de rotation θ_1

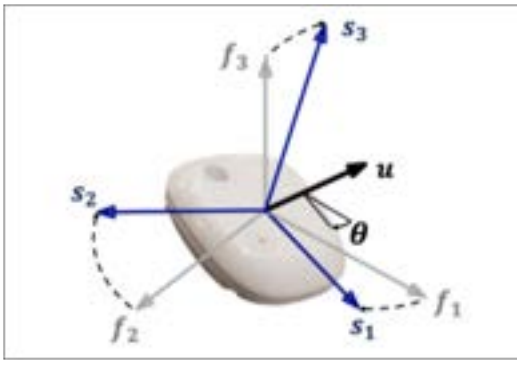


Figure 4: Capteur avec son référentiel R_s (en bleu) et R_f le référentiel terrestre (en gris)

Le groupe des quaternions unitaires, noté $\mathbb{H}_u$, a plusieurs particularités. Il est notamment doté d'une métrique particulière : la distance géodésique. Celle-ci permet de calculer la distance minimale entre deux quaternions unitaires $\mathbf{q}_1$ et $\mathbf{q}_2$:

$$d(\mathbf{q}_1, \mathbf{q}_2) = 2 \arccos Re(\mathbf{q}_1^{-1}\mathbf{q}_2)$$

Le quaternion $\mathbf{q}_1^{-1}\mathbf{q}_2$ représente la rotation nécessaire pour obtenir $\mathbf{q}_2$ à partir de $\mathbf{q}_1$.

1.4 Etude des données de marche

1.4.1 Cycle de marche

La démarche d'un individu est caractérisée par la répétition de cycles de marche au cours du temps. Un cycle de marche correspond à l'ensemble des mouvements réalisés entre deux contacts successifs d'un pied donné avec le sol. Il est constitué d'une phase d'appui et d'une phase de balancement, comme représenté sur le schéma ci-dessous.

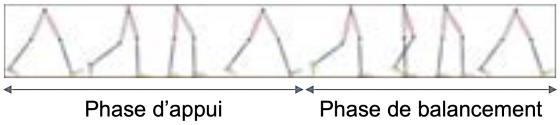


Figure 5: Description d'un cycle de marche

- La **phase d'appui** correspond à la période où le pied est en contact avec le sol. Elle débute par le contact initial du pied avec le sol, et se termine au décollement des orteils, marquant la perte de contact avec le sol.
- La **phase de balancement** correspond à la période où le pied n'est plus en contact avec le sol. C'est le décollement des orteils, jusqu'au contact initial suivant du même pied.

1.4.2 Algorithme STRIPAGE et Signature de Marche

Les données brutes récupérées par le système de capteurs sont ensuite traitées afin de générer une **signature de marche** (SDM) propre à chaque individu. La SDM est un biomarqueur représentant la démarche d'un individu et caractérisant les mouvements qu'un individu fait au cours d'un pas. Elle correspond à la rotation de la hanche d'un individu durant un cycle de marche moyen.

Le capteur permet d'obtenir une séquence de quaternions unitaires pouvant contenir plusieurs cycles de marche; on a alors une série temporelle de quaternions unitaires. Ces différents cycles de marche sont identifiés à l'aide de l'algorithme STRIPAGE (STRIde PAttern GEneration). Cet alorithme a été développé par Pierre Drouin dans sa thèse [5]. Les données étant stockées dans des fichers CSV (voir Figure 2), on retrouve plusieurs étapes pour chaque fichier :

- on lit la QTS et on la centre [13]
- on déduit des QTS la vitesse de marche

- on segmente chaque QTS

L'étape de segmentation a initialement été développée par Pierre Drouin dans sa thèse. Elle a par la suite été améliorée par Aymeric Stamm et Manon Simonot et utilise désormais un arbre de décision. Pour cette nouvelle façon de segmenter les QTS en cycles de marche, le modèle d'arbre de décision est entraîné sur un *feature space* comportant :

- la vitesse angulaire
- l'accélération angulaire
- les angles d'Euler RPY (Roll-Pitch-Yaw) [21]
- la vitesse de marche

qui sont calculés aux temps t et aux temps $t-1, ..., t-k$ avec k le paramètre de lag.

Ensuite, pour chaque patient, l'algorithme réalise les étapes suivantes :

1. Regroupement des cycles obtenus dans les différents fichiers lui correspondant (en général 4 fichiers car le patient fait normalement deux T25FW, chacun comprenant un aller-retour)
2. Calcul de la durée des cycles pour la mettre de coté
3. Ré-échantillonnage de chaque cycle pour garder 101 points
4. Définition de la grille 0:100 comme grille commune à tous les cycles
5. Détection et suppression des outliers de forme [10] [1]
6. Alignement des cycles [12]
7. Définition de la SDM comme la médiane des cycles alignés

Cet algorithme permet alors de représenter la démarche d'un individu sous la forme d'une unique séquence de quaternions unitaires, correspondant à la signature de marche.

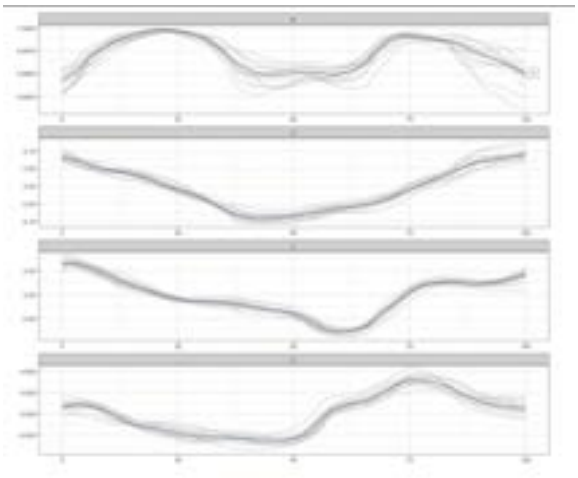


Figure 6: SDM d'un individu et ses cycles de marche

Sur la figure 6, la SDM est représentée en bleu et les cycles de marche sont en gris. Ils sont données en pourcentage de durée de cycle sur chaque composante des quaternions.

1.4.3 Paramètres spatio-temporels

Les cycles de marche et la SDM permettent le calcul de **paramètres spatio-temporels** (PST) donnant des informations sur la démarche des individus. On a :

- Paramètres temporels :
  - durée des pas
  - durée des cycles
  - durée de la phase d'appui et de la phase de balancement
  - 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 ou de la cheville au cours du cycle
- Paramètre spatio-temporel :

– vitesse de marche

On peut alors caractériser la démarche d'un individu à l'aide des paramètres spatio-temporels suivants :

- La durée moyenne d'un cycle
- L'amplitude moyenne d'un cycle
- La durée moyenne de la phase d'appui
- La variabilité des cycles
- La vitesse angulaire moyenne d'un cycle

2 Méthodes statistiques d'intégration de données multi-sources

2.1 Notation

Dans le cadre de l'intégration de données et pour la suite de ce papier nous posons les notations suivantes. On considère K sources de données, représentées par K matrices contenant les mêmes n observations. La matrice X_k contient d_k variables, pour $k = 1, ..., K$. Pour chaque matrice X_k, nous notons r_k son rang et nous considérons $F_k = (f_{k1}, ..., f_{kr_k})$ la matrice des scores et W_k la matrice des poids. Nous retrouvons la relation : $F_k = X_k W_k$.

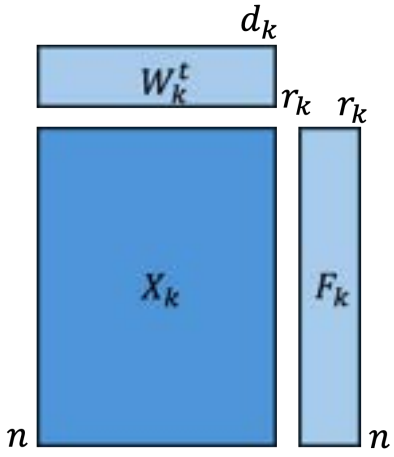


Figure 7: Notation pour la matrice X_k

Pour la suite, nous considérons également les **décompositions en valeurs singulières** (DVS) des matrices. Nous posons ainsi les notations suivantes. On considère la matrice X_k. Sa DVS se note :

$$X_k = U_k \Lambda_k^{\frac{1}{2}} V_k^t$$

Où U_k représente la matrice des vecteurs singuliers à gauche, V_m est la matrice des vecteurs singuliers à droite et la matrice $\Lambda_k^{\frac{1}{2}}$ contient les valeurs singulières sur sa diagonale, c'est-à-dire les racines carrées des valeurs propres $\lambda_1, \lambda_2, ...$

On définit également les notions de variabilité jointe et individuelle. On considère $k = 1, ..., K$ jeux de données avec les mêmes n observations et différentes variables et on s'intéresse à l'espace des scores engendré par les composantes $(f_{k1}, ..., f_{kr_k})$. Chaque jeu de données apporte de l'information qui peut être décomposée en trois types de variabilité, on observe ainsi :

- la **variabilité jointe** : c'est la variabilité commune aux blocs de données, représentée en bleu sur le schéma 8. Les composantes communes engendrent un sous-espace de $\mathbb{R}^n$ composé de la variabilité commune.
- la **variabilité individuelle** : c'est de la variabilité propre à chaque bloc. Les composantes individuelles forment des sous-espaces, un pour chaque bloc, où l'on retrouve la variabilité individuelle.
- le bruit : c'est l'information résiduelle.

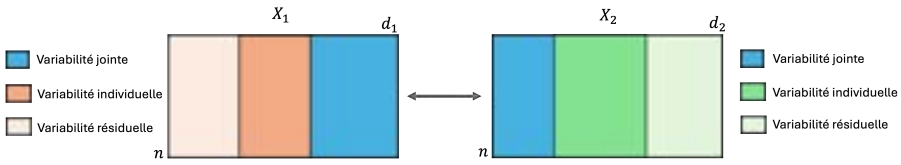


Figure 8: Décomposition de la variabilité de 2 blocs de données en 3 types de variabilité

Cette décomposition se fait donc en terme de sous-espaces et non en terme de variables, puisqu'une même variable peut fournir à la fois de la variabilité commune et de la variabilité individuelle.

