Cluster analysis with cellwise trimming and applications to robust clustering of curves

L.A. García-Escudero[a], D. Rivera-García[b], A. Mayo-Iscar[a], J. Ortega[c,*]

[a] *Universidad de Valladolid, Valladolid, España*
[b] *Centro de Investigación Coppel, Ciudad de México, México*
[c] *King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

Abstract

In this work, we propose a robust Cluster Analysis methodology based on cell trimming as an extension to a recently introduced robust version of Principal Component Analysis. This new approach allows for cellwise trimming in cluster analysis, which is more reasonable than traditional casewise trimming when the problem's dimension is large. This type of trimming avoids an unnecessary loss of information when only a few cells of the entirely trimmed observations are atypical. An algorithm is proposed to apply this approach. This algorithm is particularized to the interesting case of functional cluster analysis. Simulations and applications to real data sets are given to illustrate the proposed methods.

1. Introduction

Given a data set $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ with $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$, trimming complete observations $\boldsymbol{x}_i$ (row trimming) is the approach traditionally followed in Robust Statistics. The idea is to trim entire rows $\boldsymbol{x}_i$ as long as these rows include at least some contaminated cells x_{ij}. This type of trimming is reasonable in low dimensions but can be extreme when the dimension p grows. A very small fraction of the contaminant cells (evenly distributed) when p grows makes it necessary to trim a vast number of rows $\boldsymbol{x}_i$. This massive trimming leads to discarding much useful information in non-atypical cells, which translates into a considerable loss of efficiency. Besides, many tools in Robust Statistics are not designed to handle trimming levels greater than 50 percent of observations. These high-dimensional problems are becoming more frequent in modern statistics because of the ease of recording and storing large volumes of data that new technologies provide. Unfortunately, the presence of outliers is often the rule in most data sets, and robust methods need to be implemented that can cope with these outliers. The first work where this problem was formally addressed was [1]. Recent references of interest and possible solutions to the problem are [47], [41] and [29].

The above proposals work without assuming substructures or clusters in the data. The presence of subgroups is frequent, given the common heterogeneity that appears in many data sets. Our purpose is to introduce a robust clustering methodology that would allow for cellwise trimming. The clusters to be searched are assumed to 'live' in G related subspaces of smaller dimensions than the original space (subspace clustering). This philosophy is precisely what underlies the $G = 1$ case when applying the well known Principal Component Analysis (PCA) method for

*Corresponding author
Email addresses: lagarcia@eio.uve.es (L.A. García-Escudero), diego.rivera@coppel.com (D. Rivera-García), agustinm@eio.uva.es (A. Mayo-Iscar), joaquin.ortegasanchez@kaust.edu.sa (J. Ortega)

dimension reduction. Different proposals for the robustification of PCA, together with a comparative study of them, can be found in [15]. Among these proposals, it is worth mentioning the one obtained through (impartial) trimming in [35], which was also analyzed at the theoretical level in [24].

Subspace clustering is not new and is behind various Cluster Analysis procedures designed especially for high-dimensional problems. Interesting references in this line, including review papers, are [36], [34], [44], and [6]. However, none of these procedures are directly designed to cope with the presence of outliers.

A robust clustering procedure based on trimmings of whole observations using subspaces was proposed in [25] as a robustification of the 'linear grouping algorithm' method in [43]. [25] assumed that the intrinsic dimensions q_g were common to all the approximating subspaces and equal to $q_g = p - 1$, but it is not difficult to extend the algorithm proposed there to the case of different intrinsic dimensions q_g by groups.

Cellwise trimming in Cluster Analysis was already proposed through a modification of the trimmed k-means [13] in [16] ('snipped' in its terminology), and through modification of the TCLUST [21] in [17]. The first approach is based on groupings around centroids and does not take advantage of the structure of dependence between variables, and the second is challenging to apply in dimensions that are not very low since its complexity increases notably with dimension.

In this work, we will propose a methodology of cellwise trimming that will take into account the subspace structure in the G groups and has a feasible algorithm for its implementation. The algorithm is based on alternate regressions with weights, extending the proposals in [3] and [8] for robust PCA. A short version of the Least Trimmed Squares (LTS) algorithm based on 'concentration steps' plays a significant role in the proposed algorithm.

Subsequently, the methodology proposed is applied to the important case of the Functional Cluster Analysis, focusing on providing appropriate initializations to the procedure. The problem of Functional Cluster Analysis is receiving considerable attention, as can be seen in the review papers in [33], [27], and [46]. Given the infinite dimension that underlies these problems, it is critical to assume that clusters of curves are mostly arranged around finite-dimensional functional subspaces. This is the approach that has been adopted in works such as [10], [5], [26], [45], [32], and [18].

Obviously, in the setting of Functional Cluster Analysis, it is also important to have robust procedures that are not significantly affected by functional outliers. The possibility of enhancing robustness by trimming complete functions was considered in [19], after projecting the curves in the space generated by a B-spline basis, and [12] proposed the use of trimmings working directly in a functional $\mathcal{L}_2$ space. These two trimming approaches looked for groups around 'functional centroids' and were not designed to handle dispersion structures within complex groups not easily recognizable by an $\mathcal{L}_2$ norm. The trimming of whole atypical curves for more complex patterns was considered in [38], using 'model-based' methods arising from the 'pseudo-density' for functional data introduced in [14]. Trimming only atypical 'chunks' of the curve (analogous to cellwise trimming for the functional case) was noted as an interesting line of work in [23].

The rest of this work is organized as follows. We start considering the simplest case of PCA (G = 1) to facilitate the presentation of the methodology and introduce the necessary notation. Thus, we will briefly review some proposals for the robustification of the PCA using trimmings in Section 2. Subsequently, we will introduce the methodology proposed for Robust Cluster Analysis ($G > 1$) in Section 3, together with a feasible algorithm for its application

in Section 4. Section 5 presents the adaptation of the method to the functional data case. Some examples of its applicability and a simulation study will be shown in Section 6, together with real examples of practical application in Section 7. Finally, conclusions and possible open research problems will be provided in Section 8.

2. Robust Principal Components

Principal Component Analysis (PCA) aims to obtain $q \leq p$ unitary orthogonal vectors generating the linear subspace that provides the best approximation to a data set. This approximation will be obtained from a matrix $\boldsymbol{B}_q \in \mathbb{R}^{p\times q}$ ($\mathbb{R}^{a\times b}$ denotes the matrices with a rows and b columns) with $\boldsymbol{B}_q^T\boldsymbol{B}_q = I_q$ (orthogonal) and whose rows we will denote by $\boldsymbol{b}_j^T$ for $j = 1, ..., p$. The matrix $\boldsymbol{A}_q \in \mathbb{R}^{n\times q}$ contains the so-called 'scores' and the rows of this matrix are given by the vectors $\boldsymbol{a}_i$. Finally, $\boldsymbol{m}$ will be a vector in $\mathbb{R}^p$. The approximation to the observation $\boldsymbol{x}_i$ in the approximating subspace is written as:

$$\widehat{\boldsymbol{x}}_i := \widehat{\boldsymbol{x}}_i(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = \boldsymbol{m} + \boldsymbol{B}_q\boldsymbol{a}_i,$$

or, working by cells, with $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, ..., \widehat{x}_{ip})'$ where:

$$\widehat{x}_{ij} = m_j + \boldsymbol{a}_i^T\boldsymbol{b}_j. \tag{2.1}$$

With this notation, the problem of finding the best PCA approximating subspace is formally posed by minimizing:

$$\min_{\boldsymbol{B}_q, \boldsymbol{A}_q, m} \sum_{i=1}^{n} d_i^2(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}), \tag{2.2}$$

for

$$d_i^2(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = ||\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m})||^2,$$

over all possible orthogonal matrices $\boldsymbol{B}_q \in \mathbb{R}^{p\times q}$, matrices $\boldsymbol{A}_q \in \mathbb{R}^{n\times q}$, and vectors $\boldsymbol{m} \in \mathbb{R}^p$. This minimization can be posed in terms of cells, through the minimization of

$$\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^2,$$

where $r_{ij} = x_{ij} - \hat{x}_{ij}$ for $\hat{x}_{ij}$ defined as in (2.1).

If $\bar{\boldsymbol{x}}$ and $\mathcal{S}$ are, respectively, the sample mean and the sample covariance matrix then it is well known that the solution to the PCA problem is obtained by considering $\boldsymbol{m} = \overline{\boldsymbol{x}}$ and $\boldsymbol{B}_q$ including the eigenvectors of $\mathcal{S}$ associated with the q largest eigenvalues.

Unfortunately, classical PCA is quite sensitive to outliers. This fact is not surprising given that $\overline{\boldsymbol{x}}$ and $\mathcal{S}$ are very non-robust estimators. A single outlier x_{ij} in one of the observations can already have a very detrimental effect on determining the optimal subspace when using the classical PCA. Numerous robust proposals have been put forward in the literature to solve this problem (see, for example, [15]).

Among these robust proposals for the PCA, an LTS (Least Trimmed Squares) approach based on cutting a proportion α of observations $\boldsymbol{x}_i$ was proposed in [35]. This approach replaces the minimization of (2.2) by the minimization of

$$\widehat{\sigma}_{LTS}^2(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = \sum_{i=1}^{[n(1-\alpha)]} d_{(i:n)}^2(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}), \tag{2.3}$$

where

$$d^2_{(1:n)}(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) \leq ... \leq d^2_{(n:n)}(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m})$$

(this type of notation will always be used when referring to element ordering).

[35] provides an algorithm based on "concentration steps" that requires the search of the eigenvectors of the covariance matrices of subsets of observations in each step. Alternatively, [3] and [9] propose an iterative method of "alternating least squares with weights" to perform the minimization of (2.3) avoiding working with covariance matrices of subsets of observations (which can be problematic in high dimensions or when $n < p$).

The approach in [3] does not guarantee the orthogonality of the resulting $\boldsymbol{B}_q$ matrix, although it can be proven at the population level that the optimal $\boldsymbol{B}_q$ matrix is an orthogonal matrix for elliptical distributions as shown in [3] and [24]. However, even without orthogonality, $\boldsymbol{B}_q$ provides a reasonable estimate of the best approximating subspace for the data. The Gram-Schmidt method can be applied if orthogonality is required.

A very interesting new approach was put forward in the doctoral thesis [8]. The iterative approach of alternating least squares with weights allows incorporating null weights to some specific cells opening the door to cellwise trimming. [8] defines the estimator of least squares trimmed 'by coordinates' (Coo-LTS) minimizing

$$\widehat{\sigma}^2_{\text{Coo-LTS}}(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = \sum_{j=1}^{p} \widehat{\sigma}^2_{\text{LTS},j}(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}),$$

where

$$\widehat{\sigma}^2_{\text{LTS},j}(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = \sum_{i=1}^{[n(1-\alpha)]} r^2_{(i:n)j} = \sum_{i=1}^{n} w_{ij}(x_{ij} - m_j - \boldsymbol{a}_i^T \boldsymbol{b}_j)^2, \tag{2.4}$$

for

$$w_{ij} = \begin{cases} 1 & \text{if } r^2_{ij} \leq r^2_{([n(1-\alpha)]:n)j} \\ 0 & \text{if } r^2_{ij} > r^2_{([n(1-\alpha)]:n)j} \end{cases}. \tag{2.5}$$

These weights w_{ij} would inform us whether the cell x_{ij} is trimmed ($w_{ij} = 0$) or not ($w_{ij} = 1$). Again, this way of proceeding provides a better approximating subspace but does not guarantee that the obtained array $\boldsymbol{B}_q$ is orthogonal.

Finally, although it is not a PCA method, we will conclude this section by briefly reviewing the LTS method applied in regression [42], and which will later be used in the algorithm provided in Section 4. Given n values of a response variable $\{y_i\}_{i=1}^n$ and n vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ with the values of p explanatory variables, LTS regression looks for the coefficient vector $\tilde{\boldsymbol{b}} \in \mathbb{R}^p$ for which the following expression is minimized:

$$\sum_{i=1}^{[n(1-\alpha)]} r^2_{(i:n)}, \tag{2.6}$$

with $r_i^2 = (y_i - \tilde{\boldsymbol{b}}' \boldsymbol{x}_i)^2$. The Algorithm 1 shows a simple description of the fast-LTS algorithm [39, 40] commonly applied to solve the minimization of (2.6).

Algorithm 1 Fast-LTS algorithm

Data: $\{y_i : i = 1, ..., n\}$, $\{\boldsymbol{x}_i : i = 1, ..., n\}$, and a trimming level α
Output: Regression parameter vector $\tilde{\boldsymbol{b}}$
for $b = 1, ..., B_1$ **do** ▷ *Random initializations*
  Initialize $\boldsymbol{b}$ (usually by randomly choosing $p + 1$ observations $\boldsymbol{x}_i$)
  for $c = 1, ..., C$ **do** ▷ *'Concentration' steps*
    $r_i \leftarrow y_i - \boldsymbol{b}^T \boldsymbol{x}_i$
    $\mathcal{I} = \{i : r_i^2 \leq r^2_{([n(1-\alpha)]:n)}\}$
    Update $\boldsymbol{b}$ using regression on the subsets $\{y_i\}_{i\in\mathcal{I}}$ and $\{\boldsymbol{x}_i\}_{i\in\mathcal{I}}$
  end for
end for
Return $\tilde{\boldsymbol{b}}$, which is the $\boldsymbol{b}$ giving the smallest value of (2.6)

3. Cellwise trimming in cluster analysis

In this section we extended the 'Coo-LTS' method reviewed in Section 2 to the case of G populations. For this purpose, we consider the approximations provided by G approximating subspaces:

$$\widehat{\boldsymbol{x}}_i^g\big(\boldsymbol{B}_{q_g}^g, \boldsymbol{A}_{q_g}^g, \boldsymbol{m}^g\big) = \boldsymbol{m}^g + \boldsymbol{B}_{q_g}^g \boldsymbol{a}_i^g \text{ or, at cell level, } \widehat{x}_{ij}^g = m_j^g + (\boldsymbol{a}_i^g)^T \boldsymbol{b}_j^g,$$

with $\boldsymbol{B}_{q_g}^g \in \mathbb{R}^{p\times q_g}$ (rows given by $\boldsymbol{b}_j^g$ for $j = 1, ..., p$), $\boldsymbol{A}_{q_g}^g \in \mathbb{R}^{n\times q_g}$ (rows given by $\boldsymbol{a}_i^g$ for $i = 1, ..., n$) and $\boldsymbol{m}^g \in \mathbb{R}^p$ (elements given by m_j^g for $j = 1, ..., p$). The parameter q_g is the intrinsic dimension of the g-th approximating subspace. For each observation $\boldsymbol{x}_i$, the best approximation to that observation is unknown.

