sciendo

JAISCR, 2021, Vol. 11, No. 3, pp. 195 – 215
10.2478/jaiscr-2021-0012

LEARNING NOVELTY DETECTION OUTSIDE A CLASS OF RANDOM CURVES WITH APPLICATION TO COVID-19 GROWTH

Wojciech Rafajłowicz

Department of Control Systems and Mechatronics
Wrocław University of Science and Technology, Wroclaw, Poland

E-mail: wojciech.rafajlowicz@pwr.edu.pl

Submitted: 6th September 2020; Accepted: 21th March 2021

Abstract

Let a class of proper curves is specified by positive examples only. We aim to propose a learning novelty detection algorithm that decides whether a new curve is outside this class or not. In opposite to the majority of the literature, two sources of a curve variability are present, namely, the one inherent to curves from the proper class and observations errors'. Therefore, firstly a decision function is trained on historical data, and then, descriptors of each curve to be classified are learned from noisy observations.When the intrinsic variability is Gaussian, a decision threshold can be established from T^2 Hotelling distribution and tuned to more general cases. Expansion coefficients in a selected orthogonal series are taken as descriptors and an algorithm for their learning is proposed that follows nonparametric curve fitting approaches. Its fast version is derived for descriptors that are based on the cosine series. Additionally, the asymptotic normality of learned descriptors and the bound for the probability of their large deviations are proved. The influence of this bound on the decision threshold is also discussed.The proposed approach covers curves described as functional data projected onto a finite-dimensional subspace of a Hilbert space as well a shape sensitive description of curves, known as square-root velocity (SRV). It was tested both on synthetic data and on real-life observations of the COVID-19 growth curves.

Keywords: classification, learning, novelty detection, functional data

1 Introduction

Our aim is to propose learning algorithms for novelty detection. In opposite to the majority of papers on this topic, which are mainly focused on vector data, we concentrate on detecting curves that are outsiders in a class of random functions or curves. The specific feature of novelty detection, also known as one-class pattern recognition, is that a learning sequence contains only positive examples. The absence of negative examples makes this problem essentially different than typical classification problems (see the citations at the end of this section). It is closer in spirit to significance tests of hypothesis and this line of reasoning will be dominating in our considerations. We refer the reader to [18] for the survey on the nature and types of novelties and anomalies.

In order to be able to collect positive examples of random curves, we have to assume that there exists a certain random phenomenon that is repeatable in the sense that we can observe many different realizations of random functions over a certain interval

$[0,\tau]$, $\tau > 0$, which are generated by the underlying stationary random phenomenon. On the other hand, we admit that curves can be generated by other processes and we have to distinguish them from the former ones. In particular, batch processes in chemical, electronic, and ceramic industries are repeatable in this sense.

Examples of repeatable processes. Additional examples of such processes, which also motivate the proposed approach, include:

- **meteorological data** such as temperatures, humidity, rainfall at a given site in a selected month that are compared year to year in order to detect untypical months,
- **air and water pollution** observations, collected and compared in a similar way as above,
- **quality characteristics of products** for example, frequency characteristics of high quality headphones and loudspeakers,
- **detecting untypical signatures** of plains, ships, cars,
- **abnormalities** in: spreading epidemic diseases in various countries or in the same country at different seasons, computer network traffic, etc.

Curves as mathematical objects. From the mathematical point of view, curves arising when observing the above mentioned processes can be considered as:

1. random functions, denoted as $\mathbb{X}(t)$, $t \in [0,\tau]$ or random elements (see the next section),
2. randomly generated closed curves, described in the parametric form,
3. functions attempting to emphasize the curve's shape, such as square-root velocity (SRV) description.

All these cases are covered by our approach, since – in the most cases – closed curves can be transformed to ordinary functions by expressing them in polar coordinates, while the SRV curve description can be handled as follows (see, e.g., [63, 40, 39, 28, 71] for basic facts concerning SRV approach and its recent extensions and applications).

Shape-sensitive description of curves. Let $f(t)$, $t \in [0,\tau]$ be a differentiable function, describing a curve in the classic way. Its SRV description, denoted further as $q_f(t)$, is defined as

$$q_f(t) = sgn(f'(t))\sqrt{|f'(t)|}, \tag{1}$$

where f' stands for the derivative of f, $sgn(a)$ is the signum of a and q_f is well defined at every point of $t \in [0,\tau]$, where $f'(t) \neq 0$.

The following properties of q_f are easy to verify:

1. it can be equivalently expressed as
$$q_f(t) = \frac{f'(t)}{\sqrt{|f'(t)|}}, \tag{2}$$
2. q_f is scale invariant, i.e., $q_{(cf)}(t) = q_f(t)$ for every $c > 0$,
3. q_f is translation invariant in the vertical direction, i.e., $q_{(c+f)}(t) = q_f(t)$ for every real constant c,
4. if $|f'(t)|$ is integrable on $[0,\tau]$, then $q_f(t)$ is square integrable there and
$$\int_0^\tau q_f^2(t)\,dt = \int_0^\tau |f'(t)|\,dt \tag{3}$$

According to 4, one can define the distance $\Omega(q_f, q_g)$ between two curves q_f and q_g as follows

$$\Omega(q_f, q_g) = \int_0^\tau \left|f'(t) - g'(t)\right| dt. \tag{4}$$

In (4) it was tacitly assumed that the time scales of curves f and g are the same. When one wants to compare curves with different time scales, then the so called time warping can be applied, i.e., the time t is re-scaled by non-decreasing function λ, say. Thus, one can obtain a better match between q_f and q_g by using the following distance

$$\Omega^*(q_f, q_g) \stackrel{def}{=} \inf_\gamma \Omega(q_{f\circ\gamma}, q_g), \tag{5}$$

where $f\circ\gamma$ is the composition of functions f and γ (see [63, 64, 40, 39, 28, 71]) and the bibliography cited therein, where also an algorithm for computing γ in (5) by dynamic programming is mentioned).

Summarizing, formally, one can select $\mathbb{X}(t) = q_f(t)$ in order to apply all the results of this paper to shape-sensitive description of curves. Notice, however, that we do not have direct observations of

$q_f(t)$ and one has to use an approximation of the derivative of $f(t)$ – see [53] for the approach to nonparametric estimation of derivatives from noisy observations of $f(t)$.

However, caution is needed in applying this approach, since in some cases both differences in amplitudes and retaining original time scales may be of importance, e.g., when one compares curves of the number of infected by COVID-19.

Curves that are not covered in this paper include space-filling curves, like those constructed by Hilbert, Peano and Sierpiński, since they are not sufficiently smooth. However, space-filling curves are proved to be useful for classification problems (see [58, 59, 61]).

Outline of the approach. The crucial issue in stating novelty detection problems is: how to define a (dis-)similarty measure between processes $\mathbb{X}$, $\mathbb{Y}$, ... that are considered as elements of a certain separable Hilbert space $\mathcal{H}$ of functions defined on $[0,\tau]$. In general, this task is difficult and we confine our attention to curves that can be sufficiently accurately represented by their orthogonal projections on K-th dimensional, $1 \leq K < \infty$, subspace $\mathcal{H}_K$ of $\mathcal{H}$. If a learning sequence is sufficiently long, then one can approximate any problem of practical importance just by selecting K sufficiently large. Any $\mathbb{X} \in \mathcal{H}_K$ can be represented by its coefficients $\bar{\theta} = [\theta_1, \theta_2, \ldots, \theta_K]^T$ in a selected, countable basis of $\mathcal{H}$. We assume that these coefficients are random, but not necessarily Gaussian, with $\bar{\theta}^0$ as the mean vector and with $K \times K$ covariance matrix Σ. This variability is later interpreted as the liability intrinsic to the class of curves that we want to consider as similar and typical for processes at hand. This set of curves is later called **class "0"** of typical processes. Confining our attention to large but finite K allows to consider an arbitrary probability distribution functions of $\bar{\theta}$.

The second source of a random variability in classifying a process to the class "0" or as the novelty is caused by errors in observations of $\mathbb{X}$. Moreover in this case we do not assume that they have any particular distribution, except that they have zero mean and finite variances. Thus, the estimation of $\mathbb{X}$ to be classified is a nonparametric one, if we allow a data driven selection of K. We comment on this aspect later in the paper.

