Probabilistic $K$-means with Local Alignment for Clustering and Motif Discovery in Functional Data

Marzia A. Cremona$^{a,b,c}$ and Francesca Chiaromonte$^{c,d}$

$^a$Department of Operations and Decision Systems, Université Laval, Québec, Canada; $^b$CHU de Québec – Université Laval Research Center, Québec, Canada; $^c$Department of Statistics, The Pennsylvania State University, University Park, PA; $^d$Institute of Economics and EMbeDS, Sant’Anna School of Advanced Studies, Pisa, Italy

ABSTRACT

We develop a new method to locally cluster curves and discover functional motifs, that is, typical shapes that may recur several times along and across the curves capturing important local characteristics. In order to identify these shared curve portions, our method leverages ideas from functional data analysis (joint clustering and alignment of curves), bioinformatics (local alignment through the extension of high similarity seeds) and fuzzy clustering (curves belonging to more than one cluster, if they contain more than one typical shape). It can employ various dissimilarity measures and incorporate derivatives in the discovery process, thus exploiting complex facets of shapes. We demonstrate the performance of our method with an extensive simulation study, and show how it generalizes other clustering methods for functional data. Finally, we provide real data applications to Italian Covid-19 death curves and Omics data related to mutagenesis. Supplementary materials for this article are available online.

1. Introduction

Given a set of curves, we consider the problem of discovering functional motifs inside them, that is, typical shapes (continuous curve portions) that may recur within each curve, and across several curves in the set (see Figure 1). Some of these motifs may be present in most of the curves, but in different positions. Conversely, other motifs may characterize subgroups of curves and thus differentiate among them based on local shape similarities. We provide a novel method for functional motif discovery that aligns curves locally to identify their shared continuous portions, employing different definitions of (dis)similarity. Importantly, neither the motifs nor their number, lengths, or radii (i.e., the maximum dissimilarity between a motif shape and a portion of curve containing it) need to be known in advance; lengths and radii are specific to each motif.

Our motivating example is the study of mutagenesis, that is the processes that generate mutations in the DNA sequence of an organism. The aim is to identify "signature shapes" in multidimensional curves consisting of multiple types of mutation rates measured at high resolution along the genome (see Section 5.2). Each of these shapes, or motifs, represents a specific mutagenesis pattern, which recurs in a set of genomic regions and is characterized by a certain genomic landscape. Functional motif discovery can also be useful to study protein–DNA interactions—an area in which shape has already been shown to carry biological information (Cremona et al. 2015). For instance, functional motifs in the one-dimensional curves of ChIP-seq signals could help us distinguish different protein–DNA binding. Functional motifs are relevant in many other domains. For example, in finance, one may be interested in detecting recurrent patterns in the time series of asset prices related to different companies. This is an important problem in technical trading analysis, whose rules aim at predicting price changes based on observed patterns. A related problem is the detection of financial bubbles—which are characterized by a rapid surge in prices, not justified by the fundamentals, followed by an unexpected crash—in data such as stock market indices or exchange rates. Outside the “Omics” and financial fields, detecting functional motifs in weather time series may help us understand some aspects of climate change, while detecting functional motifs in time series generated by wearable devices, for example accelerometer data, may help us characterize patterns of physical activity.

During the last two decades, the analysis of curves has received increasing attention and interest in the statistical literature. Indeed, several functional data analysis methods have been developed and applied in many fields (see, e.g., Ramsay and Silverman 2005; Ferraty and Vieu 2006; Horváth and Kokoszka 2012). Several algorithms have been proposed to cluster aligned functional data (reviewed in Jacques and Preda 2014). Since functional data are very often misaligned, algorithms have also been proposed to simultaneously cluster and align curves (Liu and Yang 2009; Sangalli et al. 2010; Park and Ahn 2017). All these methods consider the curves globally, over their entire domain of definition. However, in many applications, separation in groups may occur only on a portion of the domain; this type of clustering structure might be missed by methods that
consider curves in their entirety. The multivariate counterpart of this domain selection problem is usually referred to as feature selection and has been widely studied (see, e.g., Friedman and Meulman 2004; Witten and Tibshirani 2010). In the functional framework, Fraiman, Gimenez, and Svarc (2016) and Floriello and Vitelli (2017) proposed methods to cluster curves while performing feature (i.e., domain) selection. More recently, Vitelli (2019) integrated curve alignment in the sparse clustering procedure.

The problem of functional motif discovery we tackle here is more general and, to the best of our knowledge, it has never been studied in the statistical literature. To identify motifs, we define clusters locally on continuous portions of the misaligned curves and allow each cluster to contain multiple portions of the same curve (i.e., multiple instances of the same functional motif). In addition, we allow each curve to belong to zero, one, or multiple clusters (i.e., to comprise zero, one, or multiple functional motifs). This problem is the continuous version of sequence motif discovery, which is ubiquitous in bioinformatics and "Omics" sciences (see, e.g., Bailey et al. 2006) and consists of searching for highly similar patterns in a set of DNA or protein sequences. While these are discrete sequences of symbols (4 nucleotides, or 20 amino acids), we consider curves that can attain any real values and can be multivariate (i.e., take values in \( \mathbb{R}^d \)). A similar problem for time series has been addressed by the data mining community (Lin et al. 2002; Mueen et al. 2009; Yeh et al. 2016, 2018) defining a motif as a pattern repeated multiple times within a single time series. Available tools generally employ the Euclidean distance or the correlation between portions of the time series. They usually require as input the length and the number of motifs to be found, although Linardi et al. (2018) recently introduced an algorithm that finds all motifs in a given range of lengths. Importantly, these tools require a user-specified minimum distance within which two portions of the time series are considered the same motif (i.e., a motif radius), and this distance is the same across motifs.

We embed the problem of functional motif discovery in a full-blown functional framework, which allows us to capture complex shape characteristics by incorporating derivatives in the discovery process. The functional framework also allows us to rigorously define variability within each motif, and to naturally reduce noise in the curves through smoothing. Our novel method, probabilistic \( K \)-means with local alignment (probKMA), leverages ideas from functional data analysis, bioinformatics, and fuzzy clustering in order to identify \( K \) shared curve portions, which represent \( K \) candidate functional motifs in the set of curves under consideration. Similar to the \( K \)-means with (global) alignment of Sangalli et al. (2010), we simultaneously perform clustering and alignment of curves. However, we employ local alignment in place of their global alignment. Also, similar to BLAST-type algorithms in bioinformatics (Altschul et al. 1990), we perform local alignment through the extension of high similarity seeds. Finally, similar to fuzzy clustering in which points can belong to multiple clusters (Bezdek 1981; Bezdek, Ehrlich, and Full 1984), curves can be associated with zero, one, or more than one cluster (if they contain zero, one, or more than one typical shape).