To simplify the notation, let us assume that the 'intrinsic dimensions' q_g are all equal to q (i.e., we assume $q_1 = \cdots = q_G = q$). The extension of the methodology to the case of different dimensions is not especially complex.

We can define residuals for each cell:

$$r_{ij}^g = r_{ij}^g(\boldsymbol{B}_q^g, \boldsymbol{A}_q^g, \boldsymbol{m}^g) = x_{ij} - \widehat{x}_{ij}^g,$$

and, looking for an approach linked to the "Coo-LTS", we consider weights w_{ij}^g for $i = 1, ..., n$, $j = 1, .., q$ and $g = 1, ..., G$, which may be seen as the contribution to the group g of the j-th coordinate from observation $\boldsymbol{x}_i$. With these weights, we would be interested in minimizing a target function of the type:

$$\min_{w_{ij}^g, B_q^g, A_q^g, \boldsymbol{m}^g} \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{G} w_{ij}^g (r_{ij}^g(\boldsymbol{B}_q^g, \boldsymbol{A}_q^g, \boldsymbol{m}^g))^2. \tag{3.1}$$

However, we must place appropriate restrictions on the minimization of (3.1) on the weights w_{ij}^g so that we can pose a reasonable problem that leads us to a useful method. Thus, from a cluster analysis point of view, a reasonable restriction is that an observation $\boldsymbol{x}_i$ is entirely assigned to one and only one of the g groups, so that different cells x_{ij} and $x_{ij'}$ (from the same observation $\boldsymbol{x}_i$) are not assigned to two different groups g and g'. In other words, if we denote by $g(i)$ the group assignment of observation $\boldsymbol{x}_i$ we require:

$$w_{ij}^g = 0 \text{ for every } g \neq g(i). \tag{3.2}$$

A second type of restriction on the weights w_{ij}^g is aimed at trimming a controlled fraction of cells. Given a trimming size α, we require that

$$\sum_{\{i:g(i)=g\}} w_{ij}^g = [n_g(1-\alpha)], \text{ for every } g \text{ and } j, \text{ where } n_g = \#\{g(i) = g\}. \tag{3.3}$$

These restrictions mean that the same proportion of cells are trimmed in each 'coordinate' and in each cluster.

The use of a more global restriction in combination with (3.2) has been also explored:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{G} w_{ij}^{g} = [np(1-\alpha)]. \tag{3.4}$$

This would mean that a proportion α of cells is trimmed from the total $n \times p$ cells of the data matrix without imposing any other restrictions. Although the idea seemed reasonable initially, we have found that the use of the restrictions in (3.3) leads to much more stable procedures than considering the restriction (3.4). For example, if α is significantly larger than the true level of contamination or if the initialization of the parameters is not practically perfect, then, frequently, the method ends up trimming all the observations into a few coordinates and the iterative process in the proposed algorithm stops.

4. Description of the algorithm

In this section, we will first present the different steps for parameter updating (assuming other parameters are known) that will be combined in a global pseudo-code in Section 4.4 as a feasible algorithm to implement the proposed cell trimming methodology.

4.1. Updating the subspace parameters assuming known weights

Suppose that the optimal weights w_{ij}^{g} were known and that we sought to optimize the rest of the parameters conditionally on these weights. Let's denote by $L\left(\{w_{ij}^{g}\}_{ij}^{g}, \{\boldsymbol{B}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{A}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{m}^{g}\}^{g}\right)$ the sum of reweighted cellwise squared errors in (3.1), that will be our target function. We can differentiate function L with respect to $\boldsymbol{a}_{i}^{g}, \boldsymbol{b}_{j}^{g}$, and m_{j}^{g} to obtain:

$$\frac{\partial}{\partial \boldsymbol{a}_{i}^{g}} L\left(\{w_{ij}^{g}\}_{ij}^{g}, \{\boldsymbol{B}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{A}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{m}^{g}\}^{g}\right) = -2\sum_{j=1}^{p} w_{ij}^{g} r_{ij}^{g}(\boldsymbol{B}_{q}^{g}, \boldsymbol{A}_{q}^{g}, \boldsymbol{m}^{g})\boldsymbol{b}_{j}^{g},$$

$$\frac{\partial}{\partial \boldsymbol{b}_{j}^{g}} L\left(\{w_{ij}^{g}\}_{ij}^{g}, \{\boldsymbol{B}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{A}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{m}^{g}\}^{g}\right) = -2\sum_{j=1}^{p} w_{ij}^{g} r_{ij}^{g}(\boldsymbol{B}_{q}^{g}, \boldsymbol{A}_{q}^{g}, \boldsymbol{m}^{g})\boldsymbol{a}_{i}^{g},$$

and

$$\frac{\partial}{\partial m_{j}^{g}} L\left(\{w_{ij}^{g}\}_{ij}^{g}, \{\boldsymbol{B}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{A}_{q}^{g}\}_{q}^{g}, \{\boldsymbol{m}^{g}\}^{g}\right) = -2\sum_{j=1}^{p} w_{ij}^{g} r_{ij}^{g}(\boldsymbol{B}_{q}^{g}, \boldsymbol{A}_{q}^{g}, \boldsymbol{m}^{g}).$$

Setting these derivatives to 0 gives the following system of equations:

$$\sum_{j=1}^{p} w_{ij}^{g}(x_{ij} - m_{j}^{g})\boldsymbol{b}_{j}^{g} = \left(\sum_{j=1}^{p} w_{ij}^{g}\boldsymbol{b}_{j}^{g}(\boldsymbol{b}_{j}^{g})^{T}\right)\boldsymbol{a}_{i}^{g}, i = 1, ..., n, \text{ y } g = 1, ..., G,$$

$$\sum_{i=1}^{n} w_{ij}^{g}(x_{ij} - m_{j}^{g})\boldsymbol{a}_{i}^{g} = \left(\sum_{i=1}^{n} w_{ij}^{g}\boldsymbol{a}_{i}^{g}(\boldsymbol{a}_{i}^{g})^{T}\right)\boldsymbol{b}_{j}^{g}, j = 1, ..., p, \text{ y } g = 1, ..., G,$$

and

$$\sum_{i=1}^{n} w_{ij}^{g}(x_{ij} - (\boldsymbol{a}_{i}^{g})^{T}\boldsymbol{b}_{j}^{g}) = \sum_{i=1}^{n} w_{ij}^{g} m_{j}^{g}, j = 1, ..., p, \text{ y } g = 1, ..., G.$$

Therefore, with known weights, the optimal parameters $\boldsymbol{B}_{q}^{g}, \boldsymbol{A}_{q}^{g}$, and $\boldsymbol{m}^{g}$ can be obtained by applying weighted least squares. Fast and computationally efficient procedures exist to solve these problems since the common intrinsic dimension q is usually much smaller than the original dimension p of the problem.

4.2. Updating weights with known group membership

Now let us assume that the group assignments are known. That is, we have a partition of $\{1, 2, ..., n\}$ into $\{\mathcal{I}_1, ..., \mathcal{I}_G\}$ with $\mathcal{I}_g = \{i : g(i) = g\} = \{i_1^g, ..., i_{n_g}^g\}$ and $n_g = \#\mathcal{I}_g$. Since the cell prediction is $\widehat{x}_{ij} = m_j^g + (\boldsymbol{a}_i^g)^T \boldsymbol{b}_j^g$ for $g = g(i)$ and the associated residuals are $R_{ij}^2 = (x_{ij} - \widehat{x}_{ij})^2$, it is easy to see that the best update of weights is provided by:

$$w_{ij}^g = \begin{cases} 1 & \text{if } g = g(i) \text{ and } R_{ij}^2 \leq R^2_{i^g_{([n_g(1-\alpha)]:n_g)}j} \\ 0 & \text{in other case} \end{cases}. \tag{4.1}$$

It is trivial to see that these weights w_{ij}^g meet the required restrictions (3.2) and (3.3), and reduce as much as possible the objective function (3.1) if the assignments $g(i)$ and the parameters $\boldsymbol{B}_q^g$, $\boldsymbol{A}_q^g$ and $\boldsymbol{m}^g$ remain constant.

4.3. Updating group membership

Even with fairly reasonable initializations of $\boldsymbol{B}_q^g$ and $\boldsymbol{m}^g$, the initialization of $\boldsymbol{A}_q^g$ by applying simple regression (taking into account (2.1)) may not be adequate. Note that $\boldsymbol{x}_i$ (the values of the response variable in these regressions) may include outliers and that a few atypical values may affect very negatively classical least-squares regression. In this initialization phase, we also do not have reliable weights w_{ij}^g (the w_{ij}^g are only reliable for the g corresponding to the group to which the observation $\boldsymbol{x}_i$ was assigned in the previous step) to decrease the weight of the outliers. Something similar occurs with group assignments since, even with well-defined approximating spaces, it is difficult to know if an observation $\boldsymbol{x}_i$ with $g(i) = g$ would have been better assigned to another group and thus have $g(i) = g' \neq g$.

A fairly reasonable idea to solve the problems discussed above might be to use the LTS robust regression reviewed in Section 2 instead of using least squares regression. LTS regression is not affected by a few atypical cells and allows us to make more reliable determinations of $\boldsymbol{A}_q^g$ and of group assignments. To clarify this proposal, we will initially assume that $\alpha = 0$. When applying the Lloyd-Forgy classic k-means algorithm, assignments are made using $g(i) = \arg\min_{g=1,...,G} \|\boldsymbol{x}_i - \boldsymbol{m}^g\|^2$ (groups around "centroids"). Similarly, when looking for groups around subspaces it seems reasonable to use $g(i) = \arg\min_{g=1,...,G} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i^g\|^2$ where $\hat{\boldsymbol{x}}_i^g = \boldsymbol{m}^g + \boldsymbol{B}_q^g \boldsymbol{a}_i^g$ would be the closest point to $\boldsymbol{x}_i$ in the g-th approximating subspace. This point $\hat{\boldsymbol{x}}^g$ can be obtained directly by applying least squares regression, modeling the p values of the vector $\boldsymbol{x}_i - \boldsymbol{m}^g$ as the optimal linear combination of the q columns of matrix $\boldsymbol{B}_q^g$. Our proposal is to find this optimal linear combination by using the robust LTS regression with a trimming fraction of size α_{LTS}. Obviously, this assumes that a fraction greater than α_{LTS} of contaminating cells in $\{x_{ij} : j = 1, ..., p\}$ is not expected.

We will denote by $\tilde{\boldsymbol{a}}_i^g$ the coefficients obtained by applying these LTS regressions and their residuals by $\tilde{r}_{ij}^g = (x_{ij} - \boldsymbol{m}^g - (\boldsymbol{b}_j^g)^T \tilde{\boldsymbol{a}}_i^g)^2$. We can define a 'distance' D_i^g between $\boldsymbol{x}_i$ and the g-th approximating subspace considering only the fraction $[p(1 - \alpha_{\text{LTS}})]$ of more 'favorable' cells. That is,

$$D_i^g = \sum_{j=1}^{[p(1-\alpha_{\text{LTS}})]} \tilde{r}^g_{i(j:p)},$$

and group assignments would be

$$g(i) = \arg \min_{g=1,...,G} D_i^g.$$

Moreover, the vectors $\tilde{\boldsymbol{a}}_i^g$ give a robust initialization of the score matrices $\boldsymbol{A}_q^g$.

Algorithm 2 Summary of the proposed algorithm

Data: $\{\mathbf{x}_i : i = 1, ..., n\}$, trimming level α, and intrinsic dimension q (could be adapted to different intrinsic dimensions q_g)
Result: Optimal values of w_{ij}^g, $\boldsymbol{B}_q^g$, $\boldsymbol{A}_q^g$, and $\boldsymbol{m}_g$ and group assigments $\{g(i) : i = 1, ..., n\}$.
for $b = 1, ..., B_2$ **do** ▷ *Random initializations*
  Initialize $\boldsymbol{B}_q^g$ and $\boldsymbol{m}_g$ (see Section 4.5)
  for $l_1 = 1, ..., L_1$ **do** ▷ *External loop*
    LTS regression to initialize $\boldsymbol{A}_q^g$ and $\{g(i) : i = 1, ..., n\}$ (see Section 4.3).
    Group assignments $g(\cdot)$ are fixed in the *internal loop*:
    for $l_2 = 1, ..., L_2$ **do** ▷ *Internal loop*
      $\widehat{x}_{ij} \leftarrow m_j^g + (\boldsymbol{a}_i^g)^T \boldsymbol{b}_j^g$ for $g = g(i)$ and $R_{ij}^2 \leftarrow (x_{ij} - \widehat{x}_{ij})^2$
      Update w_{ij}^g (Section 4.2)
      Update $\boldsymbol{B}_q^g$, $\boldsymbol{A}_q^g$, and $\boldsymbol{m}_g$ using weighted regression (see Section 4.1)
    end for
  end for
  After the loops the objective function (3.1) is calculated
end for
Return parameters and group assignment with the smallest value of (3.1)
A final improvement step can be applied (see Section 4.5)

The main problem with this LTS approach is its high computational cost. Note that a total of $G \times n$ LTS regressions must be performed (although with a moderate number of observations p and an even smaller number q of explanatory variables if the intrinsic dimension chosen is not high). To reduce this computational burden, we propose to run Algorithm 1 with a very reduced number of initializations B_1 and very few concentration steps C since the coefficients $\tilde{\boldsymbol{a}}_i^g$ can be improved in subsequent steps of the algorithm.

