Comparaison statistique de courbes - Statistical Parametric Mapping

Gael Druez

Mai 2022

gael.druez@ens-rennes.fr

Stage de fin de L3 "Mathématiques pour la recherche" Université Rennes 1

Sous la direction de Floren Colloud, professeur aux Arts et Métiers ParisTech

ENS Rennes - Institut de Biomécanique Humaine Georges Charpak

1 Objectif du stage

En sciences, et en particulier en biomécanique, l'expérience est une partie majeure du travail des chercheurs. Ces expériences produisent des données, mais une fois le protocole déterminé et les données collectées, il faut encore les analyser pour pouvoir conclure.

Or ces données sont souvent en grande quantité, et essayer de remarquer une tendance ou un modèle dans des milliers et milliers de valeurs peut très vite se révéler impossible.

Pour pallier cela, le domaine des statistiques fournit des outils puissants qui permettent de comparer de gros groupes de données, soit à des modèles, soit entre eux. On parle pour cela de *test statistique*.

Seulement, les tests statistiques usuels possèdent certaines limites, et ne vont pas être adaptés pour analyser des courbes par exemple.

Ainsi, l'objectif de ce stage est d'étudier une méthode de comparaison statistique de courbes, de comprendre pourquoi un tel outil est nécéssaire et comment une telle comparaison est possible.

L'article [Col06] servira de fil rouge à ce stage. En particulier, l'auteur compare des courbes issus d'une expérience d'aviron, et on voudrait valider ses conclusions à la lumière d'une comparaison statistique de courbes.

Table des matières

2 Article étudié

2.1 Objectif de l'article

L'objectif de l'article [Col06] de Floren COLLOUD était de comparer la rame sur ergomètre fixe d'une part, c'est-à-dire avec les pieds fixes par rapport au sol et le siège mobile, et sur ergomètre libre d'autre part, c'est-à-dire où les pieds ne sont plus fixes et donc le siège bouge beaucoup moins. L'avantage de ce second ergomètre est qu'il se rapproche plus de la rame sur l'eau, le mouvement du rameur étant un facteur d'amélioration de la performance sur l'eau mais non pris en compte sur ergomètre fixe. L'objectif était de déterminer s'il y avait une différence de technique et d'efficacité de rame entre les deux ergomètres, et si oui, comprendre où avait lieu cette différence.

2.2 Méthode

L'expérience consiste en une mesure de plusieurs cycles de rames. Au total, 25 rameurs élite ont participé, chacun sur l'ergomètre fixe et sur l'ergomètre libre. Les données ont ensuite été normalisées temporellement pour avoir en abscisse le pourcentage de cycle (permet d'effacer les légères différences de cadence entre les rameurs).

La figure 1 récapitule les forces mesurées par l'expérience lors des cycles de rame sur l'ergomètre fixe et l'ergomètre libre.

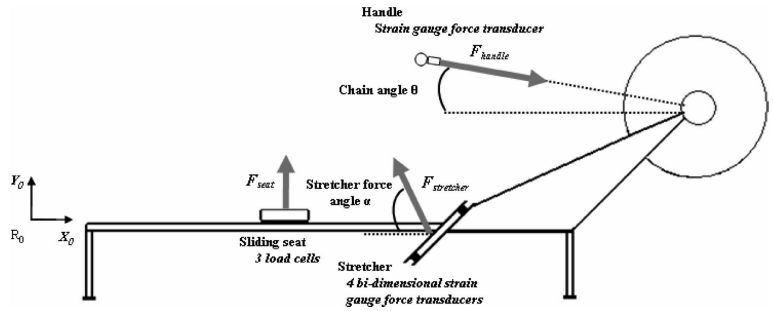


FIGURE 1 – Schéma expérience article F. COLLOUD

L'analyse est réduite au plan sagittal (X_0, Y_0) : les variations latérales sont négligées.

Différentes forces appliquées par le sujet ont été enregistrées : sur la poignée, sur le siège, sur le repose-pieds. De ces forces, ont été déterminées les puissances correspondantes fournies par les sujets.

Chacune de ces mesures fournit alors une courbe, qui donne soit la force soit la puissance en fonction du pourcentage de cycle. Ce sont ces courbes que l'on veut

comparer, et plus précisément comparer chaque courbe sur ergomètre fixe avec la courbe correspondante sur ergomètre libre.

La figure 2 montre une telle courbe issue de l'article. Celle-ci représente la puissance extérieure, c'est-à-dire la puissance totale fournie par les sujets sur un cycle. On voit en traits épais la moyenne sur les 25 sujets, en traits fins les intervalles de confiance à 95%. La courbe en traits pleins correspond à l'eromètre fixe, celle en pointillés à l'ergomètre libre.

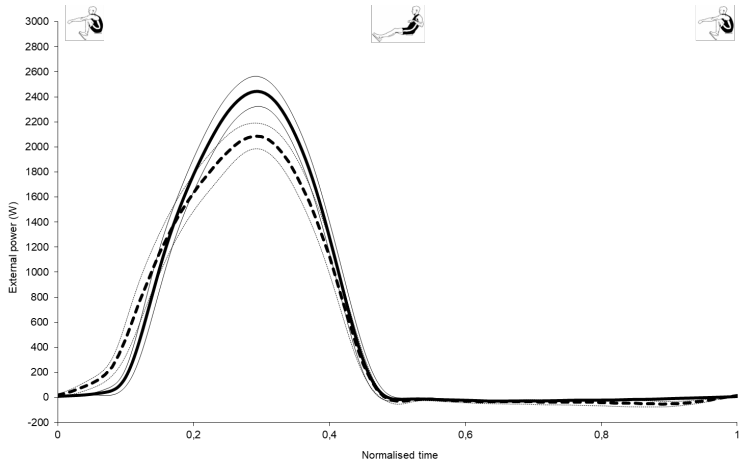


FIGURE 2 – Courbes moyenne de puissance extérieure et leur intervalle de confiance à 95%

Pour comparer les courbes entre elles, deux méthodes ont été utilisées par l'auteur :

— les tracés des courbes obtenues, avec leur intervalle de confiance à 95% : une différence significative était annoncée quand les deux intervalles de confiance étaient disjoints

— un test statistique usuel sur certains points particuliers de chaque courbe (extrema, valeurs aux moments de changement de phase, ...)

Ce sont ces courbes que l'ont va vouloir analyser d'une autre manière, avec un nouvel outil.

3 Point statistiques

Avant de parler concrètement de comparaison de courbes, revenons sur les tests statistiques usuels utilisés très fréquemment en statistique. Comme évoqué en partie 1, ils permettent de comparer des ensembles de données numériques : on se contente donc dans cette partie d'étudier des tests statistiques qui portent sur des données numériques et non des courbes.

Ces tests usuels sont souvent séparés en deux catégories :

3.1 Tests paramétriques

La première catégorie est constituée des tests dits *paramétriques*. Ces test comparent les données de l'utilisateur à un modèle choisi. Ainsi, ils nécessitent de savoir à quel modèle comparer les données.

