

Phase and amplitude-based clustering for functional data

Gerda Claeskens^{1,*}, Mia Hubert^b, Leen Slaets¹

^a*ORSTAT and Leuven Statistics Research Center, K.U.Leuven, Naamsestraat 69, 3000 Leuven, Belgium*

^b*Department of Mathematics and Leuven Statistics Research Center, K.U.Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium*

Abstract

Functional data that are not perfectly aligned in the sense of not showing peaks and valleys at the precise same locations possess phase variation. This is commonly addressed by preprocessing the data via a warping procedure. As opposed to treating phase variation as a nuisance effect, it is advantageous to recognize it as a possible important source of information for clustering. It is illustrated how results from a multiresolution warping procedure can be used for clustering. This approach allows to address detailed questions to find local clusters that differ in phase, or clusters that differ in amplitude, or both simultaneously.

Keywords: Functional data, clustering, phase variation, amplitude variation, warping.

1. Introduction

Functional data analysis studies data structures which are believed to be generated by underlying (smooth) functions. We consider samples of curves. Examples include growth curves in biology (Gasser and Kneip, 1995) and market penetration in economics (Sood et al., 2009).

Although there is an overlap between the analysis of curve data and the longitudinal data framework (Hall et al., 2006), functional data focusses more on studying variation in a sample of complex patterns, with several extremes and local amplitude variation (variation in the response values), which would call for complicated structures for the random effects in a longitudinal data analysis. Typical for functional data is phase variation, or misalignment of the curves, that is, not all features of the curves occur at the same locations, which would, for example, make a pointwise average of curve values become useless. Functional data analysis has devoted a great deal of attention to this phenomenon known as registration (Ramsay

*Corresponding author. Phone +32-16-326993, fax +32-16-326624.

Email addresses: Gerda.Claeskens@econ.kuleuven.be (Gerda Claeskens), Mia.Hubert@wis.kuleuven.be (Mia Hubert), Leen.Slaets@econ.kuleuven.be (Leen Slaets)

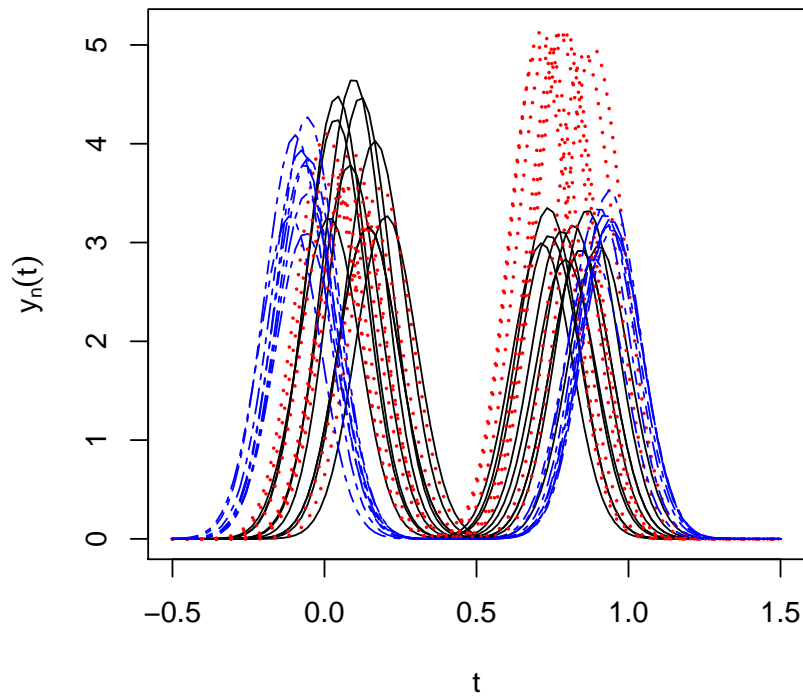


Figure 1: Illustrative data set with three clusters. The 8 (blue) dashed lines represent curves for which the distance between the peaks is larger than for the other curves. The 8 (red) dotted lines represent curves that possess a higher second peak.

and Li, 1998), time warping in engineering (Rabiner et al., 1978) and curve alignment (Wang and Gasser, 1997).

As opposed to treating phase variation as a nuisance effect and to ignoring in further analysis that it even took place, we explicitly recognize it as a source of information for clustering. The multiresolution warping method of Claeskens et al. (2010); Slaets et al. (2010) summarizes the data in a relatively small number of interpretable phase and amplitude components. These components are represented by a vector, one for each curve. We use the well-studied multivariate partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990) to cluster this multivariate data object.

We explain the new method via a simulated sample of curves, see Figure 1, that is generated according to the simulation setting of Section 4.1. This sample of 25 curves consist of three clusters. There is one group of five curves for which the distance between the two peaks is larger than for the other curves. A second group is formed by the six curves

for which the height of the second peak is larger than for the other curves. The remaining six curves form the third cluster. Precise details about how this dataset is generated can be found in Section 4. Clearly, the curves are not aligned and show variation both in phase (horizontal) and in amplitude (vertical).

Clustering functional data received quite some attention already, e.g. in the regression-mixtures framework (DeSarbo and Cron, 1988). The majority of the existing methods does not take phase variation into account and hence assumes that its presence is only limited, or else those works operate on the warped data under the assumption that the preprocessing warping stage contains no cluster information. Examples of such approaches include a functional version of k -means clustering by using functional principal components in Chiou and Li (2007), k -means clustering on fitted B-spline coefficients (Abraham et al., 2003), a robustification thereof (Garcia-Escudero and Gordaliza, 2005) and a flexible clustering model especially for sparsely sampled data (James and Sugar, 2003). In Lopez-Pintado and Romo (2005) the notion of functional depth is used as a way to robustly classify functional data. More recently, timing differences across subjects are acknowledged, but seen as a nuisance effect, rather than as a source of information. Liu and Yang (2009) incorporate shifted B-splines in their clustering model to account for phase variation within clusters, but do not use the information contained in the estimated shift coefficients in the clustering procedure. Sangelli et al. (2010) present a k -means type clustering procedure based on a similarity index between two curves, in which they optimize the similarity between strictly increasing affine transformed (warped) curves within each cluster.

Ignoring phase variation for clustering may result in a possible loss of information. James (2007) illustrates this using a weighted warping function together with the warped curves for clustering coefficients of functional principal components. The problem of clustering in the presence of complex phase variation, however, has not yet been studied.

While detecting curves with distinct functional shapes, e.g. some of the curves are missing a peak, or being of completely different form, is relatively easy, even by eye, we focus our research on samples of curves that look quite alike at a first glance, though, still contain distinct groups of curves. Even within this more difficult setting, we are able to separate clusters based on information on phase and amplitude variation. Moreover, our method provides valuable information to the user by identifying the reason (phase or amplitude or both) that clusters were formed.

The new method of this paper, the explicit incorporation of the warping function and estimated amplitude coefficients in a clustering approach is evaluated and compared with the methods by Chiou and Li (2007) and Liu and Yang (2009) in a simulation study in Section 4. A data example concerning growth curves is included in Section 5.

2. Clustering via multiresolution warping

2.1. Warplets

Multiresolution warping (Claeskens et al., 2010) uses warplets as building blocks. The warplets (see below for a definition) are local warping functions that concentrate the warping action to a certain domain and have a clear interpretation of the location and the intensity of the warp. The final warping function consists of using function composition to combine different warplets.

Warplets are strictly monotone increasing functions that deviate from the identity function in a smooth manner on the interval $[a - r_1, a + r_2] = [w_l, w_u]$. The following definition corresponds to Def. 2.2 of Claeskens et al. (2010) for asymmetric warplets.

$$\begin{aligned} \tilde{\tau}(a, \lambda, w_l, w_u; t) &= \tilde{\tau}(a, \lambda, a - r_1, a + r_2; t) \\ &= \begin{cases} a + r_1 \cdot g\left(\lambda \frac{r}{r_1}; (t - a)/r_1\right), & t \in [a - r_1, a - \frac{3\sqrt{3}}{8}\lambda r] \\ a + r_2 \cdot g\left(\lambda \frac{r}{r_2}; (t - a)/r_2\right), & t \in [a - \frac{3\sqrt{3}}{8}\lambda r, a + r_2] \\ t, & \text{otherwise,} \end{cases} \end{aligned} \quad (1)$$

with $r_1, r_2 > 0$, $r = \min(r_1, r_2)$, $\lambda \in (-1, 1)$, $g(\lambda; y) = z + \lambda K(z)$ in which z is the solution to $z - \lambda K(z) = y$, and with the quartic kernel K :

$$K(z) = \begin{cases} \frac{3\sqrt{3}}{8}(1 - z^2)^2, & z \in [-1, 1] \\ 0, & \text{otherwise.} \end{cases}$$

When $\lambda > 0$ the warplet will cause a dilation of the time points in the interval $[a - r_1, a - \frac{3\sqrt{3}}{8}\lambda r]$, followed by a compression of the time values in the interval $[a - \frac{3\sqrt{3}}{8}\lambda r, a + r_2]$, with an intensity determined by the value of λ . When $\lambda < 0$ the curve is compressed for time values in the interval $[a - r_1, a - \frac{3\sqrt{3}}{8}\lambda r]$, and dilated on the interval $[a - \frac{3\sqrt{3}}{8}\lambda r, a + r_2]$.

