Méthode non paramétrique pour l'analyse et la classification des données fonctionnelles

Papa MBAYE[1,2] Anne-Françoise YAO[1] Chafik SAMIR[2]

[1] Laboratoire de Mathématiques Blaise Pascal CNRS UMR 6620
[2] Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes CNRS UMR 6158

papa_alioune_meissa.mbaye@uca.fr, anne.yao@uca.fr, chafik.samir@uca.fr

Résumé

L'analyse de données fonctionnelles joue un rôle important dans beaucoup de domaines de la santé publique et des applications biomédicales. En particulier, de telles méthodes statistiques fournissent des outils permettant de recaler, de comparer et de modéliser des données constituées de mesures corrélées. Dans ce travail, nous présenterons une nouvelle approche d'analyse de régression pour la classification de données fonctionnelles. D'abord, nous commencerons par analyser les observations fonctionnelles en faisant un recalage temporel. Ensuite, nous allons étudier différentes représentations standards de la littérature et estimer le modèle de régression appropriée comme une fonction de densité. Enfin, un exemple d'application, constitué de personnes ayant l'Arthrite Rhumatoïde (AR) et de personnes bien portantes comme groupe de référence, est présenté.

Mots Clef

Analyse de données fonctionnelles, Régression non paramétrique, Recalage.

Abstract

Functional data analysis plays an increasingly important role in many public health and biomedical applications. In particular, such statistical methods provide tools for warping, comparing, averaging, and modeling data involving correlated measurements. In this paper, we present a new approach of regression analysis for classification of functional data. First, we analyze functional observations to capture their key spatio-temporal patterns by searching optimal warping and then estimate the regression function. Next, we investigate different standard representations from literature and estimate the appropriate regression model as a density function. Finally, an example of application involving patients with Rheumatoid Arthritis and healthy subjects as a reference group, is presented.

Keywords

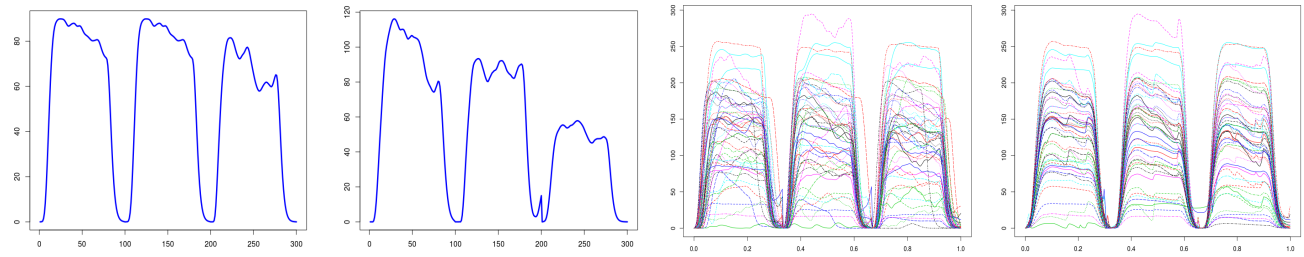
Functional Data Analysis, Nonparametric Regression, Registration, Time Warping.

1 Introduction

Analyser des données constituées de fonctions (courbes, surfaces ou d'autres fonctions), au lieu de vecteurs de scalaires, devient de plus en plus populaire [1, 2]. De tels problèmes nécessitent de considérer les courbes comme des fonctions continues et d'utiliser des représentations et analyses appropriées. Les méthodes de régression fonctionnelle ont été largement utilisées pour résoudre ce genre de problèmes [1, 3]. Récemment, différentes méthodes ont été proposées pour les régressions linéaires fonctionnelles. Cependant une étape clé pour analyser les données fonctionnelles temporelles est la capacité de capturer la variabilité temporelle, qui peut être considérée comme une transformation aléatoire du temps. En effet les variations obtenues au niveau des données collectées sont dues à plusieurs facteurs, incluant les outils de mesure et le comportement humain ; ce qui fait que les mêmes personnes observées peuvent donner lieu à différentes observations. La procédure de recalage pourrait ainsi être utilisée pour traiter cette variabilité temporelle qui est considérée comme une nuisance. Plusieurs alternatives ont été introduites pour représenter les courbes ou pour les comparer d'une manière invariante [2, 4].

L'arthrite est une maladie polymorphe qui est souvent caractérisée par un gonflement d'un ou de plusieurs articulations. D'après [5], l'arthrite est l'une des principales causes de l'incapacité physique qui affecte les jeunes et les personnes âgées, où les femmes sont plus touchées que les hommes. Malheureusement, il n'y a actuellement aucun remède pour l'arthrite et les traitements coûteux sont disponibles selon le type d'arthrite. Il y a plusieurs formes d'arthrites, dans lesquelles l'Arthrite Rhumatoïde, que l'on notera par la suite par AR, est la forme la plus commune

CEUR-WS.org/Vol-2133/rjcia-paper3.pdf



(a) (b) (c) (d)

Figure 1: Exemples de données fonctionnelles qui mesurent l'intensité de la force musculaire: (a) Fonction de la force de la main d'une personne bien portante, (b) Fonction de la force de la main d'une personne malade (AR modéré), (c) L'ensemble des fonctions musculaires avant recalage, et (d) fonctions musculaires après recalage utilisant notre méthode.

d'inflammation chronique [6].

