Investigating swimming technical skills by a double partition clustering of multivariate functional data allowing for dimension selection

Antoine Bouvet [1,2,3] & Salima El Kolei [3] & Matthieu Marbac [3]

[1] *Univ. Rennes 2, ENS Rennes, M2S Laboratory-EA 7470, France, antoine.bouvet@ens-rennes.fr*
[2] *Inria Rennes Bretagne Atlantique, MIMETIC, France*
[3] *Univ. Rennes, Ensai, CNRS, CREST—UMR 9194, France, salima.el-kolei@ensai.fr, matthieu.marbac-lourdelle@ensai.fr*

Résumé. Monitorer les compétences techniques des nageurs constitue un défi majeur en sciences du sport afin d'améliorer les performances. Cela peut être fait en analysant les données fonctionnelles multivariées mesurées par des capteurs miniaturisés tels que les centrales inertielles (IMU). Ces données sont composées de six dimensions décrivant la cinématique 3D du nageur au cours du temps à travers les accélérations et la vitesse angulaire. Pour investiguer les niveaux techniques en crawl sur la base des enregistrements IMU, un modèle de mélange produisant deux partitions complémentaires est proposé et reflète, pour chaque nageur, son pattern de nage et sa capacité à le reproduire. Contrairement aux approches habituelles de clustering de données fonctionnelles, celle-ci prend également en compte les informations présentes dans les termes d'erreur résultant de la décomposition en bases fonctionnelles. En effet, après avoir décomposé en bases fonctionnelles avec un nombre fini d'éléments à la fois le signal original (mesurant le pattern de nage) et le signal des termes d'erreur au carré (mesurant la capacité à le reproduire), la méthode ajuste la distribution conjointe des coefficients liés aux deux décompositions en tenant compte de la dépendance entre les deux partitions. La modélisation de cette dépendance est obligatoire puisque la difficulté à reproduire un pattern de nage dépend de sa forme. En outre, une décomposition éparse de la distribution au sein des composantes permet de sélectionner les dimensions pertinentes lors du clustering. Cela permet d'améliorer l'interprétation technique du modèle pour les utilisateurs. Les partitions obtenues sur les données IMU agrègent la variabilité cinématique de la nage associée aux compétences techniques et permettent d'identifier les habiletés biomécaniques pertinentes pour des sprinteurs en crawl.

Mots-clés. Clustering, Sélection de variables, Données fonctionnelles, Performance sportive

Abstract. Investigating technical skills of swimmers is a challenge for sports science to reach performance improvement. It can be achieved by analyzing multivariate functional data recorded by miniaturized sensors such as Inertial Measurement Units (IMU). These data are composed of six dimensions describing swimmer's kinematic through the 3D accelerations and angular velocity temporal records. To investigate technical levels of front-crawl swimmers, a new model-based approach is introduced to obtain two complementary partitions reflecting, for each swimmer, its swimming pattern and its ability to reproduce it based on the IMU records. Contrarily to the usual approaches for functional data clustering, the proposed approach also considers the information of the error terms resulting from the functional

basis decomposition. Indeed, after decomposing into functional basis with finite number of elements both the original signal (measuring the swimming pattern) and the signal of squared error terms (measuring the ability to reproduce it), the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modeling this dependency is mandatory since the difficulty of reproducing a swimming pattern depends on its shape. Moreover, a sparse decomposition of the distribution within components that permits a selection of the relevant dimensions during clustering is proposed. It allows the improvement of the technical interpretation of the model for users. The partitions obtained on the IMU data aggregate the kinematical stroke variability linked to swimming technical skills and allow relevant biomechanical ability for front-crawl sprinter to be identified.

Keywords. Clustering, Feature selection, Functional data, Sport performance

1 Introduction

Tracking of technical skills by comprehensive training monitoring is a main challenge for sport performance improvement, especially in swimming that requires efficient movement in aquatic environment. Swimming technique is defined by the repetition of similar but not identical stroke patterns constituted of instabilities called biomechanical variability (Fernandes et al. (2022a)). Stroke variability plays a major role in generating swimming speed because it is related to swimming efficiency and differs according to swimming performance levels (Seifert et al. (2016)). Thus, technical skills can be described by this biomechanical variability and quantified through two complementary and associated aspects of motion: the swimming pattern and the ability to reproduce it (Fernandes et al., 2022b). Development of automatic methodologies supporting on-board investigation of both of these components is promising. To do this, Inertial Measurement Units (IMU) are used to provide embedded kinematical data collection regarding sports related movements through tri-axial accelerometer and gyroscopic temporal records. However, the current literature is focused on empirical indicators describing stroke patterns without taking their functional nature into account. Hence, this leads to limited insights regarding underlying kinematical variability defining technical skills and making their conclusions weak or not representative. Their main limitation is due to the absence of a powerful statistical model able to analyze the complex multivariate dynamics of swimming patterns.