For each curve n ($n = 1, \dots, N$), the warping function consists of a function composition of warplets: $\tilde{\tau}_{n,q}$ ($q = 1, \dots, Q$) are composed in a warping function

$$\tau_n = \tilde{\tau}_{n,Q} \circ \dots \circ \tilde{\tau}_{n,2} \circ \tilde{\tau}_{n,1}.$$

Since each warplet is monotone increasing, the same property holds for the composition. Moreover, the inverse of a warplet is explicit to obtain by changing the sign of λ , that is, $\tilde{\tau}^{-1} = \tilde{\tau}(a, -\lambda, w_l, w_u; t)$, leading to the attractive property that the inverse transformation has an easy and explicit formula,

$$\tau_n^{-1} = \tilde{\tau}_{n,1}^{-1} \circ \dots \circ \tilde{\tau}_{n,Q-1}^{-1} \circ \tilde{\tau}_{n,Q}^{-1}.$$

2.2. The model for multiresolution warping

The observed function values y are noisy observations of an underlying noise-free curve F . This curve consists of a common mean function μ with added local amplitude, which can be warped through compositions of warplets, as denoted more precisely in the multiresolution warping model:

$$y_{n,j} = F_{n,j} + e_{n,j} = \mu(\tau_n(t_j + w_{\text{shift},n})) + \sum_{k=1}^K b_{n,k} \psi_k(\tau_n(t_j + w_{\text{shift},n})) + e_{n,j}. \quad (2)$$

The model incorporates for each curve a horizontal shift parameter $w_{\text{shift},n}$, which performs a global warping action prior to applying the local warplets. This is useful since the warplets are designed to model local phase variation.

The random variables $b_{n,k}$ and $e_{n,j}$ are independent realizations of respectively $\mathcal{N}(0, \sigma_k^2)$ and $\mathcal{N}(0, \sigma^2)$ for $n = 1, \dots, N$, $j = 1, \dots, T$, $k = 1, \dots, K$.

The local amplitude differences are modeled by a set of asymmetric rescaled quartic kernels ψ_k , defined as

$$\psi_k(\bar{a}_k, a_{l,k}, a_{u,k}; t) = \begin{cases} \left(1 - \left(\frac{t - \bar{a}_k}{a_{u,k} - \bar{a}_k}\right)^2\right)^2, & \bar{a}_k \leq t \leq a_{u,k} \\ \left(1 - \left(\frac{t - \bar{a}_k}{\bar{a}_k - a_{l,k}}\right)^2\right)^2, & a_{l,k} \leq t \leq \bar{a}_k. \end{cases}$$

of which the locations (lower bound $a_{l,k} <$ center $\bar{a}_k <$ upper bound $a_{u,k}$) are estimated from the data. The number of kernels K has to be provided by the user. For successive kernels we make sure that $\bar{a}_{k+1} > a_{u,k}$ and $a_{l,k+1} > \bar{a}_k$. The curve-specific kernel coefficients $b_{n,k}$ are included in the model as random effects (see also Gervini and Gasser, 2005, for using B-splines instead of kernels). The curve-specific warping functions τ_n are modeled as follows,

$$\tau_n(t_j) = \tilde{\tau}(a_Q, \lambda_{n,Q}, w_{l,Q}, w_{u,Q}) \circ \dots \circ \tilde{\tau}(a_1, \lambda_{n,1}, w_{l,1}, w_{u,1})(t_j),$$

such that the intensities λ are the only curve-specific warplet parameters. For each component they are directly comparable across curves. As an averaging constraint, the intensity parameters satisfy that

$$\lambda_{N,q} = - \sum_{n=1}^{N-1} \lambda_{n,q} \text{ for } q = 1, \dots, Q. \quad (3)$$

The multiresolution warping approach of Claeskens et al. (2010) uses model (2) to align the data by minimizing the squared error between the warped responses (without shifts), while allowing for amplitude differences by means of a limited number of local variability areas. Model (2), see also Slaets et al. (2010), includes horizontal shifts and the kernel

parameters are estimated simultaneous with the warping functions. A Bayesian Markov chain Monte Carlo (MCMC) estimation procedure is implemented to estimate model (2). This routine is programmed in the R library MRwarping and executes the time-consuming MCMC part in C. The function MRwarp in this library estimates model (2). We refer to Slaets et al. (2010) for details about the implementation and the MRwarping library. While the algorithm of Slaets et al. (2010) fits such models treating the amplitude variation as nuisance effects, we here need predictions for the amplitude coefficients $b_{n,k}$ to cluster the curves. Hence we now estimate the parameters of the warping functions τ_n as compositions of individual warplets, the parameters of the kernel functions ψ_k , as well as the coefficients $b_{n,k}$ of the kernels.

An iterative estimation procedure is proposed to obtain estimates of the kernel coefficients, after the estimation of model (2). Denote $(t, \tilde{y}_n(t))$ for the warped curves $(\tau_n(t), y_n(t))$, where

$$\tilde{y}_n(t_i) = \mu(t_i) + \sum_{k=1}^K b_{n,k} \psi_k(t_i) + e_{n,i},$$

which are obtained by means of penalized spline smoothing of the warped data points $(\tau_n(t_j), y_n(t_j))$, using the **SemiPar** package in R (Wand et al., 2005). This procedure uses the mixed model representation of penalized spline models to estimate a smooth curve to the data with simultaneous estimation of the smoothing parameter and is hence completely automatic. The iterative estimation of mean and kernel coefficients goes as follows. For 1000 equally spaced time points $t_i, i = 1, \dots, 1000$ on $[t_1, t_T]$:

1. Initialization: $\hat{\mu}(t_i) = \frac{1}{N} \sum_{n=1}^N \tilde{y}_n(t_i)$.
2. Estimate $b_{n,k}$ by using linear regression on the following models for each of the curves separately $\tilde{y}_n(t_i) = \hat{\mu}(t_i) + \sum_{k=1}^K b_{n,k} \psi_k(t_i) + e_{n,i}$.
3. Use $\hat{b}_{n,k}$ in 1. to update $\hat{\mu}(t_i) = 1/N \sum_{n=1}^N \left(\tilde{y}_n(t_i) - \sum_{k=1}^K \hat{b}_{n,k} \psi_k(t_i) \right)$.
4. Repeat 2. and 3. until the maximum absolute differences between the successive estimates are smaller than 1%, or a maximum of e.g. 50 iterations have been performed.

Once the model coefficients are estimated, these are used to cluster the curves.

2.3. Partitioning around medoids for clustering

An advantage of our method is that it reduces a sample of curves to a vector of coefficients for each curve. When this vector is used for clustering, standard multivariate cluster methods are applicable. We use the multivariate partitioning around medoids (PAM) clustering technique, as described in Kaufman and Rousseeuw (1990). We applied PAM with the Euclidian distance as dissimilarity throughout this paper.

To find M clusters, the PAM algorithm starts by choosing M subjects as cluster centers (medoids) in an initialization step. It then assigns other subjects to a cluster by minimizing the sum of the dissimilarity between each subject and the medoid. In the next step new medoids are calculated, these are the cluster subjects which minimize the average dissimilarity to its cluster members. Implementations of PAM are available in the R-package `cluster`, in S-PLUS (Struyf et al., 1997) and MATLAB (Verboven and Hubert, 2005). PAM is a popular approach for multivariate clustering, as it has several advantages over classical k -means. As it is defined through the dissimilarities only, it can be applied to any type of data by choosing an appropriate dissimilarity measure. The k -means algorithm on the other hand only uses Euclidean distances. PAM also yields the medoids, which are representative objects for each cluster, whereas the centroids of the k -means method are not observed data points. The algorithm does not depend on the order of the objects and can be easily extended to large data sets.