The article is organized as follows. In Section 2 we present probKMA’s theoretical setting, formulate it as an optimization problem, derive necessary conditions for its solution, and describe its algorithmic implementation. In Section 3 we discuss the evaluation of the clusters produced and identification of the motifs discovered. In Section 4 we provide simulation studies to evaluate probKMA and compare it to other approaches. We present real data applications in Section 5 and provide concluding remarks in Section 6.

2. Probabilistic \( K \)-means with Local Alignment

2.1. Optimization Problem and Necessary Conditions

We consider a set of \( N \) (\( d \)-dimensional) curves \( \mathbf{x}_i : \mathbb{R} \to \mathbb{R}^d \), \( i = 1, \ldots, N \). Our goal is to identify \( K \) (\( d \)-dimensional) cluster
centers \( v_k \)—representing \( K \) candidate motifs—to which the curves are, locally, highly similar with respect to a distance \( d(\cdot, \cdot) \).

Without loss of generality, we can assume that the domain of each \( v_k \) starts at 0; in symbols, \( v_k : (0, c_k) \rightarrow \mathbb{R}^d, k = 1, \ldots, K \), with unknown lengths \( c_1, \ldots, c_K \in [c_{\text{min}}, c_{\text{max}}] \). Then, each curve is aligned to each cluster center \( v_k \) as to minimize their distance in the interval \((0, c_k)\). Alignment is performed composing each curve \( x_i \) with a warping function \( h_{k,i} : \mathbb{R} \rightarrow \mathbb{R} \) from a class \( W \) so that the curve portion matching the cluster center moves to the interval \((0, c_k)\). Here we consider shifts \( W = \{ h : t \mapsto t + s; s \in \mathbb{R} \} \), but our method can be generalized to other warping functions commonly employed in the functional data analysis literature (see, e.g., Chapter 7 of Ramsay and Silverman 2005).

Because of the focus on local similarity, a curve can belong to more than one cluster; that is, different portions of a curve can belong to different portions of curves. Hence, mimicking fuzzy clustering (see, e.g., Bezdek 1981; Bezdek, Ehrlich, and Full 1984), we assign to each curve \( x_i \) a probability \( p_{k,i} \) to be a member of each cluster \( k \). We define a membership function \( p_k : \{x_1, \ldots, x_N\} \rightarrow [0, 1] \) for each \( k = 1, \ldots, K \), with \( p_k(x_i) = p_{k,i} \), requiring that \( \sum_{k=1}^K p_{k,i} = 1 \) for all \( i = 1, \ldots, N \), and that \( \sum_{i=1}^N p_{k,i} > 0 \) for all \( k = 1, \ldots, K \). Each \( p_{k,i} \) corresponds to a particular shift \( s_{k,i} \) of the curve \( x_i \); namely, the one that minimizes the distance between \( x_i \) and \( v_k \) given all constraints. We denote \( S = [s_{k,i}] \in [R^{K\times N}] \) and \( P = [p_{k,i}] \in [0, 1]^{K\times N} \), where \([R^{K\times N}]\) and \([0, 1]^{K\times N}\) indicate the space of matrices of dimension \( K \times N \) with elements in \( R \) and \([0, 1] \), respectively.

Consider the cluster center lengths \( c_1, \ldots, c_K \) as fixed (identification of \( c_k \in [c_{\text{min}}, c_{\text{max}}] \) is discussed in the next section). ProbKMA can be formulated as the following optimization problem: find \( K \) cluster centers \( v_1, \ldots, v_K \), membership probabilities \( P \) and shifts \( S \) that minimize the generalized least-squares functional

\[
J_m(P, S, v_1, \ldots, v_K) = \sum_{i=1}^N \sum_{k=1}^K (p_{k,i})^m d^2(\hat{x}_{i,s_{k,i}}, v_k) \tag{1}
\]

under the constraints \( p_{k,i} \in [0, 1], \forall i, k; \sum_{i=1}^N p_{k,i} = 1, \forall i; \) and \( \sum_{i=1}^N p_{k,i} > 0, \forall k \). Here \( m > 1 \) is a fixed parameter controlling the degree of fuzziness, and \( \hat{x}_{i,s_{k,i}}(t) = x_i(t + s_{k,i}) \) are the shifted curves. Necessary conditions for \( (\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \) to be a (local) minimizer of (1) are that each of \( \hat{P}, \hat{S} \) and \( \hat{v}_1, \ldots, \hat{v}_K \) minimizes (1) fixing all the other variables. We prove two key results (see Section S1). The first provides an explicit solution for \( \hat{P} \) given shifts and centers. Importantly, this result holds for any distance \( d(\cdot, \cdot) \) and does not rely on any regularity assumption on curves or cluster centers.

**Proposition 1.** Fix \( \hat{S} \) and \( \hat{v}_1, \ldots, \hat{v}_K \). Let \( R = \{ i \in \{1, \ldots, N\} \mid d(\hat{x}_{i,s_{k,i}}, \hat{v}_k) > 0 \text{ for all } k \} \) and suppose that \( |R| \geq K \). Then \( \hat{P} = [p_{k,i}] \) is a global minimizer of

\[
J_m(\cdot, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) : [0, 1]^{K\times N} \rightarrow \mathbb{R}, \tag{2}
\]

under the constraints \( \sum_{k=1}^K p_{k,i} = 1, \forall i \) and \( \sum_{i=1}^N p_{k,i} > 0, \forall k \) if and only if

\[
\hat{P}_{k,i} = \left[ \sum_{l=1}^K \left( \frac{d^2(\hat{x}_{i,s_{l,i}}, \hat{v}_l)}{d^2(\hat{x}_{i,s_{l,i}}, \hat{v}_l)} \right) \right]^{-1} \tag{3}
\]

for all \( i \in R \) and

\[
\hat{P}_{k,i} = \begin{cases} 0, & k : d(\hat{x}_{i,s_{k,i}}, \hat{v}_k) > 0 \\ \in [0, 1], & k : d(\hat{x}_{i,s_{k,i}}, \hat{v}_k) = 0 \end{cases} \tag{4}
\]

with \( \sum_{k=1}^K \hat{P}_{k,i} = 1, \forall i \in \mathbb{R} \).

If the \( i \)th curve has positive distance from all cluster centers, (3) states that its probability of belonging to cluster \( k \) is inversely proportional to the \((m-1)\)th root of its squared distance from the \( k \)th cluster center. Equation (4) tackles the extreme case, very seldom in practice, of a curve with distance 0 from one or more cluster centers; in this case, the probabilities are set to 0 for all clusters from which the curve has a positive distance. If a curve has distance 0 to exactly one cluster, the constraint implies that the corresponding probability is 1. If a curve has distance 0 to more than one cluster, the corresponding probabilities can be arbitrarily chosen as long as the constraint is satisfied.