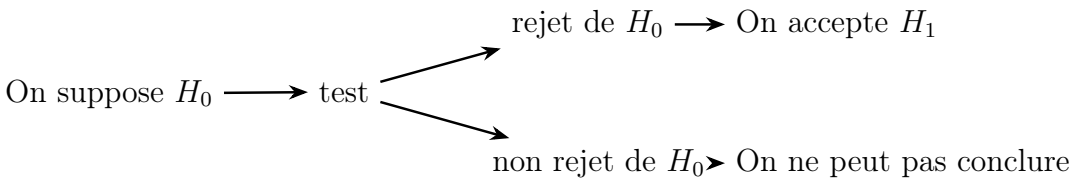
La méthode de test dépend du test effectué, mais le principe est souvent le même : on choisit un indicateur à comparer (moyenne, variance, ...), et on pose deux hypothèses : H_0 est l'*hypothèse nulle*, qui suppose que l'échantillon étudiée suit bien la loi de notre test ; H_1 est l'*hypothèse alternative*, qui suppose une différence entre la loi de l'échantillon et celle du test. Le but du test va être de savoir si on peut rejeter cette hypothèse nulle et conclure que notre échantillon est significativement différent du modèle.

On choisit également un facteur de risque α, c'est-à-dire le risque qu'on accepte de rejeter l'hypothèse nulle alors qu'elle est vraie. On va ensuite calculer la probabilité, sous l'hypothèse nulle, que l'indicateur choisie prenne une valeur au moins aussi extrême que la valeur effectivement prise par l'indicateur de notre échantillon. Cette probabilité, appelée *valeur p* (ou *p-value* en anglais), va être le résultat de notre test. Ainsi, si $p < \alpha$, on considère que l'échantillon s'écarte significativement du modèle, et on conclut en rejetant H_0. Sinon, on ne peut pas conclure.

> *Le fait qu'on ne puisse pas conclure, et en particulier qu'on ne puisse pas valider H_0 fait qu'on ne peut pas a rmer qu'un échantillon suit un modèle donné. Cela empêche de montrer, statistiquement parlant, que quelque chose* n'a pas d'impact *sur le groupe expérimental, puisqu'on ne peut pas a rmer que ce groupe suit la même distribution que la population globale ou le groupe témoin.*

On a alors le tableau suivant :

$\alpha > p$	La probabilité que l'indicateur soit si écarté du modèle est inférieure au risque accepté : on rejette l'hypothèse H_0 et on accepte H_1.
$\alpha < p$	L'échantillon n'est pas significativement écarté du modèle : on ne peut pas conclure.



Maintenant, il existe différents test paramétriques, qui correspondent à différentes statistiques de test :

— **test Z** : on teste les données contre une loi normale

— **test T de Student** : loi de Student à $n-1$ degrés de liberté, permet de "corriger" la loi normale pour de petits échantillons (la loi de Student converge en loi vers la loi normale quand $n \to \infty$)

— **test du** χ^2 : loi du χ^2

— **test F de Fisher-Snedecor** ou **ANOVA** : pour comparer les distributions de plusieurs groupes, en analysant la variance de l'ensemble des données

3.2 Tests non paramétriques

On va parfois avoir besoin de tester si un échantillon d'une population suit la même distribution qu'un autre échantillon, mais sans connaître la distribution d'aucune des deux populations dont sont extraits les échantillons en question.

On peut par exemple mentionner le **test de Wilcoxon-Mann-Whitney (WMW)**, qui ne se base pas sur les valeurs prises par les deux groupes, mais plutôt sur leur rang. En effet, sous l'hypothèse nulle, les valeurs du groupe 1 devraient être équitablement réparties, c'est-à-dire la somme des rangs des valeurs de la série 1 doit être proche de la somme attendue en cas de pafaite répartition, et donc leur différence Δ doit être petite. Il a été montré que, sous l'hypothèse nulle, Δ suit une loi normale centrée.

Le test WMW revient donc à tester $\frac{\Delta}{\text{écart-type de } \Delta}$ contre une loi normale centrée réduite, et cela permet de ne pas préjuger de l'éventuelle distribution suivie par le groupe.

3.3 Quel test utiliser ?

On se demande maintenant quel test on veut utiliser en fonction des données qu'on a. Pour cela, plusieurs points sont à prendre en considération :

- Le type des variables : on distingue les variables *quantitatives* (les "nombres", qui mesurent une donnée sur chaque élément du groupe) des variables *qualitatives* (qui décrivent une caractéristique de l'élément : les questions "Quelle est la couleur de vos yeux ?", "Aimez-vous le foot ?" ou "Cette mesure a-t-elle été enregistrée avec le traitement ou le placebo ?" fournissent une variable qualitative)
- Le nombre de groupes
- L'indicateur à comparer (moyenne, variance, valeurs extrêmes, ...)
- L'indépendance des variables entre elles
- L'indépendance des groupes entre eux

On a alors, pour la comparaison de deux séries de données, en notant :

- m_1 et m_2 les moyennes empiriques du 1er groupe et du 2nd groupe respectivement
- m_d la moyenne empirique de la différence des deux séries (pour deux séries appariées)
- μ la moyenne attendue

Type de série	Différence étudiée	Test	Condition
2 séries indépendantes	$m_1 - m_2$	Z	n_1 et $n_2 > 30$
		T de Student	normalité des distributions $n_1 \approx n_2$ ou $s_1^2 \approx s_2^2$
		WMW	n_1 et $n_2 > 10$ ex-aequo peu nombreux
2 séries appariées	$m_d - 0$	Z	paires > 30
		T de Student	normalité des différences
		WMW	paires non nulles > 20 ex-aequo peu nombreux
Comparaison entre moyenne observée et moyenne attendue	$m - \mu$	Z	$n > 30$
		T de Student	normalité distribution population

Faisons une pause dans les valeurs numériques et revenons à l'article qui nous intéresse et ses séries de courbes.

On a bien à comparer 2 séries (pour chaque courbe, on compare la série sur ergomètre fixe et la série sur ergomètre libre).

Les 2 séries sont appariées car chaque sujet est passé 2 fois, 1 fois sur chaque ergomètre. Ainsi, chaque acquisition sur ergomètre fixe possède sa "paire" sur ergomètre mobile. On peut donc considérer la série "différence" qui contient les courbes correspondant aux différences entre une courbe et sa paire.

Comme on n'a que 25 paires, on préfère utilisé un *test T apparié.*

3.4 La conclusion des tests statistiques

Après avoir mené un test statistique, on aboutit à une conclusion : soit on rejette H_0, soit on ne la rejette pas.

Quoi qu'on fasse, il y a un risque de se tromper et de faire le mauvais choix. Le risque de rejeter H_0 alors qu'elle est vraie est α : on l'a vu, ce risque est choisi par la personne qui fait le test, on peut donc fixer sa valeur en fonction du risque qu'on souhaite prendre. C'est le risque de réaliser une *erreur de type I* ou un *faux positif* : on croit voir une différence significative alors qu'il n'y a rien du tout.