Another interesting feature of PAM is its L_1 -type of objective function (summing the individual dissimilarities instead of taking sums of squares), which makes it less sensitive to outlying observations. This robustness property of PAM has for example been exploited in Barbieri et al. (2001), Bozinov and Rahnenführer (2002), Liu et al. (2003) and Douzal-Chouakria et al. (2009). In section 3.7 we illustrate the advantage that PAM can have over k -means by means of a dataset with an outlying curve. As pointed out by García-Escudero and Gordaliza (1999) and Hennig (2008), even more robust approaches can be obtained by trimming, such as the trimmed k -means method of Cuesta-Albertos et al. (1997) or the trimmed determinant criterion of Gallegos and Ritter (2005). See García-Escudero et al. (2010) for a nice overview. These methods could be considered in our procedure as well.

We note that, despite the usage of PAM, we do not consider multiresolution clustering to be a robust procedure. The multiresolution warping model (2) is not robust, neither is the estimation of the kernel coefficients.

3. Illustrative example

We now return to the illustrative example presented in Figure 1 in the introduction and show step by step how the method works.

3.1. Shifts and warplet intensities

Figure 2 shows the estimated shifts (a) and the set of horizontally shifted curves (b). These shifts form the first column in the multivariate data object, see Table 1, that will be used to cluster the data.

As explained in Section 2.2 each component of the warping function only differs across the curves in the intensity via the parameter λ . This can be clearly observed in the esti-

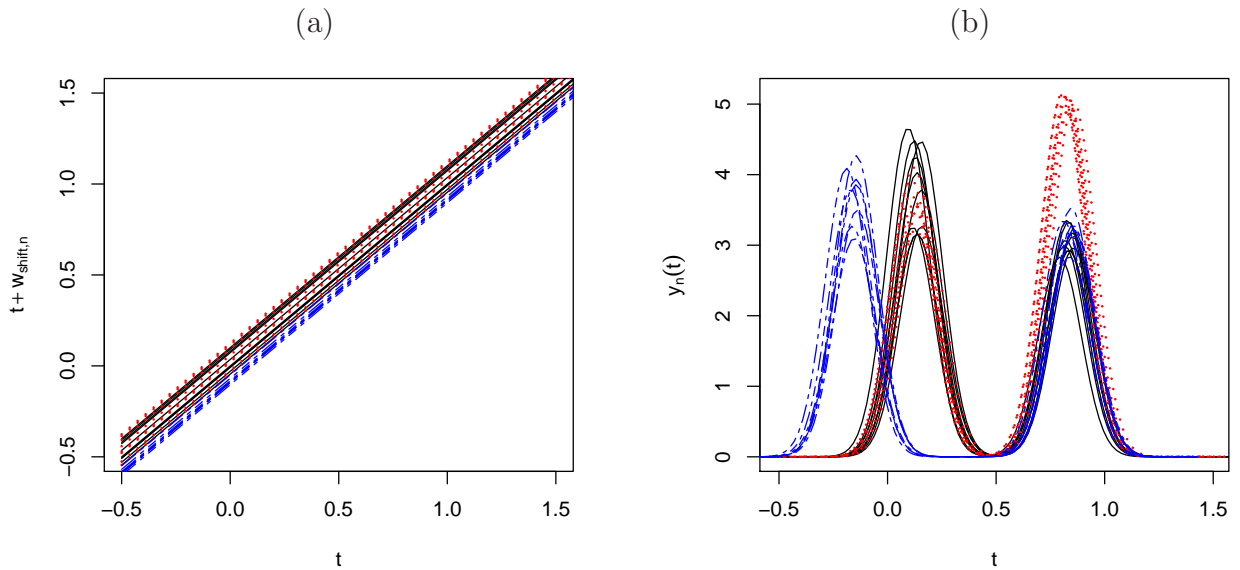


Figure 2: (a) Estimated shifts and (b) data after applying the horizontal shift ($w_{\text{shift},n}$) only.

mated warping components for the illustrative example (Figure 3, panel (a)). By letting the components operate on the same domain across curves, the warplet intensities characterize the difference between the curves with respect to the warping stage. By allowing for curve specific warping domains, not only would there be $3N - 2$ more parameters to estimate, but the averaging constraint (3) would not be satisfied and a direct comparison of the warplets would no longer be possible. In terms of computation time, which is related to the number of parameters, we are able to include about four components with a fixed domain, as compared to only one component with curve-specific domains.

We stopped the warping procedure after just one warplet, Figure 3 (b) gives a satisfactory result showing the aligned curves. An additional, second warplet was included but did not result in a noticeable warping improvement. This was also indicated by the model selection criterion discussed in Claeskens et al. (2010) and which is included in the output of the MRwarp function (Slaets et al., 2010). The warplet intensities are included in the data summary matrix, see Table 1. Figure 3 (a) illustrates clearly that the group of eight curves for which the peaks are further apart all have positive intensities for the first warplet (first a dilation, then a compression) and a large compression domain, in order to mimic a local shift to the right in a smooth way. The other curves have negative estimated intensities for the warplet to meet them in the ‘middle’.

3.2. Kernel coefficients

In this illustrative example the choice of the number of kernels, being two, was taken since each of the curves consists of two peaks of varying heights. When including more kernels

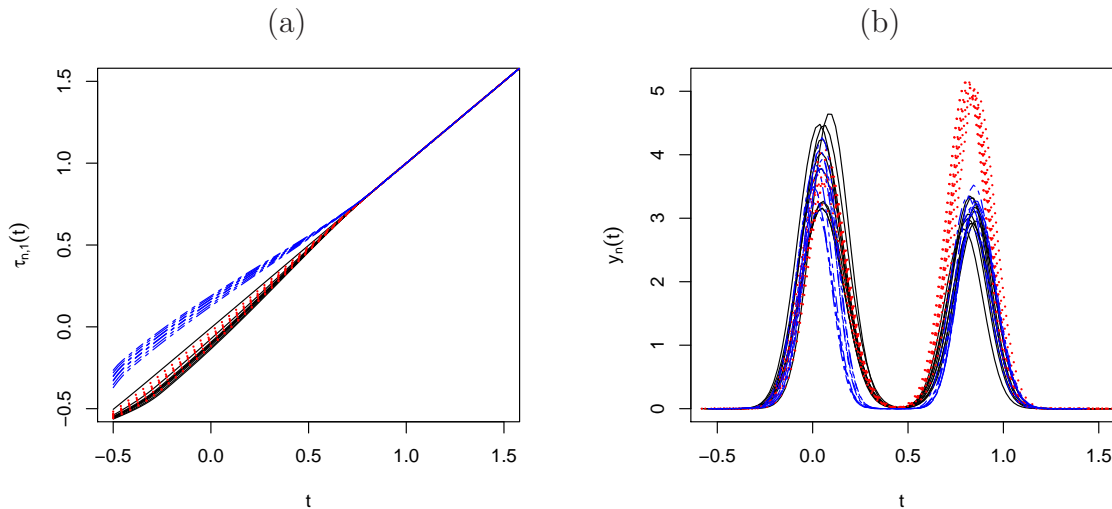


Figure 3: (a) Estimated first warplet for each of the curves, (b) warped data after applying one warplet,

than necessary, it is observed that the multiresolution warping model (2) moves them outside the range of the data. It is a good strategy to inspect the estimated kernel parameters of the output of MRwarp and re-estimate the model with fewer kernels should there be redundant ones. Figure 4 illustrates how amplitude variation is captured by means of the two kernels in the model. It displays one functional curve observation from the sample of 25 such curves in Figure 1, together with the kernel functions obtained from multiresolution warping, the estimated overall mean function and a prediction $\hat{\mu} + \hat{b}_{n,1}\psi_1 + \hat{b}_{n,2}\psi_2$ of this particular warped curve. The predicted kernel coefficients $\hat{b}_{n,1}$ and $\hat{b}_{n,2}$ contain information on whether a peak (in this example) is larger, or smaller, than the average value of the peak. These coefficients are used to summarize the amplitude variation in the data. The full multivariate summary data matrix for this example is presented in Table 1. It contains for each of the 25 curves in the sample the estimated horizontal shift parameters $w_{\text{shift},n}$, the intensities of the (first) warplet, $\lambda_{n,1}$, as well as the predicted kernel coefficients $\hat{b}_{n,1}$ and $\hat{b}_{n,2}$. The table illustrates clearly that for the group of eight curves with increased height of the second peak (curves 10–17) the predicted coefficients $\hat{b}_{n,2}$ are all positive, while for the other curves these predicted coefficients are negative.

