



HAL
open science

Multiple testing for outlier detection in functional data

Clémentine Barreyre, Béatrice Laurent, Jean-Michel Loubes, Bertrand Cabon,
Loïc Boussof

► **To cite this version:**

Clémentine Barreyre, Béatrice Laurent, Jean-Michel Loubes, Bertrand Cabon, Loïc Boussof. Multiple testing for outlier detection in functional data. 2017. hal-01651191

HAL Id: hal-01651191

<https://hal.science/hal-01651191>

Preprint submitted on 13 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multiple testing for outlier detection in functional data

C. Barreyre^a, B. Laurent^b, J.-M. Loubes^c, B. Cabon^a and L. Boussouf^a

^aAirbus Defence and Space, Z.I Palays, 31 rue des Cosmonautes, 31400 Toulouse, France

^bInstitut des Mathématiques de Toulouse (UMR 5219), INSA Toulouse, Université de Toulouse, 135 avenue de Rangueil, 31400 Toulouse, France

^cInstitut des Mathématiques de Toulouse (UMR 5219), Université Paul Sabatier, Université de Toulouse, 118 route de Narbonne, F-31062 Toulouse Cedex 9, France

ARTICLE HISTORY

Compiled December 13, 2017

ABSTRACT

We propose a novel procedure for outlier detection in functional data, in a semi-supervised framework. As the data is functional, we consider the coefficients obtained after projecting the observations onto orthonormal bases (wavelet, PCA). A multiple testing procedure based on the two-sample test is defined in order to highlight the levels of the coefficients on which the outliers appear as significantly different to the normal data. The selected coefficients are then called features for the outlier detection, on which we compute the Local Outlier Factor to highlight the outliers. This procedure to select the features is applied on simulated data that mimic the behaviour of space telemetries, and compared with existing dimension reduction techniques.

KEYWORDS

Functional data ; Two sample tests ; Outlier detection ; Local Outlier Factor ; Multiple testing

1. Introduction

In this paper, we propose a novel procedure for outlier detection in a set of functional data.

Detecting outliers has become an increasing challenge in many areas, such as network intrusion detection, fraud detection, medical anomaly detection, and failure detection, as it was described by Chandola [1]. An outlier is basically a data that is significantly different from the normal behavior. In addition, several anomalies do not necessarily exhibit similar characteristics. Hence, detecting anomalies must be done by defining the normal behavior in the first place. Then, the deviation measured between an individual and the normal behavior gives good indications of anomalousness.

However, as noticed in the same paper [1], defining a normal region that encompasses all the possible normal behaviors is sometimes really difficult. Moreover, an anomaly does not appear necessarily on all the explanatory variables, especially when the data is high-dimensional.

In the framework of this paper, the normal behavior can be partially learned thanks

to a semi-supervised approach. It means that we can isolate a subset of data that do not contain any anomaly and that will be referred as the nominal set. We therefore have two sets of data, a nominal one and another one containing a small proportion of anomalies, that we want to detect. This approach is adapted when a first set of data has already been analyzed and reviewed manually.

Semi-supervised outlier detection was already treated by [2] and [3] where the normal behaviour is learned thanks to fixed-background and Gaussian mixture models.

We want to detect outliers from functional data since this work is motivated by applications to satellites telemetries, which are measurements of thousands of health parameters through time, hence functional data. Those telemetries exhibit daily-periodic patterns, thus can be logically split into days, leading to consider functions defined on intervals.

We suppose that we have n days of a telemetry that is regularly sampled on p times each day. We assume that our observations are corrupted by independent and identically distributed (i.i.d.) Gaussian noise. This corresponds to the following model:

$$X_{i,j} = f_i(t_j) + \varepsilon_{i,j}, \quad i = 1 \dots n, \quad j = 1, \dots, p, \quad (1)$$

where f_i is originally defined on a compact set, that can be modeled, without loss of generality, $f_i : [0, 1] \mapsto \mathbb{R}$. The variables $(\varepsilon_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ are i.i.d. centered Gaussian variables with variance σ^2 which is unknown. Since the telemetries are regularly sampled, we assume that for all $j = 1, \dots, p$, $t_j = j/p$.

Outliers detection applied to functional data was already treated for instance by Hoffmann [4], where the Kernel Principal Component Analysis (KPCA) decomposition is used for detecting novelties in hands digits. Ordoñez [5] applied functional data analysis methods on GPS measurements, and the outlier detection is done thanks to the depth metrics. Ren [6] proposed to use projections coupled with high-breakdown mean function estimator to detect outliers.

Our approach relies on three main steps. Firstly, we project the data onto orthonormal bases and collect the coefficients resulting from these projections. This approach is really common to cluster functional data, such as in [7] and [8] where coefficients arising from functional basis are used: Reproducing Kernel Hilbert Space (RKHS), wavelets, Functional Principal Component Analysis (FPCA). One can also use the standard Principal Component Analysis (PCA) on the vectors of observations. We choose to project onto the Haar wavelet basis and the principal component basis.

In a second step, we select with a multiple testing procedure the projection coefficients that highlight anomalies. The selected coefficients are called features for anomaly detection.

Once the features are selected, we compute in a third step the Local Outlier Factor (LOF) [9] on these features to detect the abnormal behaviours and we compare our results with some common procedures to select features from observations of functional data. We choose this method because it is well adapted to the data we have. Other methods could be chosen, as the ones described by Barnett [10], such as One-Class SVM [11], One-Class Random Forest [12], or depth methods [13], for instance.

We then apply our procedure on simulated data that mimic the behaviour of space telemetries.

The paper is organized as follows. In the first part, we detail two orthonormal bases that are widely used : a wavelet basis and the principal component basis. In the next section, we apply two-sample tests on each level of coefficients, and then keep the ones that are rejected after controlling the false discovery rate thanks to the Benjamini-

Hochberg procedure [14]. We consider the Kolmogorov-Smirnov test, as well as two other tests built on Wasserstein metrics, defined for instance in [15].

The last section is dedicated to an application on a simulated telemetry reproducing faithfully the behaviour of a real telemetry, and containing all the different types of anomalies that can be observed. The outlier detection method is done thanks to the Local Outlier Factor (LOF) computation. We show in this section the good performances of our procedure to select the best coefficients for outlier detection. We compare it with more classical dimension reduction methods. We also apply the outlier detection on the raw-data, as it was done in [16], to show that this naive method is not powerful in this framework.

2. Definition of the features

The observations of functional data are high-dimensional, hence it is important to reduce the dimension to concentrate the information.

We recall that our observations obey to Model (1) where the variables $(\varepsilon_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ are i.i.d. centered Gaussian variables with variance σ^2 and for all j , $t_j = j/p \in [0, 1]$. We suppose that the nominal days are the n_0 first ones, where $n_0 < n$.

There exist many ways to represent functional data in a reduced dimension. Some bases are more adapted to some types of functions. Selecting the best basis is in fact a real challenge that has been widely addressed, see [7], for example.

We consider two types of features, the first ones are obtained from a projection onto an orthonormal basis of $\mathbb{L}^2([0, 1])$, the second one correspond to a Principal Components Analysis.