Another possibility to explore in the algorithm would be to remove temporarily, during updates, a preset fraction of observations $\boldsymbol{x}_i$ with the highest values of D_i^g. This step would be identical to the one performed in the k-trimmed means algorithm, [20], and would allow trimming cells and rows in a unified way.

4.4. Pseudo-code of the algorithm

Algorithm 2 shows a simplified pseudo-code of the complete iterative process being proposed, which integrates all the parameter updates presented in the previous sections.

Given the high computational cost of performing many LTS regressions, our recommendation would be that the number of external loops should not be too high (i.e., a not very high L_1). However, note that LTS regressions would only be done once in the *external loop*, and no further LTS regressions are needed inside the *internal loop*. Fortunately, as with other algorithms with an analogous philosophy to k-means, not many reassignment steps are usually necessary when starting from a 'reasonable' initialization of the parameters. The *internal loops* can therefore incorporate a 'stopping criterion' if $\boldsymbol{B}_q^g$, $\boldsymbol{A}_q^g$, and $\boldsymbol{m}_g$ do not change appreciably in consecutive iterations. The number of *internal loops* L_2 would not be critical either because a few iterations usually allow us to get an idea of the most promising solutions, and these will later be the only ones to be completely iterated. However, the number of *random iterations* B_2 and how they should be performed reasonably and efficiently is of paramount importance and discussed in more detail in Section 4.5.1.

We also propose to incorporate a final improvement step that will be detailed in Section 4.5.2. This improvement allows 'retrieval' of incorrectly trimmed cells and, also, to trim complete observations bx_i when the set of outliers in

the observation is too large for the observation to be globally reliable.

4.5. Initialization and final improvement step

4.5.1. Initialization

Numerous empirical results show that moderate values of L_1 and L_2 are typically necessary when starting from a (not necessarily 'optimal') but 'reasonable' initialization. The steps described in Sections 4.1 and 4.2 seek to ensure that the objective function is monotonically decreasing and that the weight restrictions are always satisfied. However, the algorithm may get stuck in a local minimum of the target function when starting from an unreasonable initial solution. Therefore, considering multiple random initializations is, in general, essential and, also, trying that these initializations allow suitably exploring the solution space. How to provide these initializations is not a trivial problem when G or p is large. This problem is not exclusive to this methodology and also appears when applying other more straightforward methods of Cluster Analysis (robust and not robust). For example, when applying the TCLUST method in [21], it was proposed to randomly select $G \times (p+1)$ observations from the sample, while with this methodology, based on approximating subspaces, this could be reduced to selecting $G \times (q+1)$ observations with $q \leq p$.

The initialization process would be greatly simplified if we could count on a fraction of observations that we know for sure are little or not at all contaminated or if we have a fraction of observations already correctly assigned to the possible groups (semi-supervised problem).

However, a large number of initializations B_2 will generally be required when either G or p are high. Some computation shortcuts can be proposed in these cases, such as only iterating more exhaustively those initializations that are most promising in their first steps or applying parallel computing to consider a larger number B_2 of random initializations. It also makes sense to include initializations that come from applying some method for robust cluster analysis not necessarily designed for cell outliers. A particular example of this idea to the functional case will be shown in Section 5 using TCLUST as the initialization procedure.

4.5.2. Final improvement step

The type of trimming considered in (3.3) can lead to trimming a perhaps too high proportion of atypical cells, for values of the trimming size α large or when there are no atypical cells in the coordinate j for some group g. Our recommendation is to always consider a large initial trimming size α as a precaution and to retrieve incorrectly trimmed cells later.

Recovering incorrectly trimmed cells is not a very complex task once good estimates of the approximating subspaces are available and the $\boldsymbol{x}_i$ observations have been correctly assigned to groups. As was done in Section 4.2, we can calculate $n \times p$ residuals $R_{ij}^2 = (x_{ij} - \widehat{x}_{ij})^2$ and order them globally. By examining these ordered residuals, in most cases, you can clearly distinguish cells x_{ij} with larger and extreme values of R_{ij}^2 from other cells with smaller values and a slow decreasing pattern. The trimmed cells with small values of R_{ij}^2 (or R_{ij}^2 very close to other cells that were trimmed) can be recovered in a simple step of fine-tuning.

A second possibility of improvement is based on allowing complete rows to be trimmed for observations with $\boldsymbol{x}_i$ such that $\#\{w_{ij}^{g(i)} = 0 : j = 1, ..., p\}/p > \alpha_{\text{LTS}}$. Note that, in this case, we do not have a full guarantee that the LTS returns a correct group assignment since the observation of $\boldsymbol{x}_i$ does not appear to be 'comfortably' located in that group.

5. Application to the functional case

For simplicity, we will always assume that the functions are observed in the interval $[0,1]$ in p equidistant moments of time $0 < t_1 < \cdots < t_p < 1$. Since our proposal does not require the calculation of huge variance-covariance matrices, we will explore the limits of our methodology by working directly with this curve discretization even though p may be quite large. However, smaller finite-dimensional representation of the functions on an orthonormal functional basis can be also applied when p is definitely too large.

The fundamental idea is to look for reasonable initializations of the parameters $\boldsymbol{B}_q^g \in \mathbb{R}^{p\times q}$, $\boldsymbol{A}_q^g \in \mathbb{R}^{n\times q}$, and $\boldsymbol{m}^g \in \mathbb{R}^p$ by applying a traditional robust Cluster Analysis method (trimming of complete observations) after having smoothed the curves and a reduction of their dimensionality that is done exclusively in the initialization stage. For this initialization to be satisfactory, we will assume that the approximating subspaces can be reasonably represented on a finite functional basis. Proposing reasonable initializations can be an extremely complex problem without such assumptions in this functional case. We present below a more detailed description of the initialization process we propose.

In a *first phase*, the curves are smoothed out by a robust local regression where sharp discontinuities are eliminated or smoothed out. Our proposal is to use the `lowess` method [11]. Subsequently, the new p-dimensional smoothed data is represented in a lower dimension P ($P << p$) considering a functional basis with functions $\{\phi_1, ..., \phi_P\}$. In all the examples that will be shown in this work we have used B-splines, although other functional bases -Fourier-type or wavelets- could be applied depending on the data's specific characteristics. When using B-splines with η inner nodes, the reduced dimension is $P = \eta + 4$. This representation reduces the dimension to apply finite-dimensional methods of robust cluster analysis and provides a second smoothing/regularization of the original curves. As a result of this first phase we will have some coefficients $\{\widetilde{\boldsymbol{x}}_i, i = 1, ..., n\}$ representing the curves with $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^P$ ($P << p$), that will be the input of the second phase.

In the *second phase*, we apply a robust clustering method on $\{\widetilde{\boldsymbol{x}}_i : i = 1, ..., n\} \subset \mathbb{R}^P$. Our suggestion is to use TCLUST [21] but other methods of robust cluster analysis can also be considered. After applying TCLUST with an α_{TCLUST} trimming level, which does not have to match the cell trimming level α, we obtain robust estimators of the averages $\widetilde{\boldsymbol{\mu}}_g \in \mathbb{R}^P$ and robust estimators of the covariance matrices $\widetilde{\boldsymbol{\Sigma}}_g \in \mathbb{R}^{P\times P}$ for $g = 1, \ldots, G$. If $\boldsymbol{\Phi} \in \mathbb{R}^{p\times P}$ is the matrix with values $\{\phi_l(t_j)\}_{j=1,\ldots,p}^{l=1,\ldots,P}$ we propose to use the mean vectors obtained with TCLUST to initialize the mean functions $\boldsymbol{m}^g$ as $\boldsymbol{m}^g = \boldsymbol{\Phi}\widetilde{\mu}_g \in \mathbb{R}^p$. Similarly, if $\mathbf{V}_q^g \in \mathbb{R}^{P\times q}$ is the matrix that has as columns the q eigenvectors associated to the largest q eigenvalues of $\widetilde{\boldsymbol{\Sigma}}_g$, a simple way to initialize $\boldsymbol{B}_q^g$ in the original p-dimensional space is to consider $\boldsymbol{B}_q^g = \boldsymbol{\Phi}\mathbf{V}_q^g \in \mathbb{R}^{p\times q}$.

In our experience, the combination of these two phases provides fairly reasonable initializations of $\boldsymbol{B}_q^g$ and $\boldsymbol{m}^g$. It is not intended that this procedure will directly provide the optimal solution to the problem of grouping in the original p-dimensional space, but it does provide a 'reasonable' initialization of $\boldsymbol{B}_q^g, \boldsymbol{A}_q^g$ and $\boldsymbol{m}^g$ for the iterative process. Any other method providing a robust estimate of the G approximating subspaces in the reduced dimensional space $\mathbb{R}^P$ could be used instead. Ideally, several such initializations should be incorporated whenever possible, as is done in [30] in the computation of the MCD estimator. Thus, for example, an adaptation of the procedure in [25] could be considered. Another possibility would be the procedure in [2], which adds trimming by observations to the HDDC method in [4] designed to perform robust cluster analysis at high dimensions.

As already mentioned, the possibility of trimming portions of curves was noted as an interesting line of research in the discussion of [28]. In that discussion, it was also proposed to use 'snipping' techniques as in [16, 17] after using a finite-dimensional representation of the curves on a B-spline base. Snipping can be seen as an adaptation of the k-trimmed means method to the case of cell trimming. However, the use of the subspace grouping technique introduced in this work makes it possible to explore dependency structures in groups with some parsimony and to deal with cases in a higher initial dimension p. It is also well known that B-spline bases 'expand' to nearby nodes (i.e., they are not null in adjacent nodes), and the trimming procedure based exclusively on a representation using B-splines with a moderate number of nodes, would not be as local as might sometimes be desired. The representation using B-splines is only used in the first phase of initialization of our current proposal.

6. Examples and simulation study

In this section, we give some examples and a basic simulation study showing the relevance of the proposed methodology. We have focused exclusively on the functional case because it provides a more precise and simpler illustration of the methods and allows us to explore the limit of the technique by considering cases where high dimensions appear naturally.

6.1. Examples

Consider two groups with 200 observation each, and centered around the mean functions $\mu_1(t) = 5+10\sin(4\pi t)\exp(-2t)+5\sin(\pi t/3)+2\cos(2\pi t/2)$ and $\mu_2(t) = 10+10\cos(4\pi t)$. To generate observations from the approximating functional subspaces we use the functions $\varphi_1(t) = \sqrt{2}\cos(2\pi t)$ and $\varphi_2(t) = \sqrt{2}\sin(2\pi t)$. Curves are discretized on a grid with $p = 100$ points and independent observation errors having a normal distribution are added.

Specifically, we generate $\{\boldsymbol{x}_i\}_{i=1}^{400}$ with $\boldsymbol{x}_i \in \mathbb{R}^{100}$ such that

$$\begin{aligned} x_{ij} &= m_i(t_j) + 0.5 \cdot \varepsilon_{ij} = \\ &= \mu_{g(i)}(t_j) + a_{g(i)1} z_{i1} \varphi_1(t_j) + a_{g(i)2} z_{i2} \varphi_2(t_j) + 0.5 \cdot \varepsilon_{ij}, \end{aligned}$$

for $t_j = j/101$ and $j = 1, ..., 100$. We set $g(i) = 1$ for $i = 1, ..., 200$ (first cluster) and $g(i) = 2$ for $i = 201, ..., 400$ (second cluster). The z_{i1}, z_{i2} and ε_{ij} are independent realizations of a standard normal distribution and $a_{11} = 3, a_{12} = 2$ are fixed values for the first cluster while $a_{21} = 2, a_{22} = 4$ are fixed values for the second cluster.

We will consider four different contamination scenarios that arise from replacing 3% of the cell values in all cases (i.e., 1,200 corrupt cells x_{ij} out of a total of 40,000 cells):

Contamination Scheme I (scattered): A fraction of 3% of cells are selected at random. Each of these cells is replaced with either a random value in the interval $[-20, -15]$ with probability 0.5 or by a random value in the interval $[35, 40]$ with probability 0.5.

Contamination Scheme II (consecutive): We choose 60 different rows $i_1, ..., i_{60}$ ($i_l \in \{1, ..., 400\}$) and 60 random numbers $j_1, ..., j_{60}$ with $j_l \in \{0, ..., 80\}$. Then, 20 consecutive cells $\{x_{i_l j_l+1}, ..., x_{i_l j_l+20}\}$ are replaced by a single value in the interval $[-45, -35]$ with probability 0.5 or by a random value chosen in the interval $[35, 40]$ also with probability 0.5.

