KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Phase and amplitude-based clustering for functional data

Gerda Claeskens, Mia Hubert and Leen Slaets

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1030

Phase and amplitude-based clustering for functional data

Gerda Claeskens[1,*], Mia Hubert[b], Leen Slaets[1]

[a] *ORSTAT and Leuven Statistics Research Center, Naamsestraat 69, 3000 Leuven, Belgium*
[b] *Department of Mathematics and Leuven Statistics Research Center, 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, we explicitly recognize it as a possible important source of information for clustering. We illustrate 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 simultaneous.

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 economy (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)

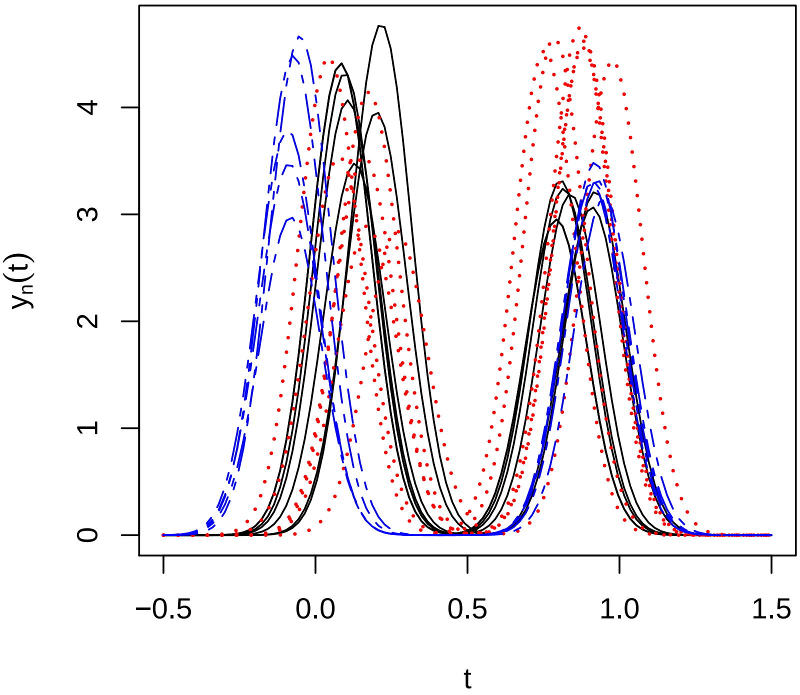

Figure 1: Illustrative data set with three clusters. The five (blue) dashed lines represent curves for which the distance between the peaks is larger than for the other curves. The six (red) dotted lines represent curves that possess a higher second peak.

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 17 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.

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

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

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

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 $[a - \frac{3\sqrt{3}}{8}\lambda r, a + r_2]$. Asymmetric domains of dilation and compression require less warplets in the final warping function than would be the case when using symmetric components (see Claeskens et al., 2010).

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

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

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,

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

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:

$$y_n(t_j) = y_{n,j} = F_{n,j} + e_{n,j} = \mu(\tau_n(t_j)) + \sum_{k=1}^{K} b_{n,k}\psi_k(\tau_n(t_j)) + e_{n,j}, \tag{2}$$

with $b_{n,k}$ and $e_{n,j}$ 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 fixed 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. 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 parameters and for each component they are directly comparable across curves. As an averaging constraint, the intensity parameters satisfy that

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

Since the warplets can only model local phase variation, we have included horizontal shifts in a preprocessing step, which allows the original curves to be shifted over a limited time range similar to the preprocessing step of Slaets et al. (2010). These shifts $w_{shift,n}$ are estimated by minimizing the sum of the squared distances between each set of two shifted curves. As a distance measure between the curves we took the average over the squared distances in each of the time points t_j. Linear interpolation is used to evaluate the shifted curves in these time points. This preprocessing step is optional.

The multiresolution warping approach of Claeskens et al. (2010) uses model (2) to align the (pre-shifted) data by minimizing the squared error between the warped responses, while allowing for amplitude differences by means of a limited number of local variability areas. A Bayesian estimation procedure is used, see Slaets et al. (2010) for details about the implementation. 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 use for clustering 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 joint mean and kernel coefficients goes as follows. For 1000 equally spaced time points t_i, $i = 1, \ldots, 1000$ 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 $\hat{\mu}(t_i) = 1/N \sum_{n=1}^{N} \Big(\tilde{y}_n(t_i) - \sum_{k=1}^{K} \hat{b}_{n,k}\psi_k(t_i)\Big)$.
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 is estimated and we have for each curve estimates of the model coefficients, these are used for clustering the curves.

2.3. Partitioning around mediods 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 mediods (PAM) clustering technique, as described in Kaufman and Rousseeuw (1990).