2.1. Projection onto the Haar basis

We assume that for all $i = 1 \dots n$, $f_i \in \mathbb{L}^2([0, 1])$, and we denote by $\langle \cdot, \cdot \rangle$ the usual scalar product in $\mathbb{L}^2([0, 1])$.

$$\langle f, g \rangle = \int_0^1 f(t)g(t)dt.$$

The functions f_i can be represented in an orthonormal functional basis of $\mathbb{L}^2([0, 1])$. See [17] for more details. If $(\phi_\lambda)_{\lambda \in \mathbb{N}^*}$ is an orthonormal basis in $\mathbb{L}^2([0, 1])$,

$$f_i(t) = \sum_{\lambda=1}^{\infty} \theta_{i,\lambda} \phi_\lambda(t), \text{ with } \theta_{i,\lambda} = \langle f_i, \phi_\lambda \rangle. \quad (2)$$

From the observations obeying to Model (1), the coefficients $\theta_{i,\lambda}$ are estimated by their empirical counterparts

$$\hat{\theta}_{i,\lambda} = \frac{1}{p} \sum_{j=1}^p X_{i,j} \phi_\lambda(t_j). \quad (3)$$

We focus here on a wavelet basis, namely the Haar basis of $\mathbb{L}^2([0, 1])$. Let us first recall its definition. We set $\phi_0 = \mathbb{1}_{[0,1]}$ and $\psi = \mathbb{1}_{[0,1/2[} - \mathbb{1}_{[1/2,1[}$. For all $l \geq 0$,

$k \in \Lambda(l) = \{0, 1, \dots, 2^l - 1\}$, let $\phi_{l,k}(x) = 2^{l/2}\psi(2^l x - k)$. The functions $(\phi_0, \phi_{l,k}, j \geq 0, k \in \Lambda(l))$ form the orthonormal Haar basis of $\mathbb{L}^2([0, 1])$.

A function $f \in \mathbb{L}^2([0, 1])$ can be represented by its expansion onto the Haar basis :

$$f(t) = \alpha\phi_0(t) + \sum_{l \geq 0} \sum_{k \in \Lambda(l)} \beta_{l,k}\phi_{l,k}(t).$$

See Daubechies [18] for more details and other examples of wavelet bases.

For the sake of simplicity, we assume that p is a power of 2, namely $p = 2^{J+1}$. Since we only have p regularly spaced observations per curve, we only keep the p first wavelet coefficients, which corresponds to all the coefficients up to the level $l = J$. We define

$$\Lambda = \{0\} \cup \{\lambda = (l, k), 0 \leq l \leq J, k \in \Lambda(l)\}, \quad (4)$$

and

$$\{\phi_\lambda, \lambda \in \Lambda\} = \{\phi_0, \phi_{l,k}, 0 \leq l \leq J, k \in \Lambda(l)\}.$$

Hence, our initial set of data from Model 1 is represented by the same number of coefficients $(\hat{\theta}_{i,\lambda})_{\lambda \in \Lambda}$, $1 \leq i \leq n$. We can of course recover exactly the initial data $(X_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ from the coefficients. For the next steps, we need the independence of the random variables $(\hat{\theta}_{i,\lambda}, \lambda \in \Lambda, 1 \leq i \leq n)$ with respect to λ and to i . This is indeed the case for the three following properties :

- The Haar basis is orthonormal with respect to the discrete scalar product :

$$\forall \lambda, \lambda' \in \Lambda, \frac{1}{p} \sum_{j=1}^p \phi_\lambda(t_j)\phi_{\lambda'}(t_j) = \delta_{\lambda,\lambda'},$$

where $\delta_{\lambda,\lambda'} = 0$ if $\lambda \neq \lambda'$ and 1 otherwise.

- For all i , the vector $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,p})'$ is a Gaussian vector.
- The vectors $(\mathbf{X}_i, 1 \leq i \leq n)$ are independent.

In this case, the feature selection will consist in finding the indexes $\lambda \in \Lambda$ of interest in the decomposition. Let us now introduce the other set of features that we consider.

2.2. Principal component basis

The principal component analysis (PCA) is a very powerful way to reduce the dimension of the data by finding the vectors that recover the best the variance of the data. As this basis depends on the data, we will apply it on a subset I_0 of the data known as nominal. We choose I_0 the set of even indexes $\{2i, 1 \leq i \leq n_0/2\}$. Indeed, in order to get independent features as previously, we will compute the principal components only on a subset of $n_0/2$ nominal days and project the remaining data on this orthonormal basis.

Let $\bar{\mathbf{X}}_{I_0} = \frac{1}{n_0/2} \sum_{i \in I_0} \mathbf{X}_i$ be the mean of the observations. The PCA finds the eigen-

vectors of the matrix

$$\mathbf{\Gamma}_{n_0/2} = \frac{1}{n_0/2} \sum_{i \in I_0} (\mathbf{X}_i - \bar{\mathbf{X}}_{I_0})(\mathbf{X}_i - \bar{\mathbf{X}}_{I_0})'$$

Let $(\Phi_\lambda)_{\lambda=1\dots p}$ be an orthonormal family of vectors of \mathbb{R}^p built with the eigen vectors of $\mathbf{\Gamma}_{n_0/2}$, ordered by decreasing eigenvalues. This family is orthonormal for the scalar product $\langle \cdot, \cdot \rangle_p$ in \mathbb{R}^p defined by :

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^p, \langle \mathbf{u}, \mathbf{v} \rangle_p = \frac{1}{p} \sum_{j=1}^p u_j v_j.$$

Finally, we project the vectors \mathbf{X}_i on the subspace generated by the vectors $(\Phi_\lambda)_{\lambda=1\dots p}$, in order to get

$$\mathbf{X}_i = \sum_{\lambda=1}^p \hat{\theta}_{i,\lambda} \Phi_\lambda$$

where $\hat{\theta}_{i,\lambda} = \langle \mathbf{X}_i, \Phi_\lambda \rangle_p = \frac{1}{p} \sum_{j=1}^p X_{i,j} \Phi_{\lambda,j}$. We still have the independence of the random variables $(\hat{\theta}_{i,\lambda}, \lambda \in \Lambda, i \notin I_0)$ with respect to λ and i . This is why we need to split the nominal set in two parts in this case.

The feature selection will consist in finding the levels λ of interest in this decomposition.

Once we have presented the features that we consider, we will now see how to reduce the dimension by selecting a small set of features, in order to have good performances for the outlier detection procedure.

3. From approximation coefficients to features selection

We have represented our initial set of data by the approximation coefficients. Our aim is now to select a small set of coefficients in order to reduce the dimension before applying an outlier detection procedure. Indeed, if we apply directly the outlier detection procedure on the whole set of coefficients, we obtain weak results since the anomalies are not enough separated from the noisy observations, as we will see in the simulation study. There are several ways to reduce the dimension. If the coefficients are naturally ordered (this is the case for the PCA), usual procedures consist in keeping the d (with $d < p$) first coefficients in order to maximize the variance explained by those features. For projections onto wavelet bases, we can consider linear procedures of dimension reduction : they consist in keeping all the coefficients up to some level $l = J_0 < J$. We can also consider nonlinear wavelet thresholding procedures that consist in keeping the d largest wavelet coefficients. These procedures have been proved to be very performant for regression function or density estimation procedures, see [19] for more details.