On the other hand, derivations of the proposed algorithms go bottom up, in the sense that we temporarily assume normality of $\bar{\theta}$, then an outline of the algorithm is inferred and generalized to non-Gaussian cases and finally, its learning version and statistical properties are presented.

In fact, two kinds of learning algorithms are considered. The first of them uses historical data concerning whole curves. Their observations are collected along $[0,\tau]$. In every case when the process at hand starts at $t = 0$ and ends at $t = \tau$ is further called **one pass**. For this reason, the procedures that operate on subsequent historical curves are called pass-to-pass learning algorithms.

When a new curve to be recognized is acquired, a learning algorithm of the second kind is activated. Its role is to learn current curve $\mathbb{X}$ from its noisy observations. Therefore, it is further called the on-line learning algorithm or learning along one pass. Finally, these two kinds of algorithms are aggregated (see Figure 1 for details).

The paper organization and summary of the results. According to this methodology, the paper and the results are organized as follows.

- The problem statement is provided in the next section together with a summary of facts concerning random functions. Additionally, a skeletal Algorithm 0 is proposed. It serves as a starting point for constructing a learning Algorithm 1 in Section 3. Algorithm 1 is dedicated for learning (from retrospective data) the covariance matrix of intrinsic variability in $\bar{\theta}$ and the corresponding decision function. It is shown that for the Gaussian variability a threshold for decision making has the T^2 Hoteling distribution. The way of tuning this threshold for unknown distribution is also discussed.

- Section 4 contains the proposition of an algorithm for learning one curve $\mathbb{X}$ from noisy observations. It is based on estimating parameters (descriptors) of the orthogonal expansion of $\mathbb{X}$ in a selected basis. Basic statistical properties of the learning method are derived using the guidelines of nonparametric regression estimators by orthogonal series. Additionally, the asymptotic distribution of errors and bounds for their large deviations are proved. Finally, a fast learning Algorithm 2 is proposed that is based on the fast

cosine transform. The results of its testing on synthetic observations are provided in Section 6.

– Algorithms 1 and 2 are aggregated in Section 5 into Algorithm 3 and then tested on synthetic data (Section 6) and on the COVID-19 curves, representing the number of infected people over time (Section 7).

Related works. In recent fifteen years or so, one can observe a growing interest of statisticians and engineers to problems of classifying curves, considered as whole entities and learning tasks arising in this context. We refer the reader to monographs on these topics: [31, 17, 62, 3]. These monographs contain also a background on results from the functional analysis and random elements that are useful in this paper. Advanced applications can be found in [51] and in the most recent monograph [54] in which even more difficult data structures, namely data streams, are considered and searching for novelty is more broadly understood – as data mining.

A number of survey papers on functional data analysis (FDA) were also published [9, 70, 42, 36].

One of the mostly considered problems in the FDA is the one of classification functional patterns to one of many classes, usually stated in the empirical, nonparametric, bayesian setting, i.e., class labels are attached to each example by an expert and it is tacitly assumed that the learning sequence contains sufficiently many examples from each class. Additionally, it is usually assumed that there exist a priori probabilities that $\mathbb{X}$ was drawn at random from each particular class (see [11] for fundamental results in the finite-dimensional setting). As the methods of solution, variants of the functional counterparts of the Parzen kernel method are developed (see [4, 10, 1, 14, 15, 20, 29] for recent contributions in this stream of research).

A fundamental for FDA problem of discernibility of probability density functions (p.d.f.), basing on infinitely growing sequence of empirical data, is considered in [12]. The discernibility is understood as the existence of a sequence of classifying rules that are able to decide, with a finite number of errors and with the probability one, whether each given p.d.f. belongs a pre-specified class of densities or not.In [12] it was shown, among other results, that classes of densities, which are unimodal, log-concave or bounded by a constant are discernible. Although we shall not further pursue this topic, it is worth to mention that these distinguishing features of densities roughly describe their shape properties. It is also related to curves that have properties of p.d.f.'s, but for curves we have observations of a different kind.

The problems of classification to one class when features are finite dimensional is considered as important for many years. We refer the reader to [38, 37] for surveys of earlier papers and to [35, 60, 8] for more recent contributions. It is also worth to mention that change detection as well as drift detection and engineering diagnostic problems are related to novelty detection. Recent papers [43, 30, 21, 13] exemplify these similarities. The reader is referred also to [66, 56] for interesting applications in a jet engine diagnostics and in image segmentation of video sequences, respectively.

Up to now, the problems of the one-class classification of curves attracted less attention. We mention recent paper [69] in which an attempt to select archetypoids for anomaly detection in big functional data was undertaken and [34] where a kernel type approach was proposed. The following papers [73, 74] are closer to ours in the sense that orthogonal expansions of functional data are considered as in our case.

It should be pointed out that the nonparametric techniques based on orthogonal expansions are developed for many years. In particular, in classic papers [24, 25] asymptotically optimal classifiers are proposed, while in [26] a nonparametric, orthogonal series type estimator of a regression function is developed for random regressors. We refer the reader to [27] for further results and an extensive bibliography of papers on nonparametric estimation of such regression functions.

Papers on nonparametric regression estimation by orthogonal expansion in the so-called fixed-design case (with deterministic regressors) are closer to our needs. We refer the reader to the papers [22, 50, 55, 52] that are closely related to our needs.

Nonparametric, orthogonal expansion techniques were also applied for constructing tests for normality [33, 67] that can be useful here.

2 Problem statement and partial results

We formulate a hierarchy of one-class classification problems, starting from a version with full knowledge of probability distributions and finishing with our main problem of learning the novelty detection algorithm from historical and current observations, trying to reduce the number of assumptions as much as possible. Such a hierarchical formulations of problems and their solutions provides hints concerning ways of solving more realistic problems.

In general, an observed random element $\check{\mathbb{X}} \in \mathcal{H}$ has to be classified to the class "0" as typical for it or as untypical for the class "0" (novelty). This problem is considered under the following assumptions that apply throughout the paper.

- **H1)** Hilbert space $\mathcal{H}$ is equipped with a scalar product $< .,. >$ and it posses a countable orthonormal basis $\mathbb{V}_k$, $k = 1, 2, \ldots$.
- **H2)** For a sufficiently large, but finite, integer $K > 1$, subspace $\mathcal{H}_K \subset \mathcal{H}$, spanned by $\mathbb{V}_k$, $k = 1, 2, \ldots K$, contains all[1] the elements of the class "0".
- **H3)** According to H2), it suffices to consider $\mathbb{X} \in \mathcal{H}_K$, defined as follows

$$\mathbb{X} = \sum_{k=1}^{K} < \check{\mathbb{X}}, \mathbb{V}_k > \mathbb{V}_k \tag{6}$$

  as the element to be classified.
- **H4)** Coefficients $\theta_k =< \mathbb{X}, \mathbb{V}_k >$, $k = 1, 2, \ldots K$ are selected at random from a certain probability distribution on $\mathcal{R}^K$ with finite second moments.

A classification method should be such that if $\mathbb{X}$ is from the class "0", then the probability of classifying it as the untypical one (novelty) is not larger than α, for a preselected significance level[2] $0 < \alpha < 1$. This formulation generalizes to functional observations the classic formulation of one-class classification. On the other hand, it can be interpreted as testing the hypothesis $H_0 : \mathbb{X} \in$ class "0" at the significance level $0 < \alpha < 1$, while the alternative is $\mathbb{X}$ is not from the class "0". However, later on, we point out also differences between this statistical formulation and the proposed algorithms that arise from the need for learning invoked by rather weak assumptions.

2.1 A link between functional data and expansion coefficients

In this subsection, we assume that a vector of coefficients $\bar{\theta} = [\theta_1, \theta_2, \ldots \theta_K]^T$ has Gaussian distribution with $\bar{\theta}^0$ as the vector of expected values. Additionally, we assume that $\theta_1, \theta_2, \ldots \theta_K$ are mutually independent[3] (see [33] for the test that simultaneously verifies the assumptions about the normality and independence of random variables). For simplicity, we also assume that θ_k's have the same variances[4] $\sigma^2 = \sigma_k^2 > 0$, $k = 1, 2, \ldots K$. Thus, $\sigma^2 \mathbf{I}_K$ is the covariance matrix of $\bar{\theta}$, where $\mathbf{I}_K$ is $K \times K$ unit matrix.