2.2 CCA et extensions

La **CCA** (Analyse Canonique des Corrélations) est une méthode qui permet d'examiner le lien entre deux ensembles de variables mesurées sur les mêmes observations, et ainsi de savoir s'ils mesurent ou non les mêmes propriétés.
Cette méthode, plus classique ne prend en entrée que deux jeux de données, les matrices X_1 et X_2 définies précédemment et renvoie les $m \in \mathbb{N}$ composantes maximisant la corrélation entre les blocs.
On cherche ainsi les composantes orthogonales $f_{1h} = X_1 w_{1h}$ et $f_{2h} = X_2 w_{2h}$, appartenant à F_1 et F_2 respectivement, maximisant : $corr\,(X_1 w_{1h}, X_2 w_{2h})$, $h = 1, ..., m$.

La CCA peut également être vue sous la forme d'un autre problème de maximisation, détaillé dans l'ouvrage de A. Smilde, T. Naes et K. Liland de 2022 "Multiblock data fusion in statistics and machine learning" [18]. Pour simplifier l'explication, nous considérons une composante de chaque bloc : $f_{11} = X_1 w_{11}$ et $f_{21} = X_2 w_{21}$.
Le principe de la CCA correspond alors au problème de maximisation suivant :

$$\max_{w_{11}, w_{21}} w_{11}^t X_1^t X_2 w_{21} \text{ sc } \|f_{11}\| = \|f_{21}\| = 1 \tag{1}$$

En réécrivant cela en utilisant les DVS de X_1 et X_2, on obtient le nouveau problème de maximisation :

$$\max_{q_1,q_2} q_1 U_1^t U_2 q_2 \text{ sc } \|q_{11}\| = \|q_{21}\| = 1 \tag{2}$$

où $q_{11} = \Lambda_1^{\frac{1}{2}} V_1^t w_{11}$ et $q_{21} = \Lambda_2^{\frac{1}{2}} V_2^t w_{21}$.

On a bien

$$1 = f_{11}^t f_{11} = w_{11}^t X_1^t X_1 w_{11} = w_{11}^t V_1 \Lambda_1^{\frac{1}{2}^t} U_1^t U_1 \Lambda_1^{\frac{1}{2}} V_1^t w_{11} = w_{11}^t V_1 \Lambda_1^{\frac{1}{2}^t} \Lambda_1^{\frac{1}{2}} V_1^t w_{11} = q_{11}^t q_{11}$$

Or, la matrice $\Lambda_k^{\frac{1}{2}}$ mesure les forces des corrélations dans le bloc X_k. En effet, elle contient les carrés des valeurs singulières, qui donnent la variation expliquée dans le bloc X_k.
Ces matrices représentent en fait la variabilité intra-bloc.
Ainsi, cette réécriture, et l'absence des matrices Λ_1 et Λ_2 nous permet de montrer que la CCA n'explique pas l'information propre à chaque bloc, mais se concentre entièrement sur la variabilité commune aux blocs.

Choix du nombre de composantes
On pose :

- $R_{11} = \frac{1}{n} X_1^t X_1$ et $R_{22} = \frac{1}{n} X_2^t X_2$ les matrices des corrélations intra-groupes
- $R_{12} = \frac{1}{n} X_1^t X_2$ la matrice des corrélations inter-groupes
- $s = rg(R_{12})$ le rang de la matrice des corrélations inter-groupes
- $r_h = corr(X_1 w_{1h}, X_2 w_{2h})$ la corrélation canonique

La statistique de Wilks est utilisé pour déterminer les composantes à retenir. Elle est définie par $\Psi_j = \prod_{h=j+1}^{s} (1 - r_h^2)$.
Le niveau de signification est donné par les approximations usuelles de Barlett ou Rao.

- La statistique de Barlett est donnée par :

$$\chi^2 = -\left\{ n - \frac{1}{2}(d_1 + d_2 + 3) \right\} \ln \Psi_j$$

- La statistique de Rao est donnée par :

$$F = \frac{1 - \Psi_j^{\frac{1}{t}}}{\Psi_j^{\frac{1}{t}}} \frac{dl_2}{dl_1}$$

où

- $dl_1 = (d_1 - j)(d_2 - j)$
- $dl_2 = vt\{(d_1 - j)(d_2 - j)\} + 1$
- $v = n - \frac{1}{2}(d_1 + d_2 + 3)$
- $t = \sqrt{\frac{(d_1-j)^2(d_2-j)^2-4}{(d_1-j)^2+(d_2-j)^2-5}}$

On peut également calculer la part de variance d'un groupe expliquée par les composantes appelée redondance) :

- Redondance de X_1 par rapport à sa composante f_{1h} :

$$Rd(X_1, f_{1h}) = \frac{1}{d_1}\sum_{j=1}^{d_1} cor(x_{1j}, f_{1h})^2$$

- Redondance de X_1 par rapport à la composante de l'autre groupe f_{2h} :

$$Rd(X_1, f_{2h}) = \frac{1}{d_1}\sum_{j=1}^{d_1} cor(x_{1j}, f_{2h})2$$

Extensions de la CCA

La **GCA** (Generalized Canonical Analysis) est une méthode généralisant la CCA à plus de deux jeux de données. Elle permet d'étudier les relations entre plusieurs blocs de données $X_1, ..., X_K$ et se base ainsi sur la maximisation du critère suivant :

$$\sum_{k=1}^{K}\sum_{l\neq k} cov(X_K W_k, X_l, W_l)$$

La **RGCCA** (Regularized Generalized Canonical Correlation Analysis) est une extension de la GCA incluant une régularisation. Elle permet ainsi de stabiliser l'estimation des vecteurs de poids et est plus adaptée aux cas de la grande dimension et de problèmes de colinéarité. Elle correspond au problème suivant :

$$\max \sum_{k=1}^{K}\sum_{l\neq k} c_{kl} cov(X_k W_k, X_l W_l) - \sum_{k=1}^{K} l_k \|W_k\|^2$$

avec :

- c_{kl} : coefficient de liaison contrôlant l'importance relative de la covariance entre les blocs k et l
- l_k : paramètre de régularisation pour le bloc k
- $\|W_k\| = Tr(W_k^t W_k)$

2.3 PLS2 et extensions

La **PLS** (Moindres Carrés Partiels) est une méthode statistique de maximisation de la variance. Elle permet notamment de prévoir une variable réponse à l'aide d'un groupe de variables explicatives. La **PLS2** correspond au cas où il y a plusieurs variables dépendantes. On souhaite alors prévoir un ensemble de variables réponses X_2 à partir des variables explicatives X_1. Cette méthode est expliquée dans le livre de M. Tenenhaus "La régression PLS" [19].

Comme la CCA, la PLS2 prend deux jeux de données en entrée, et on cherche cette fois à maximiser la covariance. On cherche m composantes orthogonales f_{1h} et f_{2h} maximisant cette fois : $cov(X_1 w_{1h}, X_2 w_{2h})$ sc $\|w_{1h}\| = \|w_{2h}\| = 1$.
En considérant une composante de chaque bloc, on retrouve un problème de maximisation similaire à la CCA :

$$\max_{w_{11}, w_{21}} w_{11}^t X_1^t X_2 w_{21} \text{ sc } \|w_{11}\| = \|w_{21}\| = 1 \tag{3}$$

On peut de nouveau réécrire le problème de maximisation à l'aide des DVS :

$$\max_{w_{11}, w_{21}} w_{11} V_1 \Lambda_1^{\frac{1}{2}} U_1^t U_2 \Lambda_2 \frac{1}{2} V_2 w_{21} \text{ sc } \|w_{11}\| = \|w_{21}\| = 1 \tag{4}$$

On pose $z_{11} = V_1^t w_{11}$ et $z_{21} = V_2^t w21$. On a alors $z_{11}^t z_{11} = w_{11}^t V_1 V_1^t w_{11} = w_{11}^t w_{11} = 1$
Et on obtient le nouveau problème de maximisation :

$$\max_{z_{11}, z_{21}} z_{11}^t \Lambda_1^{\frac{1}{2}} U_1^t U_2 \Lambda_2^{\frac{1}{2}} z_{21} \text{ sc } \|z_{11}\| = \|z_{21}\| = 1 \tag{5}$$

On observe cette fois que la solution PLS prend en compte les matrices Λ_1 et Λ_2, et considère ainsi les structures de corrélation des blocs. La solution PLS permet donc d'expliquer une certaine quantité de variabilité individuelle propre à chaque bloc, en plus de la variabilité jointe.

On peut retrouver ce résultat et cette différence entre les solutions PLS et CCA à l'aide de la formule de la covariance :

$$cov(X_1 w_{11}, X_2, w_{21} = \sqrt{var(X_1 w_{11})} \sqrt{var(X_2 w_{21})} corr(X_1 w_{11}, X_2 w_{21})$$

En maximisant la covariance, on maximise en même temps les variances.

Choix du nombre de composantes
Pour déterminer le nombre de composantes à conserver en régression PLS on utilise la validation croisée.

On définit :

- RSS_{h-1} (Residual Sum of Squares) la somme des carrés résiduelle calculée avec le modèle à $h-1$ composantes
- PRESS (Prediction Error Sum of Squares) la somme des carrés des erreurs de prévisions calculées sur les jeux-test

On définit alors le Q^2 de Stone-Geisser (ou indice de redondance) comme :

- Pour chaque variable x_{2k} :
$$Q^2_{hk} = 1 - \frac{PRESS_{kh}}{RSS_{k(h-1)}}$$
- Sur l'ensemble des variables X_2 :
$$Q^2_h = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}}$$

où on a :

$$RSS_{kh} = \sum_{i=1}^{n} (x_{2ki} - \hat{x}_{2khi})^2 \text{ et } PRESS_{kh} = \sum_{i=1}^{n} \left(x_{2ki} - \hat{x}_{2kh(-i)}\right)^2$$

Cet indice de redondance permet de mesurer l'apport marginal de chaque composante PLS f_{1h} au pouvoir prédictif du modèle.
Il existe deux règles pour décider de si l'apport de la composante f_{1h} est significatif et ainsi choisir les composantes à conserver :

- Règle 1 : L'apport de f_{1h} est significatif si $Q^2_h \geq (1 - 0.95^2) = 0.0975$
- Règle 2 : L'apport de f_{1h} est significatif si au moins un Q^2_{kh} est tel que $Q^2_{kh} \geq 0.0975$

Extensions de la régression PLS
La PLS et le PLS2 ont donné suite à d'autres méthodes. Les composantes PLS obtenues sont des combinaisons linéaires de l'ensemble des variables. Un grand nombre de variables peut donc constituer une limite à la PLS et à la bonne interprétation des résultats. La **Sparse PLS** est une extension de la PLS adaptée à cette situation. Elle va en effet limiter le nombre de variables prises pour les combinaisons linéaires donnant les composantes et permettre ainsi de réduire la complexité de l'interprétation dans le cas de la grand dimension.