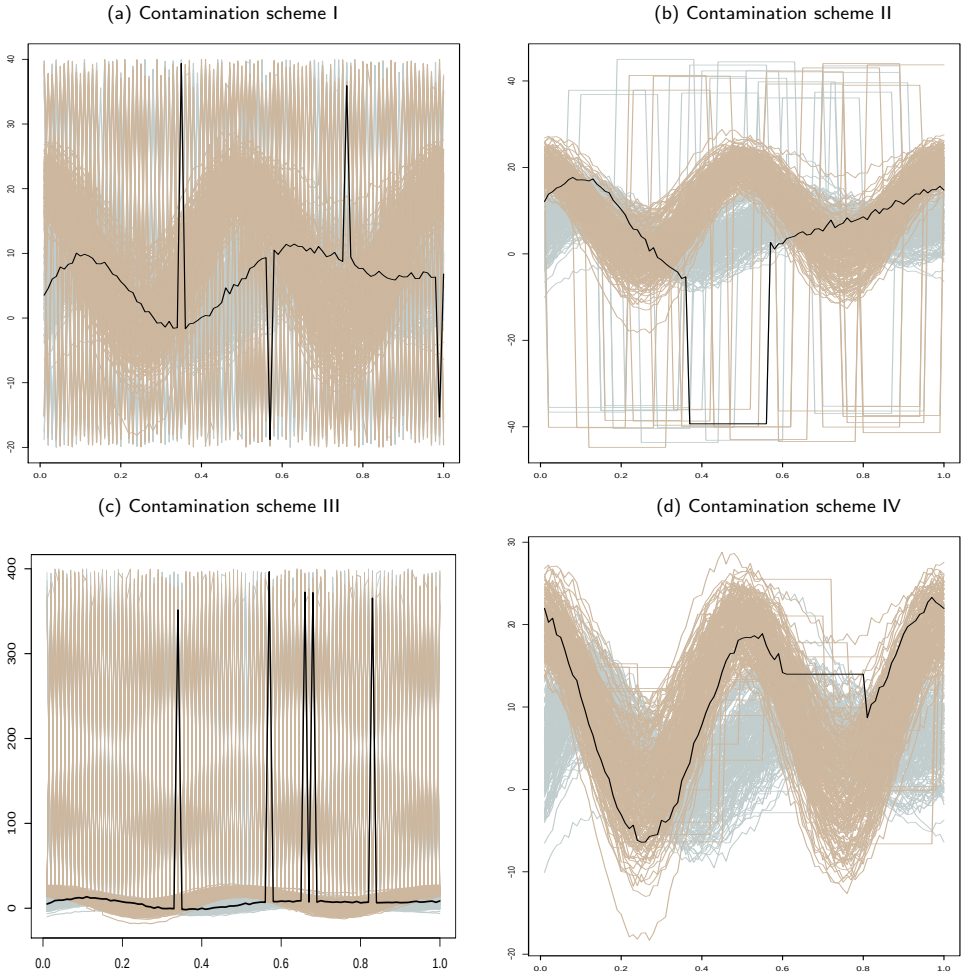


Figure 1: Examples of the four contamination schemes. In each case, one of the observations has been highlighted using a black line.

Contamination Scheme III (asymmetric and extreme): This is similar to scheme I, but now the corrupted cells are replaced by a random value in the interval $[350, 450]$. This case corresponds to corrupt cells having more extreme values than are usually observed in the data. As one would expect, this type of extreme contamination is particularly harmful.

Contamination Scheme IV (constant measurements): This is similar to scheme II but now the values $\{x_{i_l j_l+1}, ..., x_{i_l j_l+20}$ are all set at a fixed value equal to the last cell value that can be considered 'reliable' $x_{i_l j_l+1}$. This type of contamination may arise when the measuring instrument stops working properly and does not update the measurements for a certain period.

Figure 1 shows examples of the simulated functions using the four schemes (using different soft colors for the groups). In each, a curve including corrupted cells has been highlighted as a continuous black line.

Panels (a) and (b) in Figure 2 show the result of applying the proposed methodology to the same datasets as shown in Figure 1 (a) and (b) for $G = 2$ and an initial trimming size of $\alpha = 0.1$. We consider that the intrinsic dimensions are known and equal to $q_1 = q_2 = q = 2$. Both in these examples and in the simulation study, $\alpha_{\text{LTS}} = 0.3$ is considered. The initialization is based on very smooth curves by applying the `lowess` function in `R` with `f=1/5` (this parameter controls the fraction of contiguous observations to be considered in the smoothing) and, later, the smoothed curves

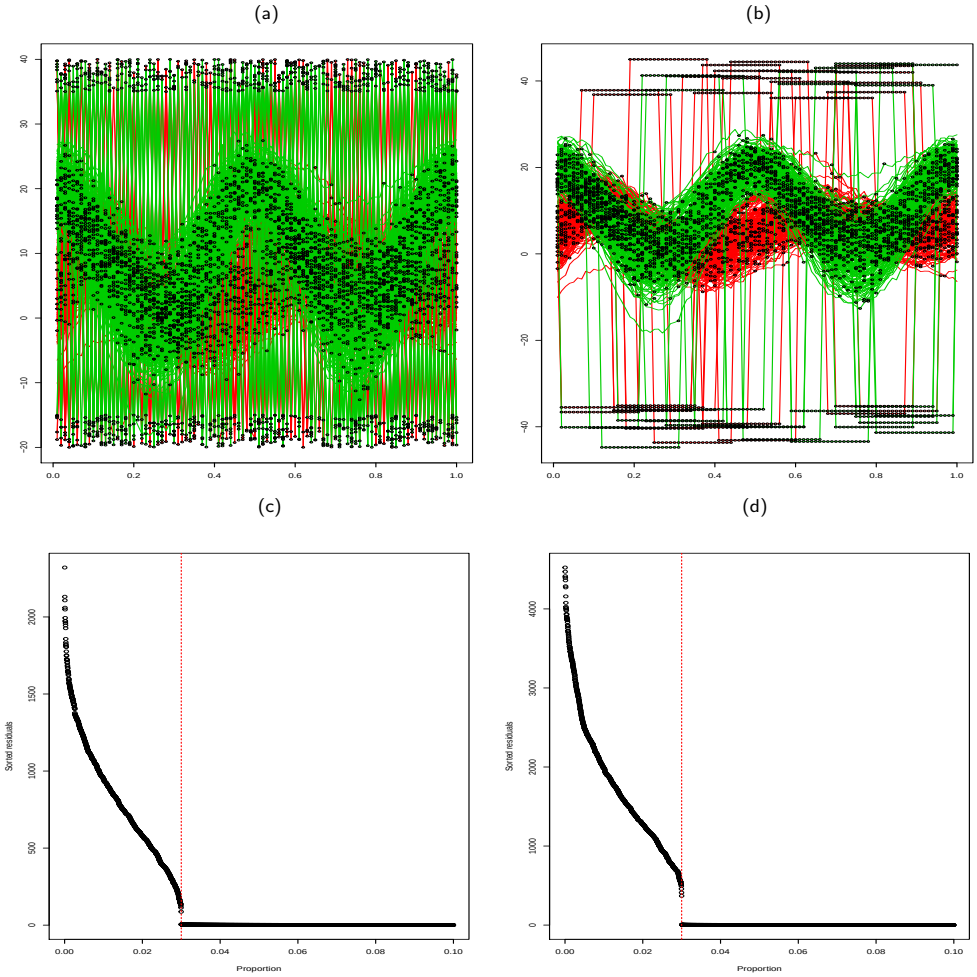


Figure 2: Results obtained by application of the methodology with $\alpha = 0.1$ (real contamination level of 3%) for the data in Figure 1(a) and (b) with trimmed cells represented by "○". Panels (c) and (d) show the ordered R_{ij}^2 and suggest a real contamination level of 3%.

are projected onto a B-spline basis with 4 nodes, a representation in dimension $P = 4 + 4 = 8$ (much lower than the original dimension of the problem $p = 100$). Finally, we apply TCLUST with $G = 2$ groups and $\alpha_{\text{TCLUST}} = 0.2$ to the coefficients $\{\widetilde{\boldsymbol{x}}_i : i = 1, ..., 400\} \subset \mathbb{R}^8$ to obtain the initialization of $\boldsymbol{B}_q^g$ and $\boldsymbol{m}_g$ by the procedure described in Section 5.

Figure 2(a) and (b) show the group assignment and the proportion of 10% of initially trimmed observations (significantly higher than the actual contamination proportion of 3%). Panels (c) and (d) in the same figure show the ordered values (from highest to lowest) of $R_{ij}^2 = (x_{ij} - \hat{x}_{ij})^2$ starting at the initial trimming value of $\alpha = 0.1$. This graph suggests that, in fact, the fraction of atypical cells should be 3% and, consequently, only the x_{ij} cells with higher values of R_{ij}^2 are finally cut.

Figure 3 shows the final result of the procedure after refinement and displays the average group curves, the group assignment, and the cells finally trimmed.

The procedure provides estimates $\hat{x}_{ij}$ for all cells, including trimmed ones. More precisely, $\hat{x}_{ij}$ should be seen as an estimation of $m_i(t_j)$. Figure 4 shows the estimated values $\hat{x}_{ij}$ (in dashed red lines) for the four individual curves highlighted in Figure 1 along with the trimmed cells in those curves marked with circles.

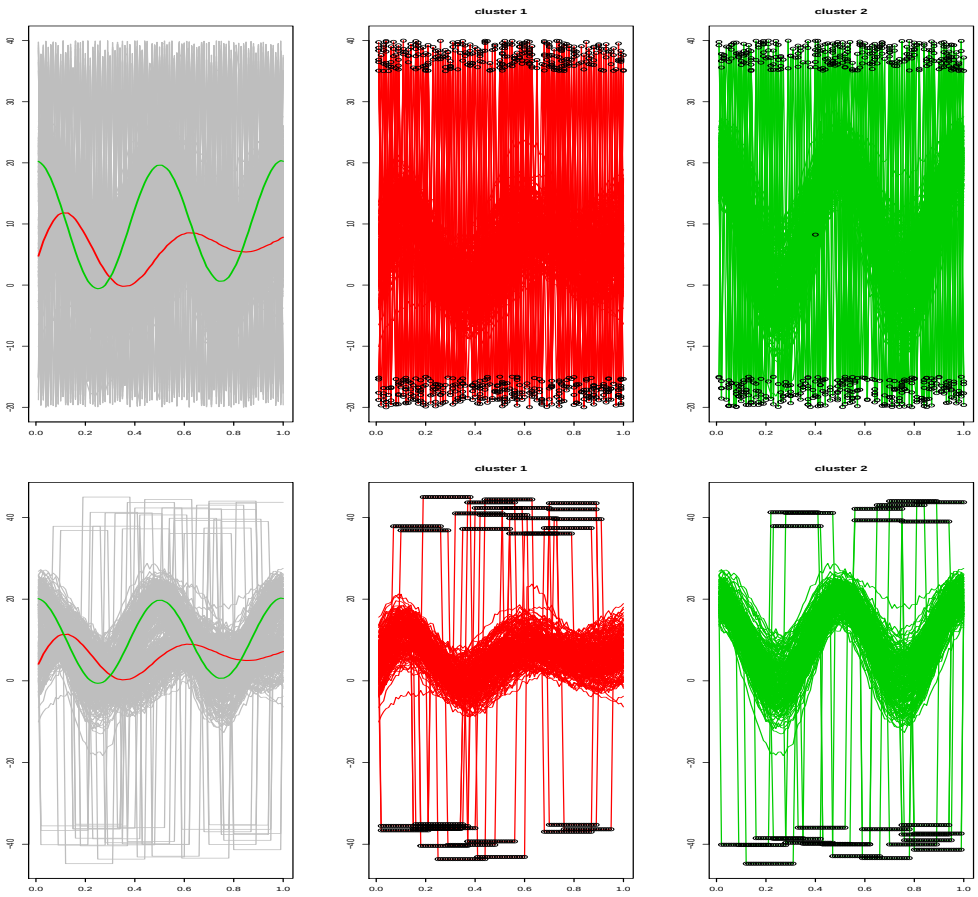


Figure 3: Average curves for the estimated groups (left), detected groups and cells finally trimmed after the refinement procedure that starts from the results in Figure 2 (center and right).