The second result provides a formula for \( \hat{v}_1, \ldots, \hat{v}_K \) given shifts and memberships, and depends on the distance employed. This provides an explicit solution for optimal centers given shifts and probabilities, for any distance \( d_0(\cdot, \cdot) \) defined as

\[
d_0^2(x, v) = \sum_{i=1}^d \frac{d_0^2(x_i, v)}{d_0^2(x_i, v)} \left[ \frac{1 - \alpha}{\int_0^{\infty} \left( x^{(v)}(t) - v^{(v)}(t) \right)^2 dt} \right. \tag{5}
\]

\[
+ \left. \frac{\alpha}{c} \int_0^{\infty} \left( x^{(v)}(t) - v^{(v)}(t) \right)^2 dt \right] \]

where \( w_x > 0 \) is the weight of the \( x \)th component of a \( d \)-dimensional curve, indicated by \( x^{(v)} \), indicates the weak derivative (see, e.g., Evans 1998, p. 254), \( (0, c) \) is the domain of \( v \), and \( \alpha \in [0, 1] \) is a parameter that defines the relative weight of the curve’s levels and derivatives. When \( \alpha = 0 \), we require \( x_i \in L^2(\mathbb{R}, \mathbb{R}^d) \) and \( v_k \in V_k = L^2((0, c_k), \mathbb{R}^d) \), where \( L^2 \) is the space of square-integrable functions. In this case we obtain an \( L^2 \)-like distance \( d_0(\cdot, \cdot) \) that focuses exclusively on the levels. When \( \alpha > 0 \), we require \( x_i \in H^{1}(\mathbb{R}, \mathbb{R}^d) \) and \( v_k \in V_k = H^1((0, c_k), \mathbb{R}^d) \), where \( H^1 \) is the Sobolev space of square-integrable functions, with square-integrable first order weak derivative (see, e.g., Evans 1998, p. 254). The choice of \( \alpha = 1 \) leads to an \( L^2 \)-like pseudo-distance \( d_1(\cdot, \cdot) \) focusing on curve variations (their slopes or trends). Finally, \( \alpha \in (0, 1) \) defines a Sobolev-like distance \( d_\alpha(\cdot, \cdot) \) that highlights more complex features of curve shapes, taking into account both levels and variations. Note that no smoothness assumption is needed for the curves nor the cluster centers.

**Proposition 2.** Fix \( \hat{P} \) and \( \hat{S} \). Consider the distance \( d_\alpha(\cdot, \cdot) \), with \( \alpha \in [0, 1] \). If \( \alpha = 0 \), assume \( x_i \in L^2(\mathbb{R}, \mathbb{R}^d) \) and \( v_k \in V_k = L^2((0, c_k), \mathbb{R}^d) \). If \( \alpha > 0 \), assume \( x_i \in H^{1}(\mathbb{R}, \mathbb{R}^d) \) and \( v_k \in V_k = H^1((0, c_k), \mathbb{R}^d) \). Then \( \hat{v}_1, \ldots, \hat{v}_K \) are the (unique) global minimizers of
\[ J_m(\hat{P}, \hat{S}, v) : V_1 \times \cdots \times V_K \rightarrow \mathbb{R} \]  

if and only if

\[ \hat{v}_k = \frac{\sum_{i=1}^{N} (\hat{p}_{k,i})^\alpha x_{ki}}{\sum_{i=1}^{N} (\hat{p}_{k,i})^\alpha} \quad \text{a.e. in } (0, c_k), \forall k. \]  

When \( \alpha = 1, \) \( \hat{v}_k \) is defined by (7) up to an additive constant.

Equation (7) defines the kth cluster center as a weighted average of the shifted curves in \((0, c_k)\). Weights are determined by memberships: the contribution of a curve to the computation of \( \hat{v}_k \) is directly proportional to its probability of belonging to cluster \( k \). Note that if a curve does not belong to any cluster, then the membership probabilities are all around \( 1/K \), hence, the curve has a reduced influence on the definition of the cluster centers. In addition, we implemented a cluster cleaning step in order to make this influence negligible (see Section S2.1).

2.2. Algorithm

Propositions 1 and 2 suggest to numerically minimize (1) through an iterative procedure that alternates: (i) identification of cluster centers with Equation (7), (ii) curve alignment (warping function selection), and (iii) computation of membership probabilities using Equations (3)–(4). We propose the following algorithm for probKMA.

Initialization Fix the number of clusters \( K \) and the cluster center lengths \( c_1, \ldots, c_K \). Consider an initial membership matrix \( P^{(0)} \) such that \( \sum_{k=1}^{K} P_{k,i}^{(0)} = 1, \forall i \) and \( \sum_{i=1}^{N} P_{k,i}^{(0)} > 0, \forall k \) (non-degenerate clusters), and an initial shift matrix \( S^{(0)} \);

Iteration Repeat the following three steps for \( \text{iter} = 1, 2, \ldots, \)

1. Identification of cluster centers. For each \( k \), compute the \( k \)th cluster center \( \hat{v}_k^{(\text{iter})} \) with Equation (7), using the shift \( \hat{s}_{k,i}^{(\text{iter}-1)} \) and memberships \( P_{k,i}^{(\text{iter}-1)} \);
2. Curve alignment. For each \( i \) and \( k \), align the curve \( x_i \) to the new cluster center \( \hat{v}_k^{(\text{iter})} \), selecting the shift \( \hat{s}_{k,i}^{(\text{iter})} \) that minimizes their distance \( d(\hat{x}_{ki}, \hat{v}_k) \);
3. Computation of membership probabilities. Compute the membership matrix \( P^{(\text{iter})} \) with Equations (3)–(4), using \( v_k^{(\text{iter})} \) and the shifts \( \hat{s}_{k,i}^{(\text{iter})} \).

Stopping criterion At each iteration, evaluate convergence using the Bhattacharyya distance \( BC \) between the membership matrices \( P^{(\text{iter})} \) and \( P^{(\text{iter}-1)} \). For each \( i \), compute

\[ BC_i = -\log \left( \sum_{k=1}^{K} P_{k,i}^{(\text{iter})} \hat{v}_k^{(\text{iter})} \right). \]

Compute \( BC \) as the maximum, mean, or order \( q \) quantile of all \( BC_i \). Repeat steps (i)–(iii) until \( BC \) reaches a given tolerance.

Remark 1. Steps (i) and (iii) are analogous to the steps of a fuzzy K-means algorithm (Bezdek, Ehrlich, and Full 1984), or of an EM algorithm for mixture models (Dempster, Laird, and Rubin 1977). Steps (i) and (ii) correspond to the functional K-means with (global) alignment (Sangalli et al. 2010).