Alternatively one could use functional principal components (Rice and Silverman, 1991) to summarize the amplitude variation.

3.3. Rescaling

The units of measurement of the responses $y_n(t_j)$ and arguments t_j determine the range of the shifts and kernel coefficients. To prevent that the choice of the units of measurement

Table 1: Multivariate summary data for the illustrative example consisting of the estimated parameters characterizing phase ($w_{\text{shift},n}$, $\lambda_{n,1}$) and amplitude variation ($b_{n,1}, b_{n,2}$). The numbers in italics for $\lambda_{n,1}$ correspond to the eight (blue) dashed curves in Figure 1, while the numbers in italics for $b_{n,2}$ correspond to the (red) dotted curves. The right-most columns contain the rescaled values.

Estimates					Rescaled estimates				
n	$w_{\text{shift},n}$	$\lambda_{n,1}$	$b_{n,1}$	$b_{n,2}$	n	$w_{\text{shift},n}$	$\lambda_{n,1}$	$b_{n,1}$	$b_{n,2}$
1	-0.008	-0.254	-0.251	-0.565	1	-0.008	-0.254	-0.097	-0.219
2	0.035	-0.280	0.808	-0.426	2	0.035	-0.280	0.313	-0.165
3	0.100	-0.184	-0.232	-0.498	3	0.100	-0.184	-0.090	-0.193
4	-0.046	-0.308	-0.100	-0.586	4	-0.046	-0.308	-0.039	-0.227
5	-0.030	-0.259	0.393	-0.264	5	-0.030	-0.259	0.153	-0.102
6	-0.001	-0.016	0.721	-0.636	6	-0.001	-0.016	0.280	-0.247
7	0.091	-0.231	0.533	-0.239	7	0.091	-0.231	0.207	-0.093
8	0.080	-0.248	0.649	-0.435	8	0.080	-0.248	0.252	-0.169
9	0.072	-0.329	0.272	-0.455	9	0.072	-0.329	0.105	-0.177
10	0.058	-0.173	0.050	<i>1.065</i>	10	0.058	-0.173	0.019	<i>0.413</i>
11	0.094	-0.074	0.092	<i>1.090</i>	11	0.094	-0.074	0.036	<i>0.423</i>
12	0.029	-0.164	0.050	<i>1.084</i>	12	0.029	-0.164	0.019	<i>0.421</i>
13	-0.044	-0.279	0.069	<i>0.944</i>	13	-0.044	-0.279	0.027	<i>0.366</i>
14	0.021	-0.135	0.250	<i>1.119</i>	14	0.021	-0.135	0.097	<i>0.434</i>
15	0.114	-0.174	0.412	<i>0.983</i>	15	0.114	-0.174	0.160	<i>0.381</i>
16	0.124	-0.292	0.218	<i>0.901</i>	16	0.124	-0.292	0.084	<i>0.349</i>
17	-0.026	-0.225	-0.215	<i>0.926</i>	17	-0.026	-0.225	-0.084	<i>0.359</i>
18	-0.104	<i>0.519</i>	-0.413	-0.532	18	-0.104	<i>0.519</i>	-0.160	-0.206
19	-0.082	<i>0.349</i>	-0.678	-0.542	19	-0.082	<i>0.349</i>	-0.263	-0.210
20	-0.092	<i>0.546</i>	-0.288	-0.384	20	-0.092	<i>0.546</i>	-0.112	-0.149
21	-0.091	<i>0.428</i>	-0.849	-0.321	21	-0.091	<i>0.428</i>	-0.329	-0.125
22	-0.069	<i>0.484</i>	-0.288	-0.520	22	-0.069	<i>0.484</i>	-0.112	-0.202
23	-0.091	<i>0.385</i>	-0.454	-0.601	23	-0.091	<i>0.385</i>	-0.176	-0.233
24	-0.088	<i>0.477</i>	-0.101	-0.562	24	-0.088	<i>0.477</i>	-0.039	-0.218
25	-0.049	<i>0.436</i>	-0.780	-0.754	25	-0.049	<i>0.436</i>	-0.303	-0.292

would cause one of the variables to dominate the dissimilarity measure, the variables are rescaled to make them comparable. One could perform a simple standardization, but this will result in a loss of information with respect to the relative importance of the effects. Instead, we proceed as follows:

- Kernel coefficients are multiplied by the same constant c_1 that would rescale the responses $y_n(t_j)$ ($n = 1, \dots, N$ and $j = 1, \dots, T$) such that the range of $c_1 y_n(t_j)$ equals 2.

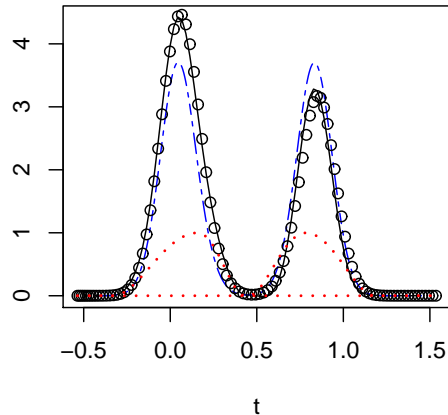


Figure 4: Shifted and warped curve observation (black dots), together with the kernels ψ_1 and ψ_2 (red dotted lines), mean curve $\hat{\mu}$ (blue dashed line) and the predictions $\hat{\mu} + \hat{b}_{1,1}\psi_1 + \hat{b}_{1,2}\psi_2$ (black solid line).

- Warplet intensities are contained in the interval $(-1, 1)$ (range 2) and are left unchanged.
- Shifts are operative on the entire interval $[t_1, t_T]$. The width of $[t_1, t_T]$ with respect to the range of the responses $y_n(t_j)$ needs to be accounted for. Similar as for the responses, the shifts are multiplied by the constant c_2 which would make the range of the rescaled time points $(c_2 t_j)$ equal to 2.

The right-hand side of Table 1 contains the rescaled values that are used in the PAM algorithm.

3.4. Searching for three clusters

We apply the PAM algorithm on the rescaled data matrix. In this illustrative example we first search for three clusters in the data. Clustering based on all of the variables in this summary matrix results for this example in a perfect identification of the three clusters. We refer to the simulation setting for a repeated experiment and for a full comparison with other clustering procedures for functional data.

Table 2 gives the clustering results for application of the new method (MRC) using all variables in the summary matrix, as well as the results from applying the competitive approaches. There are two conventional procedures (i) PAM, this is PAM directly applied on the observed data, (ii) DTW, dynamic time warping followed by PAM, and two methods

Table 2: Clustering results based on the full summary matrix (MRC) and of the methods PAM, DTW, SACK and k -centers FC for the illustrative example. Underlined results indicate misclassified curves according to the real situation with three clusters.

Method	clustering result	real clusters
MRC (all)	11111111122222222333333333	11111111122222222333333333
PAM	11 <u>2</u> 11 <u>2</u> 222222 <u>2</u> 1 <u>2</u> 221 <u>3</u> 33333333	⋮
DTW	<u>3</u> 1 <u>3</u> 3111112222222231 <u>1</u> 31113	⋮
SACK	11111111122222222333333333	⋮
k -centers FC	11 <u>2</u> 11 <u>2</u> 222222 <u>2</u> 1 <u>2</u> 221 <u>3</u> 33333333	⋮

especially for functional data (iii) k -centers functional clustering (FC) and (iv) SACK, simultaneous registration and clustering. See Section 4.2 for a description. Only MRC and SACK classify the curves correctly for this illustrative example. The k -centers FC and PAM both correctly identify the cluster with a phase difference, but do not make a correct classification according to the amplitude differences. The method DTW, which eliminates phase variation and applies PAM on the warped data, only detects the amplitude clusters as expected. Indeed, when only using the warped data the information in the data concerning phase is thrown away.

3.5. Searching for two clusters

A strong advantage of the new method is that several more specialized clustering options can be considered. If not searching for the three real clusters, but rather for two clusters, we can direct the search towards finding clusters that are characterized by differences in phase. Therefore we take the set of variables in the data summary matrix that gives information on phase variation, this corresponds to the shift parameter and the intensities of the warplet. Taking both vectors of coefficients together in the clustering algorithm PAM (Table 3, line 1) results in a correct classification according to the true situation where the set of curves with the two peaks further apart form one cluster and all other curves form another cluster. Detailed results for the illustrative data example are presented in Table 3.