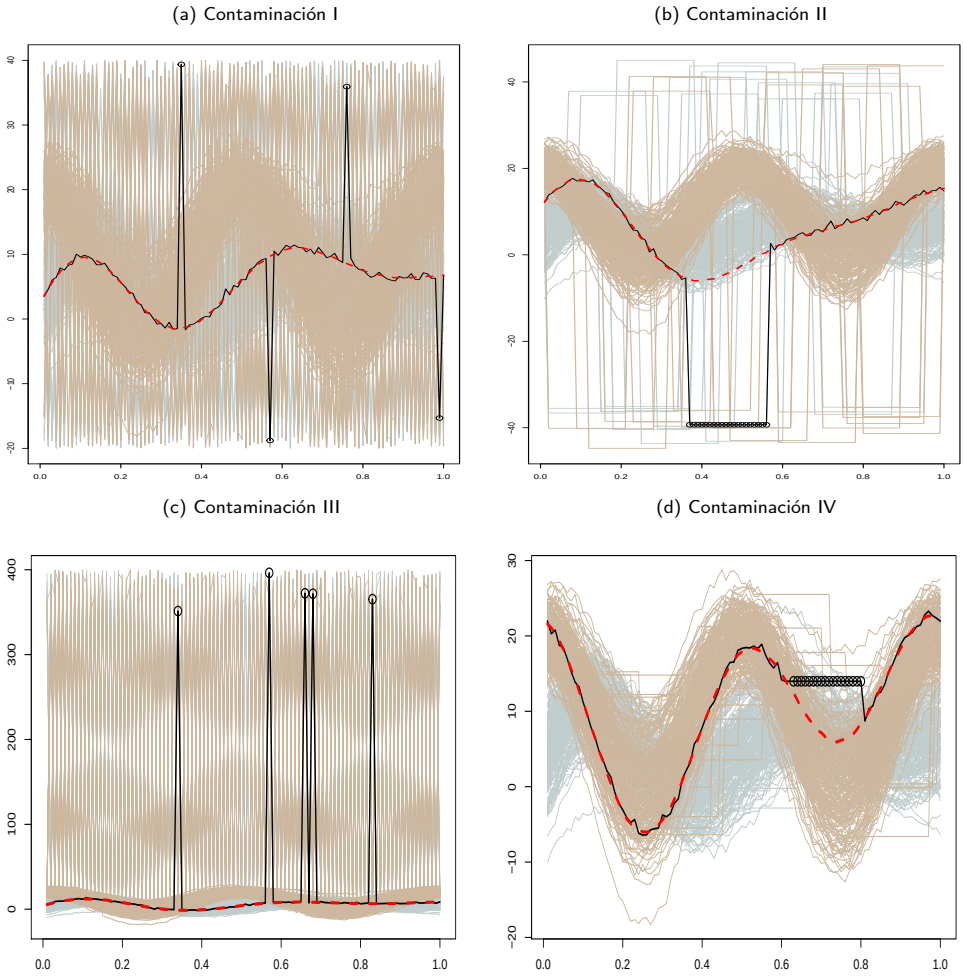


Figure 4: Estimated values $\hat{x}_{ij}$ (red dashed lines) for the highlighted curves $\boldsymbol{x}_i$ in Figure 1 along with the cells finally trimmed for those curves, represented by "o".

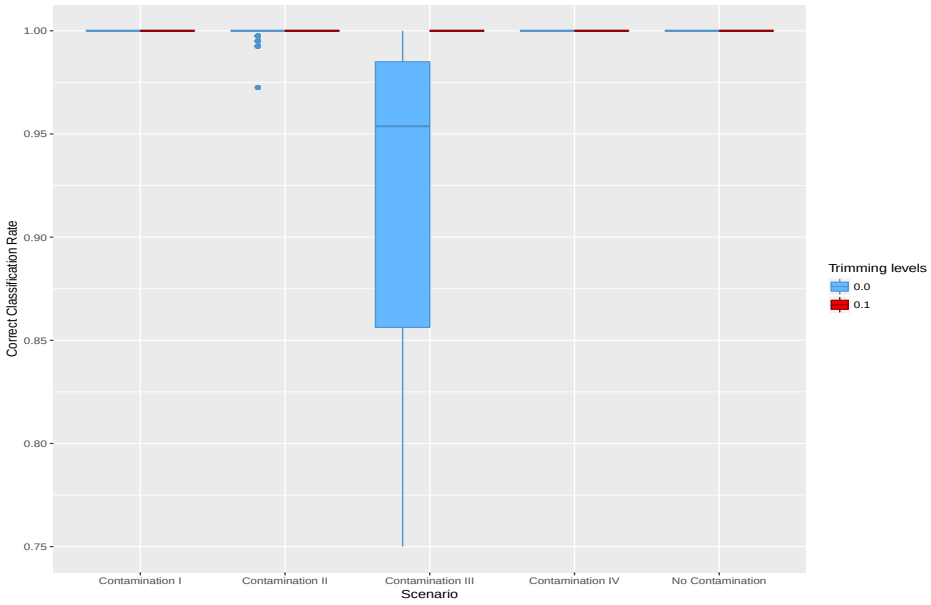


Figure 5: Correct classification rates applying the proposed methodology with $\alpha = 0$ (non robust, blue) and $\alpha = 0.1$ (robust, red).

6.2. Simulation study

In this simulation study, $B = 100$ data sets were generated, and the four mechanisms for creating contaminating cells presented in the previous section (contaminations schemes I-IV) were applied to each of them. The clustering method was applied then with no trimming $\alpha = 0$ (non-robust) and with trimming level $\alpha = 0.1$ (robust), with the same robust initialization, based on smoothing and robust cluster analysis, as in the previous section. Figure 5 shows the correct classification rate for these two approaches. It can be seen that only in the case of contamination scheme III (asymmetric contamination with extreme values) and on a few occasions for scheme II is the correct allocation of $\boldsymbol{x}_i$ changed. This is not surprising since the clusters were well separated, but we also see that very extreme contaminations (as in the case of scheme III) can be very harmful even in the case of well-separated clusters.

However, it is important to note that contaminating cells, even when they do not cause an incorrect group assignment for the whole curve, are capable of masking and not be correctly highlighted as atypical, as we will show next. Suppose that after running the procedure with $\alpha = 0$ and $\alpha = 0.1$, we decide to label as 'atypical' 3% of the cells x_{ij} with higher values of R^2_{ij}. Figure 6 shows the proportion of cells that are really atypical and correctly labeled. We can see that the rate of correctly labeled atypical cells is, in all cases, higher when working with $\alpha = 0.1$ than with $\alpha = 0$. This rate is very close to 1 in contamination schemes I, II, and III. Scheme IV is notably more complicated because it is not immediately detectable since the values at the beginning of the corruption process are very close to 'reasonable' values of the curve. However, also in this complicated case, the correct atypical labeling rate is higher when using $\alpha = 0.1$ than with $\alpha = 0$.

Finally, let $\mathcal{G}$ denote the set of indices (i, j) corresponding to the cells that were not classified as atypical in the previous step, i.e. a total of 97% of all the cells. $\mathcal{G}$ includes the $400 \cdot 100 \cdot 0.97 = 38800$ cells with smallest values of R^2_{ij}. Using this notation, we would like to have

$$SSE = \sum_{(i,j)\in\mathcal{G}} (x_{ij} - \hat{x}_{ij})^2 = \sum_{(i,j)\in\mathcal{G}} R^2_{ij},$$

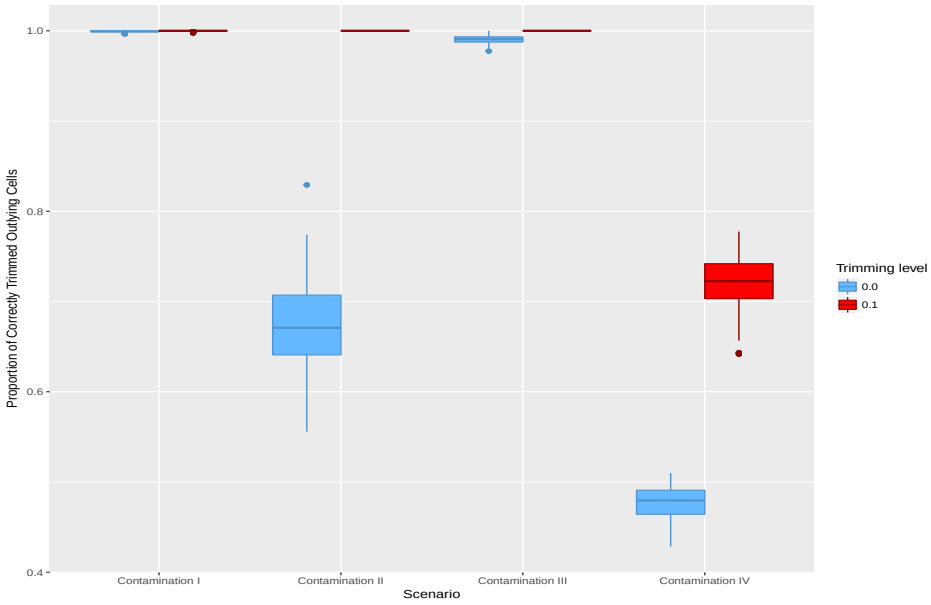


Figure 6: Proportion of correctly identified atypical cells when labeling as atypical 3% of the cells with the highest R^2_{ij} values.

small, in the sense of having good predictions for untrimmed cells. Figure 7 shows boxplots that summarize the results obtained (using a logarithmic scale) for the $B = 100$ simulated sets for each contamination scenario and also in the uncontaminated case. We see that the price paid for the robustification when $\alpha = 0.1$ is not very high in the case of uncontaminated data and that the advantage can be very large (SSE notably smaller) in the cases of contamination. Note that these SSE are never close to 0 since the error term $0.5 \cdot \varepsilon_{ij}$ was added when simulating the data and the fact that $\hat{x}_{ij}$ estimates $m_i(t_j)$.

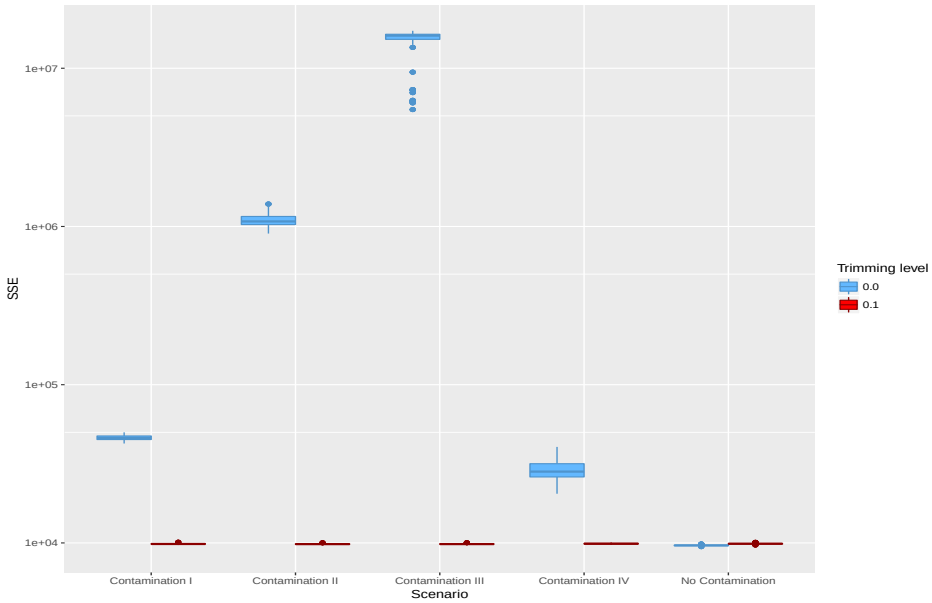


Figure 7: Sum of squared errors comparing x_{ij} and $\hat{x}_{ij}$ for the 97% of cells that were labeled as atypical.

7. Examples with real data

7.1. Mortality rates in France

In this example, we will analyze mortality data available in the Human Mortality Database (Human Mortality Database, 2013). As other authors have done before, we will focus on male mortality rates in France between 1816 and 2006. These data set is available in the `demography` package in `R`. Figure 8 presents the data, which corresponds to the logarithm of the mortality rates by age (in years) between 0 and 99 years. Lighter shades of blue correspond to earlier years.

At first glance, two clear groups can be observed. The years after 1945 show an apparent overall reduction in mortality rates. This reduction is uniform across all ages, and attributable to technological advances and improvements in the quality of life in Europe after the end of World War II. Also, a three-year transition or post-war period (1946-1948) can be seen in which the mortality curves seem to fall halfway between these two groups. In general, mortality rates decline as childhood progresses, grows again during adolescence, stabilizes at about 25 years, and finally has a smooth but continuous growth in adulthood.

This data set or parts of it have been analyzed from the functional data point of view in [31], in the thesis of H. Cevallos-Valdiviezo [8] and in the technical report that accompanies [3]. However, we think it is interesting to look at the two-group structure (pre- and post-war) when analyzing these data. For example, it is not difficult to see that a global average curve of the data with $G = 1$ would fall in a 'no man's land' between the two groups, and we would also not be able to detect differentiated 'modes of variation' in each of the groups, such as those shown in Figure 10.

The proposed methodology was applied with $G = 2$, $q = 2$, and $\alpha = 0.2$ (a high initial trimming level that will later be improved). The curves were not smoothed using `lowess` because they were already smooth enough, and the value of α_{TCLUST} was set at 0.2 and 0.3. The B-spline basis had eight nodes. Figure 9 shows a heatmap of the data matrix (191 years $\times$ 100 age groups) with warmer colors for the cells where the observed value x_{ij} is higher than the predicted value $\hat{x}_{ij}$. Values close to zero are shown in blue.