Thus, $\bar{\theta}$ and the observed random element $\mathbb{X}$ can be expressed as follows

$$\bar{\theta} = \bar{\theta}^0 + \bar{\varepsilon}, \quad \bar{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_K), \quad \mathbb{X} = \mathbb{X}^0 + \mathbb{X}_\varepsilon, \tag{7}$$

where $\bar{\varepsilon} \in \mathcal{R}^K$ is Gaussian with zero mean vector and variances σ^2, while setting $\bar{\mathbb{V}}_K = [\mathbb{V}_1, \mathbb{V}_2, \ldots, \mathbb{V}_K]^T$ we have:

$$\mathbb{X} = \bar{\theta}^T \bar{\mathbb{V}}_K, \qquad \mathbb{X}^0 = [\bar{\theta}^0]^T \bar{\mathbb{V}}_K, \qquad \mathbb{X}_\varepsilon = \bar{\varepsilon}^T \bar{\mathbb{V}}_K. \tag{8}$$

Remark 1 *Notice that the Gaussian probability measure P_K in $\mathcal{R}_K$, related to $\bar{\theta}$, induces – by applying $\mathbb{X} = \bar{\theta}^T \bar{\mathbb{V}}_K$ – the Gaussian measure, μ_K, say, in $\mathcal{H}_K$ and for the expectation we have: $E(\mathbb{X}) = \mathbb{X}^0$, while variances of the projections: $< \mathbb{X}, \mathbb{V}_k >$ are equal to σ^2 for $k = 1, 2, \ldots K$. The relationship $\mathbb{X} = \bar{\theta}^T \bar{\mathbb{V}}_K$ allows us to apply probabilities P_K directly to $\mathbb{X}$ and related random elements in $\mathcal{H}_K$.*

Under the above assumptions, the problem of classifying $\mathbb{X}$ to the class "0" or deciding that it is outside this class with the significance level $0 < \alpha < 1$ can be rephrased as follows find $\rho(\alpha) > 0$ such that

$$P_K\{||\mathbb{X} - \mathbb{X}^0||^2 > \rho(\alpha)\} \leq \alpha, \tag{9}$$

[1] We comment on the choice of K further in this paper.

[2] As in statistics, it is customary to select a relatively small α, e.g., 0.01 or 0.05.

[3] For Gaussian random variables it suffices to test the lack of correlations.

[4] Later on we comment on how to relax the assumption on the same variances of random coefficients.

where norm $||.||$ in $\mathcal{H}$ is the one induced by $<.,.>$, i.e., $||\mathbb{X}||^2 =< \mathbb{X}, \mathbb{X} >$.

Define mapping $J : \mathcal{H}_K \to \mathcal{R}^K$ as follows $\forall_{\mathbb{Z}\in\mathcal{H}_K}\ J(\mathbb{Z}) = \bar{z}$, where $\bar{z} \stackrel{def}{=} [z_1, z_2, \ldots, z_K]^T$, while $z_k =< \mathbb{Z}, \mathbb{V}_k > k = 1, 2, \ldots K$.

Corollary 1 (Isometry between $\mathcal{H}_K$ and $\mathcal{R}^K$) *Mapping J is a linear isometry between $\mathcal{H}_K$ and $\mathcal{R}^K$, which implies the following.*

- ***C1)*** *Inverse mapping J^{-1} exists and $\mathbb{Z}$ can be uniquely restored from $\bar{z} \in \mathcal{R}^K$.*
- ***C2)*** *For every $\mathbb{Z} \in \mathcal{H}_K$ we have: $||\mathbb{Z}||^2 = \sum_{k=1}^K z_k^2$.*
- ***C3)*** *For every $\rho > 0$*

$$P_K\{||\mathbb{X} - \mathbb{X}^0||^2 > \rho\} = P_K\{\sum_{k=1}^K (\theta_k - \theta_k^0)^2 > \rho\}. \tag{10}$$

Indeed, C1) follows from $\mathbb{Z} = \bar{z}^T \bar{\mathbb{V}}_K$, C2) is the consequence of the orthonormality of $\mathbb{V}_k$'s and

$$< \mathbb{Y}, \mathbb{Z} >= \sum_{k=1}^K y_k z_k = \bar{y}^T \bar{z}, \qquad \bar{y} = J(\mathbb{Y}), \tag{11}$$

by setting $\mathbb{Y} = \mathbb{Z}$, while C3) follows directly from C2).

2.2 The algorithm outline – an idealized case

From (7) it follows that the following expression:

$$\sum_{k=1}^K \left(\frac{\theta_k - \theta_k^0}{\sigma}\right)^2 \tag{12}$$

has the χ^2 distribution with K-th degree of freedom. Denote by $q(\alpha)$ the $(1-\alpha)$-quantile of this distribution.

Corollary 2 (Selecting threshold) *Comparing C3) and* (12) *we obtain:*

$$P_K\{||\mathbb{X} - \mathbb{X}^0||^2 > \sigma^2 q(\alpha)\} =$$

$$= P_K\{\sum_{k=1}^K (\theta_k - \theta_k^0)^2 > \sigma^2 q(\alpha)\} \le \alpha. \tag{13}$$

Hence, (9) *holds for* $\rho(\alpha) = \sigma^2 q(\alpha)$.

The following skeletal algorithm summarizes the above considerations. By (10) it fulfills the theoretical requirement at the expense of assumed a priori information.

Algorithm 0

Step 0) Select a significance level $0 < \alpha < 1$ and calculate $\rho(\alpha) = \sigma^2 q(\alpha)$ for $q(\alpha)$ from χ^2 distribution with K degrees of freedom. Select orthonormal basis $\mathbb{V}_k$, $k = 1, 2, \ldots, K$.

Step 1) Acquire (observe) $\mathbb{X}$ and calculate $\theta_k =< \mathbb{X}, \mathbb{V}_k >$, $k = 1, 2, \ldots, K$.

Step 2) Check the inequality:

$$||\mathbb{X} - \mathbb{X}^0||^2 = \sum_{k=1}^K (\theta_k - \theta_k^0)^2 > \rho(\alpha). \tag{14}$$

If it holds, decide that $\mathbb{X}$ is not in the class "0" (declare that $\mathbb{X}$ is a novelty). Otherwise, accept $\mathbb{X}$ as an element from class "0". Go to Step 1).

Remark 2 *If components of $\bar{\theta}$ are correlated with $K \times K$ covariance matrix Σ, then the vector $\Sigma^{-1/2}(\bar{\theta} - \bar{\theta}^0)$ has Gaussian distribution with $\mathbf{I}_K$ covariance matrix and zero mean vector. Hence, the quadratic form:*

$$\phi(\bar{\theta}) \stackrel{def}{=} (\bar{\theta} - \bar{\theta}^0)^T \Sigma^{-1} (\bar{\theta} - \bar{\theta}^0) \tag{15}$$

has the χ^2 distribution with K degrees of freedom.

3 Relaxing assumptions by learning from historical or surrogate data

Our aim in this subsection is to point out those ingredients of Algorithm 0 that can be learned from historical data: $\mathbb{X}_n \in \mathcal{H}_K$, $n = -1, -2, \ldots, -N_H$ of length $N_H \ge 1$, belonging to the class "0" only, where negative subscripts are used to indicate that these observations were collected in the past and they are statistically independent from newly coming ones. Let us assume that coefficients $\bar{\theta}_n = J(\mathbb{X}_n)$, $n = -1, -2, \ldots, -N_H$ of historical data are already calculated.

We consider also the case when one has only a minimal number of historical data, but the earlier collected knowledge on the process at hand can

be formulated as a mathematical model that can be used for generating surrogate observations playing the same role as historical data.