For our purpose, that is to recover abnormal behaviours in our set of curves, it is not necessarily relevant to keep the first principal components or the first levels in a wavelet decomposition. Moreover, the thresholding procedure is not well adapted here since the largest coefficients for one day (a value of i) are not necessarily the largest

ones for another day. We therefore propose a new procedure, based on multiple testing that is well adapted to the problem at hand.

3.1. Univariate testing at each level

For many bases, such as the PCA, the usual dimension reduction can be done by taking the d first components that sum up a given percentage of the variance. It is known that, for the clustering usecases for instance, it is a very good way to reduce the dimension, see [20] for example. It has already also been applied for anomaly detection usecases, as in [21]. However, no one can assert that an outlier will appear as different in the first principal components. The same objection also holds for projection onto wavelet bases : the first levels do not necessarily contain interesting information for outlier detection.

This is why we have defined a new way to select the features, based on statistical tests, which is adapted to the detection of outliers.

We remind that we have represented for all $1 \leq i \leq n$, our vector of data \mathbf{X}_i by the features $(\hat{\theta}_{i,\lambda})_{\lambda \in \Lambda}$ with $\Lambda = \{1, \dots, p\}$ for the PCA decomposition and Λ is defined by (4) for the Haar decomposition.

For each level $\lambda \in \Lambda$ we would like to know if the vector $\hat{\boldsymbol{\theta}}_{\cdot\lambda} = (\hat{\theta}_{1,\lambda}, \dots, \hat{\theta}_{n,\lambda})$ contains relevant information on outliers. The vector $\hat{\boldsymbol{\theta}}_{\cdot\lambda}$ is divided into two parts : $\hat{\boldsymbol{\theta}}_{0\lambda} = (\hat{\theta}_{1,\lambda}, \dots, \hat{\theta}_{n_0,\lambda})$ corresponding to the set that is known to be nominal and $\hat{\boldsymbol{\theta}}_{1\lambda} = (\hat{\theta}_{n_0+1,\lambda}, \dots, \hat{\theta}_{n,\lambda})$ for the other values. Denote n'_0 the number of individuals we keep in the nominal set. If the basis is the Haar wavelet basis, we have $n'_0 = n_0$: all the nominal individuals can be used since the basis is fixed, therefore the features from both subsets are independent. For the PCA basis, we take $n'_0 = n_0/2$ since we do not use the features arising from the data $\mathbf{X}_{2i}, i = 1, \dots, n_0/2$ that were already used to compute the principal components.

Denote $\tilde{\boldsymbol{\theta}}_{0\lambda} = \hat{\boldsymbol{\theta}}_{0\lambda}$ in the case of the Haar basis, and $\tilde{\boldsymbol{\theta}}_{0\lambda}$ is composed by the odd indexes of $\hat{\boldsymbol{\theta}}_{0\lambda}$ for the PCA.

We assume that the components of the vector $\tilde{\boldsymbol{\theta}}_{0\lambda}$ are independent and identically distributed (i.i.d.) with common distribution function $F_{\lambda}^{(0)}$ and that the components of the vector $\hat{\boldsymbol{\theta}}_{1\lambda}$ are i.i.d. with common distribution function $F_{\lambda}^{(1)}$. Both sets are independent. We will now propose testing procedures to test the null hypothesis $F_{\lambda}^{(0)} = F_{\lambda}^{(1)}$.

At first we introduce these tests for a single level λ . We will treat the problem of multiple testing in the next section.

3.1.1. Two sample tests

Let us suppose we have two independent vectors $\mathbf{X} = (X_1, \dots, X_{n_0})$ i.i.d. with common continuous cumulative distribution function F and probability distribution \mathbb{P} , and let $\mathbf{Y} = (Y_1, \dots, Y_{n_1})$ i.i.d. with common continuous cumulative distribution function G probability distribution \mathbb{Q} . The generalized inverse functions F^{-1} and G^{-1} are the also called the quantile functions. F and G are estimated by the empirical distribution functions F_{n_0} and G_{n_1} , where $\forall t \in \mathbb{R}$,

$$F_{n_0}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbb{1}_{X_i \leq t},$$

$$G_{n_1}(t) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{1}_{Y_i \leq t}.$$

Let $X_{(1)} \leq \dots \leq X_{(n_0)}$ be the ordered vector X . The quantile function F^{-1} is estimated by $F_{n_0}^{-1}$ defined as :

$$F_{n_0}^{-1}(p) = \begin{cases} X_{(1)} & \text{if } p < 1/n_0 \\ X_{(i)} & \text{if } p \in \left[\frac{i-1}{n_0}, \frac{i}{n_0} \right] \text{ and } 2 \leq i \leq n_0 \\ X_{(n_0)} & \text{if } p = 1 \end{cases}$$

and $G_{n_1}^{-1}$ is computed in the same way. We would like to test the following hypothesis :

$$\begin{cases} H_0 : F = G \\ H_1 : F \neq G \end{cases} \quad (5)$$

Many papers deal with the two-sample problem when no prior knowledge is assumed for the shape of the distributions. In this case non parametric tests are used to asses the veracity of the null assumption. The Kolmogorov-Smirnov test is a reference for this problem but other more recent tests can be also implemented built using distance that preserve the shape of the distributions. Here we use tests based on the Wasserstein metrics, reviewed for instance in [15].

3.1.2. Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test relies on the fact that under the null hypothesis H_0 , the distribution of the statistics

$$D_{n_0, n_1} = \sqrt{\frac{n_0 n_1}{n_0 + n_1}} \sup_{x \in \mathbb{R}} |F_{n_0}(x) - G_{n_1}(x)|$$

does not depend on F . The test is rejected when $D_{n_0, n_1} > c_{1-\alpha}$, where $c_{1-\alpha}$ is the defined as the $1 - \alpha$ quantile of D_{n_0, n_1} under H_0 , in order to obtain a level α for the test. See [22] for further explanation.

3.1.3. Tests based on Wasserstein distances

When testing the equality of distributions, the choice of the distance used to evaluate the statistical gap between the two samples is important. In the following we introduce a test based on Wasserstein distance. First, for $d \geq 1$, consider the set $\mathcal{W}_q(\mathbb{R}^d)$ of probabilities with finite r -th moment. For μ and ν in $\mathcal{W}_q(\mathbb{R}^d)$, we denote by $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbb{R}^d \times \mathbb{R}^d$ with first (resp. second) marginal μ (resp. ν). The L_q transportation cost between these two measures is defined as

$$W_q^q(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^q d\pi(x, y).$$

This transportation cost allows to endow the set $\mathcal{W}_q(\mathbb{R}^d)$ with the metric $W_r(\mu, \nu)$. More details on Wasserstein distances and their links with optimal transport problems can be found in [23] or [24] for instance.

The Wasserstein distance $W_q(P, Q)$ between two probability measures P and Q on \mathbb{R} with $q \geq 1$ finite moments can be easily written as

$$W_q^q(P, Q) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^q dt$$

where F^{-1} and G^{-1} are the quantile functions of P and Q respectively. In our framework, we observe two n samples of i.i.d random variables with distribution P and Q . Let P_n and Q_n be the empirical distributions, hence the Wasserstein distance between these two empirical distributions is given by

$$W_q^q(P_n, Q_n) = \frac{1}{n} \sum_{i=1}^n |X_{(i)} - Y_{(i)}|^q.$$

Testing the equality of the two distributions is equivalent to test that the Wasserstein distance $W_q^q(P, Q)$ is equal to zero. As a matter of fact, under the assumption that X_1, \dots, X_n are i.i.d. with distribution P , Y_1, \dots, Y_n are i.i.d. with distribution Q and P and Q have finite q -th moment it is easy to conclude that $W_q^q(P_n, Q_n) \rightarrow W_q^q(P, Q)$ almost surely. However, designing a test requires knowing the asymptotic distribution of a rescaled version of $W_q^q(P_n, Q_n)$ both under H_0 to estimate the level of the test and under H_1 to evaluate its power. Del Barrio et al. [25] and references therein give some insights while the case $P \neq Q$ is tackled in [26].