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

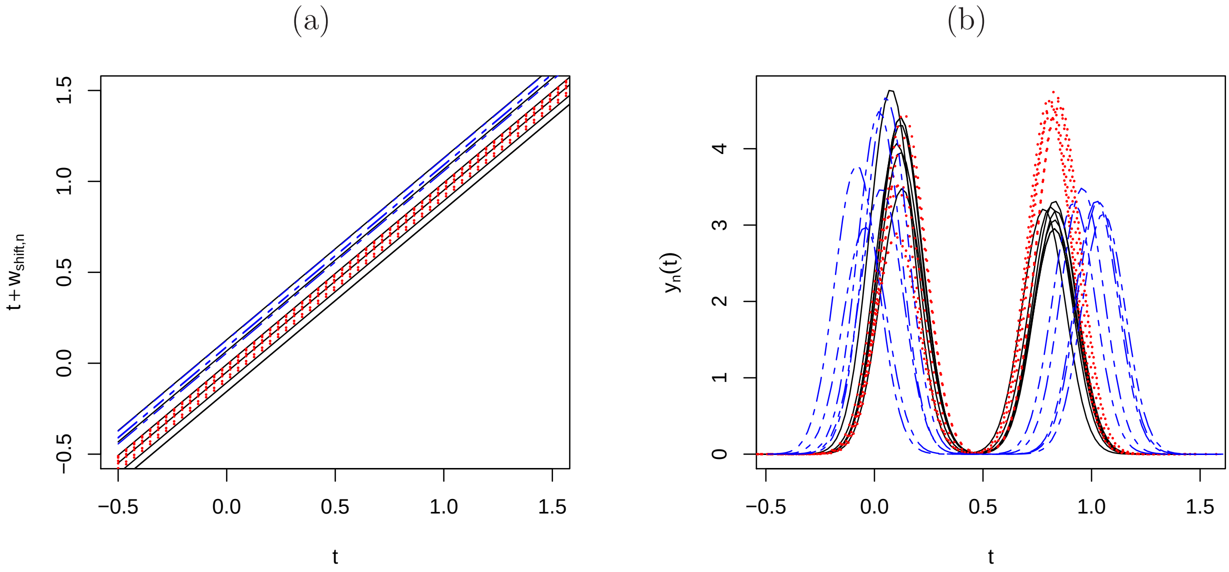

Figure 2: (a) Estimated shifts and (b) data after preprocessing with a horizontal shift only.

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. Note that PAM differs from the k-means algorithm in that a user-defined dissimilarity measure can be used, instead of the Euclidian distance only, it sums these individual distances rather than taking sums of squares, and it uses mediods rather than means. It is a popular approach for multivariate clustering, as it is less sensitive to outlying observations. Implementations of PAM are available in the R-package `cluster`, in S-PLUS (Struyf et al., 1997) and MATLAB (Verboven and Hubert, 2005).

We applied PAM with the Euclidian distance throughout this paper.

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 plots the estimated shifts (left) and the set of horizontally shifted curves (right) which result from the preprocessing step. These shifts form the first column in the multivariate data object, see Table 1, that will be used for clustering 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 estimated warping components for the illustrative example (Figure 3, panels (a) and (c)). In particular