Dans les diagnostics quotidiens, l'examen clinique est utilisé pour reconnaître les modèles spécifiques et les symptômes, et si nécessaire, il est confirmé par d'autres tests, i.e. imagerie IRM et les tests du sang. Malheureusement, de tels tests sont très chers pour les patients et longs pour les médecins. Les types de diagnostics mentionnés précédemment peuvent être utilisés pour automatiser la classification de la maladie, mais au stade précoce de l'AR, ces critères ne sont pas habituellement satisfaisants. Dans les années récentes, la recherche médicale a entraîné une nouvelle compréhension de l'AR ; en particulier, il est indiqué que les mesures de force de la main sont une technique bonne et peu coûteuse pour une évaluation préopératoire de personnes malades [7]. Bien que quelques des caractéristiques discrètes citées précédemment puissent être utile pour cet objectif, la fonction de force de la main contient plus d'informations de diagnostic et s'avère être un indicateur significatif sur la présence et le stade de la maladie. Dans cet article, nous nous concentrons sur cette nouvelle procédure de diagnostic. La fonction de force de la main d'une personne bien portante est donnée au niveau de la Figure 1(a) et celle d'une personne atteinte de l'AR au niveau de la Figure 1(b). Cette dernière montre un modèle clair de personnes malades où toutes les amplitudes de la force de la main ne sont pas très fortes et décroissent avec le deuxième et le troisième test. Cependant, en regardant les Figures 1(a) et 1(b), nous remarquons que le problème de classification entre les personnes bien portantes et les personnes malades est très difficile. Par ailleurs, pour illustrer l'importance du recalage, nous affichons les courbes originales avant recalage en 1(c) et après recalage en 1(d). Les fonctions considérées ici appartiennent à l'ensemble $\mathbb{L}^2([0,1],\mathbb{R}^+)$ car ces intensités sont enregistrées de manière continue durant un intervalle de temps $T = [0,1]$ et sont à valeurs dans $\mathbb{R}^+$. Récemment, l'analyse de données fonctionnelles a été proposée pour une étude plus générale. Bien qu'il existe une large littérature sur l'analyse statistique de fonctions, voir par exemple [2, 4, 8], quand on se limite sur l'analyse de fonctions qui nécessite le recalage temporel, la littérature est toujours relativement limitée [1, 9–13].

Dans ce travail, nous proposons un modèle de régression fonctionnelle non paramétrique pour diagnostiquer l'AR. Autrement dit, on cherchera d'abord à apprendre une fonction de régression et à partir de cette fonction utiliser un seuil pour faire la classification, c'est à dire pour prédire la présence ou l'absence de la maladie. A notre connaissance, l'analyse de la régression sur des données fonctionnelles complètes sous forme de signaux de force de la main, pour diagnostiquer l'AR, n'a pas été précédemment étudiée. Un modèle statistique approprié est nécessaire dans cette application pour modéliser ces données fonctionnelles. En particulier, nous nous intéressons à l'étude de la variabilité au sein des groupes de personnes malades et de personnes bien portantes en utilisant la méthode de régression fonctionnelle complète. Une difficulté au niveau de la main est le fait que les signaux bruts des forces de la main ne sont pas alignés dans le temps. Autrement dit, différents patients exerceront leur force à des temps différents, et alors, il devient important de découpler la quantité de force exercée (amplitude de fonction) et combien de temps la force a été exercée (phase de fonction). Ainsi, nous avons besoin d'un modèle statistique global pour l'analyse des données fonctionnelles de force de la main qui permet la séparation des variabilités d'amplitude et de phase. Le modèle récent dans [14] fournit une approche mathématique et statistique efficace pour la séparation amplitude-phase de données fonctionnelles, et par la suite l'analyse statistique de ces deux composantes. Nous adaptons cette méthode pour étudier les signaux de la force de la main et pour définir un nouveau modèle (représentation fonctionnelle couplée à un modèle de régression) basé sur des données fonctionnelles complètes dans le but de caractériser la

maladie par des méthodes d'apprentissage statistique. Donc à partir de variables (signaux), qui sont à valeurs dans un espace de fonctions, le modèle prédira si la personne est bien portante ou malade.

