Model based clustering of functional data with mild outliers

Cristina Anton and Iain Smith

Abstract We propose a procedure, called CFunHDDC, for clustering functional data with mild outliers which combines two existing clustering methods: the functional high dimensional data clustering (FunHDDC) [1] and the contaminated normal mixture (CNmixt) [3] method for multivariate data. We adapt the FunHDDC approach to data with mild outliers by considering a mixture of multivariate contaminated normal distributions. To fit the functional data in group-specific functional subspaces we extend the parsimonious models considered in FunHDDC, and we estimate the model parameters using an expectation-conditional maximization algorithm (ECM). The performance of the proposed method is illustrated for simulated and real-world functional data, and CFunHDDC outperforms FunHDDC when applied to functional data with outliers.

Keywords: Functional data, Model-based clustering, Contaminated normal distributions, EM algorithm

1 Introduction

Recently, model-based clustering for functional data has received a lot of attention. Real data are often contaminated by outliers that affect the estimations of the model parameters. Here we propose a method for clustering functional data with mild outliers. Mild outliers are usually sampled from a population different from the

Cristina Anton
MacEwan University, 10700 – 104 Avenue Edmonton, AB, T5J 4S2, Canada e-mail: popescuc@macewan.ca

Iain Smith
MacEwan University, 10700 – 104 Avenue Edmonton, AB, T5J 4S2,Canada e-mail: smithi23@mymacewan.ca

assumed model, so we need to choose a model flexible enough to accommodate them.

Functional data live in an infinite dimensional space and model-based methods for clustering are not directly available because the notion of probability density function generally does not exist for such data. A first approach is to use a two-step method and first do a discretization or a decomposition of the functional data in a basis of functions (such as Fourier series, B-splines, etc.), and then directly apply multivariate clustering methods to the discretization or the basis coefficients. A second approach, which allows the interaction between the discretization and the clustering steps, is based on a probabilistic model for the basis coefficients [1], [2].

We follow the second approach, and we propose a method, called CFunHDDC, which extends the functional high dimensional data clustering (FunHDDC) [1] to clustering functional data with mild outliers. There are several methods to detect outliers of functional data and a robust clustering methodology based on trimming is presented in [4]. Our approach does not involve trimming the outliers and it is inspired by the method CNmixt [3] for clustering multivariate data with mild outliers. We propose a model for the basis coefficients based on a mixture of contaminated multivariate normal distributions. A multivariate contaminated normal distribution is a two-component normal mixture in which the bad observations (outliers) are represented by a component with a small prior probability and an inflated covariance matrix.

In the next section we present the model and its parsimonious variants. Parameter estimation is included in section 3. In section 4 we present applications to simulated and real-world data. The last section includes the conclusions.

2 The model

We suppose that we observe n curves $\{x_1, \ldots, x_n\}$ and we want to cluster them in K homogeneous groups. For each curve x_i we have access to a finite set of values $x_{ij} = x_i(t_{ij})$, where $0 \leq t_{i1} < t_{i2} < \ldots < t_{im_i} \leq T$. We assume that the observed curves are independent realizations of a L^2- continuous stochastic process $X = \{X(t)\}_{t\in[0,T]}$ for which the sample paths are in $L^2[0,T]$. To reconstruct the functional form of the data we assume that the curves belong to a finite dimensional space spanned by a basis of functions $\{\xi_1, \ldots, \xi_p\}$, so we have the expansion for each curve

$$x_i(t) = \sum_{j=1}^{p} \gamma_{ij}\xi_j(t).$$

Here we assume that the dimension p is fixed and known. We consider a model based on a mixture of multivariate contaminated normal distributions for the coefficients vectors $\{\gamma_1, \ldots, \gamma_n\} \subset \mathbb{R}^p$, $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{ip})^\top \in \mathbb{R}^p$, $i = 1, \ldots, n$.