Every iteration can be written in a functional form as

\[ \left( P^{(\text{iter})}, S^{(\text{iter})}, v_1^{(\text{iter})}, \ldots, v_K^{(\text{iter})} \right) \in T_m \left( P^{(\text{iter}-1)}, S^{(\text{iter}-1)}, v_1^{(\text{iter}-1)}, \ldots, v_K^{(\text{iter}-1)} \right) \]

where \( T_m : Y \rightarrow Y \) is the point-to-set map defined by (i)–(iii), and \( Y \) the subset of \([0,1]^{K \times N} \times \mathbb{R}^{K \times N} \times V_1 \times \cdots \times V_K \) that satisfies \( \sum_{k=1}^{K} P_{k,i} = 1, \forall i \) and \( \sum_{i=1}^{N} P_{k,i} > 0, \forall k \). For each initialization, the algorithm generates a sequence of iterations

\[ \left( T_m^{(\text{iter})} \left( P^{(0)}, S^{(0)}, v_1^{(0)}, \ldots, v_K^{(0)} \right) \right)_{\text{iter}=1,2,\ldots}. \]

Below, we show that \( J_m \) is continuous and descends along (8). This is an important result, which mimics the one in Hathaway, Bezdek, and Tucker (1987) for fuzzy K-means (proof in Section S1).

Lemma 3. The functional \( J_m : Y \rightarrow \mathbb{R} \) is continuous.

Theorem 4. Consider \( y^{(\text{iter}-1)} \in Y. \) Then for every \( y^{(\text{iter})} \in T_m(y^{(\text{iter}-1)}) \) we have

\[ J_m(\tilde{y}^{(\text{iter})}) \leq J_m(y^{(\text{iter}-1)}), \]

that is, \( J_m \) is a descent functional for \( T_m \). Moreover, \( J_m \) descends strictly along the iterations if \( y^{(\text{iter}-1)} \notin \Omega \), where \( \Omega \subseteq Y \) is the solution set of \( \tilde{y} = (\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \in Y \) such that

\[ J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \leq J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \quad \forall P \in [0,1]^{K \times N} \sum_{k=1}^{K} P_{k,i} = 1 \quad (10) \]

\[ \sum_{i=1}^{N} P_{k,i} > 0; \quad J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \leq J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \quad \forall S \in \mathbb{R}^{K \times N}; \quad (11) \]

\[ J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \leq J_m(\hat{P}, \hat{S}, \hat{v}_1, \ldots, \hat{v}_K) \quad \forall v_k \in V_k \quad (12) \]

Remark 2. Although the previous result does not guarantee that every sequence of iterations (8) converges to a minimizer of the functional \( J_m \), it is a necessary condition for convergence and a desirable property for the algorithm.

In the previous theoretical results and algorithm, the lengths \( c_1, \ldots, c_K \in [c_{\min}, c_{\max}] \) of the cluster centers remain fixed. However, we seek to identify local similarities even when the lengths of the matching curve portions are not known a priori. This problem has been already tackled by local sequence alignment methods in bioinformatics, whose goal is to find similar stretches of unknown lengths within a collection of nucleotide or amino acid sequences. In this context, one of the most widely used algorithms is BLAST (Altschul et al. 1990).
BLAST starts by finding short stretches shared by the sequences, and uses them as seeds. It then extends the seeds on both sides to construct larger local alignments, stopping when the similarity score drops below a given threshold. Borrowing this logic, we add a center elongation step to our algorithm. This step is performed only when the algorithm is reaching convergence, to guarantee that we do not extend low-quality cluster centers. We attempt elongation on both the left and the right generation, to guarantee that we do not extend low-quality cluster centers. We attempt elongation on both the left and the right (Rousseeuw 1987). Our index is defined for portions of curves silhouette index, similar to the one used in classic clustering. To evaluate a probKMA clustering, we develop a generalized index. The flexibility provided by B-splines. We consider a B-spline basis \{Φ_l\}_l=1^L of order \(n\), with equally spaced knots \(t_1, \ldots, t_{L-n+2}\), and define each one-dimensional curve as \(x(t) = \sum_{l=1}^{L} c_l \Phi_l(t)\), where \(c_l \in \mathbb{R}\). Background coefficients \(c_{bg} \sim \text{Beta}(0.45, 0.45)\), Beta(0.45, 0.45), creating reasonably different backgrounds for both the curve and its derivative. With this flexible model, we can generate data in several scenarios, varying curve and motif lengths, as well as variability, frequencies, and positions of motifs.

Further details on probKMA implementation are provided in Section S2.

3. Cluster Evaluation and Functional Motif Discovery

To evaluate a probKMA clustering, we develop a generalized silhouette index, similar to the one used in classic clustering (Rousseeuw 1987). Our index is defined for portions of curves and measures how well each portion fits its own cluster. First, we dichotomize the membership matrix \(\tilde{\Phi}\) to transform it into a matrix of zeros and ones as explained in Section S2.1, and we extract all the curve portions belonging to a cluster, that is, for which the dichotomized membership probability is equal to 1. Next, we compute the distance \(d_j(k)\) of each portion \(j = 1, \ldots, J\) from cluster \(k\) as the mean of the distances between \(j\) itself and all the portions of cluster \(k\). We define the intra-cluster distance as \(a_j = d_j(k_j)\)—the distance of portion \(j\) from the cluster \(k_j\) it belongs to—and the inter-cluster distance as \(b_j = \min_{k \neq k_j} d_j(k)\)—the minimum distance of portion \(j\) from all the other clusters. The generalized silhouette index for portion \(j\)

\[
S_j = \frac{b_j - a_j}{\max(b_j, a_j)} \in [-1, 1].
\]

Large values of \(S_j\) indicate that \(j\) is appropriately assigned to its cluster, while low values indicate bad assignments. In particular, negative values signify that portion \(j\) is closer to a cluster different from the one it was assigned to. For each cluster \(k\), we then compute its average silhouette index \(S_k\) considering all the portions assigned to \(k\). This measures the compactness of the cluster and hence its quality. Finally, the overall average silhouette index \(S\) measures the overall quality of the clustering. Similar to classic clustering, silhouettes for portions, clusters and overall clustering are visualized in a silhouette plot (see examples in Figures S27(b) and S35(c)) that facilitates their interpretation.

Like other K-means algorithms, probKMA finds a local minimum of the functional \(J_m\) and its output heavily depends on initialization. If the goal is to locally cluster the curves in \(K\) groups, we repeat the algorithm using different initializations (and possibly different initial lengths) and we select the solution with the lowest value of \(J_m\). When \(K\) is not known, the generalized silhouette index allows one to compare the results obtained with different \(K\) and select the best one. If the goal is functional motif discovery, we run probKMA multiple times with different initializations, cluster numbers, and motif initial lengths, and form the set of candidate motifs taking the union of the solutions. We clean this set of candidate motifs using generalized silhouette indices and number of occurrences. We then merge very similar candidate motifs, as they may correspond to the same motif identified by multiple runs of probKMA (Section S3). Finally, we use a motif search algorithm to locate all instances of the discovered motifs in the input curves, that is, all portions of curve with distance lower than a given radius from each motif (Section S3).