By applying PAM clustering on each of these two phase variables separately (Table 3, lines 2 and 3), we see that the warplet intensities give the needed information for separating clusters in phase.

Alternatively, we can search for clusters that differ with regard to amplitude. Therefore we take the columns of the predicted kernel coefficients and apply the PAM algorithm to the bivariate vector (Table 3, line 4). This results in a perfect classification, where the true situation is now formed by one cluster of curves with a higher second peak, and all other

Table 3: Clustering results based on several sets of variables for the illustrative example. Underlined results in the first three rows indicate misclassified curves according to the real phase clusters. Underlined results in the next three rows indicate misclassified curves according to the real amplitude clusters.

Set of variables	clustering result	real clusters
$w_{\text{shift},n}, \lambda_{n,1}$ (all phase)	1111111111111111111122222222	1111111111111111111122222222
$w_{\text{shift},n}$	<u>2</u> 112221111111 <u>2</u> 111222222222	⋮
$\lambda_{n,1}$	1111111111111111111122222222	⋮
$b_{n,1}, b_{n,2}$ (all amplitude)	11111111122222222211111111	11111111122222222211111111
$b_{n,1}$	1 <u>2</u> 112222222222222 <u>2</u> 11111111	⋮
$b_{n,2}$	11111111122222222211111111	⋮
PAM	1111111111111111111122222222	—
DTW	11111111122222222211111111	—
SACK	1111111111111111111122222222	—
k -centers FC	211221111111211122222222	—

curves form the other group. The coefficient $b_{n,1}$ is used for the first kernel, located around the first peak in the curves, and, as expected, does not give us the wanted information (Table 3, line 5). The coefficient vector $b_{n,2}$ contains all the information concerning the amplitude variation (Table 3, line 6).

These settings illustrates how ‘local’ clusters could be sought, where areas of specific interest, e.g. the area around the second peak, can receive separate attention in the search for clusters. Figure 5 gives a graphical representation of the values in the data summary matrix, as well as the result of the clustering based on each variable separately.

In Table 3, the methods SACK and PAM are seen to identify the cluster that differs in phase (lines 7 and 9), while DTW (line 8) is able to find the amplitude clusters, as expected (see section 3.4). The method of k -centers FC makes a split in two clusters without a clear interpretation for this data (Table 3, line 10).

3.6. Selecting the number of clusters

When one is unsure about which and how many clusters to consider, one can look at the average silhouette width of the clusters (Rousseeuw, 1987), see Figure 6 for an example. The silhouette of each curve n is a measure for the similarity of this curve to the members of its own cluster, compared to its similarity to the ‘nearest’ cluster. In particular, denote $d(n, C)$ the average dissimilarity of curve n with respect to each of the members of its cluster C , and $d(n, C')$ the minimum of the average dissimilarity of curve n with respect to each of the members of any other clusters C' . The silhouette of curve n is then defined as $\{d(n, C') - d(n, C)\} / \max\{d(n, C') - d(n, C)\}$ and the silhouette width of each cluster is the average of the silhouettes of its members. The latter is a good measure for the degree of

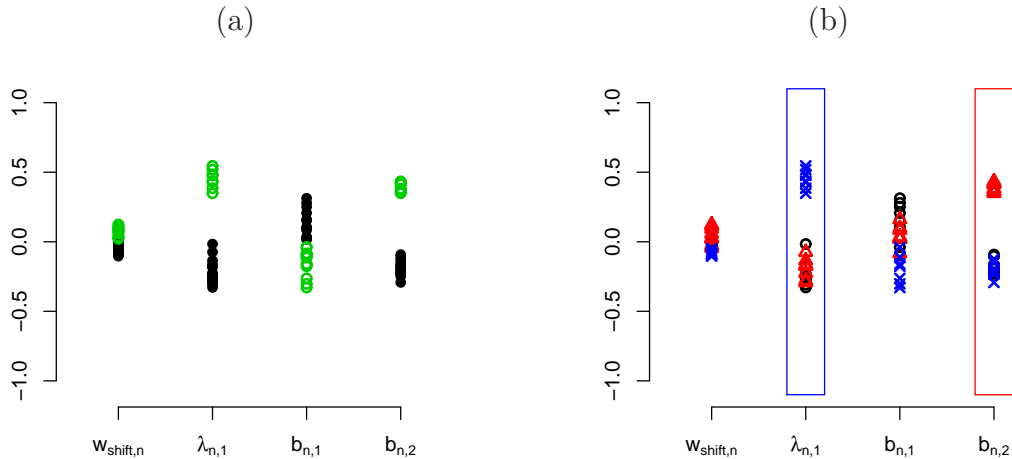


Figure 5: Scatterplots of each of the variables in the summary matrix for the illustrative example. (a) The different symbols indicate the two clusters selected by the PAM algorithm, when applied to each of the variables separately. (b) The different symbols indicate the true three underlying clusters (as in data plots), with (red) triangles representing the (red) dotted curves (clusters in amplitude) and (blue) crosses the (blue) dashed curves (clusters in phase).

separation of the clusters. When clustering is based on each of the variables ($w_{\text{shift},n}$, $\lambda_{n,1}$, $b_{n,1}$ and $b_{n,2}$) separately, in the illustrative example, the highest value of the silhouette width was each time obtained for two clusters, with the highest value for for $b_{n,2}$ (0.91), followed by that for $\lambda_{n,1}$ (0.86). For $w_{\text{shift},n}$ and $b_{n,1}$ these values were much lower (0.66 and 0.56 respectively). We see in Figure 5 that it are indeed the $b_{n,2}$ and $\lambda_{n,1}$ parameters that hold information with respect to the true clusters. When there is no information regarding which aspect is most important for clustering, we apply the MRC(all) method using all of these variables simultaneously. We compared the average silhouette widths for the resulting PAM clustering algorithm with 2, 3, 4, 5 and 6 clusters. The average silhouette width was the largest for 3 clusters (0.69), as we would expect, and decreased when increasing the number of clusters. More precisely, for this example, the silhouette widths are 0.59 for $k = 2$, 0.63 for $k = 4$, 0.53 for $k = 5$ and 0.52 for $k = 6$.

3.7. Example with one outlying curve

To investigate the effect of outlying curves, we have modified the illustrative example such that the last curve does not have the first peak (the response values are set equal to zero), only the second peak remains. We align the curves using one warplet and two kernels. The multiresolution clustering method which uses the shift parameter, warping intensities and both kernel coefficients for each of the curves, classifies the curves correctly into three groups. The outlying curve is classified with the group of curves to which it should

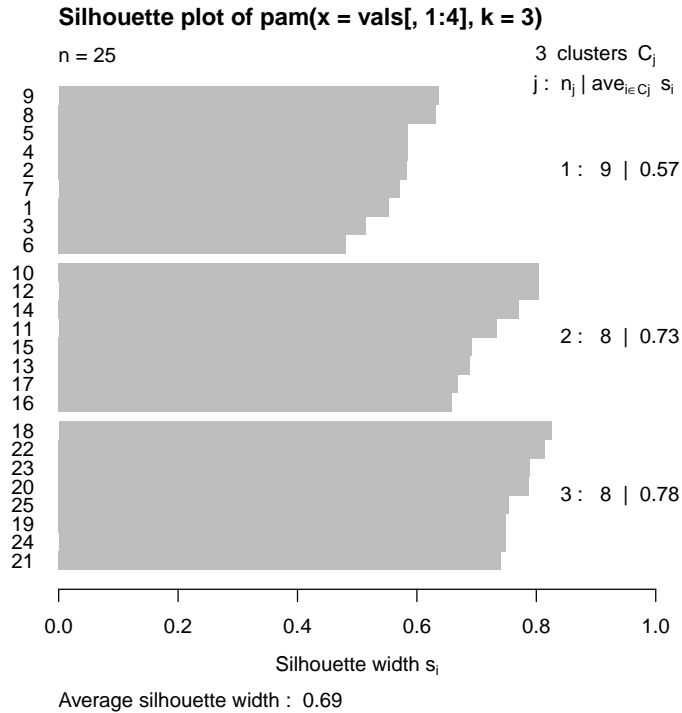


Figure 6: Silhouette plot of the PAM clustering result using the complete MRC summary matrix, for 3 clusters.

belong according to its second peak. In the silhouette plot the silhouette width of this curve is much smaller than that of the other curves, clearly indicating the outlyingness of this curve. The method of k -means applied to this same input, places the outlying curve in a separate cluster, and splits the other curves into two groups, according to their amplitude characteristics. When letting both methods classify the curves into four groups, they both give the same results, where the outlying curve is one cluster, and the other clusters are correctly identified according to the generated phase and amplitude differences.