We suppose that there exists two unobserved random variables $Z = (Z_1, \ldots, Z_K)$, $\Upsilon = (\Upsilon_1, \ldots, \Upsilon_K) \in \{0,1\}^K$ where Z indicates the cluster membership and Υ

whether an observation is good or bad (outlier). $Z_k = 1$ if $X \in k$th cluster and $Z_k = 0$ otherwise, and $\Upsilon_k = 1$ if $X \in k$th cluster and it is a good observation, and $\Upsilon_k = 0$ otherwise. For clustering we need to predict the value $z_i = (z_{i1}, \ldots, z_{iK})$ of Z, and to determine the bad observations we need to predict the value $\nu_i = (\nu_{i1}, \ldots, \nu_{iK})$ of Υ for each observed curve x_i, $i = 1, \ldots, n$.

We consider a set of n_k observed curves of the kth cluster with the coefficients $\{\gamma_1, \ldots, \gamma_{n_k}\} \subset \mathbb{R}^p$. We assume that $\{\gamma_1, \ldots, \gamma_{n_k}\}$ are independent realizations of a random vector $\Gamma \in \mathbb{R}^p$, and that the stochastic process associated with the kth cluster can be described in a lower dimensional subspace $\mathbb{E}^k[0,T] \subset L^2[0,T]$ with dimension $d_k \leq p$ and spanned by the first d_k elements of a group specific basis of functions $\{\phi_{kj}\}_{j=1,\ldots,d_k}$ that can be obtained from $\{\xi_j\}_{j=1,\ldots,p}$ by a linear transformation

$$\phi_{kj} = \sum_{l=1}^{p} q_{k,jl}\xi_l,$$

with an $p \times p$ orthogonal matrix $Q_k = (q_{k,jl})$. In [1] for FunHDDC the assumption is that the distribution of Γ for the kth cluster is $\Gamma \sim N(\mu_k, \Sigma_k)$, $\Sigma_k = Q_k \Delta_k Q_k^\top$, where

$$\Delta_k = \left(\begin{array}{ccc|ccc} a_{k1} & & 0 & & & \\ & \ddots & & & \mathbf{0} & \\ 0 & & a_{kd_k} & & & \end{array}\right\}\!\! p \quad \left.\begin{array}{ccc} & & \\ & & \\ b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{array}\right)$$

with $a_{ki} > b_k$, $i = 1, \ldots, d_k$. We can say that the variance of the actual data in the kth cluster is modeled by $a_{k1}, \ldots, a_{kd_k}$ and the parameter b_k models the variance of the noise [1].

We follow the approach in [3] and we assume that Γ for the kth cluster has the multivariate contaminated normal distribution with density

$$f(\gamma_i; \theta_k) = \alpha_k \phi(\gamma_i; \mu_k, \Sigma_k) + (1 - \alpha_k)\phi(\gamma_i; \mu_k, \eta_k \Sigma_k), \tag{1}$$

where $\alpha_k \in (0.5, 1)$, $\eta_k > 1$, $\theta_k = \{\alpha_k, \mu_k, \Sigma_k, \eta_k\}$, and $\phi(\gamma_i; \mu_k, \Sigma_k)$ is the density for the $p-$variate normal distribution $N(\mu_k, \Sigma_k)$:

$$\phi(\gamma_i; \mu_k, \Sigma_k) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(\gamma_i - \mu_k)^\top \Sigma_k^{-1}(\gamma_i - \mu_k)\right) \tag{2}$$

Here α_k defines the proportion of uncontaminated data in the $k$the cluster and η_k represents the degree of contamination. We can see η_k as an inflation parameter that measures the increase in variability due to the bad observations.

Each curve x_i has a basis expansion with coefficient γ_i such that γ_i is a random vector whose distributions is a mixture of contaminated Gaussians with density

$$p(\gamma;\theta)=\sum_{k=1}^{K}\pi_k f(\gamma;\theta_k) \tag{3}$$

where $\pi_k = P(Z_k = 1)$ is the prior probability of the kth the cluster and $\theta = \bigcup_{k=1}^{k}(\theta_k \cup \{\pi_k\})$ is the set formed by all the parameters. We refer to this model as FCLM$[a_{kj}, b_k, Q_k, d_k]$ (functional contaminated latent mixture). As in [1] we consider the parsimonious sub-models: FCLM$[a_{kj}, b, Q_k, d_k]$, FCLM$[a_k, b_k, Q_k, d_k]$, FCLM$[a, b_k, Q_k, d_k]$, FCLM$[a_k, b, Q_k, d_k]$, FCLM$[a, b, Q_k, d_k]$.