2 Modélisation et Analyse de données fonctionnelles

Nous proposons une nouvelle représentation de la force de la main qui exploite le stade de la maladie comme une distance appropriée des observations de référence. Ce travail est inspiré en premier par les diagnostics classiques basés sur le maximum des mesures de la force (et éventuellement de la vitesse d'atteinte) qui entraînaient une énorme perte d'informations pertinentes pour la classification de la maladie d'AR. Par conséquent, les précisions de la classification décroissent significativement quand la variabilité entre les personnes bien portantes croît. Ainsi pour améliorer l'analyse statistique complète, nous prenons une approche d'analyse de données fonctionnelles pour analyser les fonctions de force de la main qui représentent l'effort continu, répétitif fait par les personnes. Cette nouvelle représentation utilisée pour la classification des personnes atteintes de l'AR apporte des informations complémentaires et indispensables sur l'état de la maladie. L'intensité de la force de la main est représentée par une fonction absolument continue x définie sur un intervalle $I = [0, 1]$, pour simplifier. Comme montré dans la Figure 1, $x(t) = 0$ là où il n'y a pas d'effort : au début du test ($t = 0$), au temps de repos et à la fin du test ($t = 1$). On peut remarquer dès à présent que la variance des intensités des personnes bien portantes est faible, alors que celle des personnes malades est forte, dû aux changements progressifs causés par la maladie en évolution. Pour arriver à une telle conclusion, on a à définir un modèle approprié, qui fournit une distance appropriée et des outils d'analyses statistiques en vue d'obtenir des classifications précises (i.e. séparation du groupe des personnes bien portantes et du groupe des personnes malades). Une qualité importante d'un tel modèle est d'être capable de résumer efficacement et de capturer la variabilité dans les deux classes. De plus, on espère que la distance définie pourra fournir une mesure naturelle entre les signaux de la force de la main, permettant ainsi aux rhumatologues de quantifier le stade de gravité de la maladie d'AR, en se basant sur une personne bien portante (référence). Par la suite, nous décrirons les éléments nécessaires qui seront utilisés pour recaler les données fonctionnelles.

Supposons un échantillon de variables aléatoires $\{x_i, i = 1, ..., n\}$, où x_i est une fonction assez lisse définie dans un domaine unité de $\mathbb{R}$, et $\{y_i, i = 1, ..., n\}$ une suite de variables binaires. $y_i = 0$ si la personne est bien portante et $y_i = 1$ si la personne est malade. x_i et y_i ne sont pas généralement directement observables, au lieu de cela nous observons leurs discrétisations, avec du bruit aléatoire supplémentaire. Ainsi les données observées sont des vecteurs finis $(x_1; y_1); ...; (x_n; y_n)$. Nous supposons que ces erreurs sont gaussiennes de moyenne nulle et qu'elles sont indépendantes.

Supposons un ensemble de fonctions de force $\{x_i, i = 1, ..., n\}$, notre but est de trouver un ensemble de fonctions de reparamétrisation $\{\gamma_i^*, i = 1, ..., n\}$ (variabilité de phase) tel que les fonctions $\{x_i \circ \gamma_i^*, i = 1, ..., n\}$ soient alignées de manière optimale et alors ne varient qu'au niveau des amplitudes. γ est une fonction qui est définie par $\{\gamma : [0, 1] \to [0, 1]; \dot{\gamma} \succ 0\}$. Dans plusieurs publications précédentes, c'est la norme $\mathbb{L}^2$ pénalisée (norme $\mathbb{L}^2$ qui mesure l'écart entre deux fonctions plus une pénalisation sur la fonction de reparamétrisation) qui a été utilisée pour le recalage. Ces approches sont connues de ne pas bien fonctionner pour les fonctions de *pinching* (similaire au surapprentissage) et pour l'*asymétrie* de solutions [3]. Ce qui crée un effet sévère sur les analyses qui y découlent et cet effet vient du fait que la norme $\mathbb{L}^2$ n'est pas une métrique sur l'espace de fonctions modulo le groupe des reparamétrisations Γ. Ainsi dans ce papier, chaque fonction sera représentée par sa fonction q définie par $q(t) = \text{sign}(\dot{x}(t))\sqrt{|\dot{x}(t)|}$, où $\dot{x} = dx/dt$. Nous restreignons x d'être absolument continue parce que l'espace des résultats des fonctions q est $\mathbb{L}^2([0, 1], \mathbb{R})$, qui est l'ensemble des fonctions définies sur $[0, 1]$ et de carré intégrable. Si une fonction x est reparamétrisée par une fonction γ en $x \circ \gamma$, alors sa fonction q change et devient $(q \circ \gamma)\sqrt{\dot{\gamma}}$ et on la notera par $(q * \gamma)$. La propriété la plus importante de cette transformation est que $\|q\| = \|q * \gamma\|$ pour tout $\gamma \in \Gamma$, où $\|\cdot\|$ est la norme $\mathbb{L}^2$ de la fonction. Cette propriété permet de résoudre le problème de recalage optimal entre deux fonctions de force de la main x_1 et x_2 comme suit. Soit q_1 et q_2 leurs fonctions q. Alors la fonction de reparamétrisation optimale de x_2 à x_1 est donnée par $\gamma^* = \arg\inf_{\gamma \in \Gamma} \|q_1 - q_2 * \gamma\|$. La quantité à droite forme une distance appropriée dans l'espace quotient $\mathbb{L}^2/\Gamma$. Cette distance peut être utilisée pour définir des statistiques comprenant la moyenne de la fonction de force de la main, qui agira comme un modèle pour plusieurs recalages.