Yet the asymptotic distribution depends on the distribution P which is unknown. Hence we will use the test proposed in [15] which relies on the following property. If we consider the image by the distribution function G of the distribution P we obtain a distribution $P(G^{-1})$ with cumulative distribution function $G \circ F^{-1}$. Under the null assumption H_0 , then $F = G$ and this distribution is the uniform distribution on $[0, 1]$. Hence rather than testing the goodness of fit $P = Q$, we can use this non linear transformation to alternatively test the goodness of fit between the uniform distribution and $P(G^{-1})$. The main advantage of this setting is that the asymptotic distribution under the null assumption does not depend on the distribution P .

Assume that f and g are the density functions related to F and G . Let us suppose that there exists $C \in \mathbb{R}$ such that

$$\forall t \in \mathbb{R}, \frac{g(F^{-1}(t))}{f(F^{-1}(t))} \leq C.$$

Let $\gamma = \frac{n_0 n_1}{n_0 + n_1}$. According to Ramdas et al. [15], we know that, under the null hypothesis,

$$\gamma \times W_2^2(G_{n_1 \#} P_{n_0}, G_{\#} P) = \gamma \int_0^1 (G_{n_1}(F_{n_0}^{-1}(t)) - t)^2 dt \rightarrow \int_0^1 (\mathbb{B}(t))^2 dt$$

and

$$\sqrt{\gamma} \times W_\infty(G_{n_1 \#} P_{n_0}, G_{\#} P) = \sqrt{\gamma} \sup_{t \in [0, 1]} |G_{n_1}(F_{n_0}^{-1}(t)) - t| \rightarrow \sup_{t \in [0, 1]} |\mathbb{B}(t)|$$

Table 1. Estimated level of the 2-Wasserstein and the ∞ - Wasserstein test on Gaussian distributions.

n	2-Wasserstein test		∞ -Wasserstein test	
	$\hat{\alpha}$	$\sqrt{Var(\hat{\alpha})}$	$\hat{\alpha}$	$\sqrt{Var(\hat{\alpha})}$
50	0.052	3.1×10^{-3}	0.0394	2.8×10^{-3}
100	0.046	2.9×10^{-3}	0.0360	2.6×10^{-3}
500	0.048	3.0×10^{-3}	0.0470	3.0×10^{-3}
1000	0.045	2.9×10^{-3}	0.0514	3.1×10^{-3}
10000	0.048	3.0×10^{-3}	0.0550	3.2×10^{-3}

where $\mathbb{B}(t)$ is a Brownian bridge on $[0, 1]$.

Consequently, it is possible to build a statistical test, to check the equalities of two distributions thanks to the Wasserstein distance, by using the asymptotic distribution of the test statistics under the null hypothesis to calibrate the quantiles. The null hypothesis is rejected if

$$T_2 = \gamma \int_0^1 (G_{n_1}(F_{n_0}^{-1}(t)) - t)^2 dt > c_{2,1-\alpha}$$

for the 2-Wasserstein test, or if

$$T_\infty = \sqrt{\gamma} \sup_{t \in [0,1]} |G_{n_1}(F_{n_0}^{-1}(t)) - t| > c_{\infty,1-\alpha}$$

for the ∞ -Wasserstein test, where $c_{2,1-\alpha}$ is the $1 - \alpha$ quantile of the distribution of $\int_0^1 (\mathbb{B}(t))^2 dt$, and $c_{\infty,1-\alpha}$ is the $1 - \alpha$ quantile of the distribution of $\sup_{t \in [0,1]} |\mathbb{B}(t)|$.

Since we use the asymptotic quantiles of the test statistics under the null hypothesis, we have carried out some simulations to estimate the level of the test from a non asymptotic point of view.

Simulation study

In this example, to evaluate the non asymptotic level of the tests, we take $n_0 = n_1 = n/2$ and we simulate both samples with standard Gaussian distributions. We simulate m i.i.d. samples: for $k = 1 \dots m$, $X^k \sim \mathcal{N}_{n/2}(0, I_{n/2})$ and $Y^k \sim \mathcal{N}_{n/2}(0, I_{n/2})$, independent of X^k . At each iteration k , we test the equality of the distributions of X^k and Y^k , from which we get a p-value $p_{k,n}$. As usual, we estimate the level of our tests by the empirical estimator, namely the proportion of tests rejected at a level α among the m Wasserstein tests :

$$\hat{\alpha}(n) = \frac{1}{m} \sum_{k=1}^m \mathbb{1}_{p_{k,n} < \alpha}.$$

We repeat it for both Wasserstein tests. When n is large, we expect $\hat{\alpha}(n)$ to be close to α . We choose $\alpha = 0.05$, $m = 5000$ iterations and $n/2$ varies from 50 to 10000. The results are summarized in Table 1, showing that the level of the test is close to 5% even for small sample sizes. This shows that the asymptotic test reaches the desired level very quickly, mostly with the 2-Wasserstein test.

Table 2. Estimated level of the 2-Wasserstein and the ∞ -Wasserstein test on exponential distributions.

n	2-Wasserstein test		∞ -Wasserstein test	
	$\hat{\alpha}$	$\sqrt{\text{Var}(\hat{\alpha})}$	$\hat{\alpha}$	$\sqrt{\text{Var}(\hat{\alpha})}$
50	0.050	3.1×10^{-3}	0.0406	2.8×10^{-3}
100	0.0462	3.0×10^{-3}	0.0400	2.8×10^{-3}
500	0.0462	3.0×10^{-3}	0.0482	3.0×10^{-3}
1000	0.0470	3.0×10^{-3}	0.0536	3.2×10^{-3}
10000	0.0456	3.0×10^{-3}	0.0518	3.1×10^{-3}

We make the same work with X^k and $Y^k \sim \mathcal{E}_{n/2}(1)$, and we record the level corresponding to both Wasserstein tests. We present the results in Table 2.

The level of the test is equivalent with Exponential distributions, even for small values of n . We have carried out some simulations to compare the performances of the two tests based on T_2 and T_∞ with the Kolmogorov-Smirnov test, where the results are presented in the next section.

3.1.4. Simulation study of the power of the tests

We simulate independent samples arising from two different distributions. For $k = 1, \dots, m$, we simulate

$$X^k = (X_1^k, \dots, X_{n/2}^k) \sim \mathcal{N}_{n/2}(0, I_{n/2})$$

$$Y^k = (Y_1^k, \dots, Y_{n/2}^k) \sim \mathcal{N}_{n/2}(0, I_{n/2})$$

$$Z^k = (Z_1^k, \dots, Z_{n/2}^k) \sim \mathcal{N}_{n/2}(\mu, \sigma^2 \times I_{n/2}),$$

where $(\mu, \sigma^2) \neq (0, 1)$. We denote by F_X, F_Y and F_Z the cumulative distribution functions of X, Y and Z . We know that $F_X = F_Y$ whereas $F_X \neq F_Z$. We denote by $F_{n,X}^k, F_{n,Y}^k$ and $F_{n,Z}^k$ their empirical distribution function for the iteration k .