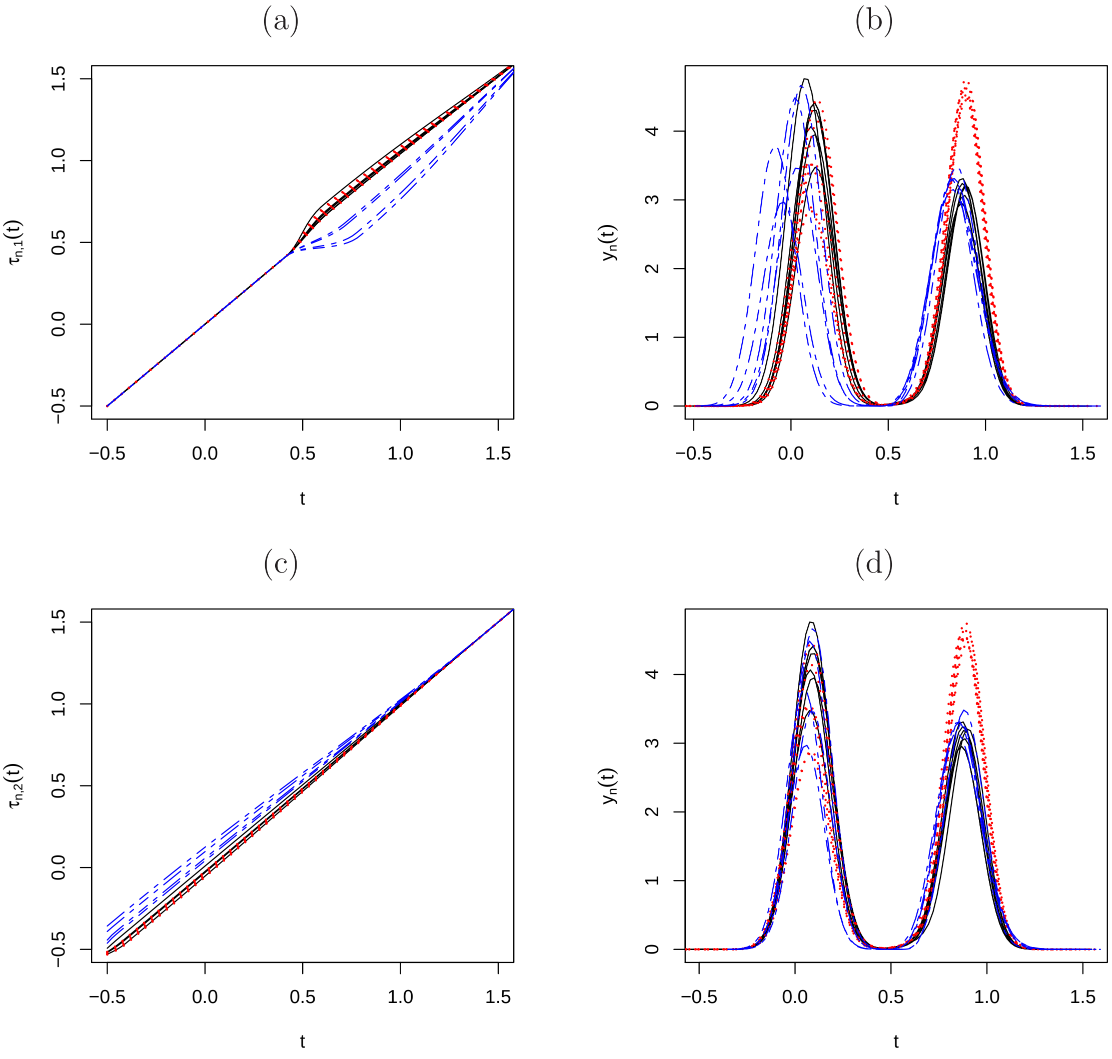

Figure 3: (a) Estimated first warplet for each of the curves, (b) warped data after applying one warplet, (c) estimated second warplet for each of the curves and (d) warped data after applying two warplets.

we note that the warplets are quite asymmetric. 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 estimated $3N$ more parameters, 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 would be 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 composing two warplets, Figure 3(d) gives a

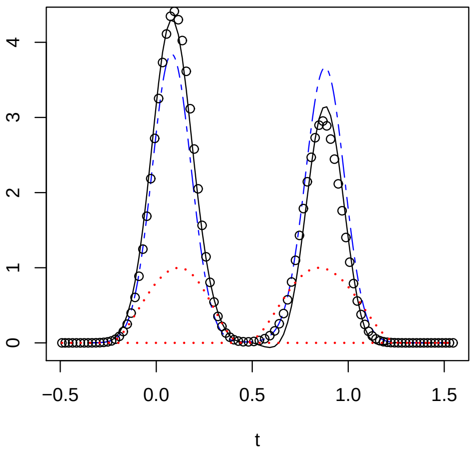

Figure 4: Shifted and warped curve observation $n = 1$ (black dots), together with the kernels ψ_1 and ψ_2 (red dotted lines), mean curve $\hat{\mu}$ (blue dashed line) and the predictions $\hat{\mu} + \hat{b}_{1,1}\psi_1 + \hat{b}_{1,2}\psi_2$ (black solid line).

satisfactory result showing the aligned curves. The warplet intensities are added in two additional columns to the data summary matrix, see Table 1, finalizing the phase variation part of the data summary matrix that will be supplied to the PAM algorithm. Figure 3 (a) and (c) illustrates clearly that the group of five curves for which the peaks are further apart all need negative first warplet intensities (first a compression, then a dilation) and a large dilation domain, to mimic a local shift to the left in a smooth way. A similar effect in the other direction is observed for the second warplet. The other curves have opposite intensities for the warplets 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. 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 17 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 constructed from the estimated mean function and the predicted kernel coefficients, as calculated in the iterative procedure described earlier. 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 amplitude variation in the data.

The full multivariate summary data matrix for this example is presented in Table 1. It

Table 1: Multivariate summary data for the illustrative example consisting of estimated parameters characterizing phase (shift, λ_1, λ_2) and amplitude variation $(b_{n,1}, b_{n,2})$. The numbers in italics for $\lambda_{n,1}$ and $\lambda_{n,2}$ characterize the five (blue) dashed curves in Figure 1, while the numbers in italics for $b_{n,2}$ correspond to the (red) dotted curves.

n	horizontal shift	$\lambda_{n,1}$	$\lambda_{n,2}$	$b_{n,1}$	$b_{n,2}$
1	0.0363	0.1564	-0.1662	0.4733	-0.5183
2	-0.0034	0.2294	-0.1354	0.1722	-0.3521
3	0.0231	0.1595	-0.1592	0.4022	-0.2707
4	-0.0812	0.1990	-0.1551	0.1218	-0.4649
5	-0.1397	0.3808	0.0504	0.6823	-0.4744
6	-0.0018	0.2049	-0.2938	-0.2478	-0.3538
7	-0.0423	0.2394	-0.2459	-0.2024	*0.7020*
8	0.0919	0.1849	-0.3009	0.5434	*0.5843*
9	0.0192	0.3266	-0.1679	-0.2490	*0.5684*
10	-0.0651	0.2394	-0.1372	-0.7163	*0.5528*
11	-0.1672	0.2933	-0.1347	-0.7569	*0.4311*
12	-0.0362	0.1457	-0.3219	0.3024	*0.6944*
13	0.0953	*-0.6543*	*0.3217*	0.3756	-0.0496
14	0.1044	*-0.7798*	*0.2341*	0.5721	-0.0737
15	0.0097	*-0.2998*	*0.7238*	-0.3260	-0.3704
16	0.1184	*-0.6694*	*0.3276*	-0.3496	-0.0489
17	0.0386	*-0.3560*	*0.5608*	-0.8316	-0.1942

contains here for each of the 17 curves in the sample the estimated horizontal shift parameters $a_{w,n}$, the intensities of the first and second warplet, $\lambda_{n,1}$, $\lambda_{n,2}$, as well as the predicted kernel coefficients $(b_{n,1}, b_{n,2})$. The table illustrates clearly that for the group of six curves with increased height of the second peak (curves 7–12) the predicted coefficients $b_{n,2}$ 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 summary the amplitude variation.

3.3. Standardization

The data matrix is columnwise standardized to obtain values that sum to zero and have sample standard deviation equal to one. This is to avoid that large values for, e.g., shifts or kernel coefficients would dominate the dissimilarity measure and hence give wrong clustering results. Table 2 contains the standardized values that are used in the PAM algorithm.