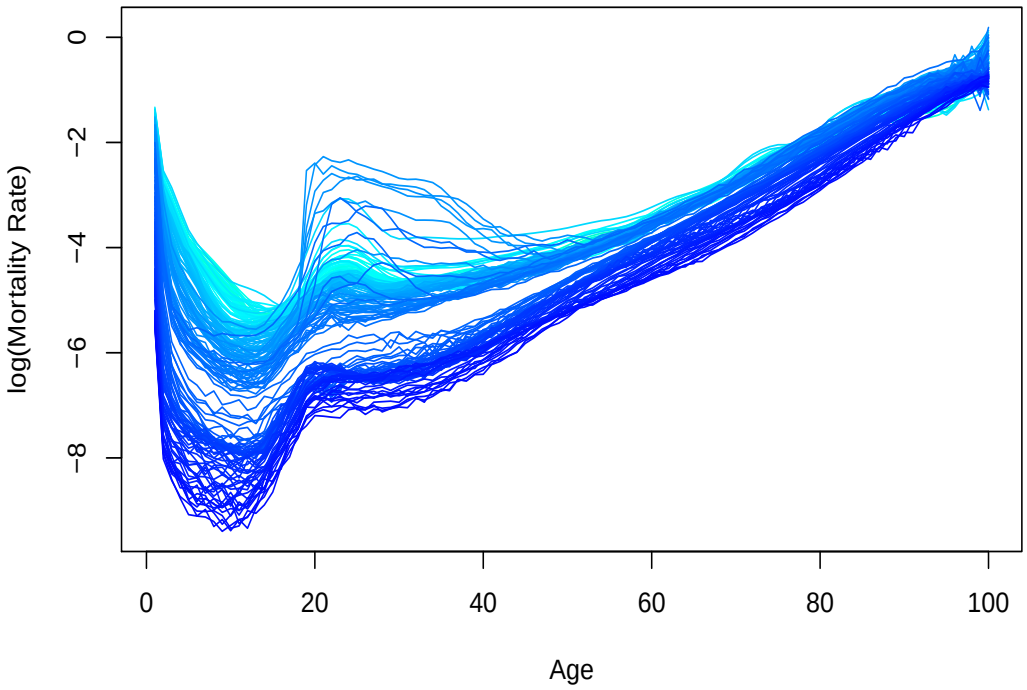


Figure 8: Male Mortality rates by age (years) in France between 1816 and 2006 (darker shade of blue for more recent years).

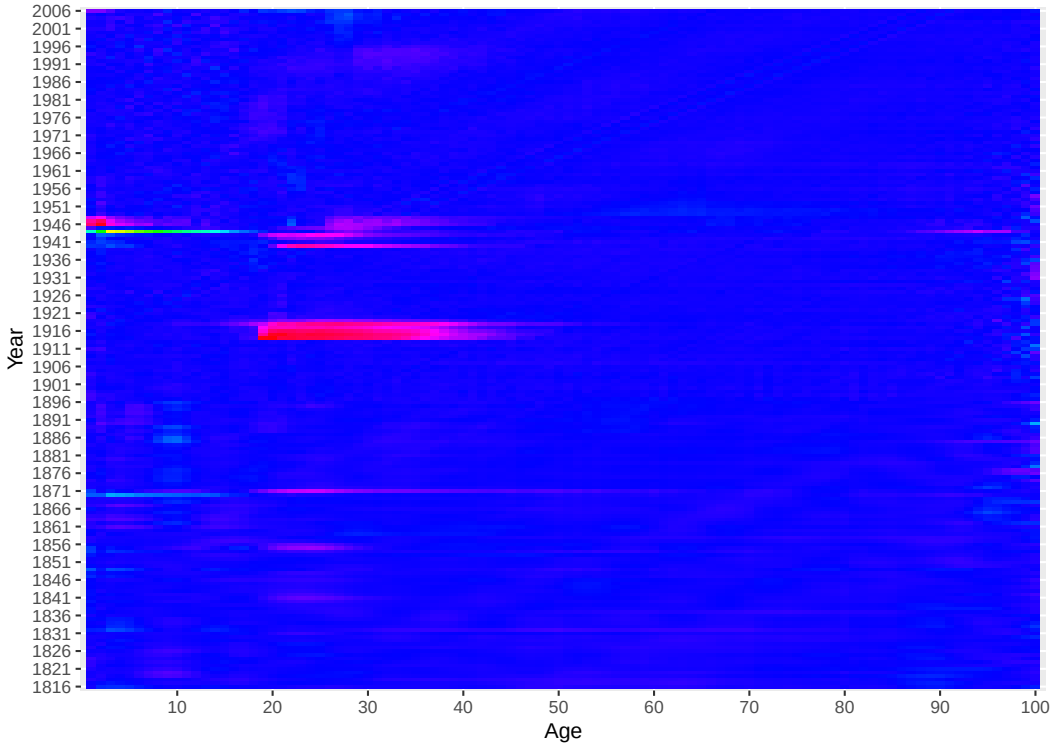


Figure 9: Heatmap for the difference $x_{ij} - \hat{x}_{ij}$ between observed and predicted values. Shades of red indicate cells in which the mortality rates are higher than expected in the age group, while shades of yellow correspond to the opposite situation.

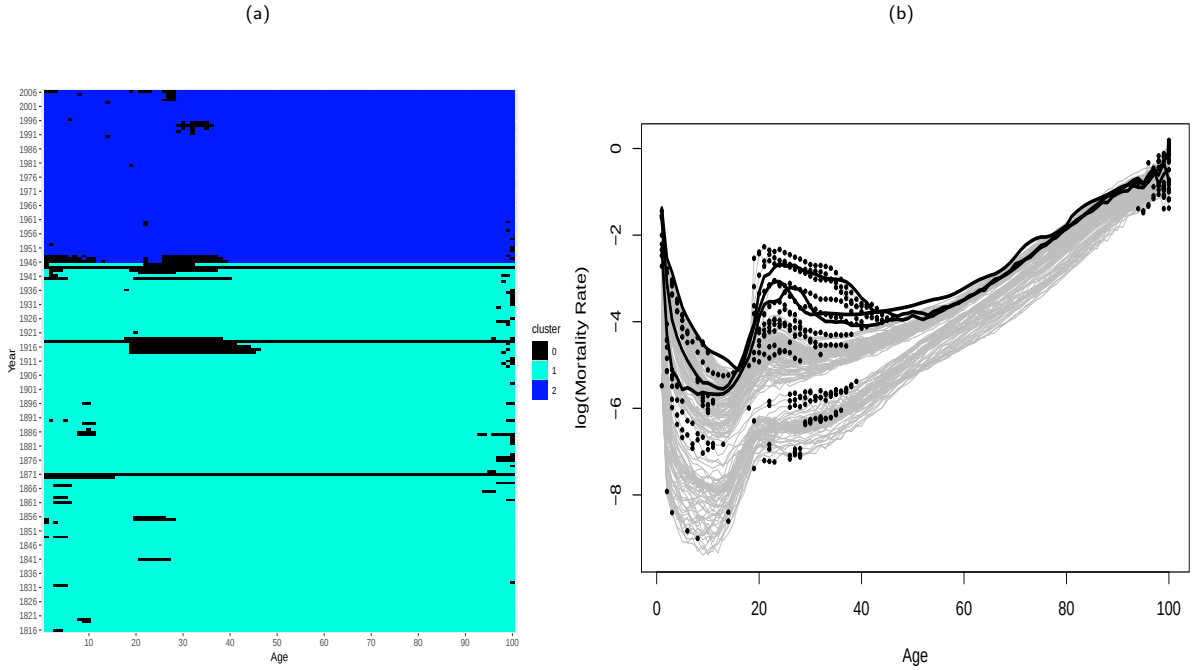


Figure 10: (a) Clusters for the mortality rate data, presented in two shades of blue. Trimmed cells are in black. (b) original mortality rate curves with trimmed cells depicted as "•". The three trimmed curves, corresponding to years 1871, 1918, and 1944, are drawn in black.

It can be seen that the periods in which mortality rates increased in France (reddish colors with higher than expected mortality rates) correspond to the Crimean War (1853-1856), the Franco-Prussian War (1870-1871), the First World War (1914-1918), and the Spanish flu (1917-1918). During the beginning of the Second World War (1939-1945) there was not a very notable increase in mortality rates in France, which was due to the rapid German occupation and the subsequent collaborationist policy. There was indeed a more marked increase in mortality in the latter part of the conflict when France was more actively involved in the war.

Figure 10(a) shows the result of applying the two types of improvement proposed. In this figure, the data matrix is depicted using two different intensities of blue to represent the final allocation to clusters, and the trimmed cells are black. It can be seen that the trimmed cells are mostly concentrated in periods of war, when mortality rates increase significantly and affect most notably the 18-40 age group fighting on the frontlines, not affecting other age groups too much. The cells trimmed from the original curves have been marked with dots in Figure 10(b). Using the second possibility of improvement that allows whole 'globally' atypical curves to be trimmed, the years 1871, 1918 and 1944 are globally trimmed and appear as continuous lines in the graph. The year 1871 corresponds to France's defeat in the Franco-Prussian War when France had to cede the territories of Alsace and Lorraine to Germany. 1918 corresponds to the First World War and the Spanish flu pandemic. The year 1918 was globally atypical because the pandemic affected all age groups, not just young soldiers fighting on the front. Finally, the year 1944 corresponds to the end of the Second World War in France, with the Normandy landing and the liberation of Paris, which implied very high mortality rates.

To facilitate the interpretation of the approximating spaces and to be able to interpret the meaning of the 'scores' better, Figure 11(a) displays the average $\boldsymbol{m}^g$ curves, showing the effect of adding (symbol '+' in red) or subtracting (symbol '−' in black) a multiple of the l-th column of the $\boldsymbol{B}_2^g$ matrix ($l = 1, 2$ in this example). This graph is

frequently used as it illustrates the result of the Principal Functional Component Analysis [37]. The l-th column of the $\boldsymbol{B}_2^g$ matrix is multiplied by twice the square root of the variance of the l-th column of the $\boldsymbol{A}_q^g$ matrix. This graph serves to summarize the variability explained by the components that generate the approximating subspaces, which we call 'modes of variation', in each of the groups detected.

The l-th column of the $\boldsymbol{A}_2^g$ matrix for $l = 1, 2$ (since $q_1 = q_2 = q = 2$) provides the coordinates or scores of each curve $\boldsymbol{x}_i$ when representing them in their approximating spaces. The interpretation of these scores will be similar to the one made in the traditional PCA except that now the approximating spaces are different for each cluster. Figure 11(b) presents a graph of the dispersion of these 'scores'. Combined with figure 11(a), valuable information can be obtained about the $\boldsymbol{x}_i$, which takes into account the structure in groups and the different modes of variation detected in them. Observations with close scores in these representations indicate similar behavior and also allow the detection of globally atypical observations within the approximation made in each cluster. Paired graphs are necessary when considering intrinsic dimensions $q_g > 2$.

We can see in this example that the largest positive values in the scores of the first component of the second group (right panel of figure 11(b), which correspond to more recent years, are associated with '+' values in the group '$g = 2$ and $l = 1$' in figure 11(a). This component seems to reflect a global evolution within cluster 2 (years 1949 to 2006) where log-mortality rates decrease steadily with the years and in a very uniform way in all age groups. Something similar can be seen when interpreting the first mode of variation $l = 1$ in cluster 1, although it is more focused on the reduction of infant mortality rates. In the graph of scores for cluster 2, the years 1946, 1947, and 1948 appear relatively isolated, and we have already commented that these represent a 'transition' between the two groups. The scores of cluster 1 allow us to visualize distinctly atypical years, such as 1944 (end of World War II) and the years 1870-1871 (Franco-Prussian war).

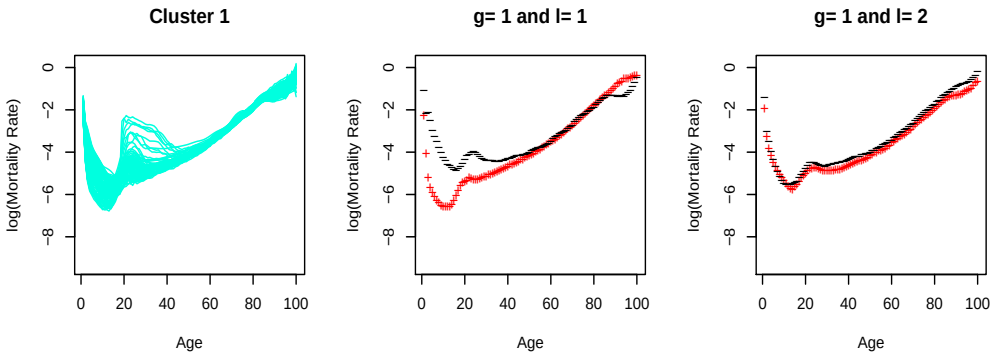
7.2. Meteorological data

We now consider the average daily temperature at 83 Spanish weather stations in the years 2007, 2008, and 2009. Data were obtained from AEMET, the Spanish State Meteorological Agency. The fact that we consider daily data over three years means that we have to work in a high dimension space ($p = 1096$).

Figure 12(a) presents a graph of these temperatures, which shows groups of curves with an approximate cyclic pattern over the three years. This graph also shows local patterns that may be related to possible 'waves' of heat or cold. These waves can be global (affecting the whole Iberian Peninsula and even the Canary Islands), but also, in many cases, these waves affect exclusively particular areas of the country (with specific climate and geographical characteristics). The proposed robust cluster analysis approach will attempt to detect anomalous seasonal temperatures taking into account the behavior within the cluster to which the observation is assigned. That is to say, for instance, if in a zone (cluster) and at a specific time there is a temporary heatwave, we will only mark as atypical data from stations within this zone that have an exaggerated temperature value compared with what is expected in the subspace model in that zone, or it may even be considered atypical if the heatwave is not very noticeable This is different from other approaches that only look for atypical values without modeling and accounting for the collective behavior of the observations within the same group.