Le problème de phase et de séparation d'amplitude est lié aux fonctions de recalage non linéaire. Supposons $x : [0, 1] \to \mathbb{R}$ une fonction absolument continue et Γ l'ensemble de toutes les frontières préservant le difféomorphisme de $[0, 1]$ à lui même. Alors pour tout $\gamma \in \Gamma$, la composition $x \circ \gamma$ représente le temps recalé de la fonction originale x. La phase est plus qu'un concept relatif. Si une fonction de reparamétrisation γ est utilisée pour recaler la fonction x_2 à x_1, alors ce γ est nommé la *phase relative* de x_1 à x_2. Notons que l'inverse de ce γ est la phase relative de x_2 à x_1. En cas de plusieurs fonctions, comme dans le cas de notre

application, les composantes de phase sont définies en cherchant une moyenne de fonction et alors en évaluant la phase relative de chaque fonction donnée par rapport à la moyenne. Voir Algorithme 1 pour plus de détails.

Data : fonctions x_i.
Result : Moyenne de Fréchet μ_f, fonction de reparamétrisation γ_i^*, fonctions recalées x_i^*.

---

1. **Initialisation:** calculer les q_i correspondant à chaque $\{x_i\}$ et $\mu_q = \frac{1}{n}\sum_{i=1}^n q_i$.
2. **Recalage:** Pour $i = 1, 2, \ldots, n$ calculer $\gamma_i^* = \arg\inf_{\gamma\in\Gamma} \|\mu_q - q_i * \gamma\|^2$.
3. **Actualisation:** Actualiser μ_q en utilisant $\mu_q \leftarrow \frac{1}{n}\sum_{i=1}^n (q_i * \gamma_i^*)$. Tant qu'il n'y a pas de convergence, on retourne à l'étape 2.
4. **Centrer:** Calculer la moyenne de la fonction de recalage $\bar{\gamma}$ et actualiser μ_q en utilisant $\mu_q \leftarrow \mu_q * \bar{\gamma}^{-1}$.
5. **Recalage final**: Répéter l'étape 2. Calculer μ_x et $x_i^* = x_i \circ \gamma_i^*$.

Algorithme 1 : Algorithme de séparation Phase-Amplitude

Ainsi, nous pourrons attribuer une amplitude et une composante de phase à chaque fonction d'un ensemble donné, et utiliser ces composantes pour définir les caractéristiques de l'AR nécessaires à la classification des personnes.
Supposons que nos fonctions x_i sont de classe C^k, $k \in \{0, 1, 2\}$. Pour le reste du papier, au lieu de x_i, nous allons utiliser une variable globale notée z_i, globale dans la mesure où elle sera utilisée pour différentes représentations comme l'intensité de la force $z_i = x_i$, sa vitesse $z_i = \dot{x}_i$, son accélération $z_i = \ddot{x}_i$ et la fonction de courbure correspondante $z_i = c_i$, pour la régression. Nous montrons dans la Figure 2 l'allure de ces différentes représentations fonctionnelles z_i . Il est important de noter que pour un signal parfait, on s'attend à ce que l'intensité de la force soit nulle au début et à la fin de chaque test. Ainsi, nous comptons sur les dérivées et la courbure pour capturer la distance entre une observation donnée et une observation ayant un comportement normal. Etant donné que les observations réelles, même les groupes des patients, ne sont pas parfaits, nous ferons un test répétitif (presque périodique) pour améliorer cette partialité.
Nous rappelons que notre objectif est d'utiliser les variables fonctionnelles d'intensité, ou une des représentations, pour prédire l'état d'une personne. Pour obtenir ceci, la méthode d'estimation de la régression fonctionnelle à noyau est utilisée. Notre analyse se fera sur des données déjà recalées, avec toutes les représentations citées précédemment.

3 Régression fonctionnelle à noyau avec réponse binaire

Différents estimateurs non paramétriques de régression ont été proposés dans la littérature quand la variable aléatoire explicative z_i prend ces valeurs dans un espace de dimension finie. Il y a beaucoup de travaux dans la littérature qui traitent les limites de ces estimateurs et d'autres questions qui y sont liées, comme la sélection de la fenêtre optimale dans les cas dépendants et indépendants. Pour plus de détails, on peut se référer aux [15, 16] et aux références citées dedans.
Les résultats asymptotiques des données fonctionnelles ont récemment eu un intérêt croissant, on peut se référer aux [17, 18] et à la récente monographie faite par Ferraty et Vieu [19] et les références citées dedans.