3.4. Searching for three clusters

We apply the PAM algorithm on the standardized data matrix. In this illustrative example we first search for three clusters in the data. Clustering results based on all variables in

Table 2: Standardized multivariate summary matrix.

n	horizontal shift	$\lambda_{n,1}$	$\lambda_{n,2}$	$b_{n,1}$	$b_{n,2}$
1	0.4429	0.4044	-0.5243	0.9600	-1.1809
2	-0.0412	0.5933	-0.4272	0.3523	-0.8172
3	0.2816	0.4125	-0.5023	0.8164	-0.6391
4	-0.9895	0.5146	-0.4892	0.2500	-1.0640
5	-1.7030	0.9846	0.1589	1.3823	-1.0848
6	-0.0215	0.5300	-0.9269	-0.4964	-0.8210
7	-0.5161	0.6190	-0.7757	-0.4047	1.4898
8	1.1177	0.4780	-0.9492	1.1018	1.2321
9	0.2352	0.8446	-0.5295	-0.4988	1.1975
10	-0.7938	0.6191	-0.4329	-1.4429	1.1631
11	-2.0387	0.7584	-0.4248	-1.5249	0.8968
12	-0.4410	0.3768	-1.0154	0.6149	1.4730
13	1.1621	-1.6920	1.0148	0.7628	-0.1551
14	1.2732	-2.0165	0.7384	1.1596	-0.2078
15	0.1184	-0.7753	2.2832	-0.6545	-0.8572
16	1.4437	-1.7309	1.0333	-0.7022	-0.1536
17	0.4703	-0.9206	1.7689	-1.6757	-0.4716

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 3 gives the clustering results for application of the new method using all variables in the summary matrix, as well as the results from applying the competitors. See Section 4.2 for a description of the simultaneous registration and clustering method (SACK) and of k-centers functional clustering (k-centers FC). Only the new method classifies the curves correctly for this illustrative example. The SACK method separates the group of curves with the peaks further apart correctly, but fails to detect the group of curves with a larger second peak. The k-centers functional clustering approach also correctly identifies the cluster with a phase difference, but likewise does not make a correct classification according to the amplitude differences.

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

Table 3: Clustering results based on the full summary matrix and of the methods 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
summary matrix–all	11111122222233333	11111122222233333
SACK	$1111112\underline{1}222233333$	$\vdots$
k-centers FC	$111\underline{22}12\underline{11}22233333$	$\vdots$

Table 4: Clustering results based on several sets of variables for the illustrative example. Underlined results in the first four 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	
shift,$\lambda_{n,1}$,$\lambda_{n,2}$ (all phase)	11111111111122222	11111111111122222	phase
shift	$\underline{222}11\underline{2}1\underline{22}11122222$	$\vdots$	
$\lambda_{n,1}$	11111111111122222	$\vdots$	
$\lambda_{n,2}$	11111111111122222	$\vdots$	
$b_{n,1}$, $b_{n,2}$ (all amplitude)	$1111112222221111\underline{2}$	11111122222211111	amplitude
$b_{n,1}$	$11111\underline{22}1\underline{222}1\underline{11}222$	$\vdots$	
$b_{n,2}$	11111122222211111	$\vdots$	
SACK	11111111111122222	—	
k-centers FC	$111\underline{22}1\underline{2}11\underline{222}22222$	—	

phase variation, this corresponds to the shift parameter and the intensities of both warplets. Taking all three vectors of coefficients together in the clustering algorithm PAM 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 4.

By applying the clustering on each of these three phase variables separately, we see that the coefficients of the warplet intensities both give the needed information for separating clusters in phase. The coefficient of the shift does not help to separate the clusters in this example, as indeed, all curves were generated with a random shift parameter.

Alternatively, we can search for clusters that differ with regard to amplitude. Therefore we take both predicted kernel coefficients and apply the PAM algorithm to the bivariate vector. This results in misclassifying one curve where the true situation is now formed by one cluster of curves with a higher second peak, and all other clusters forming the other group. The coefficient $b_{n,1}$ is used for the first kernel, located around the first peak in

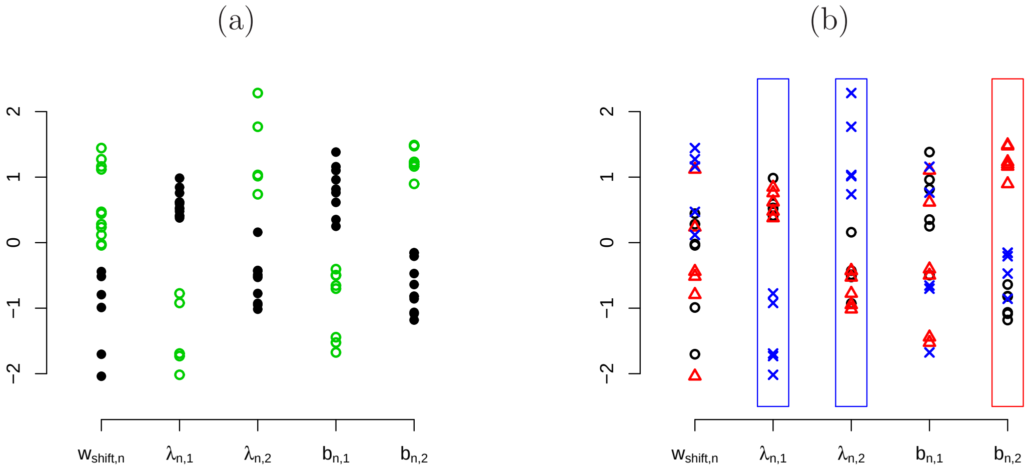

Figure 5: Scatterplots of each of the variables in the summary matrix for the illustrative example. (a) The different symbols indicate the two clusters selected by the PAM algorithm, when applied to each of the variables separately. (b) The different symbols indicate the true three underlying clusters (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).

the curves, and, as expected, does not give us the wanted information. Working with the coefficient vector $b_{n,2}$ which gives the information about the amplitude variation around the second peak gives a perfect classification of the curves in the two clusters. This 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.