4. Simulations

4.1. Generating One-Dimensional Curves in Complex Scenarios

Generating curves comprising functional motifs is a nontrivial task since we require motifs to be smoothly embedded in curves while allowing them to occur with noise. To do this, we exploit the flexibility provided by B-splines. We consider a B-spline basis \(\{\Phi_l\}_{l=1}^L\) of order \(n\), with equally spaced knots \(t_1, \ldots, t_{L-n+2}\), and define each one-dimensional curve as \(x(t) = \sum_{l=1}^{L} c_l \Phi_l(t)\), where \(c_l \in \mathbb{R}\), \(l = 1, \ldots, L\) are coefficients to be chosen. The order \(n\) controls smoothness and complexity of \(x\) (\(x\) is a curve of class \(C^{n-2}\) and a piece-wise polynomial of degree \(n - 1\)). Higher orders provide more degrees of freedom, allowing one to generate curves with more complex shapes, and smoother at the knots. Each \(\Phi_l\) has compact support—it is 0 outside an interval of length \(nT\), where \(T\) is the distance between two subsequent knots; this allows us to define a functional motif of length \(T\) fixing the values of \(n\) coefficients \(c_{m,i}, \ldots, c_{m,i+n-1}\) and repeating them multiple times within the same curve or across different curves. Longer motifs of length \(2T, 3T, \ldots\), that may result in more complex shapes, can be created similarly, fixing the values of \(n + 1, n + 2, \ldots\) subsequent coefficients. Since a single curve can embed more than one functional motif, as well as more than one occurrence of the same motif, we require motifs to be separated by at least one sub-interval \((t_i, t_{i+1})\) as not to be artificially merged (i.e., we require at least \(n\) background coefficients between them). Motif occurrences that are “the same,” both in shape and level, are generated adding Gaussian noise to their coefficients: \(\tilde{c}_{m,j} = c_{m,j} + \epsilon_j, \epsilon_j \overset{iid}{\sim} N(0, \sigma^2)\). Motif occurrences that are “the same” in shape but have different levels are obtained adding a constant \(\delta_m\) to all the coefficients that define a single occurrence (different constants for different occurrences): \(\tilde{c}_{m,j} = c_{m,j} + \delta_m + \epsilon_j, \epsilon_j \overset{iid}{\sim} N(0, \sigma^2)\). Background coefficients \(c_{bg} \in [a, b]\) (i.e., coefficients not corresponding to motifs) are generated as \((c_{bg} - a)/b \overset{iid}{\sim} \text{Beta}(0.45, 0.45)\), creating reasonably different backgrounds for both the curve and its derivative. With this flexible model, we can generate data in several scenarios, varying curve and motif lengths, as well as variability, frequencies, and positions of motifs.
4.2. Functional Motif Discovery: Varying Curve Length $\ell$ and Noise $\sigma$ in Motifs

This simulation study aims to demonstrate the performance of probKMA in discovering functional motifs embedded in a set of curves and to examine the effects of increasing curve length and the noise level comprised in motif occurrences. We consider two different scenarios, with sets of curves embedding (1) motifs that share both shapes and levels; or (2) motifs that share shapes but have different levels.

In scenario (1), we consider a set of 20 curves embedding two functional motifs, each with 12 occurrences (see Figures 2 and S3–S5). In particular, 12 curves contain only one occurrence of a motif (6 curves for each of the two motifs), 4 curves contain two occurrences of a motif (2 curves for each of the two motifs), 2 curves contain one occurrence of each of the two motifs, and 2 curves contain no motif occurrences at all. We generate data using a B-spline basis of order 3, knots at distance 10, and motifs of length 60. Coefficients defining the two motifs are randomly generated from a Beta$(0.45,0.45)$ distribution rescaled to $[-15,15]$. We consider four different curve lengths $\ell = 200, 300, 400, 500$ and four levels of noise $\sigma = 0.1, 0.5, 1, 2$, for a total of 16 simulated datasets. In order to maximize the consistency among these datasets and thus highlight the effects of different $\ell$ and $\sigma$ values, we place motif occurrences within the leftmost sub-interval of length 200 of each curve that is common to all datasets, using the same motif positions and background in all 16 cases. We treat the simulated curves as known, and we sample them on a grid of points at distance 1, so that each motif corresponds to 61 points. For each combination of $\ell$ and $\sigma$, we run our probKMA-based functional motif discovery with Sobolev-like distance $d_{0.5}(\cdot, \cdot)$. We evaluate the number of motifs found, the distance between true and estimated motifs, the estimated lengths of motifs, and the number of true and false positives. ProbKMA is run for $K = 2, 3$, minimum motif lengths $c_{\min} = 40, 50, 60$, and 20 random initializations for each $(K, c_{\min})$ pair (the maximum motif length is set to 70; see Section S4.1 for other parameters). The same initializations are employed for all $\ell$ and $\sigma$ combinations. Results for $\ell = 200$ can be found in Figure 3 and show very good performance for our method. As expected, performance slightly declines when more noise is introduced in the motif instances: some occurrences can be missed, and/or false positives can be included. However, results remain satisfactory even when $\sigma = 2$. Results for other curve lengths are shown in Figures S6–S8. They suggest the same behavior as the noise level increases and they appear rather robust across lengths. The only effect of increasing the ratio between background curve portions and curve portions occupied by motifs is a slight increase in false positives, which occurs exclusively when also the noise level is high.

In scenario (2) we consider the same curves and motifs as in scenario (1), but allow motif occurrences to have different levels (see Figures S9–S12). In particular, a random value $\delta_m \sim U(-10, 10)$ is added to all the coefficients defining each motif.
occurrence (a different value $\delta_n$ for each occurrence). For each combination of $\ell$ and $\sigma$, we run our probKMA-based discovery with the $L^2$-like pseudo-distance $d_1(\cdot, \cdot)$ to focus on curve variation. Parameters are the same as in scenario (1), and detailed results are provided in Figures S13–S16. We find again that our method has good performance, affected (as expected) by the noise level, but not much by the length of the curves. In some cases, especially when the curves are very long, we actually discover motifs that were not embedded in the simulated data. Note that, strictly speaking, these motifs are not altogether false. As one elongates the background portions of the curves, it is possible to generate by chance a few patterns that recur often enough to be identified by our algorithm. In our experiments, these additional motifs are noisier and have fewer occurrences than the two motifs originally embedded in the data.

Additional simulation studies to validate the results described above and examine the robustness of the method to the number of initializations can be found in Section S4.1, in particular in Figures S17–S22.