Statistical analysis of multivariate functional periodic IMU data supporting monitoring of kinematical variability could be achieved by a dependent double partition clustering measuring the swimming pattern (*i.e.,* first latent partition) and the ability to reproduce it (*i.e.,* second latent partition). Indeed, the type of stroke pattern directly impacts its repeatability. Moreover, dimension selection during this double clustering could allow the IMU axes to be identified for explaining disparities between swimmers and associated technical skills.

Traditionally, statistical methods allowing functional data to be clustered consider a decomposition of the data into a functional basis, and then use classical multivariate methods for

finite dimensional data directly on these basis coefficients. Considering this approach, different model-based clustering methods have been developed for univariate functional data as well as for multivariate functional data (Bouveyron et al. (2019). In sports science, model-based clustering approaches for functional data provide a useful framework for new insights on performance (Leroy (2020)) that can outpace classical approaches relying on empirical analysis that lead to limited kinematical information (Mallor et al., 2010). For all the methods considering functional basis decompositions, the loss of information due to the approximation of the original data into a functional basis is neglected. However, for investigating swimming techniques, it is crucial to keep the information of the dispersion around the swimming pattern that is traditionally lost by using the basis decomposition.

Some model-based clustering approaches have been developed to perform a selection of the variables (Marbac and Sedki, 2017). Selecting variables is very challenging in clustering because the role of a variable is defined with respect to a variable that is not observed. Thus, the selection of the variables and the clustering need to be performed simultaneously. In the context of clustering multivariate functional data, to the best of our knowledge, no methods can lead to a detection of the dimensions that are relevant for clustering. However, one can consider a natural extension of the model-based clustering approach performing feature selection. Indeed, when the functional data are decomposed into a functional basis, claiming that a dimension of the functional data is not relevant for clustering means that all the coefficients of the functional basis, related to this dimension, are not informative for the clustering.

We propose a double partition clustering approach to investigate swimming technical skills using IMU data that are periodic due to the repetition of strokes. Moreover, the proposed model-based approach allows for the identification of the discriminative dimensions for each partition. After decomposing both of the original signals (*i.e.,* measuring the swimming pattern) and the signals of squared error terms (*i.e.,* measuring the ability to reproduce the swimming pattern) into a Fourier basis, the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modelling this dependency is important because the difficulty of reproducing a swimming pattern depends on its shape and then on technical skills. The model considers that the information about the swimming pattern that measures the kinematical smoothness is contained in the coefficients arising from the decomposition of the original signal while the information about the ability of reproducing the pattern is contained in the coefficients arising from the decomposition of the signal of the squared error terms. As usual for a standard model-based approach to cluster functional data, a sparse decomposition of the distribution within components is used. Here, we consider a conditional independence between the coefficients of different dimensions within the component. This assumption allows an automatic selection to be made of the relevant dimensions during clustering that improves the accuracy of the estimates and that highlights the dimensions that are discriminative for both partitions.

This paper is organized as follows. Section 2 presents the double clustering model-based framework and dimension selection for multivariate functional data. Identifiability issues of this new model are investigated and presented. Mathematical details, model inference and numerical experiments are available in Bouvet et al. (2024) . Section 3 is devoted to the analysis of the SWIMU data and the consequent biomechanical interpretation for technical swimming skills analysis. Section 4 gives a conclusion.

2 Model-based approach for estimating a double partition from multivariate functional data

2.1 Mixture model on basis coefficients

We consider a random sample $\mathbf{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ composed of n independent and identically distributed multivariate time series. Each individual i is described by a J-dimensional discrete-time time series $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^\top, \ldots, \boldsymbol{X}_{iJ}^\top)^\top$, where $\boldsymbol{X}_{ij} = (X_{ij}(1), \ldots, X_{ij}(T_i))^\top$ and $X_{ij}(t) \in \mathbb{R}$ denotes the value of dimension j of the multivariate time series measured on subject i at time t, T_i being the length of the multivariate time series recorded on subject i. Each univariate time series admits a basis expansion leading to

$$X_{ij}(t) = \boldsymbol{Y}_{ij}^\top \boldsymbol{\psi}_j(t) + \varepsilon_{ij}(t), \tag{1}$$