The result of the trivariate clustering taking all three phase-characterizing variables together, and of the bivariate clustering taking both amplitude coefficients together, are summarized in Table 4. The PAM algorithm misclassifies none of the observation when clustering is based on phase (line 4 in Table 4), and the two true clusters are defined by separating the set of five curves with the two peaks further apart. This is because the estimated shifts do not separate the sample in any (random) way which can disturb the true clusters in the intensities. When considering the clusters defined by the set of six curves with a larger second peak, it are the predicted coefficients $b_{n,2}$ that should determine the clusters. When taking both kernel coefficients together, the influence of the first kernel coefficient slightly disturbs the clustering. The heights of the first peak are generated randomly, but due to a small sample size, an artificial cluster can arise. For this example, clustering on variable $b_{n,2}$ only gives the best results when searching for clusters that differ in amplitude.

In Table 4, the competitive method SACK for clustering functional data is seen to identify the cluster that differs in phase, but does not find the cluster that differs in amplitude. The method of k-centers FC makes a split in two clusters without clear interpretation for this data.

When one is unsure which and how many clusters to consider, one can look at the average silhouette width of the clusters (Rousseeuw, 1987). 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 separation of the clusters. For each of the variables (shift, $\lambda_{n,1}$, $\lambda_{n,2}$, $b_{n,1}$ and $b_{n,2}$) in the illustrative example, the highest value of the silhouette width was obtained for $\lambda_{n,1}$ (0.82), followed by $b_{n,2}$ (0.79) and $\lambda_{n,2}$ (0.74). When deciding on the number of clusters, We compared the average silhouette widths for the PAM clustering algorithm with 2, 3, 4 and 5 clusters. The silhouette widths for 2, 3 and 4 clusters were very similar (with the surprising 4 cluster result due to an artificial $\lambda_{n,1}$ cluster for curves 15 and 17) and went down for 5 clusters.

4. Simulation study

4.1. Generating the data

The simulated data sets consist of one hundred samples s of each $N = 17$ curves. Each sample s contains three different groups, formed by considering two different types of clusters. One cluster consists of a group of five curves with increased distance between the peaks, another cluster consists of six curves with increased hight 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 $j = 1, \ldots, 100$,

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

with $\phi\left((\mu, \sigma); t\right)$ the density function of the normal distribution with mean μ and variance

σ^2, and for each simulated sample $s = 1, \ldots, 100$:

$$\begin{array}{ccl}
\mu_{1,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(0.1, 0.08, -0.1, 0.3), & n = 1, \ldots, 12, \\
\mu_{1,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(-0.05, 0.05, -0.15, -0.05), & n = 13, \ldots, 17, \\
\mu_{2,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(0.8, 0.01, 0.55, 0.71), & n = 1, \ldots, 12, \\
\mu_{2,n}^{(s)} & = 1, & n = 13, \ldots, 17, \\
a_{1,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(1, 0.2, 0.7, 1.2), & n = 1, \ldots, 17, \\
a_{2,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(0.8, 0.2, 0.7, 0.9), & n = 1, \ldots, 6, 13, \ldots, 17, \\
a_{2,n}^{(s)} & \text{drawn from } \bar{\mathcal{N}}(1.05, 0.2, 1.1, 1.2), & n = 7, \ldots, 12,
\end{array} \tag{4}$$

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 6 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 12 versus curves 13 to 17) or two clusters based on the height of the second peak (curves 1 to 6 and 13 to 17 versus 7 to 12), which means that this is actually a setting with three clusters (curves 1 to 6 versus 7 to 12 versus 13 to 17).

4.2. Functional clustering approaches

We compare the multiresolution clustering method (MRC) with 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). Liu and Yang (2009) introduce cluster membership by means of a model with shifted B-spline basis functions and cluster-specific basis coefficients. This means clusters are characterized by similar patters, 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 $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The random shifts b_i constitute the curve registration aspect of their method.