Le risque de ne pas rejeter H_0 alors que c'est H_1 qui est vraie est noté β. Ce risque là dépend de la taille de l'échantillon et de la variance des données, ainsi pour réduire ce risque il faut augmenter la taille des données étudiées. C'est le risque de commettre une *erreur de type II* ou un *faux négatif* : on ne voit pas une différence qui existe bel et bien, car on n'a pas assez de données.

On peut résumer cela dans un tableau :

		La vérité	
		H_0 vraie	H_0 fausse
Conclusion du test	H_0 acceptée	✓	Erreur de type II (risque β)
	H_0 rejetée	Erreur de type I (risque α)	✓

On peut se demander quel risque est le plus important à éviter : est-ce qu'il vaut mieux prendre un risque α très faible, pour éviter au maximum de rejeter H_0 si elle est vraie et donc de conclure de manière érronée, ou est-il préférable de se concentrer sur β et maximiser les chances de voir une différence, quand bien même elle n'existerait pas ?

Dans la majorité des cas, c'est le risque α qu'on veut minimiser, car on préfère ne pas savoir qu'avoir tort. C'est d'ailleurs le risque choisi par la personne qui fait le test, contrairement au risque β.

Cependant, il arrive que l'erreur de type I soit moins grave que l'erreur de type II. Par exemple, si un médecin veut comparer l'efficacité de 2 traitements, il pose H_0 = "Les deux traitements sont aussi efficaces". Alors, si H_0 est vraie mais que le médecin la rejette (type I), il administrera à ses patients le traitement qu'il a trouvé plus efficace, ce que ne changera rien puisqu'H_0 est vraie donc les deux sont équivalents. En revanche, si H_0 est fausse mais qu'il ne la rejette pas (type II), il administrera potentiellement un traitement moins efficace alors qu'il disposait d'une meilleure option : il préfère donc éviter le faux négatif.

> *Notons que le but de ce genre de tests est de rejeter une hypothèse nulle qui soit fausse : si H_0 est e ectivement fausse, ce rejet arrive avec une probabilité $1-\beta$. On appelle cette probabilité la* puissance *du test (que l'on maximise de la même manière qu'on minimise β, à savoir en augmentant la taille de l'échantillon).*

4 SPM

Les tests statistiques présentés précédemment présentent plusieurs limites, dont certaines dont on a déjà parlé comme l'impossibilité de valider H_0, le risque de rejeter H_0 à tort (erreur de type I) et le risque de ne pas rejeter H_0 à tort (erreur de type II).

Une autre limite, sur laquelle on ne s'est pas arrêté, est le fait de se contenter d'un unique indicateur. Cette partie est dédiée à une méthode qui vise à s'affranchir de cette limitation à un unique indicateur dans le cas de courbes. En effet, la manière usuelle de comparer des courbes est de comparer le maximum de la différence, ou certains points particuliers comme le début ou la fin d'un cycle, mais on perd beaucoup d'informations car cela réduit une courbe à un seul point (particulier certes, mais qui ne capture pas toutes les subtilités). La méthode dite **Statistical Parametric Mapping (SPM)** pemet de faire un traitement statistique de courbes de données pour pallier ce problème, et ainsi tirer toute l'information des courbes 1D tout en réduisant le risque de commettre une erreur de première espèce.

4.1 Précisions sur les dimensions des données

Détaillons d'abord un point de terminologie relatif aux analyses statistiques que nous allons mener.

Les tests statistiques dont nous avons parlé précédemment étaient des tests dits *0D*, c'est-à-dire qu'ils analyse des valeurs ponctuelles. Par opposition, on va étudier les tests *1D*, qui analysent des données sous forme de courbe.

Pour être plus précis, il faudrait parler du format nDmD des données, où n est la dimension du domaine de mesure et m celle des variables indépendantes.

Ainsi, l'analyse du maximum local (0D) d'une courbe (1D) est une analyse 0D1D : c'est le format le plus courant quant il s'agit d'effectuer des tests statistiques.

L'analyse d'une courbe (1D) issue d'un mouvement dans l'espace (3D) (par exemple, la force antéro-postérieure exercée sur la poignée d'un ergomètre pendant un cycle d'aviron) est donc un format 1D3D. Le format 1DmD est celui qui est souvent obtenu lors d'expériences de biomécanique, où un (ou plusieurs) mouvement(s) dans l'espace fourni(ssen)t une courbe à analyser.

[Pat16] fournit des exemples, dont quelques-uns sont retranscris dans le tableau suivant :

		Exemples	
n	m	Données analysées (de dimension n)	Données mesurées (de dimension m)
0	1	instant de l'impact	angle de flexion du genou
0	3	instant de propulsion maximale	force appliquée sur la plateforme
1	1	force verticale appliquée sur un siège	longueur de poignée tirée
1	3	vitesse horizontale d'un objet	mouvement de l'objet dans l'espace
2	1	surface de contact du pied	pression
3	1	fémur	force verticale appliquée au point A

Dans la suite, on utilisera 0D pour 0D1D et 1D pour 1DmD.

4.2 Le problème de l'augmentation du risque de faux positif

Cette distinction entre les dimensions d'étude est importante, car elle peut amener à des écarts (potentiellement importants) dans l'estimation du risque de première espèce α.

En effet, comme expliqué dans [Pat16], analyser des données 1D avec les tests usuels 0D augmente le risque d'obtenir des faux positifs, c'est-à-dire de rejeter H_0 alors qu'elle est vraie.

Par exemple, on considère une série de courbes dont on veut vérifier si elles suivent un certain modèle. On peut analyser leur maximum, et comparer cette série de points à un maximum théorique prévu par H_0 : on réduit notre courbe 1D à une série de données 0D. Alors, on applique le test avec un risque α fixé, par exemple $\alpha = 5\%$.

Ensuite, on peut appliquer un nouveau test sur nos courbes initiales, afin de déterminer le risque que des courbes suivant l'hypothèse nulle soient au moins aussi extrêmes que les notres. Cette valeur est fournie par la théorie des champs aléatoires (nous reviendrons rapidement là-dessus dans la section suivante) et dépend de plusieurs paramètres tels que la régularité de nos courbes et de la dimension de notre étude (le m de 1DmD). Mais on peut noter que dans un cas favorable (mesure 1D et plutôt régulière), on a quand même une valeur p proche de 0.15, soit déjà trois fois plus que ce qu'on voulait. Dans les cas moyens (mesure 3D et régularité moyenne), on monte à p=0.75. Dans les pires cas (k vecteurs soit $3k$D et faible régularité), p

excède 0.999. Toutes ces valeurs, fournies par la théorie des champs aléatoires, ont été confirmées par [Pat16] en prduisant un grand nombre de champs gaussien et en regardant la proportion de ceux qui franchissaient un certain seuil.

Une solution envisageable serait de faire une analyse 0D sur un grand nombre de points, par exemple subdiviser l'intervalle d'étude en N segments et analyser en 0D les valeurs prises par les courbes aux extrémités de ces segments.