where $\boldsymbol{Y}_{ij} \in \mathbb{R}^{G_j}$ is a G_j-variate random variable that groups the G_j basis coefficients, $\boldsymbol{\psi}_j(t) = (\psi_{j1}(t), \ldots, \psi_{jG_j}(t))^\top$ is the vector containing the values of the G_j basis functions evaluated at time t and where the term of random error $\varepsilon_{ij}(t)$ is supposed to be centered given the natural filtration $\mathcal{F}_i(t)$ (*i.e.,* $\mathbb{E}[\varepsilon_{ij}(t) \mid \mathcal{F}_i(t-1)] = 0$). Let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^\top, \ldots, \boldsymbol{Y}_{iJ}^\top)^\top \in \mathbb{R}^G$ be the vector of length $G = \sum_{j=1}^J G_j$ that gathers the basis coefficients of subject i for the J dimensions and let $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}^\top, \ldots, \boldsymbol{\varepsilon}_{iJ}^\top)^\top$ be the vector of the error terms of individual i where $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij}(1), \ldots, \varepsilon_{ij}(T_i))^\top$. In addition, we express each univariate time series of the squared error term in a functional basis with H_j elements as follows

$$\varepsilon_{ij}^2(t) = \boldsymbol{Z}_{ij}^\top \boldsymbol{\Pi}_j(t) + \xi_{ij}(t), \tag{2}$$

where $\boldsymbol{Z}_{ij} \in \mathbb{R}^{H_j}$ is a H_j-variate random variable that groups the H_j basis coefficients, $\boldsymbol{\Pi}_j(t) = (\Pi_{j1}(t), \ldots, \Pi_{jH_j}(t))^\top$ is the vector containing the values of the H_j basis functions evaluated at time t, $\mathbb{E}[\xi_{ij}(t) \mid \mathcal{F}_i(t-1)] = 0$.
The decomposition defined by (1)-(2) considers two general functional basis with intrinsic dimension assumed to be finite. To model the cyclical pattern of the swimmers' motion, it seems appropriate to use a Fourier basis for the decomposition of each time series. The period is the same among the dimensions of the functional data but the degrees of the Fourier basis can be different. Also, although the period of the basis differs for each swimmer due to the different swimming periods between the swimmers. This follows that, if swimmer i has the same period ζ_i, we have $\boldsymbol{\psi}_j(t) := \boldsymbol{\psi}_j(t;\zeta_i)$ where $\boldsymbol{\psi}_j(t;\zeta_i) = (\boldsymbol{\psi}_{j1}(t;\zeta_i), \ldots, \boldsymbol{\psi}_{jG_j}(t;\zeta_i))^\top$, $\boldsymbol{\psi}_{j1}(t;\zeta_i) = 1$ and $\boldsymbol{\psi}_{j(2\ell)}(t;\zeta_i) = \cos(2\pi\ell t/\zeta_i)$ and $\boldsymbol{\psi}_{j(2\ell+1)}(t;\zeta_i) = \sin(2\pi\ell t/\zeta_i)$, for $\ell = 1, \ldots, (G_j-1)/2$, and we have $\boldsymbol{\Pi}_j(t) := \boldsymbol{\Pi}_j(t;\zeta_i)$ where $\boldsymbol{\Pi}_j(t;\zeta_i) = (\boldsymbol{\Pi}_{j1}(t;\zeta_i), \ldots, \boldsymbol{\Pi}_{jH_j}(t;\zeta_i))^\top$, $\boldsymbol{\Pi}_{j1}(t;\zeta_i) = 1$ and $\boldsymbol{\Pi}_{j(2\ell)}(t;\zeta_i) = \cos(2\pi\ell t/\zeta_i)$ and $\boldsymbol{\Pi}_{j(2\ell+1)}(t;\zeta_i) = \sin(2\pi\ell t/\zeta_i)$, for $\ell = 1, \ldots, (H_j-1)/2$.
We aim at grouping the swimmers according to their swimming patterns as well as their abilities to reproduce them. This implies the estimation of two latent categorical variables $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{iK})^\top$ and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{1L})^\top$ that indicate the swimming pattern and the dispersion around this pattern for swimmer i respectively, K and L denoting the number of different swimming patterns and the number of types of dispersion around a pattern. Since the basis coefficients $\boldsymbol{Y}_i$ depend on the mean behavior of $\boldsymbol{X}_i$, we consider that this vector contains all the information about $\boldsymbol{V}_i$. Similarly, since the basis coefficients $\boldsymbol{Z}_i$ depend on the