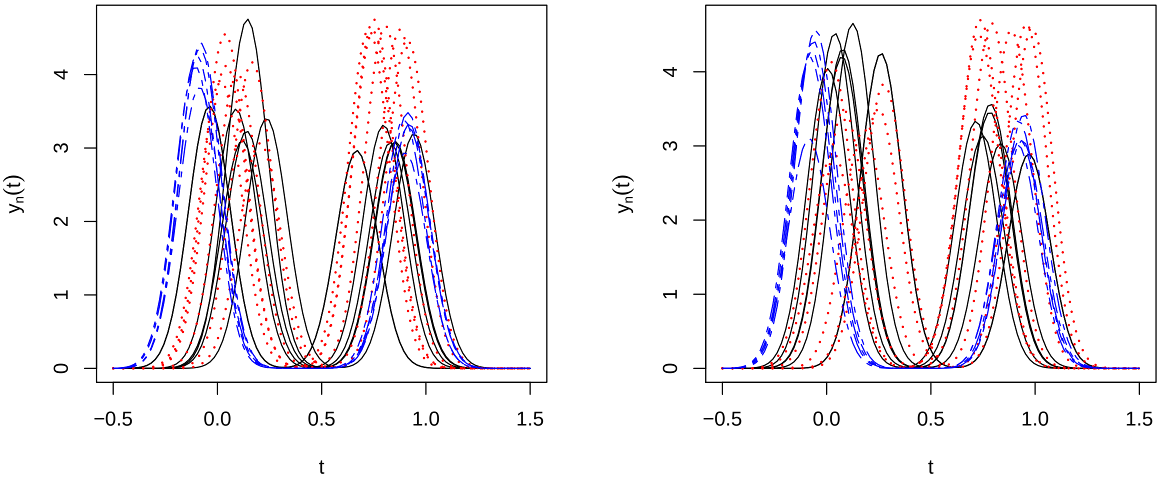

Figure 6: Two examples of simulated data samples in setting 1. Five curves with increased distance between the peaks (blue) dashed lines and 6 curves with increased height of second peak (red) dotted lines.

In Chiou and Li (2007), an advanced functional version of the popular multivariate k-means clustering algorithm is introduced via a truncated Karhunun-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 ($\langle \mathrm{cov}[Y(s), Y(t)], \rho_j^c \rangle = \lambda_j^c \rho_j^c(t)$) 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 Karhunun-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 classification error rate and the percentage of correct classifications in the simulation setting to evaluate the performance of MRC and that of the competitors. The classification error rate is the smallest percentage of cluster membership changes to be made for the clustering result to match the real configuration.

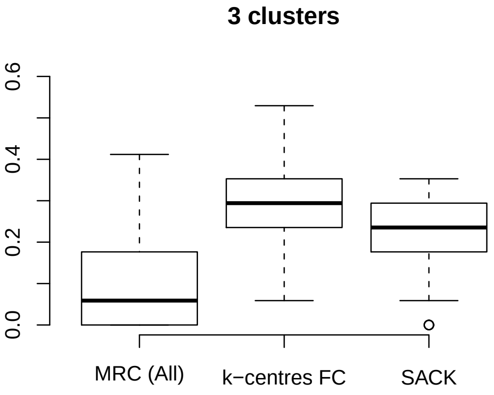

Figure 7: Boxplots of the classification error rate with respect to the three true clusters for resp. MRC based on all variables in the summary matrix (MRC All), k-centers FC and SACK.

We start by searching for three groups in the data. For the MRC method, we use the information of all coefficients. Figure 7 shows boxplots of the classification error rate 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. Similar as in the illustrative data example in Section 5, the k-centers FC method has difficulties with detecting the correct clusters, especially those curves with amplitude differences cannot be separated. In a lesser extend this holds true for the SACK method as well. In the considered simulation setting it is expected that false classifications might occur. Due to a small sample size, chance-induced clusters can arise in a shift or with fitting the first kernel. The k-centers FC method however had not a single entirely correct classification over the simulation runs, the SACK method reached 4% of correct classifications, while the new MRC method taking all variables together was able to correctly cluster the sample of curves in 28% of the simulation runs.

We now 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 five curves for which the average distance between the peaks is larger than for the other curves. The multiresolution warping method contains this information in the intensities of the first warplet $\lambda_{n,1}$. This is confirmed by inspecting the boxplots in Figure 8; clustering based on the $\lambda_{n,1}$ values gives indeed the best classification result within the MRC approaches.

When using the three phase variables (horizontal shifts and intensities of both warplets),

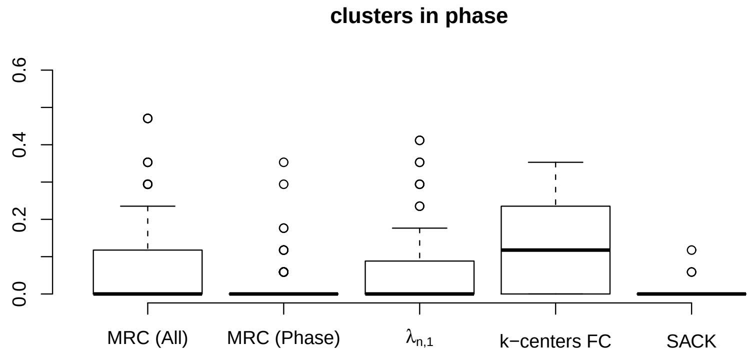

Figure 8: Boxplots of the classification error rate with respect to the two clusters based on the distance between peaks for resp. MRC based on all variables in the summary matrix (all), MRC based on the variables related to phase (phase), MRC based on the intensities of the first warping component only ($\lambda_{n,1}$), k-centers FC and SACK.

an equally good result is obtained, as expected. This analysis does not get disturbed by the distraction provided by the clusters in amplitude (height differences of the second peak). When considering all variables resulting from the MRC procedure, thus also including the estimated amplitude coefficients, we still get a reasonable result when searching for the two clusters defined by phase characteristics, which indicates that the clusters originating from phase differences might dominate the clusters originating from amplitude differences. The simultaneous registration and clustering approach SACK excels here, while the k-centers FC method, which is not designed to handle phase variation, has the worst performance.

In this example it is easier for the methods to detect the clusters in phase, than to detect the clusters in amplitude. The boxplots of the classification error rate in Figure 9 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. For all three methods the median classification error is around 35%.