Un autre point important concernant la PLS2 est qu'elle ne permet de ne prendre en compte que deux jeux de données. Une extension de cette méthode est la méthode **MB-PLS** (Multiblock PLS). C'est une méthode d'analyse de données multiblocs surpervisée, elle est donc adaptée à l'utilisation de K blocs de données $X_1, ..., X_K$. Cette méthode permet

d'étudier les relations entre ces K blocs de données. Elle est basée sur le problème de maximisation suivant :

$$\max \sum_{k=1}^{K} \sum_{l=1}^{K} cov(X_k W_k, X_l W_l) = \max \sum_{k=1}^{K} \sum_{l=1}^{K} cov(F_k, F_l)$$

2.4 AJIVE

Une autre méthode d'intégration de données appelée **AJIVE** (**Angle-based Joint and Individual Variation Explained**) a été introduite en 2018 et permet une approche différente des précédentes [8].

L'algorithme AJIVE permet d'extraire simultanément l'information commune à plusieurs jeux de données et l'information spécifique à chaque jeu de données. On considère ainsi désormais K matrices $X_1, ..., X_K$.

Plus précisément, AJIVE permet d'extraire un sous-espace avec la variabilité commune aux blocs de données, appelée variabilité jointe, engendré par les composantes communes, et des sous-espaces avec une variabilité propre à chaque bloc, appelée variabilité individuelle, engendrés par les composantes distinctes. Contrairement à précédemment, la même importance est donc donnée à chacun des jeux de données.

On s'intéresse, pour chaque jeu de données X_k, à l'espace des scores, que l'on a défini précédemment. La décomposition du block X_k se fait alors sous la forme :

$$X_k = A_k + E_k = J_k + I_k + E_k \tag{6}$$

A_k est de rang r_{A_k}, et correspond au signal, c'est-à-dire aux données sans le bruit E_k, et avec J_k correspondant au sous-espace joint des scores, et I_k correspondant au sous-espace individuel.

On note ainsi :

- col(A_k) l'espace des scores de A_k, de dimension r_{A_k}
- col(J_k) l'espace des scores de J_k, de dimension r_{J_k}
- $col(I_k)$ l'espace des scores de I_k, de dimension r_{I_k}

Plusieurs propriétés permettent de définir ces espaces :

1. L'équation (6) nous donne la propriété suivante sur les rangs : $r_{A_k} = r_{J_k} + r_{I_k}$
2. On note $col(J) = col(J_1) = ... = col(J_K)$ l'espace des scores communs, représentant toute l'information commune. Ainsi $r_J = r_{J_1} = ... = r_{J_K}$.
3. Pour $k = 1, ..., K$, on a $col(J_k) = col(J) \subset col(A_k)$.

4. L'information individuelle est propre à chaque jeu de données, il n'y a pas d'information commune. Cela se définit par la propriété : $\bigcap_{k=1}^{K} col(I_k) = \left\{\vec{0}\right\}$
5. Les espaces joints et individuels sont distincts : $col(J) \perp col(I_k)$, $k = 1, ..., K$.
6. Pour $k = 1, ..., K$, $col(I_k) \subset col(A_k)$.

La décomposition de AJIVE donnée en (6) se fait en 3 étapes distinctes.

Etape 1: Extraction de l'espace du signal
La première étape a pour but l'extraction du signal A_k pour chaque bloc de données X_k. On enlève ainsi l'information résiduelle.

Pour ce faire, on réalise une approximation de bas rang de chaque bloc X_k afin d'obtenir une nouvelle matrice simplifiée, et préservant les informations essentielles. On réalise pour chaque bloc une DVS tronquée à un seuil $\lambda_k^{\frac{1}{2}}$ à définir afin d'extraire le signal. On obtient ainsi pour chaque bloc k :

$$\tilde{A}_k = \tilde{U}_k \tilde{\Lambda}_k^{\frac{1}{2}} \tilde{V}_k^t$$

$\tilde{A}_k$ est une approximation de A_k, de rang $\tilde{r}_{A_k}$. On conserve ainsi dans la matrice $\tilde{\Lambda}_k^{\frac{1}{2}}$ les $\tilde{r}_{A_k}$ plus grandes valeurs propres. Les valeurs singulières sous le seuil $\lambda_k^{\frac{1}{2}}$ sont mises à zéro.

L'algorithme comprend également une estimation de la précision de cette approximation. Pour cela, on veut estimer la distance entre les sous-espaces $col(A_k)$ et $col(\tilde{A}_k)$. On introduit la pseudo-métrique suivante :

$$\rho\left\{col(A_k), col(\tilde{A}_k)\right\} = \left\|P_{A_k} - P_{\tilde{A}_k}\right\|_2 \tag{7}$$

où P_{A_k} et $P_{\tilde{A}_k}$ sont les matrices de projection sur $col(A_k)$ et $col(\tilde{A}_k)$ respectivement.
ρ correspond au sinus de l'angle principal maximal entre les deux sous-espaces. En effet, le plus grand angle principal entre deux sous-espaces mesure leur proximité, et donc leur distance.
On définit alors une borne pour la distance entre les sous-espaces singuliers de A_k et X_k, c'est la **borne Wedin**. Elle permet de quantifier comment les sous-espaces singuliers théoriques sont affectés par le bruit.
Cette borne est définie comme suit :

$$\rho\left\{col(A_k), col(\tilde{A}_k)\right\} \leq \frac{\max\left(\|E_k\tilde{V}_k\|, \|E_k^t\tilde{U}_k\|\right)}{\sigma_{\min}(\tilde{A}_k)} \wedge 1 \tag{8}$$

Cependant la borne Wedin utilise les matrices d'erreur E_k, qui ne sont pas observables. On utilise donc une estimation de cette borne à partir d'un ré-échantillonnage des signaux et de la relation $\tilde{E}_k = X_k - \tilde{A}_k$.

Etape 2: Segmentation de l'espace des scores
La deuxième étape de l'algorithme consiste à segmenter l'espace des scores en composantes jointes et individuelles. On extrait les composantes jointes J_k du signal A_k.
Cette segmentation repose sur l'analyse d'angles principaux. En effet, en considérant les signaux $\tilde{A}_1$ et $\tilde{A}_2$, on utilise l'idée selon laquelle on doit observer un angle assez petit entre les composantes de $col(\tilde{A}_1)$ et $col(\tilde{A}_2)$ correspondant à l'espace joint. A l'inverse, on s'attend à un angle assez large entre les composantes de $col(\tilde{A}_1)$ et $col(\tilde{A}_2)$ correspondant aux espaces individuels.

Les angles peuvent être obtenus à l'aide d'une DVS sur la concaténation des matrices de vecteurs singuliers $\tilde{V}_1$ et $\tilde{V}_2$:

$$M \triangleq \begin{bmatrix} \tilde{V}_1^t \\ \tilde{V}_2^2 \end{bmatrix} = U_M \Sigma_M^{\frac{1}{2}} V_M^t \tag{9}$$

Ici $\Sigma_M^{\frac{1}{2}}$ contient sur sa diagonale les valeurs singulières, notées $\sigma_{M,i}, i \in \mathbb{R}$, qu'on ordonne dans l'ordre décroissant et qui permettent d'obtenir les angles principaux. La plus grande valeur singulière donne ainsi le plus petit angle principal :

$$\phi_i = \arccos\left\{(\sigma_{M,i})^2 - 1\right\} \tag{10}$$

On cherche alors à déterminer les composantes jointes, les $\tilde{r}_J$ plus petits angles pouvant être considérés comme composantes jointes.
On utilise des bornes sur le plus petit et le plus grand angle correspondant aux composantes jointes, afin de ne pas retenir d'angles liés au bruit ni aux composantes individuelles.

Etape 3: Segmentation finale de l'espace La dernière étape de l'algorithme va permettre d'obtenir les sorties finales en vérifiant le respect des conditions initiales : la contrainte du seuil déterminé en étape 1.
Soit $\tilde{V}_J = \begin{bmatrix} \vec{v}_{M,1} & ... & \vec{v}_{M,\tilde{r}_J} \end{bmatrix}$ la matrice obtenue dans la 2ème étape, où $\vec{v}_{M,i}$ est la $i^{\text{ème}}$ colonne de V_M.
Le respect des conditions initiales est ensuite vérifié pour chaque bloc de données $k = 1, ..., K$. Les vecteurs ne vérifiant pas la contrainte d'identifiabilité suivante sont enlevés : $\|X_k \vec{v}_{M,i}\| > \lambda_k^{\frac{1}{2}}$.
On obtient alors $\hat{V}_M$ la matrice finale et $\hat{r}_J$ le rang joint final.
Enfin, on utilise l'orthogonalité des espaces jointes et individuelles pour obtenir les composantes individuelles, et on vérifie de nouveau le respect du seuil.

2.5 Synthèse des méthodes

Les méthodes CCA, PLS et AJIVE permettent d'étudier les relations entre des blocs de données, avec notamment l'information commune apportée par ces blocs. Elles ont cependant de nombreuses différences. En effet, PLS et CCA sont basées sur des critères de maximisation de corrélation et de covariance mais ne s'appliquent qu'à deux blocs de données. Au contraire, AJIVE est une méthode cherchant à maximiser la distance entre les espaces individuels et minimiser celle entre les espaces joints de chaque bloc, ainsi adaptée à plus de deux blocs. De plus, AJIVE est la seule des ces trois méthodes permettant d'obtenir simultanément les variabilités jointe et individuelles. Pour cela, la même importance est donnée à chaque bloc.
La CCA permet d'obtenir la variabilité jointe à deux blocs, tout en leur donnant la même importance. Enfin, PLS2 fait une distinction entre les deux blocs puisqu'elle permet d'expliquer un groupe de variables à l'aide d'un autre groupe de variables. Cette méthode apporte de l'information individuelle et commune aux deux blocs. Les résultats sont cependant plus difficiles à interpréter.