Pour formuler le problème de l'estimateur de la régression fonctionnelle, supposons $(z_i, y_i)_{i\in\mathbb{N}}$ une séquence de couple de variables aléatoires (Z, Y) où z_i prend ces valeurs dans un espace métrique $(E, d(.,.))$ et y_i est binaire. Nous considérons le modèle

$$Y = r(Z) + \epsilon \tag{1}$$

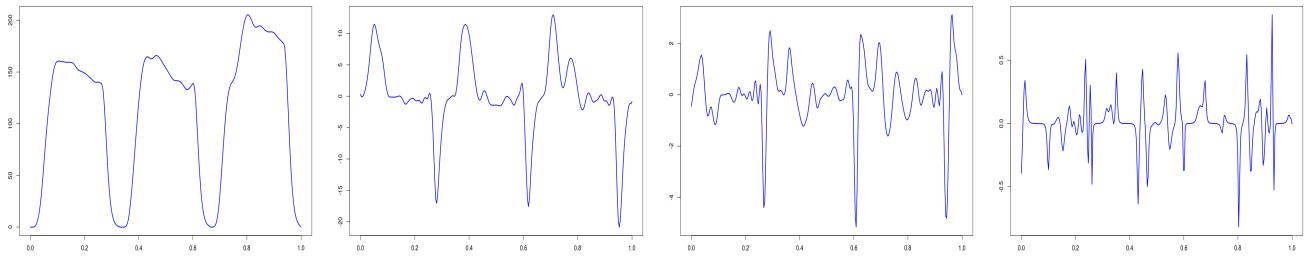
D'après (1), $r(z_i) = \mathbb{E}[Y|Z = z_i]$. Considérons d'abord E comme étant un espace d'Hilbert $\mathcal{H}$ muni de sa métrique associée d. z_i étant de dimension infinie, nous allons la décomposer dans la base de fonction $\phi = (\phi_1(t), ..., \phi_p(t))$: $z_i(t) = \sum_{j=1}^p \alpha_{ij}\phi_j(t) = \alpha_i^T\phi$ avec $\alpha_i = (\alpha_{i1}, ..., \alpha_{ip})$.
Pour des raisons pratiques, au lieu de travailler avec les z_i, nous allons travailler avec les coefficients α_i, qui sont de dimension finie, issus des décompositions des z_i dans la base de fonction ϕ.

L'estimateur de type Nadaraya-Watson a été introduit par Ferraty et Vieu [20]. Dans notre cas, il est défini par :

$$\hat{r}_n(z_i) = \frac{\sum_j y_j K_h(d(\alpha_i, \alpha_j))}{\sum_j K_h(d(\alpha_i, \alpha_j))}$$

où le dénominateur est différent de zéro et $K_h(d(\alpha_i, \alpha_j)) = K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right)$. Ici K est une fonction noyau à valeurs réelles, h est le paramètre de la fenêtre (qui tend vers zéro quand n tend vers l'infini) et d est la métrique associée à $\mathcal{H}$.
Puisque Y est binaire, on cherchera plutôt à modéliser

$$g(Y) = r(Z) + \epsilon \tag{2}$$



(a) (b) (c) (d)

Figure 2: Exemples de différentes fonctions représentant l'intensité de la force de la main: (a) une courbe originale, (b) la vitesse, (c) l'accélération, et (d) la courbure.

où g est la fonction logit. La fonction réciproque de cette logit, appliquée à $\hat{r}_n(z_i)$, renvoie des valeurs de probabilités auxquelles nous allons appliquer un seuil pour faire la classification.
Les taux de convergence presque sûre, sur un ensemble compact de l'estimateur $\hat{r}_n$, sont établis dans [21] pour les processus asymptotiquement indépendants, alors que Masry [22] obtient la convergence de la moyenne quadratique. De plus la normalité asymptotique a été obtenue par Ferraty et al. [23]
Toujours dans l'optique de trouver le meilleur modèle, nous changeons d'espace et on choisit E comme étant la sphère de Hilbert. $\alpha_i \in \mathbb{R}^p$, nous nous restreignons à la sphère $\mathbb{S}^{p-1}$. Ainsi nous avons utilisé la distance géodésique s définie sur cette sphère par :

$$s(\alpha_i, \alpha_j) = arcos(\frac{\alpha_i^T \alpha_j}{\|\alpha_i\| \|\alpha_j\|}).$$

La question qui se pose maintenant c'est quelles sont les valeurs optimales de h et de seuil qu'il faut prendre pour classer les malades et les personnes bien portantes. Dans la Figure 3, nous affichons des exemples de distribution des distances géodésiques sur la sphère et les h optimales retenues. Pour calibrer la performance de notre modèle d'estimation (régression fonctionnelle à noyau), nous considérons les critères : MSE (Mean Squared Error) et MCC (Matthews Coefficient Correlation).

- **MSE:** C'est l'erreur quadratique moyenne. Elle est définie par :

$$\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

avec n le nombre d'observation prédite et $\hat{y}_i$ la valeur prédite de la i-ème observation.