At each iteration k , we generate X^k, Y^k and Z^k and, based on these observations, we test the equality of the distribution F_X and F_Y , and then the equality of F_X and F_Z . The first test should be accepted whereas the second should be rejected.

Let $p_k^{(0)}, k = 1, \dots, m$ be the p-values corresponding to the test $\{F_X = F_Y\}$, and $p_k^{(1)}, k = 1, \dots, m$ be the p-values corresponding to the test $\{F_X = F_Z\}$ at the iteration k . From these p-values we can compute the true positive rate (TPR) and false positive rate (FPR) for each level α of the test, where $\text{TPR}(\alpha) = \frac{1}{m} \sum_{k=1}^m \mathbb{1}_{p_k^{(0)} > \alpha}$ and $\text{FPR}(\alpha) = \frac{1}{m} \sum_{k=1}^m \mathbb{1}_{p_k^{(1)} > \alpha}$. The values reported draw a ROC curve from those simulations.

We test different values for n, μ, σ^2 in order to get the behaviours of each test. We simulate $m = 2000$ samples, thus we have 4000 results of tests in total.

As we can see on the figures 1 and 2, in all the simulations, the 2-Wasserstein test has the best power.

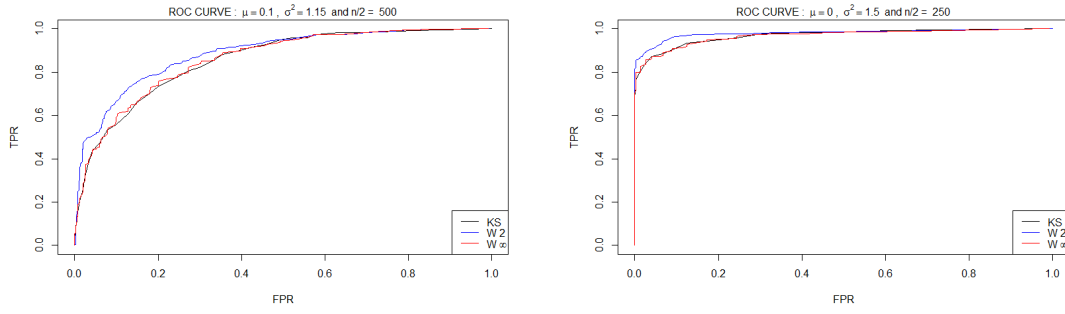


Figure 1. On the left : ROC curves with $\mu = 0.1, \sigma^2 = 1.15, n = 1000$, on the right : ROC curves, with $\mu = 0, \sigma^2 = 1.5, n = 500$

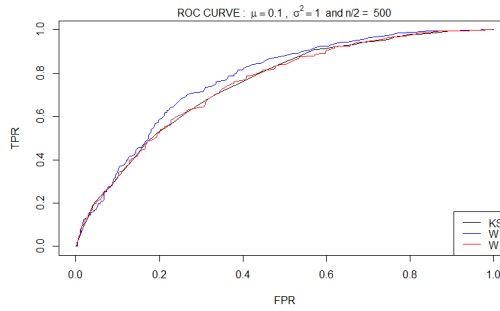


Figure 2. ROC curves, with $\mu = 0.1, \sigma^2 = 1, n = 1000$

3.2. Features selection while controlling the false discovery rate

3.2.1. A multiple testing framework

We remind that in both cases described in Sections 2.1 and 2.2, we have defined features $(\hat{\theta}_{\cdot\lambda})_{\lambda \in \Lambda}$ to represent our data. Each feature $\hat{\theta}_{\cdot\lambda}$ is divided in two parts : $\hat{\theta}_{0\lambda} = (\hat{\theta}_{1,\lambda}, \dots, \hat{\theta}_{n_0,\lambda})$ corresponding to the set that is known to be nominal which are assumed to be i.i.d. with distribution function $F_{\lambda}^{(0)}$, estimated from the features $\tilde{\theta}_{0\lambda}$, and $\hat{\theta}_{1\lambda} = (\hat{\theta}_{n_0+1,\lambda}, \dots, \hat{\theta}_{n,\lambda})$ for the other values, assumed to be i.i.d. with distribution function $F_{\lambda}^{(1)}$. In order to select interesting features, we use the 2-Wasserstein test to test, for each feature, the null hypothesis

$$H_{0,\lambda} : \{F_{\lambda}^{(0)} = F_{\lambda}^{(1)}\}.$$

We are therefore dealing with a multiple testing problem since we test in both cases (wavelet or PCA decomposition), $|\Lambda| = p$ null hypotheses. Under $H_{0,\lambda}$, the p-value p_{λ} of the test is expected to be (asymptotically) uniformly distributed on $[0, 1]$ whereas, under the alternative, it is expected to be close to zero. The null hypothesis is rejected at level α for a p-value smaller than α . Nevertheless, it is well known that, when we deal with many hypotheses, the probability to have at least one p-value smaller than the level α can be very large. Here, for both types of features, we test p hypotheses

Table 3. Multiple testing procedure

	Declared non significant	Declared significant	Total
True Null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

and for any $t > 0$,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{\lambda \in \Lambda} \{p_\lambda < t\}\right) &= 1 - \mathbb{P}\left(\bigcap_{\lambda \in \Lambda} \{p_\lambda > t\}\right) \\ &\sim 1 - (1 - t)^p \\ &\xrightarrow[p \rightarrow \infty]{} 1 \end{aligned}$$

We have used the independence of the random variables $(p_\lambda, \lambda \in \Lambda)$. Indeed, in both cases described in Sections 2.1 and 2.2, the basis $(\phi_\lambda)_{\lambda \in \Lambda}$ is orthonormal for the scalar product $\langle \cdot, \cdot \rangle_p$ in \mathbb{R}^p . Since the vectors \mathbf{X}_i are Gaussian vectors and since $\hat{\theta}_{i\lambda} = \langle \mathbf{X}_i, \phi_\lambda \rangle_p$ for any $\lambda \in \Lambda$, we get that in both cases, the p vectors $(\hat{\theta}_{\cdot\lambda})$ for $\lambda \in \Lambda$ are stochastically independent.

For example, we know that for a desired level $\alpha = 5\%$, the probability to reject at least one test among 50 if the null hypotheses are all true is already larger than 90%. Thus, we will use the procedure proposed by Benjamini and Hochberg [14] to control the false discovery rate. Let us first give some definitions. Consider that we have m hypotheses to test where m_0 hypotheses are true. R is the total number of rejections. The table 3 summarizes the situation.

In this table, only R and m are known. The false discovery rate (FDR) is defined by

$$FDR = \mathbb{E}\left(\frac{V}{\max(R, 1)}\right).$$

The Benjamini and Hochberg [14] procedure allows to control the FDR.