squared of the error terms $\boldsymbol{\varepsilon}_i$, we consider that this vector contains all the information about $\boldsymbol{W}_i$. Thus, the model assumes conditional independence between $\boldsymbol{Y}_i$ and $\boldsymbol{W}_i$ given $\boldsymbol{V}_i$ and conditional independence between $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{V}_i$ given $\boldsymbol{W}_i$. Finally, it allows for dependency between $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ and it assumes conditional independence between $\boldsymbol{Y}_i$ and $\boldsymbol{\varepsilon}_i$ given $\boldsymbol{W}_i$ and $\boldsymbol{V}_i$. Thus, the basis coefficients follow a specific mixture model with $K \times L$ components defined by the following probability distribution function (pdf)

$$p(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \pi_{k\ell} f_k(\boldsymbol{y}_i; \boldsymbol{\alpha}_k) g_\ell(\boldsymbol{z}_i; \boldsymbol{\beta}_\ell), \tag{3}$$

where $\pi_{\ell k} := \mathbb{P}(V_{ik} = 1, W_{i\ell} = 1) > 0$, $\sum_{k=1}^{K} \sum_{\ell=1}^{L} \pi_{k\ell} = 1$, $\boldsymbol{\theta}$ groups all parameters of model $\boldsymbol{m}$, f_k is the pdf of cluster k for the swimming pattern parameterized by $\boldsymbol{\alpha}_k$ that defines the conditional distribution of $\boldsymbol{Y}_i$ given $V_{ik} = 1$ and g_ℓ is the pdf of cluster ℓ for the ability of reproducing the swimming pattern parameterized by $\boldsymbol{\beta}_\ell$ that defines the conditional distribution of $\boldsymbol{Z}_i$ given $W_{i\ell} = 1$.
The model defined by (3) is a parsimonious mixture model that imposes equality constraints between the parameters of some pdfs of its components. It implies that the marginal distribution of $\boldsymbol{Y}_i$ is a mixture model with K components and that the marginal distribution of $\boldsymbol{Z}_i$ is a mixture model with L components. Note that dependency between $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ is considered by (3) and thus this model is not equivalent to a product of two mixture models modeling the marginal distributions of $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$. Finally, we present two properties of the model. Detailed proofs for lemmas 1 and 2 are presented in Bouvet et al. (2024).

Lemma 1. *If the parameters of the marginal distributions $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ are identifiable, then the parameters of model defined by* (3) *are identifiable.*

The second property states that considering the dependency between the two latent partitions, and thus using the joint distribution of $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ for estimating the two partitions leads to a better estimator of both partitions than considering an estimator of $\boldsymbol{V}_i$ based on $\boldsymbol{Y}_i$ and an estimator of $\boldsymbol{W}_i$ based on $\boldsymbol{Z}_i$ when $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are not independent. Let $\Upsilon_V(\boldsymbol{Y}_i)$ and $\Upsilon_V(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ be the applications that associate an estimator of $\boldsymbol{V}_i$ (*i.e.,* a vector of length K composed by zeros except for one coordinate that is equal to one) by using the *MAP* rule (*i.e.,* affecting an observation to the most likely cluster) based on the distribution of $\boldsymbol{Y}_i$ and $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ respectively. Let $\Upsilon_W(\boldsymbol{Z}_i)$ and $\Upsilon_W(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ be the applications that associate an estimator of $\boldsymbol{W}_i$ (*i.e.,* a vector of length L composed by zeros except for one coordinate that is equal to one) by using the *MAP* rule based on the distribution of $\boldsymbol{Z}_i$ and $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ respectively.

Lemma 2. *If the model* (3) *holds true and if the $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are not independent then, under the assumption of Lemma 1 and if the densities f_k and g_ℓ are continuous for any k and ℓ, then*

$$\mathbb{E}[\Upsilon_V(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \neq \boldsymbol{V}_i] < \mathbb{E}[\Upsilon_V(\boldsymbol{Y}_i) \neq \boldsymbol{V}_i] \text{ and } \mathbb{E}[\Upsilon_W(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \neq \boldsymbol{W}_i] < \mathbb{E}[\Upsilon_W(\boldsymbol{Z}_i) \neq \boldsymbol{W}_i].$$