Cela présente le double inconvénient de devoir faire tendre N vers $+\infty$, ce qui augmente à la fois le coût de calcul et le risque de première espèce (en effet, plus on fait de tests sur une même donnée, plus il y a de chance de tomber sur un cas particulier où la valeur est extrême malgré la validité de H_0, ce qui amène à rejeter H_0 à tort).

Tout cela montre que la réduction de données 1D à une analyse 0D est à manier avec précaution, et qu'à l'exception de cas où les données sont fondamentalement 0D, il est préféable de les analyser sous la forme la plus proche de celle qu'elles avaient quand elles ont été collectées.

4.3 La théorie derrière la méthode SPM

La méthode SPM se base sur un modèle général linéaire, c'est-à-dire qu'elle représente les données et les paramètres sous forme matricielle.

Le cas particulier sur lequel on se concentre ici (et qui est le plus utile en pratique) est le cas de données 1D, c'est-à-dire des courbes dans le plan. [Pat11] décrit la théorie derrière ce cas particulier.

Le principe est donc d'écrire nos données comme des perturbations d'une système linéaire, système qui est décrit par les paramètres et la manière dont ils influencent les courbes. On pose donc :

$$Y = X\beta + \epsilon$$

avec :

— Y une matrice de taille $I \times K$ qui contient les I enregistrements, chaque enregistrement contenant K points

— X une matrice de taille $I \times J$ qui contient les J paramètres étudiés et comment ils influencent les I mesures

— β une matrice $J \times K$ de paramètres de régression inconnus

— ϵ une matrice résiduelle $I \times K$, qui contient l'écart entre les données et la courbe théorique fournie par les paramètres dans X et β

Le but du test va être de déterminer à quel point les données enregistrées Y s'écartent du modèle sous H_0. Cet écart est caractérisé par ϵ. On va ensuite appliquer un test à cet écart pour déterminer notre valeur p, à savoir la probabilité que, sous H_0, des mesures présentent un écart au moins aussi important que nos données mesurées.

Les éléments connus sont donc Y (les données enregistrés) et X (les paramètres de l'expérience), et on cherche (β afin de déterminer) ϵ.

On obtient $\hat{\beta}$, solution au sens des moindres carrées, grâce à $X^\dagger$ pseudo-inverse de Moore-Penrose de X :

$$\hat{\beta} = X^\dagger Y = (X^T X)^{-1} X^T Y$$

La matrice $\hat{\beta}$ contient alors les courbes moyennes pour chacun des J paramètres dans X, ou encore la courbe suivie par les données si celles-ci suivent H_0.

On détermine ensuite ϵ et on obtient les I courbes résiduelles, c'est-à-dire les courbes réelles auxquelles on a soustrait les courbes théoriques fournies par le modèle paramétrique (sous H_0) :

$$\epsilon = Y - X\hat{\beta}$$

Et on peut donc obtenir la courbe de variance σ^2, un vecteur à K composantes qui correspond à une courbe repésentant la variance des courbes au point considéré :

$$\hat{\sigma}^2 = \frac{\mathrm{diag}(\epsilon^T \epsilon)}{I - \mathrm{rg}(X)}$$

Cette courbe de variance donne une variance à chaque point, variance entre les conditions et les sujets, mais **pas** entre les points d'une même courbe ([Fri95]).

Enfin, on utilise un vecteur c (qui dépend du test utilisé) qui va pondérer chaque paramètre en fonction de notre hypothèse nulle.

Ce vecteur c va contenir la courbe finale du test, notée SPM{t} pour un test T, et appelée statistique de test.

Le résultat de ces manipulation matricielles nous fournit ainsi une courbe, tracée dans le même domaine que les données, qui caractérise l'écart entre les deux séries de courbes.

On utilise ensuite la théorie des champs aléatoires (Random Field Theory, RFT) pour estimer la probabilité qu'une courbe semblable à la statistique de test soit aussi extrême.

Pour cela, on doit d'abord déterminer la largeur à mi-hauteur (Full Width at Half Maximum, FWHM) de notre courbe SPM{t}, qui est un indicateur en lien avec

sa régularité.

Ensuite, la RFT nous founit une estimation d'avoir un champ gaussien de même FWHM qui dépasse un certain seuil u. La RFT est même plus précise car elle donne la probabilité d'un dépassement en fonction de la taille de celui-ci, c'est-à-dire la probabilité d'avoir un dépassement d'une largeur donnée. Ainsi, la théorie nous donne

$P(u, k, c) = P(\text{"Avoir au moins } c \text{ clusters de volume au moins } k \text{ supérieurs au seuil u"})$

Ce qui nous intéresse principalement est $P(u, 1, 0)$ (probabilité d'avoir un dépassement) mais la formule générale sera utile dans l'implémentation pour calculer des valeurs p pour chaque cluster.

> *Ainsi, on réalise le test statistique contre un champ gaussien de même FWHM que notre statistique de test : ce champ gaussien général représente l'erreur d'une courbe aléatoire qui suit H_0, et on mesure l'éloignement de notre courbe par rapport à un comportement gaussien. On est donc bien en train de tester nos données contre un* champ *aléatoire, et non plus contre une* variable *aléatoire, et ce malgré le fait qu'on ne dispose que d'un nombre fini de points sur nos "courbes"*

On reviendra plus en détail sur la RFT en partie 5, notamment sur comment se faire une intuition des outils manipulés.

4.4 Implémentation de SPM1D en Python

Durant ce stage, l'algorithme SPM a été complètement recodé en Python, dans le cas particulier du test T apparié. L'objectif était de comprendre précisément comment fonctionne l'alogrithme, avec la procédure décrite dans la sous-partie précédente. Le recours à un module de RFT a été nécessaire pour quelques options (calcul des p-values associées à chaque cluster, calcul de FWHM, calcul de seuil à 5%). Il aurait pu en partie être évité avec de la simulation d'un grand nombre de champs gaussiens (*cf.* partie 5), mais cela aurait considérablement alourdi le programme.

La figure 3 montre que ce code Python fournit les mêmes résultats que le module Matlab utilisé lors du stage pour l'analyse SPM des courbes.