3 Model inference

To fit the models we use the ECM algorithm [3], which is a variant of the EM algorithm. In the ECM algorithm we replace the M-step in the EM algorithm by two simpler CM-steps given by the partition of the set with the parameters $\theta = \{\Psi_1, \Psi_2\}$, where $\Psi_1 = \{\pi_k, \alpha_k, \mu_k, a_{kj}, b_k, q_{kj}, k = 1,\ldots,K, j = 1,\ldots,d_k\}$, $\Psi_2 = \{\eta_k, k = 1,\ldots,K\}$, and q_{kj} is the jth column of Q_k.

We have two sources of missing data: the clusters' labels and the type of observation (good or bad). Thus the complete data are given by $S = \{\gamma_i, z_i, \nu_i\}_{i=1,\ldots,n}$, and the complete-data likelihood is

$$L_c(\theta;S)=\prod_{i=1}^{N}\prod_{k=1}^{K}\left\{\pi_k\left[\alpha_k\phi(\gamma_i;\mu_k,\Sigma_k)\right]^{\nu_{ik}}\left[(1-\alpha_k)\phi(\gamma_i;\mu_k,\eta_k\Sigma_k)\right]^{1-\nu_{ik}}\right\}^{z_{ik}}$$

We denote the complete-data log-likelihood by $l_c(\theta;S) = \log(L_c(\theta;S))$.

Next we present the ECM algorithm for the model FCLM$[a_{kj}, b_k, Q_k, d_k]$. At the q iteration of the ECM algorithm in the E-step we calculate $E[l_c(\theta^{(q-1)};S)|\gamma_1, \ldots,\gamma_n,\theta^{(q-1)}]$, given the current values of the parameters $\theta^{(q-1)}$. This reduces to the calculation of $z_{ik}^{(q)} := E[Z_{ik}|\gamma_i,\theta^{(q-1)}]$, $\nu_{ik}^{(q)} := E[\Upsilon_{ik}|\gamma_i, z_i,\theta^{(q-1)}]$.

In the first CM step in the q iteration of the ECM algorithm we calculate $\Psi_1^{(q)}$ as the value of Ψ_1 that maximize $l_c^{(q-1)}$ with Ψ_2 fixed at $\Psi_2^{(q-1)}$. We obtain

$$\pi_k^{(q)}=\frac{\sum_{i=1}^{n} z_{ik}^{(q)}}{n},\quad \alpha_k^{(q)}=\frac{\sum_{i=1}^{n} z_{ik}^{(q)}\nu_{ik}^{(q)}}{\sum_{i=1}^{n} z_{ik}^{(q)}},\quad \mu_k^{(q)}=\frac{\sum_{i=1}^{n} z_{ik}^{(q)}\left(\nu_{ik}^{(q)}+\frac{1-\nu_{ik}^{(q)}}{\eta_k^{(q-1)}}\right)\gamma_i}{\sum_{i=1}^{n} z_{ik}^{(q)}\left(\nu_{ik}^{(q)}+\frac{1-\nu_{ik}^{(q)}}{\eta_k^{(q-1)}}\right)} \tag{4}$$

$$\Sigma_k^{(q)}=\frac{1}{\sum_{i=1}^{n} z_{ik}^{(q)}}\sum_{i=1}^{n} z_{ik}^{(q)}\left(\nu_{ik}^{(q)}+\frac{1-\nu_{ik}^{(q)}}{\eta_k^{(q-1)}}\right)(\gamma_i-\mu_k^{(q)})(\gamma_i-\mu_k^{(q)})^{\top} \tag{5}$$

We introduce a value α^* and we constrain $\alpha_k \in (\alpha^*, 1)$. If the estimation $\alpha_k^{(q)}$ in (4) is less than α^*, we use the *optimize*() function in the *stats* package in R to do a numerical search for $\alpha_k^{(q)}$.