2.2 Parsimonious model for detecting the relevant dimensions

Since the decompositions into the functional basis can produce high-dimensional vectors, it is usual to assume parsimonious constraints on the dependency within components. Here, we consider that the mixture components belong to the same parametric family and that the coefficients related to different dimensions are conditionally independent given the latent variables. This leads to $\boldsymbol{Y}_{ij} \perp \boldsymbol{Y}_{ij'}|\boldsymbol{V}_i$ and $\boldsymbol{Z}_{ij} \perp \boldsymbol{Z}_{ij'}|\boldsymbol{W}_i$, for $j \neq j'$. We denote by $\phi_j(\cdot;\boldsymbol{\alpha}_{kj})$ the G_j-dimensional density of $\boldsymbol{Y}_{ij}$ within component k parameterized by $\boldsymbol{\alpha}_{kj}$ and by $\varphi_j(\cdot;\boldsymbol{\beta}_{\ell j})$ the H_j-dimensional density of $\boldsymbol{Z}_{ij}$ within component ℓ parameterized by $\boldsymbol{\beta}_{\ell j}$. Therefore, we have

$$f_k(\boldsymbol{y}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{J} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{kj}) \text{ and } g_\ell(\boldsymbol{z}_i;\boldsymbol{\beta}_\ell) = \prod_{j=1}^{J} \varphi_j(\boldsymbol{z}_{ij};\boldsymbol{\beta}_{\ell j}). \tag{4}$$

The model can consider any parametric multivariate density for ϕ_j and φ_j and thus allows the usual parametric assumptions to be made to cluster functional data based on the coefficients of their basis extension. The main benefits of (4) is that it easily permits a selection of the dimensions that are relevant for clustering in (3) and so allows to extend variable selection methods for clustering to functional data. In our context, a dimension is relevant for estimating one partition if the coefficients related to this dimension do not have the same distribution among the mixture components. Therefore, dimension j is irrelevant for estimating the swimming pattern if $\boldsymbol{\alpha}_{1j} = \ldots = \boldsymbol{\alpha}_{Kj}$ while this dimension is irrelevant for estimating the ability of reproducing the swimming pattern if $\boldsymbol{\beta}_{1j} = \ldots = \boldsymbol{\beta}_{Lj}$. We denote by $\boldsymbol{\Omega} \subseteq \{1,\ldots,J\}$ and $\boldsymbol{\Gamma} \subseteq \{1,\ldots,J\}$ the indexes of the dimensions that are relevant for estimating the swimming pattern and for estimating the ability of reproducing the swimming pattern respectively. Thus, for a fixed model $\boldsymbol{m} = \{K, L, \boldsymbol{\Omega}, \boldsymbol{\Gamma}\}$, using (3)-(4) and the definition of the relevant dimensions, the pdf of the observed data is defined by

$$f(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \left[\prod_{j\in\boldsymbol{\Omega}^c} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{1j}) \prod_{j'\in\boldsymbol{\Gamma}^c} \varphi_{j'}(\boldsymbol{z}_{ij'};\boldsymbol{\beta}_{1j'})\right] \times \sum_{k=1}^{K}\sum_{\ell=1}^{L} \pi_{k\ell} \prod_{j\in\boldsymbol{\Omega}} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{kj}) \prod_{j'\in\boldsymbol{\Gamma}} \varphi_{j'}(\boldsymbol{z}_{ij'};\boldsymbol{\beta}_{\ell j'}). \tag{5}$$

The following lemma gives sufficient conditions to state the identifiability of the parameters of the proposed model (5). Detailed proof for lemma 3 is presented in Bouvet et al. (2024).

Lemma 3. *If card*$(\boldsymbol{\Omega}) \geq 1$, *card*$(\boldsymbol{\Gamma}) \geq 1$, *exist* $j \in \boldsymbol{\Omega}$ *and* $j' \in \boldsymbol{\Gamma}$ *such that the marginal distribution of* $\boldsymbol{Y}_{ij}$ *and* $\boldsymbol{Z}_{ij'}$ *is identifiable, then the parameters of model* (5) *are identifiable.*

3 Analysis of SWIMU data

The SWIMU database used in this study includes $n = 68$ all-out 25m front-crawl from recreational to world-class swimmers. The participants were instrumented with one IMU located

on the sacrum. To deal with the issue of curve alignment, data pre-processing is conducted in order to set the first frame at the beginning of a stroke by zero-crossing on second-order Butterworth band-pass filtered between 0.1 and 1 Hz on mediolateral acceleration. Since the IMU swimming records are periodic, we decompose these multivariate functional data into a Fourier basis. The period is identified using the previously described zero-crossing. We select the degree of the Fourier basis G_j and H_j used for the decompositions (1) and (2) of each dimension j that minimizes the least square error obtained by leave-one-out cross-validation. Thus, the selected degrees are between 12 and 18 for the original signal decomposition and between 8 and 12 for the decomposition of the squared error terms.