To illustrate the methodology, clearly atypical artificial cell values have been introduced in the data set. The observed temperatures have been replaced by values equal to 0°C. This simulates cases in which measuring instruments

(a)



(b)

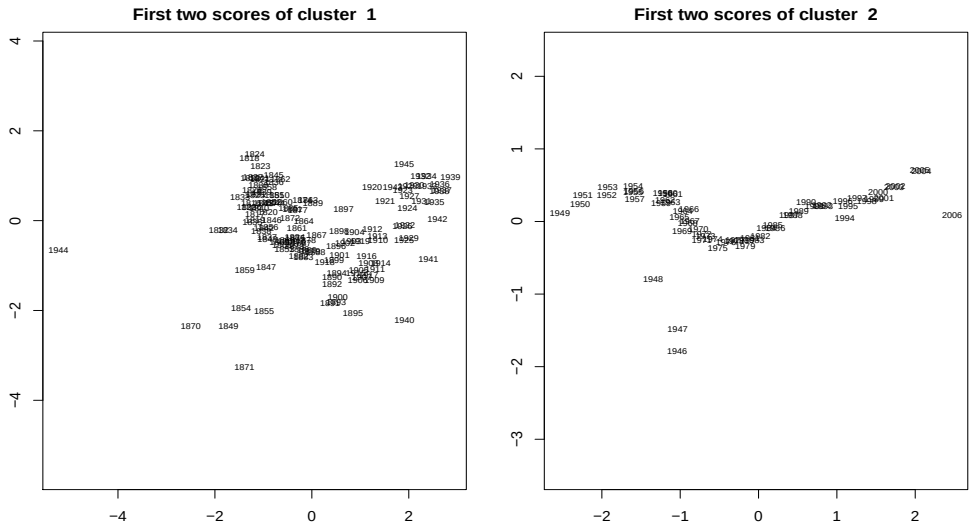


Figure 11: (a) Clusters and different modes of variation. (b) dispersion graph for scores.

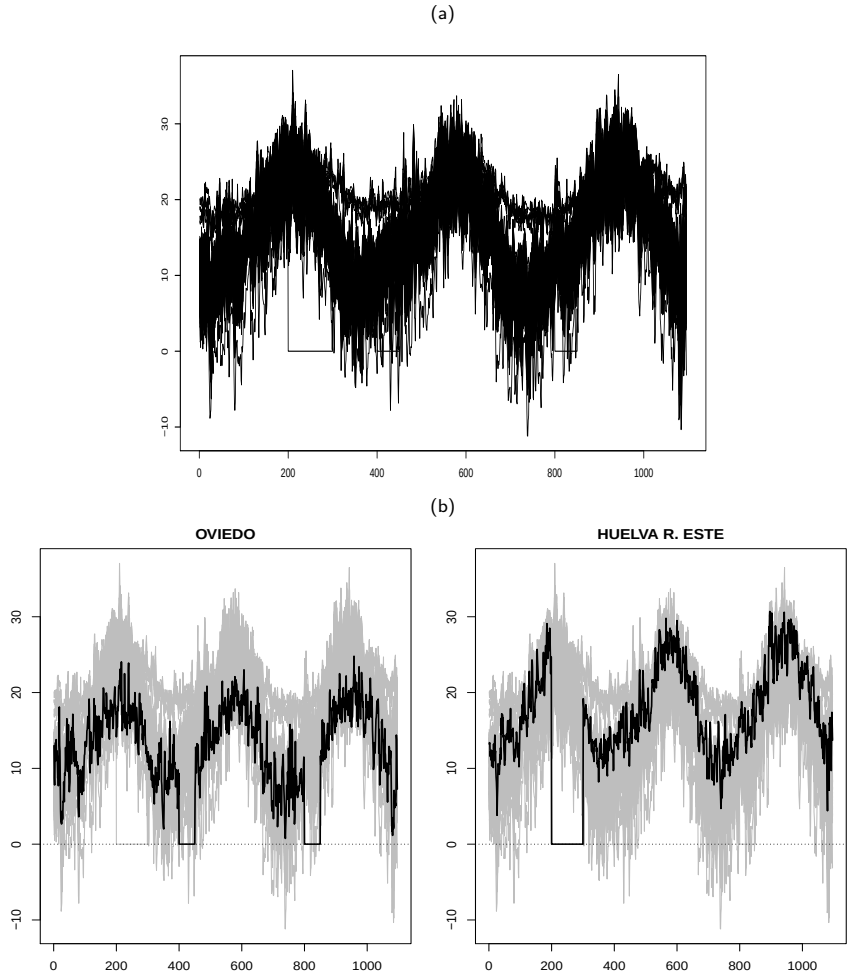


Figure 12: Daily average temperature at 93 Spanish weather stations in 2007, 2008, and 2009 with modified values of 0° for two curves. These artificially modified values for the stations of 'Huelva R.este' and 'Oviedo' are better appreciated in panel (b).

do not work properly and return a default value, for example, equal to 0°C. In particular, the measurements of 100 consecutive days (slightly more than three months) in autumn 2007 at the 'Huelva R. este' weather station have been replaced. This has also been done in the 'Oviedo' station in two different periods, changing 50 consecutive days (a little more than a month and a half) in 2008 and another 50 consecutive days in 2009, also by 0°C. In figure 12(b), these replaced values can be seen for the stations of 'Huelva R. este' and 'Oviedo'. The idea is to check whether the procedure is capable of detecting these 'altered' measurements and whether it is capable of reasonably approximating the real values, taking into account temperature observations from other stations in the same group, and information from the untrimmed cells in that temperature curve.

We applied the procedure with $G = 4$, $q = 2$, and $\alpha = 0.1$, together with a `lowess` smoothing with a window of 10% of contiguous observations, 10 internal nodes ($P = 14$), $\alpha_{\text{TCLUST}} = 0.03$ and $\alpha_{\text{LTS}} = 0.03$. The average curves $\boldsymbol{m}^g$ for the four groups are shown in Figure 13(a) and the geographical position of the weather stations, using different colors to show the cluster assignment, in Figure 13(b). The 4 clusters found correspond essentially to

1. *Cluster 1* (red): Weather stations with cold winters (sometimes you can see average daily winter temperatures that can even be negative) but with hot summers. Later we will see that they correspond essentially to stations in the northern and southern plateaus with a relatively extreme continental climate in the winters and summers.

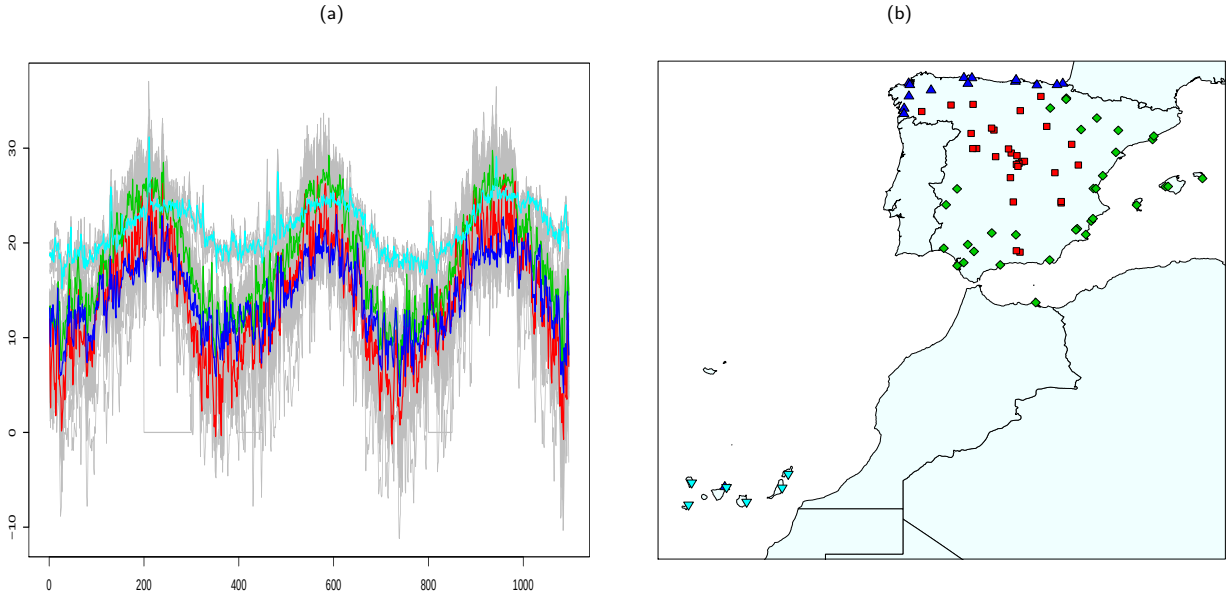


Figure 13: (a) Average curves for the four clusters. (b) Geographic position of the weather stations with their cluster assignment, using the same colors as the corresponding average curves of the cluster to which they were assigned.

2. *Cluster 2* (green): Stations with higher maximum and minimum temperatures where the frosts are unusual and with quite high temperatures in the summer. These stations correspond to areas of southern Spain or with a Mediterranean climate.
3. *Cluster 3* (blue): Stations located in the north of Spain, most of them on the Cantabrian coast, with an Atlantic climate, with winters that are not as cold as in Cluster 1 and summers that are milder than in Clusters 1 and 2. The proximity to the sea and abundant rainfall temper these temperatures
4. *Cluster 4* (cyan): This group includes the stations of the Canary Islands, where the average temperature is mild and temperate throughout the year between 15 and 25°C with less variation attributable to the weather station.

There are some exceptions to this described weather behavior. The station of 'Tenerife Norte' is assigned to cluster 3 of the Atlantic-Cantabrian climate, to which it does not correspond geographically, but when this curve is represented, we can see that its behavior is quite close to this group and that it is different to the other stations in cluster 4. Something similar happens with the stations 'Granada' and 'Granada Air (airport)', which do not have warm winters, despite being in the south of Spain, due to their height and proximity to the Sierra Nevada, and this makes them end up assigned to cluster 1. Figure 14 shows the stations in a relief graph in which the height of the weather stations in the country can be seen.

Figure 15 describes the assignment of weather stations to clusters using the same color scheme as in other graphs and marking the trimmed cells after the final improvement process in black. Although the initial trimming level $\alpha = 0.1$ is high, using the ordered graphic of the R_{ij}, we would see that a fraction of 0.3% of cells to be cut would be a more reasonable choice. This proportion of 0.3% trimmed cells includes all artificially introduced values for the 'Huelva R. Este' and 'Oviedo' stations (larger black regions), which are therefore not used in determining the approximate subspaces of their respective clusters. Other cells have been trimmed, which are due to individual outliers (from the approximately adjusted subspace) and may require attention. Many of these cells detected as outliers, for example, are concentrated in the station 'Puerto de Navacerrada', which is close to Madrid, but can be quite particular due to

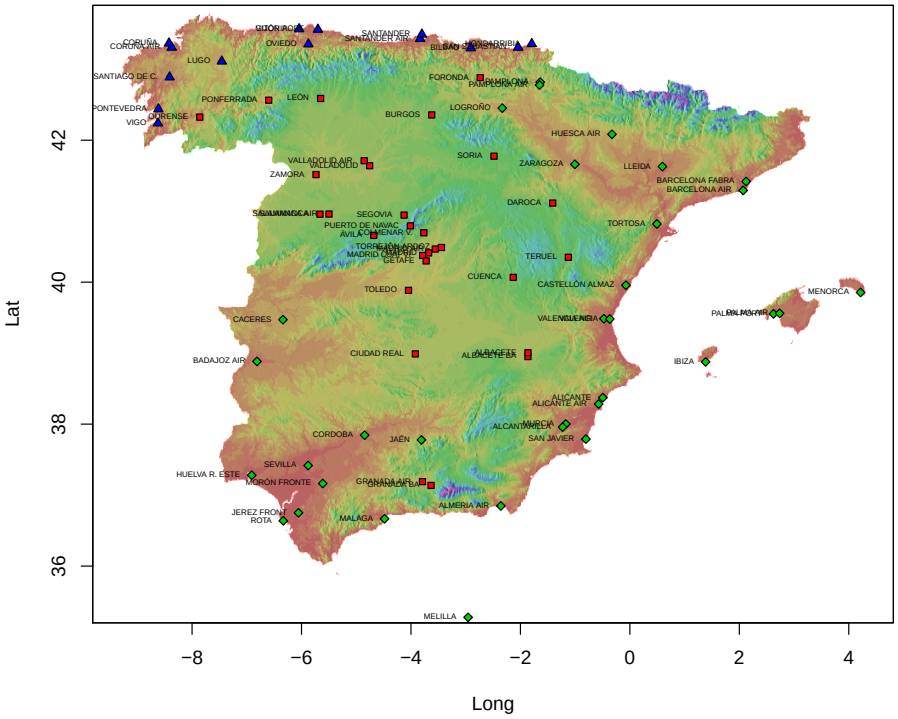


Figure 14: Allocation of weather stations (excluding the Canary Islands) to clusters on a map showing the height at which the station is located

its situation at a higher altitude, with more extreme and changing weather.

Figure 15 presents the scores of the 83 meteorological stations. We can see that stations with similar temperatures (often due to apparent geographical proximity) are usually in the same group and have similar scores. We can also distinguish some more atypical stations within the groups, such as 'Puerto de Navacerrada' in cluster 1 and 'Tenerife Norte' in cluster 3.

Figure 17 shows the predicted curve (using $\widehat{x}_{ij}$), and the observed values for 'Oviedo', predicted values in red, actual values in green. We can see that the two curves are very close, even for the segments that were artificially replaced with 0° values. The right-hand panels give a closer look at these two intervals. It is important to note that for this reconstruction, the score has been used, which is robustly determined with the untrimmed cells. This imputation, dependent on the cluster structure and the approximating spaces, seems more reasonable than a global imputation that ignores the group structure and its different 'modes of variation'.