As in [1] we get the updated values $a_{kj}^{(q)}, b_k^{(q)}, q_{kj}^{(q)}, k = 1, \ldots, K, j = 1, \ldots, d_k$ from the sample covariance matrix $\Sigma_k^{(q)}$ of cluster k, using also the matrix of inner products between the basis functions $W = (w_{jl})_{1 \le j,l \le p}$, where $w_{jl} = \int_0^T \xi_j(t)\xi_l(t)dt$.

In the second CM step in the q iteration of the ECM algorithm we calculate $\eta_k^{(q)}$ as the value that maximize $l_c^{(q-1)}$ with Ψ_1 fixed at $\Psi_1^{(q)}$.

At the end of the ECM algorithm, we do a two-step classification to provide the expected clustering. If q_f is the last iteration of the algorithm before convergence, an observation $\gamma_i \in \mathbb{R}^p$ is assigned to the cluster $k_0 \in \{1, \ldots, K\}$ with the largest $z_{ik}^{(q_f)}$. Next, an observation γ_i that was assigned to the cluster k_0 is considered good if $v_{ik_0}^{(q_f)} > 0.5$, and it is considered bad otherwise. After the classification step we can eliminate the bad observations and run FunHDDC to re-cluster the remaining observations.

The class specific dimension d_k is selected through the scree-test of Cattell by comparison of the difference between eigenvalues with a given threshold [1]. The number of clusters K as well as the parsimonious model are selected using the BIC criterion.

4 Applications

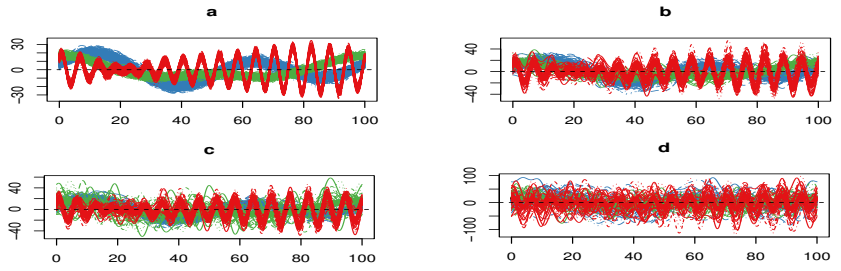

Fig. 1 Smooth data simulated without oultiers (a), according to scenario A (b), scenarion B (c), and scenario C (d), coloured by group for one simulation.

We simulate 1000 curves based on the model FCLM$[a_k, b_k, Q_k, d_k]$. The number of clusters is fixed to $K = 3$ and the mixing proportions are equal $\pi_1 = \pi_2 = \pi_3 = 1/3$. We consider the following values of the parameters

Group 1: $d = 5$, $a = 150$, $b = 5$, $\mu = (1, 0, 50, 100, 0, \ldots, 0)$
Group 2: $d = 20$, $a = 15$, $b = 8$, $\mu = (0, 0, 80, 0, 40, 2, 0, \ldots, 0)$
Group 3: $d = 10$, $a = 30$, $b = 10$, $\mu = (0, \ldots, 0, 20, 0, 80, 0, 0, 100)$,

where d is the intrinsic dimension of the subgroups, μ is the mean vector of size 70, a is the value of the d-first diagonal elements of Δ, and b the value of the $70 - d$- last ones. Curves are smoothed using 35 Fourier basis functions. We repeat the simulation 100 times. A sample of theses data is plotted in Figure 1 a. We consider the following contamination schemes where the scores are simulated from contaminated normal distributions with the previous parameters and

A: $\alpha_i = 0.9$, $i = 1, \ldots, 3$, and $\eta_1 = 7$, $\eta_2 = 10$, $\eta_3 = 17$.
B: $\alpha_i = 0.9$, $i = 1, \ldots, 3$, and $\eta_1 = 5$, $\eta_2 = 50$, $\eta_3 = 15$.
C: $\alpha_i = 0.9$, $i = 1, \ldots, 3$, and $\eta_1 = 100$, $\eta_2 = 70$, $\eta_3 = 170$.