In this application, we consider a Gaussian distribution within components with a diagonal matrix *i.e.,* ϕ_j is the pdf of Gaussian distribution with mean $\boldsymbol{\mu}_{jk} = (\mu_{jk1}, \ldots, \mu_{jkG_j})^\top$ and a diagonal covariance matrix. Similarly, φ_j is the pdf of Gaussian distribution with mean $\boldsymbol{\nu}_{j\ell} = (\nu_{j\ell 1}, \ldots, \nu_{j\ell H_j})^\top$ and a diagonal covariance matrix. Best model according to the BIC is composed of 2 clusters for the swimming pattern (*e.g.,* $K = 2$), 3 clusters for the ability of reproducing the pattern (*e.g.,* $L = 3$). Moreover, five among the six dimensions are selected for the swimming pattern (all the dimensions but the mediolateral angular velocity) and all the dimensions are selected for the ability of reproducing the pattern. Mediolateral angular velocity mainly reflects the pitch movement of the swimmers and so is not a discriminant constitutive motion of technical abilities for front-crawl sprint.

Figure 1 allows for an easy interpretation of the results. Indeed, it summarizes the swimming patterns by presenting in black plain lines, for each dimension j, the mean curves defined, for any $k \in \{1, 2\}$ and $t \in [0, 1[$, by $\overline{X}_{jk}(t) := \widehat{\boldsymbol{\mu}}_{jk}^\top \boldsymbol{\psi}_j(t; 1)$. Moreover, the three clusters of abilities for reproducing the swimming pattern are summarized, for each dimension j, by the repeatability region, at each time t, $[\overline{X}_{jk}(t) - 2\overline{\varepsilon}_{j\ell}(t), \overline{X}_{jk}(t) + 2\overline{\varepsilon}_{j\ell}(t)]$ defined as the area that differs from two standard deviations at each time t from the mean curve where, for any $\ell \in \{1, 2, 3\}$ and $t \in [0, 1[$, $\overline{\varepsilon}^2_{j\ell}(t) := \widehat{\boldsymbol{\nu}}_{j\ell}^\top \boldsymbol{\Pi}_j(t; 1)$.

The partition of swimming patterns is composed of a majority class (81%) presenting a smooth acceleration pattern with continuity of propulsive actions during the stroke cycle. This class, called the *smoothy* class is characterized by a lower acceleration variation on the propulsive axis (*i.e.,* longitudinal and anteroposterior acceleration) and a higher angular velocity variation on both longitudinal and anteroposterior axes in addition to mediolateral acceleration. The second class of swimming patterns (19%) is composed of more explosive and irregular stroke patterns. It includes higher acceleration variations than in the *smoothy* class particularly on the propulsive axis, associated with reduced mediolateral acceleration and angular velocity variations inducing more stability and a better alignment of the body. This steadiness especially occurs during the peak propelling phases of the cycle (*i.e.,* peaks on longitudinal acceleration). This class is then called the *jerky* class.

The partition of abilities to reproduce the swimming patterns is composed of two classes having equal proportions (37%) characterized by moderate and high stroke pattern repeatability, and of a third class (26%) with lower level. We respectively call them the *moderate*, *high* and *low repeatability* classes (see colors in Figure 1). The repeatability region around mean curves are not constant indicating unequal variance. In this way, this functional visualisation provides a useful understanding of kinematical variability during specific stroke cycle phases. The assumption of independence between both partitions is rejected according to a Pearson's Chi-squared test (χ^2=15.62, df=2, $p \leq$0.01). This confirms the biomechanical association