### 4.3. Comparison with Time Series Motif Discovery

We compare our probKMA-based functional motif discovery to time series motif discovery. In particular, we consider the recent Matrix Profile (Yeh et al. 2016, 2018). This tool discovers the top motif pairs in a time series and, for each of these pairs, provides all the neighboring subseries, that is, all the subseries with distance less than $R$ from the motif pair (see Section S4.2). We consider two specifications of the simulation scenarios introduced in Section 4.2: the simple case of short curves and low noise level ($\ell = 200$ and $\sigma = 0.1$), and the complex case of long curves and high noise level ($\ell = 500$ and $\sigma = 2$). Both probKMA-based motif discovery and Matrix Profile discover the two motifs in the simple case. However, when curves are longer and motifs noisier (complex case), Matrix Profile fails to find Motif 1 and includes many false positives in Motif 2. When the radius is large, it does correctly identify a small number of occurrences of Motif 1, but it also reports a very large number of false positives (for both motifs). On the contrary, probKMA-based motif discovery remains robust to noise level and curve length, and is able to identify both motifs with a very small number of false positives (see Tables S1–S3).

### 4.4. Comparison with Nonsparse and Sparse Functional Clustering

We perform simulation experiments to compare probKMA, meant as a clustering method and separate from its motif discovery purpose, to other functional clustering methods: the standard functional $K$-means (Tarpey and Kinateder 2003), the
K-means with (global) alignment of Sangalli et al. (2010), and
the sparse clustering technique of Floriello and Vitelli (2017)
(see details in Section S4.3). We consider the following scenarios:
(a) curves in the two clusters are aligned and they differ on
the entire domain; (b) curves in the two clusters are misaligned
and they differ on the entire domain; (c) curves in the two clusters
differ on a portion of the domain and this portion is aligned; (d)
curves in the two clusters differ on a portion of the domain and
this portion is misaligned.

Running all methods with Euclidean distance and $K = 2$,
you all correctly classify curves in scenario (a). $K$-means only
works in this scenario, while $K$-means with (global) alignment
performs well in scenarios (a) and (b), and sparse clustering per-
forms well in scenarios (a) and (c). Interestingly, when the noise
level is small, sparse clustering also achieves a good performance
in scenario (b). ProbKMA performs very well in all scenarios
(Tables S4–S5).

5. Real Data Applications

ProbKMA is very flexible and it can be applied to any kind of
functional data, from any domain. As shown in the previous
section, it can be employed not only to discover functional
motifs but also as a (probabilistic) local clustering method. In
this section, we provide two detailed real data applications which
illustrate these two possible uses of probKMA. An additional
application to a well-known dataset in the functional clustering
literature, the Berkeley Growth Study curves, is provided in
Section S5.1.

5.1. Local Clustering of Italian Covid-19 Excess Mortality
Curves

Italy was the first European country to be hit by the Covid-19
pandemic, with the first confirmed cases around mid-February
2020. Italian regions were hit at different times and with different
strength, and local authorities implemented different responses,
especially in the initial stages. Comparing the pandemic evo-
lution across regions can therefore provide important insights
on the role of underlying factors and different containment
measures. We estimate excess mortality due to Covid-19 in Italy
using the mortality data (due to all causes) from the Italian
Institute of Statistics (ISTAT). The dataset contains the daily
number of deaths for 7270 municipalities (covering about 93.5%
of the Italian population) from January 1st to April 30th, for
the years 2015–2020. We aggregate data by region and we com-
pute the excess mortality rate curves as the daily difference
between 2020 deaths and average deaths in the period 2015–
2019, divided by the population of the considered municipalities
(see Section S5.2). In order to focus on the Covid-19 period, we
only consider data starting from February 16th. To reduce noise,
we smooth the curves using B-spline smoothing (cubic splines,
knots at each day, roughness penalty on the second derivative,
and smoothing parameter chosen by average generalized cross-
validation). Smoothed curves are shown in Figure 4(a), while
raw data are in Figure S29.

We cluster the 20 Italian regions according to their excess
mortality rate curves to assess if some regions are sharing similar
patterns (see also Boschi et al. 2021). We are interested in the
entirety of the curves—possibly excluding the extremes of their
domains—but we allow shifts in their alignment to take into
consideration possible differences in the time when the (shared)
patterns began in each region. We employ probKMA as a local
clustering method, with $L^2$–like distance $d_0(\cdot, \cdot)$ and cluster cen-
ters of fixed length $c = 65$ days (hence, allowing for a maximum
shift of 10 days). Figure 4(b)–(c) shows probKMA results for
$K = 2$, when assigning each curve to the cluster with highest
membership probability. Cluster 2 contains the regions (mainly
located in the north of Italy) where Covid-19 hit the hardest.
Lombardia is the region with earliest Covid-19 related deaths,
followed by Emilia Romagna, Marche, Liguria, Piemonte and
Trento/Bolzano, and last Valle d’Aosta (with a delay of 7 days).
Cluster 1 contains the regions with milder epidemic patterns.
Interestingly, Veneto is placed in Cluster 1 despite being the
first region, together with Lombardia, to report Covid-19 cases.
This suggests that Veneto successfully managed to flatten the
curve with its early mass testing and contact tracing response
(Mugnai and Bilato 2020). In contrast, the pattern in Lombardia
is so stark that it does not seem to fit properly even in Cluster 1
and shows a large distance from the cluster center (Figure S30).
Indeed, repeating the clustering with $K = 3$, Lombardia is
placed in a cluster of its own, while the other two clusters and
the alignments within them do not change (see Figure S31).

5.2. Motif Discovery in Mutagenesis Data

To fully illustrate the proposed method in its motif discovery
purpose, we apply it to a mutagenesis dataset adapted from
Kuruppumullage Don et al. (2013) that we provide at https://
github.com/marziacremona/mutagenesis_data. Mutagenesis
comprises all the processes by which mutations are generated
in DNA, it is one of the major evolutionary forces and is central
to causing many human diseases (e.g., cancer). Understanding
mutagenesis and how it is influenced by the genomic landscape
is key to shedding light on genome dynamics (Makova and
Hardison 2015). Kuruppumullage Don et al. (2013) estimated
different types of neutral (i.e., not affected by selection) mutation
rates in nonoverlapping windows along the human genome
comparing it with primates, and employed Hidden Markov
Models to define six divergence states and segment the genome
accordingly. One of the states is of particular interest: it
comprises hot regions with very high rates for substitutions,
small insertions, and small deletions, which are associated with
high GC (guanine-cytosine) content, early replication timing,
and open chromatin. Since these results were obtained at a
rather large scale (1-Mb windows), investigating rates at a finer
resolution within the hot regions may reveal more specific
trends and patterns of variation. Note that, in general, we
could aim at discovering three-dimensional motifs, that is, joint
patterns of substitutions, insertions, and deletions. However,
for simplicity, we consider only one-dimensional substitution
rate curves. Estimating high-resolution substitution rates in 1-
kb windows within each hot region (with the same pipeline
as in Kuruppumullage Don et al. 2013, see Section S5.3 and
Figure S32), we generate a dataset of 43 curves, varying in length
from 1 Mb (corresponding to a grid of 1000 points) to 22 Mb
(22,000 points). The curves are very noisy and contain several
missing or inaccurate values since in many 1-kb windows the
Figure 4. Covid-19 excess mortality curves and probKMA results. (a) Smoothed curves. Vertical solid lines represent national lockdown (March 9th) and closure of all non-essential economic activities (March 23rd); (b) Cluster centers (thick curves) with aligned portions of curves; (c) Alignment between portions of curves within clusters (start day of each portion).

information needed to estimate rates is scarce (see Figure S33). In particular, substitution rates can be reliably estimated only in 60% of the 1-kb windows. After pre-processing with stochastic regression imputation and local smoothing, missing values are reduced to 17% of the windows (see Section S5.3).