Samples for data generated according to scenarios A, B, C are plotted in Figure 1 b, c, d, respectively. We notice that there is more overlapping between the 3 groups when we increase the values of η.

Table 1 Mean (and standard deviation) of ARI for BIC best model on 100 simulations. Bold values indicates the highest value for each method.

Scenario	Method	α^*	ϵ	ARI	ARI Outliers
A	FunHDDC	-	0.05	**0.519 (0.11)**	-
A	FunHDDC	-	0.1	0.499(0.05)	-
A	FunHDDC	-	0.2	0.494 (0.01)	-
A	CFunHDDC	0.75	0.05	0.769 (0.23)	0.959(0.04)
A	CFunHDDC	0.75	0.1	0.986(0.08)	0.998(0.01)
A	CFunHDDC	0.75	0.2	**0.9995 (0.001)**	**1 (0)**
B	FunHDDC	-	0.05	**0.861 (0.23)**	-
B	FunHDDC	-	0.1	0.754(0.25)	-
B	FunHDDC	-	0.2	0.52 (0.09)	-
B	CFunHDDC	0.75	0.05	0.807 (0.22)	0.961(0.05)
B	CFunHDDC	0.75	0.1	0.948 (0.14)	0.99(0.03)
B	CFunHDDC	0.75	0.2	**0.990 (0.062)**	**0.971 (0.149)**
C	FunHDDC	-	0.05	0.490 (0.02)	-
C	FunHDDC	-	0.1	0.491(0.02)	-
C	FunHDDC	-	0.2	**0.494 (0.01)**	-
C	CFunHDDC	0.75	0.05	0.736 (0.23)	0.928(0.10)
C	CFunHDDC	0.75	0.1	0.911 (0.18)	0.958(0.15)
C	CFunHDDC	0.75	0.2	**0.965 (0.11)**	**0.994 (0.03)**

The quality of the estimated partitions obtained using FunHDDC and CFunHDDC is evaluated using the Adjusted Rand Index (ARI) [3], and the results are included in Table 1. For FunHDDC we use the library *funHDDC* in R. We run both algorithms for $K = 3$ with all 6 sub-models and the best solution in terms of the highest BIC

Table 2 Correct classification rates for each method.

Method	ϵ	CCR	Method	α^*	ϵ	CCR	Method	α^*	CCR
FunHDDC	0.01	0.68	CFunHDDC	0.85	0.01	0.67	CNmixt	0.5	0.67
FunHDDC	0.05	0.64	CFunHDDC	0.85	0.05	0.70	CNmixt	0.75	0.66
FunHDDC	0.1	0.59	CFunHDDC	0.85	0.1	0.70	CNmixt	0.85	0.67
FunHDDC	0.2	0.57	CFunHDDC	0.85	0.2	0.6	CNmixt	0.9	0.66

value for all those submodels is returned. The initialization is done with the *kmeans* strategy with 50 repetitions, and the maximum number of iterations is 200 for the stopping criterion. We use $\epsilon \in \{0.05, 0.1, 0.2\}$ in the Cattell test.

We notice that CFunHDDC outperforms FunHDDC, and it gives excellent results even in Scenario C. For CFunHDDC the best results are obtained for $\epsilon = 0.2$ in the Catell test, and the values of the ARI are close to 1.

Next, we consider the NOx data available in the *fda.usc* library in R and representing daily curves of Nitrogen Oxides (NOx) emissions in the neighborhood of the industrial area of Poblenou, Barcelona (Spain). The measurements of NOx (in $\mu g/m^3$) were taken hourly resulting in 76 curves for "working days" and 39 curves for "non-working days" (see Figure 2 a). Since NOx is a contaminant agent, the detection of outlying emission is useful for environmental protection. This data set has been used for testing methods for the detection of outliers and to illustrate robust clustering based on trimming for functional data [4].