4. Simulation study

4.1. Generating the data

The simulated data sets consist of one hundred samples s of each $N = 25$ curves. Each sample s contains three different groups, formed by considering two different types of clusters. One cluster consists of a group of eight curves with increased distance between the peaks, another cluster consists of eight curves with increased height of the second peak. See below for details. The objective is in the first place to detect the three groups, though we also demonstrate that it is possible to uncover the two different sets of clusters when looking for two groups only.

The samples are generated as follows. For time point subscripts $j = 1, \dots, 100$,

$$y_n^{(s)}(t_j) = a_{1,n}^{(s)}\phi((\mu_{1,n}^{(s)}, 0.1); t_j) + a_{2,n}^{(s)}\phi((\mu_{1,n}^{(s)} + \mu_{2,n}^{(s)}, 0.1); t_j), \quad t_j = -0.5 + (j - 1) \cdot 0.02,$$

with $\phi((\mu, \sigma); t)$ the density function of the normal distribution with mean μ and variance σ^2 , and for each simulated sample $s = 1, \dots, 100$:

$$\begin{aligned} \mu_{1,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(0.1, 0.08, -0.1, 0.3), & n = 1, \dots, 17, \\ \mu_{1,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(-0.05, 0.05, -0.15, -0.05), & n = 18, \dots, 25, \\ \mu_{2,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(0.8, 0.01, 0.55, 0.70), & n = 1, \dots, 17, \\ \mu_{2,n}^{(s)} & = 1, & n = 18, \dots, 25, \\ a_{1,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(1, 0.2, 0.7, 1.2), & n = 1, \dots, 25, \\ a_{2,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(0.8, 0.2, 0.7, 0.9), & n = 1, \dots, 9, 18, \dots, 25, \\ a_{2,n}^{(s)} & \text{ drawn from } \bar{\mathcal{N}}(1.05, 0.2, 1.2, 1.3), & n = 10, \dots, 17, \end{aligned} \tag{4}$$

with $\bar{\mathcal{N}}(\mu, \sigma, a, b)$ the truncated normal distribution with mean μ , variance σ^2 and lower and upper bound resp. a and b .

Figure 7 illustrates this setting with two examples of generated samples.

Each curve in the sample follows the same pattern: there are two peaks with random heights and there is a random horizontal shift. For five of the curves the mean distance between the curves is larger than for the other twelve curves, in this way creating a first set of two clusters based on $\mu_{1,n}^{(s)}$ and $\mu_{2,n}^{(s)}$ in (4). An additional set of clusters is constructed in the first set of twelve curves by increasing the average height of the second peak, as indicated by $a_{2,n}^{(s)}$ in (4). Thus we can distinguish two clusters based on the distance between peaks (curves 1 to 17 versus curves 18 to 25) or two clusters based on the height of the second peak (curves 1 to 9 and 18 to 25 versus 10 to 17), which means that this is actually a setting with three clusters (curves 1 to 9; curves 10 to 17; curves 18 to 25).

4.2. Functional clustering approaches

We compare the multiresolution clustering method (MRC) with two conventional methods (which we denote by PAM and DTW) and two advanced functional clustering models: the SACK model (simultaneous alignment and clustering k -centers) (Liu and Yang, 2009) and k -centers functional clustering (k -centers FC) (Chiou and Li, 2007). We briefly explain each one.

Naively, one could consider the raw functional data as multivariate data objects where the columns (time points) are treated as different variables. Therefore, the original data can be directly supplied to the PAM algorithm (based on Euclidian distances). We refer to this first approach by the PAM method.

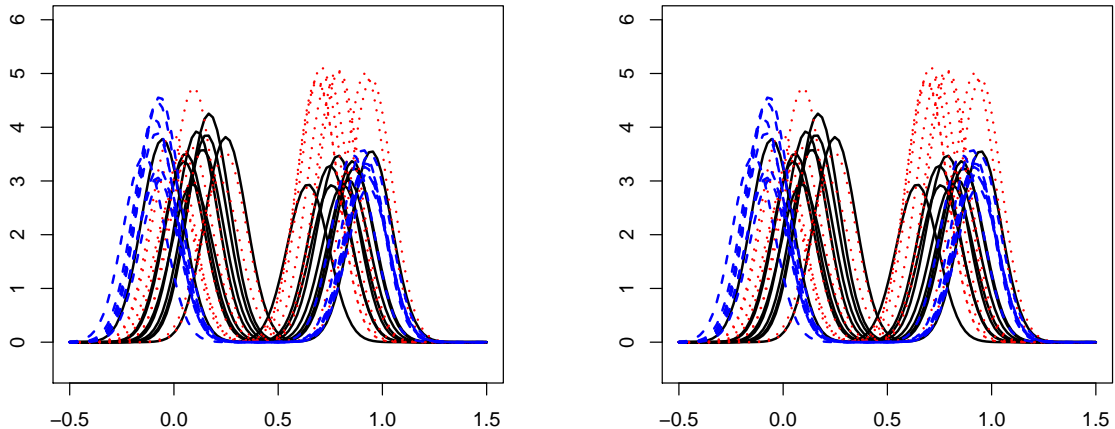


Figure 7: Two examples of simulated data samples in setting 1. Five curves with increased distance between the peaks (blue) dashed lines and 6 curves with increased height of second peak (red) dotted lines.

A popular method for aligning curves is dynamic time warping (DWT, Sakoe and Chiba, 1978). In the DTW clustering approach we first warp the data using the R library dtw (Giorgino, 2009). Afterwards, the warped data, in which the phase variation was eliminated, are used in the PAM algorithm (based on Euclidian distances).

Liu and Yang (2009) (SACK) introduce cluster membership by means of a model with shifted B-spline basis functions and cluster-specific basis coefficients. This means that clusters are characterized by similar patterns, while allowing for random phase differences (basis functions shifts b_i) and amplitude differences (shifts d_i) within clusters.

$$Y_{ij} = d_i + \sum_{l=1}^l \beta_l^{(k)} B_l(b_i + t_{ij}) + \epsilon_{ij} \approx d_i + \sum_{l=1}^l \beta_l^{(k)} B_l(b_i) + b_i B_l'(t_{ij}) + \epsilon_{ij},$$

with the B_l cubic B-spline basis functions, $d_i \sim \mathcal{N}(0, \sigma_d^2)$, $b_i \sim \mathcal{N}(0, \sigma_b^2)$ and $\epsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The random shifts b_i constitute the curve registration aspect of their method.

In Chiou and Li (2007) (k -centers FC), an advanced functional version of the popular multivariate k -means clustering algorithm is introduced via a truncated Karhunen-Loève representation of each cluster c of curves Y in each step of the algorithm

$$\tilde{Y}^{(c)}(t) = \mu^{(c)}(t) + \sum_{j=1}^{M_c} \xi_j^c(Y) \rho_j^c(t),$$

with eigenfunctions ρ_j associated with the covariance ($\langle \text{cov}[Y(s), Y(t)], \rho_j^c \rangle = \lambda_j^c \rho_j^c(t)$) and

$\xi_j^c(Y)$ uncorrelated random variables with zero mean and variance λ_j^c such that $\xi_j^c(Y) = \langle Y - \mu^c, \rho_j^c \rangle$. We refer to the mentioned paper for more details. Briefly, if in iteration i Y is wrongly classified and does not belong to cluster c , discrepancies exist between $Y^{(c)}$ and Y which can favor a change of cluster membership by comparing Y with its truncated expansion with respect to the Karhunen-Loève eigenbases of the other clusters. These bases maximize the percentage of total variation explained in the cluster curves, solving the issue of having to choose the one set of proper basis functions (e.g. B-splines with equidistant knots).

4.3. Clustering: simulation results

We use the adjusted rand index (Hubert and Arabie, 1985) to evaluate the performance of MRC and that of the competitors. The adjusted rand index is a measure of the similarity between the clustering result and the true configuration, taking into account the chance mechanism. It has a maximum value of 1, which indicates a perfect clustering result.

We start by searching for three groups in the data. For the MRC method, we use the information of all coefficients. Figure 8 shows boxplots of the adjusted rand index over the simulation runs. The true situation is one where there are three clusters. The good performance of the new method which is based on the information contained in the warped curves, is clearly visible. The PAM, DTW and k -centers FC method have difficulties with detecting the three correct clusters. In a lesser extend this holds for the SACK method as well.