Chaque méthode a ainsi plusieurs avantages et inconvénients que l'on peut résumer dans le tableau suivant, dans le cas où l'on considère deux blocs de données à analyser.

	CCA	**PLS2**	**AJIVE**
Avantages	• Variabilité commune • Permet de bien comprendre les relations entre les blocs • Interprétation des composantes + simple que PLS • Méthode descriptive / exploratoire	• Variabilité commune + individuelle • Robuste à la colinéarité • Utile pour la construction de modèles prédictifs • Capacité à gérer la grande dimension • Méthode explicative	• Variabilité individuelle • Variabilité commune • Structure détaillée des données et des relations entre les ensembles • Méthode descriptive/exploratoire • Plusieurs jeux de données
Inconvénients	• Pas de variabilité individuelle • Sensible à la colinéarité • Pas adapté à la grande dimension • Choix du nombre de composantes • 2 jeux de données	• Difficulté d'interprétation • Sensibilité au bruit • Moins sensible à la variabilité jointe • Choix du nombre de composantes • 2 jeux de données	• Peut être coûteux, en particulier pour de très grands ensembles de données • Choix du nombre de composantes

Table 1: Tableau récapitulatif des 3 méthodes multiblocs

3 Matériel

Dans le cadre de ce projet, les bases de données de deux études sont utilisées et étudiées, toutes deux regroupant des patients atteints de sclérose en plaques.

3.1 Données AMIES

3.1.1 Description des données

Dans un premier temps, nous avons utilisé des donnés provenant de l'étude AMIES. Celles-ci sont issues d'une étude mono-centrique de l'hôpital universitaire Laennec de Nantes. Les patients considérés ici ont donné leur accord pour rejoindre la cohorte OFSEP-HD de l'Observatoire Français de la Sclérose en Plaques (OFSEP). L'inclusion des patients a été réalisée sur les années 2021 et 2022 sur 39 patients de l'hôpital de Nantes atteints de sclérose en plaque. Pour être inclus dans cette cohorte, les patients ne doivent pas nécessiter d'un usage permanent d'un fauteuil roulant, ils doivent donc avoir un score EDSS inférieur à 7 et ils doivent être agés d'au moins 15 ans.

Les données AMIES regroupent de nombreuses informations sur les patients tels que l'âge, le sexe, le score EDSS ou encore le temps réalisé lors du T25FW. Toutes les variables disponibles dans les données AMIES sont détaillées en annexe. Parmi ces variables, deux des indicateurs de sévérité de la maladie sont disponibles : les scores EDSS et la SDM. Ce sont sur ces variables que nous travaillons.
Pour la suite, nous considérons alors 2 jeux de données, tirées des données initiales :

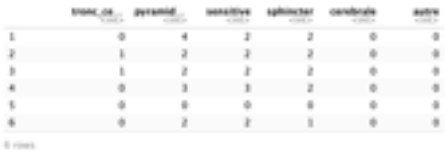
- les données cliniques : jeu de données contenant les sous-scores EDSS en colonne et les 39 patients en ligne
- les données de marche : jeu de données contenant les SDM sous forme de séries temporelles de quaternions

3.1.2 Pré-traitement des données

Les données AMIES sont constituées de 39 patients, avec d'une part les 8 sous-scores EDSS donnés en colonne pour chaque patient, et de 39 tableaux tibbles contenant la SDM et les coordonnées à chaque temps, pour chaque patient.

On s'intéresse d'abord aux données cliniques. Une première étape est la vérification de données manquantes. Certains patients ont des scores notés "X", indiquant une difficulté de la part du neurologue d'indiquer un score. Ces données ne correspondent pas à des valeurs numériques que nous pouvons utiliser pour la suite, et nous les considérons comme des données manquantes. Nous avons ainsi des données manquantes pour 3 différents patients sur les fonctions visuelles et cérébrales. Nous faisons le choix de supprimer les 2 variables correspondant à ces 2 sous-scores, bien qu'une alternative était de supprimer les patients concernés.

Concernant les données de marche, elles sont sous la forme de séries temporelles de quaternions et n'ont donc pas la même forme que des données habituellement traitées, de la forme lignes × colonnes. Un travail de pré-traitement et de transformation des données est ainsi nécessaire avant de poursuivre notre étude et d'appliquer les méthodes détaillées plus tôt.
Les figures 11 et 12 montrent les données originales avant le travail de transformation des données.



	tronc_cx	pyramid	sensitive	sphincter	cerebrale	autre
1	0	4	2	2	0	0
2	1	2	2	2	0	0
3	1	2	2	2	0	0
4	0	3	3	2	0	0
5	0	0	0	0	0	0
6	0	2	2	1	0	0

6 rows

Figure 9: Premières observations du jeu de données des données cliniques

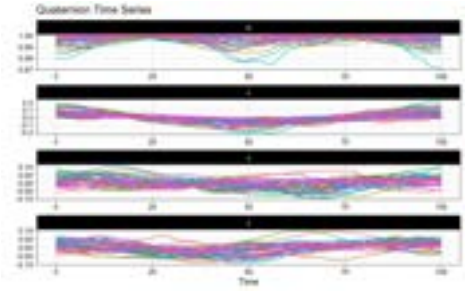


Figure 10: SDM contenues dans le jeu des données de marche

L'approche utilisée pour continuer notre analyse est l'utilisation de l'**Analyse en Coordonnées Principales** (PCoA), méthode détaillée en annexe. Nous décidons d'appliquer cette méthode aux deux jeux de données afin d'avoir des données homogènes. Différentes étapes sont appliquées aux jeux de données initiaux avant d'obtenir les données finales :

1. Calcul de la matrice des distances
2. Application de la PCoA avec correction Cailliez, qui prend en entrée la matrice de distance

La matrice de distance pour les données cliniques est calculée à l'aide de la distance de Gower [9] tandis que la distance entre QTS est donnée par la distance elastic shape [12].

PCoA :
Une fois les matrices de distance obtenues, nous réalisons une PCoA pour chaque matrice. Pour le choix des composantes, nous décidons de conserver 80% de l'inertie dans chacun des cas. Cela correspond à 13 composantes conservées pour les données cliniques, et 21 composantes conservées pour les données de marche, comme nous pouvons le voir sur les graphiques ci dessous (Figures 13 et 14).

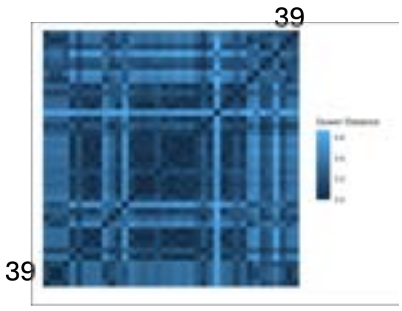


Figure 11: Matrice de distance des données cliniques

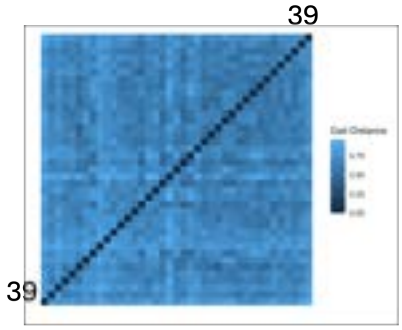


Figure 12: Matrice de distance des données de marche

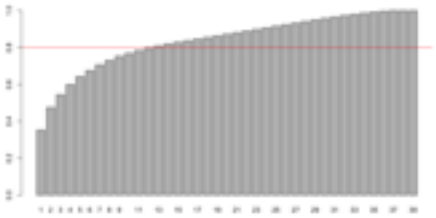


Figure 13: Pourcentage de variance expliquée : données cliniques

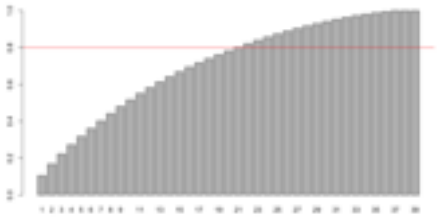


Figure 14: Pourcentage de variance expliquée : données de marche

Les composantes ainsi conservées constituent nos matrices finales que nous utiliserons pour la suite. Une rapide interprétation des composantes de la PCoA est possible dans chaque cas en regardant les corrélations avec les variables initiales. Les matrices des corrélations sont données en Figures 15 et 16.

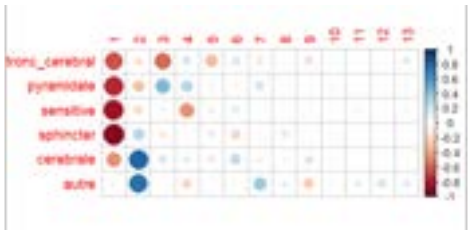


Figure 15

Figure 16

Dans le cas des données cliniques on observe que la première composante représente les fonctions sphinctérienne, sensitive, pyramidale, du tronc cérébral et cérébrale tandis que la deuxième composante représente la fonction cérébrale et les autres fonctions neurologiques.

Pour les données de marche, on s'intéresse aux paramètres spatio-temporels. On observe notamment que la première composante est corrélée à la vitesse angulaire moyenne d'un cycle de marche et à l'amplitude moyenne. La deuxième composante est corrélée à la durée moyenne de la phase d'appui.

Les étapes de transformation des données sont résumées dans le schéma ci-dessous :

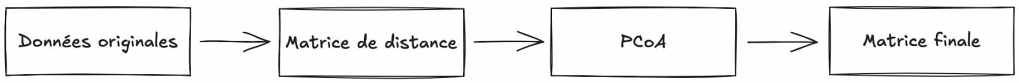


Figure 17: Etapes de transformation des données

Et les données finales que nous utilisons pour la suite sont des matrices de la forme :

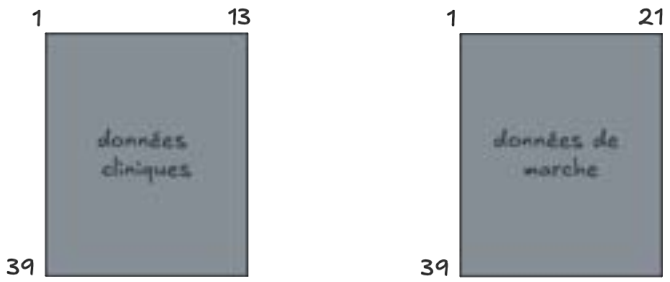


Figure 18: Matrices finales représentant les données cliniques et les données de marche

3.2 Données MS-CSI

3.2.1 Description des données

Une nouvelle étude a été menée par la suite. Intitulée MS-CSI, cette étude multicentrique inclut cette fois des patients nantais et rennais dans la cohorte OFSEP-HD. Les critères d'inclusion sont les mêmes que précédemment et un total de 100 patients répartis sur les deux sites est attendu pour l'inclusion.