We apply CFunHDDC, FunHDDC, and CNmixt to the NOx data. Curves are smoothed using a basis of 8 Fourier functions, and we run the algorithms for $K = 2$ clusters. For CFunHDDC, FunHDDC we use $\epsilon \in \{0.001, 0.05, 0.1, 0.2\}$ in the Cattell test and the rest of the settings are the same as in the simulation study. We run CNmixt for all 14 models from the *ContaminatedMixt* R library, based on the coefficients in the Fourier basis, with 1000 iterations for the stopping criteria, and initialization done with the *kmeans* method. The correct classification rates (CCR) are reported in Table 2.

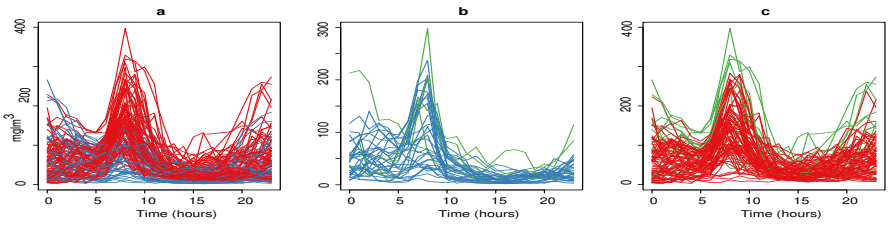

Fig. 2 a.Daily NOx curves for 115 days; b. c. Clustering obtained with CFunHDDC, $\epsilon = 0, 05, \alpha^* = 0.85$; Non-working days (blue), working days (red), outliers (green)

The CCR for CFunHDDC are slightly better than the ones for FunHDDC and CNmixt, and are comparable with the ones reported in Table 1 in [4] for Funclust, RFC, and TrimK. In Figure 2 b, c we present the clusters and the detected outliers for $\epsilon = 0.05$ and $\alpha^* = 0.85$. The curves that are detected as outliers (green lines) exhibit different patterns from the rest of the curves.

One of the advantages of extending the FunHDDC to CFunHDDC is the outlier detection. For $\alpha^* = 0.85$ and $\epsilon = 0.05$, CFunHDDC detects 16 outliers, which are the same with the outliers mentioned in [4]. For the data without outliers, CFunHDDC becomes equivalent to FunHDDC, and for the trimmed data the CCR increases to 0.79.

5 Conclusion

We propose a new method, CFunHDDC, that extends the FunHDDC functional clustering method to data with mild outliers. Unlike other robust functional clustering algorithms, CFunHDDC does not involve trimming the data. CFunHDDC is based on a model formed by a mixture of contaminated multivariate normal distributions, which makes parameter estimation more difficult than for FunHDDC, so we use an ECM instead of an EM algorithm. The clustering and outlier detection performance of CFunHDDC is tested for simulated data and the NOx data and it always outperforms FunHDDC. Moreover, CFunHDDC has a comparable performance with robust functional clustering methods based on trimming, such as RFC and TrimK, and it has similar or better performance when compared to a two-step method based on CNmixt. Although there are several model-based methods for multivariate data with outliers that can be used to construct two-step methods for functional data, as observed in [1], these two-step methods always suffers from the difficulty to choose the best discretization. CFunHDDC can be extended to multivariate functional data, and recently, independently of our work, a similar approach was followed in [5], but without considering the parsimonious models and the value α^*.

References

1. Bouveyron, C., Jacques, J.: Model-based clustering of time series in group-specific functional subspaces. Adv. Data. Anal. Classif. **5**(4), 281–300 (2011)
2. Jacques, J., Preda, C.: Funclust: a curves clustering method using functional random variables density approximation. Neurocomputing **112**, 164–171 (2013)
3. Punzo, A., McNicholas, P.D.: Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J. **58**, 1506–1537 (2016)
4. Rivera-Garcia, D., Garcia-Escudero, L.A., Mayo-Iscar, A., Ortega, J.: Robust clustering for functional data based on trimming and constraints. Adv. Data Anal. Classif. **13**, 201–225 (2019)
5. Amovin-Assagba, M., Gannaz, I., Jacques, J.: Outlier detection in multivariate functional data through a contaminated mixture model. (2021) https://doi.org/10.48550/arXiv.2106.07222