We employ our probKMA-based functional motif discovery on the 43 curves using the Sobolev-like distance $\tilde{d}_{0.5}(\cdot, \cdot)$ (the generalized version which can accommodate large gaps; see Section S2.2). We look for motifs with minimum lengths $c_{\min} = 40, 50, 60, 70$ (maximum length $c_{\max} = 150$), and we run probKMA for $K = 2, 3, 4, 5$ using 10 random initialization for each $(K, c_{\min})$ pair. We employ our generalized silhouette index to evaluate each probKMA run and to filter the set of candidate motifs, and we select motif-specific radii based on probKMA results (see details in Section S5.3). We identify 13 functional motifs that differ substantially in length (40–104 kb), levels and shapes (see Figure 5(a)). The motifs also differ in frequency (i.e., number of occurrences in the data) and level of variability (see Table 1). This highlights the advantage of employing a motif discovery methodology able to learn motif-specific length, frequency, and variability from the data.

At least four of the motifs found are of biological interest: Motif 12 corresponds to eight long sub-regions (about 100 kb) with extremely high substitution rate (an elevation of 10%–20% relative to the mean level across all hot regions, which is already elevated in comparison to the genome at large). Motif 4 and
Figure 5. ProbKMA-based functional motif discovery in substitution rate curves. (a) Motifs found, plotted as percent changes with respect to the mean substitution rate across all hot regions; (b) Genomic landscape of the motifs, with color intensity proportional to the significance ($-\log_{10}(p)$) of a mean difference two-sided test contrasting motif occurrences and hot regions at large; red, blue, and white represent positive, negative and nonsignificant ($p > 0.1$) differences, respectively. Rimmed columns and arrows show four particularly interesting motifs.

Table 1. ProbKMA-based functional motif discovery in substitution rate curves. For each motif found, we report the number of occurrences and their mean distance from the motif.

| Motif | Number | Mean dist |
|-------|--------|-----------|
| 1     | 19     | 1.9       |
| 2     | 12     | 1.9       |
| 3     | 27     | 3.5       |
| 4     | 37     | 5.1       |
| 5     | 63     | 6.5       |
| 6     | 12     | 3.0       |
| 7     | 72     | 8.7       |
| 8     | 47     | 7.4       |
| 9     | 14     | 5.2       |
| 10    | 11     | 3.0       |
| 11    | 9      | 5.5       |
| 12    | 8      | 18.5      |
| 13    | 6      | 17.0      |

Motif 8 also present very high substitution rate and opposite patterns. In Motif 4, rate is about 10% above the overall hot regions mean for the initial ~20 kb, and then decrease. In Motif 8 rate increases and then stabilizes at about 10% above the mean for ~20 kb. The two motifs have similar variability and are both very frequent (37 and 47 occurrences, respectively). Finally, Motif 13 corresponds to six long sub-regions with a substitution rate 20%–30% below the mean. These portions of the hot regions are in fact not hot; substitutions rate is similar to that of the rest of the genome. To investigate the genomic landscape of the motifs found, we consider a set of 35 genomic features measured in each of the 1-kb windows constituting the hot regions. These features represent biological contexts that have an interplay with mutagenesis, such as DNA conformation, DNA sequence, replication, recombination, chromatin openness and modifications (see Table S6). We then compare, independently for each genomic feature and each motif, the mean of the measurements in motif occurrences with the mean across all hot regions. We perform a simulation-based two-sided test for mean difference, where the empirical null distribution is obtained from 1000 datasets generated by randomly relocating motif occurrences within the set of curves. Figure 5(b) shows that each motif has a characteristic genomic landscape, which helps in its biological interpretation. For example, occurrences of Motif 13 are enriched in exons and conserved elements compared to hot regions in general; their lowered substitution rate may correlate with such enrichments.

6. Discussion

This article, for the first time to the best of our knowledge, tackles the problem of functional motif discovery from a statistical perspective. We proposed probKMA for discovering candidate motifs in a set of curves, incorporating ideas from functional data analysis, bioinformatics and fuzzy clustering. In addition, we proposed a generalized silhouette index to evaluate probKMA results, and implemented a post-processing for merging candidate motifs and searching motif occurrences along the curves. Although many alternative strategies can be employed in post-processing, each with pros and cons, results on simulated and real data suggest that our implementation is effective in a range of scenarios.

ProbKMA employs a flexible definition of curve similarity, which incorporates both levels and derivatives. In addition, similarity is defined locally, in a way that tolerates large gaps in the curves. This broadens the application scope of our methodology. ProbKMA can also be applied to multivariate curves and does not require the user to specify the exact motif lengths or the motif variability levels at the outset. These are learned from the data—the user only needs to specify the minimum and the maximum lengths of the motifs to discover—substantially improving performance with respect to approaches where lengths and/or radii are fixed. Real data applications usually require some pre-processing steps—such as smoothing to estimate the curves from discretely observed data—which might artificially introduce “false” motifs in the curves (see, e.g., Section S5.3). As a consequence, the user must carefully select minimum motif lengths which are compatible with the pre-processing, in particular with the choice of smoothing parameters. The mini-
mum motif length ought to be larger than the length of potential artificially-introduced motifs which, intuitively, is larger the more smoothing has been applied to the data.

In our experience, motif discovery with probKMA can fail when motifs are too similar to one another or when they are too similar to background portions of the curves. This can happen by chance when motifs are very noisy. Relatedly, simulations show that, when motifs are very noisy and/or dispersed in very long curves, our method can identify motifs that were not intentionally introduced in the data, but rather randomly created when generating background portions of the curves. In a way, these additional motifs may be considered as unintentional and yet true (as opposed to false) positives; they do recur in the curves in a way that is detectable by the algorithm. Nevertheless, in our simulations they are noisier and have fewer occurrences. This observation underscores the need for further work addressing the statistical significance of the motifs. The flexible model that we introduced to generate simulation data may play an important role in this context, providing a way to estimate the likelihood of discovering motifs in background curves.