3.2.2. Control the FDR with the Benjamini-Hochberg procedure

Benjamini and Hochberg [14] proposed a simple way to control the false discovery rate. Assume that all the p-values $(p_k)_{1 \leq k \leq m}$ are independent random variables, and let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the ordered p-values. Let k^* be the largest k for which $p_{(k)} \leq k\alpha/m$, then reject all the hypotheses $H_{0,(k)}$, $k = 1, \dots, k^*$. Benjamini and Hochberg proved that the FDR for this procedure does not exceed the level α .

Other procedures exist, and many of them are detailed in [27] but we limit ourselves to this procedure since it is very easy to implement. Note that it is justified in our cases (wavelet or PCA decompositions) since, as explained above, our p-values are independent. This procedure has been chosen since it is really easy to implement, but other procedures can be chosen, where many of them can be found in [27].

4. Outlier detection with the Local Outlier Factor

Once we have selected the features that isolate the best the anomalies in the sense of the comparison of the distributions, we are ready to apply an outlier detection technique on those features. There exists many unsupervised outlier detection methods. Chandola et al. [1] detailed many of them. They are often categorized into two categories : distance-based and density-based methods. For instance, the One-Class SVM developed by Schölkopf et al. [11] is a density-based method that is widely used for outlier detection purposes. However, most of time, the telemetry data we are dealing with evolve slowly because of seasonality effects. As a result, there is no immediate separation between the anomalies and the nominal data. Hence, the optimal conditions to use the One-Class SVM are not satisfied.

The local outlier factor (LOF) is a score introduced by Breuning et al. [9] to detect outlier data. In addition of detecting outliers, it returns a score of anomalousness. This method is mixing the density-based and distance-based points of view, and has already been tested on space telemetry data since this method inspired the ESA in the Novelty software [28]. It is a local method since this factor depends on how the object is isolated with respect to the surrounding neighbourhood.

Suppose we have n objects x_1, \dots, x_n to cluster. To simplify the notations, we suppose that for x_1, x_2, x_3 all different, $d(x_1, x_2) \neq d(x_1, x_3)$.

Let $x \in \{x_1, \dots, x_n\}$ and let $k < n$ be the number of neighbours that we consider. Choosing the best value for k is not so easy. In [9], different values are tested from 10 to 50, and the performance depends on how the data is distributed (different clusters, statistical fluctuations...). As our data depends on seasonality effects, we opt for considering small values for k .

- Let $d^k(x)$ be the k -distance of x , which means that for k objects among x_1, \dots, x_n , the distance to x is closer than $d^k(x)$, and the other $n - k$ points are situated further.

Let $N_k(x)$ be the set of k nearest neighbours of x .

- The reachability distance of x with respect to an object o is then defined as $r_k(x, o) = \max\{d^k(o), d(x, o)\}$. If x and o are sufficiently close, the distance between them is replaced by the k -distance of o .

- Then the local reachability density of x is defined as

$$lr_k(x) = \frac{k}{\sum_{o \in N_k(x)} r_k(x, o)}.$$

- The local outlier factor is then defined as

$$LOF_k(x) = \frac{1}{k} \sum_{o \in N_k(x)} \frac{lr_k(o)}{lr_k(x)}.$$

In other words, the *LOF* compares the nearest neighbours density distance of a given object with the nearest neighbours density distance of its nearest neighbours. When this score is close to 1, it means that the object is distributed in the same way as its neighbours. When having a large *LOF*, the corresponding object is likely to be an outlier.

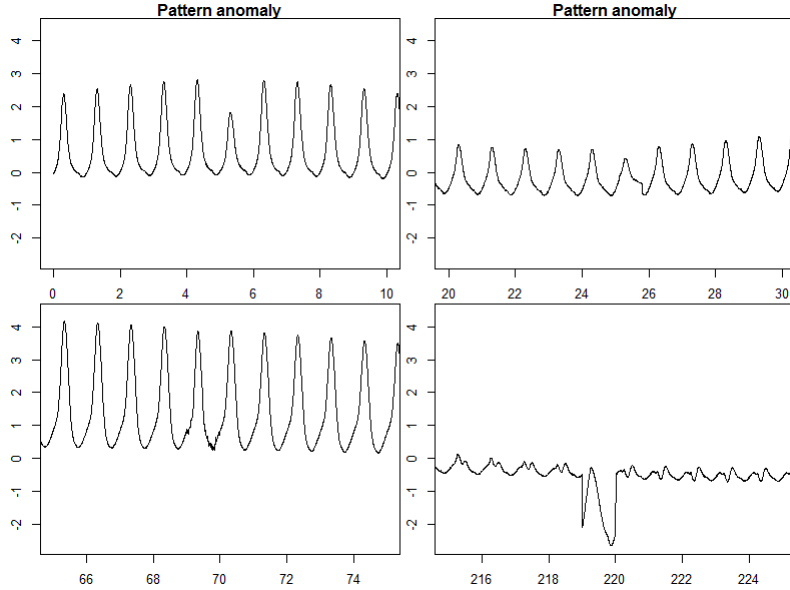


Figure 3. Portions of the signal containing the 4 pattern anomalies

5. Application to simulated data

We apply our outlier detection method on simulated telemetry that is really close to what we can have on real space telemetries. This simulation example has been created in order to ease the validation of each method. This example was inspired by geostationary satellites telemetries that have daily periodicity as well as yearly periodicity.

We used a combination of periodic signals and given patterns to generate our telemetry. We simulated $n = 480$ days of telemetry to get a significant number of signals after splitting the signal into days. We have $p = 256$ measurements per day. Each day of telemetry corresponds to an observation. The total signal symbolize two year of telemetry, where a year lasts 240 days in this example. We added a Gaussian noise to our observations.

We introduced eight anomalies of several types, that represent a complete panel to what can be observed on real telemetries. These anomalies are situated only in the first 240 days.

The anomalies that are introduced are the following: 4 pattern anomalies (change in the pattern or in the amplitude of the data), 3 local anomalies (noise, spikes, data set to default value...) and one periodicity anomaly (two patterns instead of one). The pattern anomalies occur on days 6, 26, 70 and 220, the local anomalies on days 134, 156 and 201, and the periodicity anomaly on day 98.

The figures 3 and 4 show the portions of the signal containing anomalies, where the pattern anomalies are all in Figure 3. Some of the anomalies seem obvious (days 219, 97), and some other are less pronounced (days 6, 26, 134).

The aim of the study is to retrieve the days with abnormal behaviours.