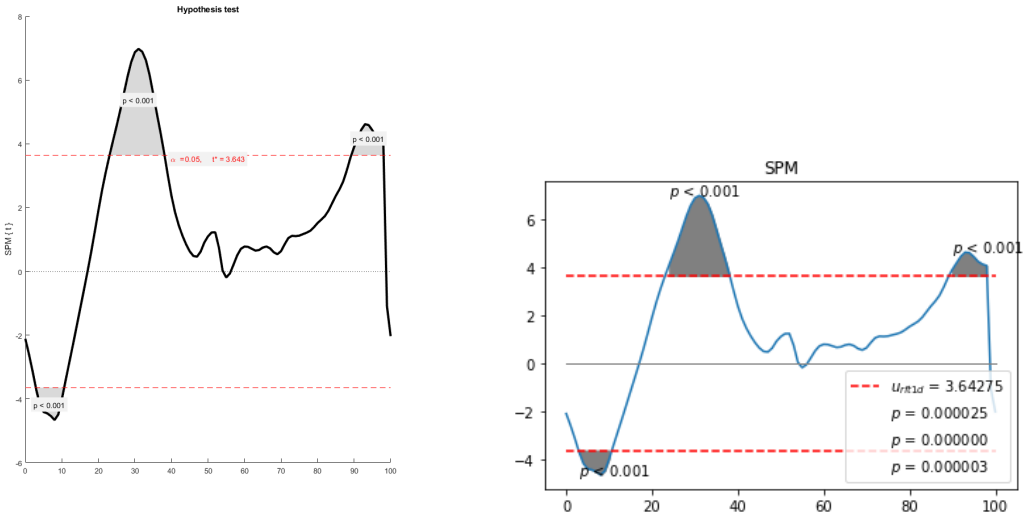


FIGURE 3 – Exemple d'analyse SPM avec le module usuel (à gauche) et le code Python (à droite)

5 Théorie des champs aléatoires

Revenons sur les outils de théorie des champs aléatoires qui sont utiles lors d'une analyse SPM.

Le principal objet est le champ gaussien aléatoire. En effet, l'analyse SPM consiste à comparer nos données à un champ gaussien, car on manipule plus facilement de genre d'outil. Cette partie sera l'occasion d'étudier plus en détail la manière dont on peut, de manière appliquée, s'en servir dans une analyse, avec une définition constructive des champs gaussiens.

5.1 Produit de convolution

Pour se faire une intuition d'un champ gaussien, on va considérer une manière d'en construire un, en utilisant un produit de convolution contre une fonction gaussienne.

Dans cette section, une fonction gaussienne est une fonction de la forme $x \mapsto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. On peut ajuster sa forme grâce à des parametres (sa moyenne μ, qui n'aura ici que peu d'importance et qu'on prendra nulle, et son écart-type σ).

Prendre le produit de convolution de deux fonctions f et g, c'est considérer la fonction $f * g : x \mapsto \int_{-\infty}^{+\infty} f(t)g(x-t)dt$. Ce produit est commutatif avec les bonnes conditions, mais dans notre cas particulier, on n'accorde pas la même importance aux deux fonctions : il y a notre fonction d'entrée, et une fonction gaussienne par laquelle on "multiplie" (au sens du produit de convolution) cette première fonction. Ainsi, on peut voir cette opération comme une courbe gaussienne qui se déplace sur la courbe de notre fonction et qui, à chaque point, associe l'aire recouverte par les deux courbes.

Maintenant qu'on a vu ce rappel sur le produit de convolution, voyons comment on s'en sert pour construire des champs gaussiens

Si on considère une suite de Q (par exemple $Q = 101$) valeurs gaussiennes non corrélées (par exemple, si elles sont indépendantes), et qu'on trace ces valeurs de 0 à 100, on obtient une courbe qui ressemble à du bruit. Maintenant, si on fait "passer" une courbe gaussienne dessus, cela va agir comme un filtre (appelé *filtre gaussien*) qui va faire une moyenne sur certaines valeurs, en fonction de σ.

On obtient ainsi un signal "lissé", car moyenné, qui dépend directement de nos valeurs aléatoires et du σ qu'on a choisi : c'est notre champ gaussien aléatoire. Il est donc décrit par 3 paramètres : la moyenne et l'écart-type de la variable gaussienne qui a fournit les valeurs, et le σ de notre fonction filtre.

5.2 Variables caractéristiques d'un champ gaussien

Revenons sur les variables utilisées, car une confusion est possible.

En effet, parmi les 3 paramètres, les 2 derniers peuvent sembler similaires. Ils ont pourtant un effet bien différent, car l'écart-type de la distribution va influencer les valeurs obtenues tandis que le σ de la fonction filtre g va seulement modifier la façon dont on traite ces variables.

Pour les différencier, on ne va pas travailler exactement avec le σ de la fonction, mais avec sa FWHM (*Full Width at Half Maximum*) qui, comme son nom l'indique et comme on l'a vu précdemment, donne la largeur de la courbe à une hauteur égale à la moitié de son maximum. On peut relier rapidement cette FWHM à σ en écrivant l'égalité que vérifient les points x_1 et x_2 où g atteint la moitié de son maximum, soit $\frac{a}{2}$:

$$
\begin{aligned}
f(x_{12}) = \tfrac{a}{2} \quad &\Longrightarrow \quad e^{-\frac{(x_{12}-\mu)^2}{2\sigma^2}} = \tfrac{1}{2} \\
&\Longrightarrow \quad \tfrac{(x_{12}-\mu)^2}{2\sigma^2} = 2\ln(2) \\
&\Longrightarrow \quad \left| \begin{array}{l} x_1 = -\sqrt{2\ln(2)}\sigma + \mu \\ x_2 = \sqrt{2\ln(2)}\sigma + \mu \end{array} \right. \\
&\Longrightarrow \quad \text{FWMH} = x_2 - x_1 = 2\sqrt{2\ln(2)}\sigma \approx 2,35\sigma
\end{aligned}
$$

Un champ gaussien est donc déterminé par 3 paramètres : sa moyenne μ, son écart-type σ et sa FWHM. À partir de maintenant, on notera σ l'écart-type de la distribution et non plus de la fonction, en gardant uniquement la FWHM.

On peut remarquer que :

1. la FWHM traduit la régularité du champ obtenu : en effet, plus la FWHM est grande, plus un grand nombre de valeurs sera pris en compte dans le moyennage, plus les variations seront lentes
2. à la limite où FWHM $\to \infty$, on a un champ contant qui équivant à tirer une valeur suivant une loi $\mathcal{N}(\mu, \sigma)$ (cette dernière affirmation peut être montrée en simulant un grand nombre de champs gaussien de FWHM très grande (donc des champs quasi-constants) et montrer qu'ils se répartissent selon une loi normale)

5.3 Génération de champ gaussien par produit matriciel

En pratique, pour générer un champ gaussien, on peut utiliser le principe de convolution décrit jusque-là, mais il y a une méthode plus simple.

Celle-ci consiste à utiliser un produit matriciel, et va donc simplement consister à calculer notre champ gaussien y (un vecteur $Q \times 1$) ainsi : $y = Cz$.

z représente ici nos Q valeurs non corrélées (avec toujours comme exemple $Q = 101$), de dimensions $Q \times 1$.

La matrice C, de dimensions $Q \times Q$, est légèrement plus difficile à déterminer. Elle correspond à la matrice de convolution, qui code la manière dont chaque point "influence" les autres.

Pour l'obtenir, on code d'abord une matrice D de distance, telle que $D[i, j] = \frac{j-i}{Q-1}$ (*cf.* figure 4).