3.1 Learning from retrospective observations

When N_H is sufficiently large, then one can replace a priori knowledge concerning $\bar{\theta}^0$ by

$$\hat{\bar{\theta}}^0 = N_H^{-1} \sum_{k=-N_H}^{-1} \bar{\theta}_n \tag{16}$$

or by its well-known recurrent version. Analogously, Σ can be replaced by its empirical counterpart:

$$\hat{\Sigma} = \frac{1}{N_H - 1} \sum_{k=-N_H}^{-1} (\bar{\theta}_n - \hat{\bar{\theta}}^0)(\bar{\theta}_n - \hat{\bar{\theta}}^0)^T. \tag{17}$$

This version of learning $\hat{\Sigma}$ is the basic one, but for computational purposes more advanced and more accurate algorithms are advised [23], especially for large covariance matrices [65].

Assume that $N_H \geq K$. Then, for N_H sufficiently large, one can also assume that $\hat{\Sigma}$ is invertible (see, e.g., [2] for a discussion on this topic). In such a case, the empirical version of the quadratic form (15) is given by

$$\hat{\phi}(\bar{\theta}) \stackrel{def}{=} (\bar{\theta} - \hat{\bar{\theta}}^0)^T \hat{\Sigma}^{-1} (\bar{\theta} - \hat{\bar{\theta}}^0), \tag{18}$$

where $\hat{\phi}$ depends also on K and on the learning sequence, but this is not displayed in the notation.

Remark 3 *It is known (see, e.g., [41]) that $\hat{\phi}(\bar{\theta})$ has the Hotelling T^2 distribution (or appropriately rescaled F-distribution) with the degrees of freedom depending on K and N_H. This distribution is directly used for calculating the threshold $\rho(\alpha)$ (see also [49]).*

Then, similarly as in the theory of designing statistical control charts, it is reasonable to to confront this threshold with learning data and – if necessary – to tune it appropriately. Thus, the following two phase algorithm can be proposed.

Algorithm 1

Learning phase

Step 1 Collect learning data: $\mathbb{X}_n \in \mathcal{H}_K$, $n = -1, -2, \ldots, -N_H$.

Step 2 Compute $\hat{\bar{\theta}}^0$ and $\hat{\Sigma}$.

Step 3 Tuning the threshold: if $\hat{\phi}(\bar{\theta}_n) \leq \rho(\alpha)$ for all $n = -1, -2, \ldots, -N_H$, then set $\hat{\rho} = \rho(\alpha)$ and go to the Application phase. Otherwise, set

$$\hat{\rho} = \max_{-N_H \leq n \leq -1} \left[\hat{\phi}(\bar{\theta}_n)\right]. \tag{19}$$

Application phase

Acquire (observe) $\mathbb{X}$ and calculate elements of $\bar{\theta}$ as: $\theta_k =< \mathbb{X}, \mathbb{V}_k >$, $k = 1, 2, \ldots, K$.

Check the inequality: $\hat{\phi}(\bar{\theta}) > \hat{\rho}$ and if it holds, decide that $\mathbb{X}$ is not in the class "0" (declare that $\mathbb{X}$ is a novelty). Otherwise, accept $\mathbb{X}$ as an element from the class "0" and acquire new $\mathbb{X}$ for testing.

Corollary 3 (Algorithm 1 correctness) *For the Gaussian distribution of historical and current observations, if $\hat{\Sigma}$ is nonsingular, then – by construction – Algorithm 1 may erroneously reject $\mathbb{X}$ from the class "0" with the probability not larger than preselected $0 < \alpha < 1$.*

If observations are not Gaussian, then one can replace the test statistics (18) by the one recently proposed in [32]. This statistic is based on the ranks and it is a nonparametric one.

Algorithm 1 is ready for use with one exception, namely, one has to point out how to evaluate $\theta_k =< \mathbb{X}, \mathbb{V}_k >$, $k = 1, 2, \ldots, K$ for current and historical data. This topic is discussed in the next section.

3.2 Learning from model-based surrogate data

It may happen that we do not have enough historical observations. In the extreme case, $N_H = 1$, as in the case study presented in the final section. However, if we have a mathematical model of the process at

hand, then it is still possible to use Algorithm 1 by generating surrogate observations.

The outline of this approach is the following. Assume that we have a model $\mathbb{F}(\bar{a}) \in \mathcal{H}_K$, or in a more detailed form:

$$\mathbb{Y}(t) = \mathbb{F}(\bar{a})(t), \qquad t \in [0,\tau], \tag{20}$$

where $\bar{a} \in \mathcal{R}^d$, $d \geq 1$ is the vector of tunable parameters. This model can be given explicitly as in (20) or implicitly, e.g., as a solution of a certain differential equation.

Assume also that for certain $\bar{a}^0 \in \mathcal{R}^d$ we are able to generate the central element of the class "0", i.e., $\mathbb{X}^0 = \mathbb{F}(\bar{a}^0)$.

Method of generating surrogate data

Stage 1 Model tuning: having historical observations $\mathbb{X}_n \in \mathcal{H}_K$, $n = -1, -2, \ldots, -N_H$ at our disposal, one can estimate $\bar{a}^0$ in the obvious, but not always computationally simple, way as

$$\hat{\bar{a}}^0 = \arg\min_{\bar{a}} \sum_{n=-1}^{-N_H} ||\mathbb{X}_n - \mathbb{F}(\bar{a})||^2. \tag{21}$$

Stage 2 Data generation: select number $N_S \geq 1$ of surrogate observations $\hat{\mathbb{X}}_n$ and for $n = (-N_H - 1), (-N_H - 2), \ldots, (-N_H - N_S)$ calculate

$$\hat{\mathbb{X}}_n = \mathbb{F}(\hat{\bar{a}}^0 + \bar{\eta}_n), \tag{22}$$

where $\bar{\eta}_n \in \mathcal{R}^d$ are i.i.d. samples from a certain, e.g., Gaussian, distribution with zero mean and the unit covariance matrix.

Stage 3 Concatenate the historical data $\mathbb{X}_n$, $n = -1, -2, \ldots, -N_H$ and the surrogate data $\hat{\mathbb{X}}_n, n = (-N_H - 1), (-N_H - 2), \ldots, (-N_H - N_S)$ into the learning sequence.

After extending the learning sequence, feed it as the input of Algorithm 1 or as inputs of algorithms that are described in the next sections.

4 Learning algorithm from data along one pass

As it was assumed in the Introduction, the process under consideration is a repetitive one. It consists of passes for $t \in [0,\tau]$. Along each pass we observe random element $\mathbb{X}(t)$, $t \in [0,\tau]$ and similar functions, e.g., retrospective data. The aim of this section is to propose a learning algorithm that enters into details of estimating $\theta_k =< \mathbb{X}, \mathbb{V}_k >$'s and related expansion coefficients that are needed in Algorithm 1, where $\mathbb{X}$ is current random element to be classified.

We would like to derive a learning algorithm that is convergent to $\bar{\theta}$ and we can asses its accuracy, at least for a sufficiently large number of observations. Here, we mention only that in our case $\bar{\theta}$ and related expansion coefficients are random. Thus, expectations, variances etc., that appear in our derivations will be conditioned on $\bar{\theta}$. Hence, also our conclusions will be conditionally PAC.

4.1 Learning of one curve – theoretical foundations

In order to derive a learning algorithm, we have to be more detailed in specifying $\mathcal{H}_K$ and observations.

4.1.1 More specialized assumptions

The following assumptions specialize or extend those listed as H1) - H4).

h1) As Hilbert space $\mathcal{H}$ we take $L_2(0,\tau)$ with the inner product and the norm:

$$< \mathbb{Y}, \mathbb{Z} >= \int_0^\tau \mathbb{Y}(t)\,\mathbb{Z}(t)\,dt, \; ||\mathbb{Y}||^2 =< \mathbb{Y}, \mathbb{Y} > .$$

The orthonormal and complete in $L_2(0,\tau)$ sequence $\mathbb{V}_k$, $k = 1, 2, \ldots$ consists of absolutely continuous functions that are commonly[5] bounded, i.e., there exists $0 < \omega < \infty$ such that $\forall_k \forall_{t\in[0,\tau]} |\mathbb{V}_k(t)| \leq \omega$.

h2) Element $\mathbb{X}$ is either from $\mathcal{H}_K$ or

$$||\mathbb{X} - \sum_{k=1}^{K} \theta_k \mathbb{V}_k||^2 = \sum_{k=K+1}^{\infty} \theta_k^2 \leq C/K \tag{23}$$

for a certain constant $0 < C < \infty$ that may depend on $\mathbb{X}$, but not on K, while $\theta_k =< \mathbb{X}, \mathbb{V}_k >$.