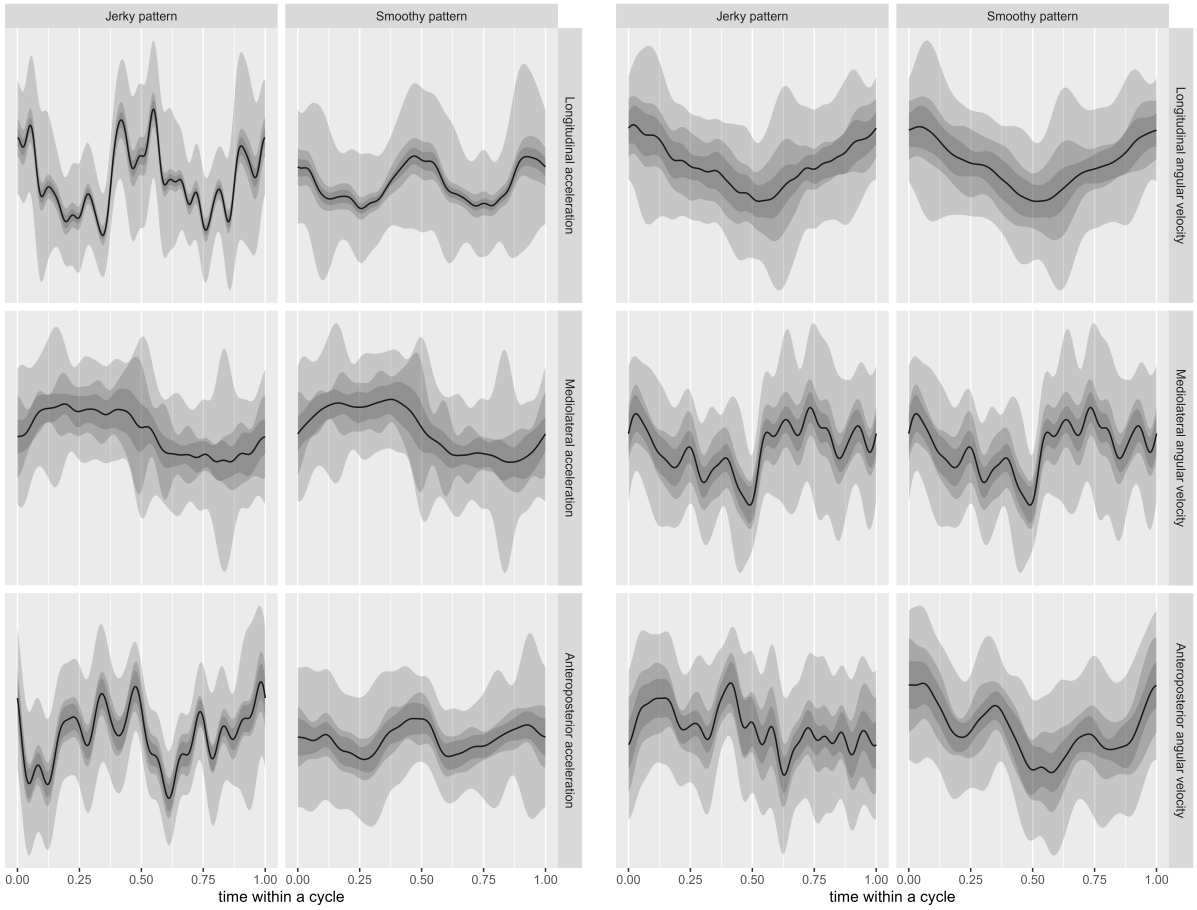

Figure 1: Description of the double partition clustering obtained on the SWIMU data for the six dimensions : columns correspond to the partition defined by the two swimming patterns (*jerky* and *smoothy*) that are represented by their mean curves $\overline{X}_{jk}(t)$ plotted in black plain lines, colors correspond to the partition defined by the three clusters of abilities to reproduce the swimming patterns (dark gray: high repeatability, gray: moderated repeatability, light gray: low repeatability) that are represented by the repeatability regions $[\overline{X}_{jk}(t) - 2\overline{\varepsilon}_{j\ell}(t), \overline{X}_{jk}(t) + 2\overline{\varepsilon}_{j\ell}(t)]$ defined by the dashed lines.

between the swimming pattern and the ability to reproduce it as a discriminant feature of technical skills for front-crawl sprint swimming.

The two partitions clustering allows technical skills to be measured. We now investigate the relation between the technical skills reflected by estimated partitions and performance (see Figure 2). There is a significant effect of gathered partitions on swimming speed as shown by a Fisher test ($F(5.62)=10.1$, $p \leq 0.001$, $\eta^2=0.45$). The *jerky* swimming pattern associated with *low repeatability* is the fastest biomechanical strategy (1.86 ± 0.10 m/s). Indeed, all its pairwise comparisons with others clusters are significant considering a nominal level of 0.05. There are no other significant pairwise comparisons between all kinds of other clusters.

Our approach allows us to discriminate the performance level since there is a clear speed trend for the *jerky+low repeatability* cluster. Thus, it enables to group swimmers of homogeneous technical skills regarding performance. Furthermore this approach allows to perform