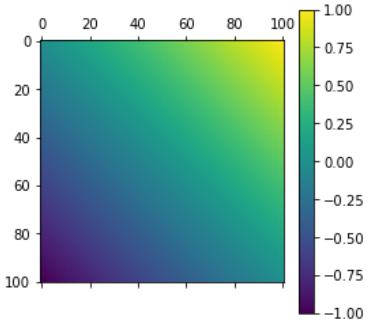


FIGURE 4 – Matrice D

On calcule ensuite l'image de chaque coefficient par un filtre gaussien tel que ceux décrits précedemment, pour que le coefficient $A[i, j]$ corresponde à la "distance exponentielle" (à savoir l'image de la distance par une fonction gaussienne de FWHM voulue) entre les points i et j (*cf.* figure 5) : plus des points sont proches, plus ils vont s'influencer l'un l'autre, plus le coefficient correspondant dans A sera proche de 1, et plus les points sont éloignés, plus le coefficient sera proche de 0.

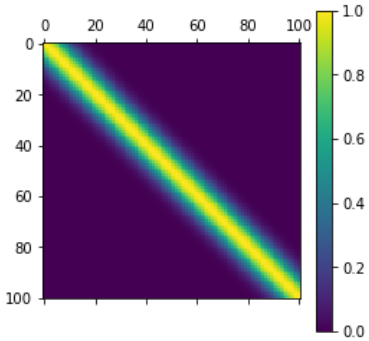


FIGURE 5 – Matrice A

Cette matrice correspond à la matrice de covariance entre les différentes valeurs de notre champ gaussien.

Comme elle est symétrique réelle, d'après le théorème spectral, elle est diagonalisable en base orthonormée. Notons $A = VUV^T$ sa décomposition en base orthonormée, avec U une matrice diagonale contenant les valeurs propres de A.

On va alors construire $C = V\sqrt{U}V^T$, où $\sqrt{U}$ est la matrice diagonale dont les coefficients diagonaux sont la racine carrée de chaque élément diagonal positif de U.

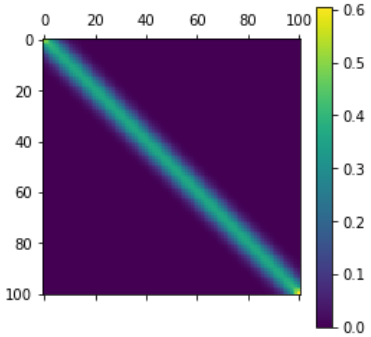


FIGURE 6 – Matrice C

Une fois cette matrice de convolution obtenue, on peut générer des champs gaussien de cette FWHM en multipliant un vecteur gaussien aléatoire par C.

Il n'est pas nécéssaire de recalculer C à chaque fois, seulement si on change de FWHM.

Pour résumer :

- on calcule Q valeurs non corrélées suivant la loi $\mathcal{N}(\mu, \sigma)$

- on calcule la matrice de covariace C, qui code dans quelle mesure chaque valeur influence chaque autre valeur, en fonction de notre FWHM

- on fait le produit des deux pour avoir un vecteur où chaque point dépend effectivement de ses voisins plus ou moins proches

OU

- on fait le produit de convultion de la courbe obtenue par un filtre gaussien de FWHM voulue

5.4 Exemple

Illustrons ce procédé :

On considère nos 101 valeurs aléatoires dans un vecteur x (*cf.* figure 7).

On peut ensuite calculer la matrice C (*cf.* figure 6), dépendant de la FWHM, et faire le produit matriel $C \times x$ pour obtenir les courbes orange de la figure 7.

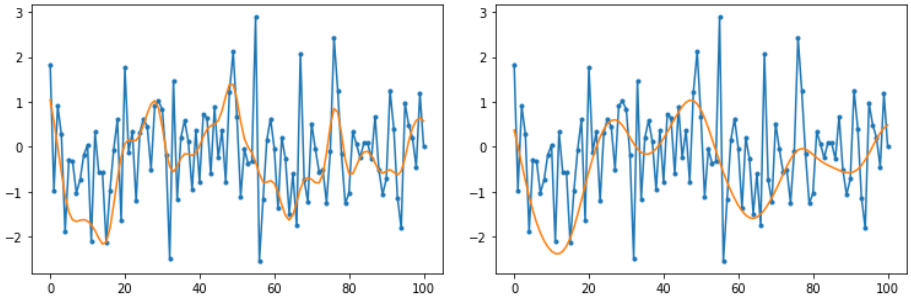


FIGURE 7 – Vecteur x (en bleu) qui donne des champs gaussiens (en orange) de FWHM = 5 (gauche) et FWHM = 10 (droite)

Ainsi, on obtient des champs gaussiens aléatoires générés très rapidement. Cela va nous permettre de retrouver les valeurs de seuil à 5% fournies par la théories des champs aléatoires. Ces valeurs sont calculées par une méthode trop compliquée pour être présentée ici car elle utilise des outils trop avancés, mais en générant un grand nombre de champs, on peut déterminer le seuil u tel que 5% de nos champs dépassent ce seuil.

À titre d'exemple, estimons le seuil u à 5% pour un champ de FWHM 10. Pour cela, on génère un grand nombre (N = 10000) champs gaussiens de FWHM 10, on stocke la valeur la plus extrême, et on trace l'histogramme correspondant. (*cf.* figure 8 pour un exemple des champs de FWHM = 10 et de leur extremum, avec N = 10 dans un soucis de visibilité)

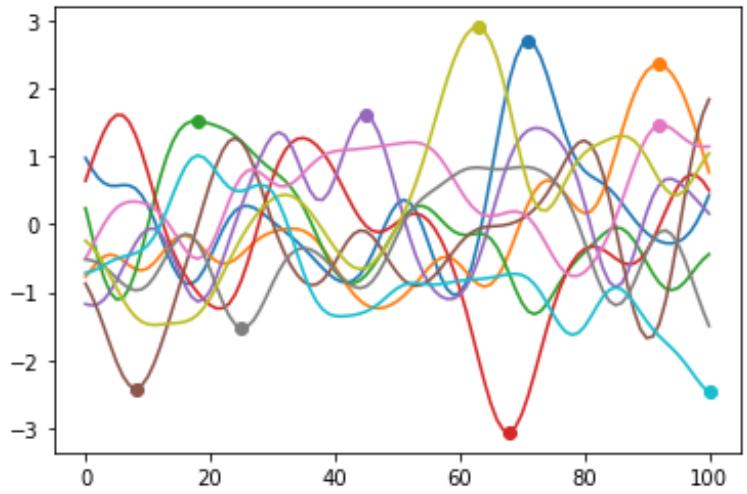


FIGURE 8 – 10 champs gaussiens et leur extremum

Si on fait la même opération avec 10000 champs générés, on peut ainsi tracer l'histogramme des valeurs extrêmes (on travaille avec les valeurs absolues, donc les valeurs extrêmes peuvent être des maxima ou des minima). On peut également y ajouter le tracé de la densité de probabilité inversée, à savoir la fonction $u \mapsto$

$P(z_{extreme} > u)$.

Les figures 9 et 10 donnent les tracés pour 3 tirages aléatoires de 10000 champs gaussiens, pour une FWHM égale à 10.