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. Using the estimated coefficients for the second kernel in the model used for warping as a basis for clustering, results in the lowest classification error rate amongst the considered methods. Including the predicted coefficients for the first kernel, which does not add information in this example, gives a slightly higher classification error rate, but performs still better than the other considered methods.

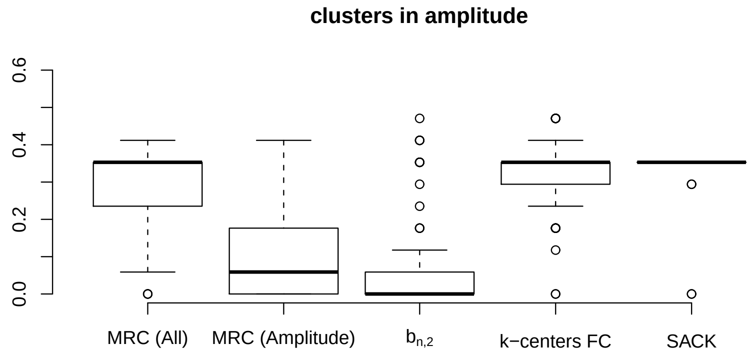

Figure 9: Boxplots of the classification error rate with respect to the two clusters based on the height of the second peak for resp. MRC based on all variables in the summary matrix (all), MRC based on the variables related to amplitude (amplitude), MRC based on the coefficients for the second kernel ($b_{n,2}$), k-centers FC and SACK.

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 10 and 11.

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 10 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

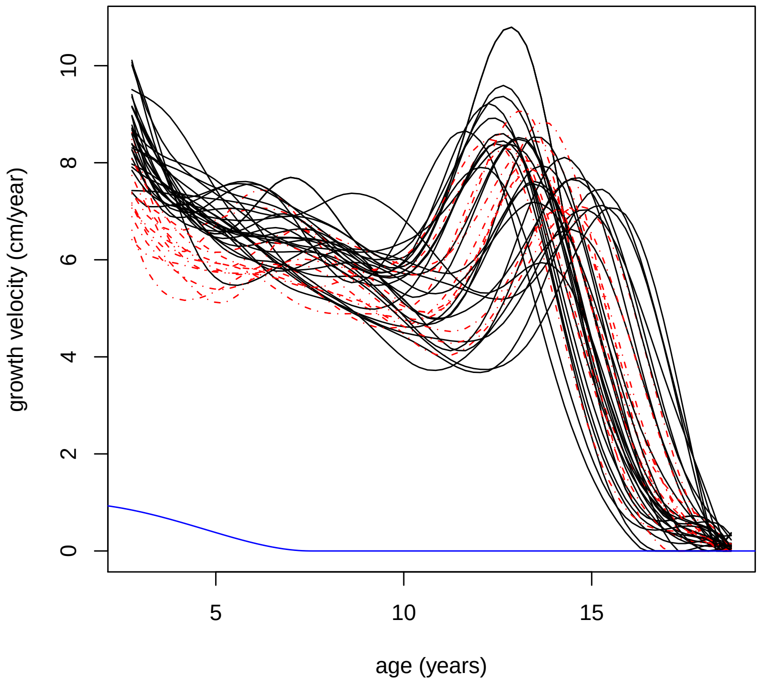

Figure 10: Growth velocity curves for 39 boys, clusters based on kernel 1 only (red dashed lines) and kernel 1 (blue line).

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 11 (a)) and the warping component (Figure 11 (b)). 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 12 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.

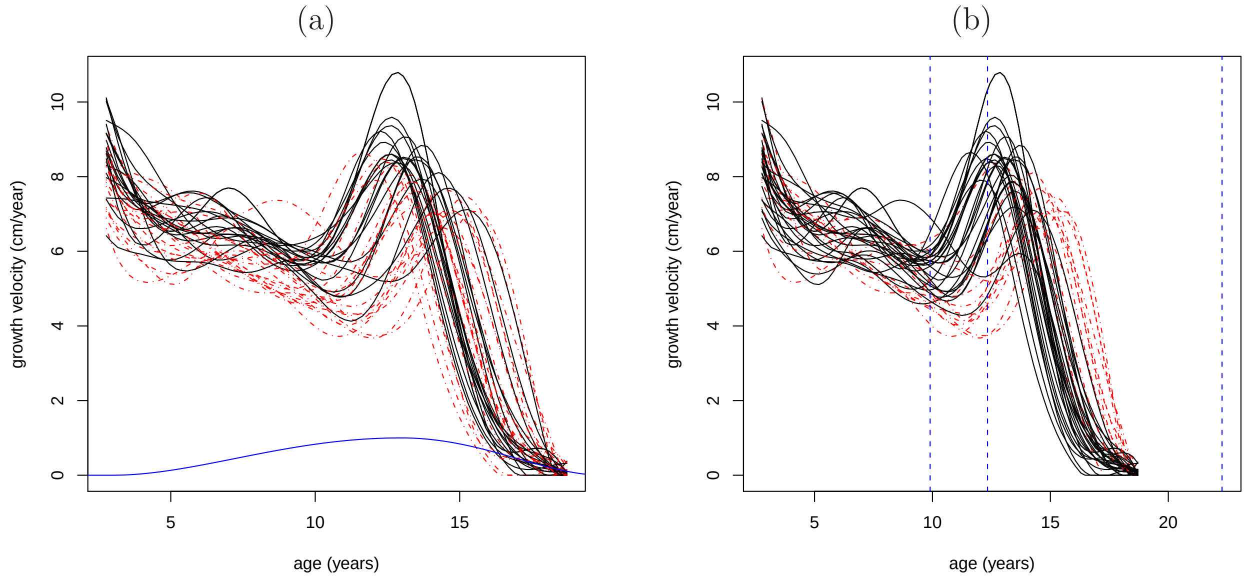

Figure 11: Growth velocity curves for 39 boys. (a) Clusters based on kernel 2 only (red, dashed) and the estimated kernel function (blue). (b) Clusters based on the warplet only (red, dashed), together with the warping bounds and warping center (blue, dashed).