Les données récupérées regroupent 49 patients pour lesquels des données cliniques, des données de marche et des données d'imagerie cérébrale et médulaire ont été récoltées.
L'évaluation clinique réalisée par les neurologues permet l'évaluation des scores EDSS et donne plusieurs autres informations : l'âge, le sexe, la taille, le poids et les temps réalisés par le patient lors du T25FW. Les signatures de marche ont également pu être récupérées lors de l'exécution du T25FW.

Enfin des acquisitions IRM ont été effectuées sur ces patients et nous permettent de rajouter une source de données à notre étude. Les lésions au niveau du cerveau et de la

moelle épinière sont localisées et retranscrites dans un tableau de données. On distingue les lésions cervicales et thoraciques de la moelle épinière.
Cette étude nous permet donc de considérer trois jeux de données distincts :

- Les données cliniques
- Les données de marche
- Les données IRM

Les données cliniques et les données de marche sont similaires à celles de l'étude AMIES. Les données IRM contiennent 26 variables, détaillées en annexe. On a l'ensemble des lésions au niveau du cerveau, de la moelle cervicale et de la moelle thoracique, ainsi que les lésions pour les tract cortico-spinaux gauche et droit de chaque région d'intérêt. Le volume total de la substance blanche est aussi donné à chaque fois.

3.2.2 Pré-traitement des données

Comme pour les données AMIES, un travail de pré-traitement des données est nécessaire. D'abord, la segmentation pour la moelle thoracique n'a pas bien fonctionné et les lésions n'ont quasiment pas été détectées, nous retirons donc des données toute la partie thoracique de la moelle. De plus, certains patients n'ont pas réalisé l'examen clinique, l'IRM ou bien n'ont pas de SDM, nous les retirons donc de l'analyse. Nous enlevons également les patients avec des données IRM de mauvaise qualité ou ayant des données manquantes, ce qui nous donne un total de 40 patients.

Comme pour les données AMIES, certains patients ont des sous-scores EDSS de "X". Deux fonctions neurologiques sont concernées : la fonction cérébelleuse et les autres fonctions neurologiques. Afin de ne pas retirer davantage de patients, nous supprimons ces fonctions de nos données cliniques.
On peut alors visualiser les trois types de données utilisées :

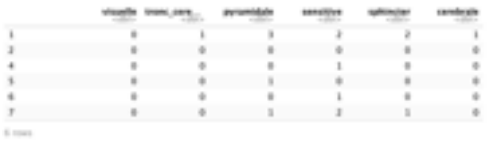


Figure 19: Premières observations du jeu des données cliniques

Figure 20: Premières observations du jeu des données IRM

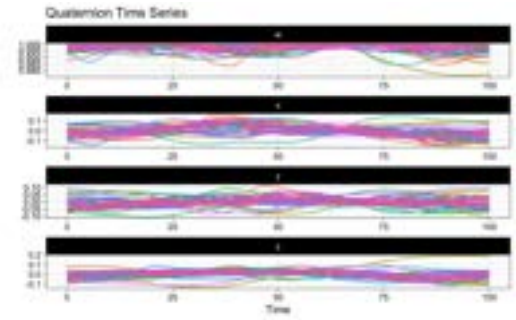


Figure 21: SDM contenues dans le jeu des données de marche

Nous réalisons de nouveau une PCoA sur les données de marche mais nous décidons de ne pas l'appliquer aux données cliniques et IRM afin de permettre une interprétation plus simple pour la suite. Comme précédemment, la distance elastic shape est utilisée comme distance entre les QTS. Nous retenons 80% de l'inertie expliquée, ce qui correspondant aux 20 premières composantes de la PCoA (voir Figure 23).

Les données de marche sont donc donnés par la matrice de dimension 40×20, contenant les 20 composantes de la PCoA.

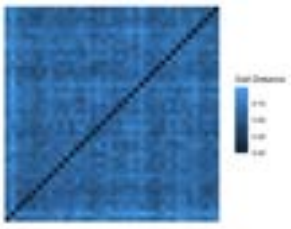


Figure 22: Matrice de distance entre QTS

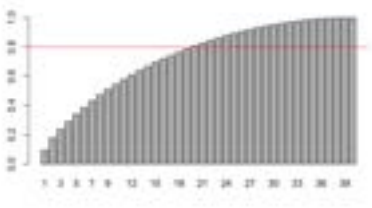


Figure 23: Pourcentage de variation expliquée par les composantes de la PCoA

Les données finales que nous utiliserons par la suite ont la forme suivante :

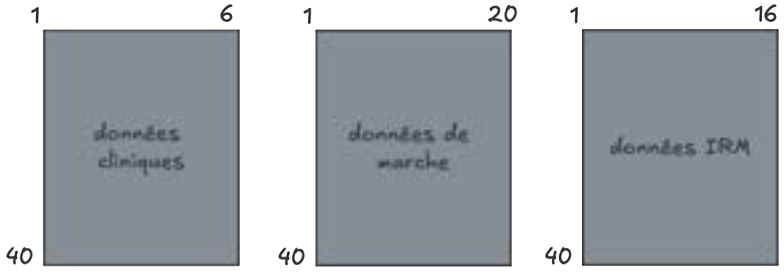


Figure 24: Matrices finales représentant les données cliniques, de marche et IRM

4 Résultats et discussion

4.1 CCA

Nous utilisons la fonction `cca` du package `multiblock` [3] sur les données de cliniques et les données de marche.
Tout d'abord la statistique de Wilks testant l'hypothèse de nullité des corrélations nous permet de déterminer le nombre de composantes à conserver. On obtient les résultats suivants :

Composantes conservées	p-value
1	0.28
2	0.71
3	0.91
4	0.99

Table 2: Résultats de la statistique de Wilks

On observe des résultats peu concluants puisqu'on ne rejette l'hypothèse nulle dans aucun cas. On conserve alors une composante de chaque bloc pour la suite, que l'on note f_{11} pour le bloc X_1 des EDSS, et f_{21} pour le bloc X_2 des SDM.

Le calcul des redondances nous indique que la composante f_{11} n'explique que 7.69% de son propre groupe et 4.64% de l'autre groupe, tandis que la composante f_{21} n'explique que 4.76% de son propre groupe et 7.49% de l'autre groupe. Nous cherchons alors l'information apportée par ces composantes.
La Figure 25 montre une corrélation entre la composante f_{11} et la fonction sensitive. De plus, en s'intéressant aux autres variables, on observe une corrélation avec le temps réalisé lors du T25FW (voir Annexe).

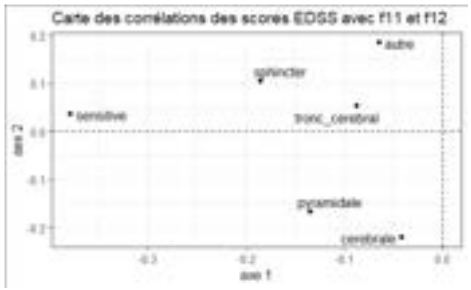


Figure 25: Carte des corrélations des composantes avec les scores EDSS

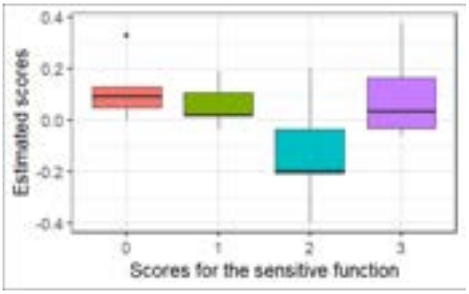


Figure 26: Distribution des scores pour la fonction sensitive

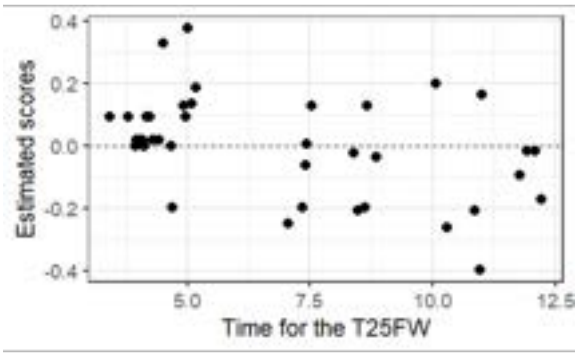


Figure 27: Distribution des scores pour le temps réalisé lors du T25FW

La CCA renvoie la variabilité commune aux deux blocs, on peut donc en déduire que les EDSS et les SDM donnent de l'information relative à la vitesse de marche lors de la réalisation du T25FW et à la fonction sensitive. Rappelons que la fonction sensitive est liée à la proprioception et la douleur, un score élevé pour cette fonction peut ainsi impacter la démarche d'un individu atteint de SEP, et peut être mis en lien avec sa vitesse de marche. Notons cependant que les composantes sélectionnées ne sont pas significatives, les résultats doivent donc être interprétés avec prudence.

4.2 PLS2

Nous utilisons la fonction `plsreg2` du package `plsdepot` [4] afin d'expliquer les données de marche, c'est-à-dire les SDM, à l'aide des scores EDSS.
Le Q2 de Stone-Geisser est un indicateur du choix des composantes à conserver. Nous n'obtenons cependant pas de composante significative, nous décidons de conserver une composante de chaque groupe : f_{11} et f_{21}. Comme précédement, on obtient la part de variance expliquée par chaque composante; f_{11} explique 7.69% de son groupe et 4.63% de l'autre groupe des SDM. f_{21} explique 4.44% de son groupe et 7.69% de l'autre groupe. Comme pour la CCA, on retrouve peu de variabilité expliquée par ces composantes.

La PLS2 permet d'obtenir des composantes liées au temps du T25FW. On peut en effet observer sur la Figure 28 une séparation entre les patients ayant mis plus de temps à réaliser le T25FW et ceux qui sont plus rapides.