[5] This assumption is made for simplicity of formulas. It can be relaxed by admitting that ω grows, at most polynomially, with k, as it is in the case of the Legendre polynomials (see [55] for more examples).

h3) Element $\mathbb{X}$ to be classified has the form: $\mathbb{X} = \bar{\theta}^T \bar{\mathbb{V}}_K,$, $\bar{\theta} = \bar{\theta}^0 + \bar{\varepsilon}$, and

$$\mathbb{X} = \mathbb{X}^0 + \mathbb{X}_\varepsilon, \ \mathbb{X}^0 \stackrel{def}{=} [\bar{\theta}^0]^T \bar{\mathbb{V}}_K, \ \mathbb{X}_\varepsilon \stackrel{def}{=} \bar{\varepsilon}^T \bar{\mathbb{V}}_K,$$

where $\bar{\varepsilon} \in \mathcal{R}^K$ is a random vector with zero mean and mutually uncorrelated Gaussian components, having variances $\sigma^2 > 0$.

h4) Available observations x_i, $i = 1, 2, \ldots, m$ of $\mathbb{X}$ are of the following form:

$$x_i = \mathbb{X}(t_i) + \zeta_i, \qquad i = 1, 2, \ldots, m, \quad (24)$$

where ζ_i's are zero mean, finite variance, i.i.d., random variables that are not necessarily Gaussian. Observation errors ζ_i's are defined on the same probability space as $\bar{\theta}$ and they are mutually independent. Observation points t_i's are placed equidistantly in $[0, \tau]$ with the distance $\Delta_m = \tau/m$, assuming $t_1 = 0$.

Assumptions h1)-h4) apply also for historical data $\mathbb{X}_{-n}$'s. In such cases, index n will attached to x_i's.

Remark 4 *It is not difficult to specify classes of functions $\mathbb{X}(t)$, $t \in [0, \tau]$ and orthonormal systems for which assumption h2) holds. For example, if $\mathbb{V}_k$'s form the trigonometric basis and θ_k's decay as k^{-1}, then h2) holds. The required rate k^{-1} can be assured if, e.g., $\mathbb{X}$ is absolutely continuous.*

Remark 5 *As is known, see [50, 45, 46], in non-parametric regression estimation, it is desirable to observe an estimated function either at t_i's that are equidistributed in $[0, \tau]$ or at the nodes of a highly accurate quadrature formulas, but it is not always possible. For this reason, we confine our attention to the most frequent case of equidistant observations.*

In the problem of learning $\mathbb{X}$, hence also $\bar{\theta}$, from (24) we assume that $\bar{\theta}^0$ is known from historical data. If drawn according to h3), $\bar{\theta}$ remains unchanged when observations h4) are acquired. When a decision concerning $\mathbb{X}$ is made, a new pass starts from drawing the next $\mathbb{X}$. The proposed algorithm is designed to learn one $\mathbb{X}$ from one pass[6] of observations (24) only.

4.1.2 Learning expansion coefficients

When $\mathcal{H}_K \subset L_2(0, \tau)$, then parameters are given explicitly as:

$$\theta_k = \int_0^\tau \mathbb{X}(t)\, \mathbb{V}_k(t)\, dt, \qquad k = 1, 2, \ldots, K. \quad (25)$$

Thus, it is natural to estimate them as follows

$$\hat{\theta}_k = \frac{\tau}{m} \sum_{i=1}^{m} x_i\, \mathbb{V}_k(t_i), \qquad k = 1, 2, \ldots, K. \quad (26)$$

We shall occasionally write $\hat{\theta}_k(m)$ when the role of the number of observations should be underlined, while a vector of $\hat{\theta}_k$'s is denoted as $\hat{\bar{\theta}}$ or as $\hat{\bar{\theta}}(m)$.

Corollary 4 (Asymptotic unbiasedness) *If $\mathbb{X}(t)$ is Lipschitz continuous in $[0, \tau]$ with constant $L > 0$ and h1)-h4) hold, then for the bias of $\hat{\theta}_k$ we have*

$$|E_\zeta[\hat{\theta}_k] - \theta_k| \leq L\tau/m, \quad (27)$$

where E_ζ stands for the expectation with respect to ζ_i's, assuming θ_k's to be fixed. Hence, $\hat{\theta}_k$ is asymptotically unbiased as $m \to \infty$.

Indeed, $E_\zeta[\hat{\theta}_k] = \frac{\tau}{m} \sum_{i=1}^{m} \mathbb{X}(t_i)\, \mathbb{V}_k(t_i)$ and the bias is equal to the error of the Riemann sum quadrature formula.

Remark 6 *It is known that using more sophisticated quadrature formulas in* (26) *is beneficial [50] for the estimation accuracy of $\hat{\theta}_k$, but for the purposes of on-line novelty detection that is proposed in the next section, version* (26) *is preferable. Nevertheless, even for equidistant sampling* (26) *can be unbiased when $\mathbb{V}_k$'s is the trigonometric basis (see Section 5). Also, the proofs of properties of* (26) *are more informative.*

By h4) and h1), for the variance $Var_\zeta[\hat{\theta}_k]$ we obtain

$$Var_\zeta[\hat{\theta}_k] = \frac{\tau^2}{m^2} \sum_{i=1}^{m} E_\zeta[x_i - \mathbb{X}(t_i)]^2\, \mathbb{V}_k^2(t_i) = \quad (28)$$

$$= \frac{\tau^2 \sigma^2}{m^2} \sum_{i=1}^{m} \mathbb{V}_k^2(t_i) \leq \frac{\tau^2 \sigma^2 \omega^2}{m}.$$

This bound has the exact m^{-1} order, since for large m we have $(\tau/m) \sum_{i=1}^{m} \mathbb{V}_k^2(t_i) \approx 1$.

[5] One can allow $\bar{\theta}$ to have a general covariance matrix and to apply de-correlation, described in the previous section. The assumption about normality is made for theoretical purposes only. The algorithm proposed in this section is applicable without this assumption.

[6] Learning from pass-to-pass data is also possible, but it is too complicated to be described here.

4.1.3 MSE bound and consistency

Summarizing, for the mean square error (MSE) of $\hat{\theta}_k$ one gets

$$E_\zeta[\hat{\theta}_k-\theta_k]^2 = Var_\zeta[\hat{\theta}_k]+|E_\zeta[\hat{\theta}_k]-\theta_k|^2 \leq \quad (29)$$

$$\leq \frac{\tau^2\sigma^2\omega^2}{m}+\frac{L^2\tau^2}{m^2}.$$

Corollary 5 (MSE upper bound) *Under the same assumptions as in Corollary 4, for $\mathbb{X}\in\mathcal{H}_K$ and for $\hat{\mathbb{X}} \stackrel{def}{=} \hat{\bar{\theta}}^T\bar{\mathbb{V}}_K$ we obtain*

$$E_\zeta||\mathbb{X}-\hat{\mathbb{X}}||^2 = E_\zeta||\bar{\theta}-\hat{\bar{\theta}}||_K^2 \leq \quad (30)$$

$$\leq \frac{K}{m}\left[\tau^2\sigma^2\omega^2+\frac{L^2\tau^2}{m}\right],$$

where $||.||_K$ is the Euclidean norm in $\mathcal{R}^K$.

Notice that if $\mathbb{X}\in L_2(0,\tau)$, but $\mathbb{X}\notin\mathcal{H}_K$, then – according to h2) – one has to add C/K to the right hand side of (30).