For each day of telemetry, we calculate the features based on the Haar wavelet basis and the PCA basis, and keep the raw-data as a reference result. We then compute the Local Outlier Factor on the features that are not labeled as nominal. We will test several feature sets:

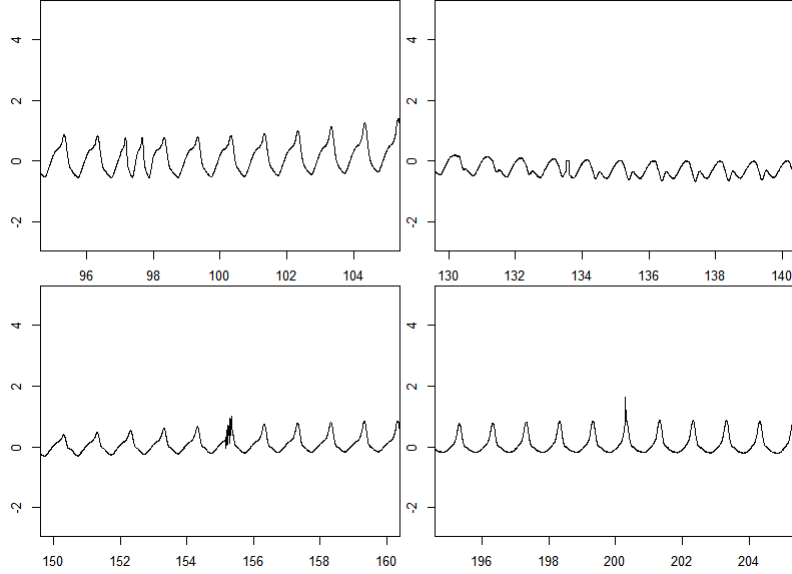


Figure 4. Portions of the signal containing the periodicity anomaly (top left) and the 3 local anomalies

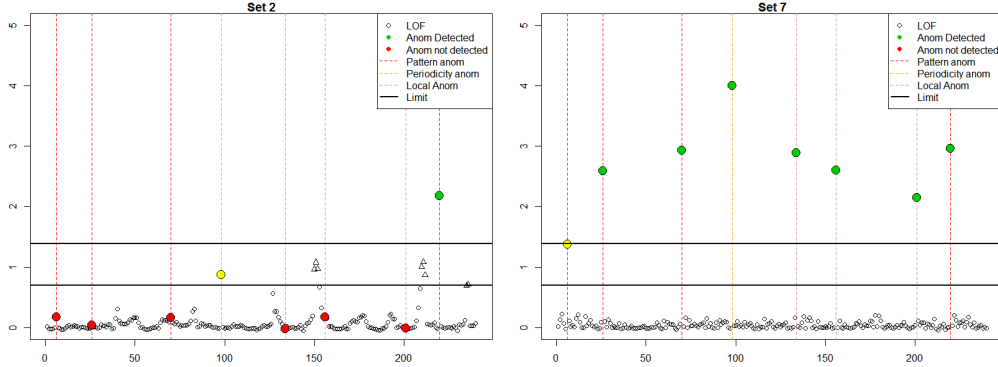
- (0) The raw data.
- (1) The full PCA coefficients.
- (2) The first d coefficients of the PCA representing at least 95% of the variance.
- (3) The full Haar wavelet coefficients.
- (4) The 8 coefficients from the levels 0, 1, 2 of a Haar wavelet basis.
- (5) The 8 coefficients from the third level of a Haar wavelet basis.
- (6) The 16 coefficients from the fourth level of a Haar wavelet basis.
- (7) The PCA components resulting from the feature selection based on the 2-Wasserstein test.
- (8) The PCA components resulting from the feature selection based on the ∞ -Wasserstein test.
- (9) The wavelet coefficients resulting from the feature selection based on the 2-Wasserstein test.
- (10) The wavelet coefficients resulting from the feature selection based on the ∞ -Wasserstein test.

We do not present the feature selection on Kolmogorov-Smirnov test, because this test is really close to the ∞ -Wasserstein test, hence returns the same feature selection. After selecting our feature sets, we compute the local outlier factor. We have two parameters to calibrate : the number of neighbours to compute the local outlier factor. and the threshold for the value of the local outlier factor to detect outliers. For the first parameter, we choose $k = 10$ neighbours. As mentionned earlier, $k = 10$ seems to be a good choice since the data has a yearly trend, and it is better to consider a number of neighbours that is not too large.

For the threshold, we report the days where the LOF is greater than 2, and the ones that are larger than 4. The results are summarized in the table 4. There are 8 anomalies to detect, and the local anomalies are expected to be harder to spot than the other anomalies.

Table 4. Anomaly found for each feature set

Feature	Anom LOF>2	False alarm	Anom LOF>4	False alarm	Nb features
0 - Raw data	2/8	7	1/8	0	256
1 - PCA - full	2/8	7	1/8	0	256
2 - PCA - 95%	2/8	8	1/8	0	3
3 - Haar - full	3/8	0	1/8	0	256
4 - Haar (lev 0,1,2)	3/8	0	1/8	0	8
5 - Haar (lev 3)	3/8	0	2/8	0	8
6 - Haar (lev 4)	5/8	0	3/8	0	16
7 - PCA - W2	8/8	0	7/8	0	17
8 - PCA - W ∞	8/8	0	7/8	0	14
9 - Haar - W2	6/8	2	5/8	0	4
10 - Haar - W ∞	6/8	2	5/8	0	4

**Figure 5.** LOF for both PCA coefficients selection - common (set2) and novel (set7), and limit (LOF=2,4)

A first constat is that the Local Outlier Factor computed on the raw-data gives bad results, the same as the ones provides on the full PCA coefficients. There is too many redondant information in the raw-data, and the values are maybe too close to each other in general, hence do not allow to detect the outliers. This reinforces the fact that projections are really helpful for highlighting outliers.

5.1. Comments on the PCA results

At first, one can notice that both two-sample tests generates almost the same feature selection, with, consequently, the same results. The set 2, containing the first PCA coefficients, does not contains major information on the outliers. In fact, the 4 first coefficients are not selected when we use the novel feature selection (sets 7 and 8). The selected features that are common from both tests are the ones for which $\lambda \in \{10, 17, 21, 29, 31, 27, 41, 57, 58, 65, 68, 69, 88, 94\}$. It indicates that resuming the full data is not the best way to detect outliers. In fact, the information contained on outliers is unlikely to appear in the first components since the anomaly is rare, thus not representative of a large portion of the variance of the data. The figure 5 shows clearly how performant our procedure is comparatively to the common way to proceed. It enables to isolate better the anomalies to the nominal data. Without controlling the FDR, we would have retained 20 features for the 2-Wasserstein test, and 19 for the ∞ -Wasserstein test, instead of 17 and 14. As we have chosen $p = 256$, it enables to reduce even more the dimension of the data, with better results. In fact, without controlling the FDR, we would have missed one anomaly.

An important advantage of our procedure based on the feature selection is that it al-

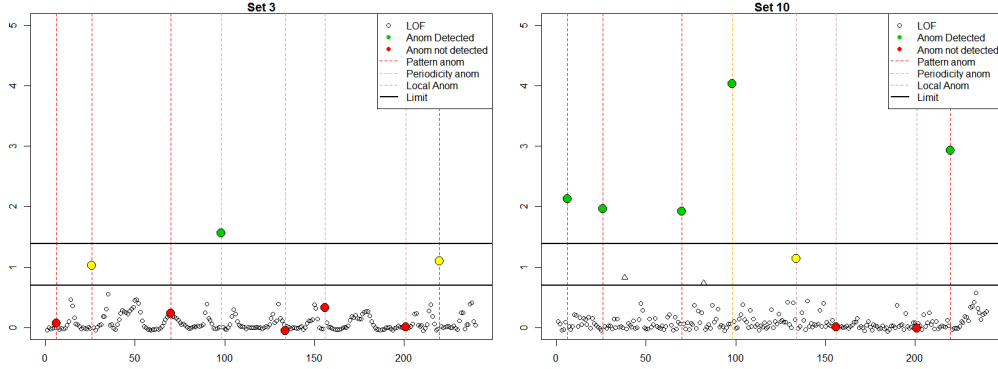


Figure 6. LOF for two Wavelet sets coefficients selection -levels where $j \leq 2$ (set3), novel procedure (set10), and limit ($LOF > 2, LOF > 4$)

lows to well separate the values of the LOF of the outliers from the values of the LOF for nominal data and therefore it is not very sensitive to the threshold : with the value 2 and 4 of the threshold, we detect almost the same anomalies for the sets 7, 8, 9, 10. This is an important property because for the other cases, the set of detected outliers is very sensitive to the threshold.