Comme en CCA, les composantes conservées ne sont pas significatives, leur interprétation est donc à réaliser avec précaution.

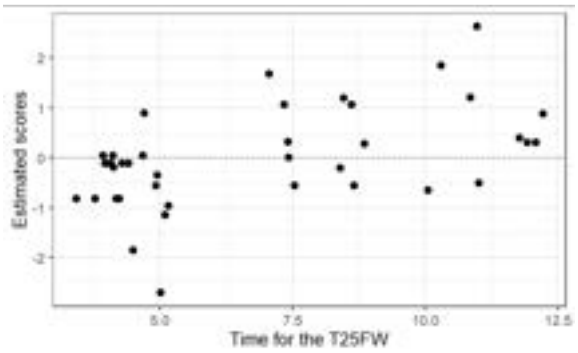


Figure 28: Distribution des scores pour le temps lors du T25FW

4.3 AJIVE

Nous appliquons mainstenant l'algorithme AJIVE à nos données. La première étape est celle de l'extraction du signal et donc du choix des rangs initiaux. On obtient les screeplots suivants, sur lesquels sont représentées les valeurs singulières :

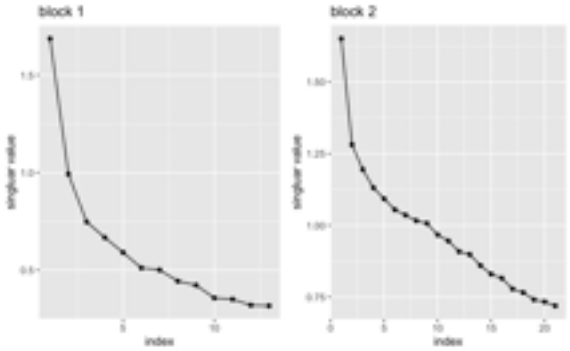


Figure 29: Screeplot des données cliniques et de marche

On observe plusieurs sauts sur ces graphiques, plusieurs rangs initiaux sont alors possibles.

- Données cliniques : on peut conserver 3, 6, 8 ou 10 valeurs singulières.
- Données de marche : on peut conserver 3, 9, 11 ou 13 valeurs singulières.

On a alors 16 combinaisons possibles de rangs initiaux. Pour chacune de ces combinaisons possibles, on obtient un espace joint à une dimension. Les composantes jointes obtenues par chaque combinaison sont représentées sur la figure 30. On observe une certaine stabilité du sous-espace joint obtenu.

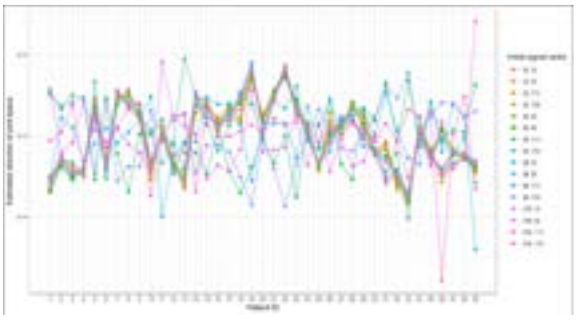


Figure 30: Composantes jointes estimées pour chacune des 16 combinaisons

Le choix final du rang du signal reste complexe car il n'existe pas de méthode claire et définie pour le déterminer. Nous faisons le choix de conserver un rang initial de 3 pour les données cliniques et un rang initial de 11 pour les données de marche, afin de nous concentrer sur l'information apportée par les premières composantes qui sont aussi celles avec le plus de variabilité car elles sont obtenues par PCoA.

Variabilité jointe

On cherche maintenant à savoir l'information apportée par la composante jointe obtenue. Pour cela nous cherchons quelles variables parmi les variables initiales sont liées à cette composante. Des tests de corrélation et d'analyse de la variance (ANOVA) sont donnés en annexe.

Nous observons que la composante jointe est corrélée avec le temps réalisé lors du T25FW.

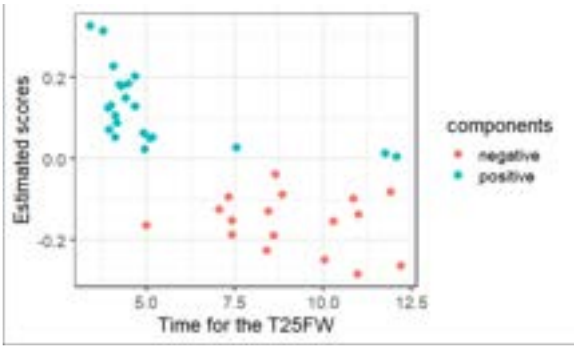


Figure 31: Représentation du T25FW dans l'espace joint

Cette composante oppose ainsi les patients marchant lentement, et qui mettent donc plus de temps à faire le test, lors du T25FW à ceux marchant plus vite. On distingue ainsi les patients plus atteints par la maladie d'un côté et ceux avec moins de problème de marche de l'autre.

De même, la composante jointe est liée à plusieurs sous-scores EDSS : elle implique la

fonction sensitive, la fonction sphinctérienne et la fonction pyramidale. Enfin, elle est aussi liée à l'un des paramètres spatio-temporels : l'amplitude moyenne d'un cycle.

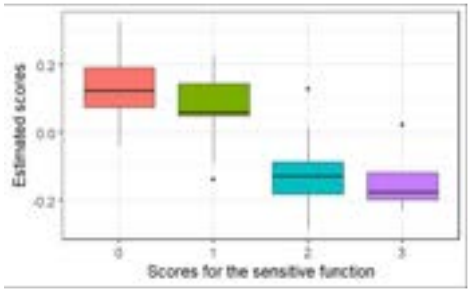


Figure 32: Distribution des scores pour la fonction sensitive

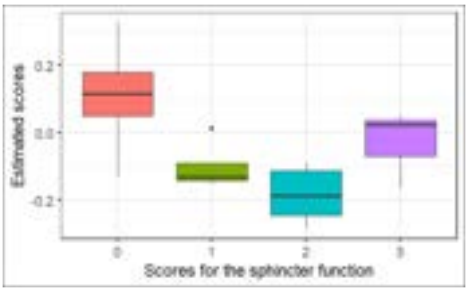


Figure 33: Distribution des scores pour la fonction sphinctérienne

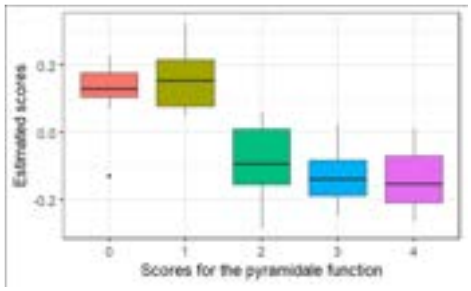


Figure 34: Distribution des scores pour la fonction pyramidale

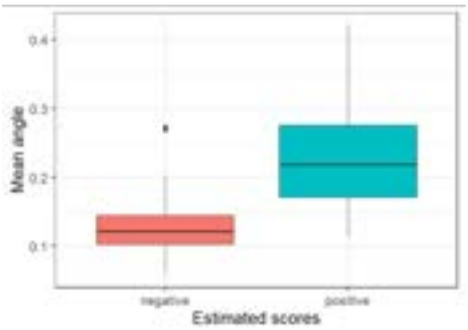


Figure 35: Distribution des scores pour la l'amplitude moyenne d'un cycle

La composante jointe oppose ainsi les patients avec des sous-scores élevés, donc plus malades, à ceux avec des sous-scores faibles pour les fonctions sensitive, sphinctérienne et pyramidale. Ces fonctions sont liées à la marche, elles concernent la motricité, la proprioception, ou encore la contraction des muscles.

Finalement, l'information commune aux scores EDSS et à la SDM concerne principalement directement la démarche et les difficultés que peuvent avoir les malades à marcher. D'un côté nous retrouvons les patients avec des scores plus élevés (scores de 2, 3, voire 4) pour les fonctions concernées qui ont ainsi plus de difficultés à marcher : leur vitesse de marche est plus lente et l'amplitude lors de la marche est plus faible. De l'autre, les patients semblent avoir moins de difficultés lors de la marche, avec une amplitude et une vitesse de marche plus élevées.

Variabilité individuelle des données cliniques

Nous nous intéressons maintenant à l'information spécifique aux scores EDSS. Le sous-espace

individuelle relatif aux données cliniques est un espace à 3 dimensions, nous cherchons donc l'information apportée par chacune des composantes individuelles.

La première composante de l'espace individuelle est principalement liée à la fonction du tronc cérébral (voir 36) tandis que la 2ème composante est liée aux fonctions pyramidale et sphinctérienne. Enfin, la 3ème composante représente l'aide lors de la marche ainsi que la fonction du tronc cérébral.