Next we ask some more specialized questions. We wish to specify clusters of curves that distinguish from other curves in showing more phase variation. This implies that in a warping action, these curves need more severe warping than other curves. For the simulation setting this corresponds to identifying the eight curves for which the average distance between the peaks is larger than for the other curves. The multiresolution warping method can do this by focussing at the phase parameters ($w_{\text{shift},n}$ and $\lambda_{n,1}$) in the summary matrix, see Figure 8(b). The SACK approach is also able to detect the phase clusters, while the PAM and k -centers FC, which are not designed to handle phase variation, perform not so good. Worst in this scenario is the performance by DTW. This is no surprise as this methods eliminates all phase information prior to the clustering.

In this example it is easier for the methods, apart from DTW, to detect the clusters in phase, than to detect the clusters in amplitude. The boxplots in Figure 8(c) show that the MRC method (using all variables), k -centers FC as well as the SACK method fail to find the clusters defined by amplitude differences. A strong characteristic of the MRC method is that we can specify which variables should be used for clustering, which is not possible for any of the other methods. By focussing on the amplitude coefficients only, we are able to

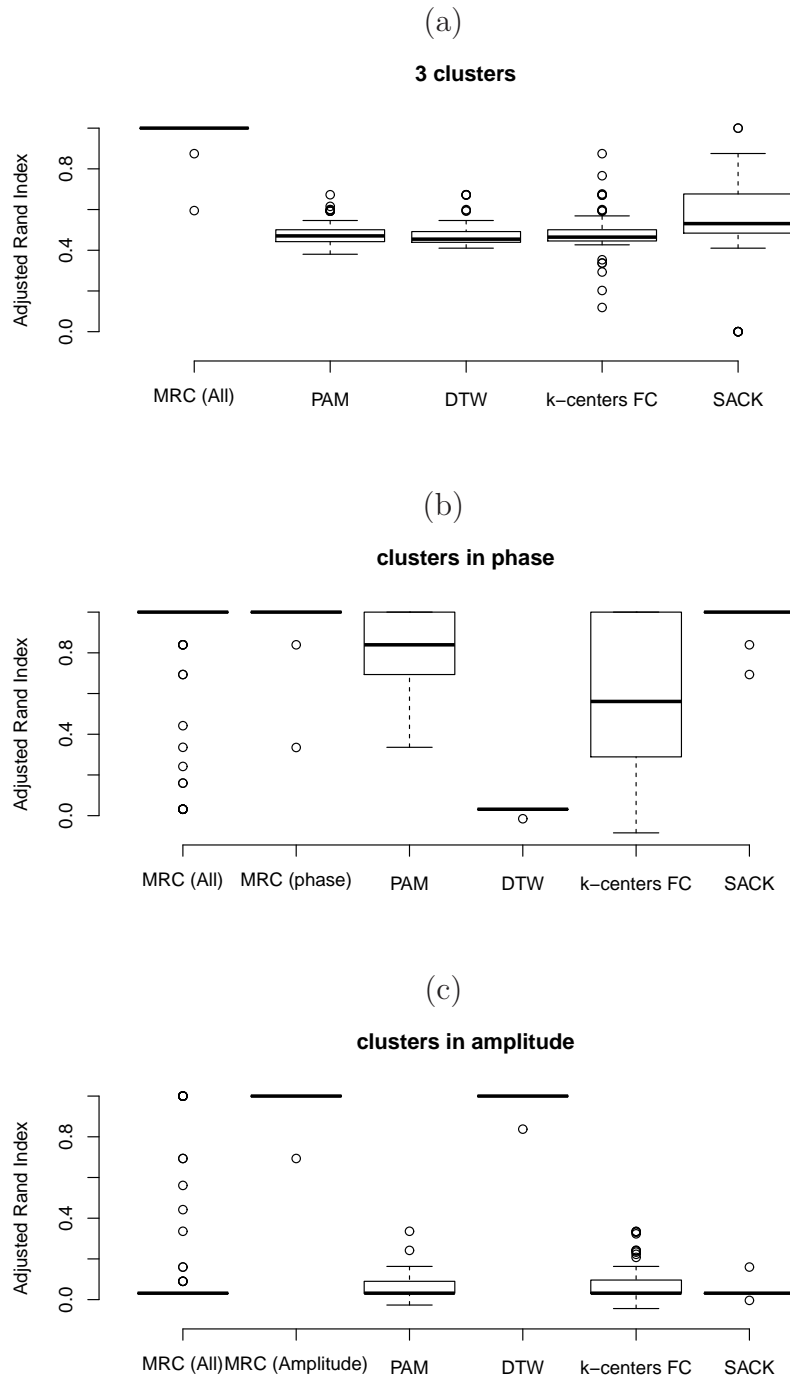


Figure 8: Boxplots of the adjusted rand index with respect to (a) the three true clusters, (b) two clusters based on the distance between the peaks, and (c) two clusters based on the height of the second peak, for resp. MRC based on all variables in the summary matrix (MRC All), PAM, DTW, k -centers FC and SACK.

identify the correct clusters in amplitude.

5. Clustering the Berkeley growth data

The Berkeley growth data (Tuddenham and Snyder, 1954) encompasses 31 height measurements of boys and girls over a period of 17 years (from age 1 to 18). Important features of human growth are easily observed when looking at growth velocity or acceleration; the derivatives of the originally observed process. Hence growth curves are processes which have intrinsic functional features. Instead of clustering the entire sample, with obvious differences between boys and girls, we focus on the more homogeneous sample of velocity curves for the 39 boys, as shown in Figures 9.

Multiresolution warping was performed with one warplet and two kernel functions. The warplet takes action on the interval (9.90, 22.28) with a center at 12.33. We note that the warping and kernel bounds are allowed to extend beyond the observation domain to incorporate possible increased variability at the boundaries. The first kernel acts on the interval (0.03, 7.53) with a center at 0.86, while the second kernel focusses on amplitude variation in the (2.98, 19.95) region with the center at 12.98. We immediately see that the first warping component and the second kernel function focus on the variation around age 12. This leads to the interpretation that the variation in the pubertal growth spurt is related both to phase (timing) and amplitude (magnitude).

Figure 9(b) shows the clustering result when applying the PAM algorithm for constructing two clusters after multiresolution warping. It selects those boys who have a strong downfall in growth velocity, right after the initial strong growth peak after birth.

Since the MRC method allows to search for local clusters in functional data, we decide to change the focus from the early age growth variation towards that at a later age. Therefore we concentrate on the information contained in the second kernel (Figure 9 (c)) and the warping component (Figure 9 (a)). The second kernel detects overall lower velocities in the 8–16 age range, while the warping component forms clusters based on the timing of the pubertal growth spurt. Finally, Figure 9(d) and (e) shows the clustering results for the two competitive methods considered in section 4: the SACK model and k -centers FC. For the SACK model it is not clear which feature exactly characterizes the clusters, but it seems to be a joint late and small pubertal growth spurt effect. For k -centers FC the groups are based on overall low versus high velocity curves, lacking any phase perspective and mixing information from several growth periods. As with most clustering methods, we get only one clustering result and no additional information on what caused the method to build these groups, contrary to MRC, which supplies us with three options and allows for meaningful interpretations.