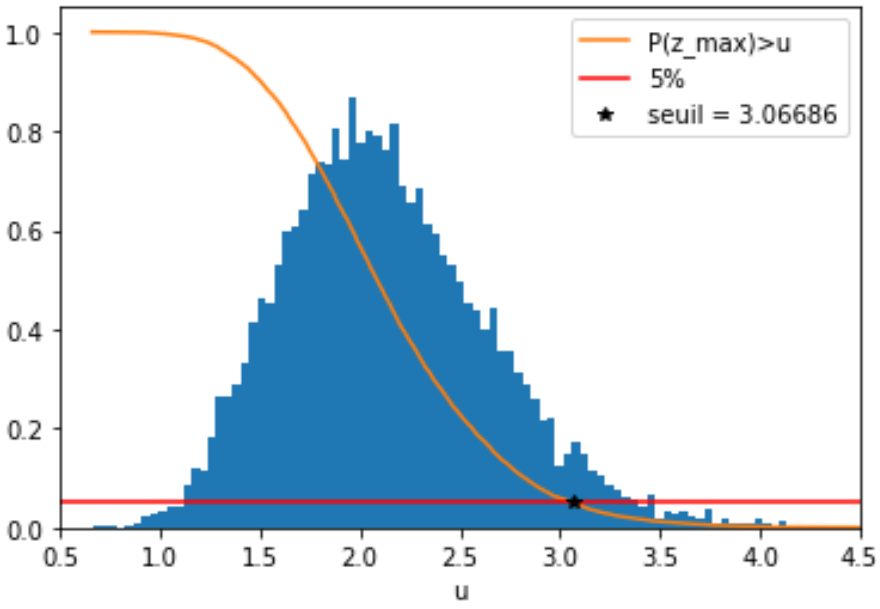


FIGURE 9 – Histogramme et détermination du seuil pour FWHM = 10

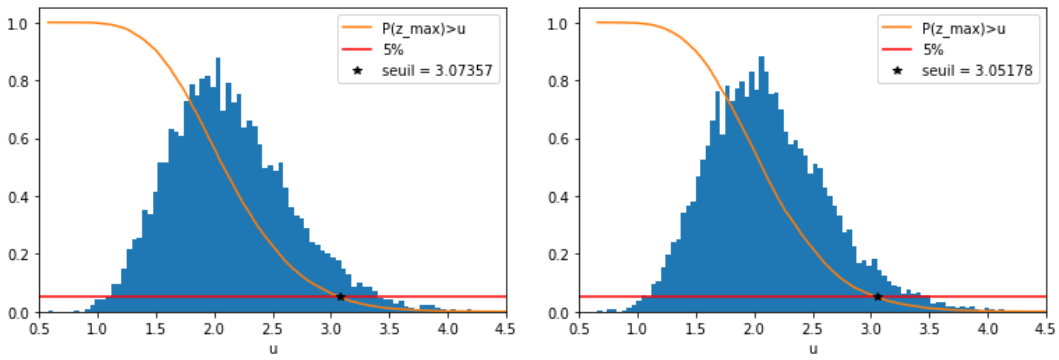


FIGURE 10 – Histogramme et détermination du seuil pour FWHM = 10 - deux autre applications

On peut lire à chaque fois l'intersection de la courbe orange et du seuil rouge : on trouve que le seuil de significativité vaut autour de $u = 3,06$.

Cela correspond avec la valeur théorique issue de la théorie des champs aléatoires. En effet, en utilisant le module rft1d sur lequel s'appuie le code, et qui a été utilisé pendant ce stage, on trouve une valeur seuil $u = 3.0651$ pour FWHM = 10.

On peut réaliser la même opération pour une FWHM égale à 5 : on obtient un seuil proche de la valeur théorique prédite, à savoir $u = 3.2767$ pour FWHM = 5.

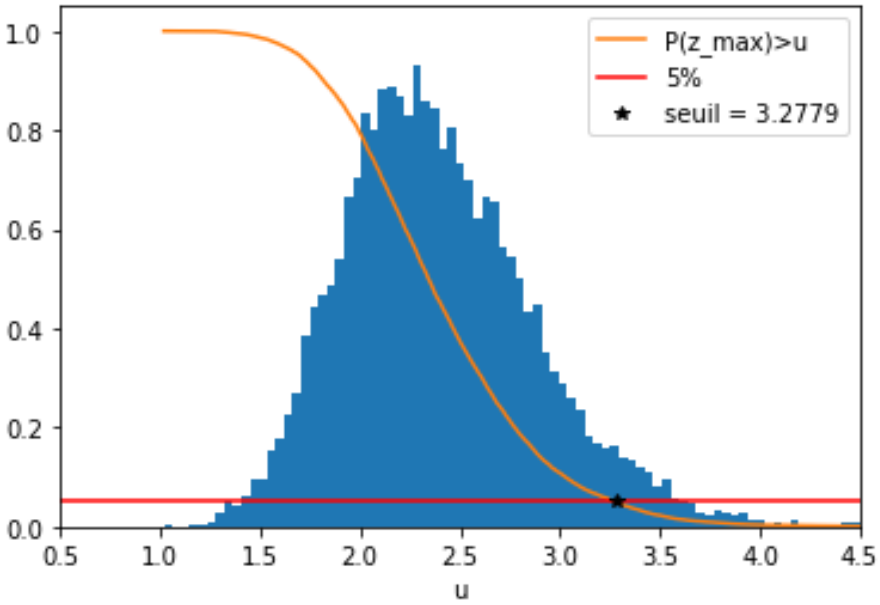


FIGURE 11 – Histogramme et détermination du seuil pour FWHM = 5

6 Retour à l'article et conclusion

On peut ainsi faire l'analyse SPM des données issues de l'expérience et comparer le résultat à l'analyse faite par l'auteur de l'article.

Dans la grande majorité des cas, on trouve la même chose que l'auteur, comme par exemple pour les courbes de puissance extérieure du début de l'article.

En effet, si on effectue l'analyse SPM, on obtient la courbe de droite sur la figure 12. Le seuil de significativité est indiqué par les deux barres horizontales rouges, et les clusters de différence significative sont grisés.