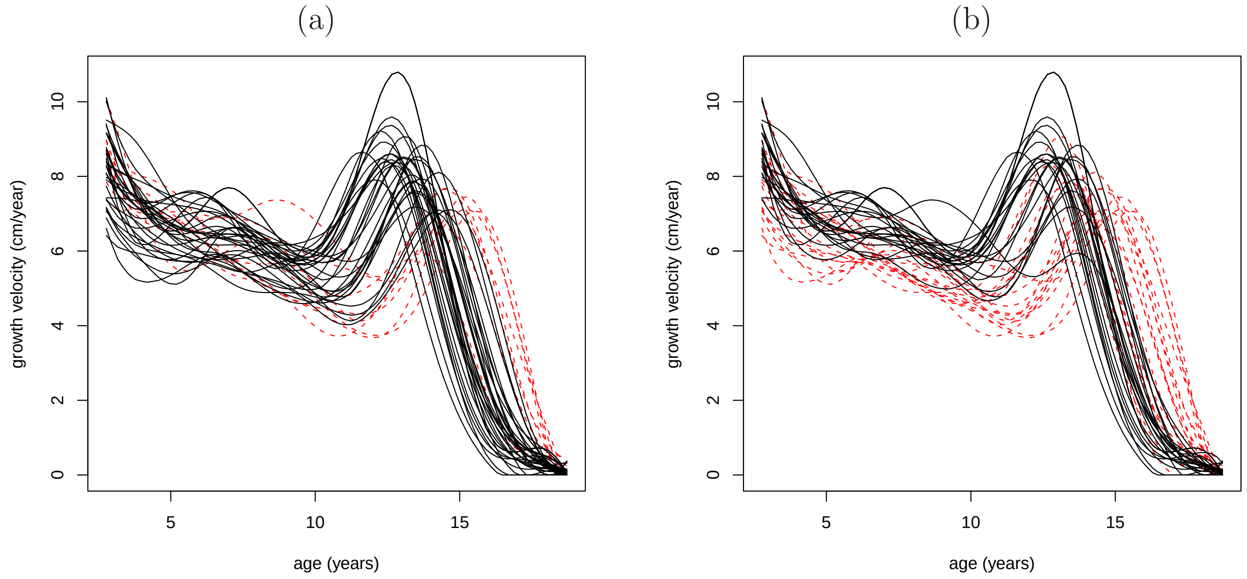

Figure 12: Growth velocity curves for 39 boys. (a) Clusters based SACK model (red, dashed). (b) Clusters based k-centers FC (red, dashed).

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.

Appendix A. R code for the multiresolution clustering

The R library MRwarp 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 full in the file MRWclust.examplecode.R, where also the function kernelcoeff is made available.

```
library(MRwarp); library(SemiPar); library(cluster); library(pls)

# multiresolution warping step

output <- MRwarp(Xdata,Ydata,chain=1500,thin=5,burnin=1/3,kernel.s=c(-0.2,0.2,0.5,0.6,0.9,1.2),
components=2,selection="FIXED",thresh=0.8,threshd=1/20,prepr=c(1,0,0,0),outputfit=0)
```

```
parsk = as.matrix(read.table("amcmcker2.txt",header=FALSE))
K <- parsk[output$index[1],]

ll <- min(Xdata);   ul <- max(Xdata)
S <- dim(Xdata)[1]; N <- dim(Xdata)[2]

A <- output$A
Ll <- output$Ll
Ul <- output$Ul
Lambda <- output$Lambda
shift.w <- output$shift.w

valvec <- matrix(0,S,1+length(K)/3+dim(Lambda)[2]) #data summary matrix

# shift
valvec[,1] <- shift.w
Wx <- Xdata + shift.w

# warplets
valvec[,2:(1+dim(Lambda)[2])] <- Lambda
for (ii in 1:2)
    {for (j in 1:S)
     {Wx[j,] <- warp(A[ii],Lambda[j,ii],A[ii]-Ll[ii],Ul[ii]-A[ii],Wx[j,]) }
    }

#kernels
kc <- kernelcoeff(Wx,Ydata,K,ll,ul,maxiter=50,prec=0.01)
valvec[,(2+dim(Lambda)[2]):(1+dim(Lambda)[2]+length(K)/3)] <- kc$coeff

#standardize the summary matrix
valvec.st <- stdize(valvec)

#perform PAM procedure
pam(valvec.st[,1:3],2)$cluster
pam(valvec.st[,5],2)$cluster
pam(valvec.st[,1:5],3)$cluster
```

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.

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.

DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for clusterwize linear regression. *Journal of Classification*, 5:249–282.

Garcia-Escudero, L. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, 22:185–201.

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.

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.

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, 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 Estadstica y Econometra.

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.

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, 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. *Chemometrics 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.