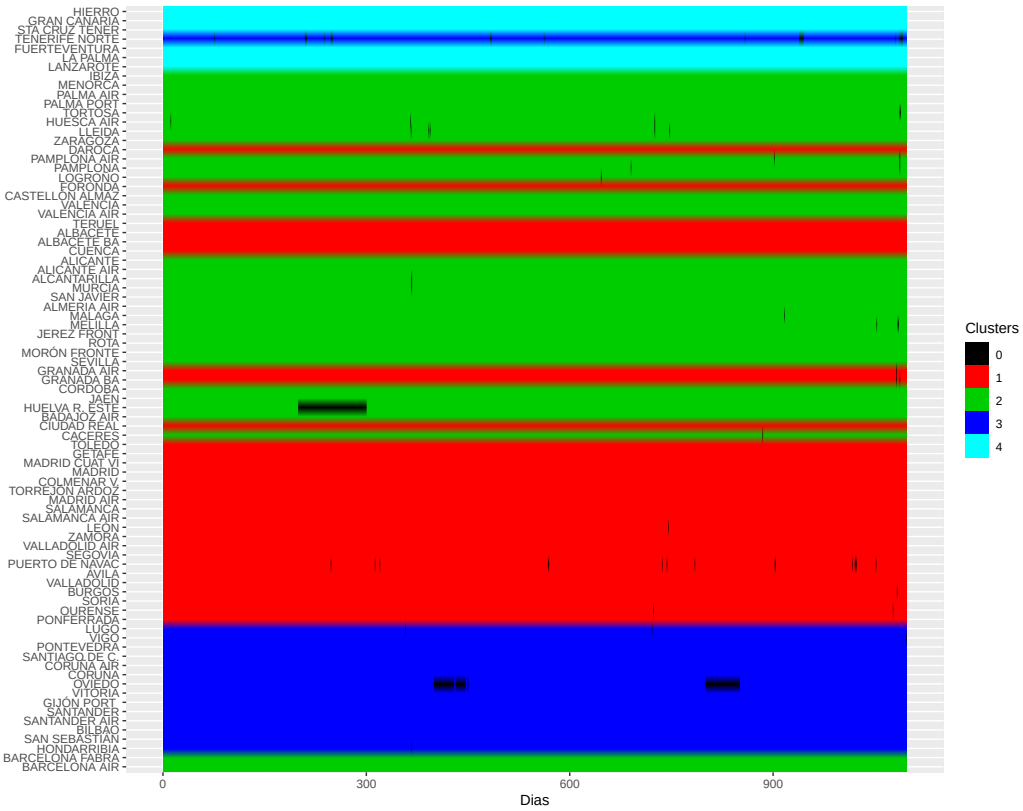


Figure 15: Allocation of weather stations to clusters (using the same color scheme). The trimmed cells after the refinement process are marked in black.

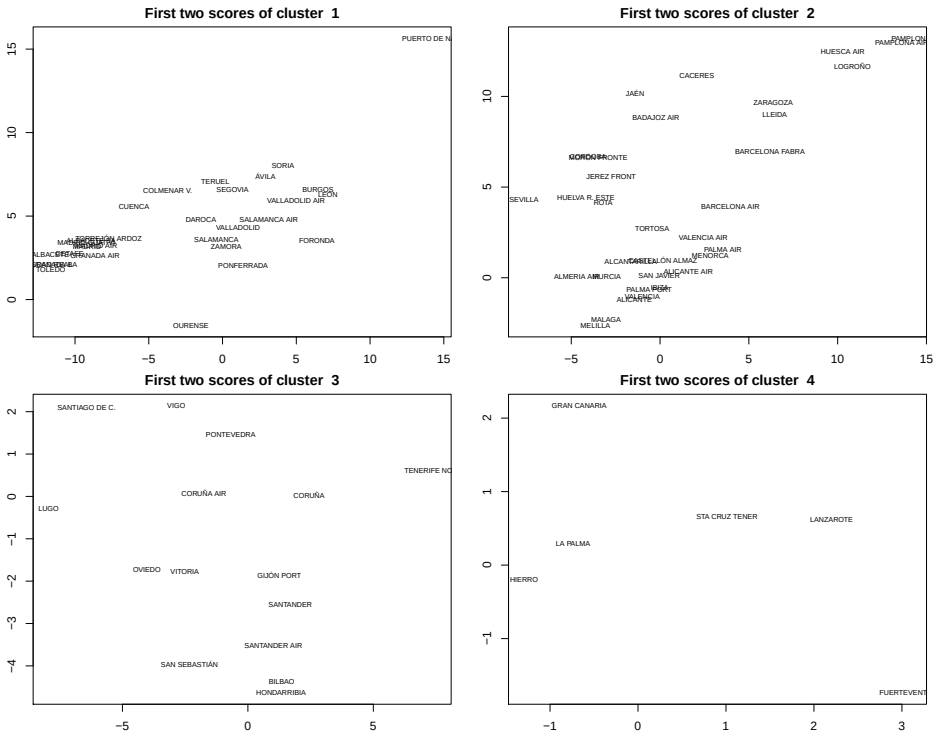


Figure 16: Scores of the 83 weather stations in their corresponding clusters.

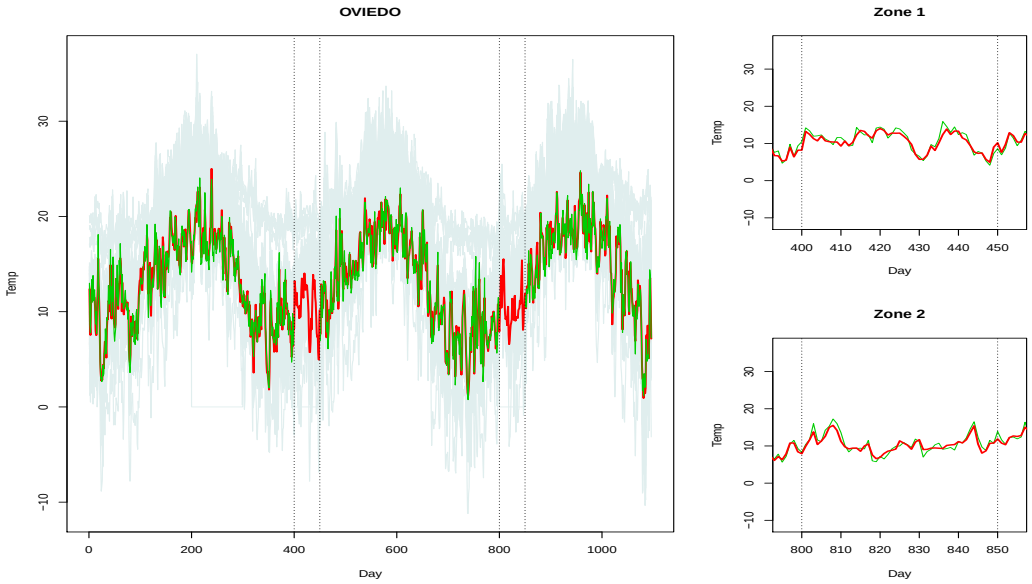


Figure 17: In all panels, the red line shows the predicted temperature curve for the Oviedo station. The green line shows the actual values that were used by the method (left panel) and also those not used because they were replaced by 0°C values (right panels).

8. Conclusions and future work

We have presented a robust Cluster Analysis methodology based on cell trimming. This approach is particularly interesting in problems of large or moderate size since it avoids the important loss of information that occurs when trimming entire observations with few outliers. Using cluster membership information seems reasonable when marking these outliers. Robust fitting of approximating subspaces, which serve to describe the structure of variability within each cluster, is very useful in this automated detection, and it seems logical that the determination of these subspaces and outliers be done in a unified way because the two problems are interrelated. An algorithm has been proposed along with examples of the applicability of the algorithm in simulated and real data.

This work is a first approach to the problem and there are many open lines of work that need to be addressed in the future. For example, although the iterative procedure by alternating regressions with weights is computationally feasible for not too high intrinsic dimensions, it has been verified that its effectiveness requires reasonable initializations. A proposal for initialisation in the functional case has been provided but it would be interesting to establish alternative procedures in other situations.

Another interesting line of work is to implement procedures that help the user to set the multiple parameters needed: the number of groups G, the trimming size α and the intrinsic dimensions q_g of the approximating subspaces, and other tuning parameters of the algorithm.

The problem of determining the number of groups G is obviously complex, as is already the case with simpler (robust and non-robust) cluster analysis problems. Furthermore, it is known that this determination of G is, in many cases, dependent on the final goal for the user. The problem is even more involved in the case of trimming because the α and G parameters could interact. Techniques based on monitoring changes in the target function (3.1) when moving α and G, such as those already applied in [20] and [22], could be usefully adopted. It has been shown in this work that 'retrieving' incorrectly trimmed cells in a final phase of improvement is a possibility to take into account

and that it makes the choice α less critical. The determination of intrinsic dimensions is another important and not at all trivial problem. Monitoring the sample variances of the columns of the $\boldsymbol{A}_{q_g}^g$ matrices may be useful, as is the case in the classical PCA when the [7] procedure is followed.

Another interesting line of work would be to explore the use of information in the scores to temporarily trim the 'less reliable' observations (because they have poorly integrated scores in their groups) in the iterative process of the algorithm. This would result in an algorithm that would combine the proposed methodology with the strengths of the trimmed k-means algorithm. Considering regularization techniques in the iterative process could also be useful in applications where it is desired to have more interpretable 'modes of variation', which are estimated with less variability.

It would be important to implement a library (in `R`) to apply this methodology in a user-friendly way that presents all the resulting information and allows the user to explore the effect of the choice of parameters.

Acknowledgements

This research was partially supported by Spanish Ministerio de Economía y Competitividad, Grant MTM2017-86061-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León and FEDER, Grant VA005P17 and VA002G18.

References

[1] Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37:311–331.

[2] Bellas, A., Bouveyron, C., Cottrell, M., and Lacaille, J. (2012). Robust clustering of high-dimensional data. *ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 25–27.

[3] Boente, G. and Salibian-Barrera, M. (2015). S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, 110:1100–1111.

[4] Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519.

[5] Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5:281–300.

[6] Brunet-Saumard, C. and Bouveyron, C. (2014). Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*, 29:489–513.

[7] Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behaviour Research*, 1:245–276.

[8] Cevallos-Valdiviezo, H. (2016). *On methods for prediction based on complex data with missing values and robust principal component analysis*. PhD thesis, Ghent University.

[9] Cevallos-Valdiviezo, H. and Van Aelst, S. (2019). Fast computation of robust subspace estimators. *Computational Statistics & Data Analysis*, 134:171–185.

[10] 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.

[11] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

[12] Cuesta-Albertos, J. and Fraiman, R. (2007). Impartial trimmed k-means for functional data. *Computational Statistics & Data Analysis*, 51:4864–4877.

[13] Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25:553–576.

[14] Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38:1171–1193.

[15] Engelen, S., Hubert, M., and Branden, K. V. (2005). A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics*, 34:117–126.

[16] Farcomeni, A. (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, 56:102–111.

[17] Farcomeni, A. (2014b). Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing*, 24:907–919.

[18] Fauvel, M., Bouveyron, C., and Girard, S. (2015). Parsimonious gaussian process models for the classification of hyperspectral remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 12:2423–2427.

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

[20] García-Escudero, L. A., Gordaliza, A., and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statitistics*, 12:434–449.

[21] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36:1324–1345.
[22] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, 21:585–599.
[23] García-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. (2015). Comments on "Multivariate functional outlier detection" by M. Hubert, P. Rousseeuw and P. Segaert. *Statistical Methods and Applications*, 24:233 – 235.
[24] García-Escudero, L. A., Gordaliza, A., Ruwet, C., and Martín, S. (2017). Robust principal component analysis based on trimming around affine subspaces. *Statistica Sinica*, 27:1437–1459.
[25] García-Escudero, L. A., Gordaliza, A., San Martín, R., Van Aelst, S., and Zamar, R. (2009). Robust linear clustering. *Journal of the Royal Statistical Society. Series B*, 71:301–318.
[26] Gattone, S. A. and Rocci, R. (2012). Clustering curves on a reduced subspace. *Journal of Computational and Graphical Statistics*, 21:361–379.
[27] Hitchcock, D. and Greenwood, M. (2015). Clustering functional data. In C. Hennig, M. Meila, F. M. and Rocci, R., editors, *Handbook of Cluster Analysis*. Chapman and Hall/CRC.
[28] Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*, 24:177–202.
[29] Hubert, M., Rousseeuw, P. J., and Van den Bossche, W. (2019). MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, 61:459–473.
[30] Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21:618–637.
[31] Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19:29–45.
[32] Jacques, J. and Preda, C. (2013). funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171.
[33] Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8:231–255.
[34] Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3:1–58.
[35] Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics*, 47:264–273.
[36] Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: A review. *ACM SIGKDD explorations newsletter*, 6:90–105.
[37] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York.
[38] Rivera-García, D., García-Escudero, L. A., Mayo-Iscar, A., and Ortega, J. (2019). Robust clustering for functional data based on trimming and constraints. *Advances in Data Analysis and Classification*, 13:201–225.
[39] Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.
[40] Rousseeuw, P. and Van Driessen, K. (2000). An algorithm for positive-breakdown methods based on concentration steps. In *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346, New York.
[41] Rousseeuw, P. J. and Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, 60:135–145.
[42] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York.
[43] Van Aelst, S., Wang, X., Zamar, R., and Zhu, R. (2006). Linear grouping using orthogonal regression. *Computational Statistics & Data Analysis*, 50:1287–1312.
[44] Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, 28:52–68.
[45] Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, 6:219–247.
[46] Yassouridis, C. and Leisch, F. (2017). Benchmarking different clustering algorithms on functional data. *Advances in Data Analysis and Classification*, 11:467–492.
[47] Zamar, R. H., Yohai, V. J., Leung, A., and Agostinelli, C. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, 24:441–461.