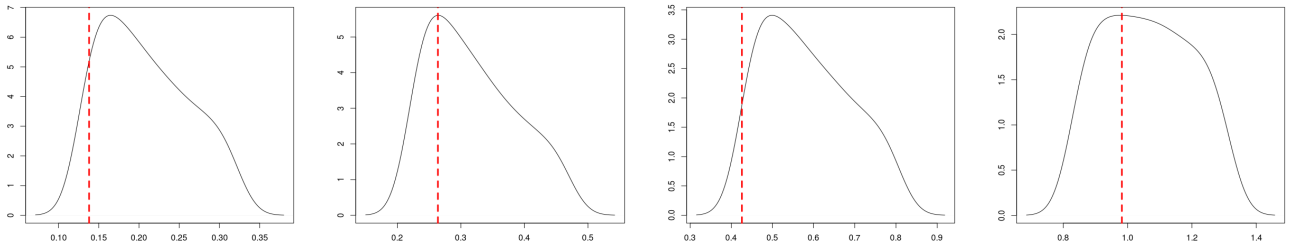
- **MCC:** Basé sur les Vrais et Faux Positifs (VP, FP), et sur les Vrais et Faux Négatifs (VN, FN), il est généralement considéré comme une mesure équilibrée [24]. MCC:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC ou coefficient de corrélation entre les valeurs observées et les valeurs prédites, renvoie des valeurs comprises entre -1 et 1. Plus la valeur est proche de +1, plus la classification est bonne. Plus elle est proche de -1, plus la classification est mauvaise.

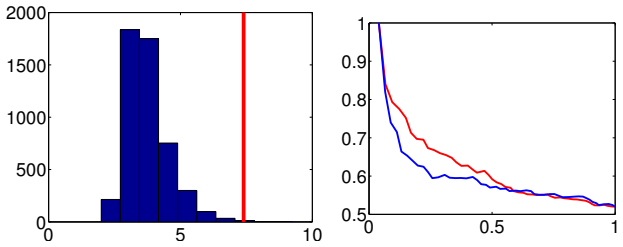
4 Résultats expérimentaux

Dans cette section, nous décrivons notre approche qui vise à classifier les observations dans le groupe des personnes malades ou dans celui des personnes bien portantes, en utilisant les signaux de force de la main. Les résultats présentés dans cette section sont les résultats moyens obtenus après 100 itérations. Chaque itération consiste à générer aléatoirement 1500 observations composées de personnes malades et de personnes bien portantes, dont 900 constituent la base d'apprentissage et de validation et les 600 restantes la base test. Nous utilisons un modèle de régression fonctionnelle avec différents critères et différentes représentations. Chaque observation est représentée par une seule fonction de force de la main, combinant les 3 tests consécutifs. De ces fonctions, dérivent différentes représentations utiles pour la classification. Notre modèle de régression fonctionnelle à noyau utilise le noyau gaussien. Ces paramètres sont choisis grâce à la base d'apprentissage et les optimaux sont retenus grâce à la base de validation, avec le critère MCC. Cela assure et améliore la précision de la classification. Avant de présenter les principaux résultats de ce travaux, nous montrons une comparaison d'une méthode proposée et une simple approche d'analyse de données fonctionnelles, qui utilise la métrique $\mathbb{L}^2$ entre les fonctions et qui ne tient pas en compte des variabilités de phase. Nous calculons la matrice de distance



(a) (b) (c) (d)

Figure 3: Exemples de densités des distances géodésiques sur la sphère pour chaque représentation, en rouge la valeur de la h optimale retenue pour le modèle : (a) Initial, (b) Vitesse, (c) Accélération, et (d) Courbure.



(a) (b)

Figure 4: (a) Distribution empirique du test statistique de distance d'amplitude, avec la distance entre les 2 groupes marquée en rouge. (b) Précision (ordonnée) vs. Rappel (abscisse) la courbe de la méthode proposée (rouge) et la courbe obtenue avec la méthode d'analyse de données fonctionnelles non élastique basée sur la métrique $\mathbb{L}^2$ (blue).

pour chaque méthode et nous affichons la courbe Rappel/Précision dans la Figure 4(b). De cette figure, on peut dire que le fait de prendre en compte la variabilité de phase de signaux de force de la main est important et a le potentiel d'améliorer considérablement la performance de classification.