We used a deliberately general, data-driven and nonparametric notion of functional motif—in line with those used in, for example, the bioinformatics and data mining literature (Lin et al. 2002; Bailey et al. 2006). Albeit beyond the scope of this article, it would be of utmost interest to formulate a parametric definition of functional motif and develop a rigorous statistical theory for its estimation.

Separately from its motif discovery purpose, probKMA can also be employed for probabilistic clustering of misaligned functional data based on local similarities. In this respect, it also represents a generalization of sparse clustering procedures recently proposed in functional data analysis (Fraiman, Gimenez, and Svarc 2016; Floriello and Vitelli 2017). In the limit, when the minimum motif length is close to the length of the curves under consideration, probKMA becomes a probabilistic version of K-means with (global) alignment (Sangalli et al. 2010).

Supplementary Materials

Supplementary material includes proofs, additional methods and results. An R implementation (with examples) is available at https://github.com/mariacremona/ProbKMA-FMD.

Acknowledgments

We thank Matthew Reinherr and Piercesare Secchi for discussions about functional data methodology; Katelyna D. Makova and Di (Bruce) Chen for help with the mutagenesis application; Valeria Vitelli and Davide Floriello for their sparse functional clustering code.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was partially funded by the Eberly College of Science, the Institute for Computational and Data Sciences and the Huck Institutes of the Life Sciences (Penn State University); NSF award DMS-1407639; and Tobacco Settlement and CURE funds of the PA Department of Health. M.A. Cremona acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

ORCID

Marzia A. Cremona https://orcid.org/0000-0001-8899-7316

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990), "Basic Local Alignment Search Tool," Journal of Molecular Biology, 215, 403–410. [1120,1122]
Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006), "MEME: Discovering and Analyzing DNA and Protein Sequence Motifs," Nucleic Acids Research, 34, W369–W373. [1120,1129]
Bezdek, J. C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, New York: Springer. [1120,1121]
Bezdek, J. C., Ehrlich, R., and Full, W. (1984), "FCM: The Fuzzy c-means Clustering Algorithm," Computers & Geosciences, 10, 191–203. [1120,1121,1122]
Boschi, T., Di Iorio, J., Testa, L., Cremona, M. A., and Chiaromonte, F. (2021), "Functional Data Analysis Characterizes the Shapes of the First COVID-19 Epidemic Wave in Italy," Scientific Reports, 11. [1126]
Cremona, M. A., Sangalli, L. M., Vantini, S., Dellino, G. I., Pelici, P. G., Secchi, P., and Riva, L. (2015), "Peak Shape Clustering Reveals Biological Insights," BMC Bioinformatics, 16. [1119]
Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, 39, 1–38. [1122]
Evans, L. C. (1998). Partial Differential Equations, Providence, RI: American Mathematical Society. [1121]
Ferraty, F., and Vieu, P. (2006), Nonparametric Functional Data Analysis: Theory and Practice, New York: Springer. [1119]
Florio, P., and Vieu, P. (2017), "Sparse Clustering of Functional Data," Journal of Multivariate Analysis, 154, 1–18. [1120,1126,1129]
Fraiman, R., Gimenez, Y., and Svarc, M. (2016), "Feature Selection for Functional Data," Journal of Multivariate Analysis, 146, 191–208. [1120,1129]
Friedman, J. H., and Meulman, J. J. (2004), "Clustering Objects on Subsets of Attributes," Journal of the Royal Statistical Society, Series B, 66, 815–849. [1120]
Hathaway, R., Bezdek, J., and Tucker, W. (1987), "An Improved Convergence Theory for the Fuzzy Isodata Clustering Algorithms," Analysis of Fuzzy Information, 3, 123–132. [1122]
Horváth, L., and Kokoszka, P. (2012), Inference for Functional Data with Applications (Vol. 200), New York: Springer. [1119]
Jacques, J., and Preda, C. (2014), "Functional Data Clustering: A Survey," Advances in Data Analysis and Classification, 8, 231–255. [1119]
Kuruppumullage Don, P., Ananda, G., Chiaromonte, F., and Makova, K. D. (2013), "Segmenting the Human Genome based on States of Neutral Genetic Divergence," Proceedings of the National Academy of Sciences, 110, 14699–14704. [1126]
Lin, J., Keogh, E., Lonardi, S., and Patel, P. (2002), "Finding Motifs in Time Series," in 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada. [1120,1129]
Linard, M., Zhu, Y., Palpanas, T., and Keogh, E. (2018), "Matrix Profile X: VALMODO - Scalable Discovery of Variable-Length Motifs in Data Series," in ACM SIGMOD/PODS International Conference on Management of Data / Principles of Database Systems, Houston, Texas, USA. [1120]
Liu, X., and Yang, M. C. (2009), "Simultaneous Curve Registration and Clustering for Functional Data," Computational Statistics & Data Analysis, 53, 1361–1376. [1119]
Makova, K. D., and Hardison, R. C. (2015), "The Effects of Chromatin Organization on Variation in Mutation Rates in the Genome," Nature Reviews Genetics, 16, 213–223. [1126]
Mueen, A., Keogh, E., Zhu, Q., Cash, S., and Westover, B. (2009), "Exact Discovery of Time Series Motifs," in SIAM International Conference on Data Mining, Sparks, Nevada, USA. [1120]
Mugnai, G., and Bilatom C. (2020), “Covid-19 in Italy: Lesson from the Veneto Region,” *European Journal of Internal Medicine*, 77, 161–162. [1126]

Park, J., and Ahn, J. (2017), “Clustering Multivariate Functional Data with Phase Variation,” *Biometrics*, 73, 324–333. [1119]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, (2th ed.), New York: Springer. [1119,1121]

Rousseeuw, P. J. (1987), “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” *Journal of Computational and Applied Mathematics*, 20, 53–65. [1123]

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010), “K-mean Alignment for Curve Clustering,” *Computational Statistics & Data Analysis*, 54, 1219–1233. [1119,1120,1122,1126,1129]

Tarpey, T., and Kinateder, K. K. (2003), “Clustering Functional Data,” *Journal of Classification*, 20, 93–114. [1125]

Vitelli, V. (2019), “A Novel Framework for Joint Sparse Clustering and Alignment of Functional Data,” arXiv, 1912.00687. [1120]

Witten, D. M., and Tibshirani, R. (2010), “A Framework for Feature Selection in Clustering,” *Journal of the American Statistical Association*, 105, 713–726. [1120]

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Zimmerman, Z., Silva, D. F., Mueen, A., and Keogh, E. (2018), “Time Series Joins, Motifs, Discords and Shapelets: A Unifying View that Exploits the Matrix Profile,” *Data Mining and Knowledge Discovery*, 32, 83–123. [1120,1125]

Yeh, C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., and Keogh, E. (2016), “Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets,” in *IEEE 16th International Conference on Data Mining, Barcelona, Spain*. [1120,1125]