5.2. Comments on wavelet results

We have similar results with the wavelet decomposition. The levels $l \leq 2$ do not capture any information on the local events. The larger levels, like the level $l = 4$ exhibits easier local events. However, the results are even better thanks to the automatic selection of wavelet coefficients, as we can see on Figure 6. Once again, our procedure to select the features enables to isolate better the anomalies to the nominal data. Such results can help to calibrate the level of rejection of the Local outlier factor, since a larger margin will lead to the optimum of the method. If we look at the 4 wavelet coefficients that were selected after the novel feature selection procedure, with the BH procedure at level 5%, we have:

- 1 position over 4 for $l = 2$,
- 1 positions over 8 for $l = 3$,
- 1 positions over 16 for $l = 4$,
- 1 positions over 32 for $l = 5$,

The medium levels are well represented here. Without the BH procedure to control the FDR, we would have retained 12 features, instead of 4. The results are better without controlling the FDR: we would have detected one additional anomaly at the level $LOF = 4$, while removing the false alarms for $LOF > 2$.

6. Conclusion

This original features selection procedure is really relevant for the detection of outliers. It reinforces the idea that clustering and outlier detection cannot be addressed by similar methodologies. In the case of the PCA, only the first principal components are usually retained for clustering. What we have shown in this paper is that, even

if it is the best method to reduce the dimension of the data, it is not the best one to exhibit abnormal behaviours. In fact, those unexpected events do not represent a large portion of the variance of the data, since the outliers are rare and do not have a repetitive signature. Consequently, the anomalies are unlikely to appear clearly as outliers in the very first components, but some further components can get such information. This method is really adapted for the outlier detection in periodical time series, as the space telemetry data.

We have also mentioned the case of the Haar wavelet basis. In fact, it is not easy to know which are the levels to retain for such analysis. One can be interested by concentrating the global information, and some other to get details as well. The selection we develop enables to guarantee both types of information on the data.

The PCA decomposition with our feature selection procedure gives the best results in this study. The wavelet decomposition combined with the original feature selection procedure gives also good results. An important advantage of the wavelet decomposition compared with the PCA decomposition is that it is a fixed basis, whereas the PCA is based on a data dependent basis and to keep independence properties, we had to isolate some data to compute the principal components. Moreover, if we want to implement online procedures, which is often the case when we deal with bigdata sets, a fixed basis is much more relevant.

Our approach requires to have some knowledge on the data because a set of data that does not contain any anomaly has to be isolated. The difficulty to implement a total unsupervised approach comes from the fact that the distribution of the features for nominal data is unknown.

In this direction, a test has been developed by Candelon et al. [29], their method is based on bootstrap permutations, which increases a lot the computation time, and it is not suitable for multiple testing, as in our situation.

References

- [1] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*. 2009;41(3):15.
- [2] Vatanen T, Kuusela M, Malmi E, et al. Semi-supervised detection of collective anomalies with an application in high energy particle physics. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on; IEEE; 2012*. p. 1–8.
- [3] Sillito RR, Fisher RB. Semi-supervised learning for anomalous trajectory detection. In: *BMVC; Vol. 27; 2008*. p. 1025–1044.
- [4] Hoffmann H. Kernel PCA for novelty detection. *Pattern Recognition*. 2007;.
- [5] Ordoñez C, Martínez J, Rodríguez-Pérez J, et al. Detection of outliers in gps measurements by using functional-data analysis. *Journal of Surveying Engineering*. 2011; 137(4):150–155.
- [6] Ren H, Chen N, Zou C. Projection-based outlier detection in functional data. *Biometrika*. 2017;104(2):411–423.
- [7] Auder B, Fischer A. Projection-based curve clustering. *J Stat Comput Simul*. 2012; 82(8):1145–1168.
- [8] Antoniadis A, Brossat X, Cugliari J, et al. Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*. 2013; 11(01):1350003.
- [9] Breunig MM, Kriegel HP, Ng RT, et al. LOF: identifying density-based local outliers. In: *ACM sigmod record; Vol. 29-2; ACM; 2000*. p. 93–104.
- [10] Barnett V, Lewis T, et al. *Outliers in statistical data*. Vol. 3. Wiley New York; 1994.

- [11] Schölkopf B, Platt JC, Shawe-Taylor JC, et al. Estimating the support of a high-dimensional distribution. *Neural Comput.* 2001 Jul;13(7):1443–1471.
- [12] Dsir C, Bernard S, Petitjean C, et al. One class random forests. *Pattern Recognition.* 2013;.
- [13] Febrero M, Galeano P, González-Manteiga W. Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics.* 2008; 19(4):331–345.
- [14] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B.* 1995;57(1):289–300.
- [15] Ramdas A, García Trillos N, Cuturi M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy.* 2017;19(2):Paper No. 47, 15.
- [16] Shawe-Taylor J, BZlicar. Novelty detection with one-class support vector machines. *Advances in Statistical Models for Data Analysis.* 2015;.
- [17] Ramsay JO, Silverman BW. *Functional data analysis.* 2nd ed. Springer, New York; 2005. Springer Series in Statistics.
- [18] Daubechies I. *Ten lectures on wavelets.* (CBMS-NSF Regional Conference Series in Applied Mathematics; Vol. 61). Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; 1992.
- [19] Donoho DL, Johnstone IM, Kerkyacharian G, et al. Density estimation by wavelet thresholding. *Ann Statist.* 1996;24(2):508–539.
- [20] Ding C, He X. K-means clustering via principal component analysis. In: *Proceedings of the twenty-first international conference on Machine learning;* ACM; 2004. p. 29.
- [21] Shyu ML, Chen SC, Sarinnapakorn K, et al. A novel anomaly detection scheme based on principal component classifier. *DTIC Document;* 2003.
- [22] Wilcox R. Kolmogorov–smirnov test. *Encyclopedia of biostatistics.* 2005;.
- [23] Rachev ST. The monge-kantorovich problem on mass transfer and its applications in stochastics. *Teor Veroyatnost i Primenen.* 1984;29(4):625–653.
- [24] Villani C. *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media; 2008.
- [25] del Barrio E, Cuesta-Albertos JA, Matrán C, et al. Tests of goodness of fit based on the L_2 -Wasserstein distance. *Ann Statist.* 1999;27(4):1230–1239.
- [26] Del Barrio E, Loubes JM. Central limit theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:170501299.* 2017;.
- [27] Roquain E. Type I error rate for testing many hypotheses: a survey with proofs. *Journal de la Société Française de Statistique.* 2011;152(2):3–38.
- [28] Martínez-Heras JA, Donati A, Kirsch MG, et al. New telemetry monitoring paradigm with novelty detection. In: *SpaceOps 2012 Conference, Stockholm, Sweden;* 2012. p. 11–15.
- [29] Candelon B, Metiu N. A distribution-free test for outliers. *Deutsche Bundesbank, Research Centre.* 2013;(02/2013).