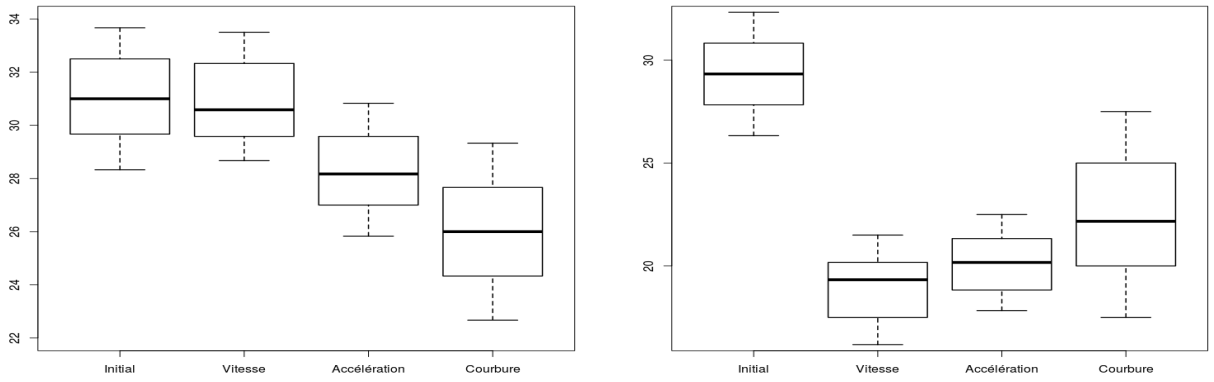
Nous évaluons maintenant la performance de notre modèle en calculant, après avoir prédit les variables réponses de la base test, les valeurs du critère utilisé (MCC).

Si on utilise la métrique L2 dans notre modèle, c'est avec la représentation courbure qu'on obtient un plus petit taux d'erreur, comme on peut le visualiser au niveau de la Figure 5(a). Par ailleurs, si on utilise la distance géodésique sur la sphère, notée ici par s, c'est la vitesse qui nous donne une meilleure classification des deux groupes, voir Figure 5(b). Nous pouvons aussi remarquer qu'avec la métrique s, c'est la représentation vitesse qui nous donne la meilleure valeur de spécificité (plus petite erreur de première espèce) et la courbure nous donne une meilleure valeur de sensibilité (plus grande valeur de la puissance du test), voir Figure 6.

Comme nous l'avons énoncé précédemment, les personnes ayant un AR avancé montrent une décroissance significative de leurs forces de main durant les tests, comparés aux personnes bien portantes. Et cet aspect était le plus utilisé par les rhumatologues dans leurs diagnostics. Cependant, une telle procédure n'est pas applicable pour toutes les personnes malades à cause de différents facteurs comme l'âge, le genre, et plus important encore, le niveau de sévérité de la maladie. Les patients ayant un niveau d'AR moyen étaient difficiles à détecter avec le diagnostic classique. Ainsi il est important de rappeler que le fait d'utiliser les mesures continues de force de la main est une méthode bénéfique, rapide et facile, et plus encore, il est très efficace pour diagnostiquer le degré de la maladie. De plus, les informations extraites de la force de la main ont une interprétation clinique naturelle et donc plus intéressantes pour les médecins.

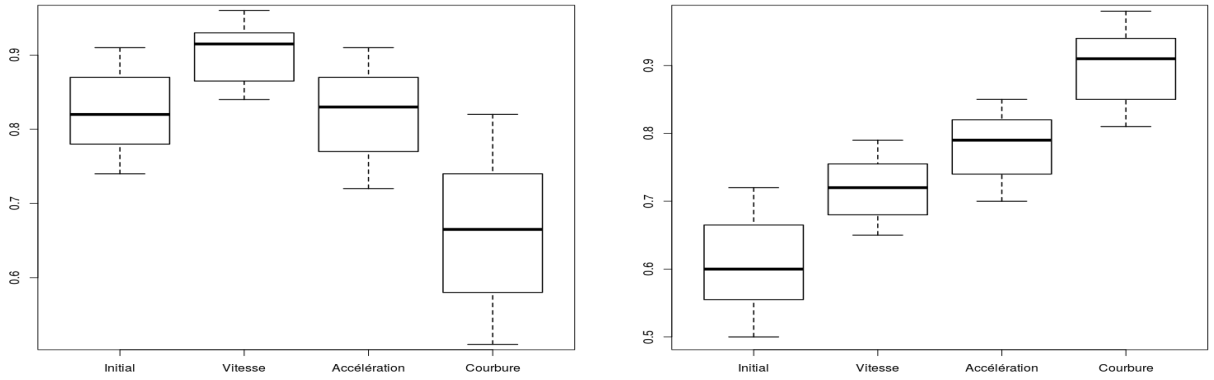
5 Conclusion

Ce travail présente une nouvelle approche permettant de caractériser les données fonctionnelles pour la classification de l'Arthrite Rhumatoïde (AR). Cette méthode a l'avantage d'utiliser les courbes recalées et de capturer ainsi plus d'informations des signaux, contrairement aux diagnostics classiques utilisés précédemment. Une fois que les courbes sont recalées, différentes représentations fonctionnelles ont été utilisées et la fonction de densité conditionnelle a été utilisée pour estimer la régression. Que ça soit la métrique d ou s, le fait d'utiliser la représentation standard (courbes initiales) ne nous permet pas d'avoir une meilleure classification. Ceci est dû au fait que la représentation standard ne capte pas bien la variabilité des différences de forces émises par les personnes. D'où l'importance d'utiliser d'autres représentations fonctionnelles, comme la vitesse, l'accélération ou la cour-



(a) (b)

Figure 5: Dans chaque figure, on a représenté les taux d'erreurs obtenus en fonction des représentations fonctionnelles utilisées : (a) en utilisant la métrique d, i.e. la métrique $\mathbb{L}^2$ et (b) en utilisant la métrique s.