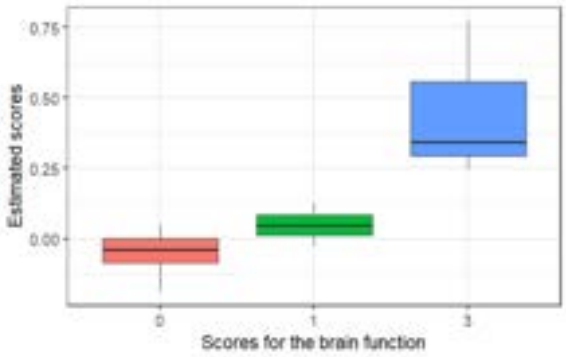


Figure 36: Distribution des scores pour la fonction cérébrale

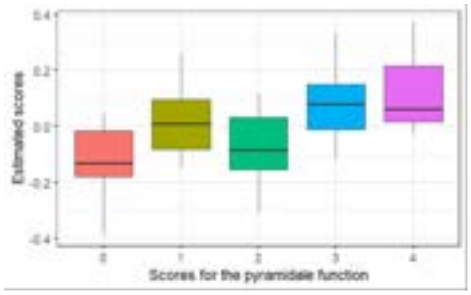


Figure 37: Distribution des scores pour la fonction pyramidale

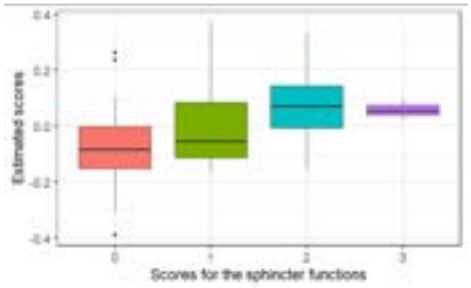


Figure 38: Distribution des scores pour la fonction sphinctérienne

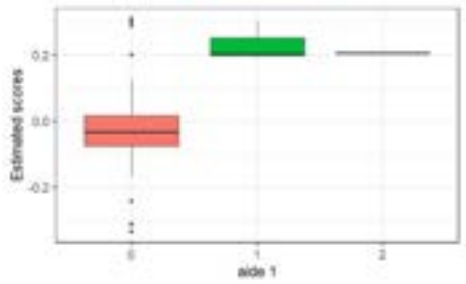


Figure 39: Distribution des scores selon l'aide lors de la marche

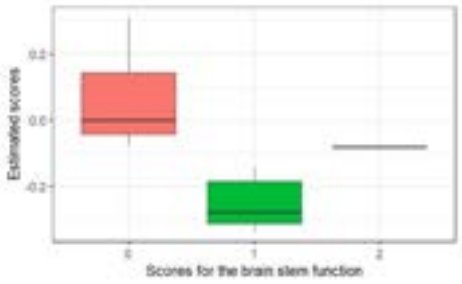


Figure 40: Distribution des scores pour la fonction du tronc cérébral

L'information spécifique aux scores EDSS concerne toujours la marche des patients, avec les fonctions pyramidale et sphinctérienne, mais illustre certaines difficultés lors de la marche qui n'apparaissent pas avec la SDM, avec l'aide lors de la marche notamment. L'information apportée concerne également d'autres fonctions neurologiques moins liées à la marche et gérant le rythme cardiaque, la respiration ou encore la mémoire et l'humeur.

Variabilité individuelle des données de marche

Le sous-espace individuelle propre aux données de marche, c'est-à-dire à la SDM, est un espace à 11 dimensions. Plusieurs des composantes engendrant cet espace sont liées aux paramètres spatio-temporels, comme le montre la matrice des corrélations ci-dessous.

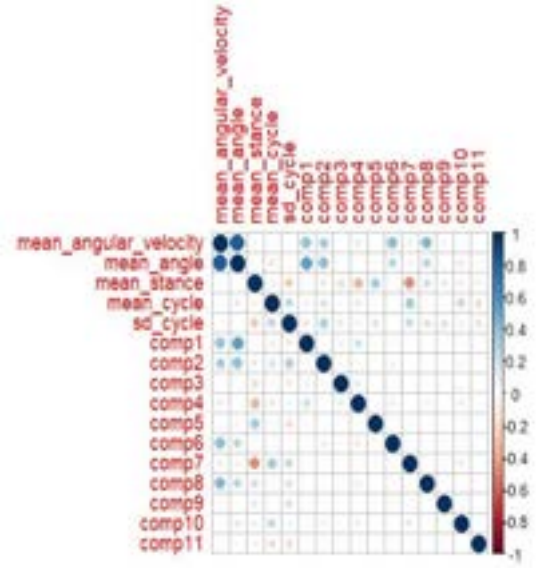


Figure 41: Matrice des corrélations entre les PST et les composantes

Les SDM calculées pour chaque patient semble ainsi donner majoritairement des informations sur les paramètres spatio-temporels, avec l'amplitude et la vitesse angulaire. On retrouve alors d'une part des patients avec une amplitude et une vitesse angulaire importante, avec une durée de la phase d'appui plus faible, indiquant des patients avec des difficultés à marcher. De l'autre, ce sont des patients moins impactés par la maladie au niveau de la marche.

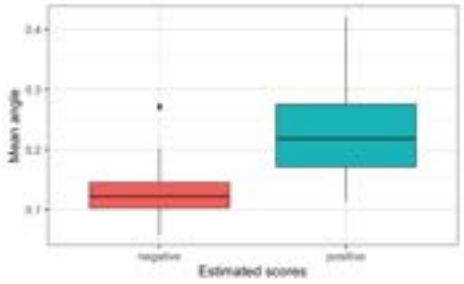


Figure 42: Distribution des scores selon l'amplitude moyenne d'un cycle

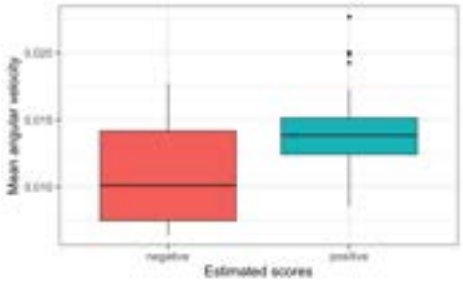


Figure 43: Distribution des scores selon la vitesse angulaire moyenne

Finalement, les scores EDSS et la SDM apportent de l'information commune et permettent tous deux d'étudier la démarche des patients et l'impact de la maladie sur celle-ci. Ils donnent également une certaine information non apportée par les autres données, nous indiquant donc l'importance de considérer les deux approches pour suivre l'évolution de la SEP chez les patients.

Conclusion et perspectives

L'analyse des données AMIES a permis d'étudier différentes méthodes d'intégration de données et de les prendre en main, mais aussi d'adapter chaque méthode aux données disponibles. Il ressort de ces trois applications la présence d'information commune entre les trois sources de données permettant le suivi de l'atteinte de la marche des patients.

Bien que les méthodes PLS2 et CCA nous permettent de conclure sur cette information jointe, ces méthodes semblent moins adaptées au format de nos données, et plus particulièrement des données de marche obtenues avec la SDM.
La méthode AJIVE permet quand à elle de vérifier à la fois la présence d'information jointe entre les données cliniques et les données de marche, confirmant l'intérêt de l'utilisation de la SDM comme moyen de suivi des troubles ambulatoires, et d'information propre à chaque jeu de données. L'information individuelle confirme également que la SDM permet d'analyser l'atteinte de la marche et complète ainsi l'utilisation des scores EDSS.
Les résultats ainsi obtenus sont satisfaisants mais ils sont à approfondir en améliorant notamment la sélection des rangs initiaux des signaux dans l'algorithme AJIVE, et en appliquant d'autres méthodes multiblocs à nos données.

L'application des méthodes aux données MS-CSI n'a pas été traitée dans ce rapport, l'étude de ces données comprenant les IRM s'arrêtant à l'étape de pré-traitement, elle sera réalisée à l'aide de la méthode AJIVE. L'utilisation d'autres méthodes telles que ComDim ou MB-PLS, a été envisagée mais sera peut-être, selon les résultats, limitée aux données AMIES.

Une fois l'évaluation de l'association entre la SDM, l'EDSS et les données IRM réalisée, nous construirons des groupes homogènes basés sur la sévérité d'atteinte de la marche, correspondant aux profils type de SDM.

Références

[1] Ana Arribas-Gil and Juan Romo. "Shape outlier detection and visualization for functional data: the outliergram". In: *Biostatistics* 15.4 (2014), pp. 603–619.

[2] Arsep. *Symptômes et poussées*. URL: `https://www.arsep.org/fr/170-symptomes%20et%20pouss%C3%A9e.html`.

[3] *Package multiblock: Multiblock Data Fusion in Statistics and Machine Learning*. URL: `https://cran.r-project.org/web/packages/multiblock/index.html`.

[4] *Package plsdepot: Partial Least Squares (PLS) Data Analysis Methods*. URL: `https://cran.r-project.org/web/packages/plsdepot/index.html`.

[5] Pierre Drouin. "Amélioration du suivi des patients atteints de maladies neurodégénératives à l'aide d'objets connectés". PhD thesis. Nantes université, 2022.

[6] Pierre Drouin et al. "Gait impairment monitoring in multiple sclerosis using a wearable motion sensor". In: *Medical Case reports and Reviews* 5 (2022), pp. 1–5.

[7] Sep Ensemble. *L'EDSS*. URL: `https://www.sep-ensemble.fr/symptomes-diagnostic/quest-ce-que-iedss`.

[8] Qing Feng et al. "Angle-based joint and individual variation explained". In: *Journal of multivariate analysis* 166 (2018), pp. 241–265.

[9] John C Gower. "A general coefficient of similarity and some of its properties". In: *Biometrics* (1971), pp. 857–871.

[10] Francesca Ieva and Anna Maria Paganoni. "Component-wise outlier detection methods for robustifying multivariate functional samples". In: *Statistical Papers* 61.2 (2020), pp. 595–614.

[11] Inserm. *Sclérose en plaques*. URL: `https://www.inserm.fr/dossier/sclerose-en-plaques-sep/`.

[12] Sebastian Kurtek et al. "Statistical modeling of curves using shapes and related features". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1152–1165.

[13] Klervi Le Gall et al. "Generation of synthetic gait data: application to multiple sclerosis patients' gait patterns". In: *The International Journal of Biostatistics* (2024).

[14] Jianming Miao and Adi Ben-Israel. "On principal angles between subspaces in Rn". In: *Linear algebra and its applications* 171 (1992), pp. 81–98.

[15] Erica Ponzi, Magne Thoresen, and Abhik Ghosh. "RaJIVE: Robust Angle Based JIVE for Integrating Noisy Multi-Source Data". In: *arXiv preprint arXiv:2101.09110* (2021).

[16] Ministère de la Santé. *La sclérose en plaques*. URL: https://sante.gouv.fr/soins-et-maladies/maladies/maladies-neurodegeneratives/article/la-sclerose-en-plaques.

[17] Gilbert Saporta. *Probabilités, analyse des données et statistique*. Editions technip, 2006.

[18] Age K Smilde, Tormod Næs, and Kristian Hovde Liland. *Multiblock data fusion in statistics and machine learning: Applications in the natural and life sciences*. John Wiley & Sons, 2022.

[19] Michel Tenenhaus. *La régression PLS: théorie et pratique*. Editions technip, 1998.

[20] John Voight. *Quaternion algebras*. Springer Nature, 2021.

[21] Wikipédia. *Angles d'Euler*. URL: https://fr.wikipedia.org/wiki/Angles_d%27Euler.

Annexes

Jeu de données AMIES

Description

Le fichier `AMIES.rds` correspond au jeu de données AMIES et contient 39 observations correspondant à des patients atteints de SEP, ainsi que plusieurs variables numériques catégorielles, ordinales et fonctionnelles.