Corollary 6 *Under the same assumptions as in Corollary 5 and selected $0<\alpha<1$, for every $\varepsilon>0$ one can find $m^*(\varepsilon)$ such that for $m\geq m*(\varepsilon)$ we have*

$$P_\zeta\{||\mathbb{X}-\hat{\mathbb{X}}||>\varepsilon\} = P_\zeta\{||\bar{\theta}-\hat{\bar{\theta}}||_K>\varepsilon\}\leq\alpha \quad (31)$$

and $m^*(\varepsilon)$ is given by

$$m^*(\varepsilon)=\frac{\sqrt{K}\sqrt{\tau}\sqrt{4\alpha\varepsilon L^2\tau^2+K\omega^4\sigma^4\tau^5}+K\omega^2\sigma^2\tau^3}{2\alpha\varepsilon}.$$

For the proof, we firstly apply the Markov inequality to $P_\zeta\{||\bar{\theta}-\hat{\bar{\theta}}||_K>\varepsilon\}$, which yields

$$P_\zeta\{||\bar{\theta}-\hat{\bar{\theta}}||_K>\varepsilon\} \leq \varepsilon^{-1}E_\zeta||\bar{\theta}-\hat{\bar{\theta}}||_K \leq \quad (32)$$

$$\leq \tau\varepsilon^{-1}E_\zeta||\bar{\theta}-\hat{\bar{\theta}}||_K^2,$$

where the second inequality follows from the Schwarz inequality. Applying Corollary 5 the last term in (32) one gets

$$P_\zeta\{||\bar{\theta}-\hat{\bar{\theta}}||_K>\varepsilon\} \leq \tau^3\varepsilon^{-1}\frac{K}{m}\left[\sigma^2\omega^2+\frac{L^2}{m}\right]. \quad (33)$$

Now, it suffices to require that the last term in (33) is equal to α and to solve the resulting equation for m to obtain $m^*(\varepsilon)$, by selecting the larger of the two solutions.

This bound is in the spirit of the well-known (see, e.g., [16, 68]) PAC-learning (probably approximately correct learning), but we shall not follow this line of research later in this paper.

If $\mathbb{X}\in L_2(0,\tau)$, but $\mathbb{X}\notin\mathcal{H}_K$, then (30) holds for the orthogonal projection of $\mathbb{X}$ onto $\mathcal{H}_K$, further denoted as $O_K(\mathbb{X})$.

Corollary 7 (MSE consistency)
1) Under assumptions of Corollary 5,

$$E_\zeta||\mathbb{X}-\hat{\mathbb{X}}(m)||\to 0 \text{ as } m\to\infty.$$

2) If $\mathbb{X}\notin\mathcal{H}_K$, then $E_\zeta||O_K(\mathbb{X})-\hat{\mathbb{X}}(m)||\to 0$ as $m\to\infty$.

It is also well known [55] that if K depends on m in such a way that $K(m)\to\infty$ and $K(m)/m\to 0$ as $m\to\infty$, then $\hat{\mathbb{X}}(m)$ is also MSE consistent for $\mathbb{X}\notin\mathcal{H}_K$.

Remark 7 (MSE convergence rate) *Adding $C/K(m)$ to the right hand side of* (30) *and optimizing the result with respect to K we obtain MSE convergence rate $O(m^{-1/2})$, which is attained for $c_1 m^{-1/2}\leq K(m)\leq c_2 m^{-1/2}$, where $0<c_1<c_2<\infty$. This rate is slightly slower than the best possible one, namely, $O(m^{-2/3})$. This rate is attainable by $\hat{\mathbb{X}}(m)$ for $\mathbb{V}_k$'s being a trigonometric system of cosine type, even for $\mathbb{X}\notin\mathcal{H}_K$, but assuming $\mathbb{X}(t)$ to be periodic, continuously differentiable in $[0,\tau]$ with periodic derivative (see [55] and Section 5 in which discrete cosine transform of the type II/III is applied, since they have orthogonal transformation matrices that implies the equality of integrals and discrete sums of functions from $\mathcal{H}_K$).*

However, later on we are more interested in the behavior of $\hat{\mathbb{X}}(m)$ for m finite, although maybe large.

4.1.4 Asymptotic distribution of learning error

It is clear that for Gaussian ζ_i's also errors $(\hat{\theta}_k(m)-\theta_k)$'s are normally distributed for any $m\geq 1$. If m is large, one can derive the asymptotic normality of these errors for an arbitrary distribution of ζ_i's, assuming only that they are zero mean, finite variance i.i.d. random variables.

Corollary 8 (Asymptotic normality) *Under the same assumptions as in Corollary 4, for $k=$*

$1, 2, \ldots, K$

$$\frac{\sqrt{m}}{\sigma\sqrt{\tau}}\left[\hat{\theta}_k(m)-\theta_k\right] \xrightarrow{d} \mathcal{N}(0,1) \text{ as } m\to\infty, \tag{34}$$

where $\xrightarrow{d}$ *denotes the convergence in distribution to the standard normal distribution.*

For the proof, consider the following expression

$$\hat{\Theta}_k(m) \stackrel{def}{=} \frac{\hat{\theta}_k - E_\zeta[\hat{\theta}_k(m)]}{\sqrt{Var_\zeta[\hat{\theta}_k(m)]}} = \frac{\frac{\tau}{m}\sum_{i=1}^m \zeta_i \mathbb{V}_k(t_i)}{\sqrt{Var_\zeta[\hat{\theta}_k(m)]}}. \tag{35}$$

In order to verify that the Lindenberg condition (see, e.g., [57]) holds in this case, consider the variance of one summand in the numerator of (35) to the overall variance $Var_\zeta[\hat{\theta}_k(m)]$ of this sum, which yields

$$\frac{\mathbb{V}_k^2(t_j)}{\sum_{i=1}^m \mathbb{V}_k^2(t_i)} \le \frac{\omega^2}{\sum_{i=1}^m \mathbb{V}_k^2(t_i)} \to 0 \text{ as } m\to\infty. \tag{36}$$

Thus, also the maximum of this expression w.r.t. $j = 1, 2, \ldots, m$ is convergent to zero, which implies that the Lindenderg condition hold. Thus, from the Lindenberg-Feller central limit theorem (CLT) holds [57] and $\hat{\Theta}_k(m) \xrightarrow{d} \mathcal{N}(0,1)$ as $m\to\infty$. From Corollary 4 we have $E_\zeta[\hat{\theta}_k(m)\to\theta_k$. Notice that $\frac{\tau}{m}\sum_{i=1}^m \mathbb{V}_k^2(t_i)\to 1$ as $m\to\infty$. Hence, for m sufficiently large $Var_\zeta[\hat{\theta}_k]$ can be approximated by $\tau\sigma^2/m$. Thus, from the Slutsky's theorem (see, e.g., [5]) we know that one can replace the corresponding terms in $\hat{\Theta}_k(m)$ by θ_k and $\tau\sigma^2/m$, respectively, still keeping its asymptotic normality.

Corollary 9 (Asymptotic uncorrelatedness) *Under the same assumptions as in Corollary 4, for* $k, l = 1, 2, \ldots, K$, *if* $k\neq l$, *then* $\hat{\theta}_k(m)$ *and* $\hat{\theta}_l(m)$ *are asymptotically uncorrelated, since*

$$Cov_\zeta(\hat{\theta}_k(m),\hat{\theta}_l(m))\to 0 \text{ as } m\to\infty \tag{37}$$

Indeed,

$$Cov_\zeta(\hat{\theta}_k(m),\hat{\theta}_l(m)) = \frac{\tau^2}{m^2}\sum_{i,j=1}^m E_\zeta[\zeta_i\zeta_j]\,\mathbb{V}_k(t_i)\,\mathbb{V}_l(t_j)$$

$$= \frac{\sigma^2\tau}{m}\frac{\tau}{m}\sum_{i=1}^m \mathbb{V}_k(t_i)\,\mathbb{V}_l(t_i) = o(m^{-1}),$$

because, for $k\neq l$, as $m\to\infty$ we have

$$\frac{\tau}{m}\sum_{i=1}^m \mathbb{V}_k(t_i)\,\mathbb{V}_l(t_i) \to \int_0^\tau \mathbb{V}_k(t)\,\mathbb{V}_l(t)\,dt = 0. \tag{38}$$

From Corollary 8 and 9 we immediately obtain.