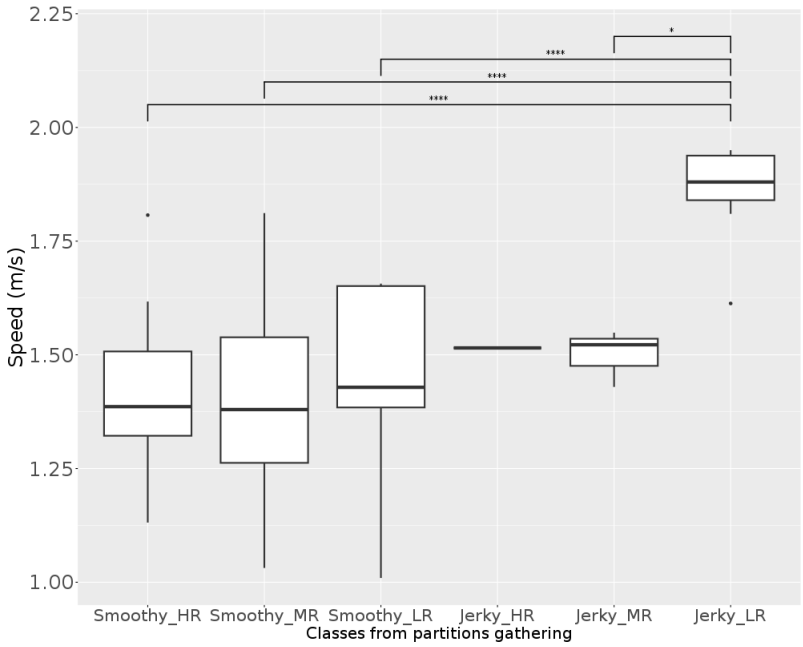

Figure 2: Boxplot of sport performance (*i.e.,* swimming speed) according to gathered partitions from clustering defining swimming pattern and ability to reproduce it. Stars indicate significant pairwise comparisons between classes: * $p \leq 0.05$, **** $p \leq 0.0001$

technical skills evaluation according to stroke cycle kinematics functional modelling on the micro-scale, instead of the classic macro-scale spatio-temporal parameters of literature. In this way, this double partition clustering provides a valuable sports-related outcome.

4 Conclusion

We have developed a model-based approach that provides two complementary partitions allowing technical skills to be tracked in swimming based on multivariate functional data. The model considers both the kinematical information from the original signals and the squared error term resulting from the functional basis decomposition and allows for dependency of the two partitions. Otherwise, the method allows for dimension selection to better establish the biomechanical contribution to technical skills.
The results of the application confirm the double partition model's sensitivity to aggregate kinematical variability defining swimming technical skills. Clustering IMU swimming data highlights specific kinds of stroke kinematics related to speed. In this way it allows us to identify a relevant biomechanical ability for front-crawl sprint performance relying on a jerky unstable pattern. Technical skills are driven by management of kinematical variability through, on the one hand, specific swimming patterns linked to stroke smoothness defining continuity of propulsive actions and energy expenditure, and on the other hand, repeatability defining pattern stability and ability to reproduce swimming strokes.
The development of this procedure expands traditional technique monitoring by avoiding only relying on human-based observations of experts or on macro-scale spatio-temporal parame-

ters recorded by a stopwatch. Hence it gives to coaches a complementary data-driven method less subjective to current eyes-based method. Firstly, to better establish skills evaluation in environmental conditions of daily training and secondly to better characterize technical levels of front-crawl swimmers through an automatic user-friendly framework. The proposed model could be applied to a wide population with different characteristics and leads to biomechanical profiling of swimmers. It provides the first modelling of swimming IMU data in the literature.

References

Bouvet, A., Kolei, S. E., and Marbac, M. (2024). Investigating swimming technical skills by a double partition clustering of multivariate functional data allowing for dimension selection. *Ann. Appl. Statist.*, 18(2):1750–1772.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Fernandes, A., Goethel, M., Marinho, D. A., Mezêncio, B., Vilas-Boas, J. P., and Fernandes, R. J. (2022a). Velocity variability and performance in backstroke in elite and good-level swimmers. *International Journal of Environmental Research and Public Health*, 19(11):6744.

Fernandes, A., Mezêncio, B., Soares, S., Duarte Carvalho, D., Silva, A., Vilas-Boas, J. P., and Fernandes, R. J. (2022b). Intra-and inter-cycle velocity variations in sprint front crawl swimming. *Sports Biomechanics*, pages 1–14.

Leroy, A. (2020). *Multi-task learning models for functional data and application to the prediction of sports performances*. PhD thesis, Université de Paris.

Mallor, F., Leon, T., Gaston, M., and Izquierdo, M. (2010). Changes in power curve shapes as an indicator of fatigue during dynamic contractions. *Journal of biomechanics*, 43(8):1627–1631.

Marbac, M. and Sedki, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27(4):1049–1063.

Seifert, L., De Jesus, K., Komar, J., Ribeiro, J., Abraldes, J., Figueiredo, P., Vilas-Boas, J., and Fernandes, R. (2016). Behavioural variability and motor performance: Effect of practice specialization in front crawl swimming. *Human movement science*, 47:141–150.