Nom	Description
`id`	ID du patient
`num_patient`	numéro d'inclusion du patient
`age`	âge du patient
`sexe`	sexe du patient
`taille_cm`	taille du patient (en cm)
`poigs_kg`	poids du patient (en kg)
`lat_ms`	latéralisation du membre supérieur (1=Droite, 0=Gauche, 2=Ambilatéralité)
`lat_mi_pied_appui`	latéralisation du membre inférieur (1=Droite, 0=Gauche, 2=Ambilatéralité)
`duree_ms_ans`	durée de la pathologie (en années)
`annee_derniere_pousse`	année de la dernière année de poussée de la maladie
`nb_poussees_n_1`	nombre de poussées de symptômes au cours de l'année précédente
`visuelle`	fonction visuelle (de 0 à 6)
`tronc_cerebral`	fonction du tronc cérébral (de 0 à 5)
`pyramidale`	fonction pyramidale (de 0 à 4)
`cerebelleuse`	fonction cérébelleuse (de 1 à 5)
`sensitive`	fonction sensitive (de 0 à 6)
`spincter`	fonction sphinctérienne (de 0 à 6)
`cerebrale`	fonction cérébrale (de 0 à 5)
`autre`	autres fonctions neurologiques touchées (1=Oui, 0=Non)
`edss`	score EDSS (de 0 à 10)
`aide_1`	aide lors de la marche pour le premier aller-retour lors du T25FW (0=Aucune, 1=Unilatérale, 2=Bilatérale)

`aide_2`	aide lors de la marche pour le deuxième aller-retour lors du T25FW (0=Aucune, 1=Unilatérale, 2=Bilatérale)
`dmt_traitement_de_fond`	nom du traitement de fond (0 si aucun)
`fampyra`	utilisation d'un médicament utilisé dans les troubles de la marche (1=Oui, 0=Non)
`forme_sep`	forme de la maladie (0=RR, 1=SP, 2=PP)
`annee_progression`	année de la progression de la forme RR en forme SP
`T1`	temps réalisé lors de l'aller pour le premier aller-retour (en seconde)
`T2`	temps réalisé lors du retour pour le premier aller-retour (en seconde)
`T3`	temps réalisé lors de l'aller pour le deuxième aller-retour (en seconde)
`T4`	temps réalisé lors du retour pour le deuxième aller-retour (en seconde)
`T25FW_mean`	moyenne du temps réalisé lors des 4 aller-retour (en seconde)
`qts`	SDM sous forme de tibble contenant les 4 composantes w, x, y, z

Table 3: Variables des données AMIES

Résultats

AJIVE :

Variabilité jointe :

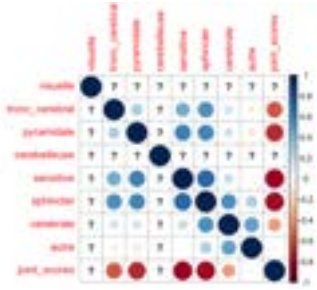


Figure 44: Corrélation entre les sous-scores et la composante jointe

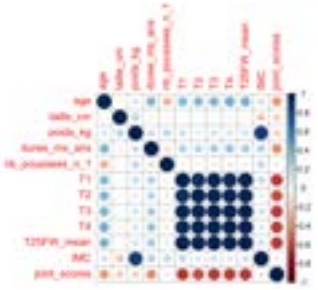


Figure 45: Corrélation entre les autres données cliniques et la composante jointe

Ces matrices permettent de visualiser les corrélations positives et négatives de la composante jointe avec les autres variables.

Variabilité individuelle des données cliniques :

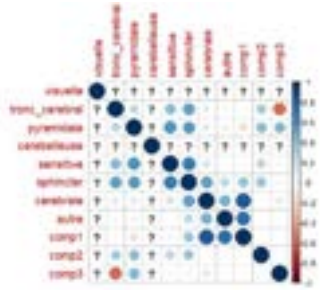


Figure 46: Corrélation entre les sous-scores et les composantes individuelles

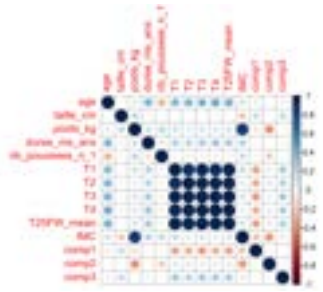


Figure 47: Corrélation entre les autres données cliniques et les composantes individuelles

On peut visualiser les corrélations entre chacune des composantes et les différentes variables quantitatives.

Jeu de données MS-CSI

Les données cliniques de l'étude MS-CSI contiennent les variables suivantes :

Nom	Description
`id`	ID du patient
`num_patient`	numéro d'inclusion du patient
`age`	âge du patient
`sexe`	sexe du patient
`taille_cm`	taille du patient (en cm)
`poigs_kg`	poids du patient (en kg)
`visuelle`	fonction visuelle (de 0 à 6)
`tronc_cerebral`	fonction du tronc cérébral (de 0 à 5)
`pyramidale`	fonction pyramidale (de 0 à 4)
`cerebelleuse`	fonction cérébelleuse (de 1 à 5)
`sensitive`	fonction sensitive (de 0 à 6)
`spincter`	fonction sphinctérienne (de 0 à 6)
`cerebrale`	fonction cérébrale (de 0 à 5)
`autre`	autres fonctions neurologiques touchées (1=Oui, 0=Non)
`edss`	score EDSS (de 0 à 10)
`T1`	temps réalisé lors de l'aller pour le premier aller-retour (en seconde)
`T2`	temps réalisé lors du retour pour le premier aller-retour (en seconde)
`T3`	temps réalisé lors de l'aller pour le deuxième aller-retour (en seconde)
`T4`	temps réalisé lors du retour pour le deuxième aller-retour (en seconde)
`T25FW_mean`	moyenne du temps réalisé lors des 4 aller-retour (en seconde)

Table 4: Variables des données cliniques de MS-CSI

Les données IRM contiennent pour chacune des zones cerveau, moelle cervicale et moelle thoracique, les variables suivantes :

- `lesion vol whole WM` : volume lésionnel de la substance blanche

- `vol lesion CST L` : volume lésionnel au niveau du tract corticospinal gauche

- `vol lesion CST R` : volume lésionnel au niveau du tract corticospinal droit

- `vol lesion CST L+R` : volume lésionnel au niveau des tracts corticospinaux gauche et droit
- `vol whole WM` : volume total de la substance blanche
- `vol CST L` : volume total du tract corticospinal gauche
- `vol CST R` : volume total du tract corticospinal droit
- `vol CST L+R` : volume total des tracts corticospinaux gauche et droit

Analyse en Coordonnées Principales

La PCoA est une méthode de réduction de dimension permettant d'obtenir une représentation euclidienne des dissimilarités.
On considère une matrice de dissimilarité D symétrique, à termes réels et de dimension $\mathbb{N} \times \mathbb{N}$:

$$\begin{pmatrix} d_{11} & \dots & d_{1N} \\ \vdots & \ddots & \vdots \\ d_{N1} & \dots & d_{NN} \end{pmatrix}$$

où d_{ij} est la valeur de la dissimilarité entre l'élément i et l'élément j.

On construit alors la matrice A :

$$A = -\frac{1}{2}[d_{ij}^2] = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{pmatrix}$$

On pose $W = [w_{ij}]$ la matrice des produits scalaires.

$$\forall i, j \in 1, ..., N, w_{ij} = [a_{ij} - a_{i.} - a_{j.} + a_{..}]$$

avec :

- $a_{i.} = \frac{1}{N} \sum_{j=1}^{N} a_{ij}$: moyenne des colonnes de A
- $a_{.j} = \frac{1}{N} \sum_{i=1}^{N} a_{ij}$: moyenne des lignes de A
- $a_{..} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}$: moyenne générale de A

L'étape suivante est le calcul des valeurs propres $\lambda_1, ..., \lambda_{N-1}$ et des vecteurs propres $[v_1, ..., v_{N-1}]$ de W. On a alors deux cas possibles :

- Soit toutes les valeurs propres sont positives : dans ce cas D est une matrice de distance euclidienne et on peut réaliser une PCoA sans correction.
- Soit il existe au moins une valeur propre négative : celles-ci ne peuvent pas être représentées dans un espace euclidien de dimension inférieure. On peut alors :
  - Conserver uniquement les valeurs propres positives pour obtenir une représentation dans un espace euclidien. On peut alors perdre des informations importantes sur les données.
  - Ou alors corriger également les valeurs propres négatives

Les coordonnées principales sont données par :

$$x_{ik} = \sqrt{\lambda_k} v_{ik} \text{ pour l'individu } i \text{ et la dimension } k$$

et on note $Y = \begin{pmatrix} x_1 & x_2 & \dots \end{pmatrix}$ la matrice des coordonnées principales.

Dans le cas de valeurs propres négatives et de leur correction, la méthode repose sur la transformation de la matrice de dissimilarité non euclidienne D en une matrice de distance euclidienne. C'est ce qu'on appelle une PCoA avec correction. Il existe trois types de correction.

1. La correction Lingoes :
   Les termes non diagonaux de la matrice D prennent les valeurs $(d_{ij}^2 + h)^{\frac{1}{2}}$ où h est la valeur absolue de la plus petite valeur propre de la matrice $\Delta_1 = (I - \frac{1}{N}\mathbb{1}\mathbb{1}^t)A(I - \frac{1}{N}\mathbb{1}\mathbb{1}^t)$ où $A = -\frac{1}{2}[d_{ij}^2]$.

2. La correction Cailliez :
   Les termes non diagonaux de la matrice D prennent les valeurs $d_{ij} + k$ où k est la plus grande valeur propre de la matrice $\begin{pmatrix} 0 & 2\Delta_1 \\ -I & 2\Delta_2 \end{pmatrix}$ et $\Delta_2 = (I - \frac{1}{N}\mathbb{1}\mathbb{1}^t)B(I - \frac{1}{N}\mathbb{1}\mathbb{1}^t)$ et $B = -\frac{1}{2}[d_{ij}]$.

3. La correction racine carrée :
   Cette méthode consiste à appliquer la fonction racine carrée à la matrice D. Cette méthode ne garantit cependant pas l'obtention d'une matrice de distance euclidienne, même en appliquant la racine carrée plusieurs fois à la matrice D.