Corollary 10 (Asymptotic squared error distribution) *Under the same assumptions as in Corollary 4 we have:*
1) for χ^2_K *denoting the standard* χ^2 *distribution with K degrees of freedom,*

$$\frac{m}{\sigma^2\tau}\sum_{k=1}^K \left[\hat{\theta}_k(m)-\theta_k\right]^2 \xrightarrow{d} \chi^2_K \text{ as } m\to\infty, \tag{39}$$

2) for the selected significance level $0<\alpha<1$ *and* $q(\alpha)$ *being* $(1-\alpha)$ *quantile of the above distribution we asymptotically have*

$$P_\zeta\{||\hat{\bar{\theta}}-\bar{\theta}||^2_K > \kappa\} \le \alpha, \tag{40}$$

where $\kappa = q(\alpha)\sigma^2\tau/m$, *while* P_ζ *is the probability measure on* $\mathcal{R}^K$ *that is induced by* ζ_i*,s and conditioned by* $\bar{\theta}$.

When σ is estimated from the same observations as $\hat{\bar{\theta}}$, then $q(\alpha)$ should be selected from the T^2 Hotelling distribution.

4.2 Fast algorithm for learning one curve

For fixed $K\ge 1$, our starting point is the formula (26) that can be rewritten in the vector form as follows

$$\hat{\bar{\theta}} = \frac{\tau}{m}\sum_{i=1}^m x_i\,\bar{\mathbb{V}}(t_i) = \frac{\tau}{m}\overrightarrow{\mathbb{V}}\,\bar{x}, \tag{41}$$

where $\overrightarrow{\mathbb{V}}$ is $K\times m$ matrix composed by stacking columns $\bar{\mathbb{V}}(t_i)$'s, while $\bar{x}$ is $m\times 1$ vector of x_i's.

For many linear transformations of the form (41) there exist fast algorithms for calculating $\hat{\bar{\theta}}$ in $O(m\,log(m)$ operations, the most notable being the fast Fourier transform (see [47] for the FFT algorithm used in nonparametric regression function estimation). For our purposes we choose $\mathbb{V}_k$'s forming the cosine series and its discrete counterpart, known as the discrete cosine transform (DCT), since it has the fast implementation. Advantages of the DCT made it the most popular algorithm in audio, video and image processing (see, e.g., [7]). In our computational tests the so-called type II DCT (or DCT-II) was used (see [44] for recent contribution and the bibliography cited therein), since it is even with respect to zero and the inverse of DCT-II is the same as DCT-III, up to a scaling factor.

The following algorithm provides the estimates: $\hat{\tilde{\theta}}$ and

$$\hat{x}_i \stackrel{def}{=} \hat{\mathbb{X}}(t_i) = \hat{\tilde{\theta}}^T \bar{\mathbb{V}}(t_i), \quad i = 1, 2, \ldots, m. \quad (42)$$

Algorithm 2

Input: K, m, $K < m$, τ and $\{x_i\}$, considered as the whole sequence $\bar{x} = [x_i, i = 1, 2, \ldots, m]$.

Step 1 Calculate $m \times 1$ vector $\tilde{\bar{\theta}}$, say, from $\{x_i\}$, using the fast version of DCT-II.

Step 2 Select K first elements of $\tilde{\bar{\theta}}$ and form $\hat{\tilde{\theta}}$ from them.

Step 3 Set to zero elements indexed as $(K+1), (K+2), \ldots, m$ of $\tilde{\bar{\theta}}$ and feed this vector as the input of the inverse of DCT-II algorithm. The output of the inverse of DCT-II algorithm is $\{\hat{x}_i\}$ sequence.

Output: $\hat{\tilde{\theta}}$ and $\{\hat{x}_i\}$ sequence.

The action of Algorithm 2 on $\{x_i\}$ is further denoted as $[\hat{\tilde{\theta}}, \{\hat{x}_i\}] = A(\{x_i\}, K)$, omitting τ, m arguments for brevity. Algorithm 2 plays a crucial role in classifying current $\mathbb{X}$, since it provides the estimate $\hat{\tilde{\theta}}$ of feature vector $\hat{\theta}$. Its role in the learning path from historical data is different, namely, it provides $\bar{\theta}_{-n}$'s for learning de-correlation matrix (see Fig. 1). The second ingredient of the output of Algorithm 2 is $\{\hat{x}_i\}$ that can be used for the selection of K, if it is not a priori specified. To this end, one can use Akaike's information criterion, the Bayesian information criterion and others. We also refer the reader to [33] and [72] and the bibliographies cited therein for methods of data driven selection of K, in the related problems of hypothesis testing when smooth alternatives are specified by orthogonal expansions.

Notice, however, that in our case we have to select K which is suitable for the whole family of functions from the class "0". This topic is outside the scope of this paper.

5 Aggregated learning algorithm

Algorithms 1 and 2 are designed in such a way that they may (and in some cases) should be used together. In this section, we outline an aggregated algorithm (Algorithm 4) of their cooperation. Already at the beginning, we turn the reader attention that Algorithm 2 may appear at different steps of the aggregated algorithm. On the other hand, in some cases, as described in the next subsection, it suffices to use Algorithm 2 only once.

5.1 Uncorrelated descriptors

If descriptors θ_k, $k = 1, 2, \ldots, m$ are uncorrelated, it suffices to apply Algorithm 2 for deciding whether each new $\mathbb{X}$ is typical or not.

Algorithm 3

Input: K, m, $K < m$, τ, $\bar{\theta}^0$ (or $\hat{\bar{\theta}}^0$, if it results from learning that is based on historical data). Establish the threshold $\rho > 0$ (see a discussion below).

Step 1 Acquire observations $\{x_i\}$ of $\mathbb{X}$, considered as the whole sequence x_i, $i = 1, 2, \ldots, m$.

Step 3 Execute Algorithm 2 in order to obtain: $[\hat{\tilde{\theta}}, \{\hat{x}_i\}] = A(\{x_i\}, K)$.

Step 4 If $||\hat{\tilde{\theta}} - \bar{\theta}^0||^2_K > \rho$, declare $\hat{\tilde{\theta}}$ (hence, also $\mathbb{X}$) to be a novelty. Otherwise, classify them to the class "0" and go to Step 1.

Data $\{\hat{x}_i\}$ obtained Step 3 might be useful in exploring the reasons why curve $\mathbb{X}$ at hand was declared to be a novelty. One may also consider $m^{-1} \sum_{i=1}^m [\mathbb{X}^0(t_i) - \hat{x}_i]^2$ as an alternative criterion for novelty detection, but this topic is outside the scope of this paper.

It remains to point out how to select ρ. To this end, it suffices to collect facts already established.

The difference $\hat{\tilde{\theta}} - \bar{\theta}^0$ has two random components that induce two kinds of variability, namely,

1. $\hat{\tilde{\theta}} - \bar{\theta}$ – variability introduced by random errors in x_i's.
2. $\bar{\theta} - \bar{\theta}^0$ – variability that is inherent for $\mathbb{X}$ coming from the class "0".

Concerning case 1., we know that the bias of $\hat{\tilde{\theta}}$ is zero, if DCT-II orthogonal sequence is used, since then $\int_0^\tau \mathbb{V}_k(t)\, \mathbb{V}_l(t)\, dt = 0$ for $k \neq l$ and simultaneously $\frac{\tau}{m} \sum_{i=1}^m \mathbb{V}_k(t_i)\, \mathbb{V}_l(t_i) = 0$. For other basis we known from Corollary 4 that $|E_\zeta[\hat{\theta}_k] - \theta_k|^2$ is of the

order $o(m^{-1})$ and it can be neglected. From Corollary 5 we also know that the variance term of $\hat{\bar{\theta}} - \bar{\theta}$ is of the order K/m and it can be made arbitrarily small by selecting m sufficiently large, both for $K = nst$ and for $K(m)/m \to 0$. Thus, we can omit the variability described in case 1.

The second source of variability is crucial for selecting $\rho > 0$, since it cannot be reduced. When $\bar{\theta} - \bar{\theta}^0$ is the main source of variability and $\bar{\theta}$ is Gaussian with uncorrelated components, then we already have the rule of selecting ρ, namely, apply Corollary 2.