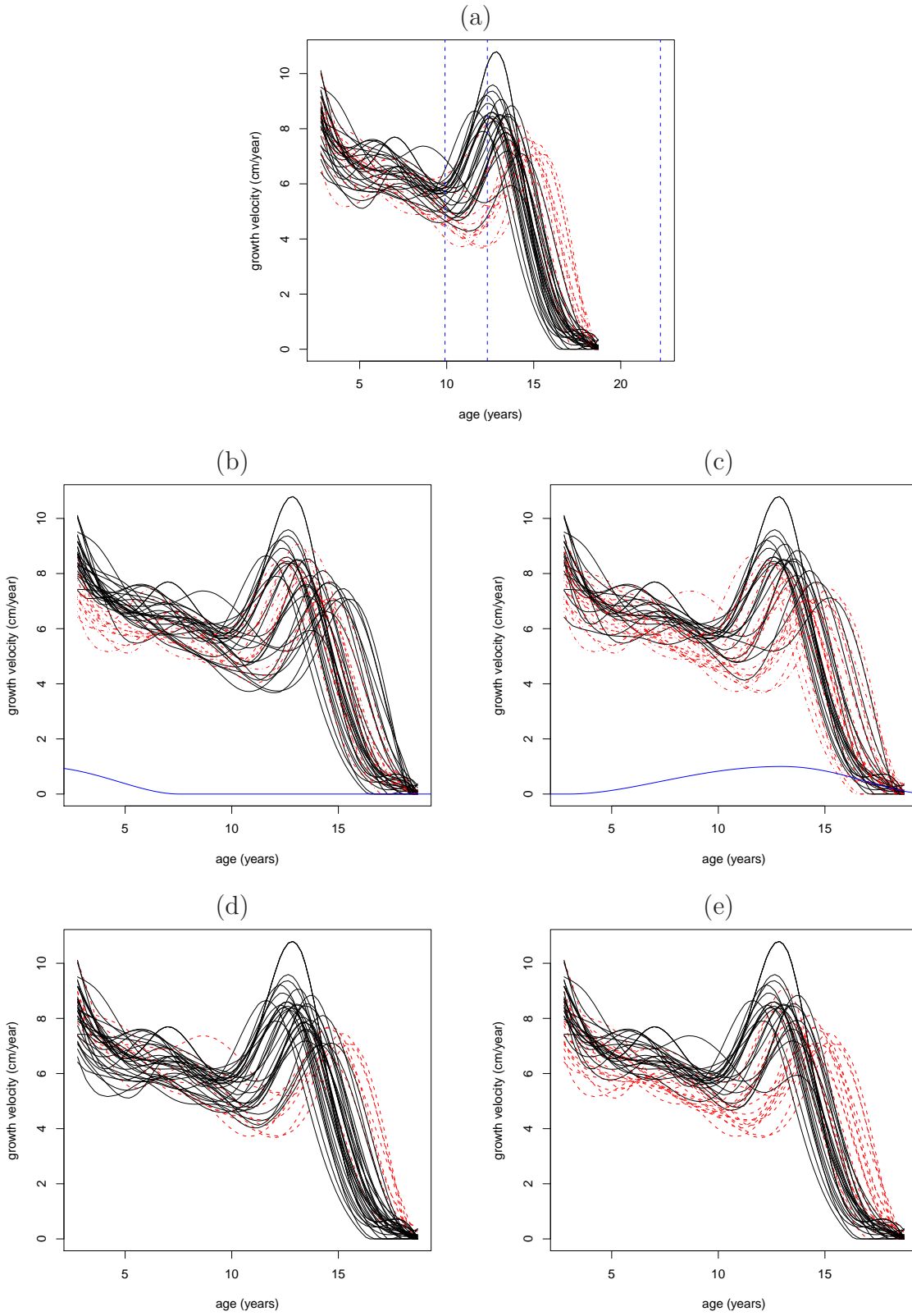


Figure 9: Growth velocity curves for 39 boys. The red, dashed lines are for one cluster and the black solid lines are for the other one. Two clusters based on (a) the warplet only, together with (as vertical blue dashed lines) the warping bounds and the warping center, (b) kernel 1 only, (c) kernel 2 only, (d) SACK, and (e) k -centers FC. Parts (b) and (c) also show the estimated kernel function (blue).

6. Discussion

The important contribution of this research is that we can explicitly use the information from warping functional data in a further analysis, this is otherwise not actively done in the available research. In particular, this information can be used for (i) clustering of functional data, (ii) outlier detection, (iii) classification. While the topic of the current paper is about clustering, current research considers the other two methods.

The dimension reduction provided by the warping facilitates the use of existing clustering approaches, without the need to develop new methods for the sole purpose of clustering functional data.

A strong point of the proposed multiresolution approach is that we can direct the search towards finding clusters with phase differences, or clusters with amplitude differences, or do not specify any option, and look simultaneously for clusters with might differ in phase and/or amplitude.

When local or some specific features are of interest, the available information can be explored in more detail. Multivariate tools can aid one in the search for clusters which yield the best separation of the curves. For instance, the warplets are built from a multiresolution approach and have a clear interpretation with respect to both location and intensity. The search for local clusters in a set of curves with similar shapes can be pursued by considering the effects of the warplets related to a specific region. This leads towards searching for clusters within a specific time frame.

While the MRC-all approach takes all of the standardized variables of the multiresolution warping method, one could weigh each of the variables according to the amount of variation, according to a chosen measure, that they explain in the curve sample. For example, when a set of curves displays much more phase variation than amplitude variation, the variables related to phase could be decided to receive a larger weight.

A simulation study showed improved clustering performance with respect to two advanced competitive methods, in a complex phase-amplitude setting. The method is also illustrated by means of the well-known Berkeley growth curves, where we considered the more homogeneous sample of velocity curves for boys.

Acknowledgements

The authors wish to express their thanks to all the reviewers of this paper, whose helpful comments have resulted in an improved paper.

Appendix A. R code for the multiresolution clustering

The R library MRwarping can be accessed from the webpage <http://perswww.kuleuven.be/gerda.claeskens/software>. The code for the analysis of the illustrative data example is included in the file MRClust.examplecode.R.

References

- Abraham, C., Cornillon, P., Matzner-Lober, E., and Molinari, N. (2003). Unsupervised clustering using B-splines. *Scandinavian Journal of Statistics*, 30(3):581–595.
- Barbieri, P., Adami, G., Favretto, A., Lutman, A., Avoscan, W., and Reisenhofer, E. (2001). Robust cluster analysis for detecting physico-chemical typologies of freshwater from wells of the plain of friuli (northeastern italy). *Analytica Chimica Acta*, 440(2):161–170.
- Bozinov, D. and Rahnenführer, J. (2002). Unsupervised technique for robust target separation and analysis of dna microarray spots through adaptive pixel clustering. *Bioinformatics*, 18(5):747–756.
- Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society, series B*, 69:679–699.
- Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy. *Journal of the Royal Statistical Society. Series B*, 72(5):673–694.
- Cuesta-Albertos, J., Gordaliza, A., and Matrán, C. (1997). Trimmed k -means: An attempt to robustify quantizers. *The Annals of Statistics*, 25:553–576.
- DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5:249–282.
- Douzal-Chouakria, A., Diallo, A., and Giroud, F. (2009). Adaptive clustering for time series: application for identifying cell cycle expressed genes. *Computational Statistics & Data Analysis*, 53(4):1414–1426.
- Gallegos, M. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, 33:347–380.
- García-Escudero, L. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969.

- García-Escudero, L. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, 22:185–201.
- García-Escudero, L., Gordaliza, A., Matrán, C., and Mayo-Isar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4:89–109.
- Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association*, 90:1179–1188.
- Gervini, D. and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika*, 92(4):801–820.
- Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7):1–24.
- Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517.
- Hennig, C. (2008). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *J. Multivar. Anal.*, 99:1154–1176.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.
- James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408.
- James, G. M. (2007). Curve alignment by moments. *The Annals of Applied Statistics*, 1(2):480–501.
- Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.
- Liu, W.-m., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T. B., Webster, T. A., Dong, S., Liu, G., Jones, K. W., Kennedy, G. C., and Kulp, D. (2003). Algorithms for large-scale genotyping microarrays. *Bioinformatics*, 19(18):2397–2403.
- Liu, X. and Yang, M. (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53:1361–1376.
- Lopez-Pintado, S. and Romo, J. (2005). Depth-based classification for functional data. Statistics and econometrics working papers, Universidad Carlos III, Departamento de Estadística y Econometría.

- Rabiner, L., Rosenberg, A., and Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE TRans. on Acoustics, Speech and Signal Processing*, ASSP-26(6).
- Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B*, 60(2):351–363.
- Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B*, 53:233–243.
- Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:43–49.
- Sangelli, L., Secchu, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54:1219–1233.
- Slaets, L., Claeskens, G., and Silverman, B. W. (2010). Warping functional data in R and C via a Bayesian multiresolution approach. KBI-report 1013, Revised manuscript submitted, K.U.Leuven.
- Sood, A., James, G., and Tellis, G. (2009). Functional regression: A new model for predicting market penetration of new products. *Marketing Science*, 28:36–51.
- Struyf, A., Hubert, M., and Rousseeuw, P. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics and Data Analysis*, 26:17–37.
- Tuddenham, R. and Snyder, M. (1954). Physical growth of california boys and girls from birth to eighteen years. *University of California Publications in Child development*, 1:183–364.
- Verboven, S. and Hubert, M. (2005). LIBRA: a Matlab library for robust analysis. *Chemo-metrics and Intelligent Laboratory Systems*, 75:127–136.
- Wand, M., Coull, B., French, J., Ganguli, B., Kammann, E., Staudenmayer, J., and Zanobetti, A. (2005). *SemiPar 1.0. R package*.
- Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276.