(a) (b)

Figure 6: Ces résultats sont obtenus avec la métrique s (distance géodésique sur la sphère) : (a) Spécificité en fonction des représentations et (b) Sensibilité en fonction des représentations.

bure. On voit par exemple qu'en utilisant la métrique d ($\mathbb{L}^2$), c'est la courbure qui nous donne la meilleure classification. Et si on utilise la distance géodésique sur la sphère (s), c'est la vitesse qui nous donne les meilleurs résultats de classification. Ces résultats expérimentaux nous montrent que les diagnostics utilisés précédemment sont insuffisants et que notre modèle est très prometteur pour ce sujet.

References

[1] G. James, "Curve alignment by moments," *Annals of Applied Statistics*, pp. 480–501, 2007.

[2] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis, Second Edition.* Springer Series in Statistics, 2005.

[3] J. D. Tucker, "Functional component analysis and regression using elastic methods," *Electronic Theses, Treatises and Dissertations, Florida State University*, 2014.

[4] R. Tang and H. G. Müller, "Pairwise curve synchronization for functional data," *Biometrika*, vol. 95, no. 4, pp. 875–889, 2008.

[5] D. L. Scott, F. Wolfe, and T. W. Huizinga, "Rheumatoid arthritis," *The Lancet*, vol. 376, no. 9746, pp. 1094–1108, 2010.

[6] S. J. Bigos, J. Holland, C. H. ands J. S. Webster, M. Battie, and J. A. Malmgren, "High-quality controlled trials on preventing episodes of back problems: systematic literature review in working-age adults," *Spine Journal*, vol. 9, no. 2, pp. 147–68, 2009.

[7] G. Michael and W. Richard, "A systematic exploration of distal arm muscle activity and perceived exertion while applying external forces and moments," *Ergonomics*, vol. 51, no. 8, pp. 1238–1257, 2008.

[8] A. Kneip and T. Gasser, "Statistical tools to analyze data representing a sample of curves," *The Annals of Statistics*, vol. 20, pp. 1266–1305, 1992.

[9] C. Samir, S. Kurtek, A. Srivastava, and N. Borges, "An elastic functional data analysis framework for preoperative evaluation of patients with rheumatoid arthritis," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[10] J. O. Ramsay and X. Li, "Curve registration," *Journal of the Royal Statistical Society, Series B*, vol. 60, p. 351–363, 1998.

[11] D. Gervini and T. Gasser, "Self-modeling warping functions," *Journal of the Royal Statistical Society, Series B*, vol. 66, pp. 959–971, 2004.

[12] X. Liu and H. G. Müller, "Functional convex averaging and synchronization for time-warped random curves," *Journal of the American Statistical Association*, vol. 99, pp. 687–699, 2004.

[13] A. Kneip and J. O. Ramsay, "Combining registration and fitting for functional models," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1155–1165.

[14] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron, "Registration of functional data using fisher-rao metric," *arXiv: 1103.3817v2*, 2011.

[15] L. T. Tran, "Density estimation for time series by histograms," *Journal of statistical planning and inference*, vol. 40, no. 1, pp. 61–79, 1994.

[16] N. Laı *et al.*, "Kernel estimates of the mean and the volatility functions in a nonlinear autoregressive model with arch errors," *Journal of statistical planning and inference*, vol. 134, no. 1, pp. 116–139, 2005.

[17] D. N. Politis and J. P. Romano, "Limit theorems for weakly dependent hilbert space valued random variables with application to the stationary bootstrap," *Statistica Sinica*, pp. 461–476, 1994.

[18] J. O. Ramsay and B. W. Silverman, *Applied functional data analysis: methods and case studies*. Springer New York, 2002, vol. 77.

[19] F. Ferraty and P. Vieu, *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[20] ——, "Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés," *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, vol. 330, no. 2, pp. 139–142, 2000.

[21] ——, "Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination," *Nonparametric Statistics*, vol. 16, no. 1-2, pp. 111–125, 2004.

[22] E. Masry, "Nonparametric regression estimation for dependent functional data: asymptotic normality," *Stochastic Processes and their Applications*, vol. 115, no. 1, pp. 155–177, 2005.

[23] F. Ferraty, A. Mas, and P. Vieu, "Nonparametric regression on functional data: inference and practical aspects," *Australian & New Zealand Journal of Statistics*, vol. 49, no. 3, pp. 267–286, 2007.

[24] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.