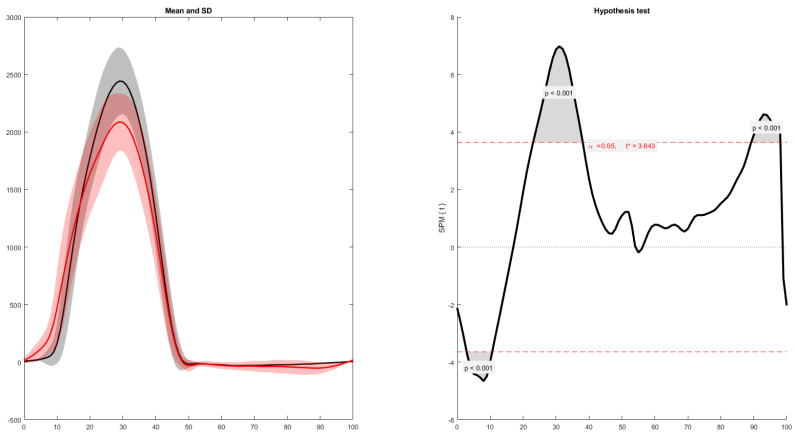


FIGURE 12 – Analyse SPM Pext2

L'article fait l'analyse suivante de ces courbes :

> The external power on the fixed stretcher was significantly lower for the first 20% or so of the cycle length in the propulsion phase (handle positions : catch position to -0.06 m). Then, the rowers produced an external power significantly greater for about 30% of the cycle length (handle positions : -0.46 to -0.94 m) and for the last 20% of the cycle length in the recovery phase (handle positions : -0.06 to catch position). [Col06]

→ Avec l'analyse SPM, on voit qu'on a bien le pic significatif dans un sens au début du cycle de propulsion, et les deux pics dans l'autre sens en milieu/fin de cycle de propulsion et en fin de cycle de récupération.

Mais ce n'est pas toujours le cas, et il arrive que sur des cas à la limite de la significativité, l'analyse SPM fournisse un résultat différent de ce qui avait été conclu en premier lieu.

Considérons pour cela les courbes de la force verticale appliquée par le sujet sur le siège, données figure 13.

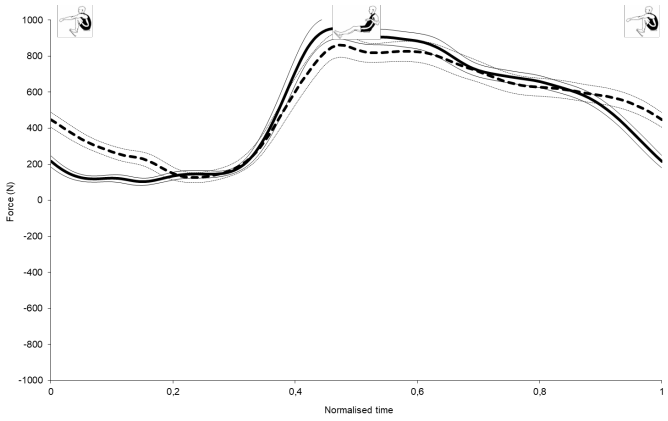


FIGURE 13 – Courbes moyenne de la force verticale sur le siège

On peut également réaliser l'analyse SPM des données pour obtenir la courbe figure 14.

L'article dit :

> "The finish value was 11.4% lower ($P < 0.01$)." [Col06]

> *Ce que est appelé* finish value *est la valeur en milieu de cycle, à savoir la transition entre la phase de propulsion et la phase de retour. Ici, c'est la valeur maximale des courbes, donc celle autour de 45% de cycle.*

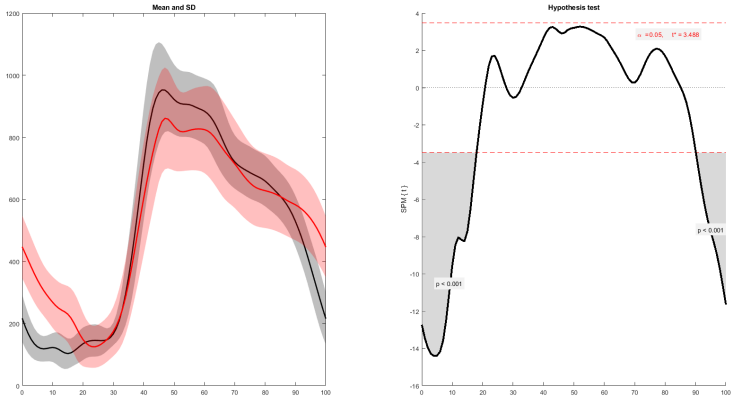


FIGURE 14 – Analyse SPM SeatFZ

→ Ici, on est dans le cas limite où le modèle 0D voit une différence significative alors que l'analyse SPM 1D ne conclut pas à une telle différence. En effet, l'article trouve une p-value très significative alors que l'analyse SPM montre une courbe qui est juste sous le seuil de significativité. Ainsi, les deux courbes étudiées (ergomètre fixe contre libre) sont différentes, avec une p-value proche du seuil choisi (5%), mais c'est loin de la significativité annoncée dans l'article.

En revanche, on peut remarquer que l'analyse SPM fournit des informations sur la courbe entière : entre 40% et 60% du cycle, les courbe présentent une différence proche du seuil, et pas uniquement au niveau de leur maximum. Ainsi, si l'on s'intéresse à cette partie dans l'objectif d'améliorer les performances des ergomètres ou des sportifs, on voudra considérer les 20% correspondants du cycle, et pas juste un extremum.

On a donc bien une différence entre la comparaison "usuelle" de courbes et la comparaison statistique SPM. Cette différence est assez minime (la grande majorité des courbes aboutissent au même résultat indépendemment de la méthode), mais elle montre quand même les avantages de la méthode SPM, en plus de la rapidité d'analyse (trier les données, extraire de chaque courbe les points significatifs et tester tous les points qui semblent differents représente une quantité de travail qui peut vite devenir imortante) et de la rigueur d'analyser un objet avec un outil parfaitement adapté.

7 Bibliographie et remerciements

Références

[Fri95] Karl J Friston. “Statistical Parametric Maps in Functional Imaging : A General Linear Approach.” In : *Human Brain Mapping 2* (1995). doi : http://dx.doi.org/10.1002/hbm.460020402.

[Col06] Floren Colloud. “Fixed versus free-floating stretcher mechanism in rowing ergometers : Mechanical aspects”. In : *Journal of Sports Sciences* (2006). doi : http://dx.doi.org/10.1080/02640410500189256.

[Pat11] Todd Pataky. “One-dimensional statistical parametric mapping in Python.” In : *Computer Methods in Biomechanics and Biomedical Engineering* (2011). doi : http://dx.doi.org/10.1080/10255842.2010.527837.

[Pat16] Todd Pataky. “The probability of false positives in zero-dimensional analyses of one dimensional kinematic, force and EMG trajectories.” In : *Journal of Biomechanics* (2016). doi : http://dx.doi.org/10.1016/j.jbiomech.2016.03.032.

Remerciements

Je tiens à remercier l’équipe de l’IBHGC qui m’a accueilli malgré mon profil un peu atypique, des chercheurs aux autres stagiaires en passant par les doctorants et les conférenciers, ainsi que pour les échanges et apprentissages très intéressants que j’ai pu faire durant mon stage.

Merci tout particulièrement à Floren Colloud de m’avoir permis de réaliser ce stage, de m’avoir fait découvrir ce monde de la biomécanique, de m’avoir permis de participer et d’assister à des expérimentations qui ont été un vrai plus tout au long de mon stage.

Une pensée particulière pour Thomas Provot et Laura Valdes Tamayo, grâce à qui j’ai pu faire le lien entre mes études et ma passion.

Ce stage a été important pour moi, et aura peut-être une influence sur la suite de mes études, et c’est en grande partie grâce à ces 3 personnes.