When $\bar{\theta}$ has an unknown distribution, then the proposed way of selecting *varrho* is based on historical observations $\mathbb{X}_{-n}$ that are transformed to $\hat{\bar{\theta}}_{-n}$, $n = 1, 2, \ldots, N_H$ by Algorithm 2. It has the following form:

$$\rho = \max_{n=1,2,\ldots,N_H} ||\bar{\theta}^0 - \hat{\bar{\theta}}_{-n}||^2 - \delta, \qquad (43)$$

where $\delta \geq 0$ is a tuning parameter. By setting $\delta = 0$ we decide to include all the learning examples to the class "0", according to their formal definition. However, if one has doubts whether examples having relatively large values of $||\bar{\theta}^0 - \hat{\bar{\theta}}_{-n}||^2$ should indeed be included, then one can select $\delta > 0$ in such a way that a prescribed fraction $\alpha > 0$ of them is excluded from further computations.

Algorithm 3 includes the main path of decision making that is depicted in Figure 1. It consists of blocks labelled as E) and F) in this figure. For uncorrelated descriptors, block D is not executed. Its role is described in the next subsection.

5.2 Correlated descriptors

When components of $\bar{\theta}$ are correlated, then we have two cases:

- Σ *is known* e.g., from a mathematical model or from a very large number of historical data, then apply Remark 2, i.e., use (15) as a measure of departure between $\bar{\theta}$ and $\bar{\theta}^0$ and – if data are Gaussian – read out threshold ρ from χ^2 distribution with K degrees of freedom. For non-Gaussian data establish ρ as in (19), replacing $\hat{\phi}(\bar{\theta}_n)$ by $\phi(\bar{\theta}_n)$.

- Σ *is unknown* but empirical covariance matrix $\hat{\Sigma}$ can be obtained by learning from historical data of a small or moderate length. In this case we have to appropriately combine Algorithms 1, 2 and 3, as it is explained in this subsection (see also Figure 1.

Meta-algorithm (for correlated descriptors)

Stage I Apply the Learning phase of Algorithm 1, using Algorithm 2 in its Step 2.

Stage II Apply Algorithm 3, replacing the condition $||\hat{\bar{\theta}} - \bar{\theta}^0||^2_K > \rho$ in its Step 4 by the following one: $\hat{\phi}(\bar{\theta}) > \rho$.

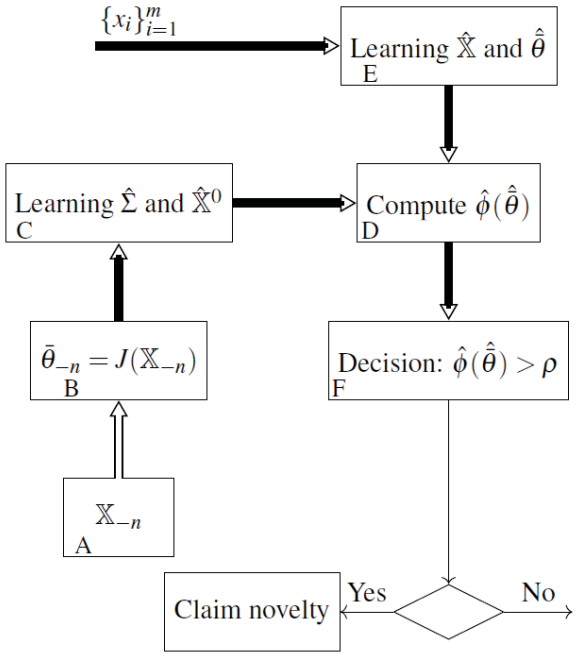

Figure 1. Flow chart of learning and decision making when descriptors are correlated. A) data base of historical curves or surrogate data, B) extraction of curves' coefficients (apply Algorithm 2 when observations are noisy), C) Learning $\hat{\bar{\theta}}^0$ (also $\hat{\mathbb{X}}^0$) and $\hat{\Sigma}^{-1}$ – Algorithm 1, D) Compute decision function $\hat{\phi}$, E) Learning current $\hat{\bar{\theta}}$ (Algorithm 2), F) Classifying current $\hat{\bar{\theta}}$. $\{x_i\}_{i=1}^m$ stands for the acquisition of observations from a new curve to be classified.

Observe that the learning stage (Stage I, see also blocks A), B), C) in Figure 1) is executed only once on historical or surrogate data. One may also consider a version of the Meta-algorithm with learning permanently, also from currently incoming curves, if we are convinced that they are correctly classified, but this is outside the scope of this paper.

On the other hand, its vertical branch (blocks E), D), F)) is executed each time when a new curve is to be classified. Arrows in Figure 1 are marked as white, if they transmit a curve as a whole, while black arrows convey data of the vector type. Thin arrows serve to pass decisions between blocks.

6 Testing on synthetic curves

In this section, our methodology of testing the algorithms and summary of the results are provided. Their aim is not only to test the algorithms, but also to check their robustness against assumption violation. In particular, the robustness of the choice of threshold ρ is of special importance. We start from the case when the components of $\bar{\theta}$ are uncorrelated and then modifications that are necessary to cover also more general case are described.

6.1 Testing – uncorrelated descriptors case

Before running the testing procedure, we have to select several important ingredients, namely,

1. K – the number of descriptors (dimension of $\bar{\theta}$), describing curves from the class "0",
2. the orthonormal basis spanning curves (here, the cosine series),
3. τ – the observation horizon and $m > K$ – the number of equidistant samples in $[0, \tau]$,
4. $0 < \alpha < 1$ – the significance level – an admissible level of rejecting a member of the class "0",
5. $\rho > 0$ – the decision threshold, initially read out as $(1-\alpha)$ quantile of χ^2 distribution with K degrees of freedom and possibly later corrected,
6. $\bar{\theta}^0$ – $K \times 1$ vector – a center of the class "0" and a probability distribution of $\bar{\varepsilon} \in \mathcal{R}^K$ with zero mean, uncorrelated components and variances $\sigma_k^2 > 0$, $k = 1, 2, \ldots, K$ that specify
$$\bar{\theta} = \bar{\theta}^0 + \bar{\varepsilon}, \tag{44}$$
i.e., coefficients of curves belonging to the class "0",
7. select a probability distribution of random errors ζ_i's in
$$x_i = \mathbb{X}(t_i) + \zeta_i, \qquad i = 1, 2, \ldots, m, \tag{45}$$
according to Assumption h4), where $\mathbb{X}$ is a curve with coefficients generated according to (44),
8. choose $N_T > 1$ as the number of repetitions of simulations.

Simulation process I

a) Set: simulation number $ns = 1$ and the number of improper classifications $n_{imp} = 0$. Prepare two $m \times 1$ vectors for intermediate data.

b) Generate $\bar{\theta}$ according to (44) and extend it to $m \times 1$ vector by padding $m - K$ zeros after $\bar{\theta}$. Store the result in $\bar{\theta}_{ext}$ vector.

c) Generate $\mathbb{X}(t_i)$'s by applying the inverse of the fast DCT to $\bar{\theta}_{ext}$ vector and use the resulting sequence in (45) to simulate observations x_i, $i = 1, 2, \ldots, m$ and store them as $\bar{x}_{sym}$.

d) Feed $\bar{x}_{sym}$ as the input of Algorithm 2 and compare its output $\hat{\bar{\theta}}$ with $\bar{\theta}^0$, by checking the condition:
$$||\hat{\bar{\theta}} - \bar{\theta}^0||^2 > \rho. \tag{46}$$
If this condition holds, set $n_{imp} = n_{imp} + 1$ and go to e). Otherwise, go directly to e).

e) If $ns < N_T$, set $ns = ns + 1$ and go to b). Otherwise, STOP – provide n_{imp} as the output.

Several remarks are in order concerning the above simulation methodology.

1. If n_{imp}/N_T is essentially larger than α, consider the reduction of ρ and repeat the Simulation process I.
2. If n_{imp}/N_T is much smaller than α, increase ρ slightly and repeat the Simulation process I.
3. To asses the robustness of the classifier, run the whole Simulation process I many times, changing the variances σ_k^2's, ζ_i's, the number of samples m etc.
4. For diagnostic and illustrative purposes, it may be useful to store the sequences $\hat{x}_i$'s that are generated as the second output of Algorithm 2.
5. Choose N_T relatively large (e.g., 10^3 or even 10^4), since the experience gained so far (see

[33], [48]) indicates that it is necessary for reducing a large variability of results when simulation experiments are used for testing methods that are based on thresholding.

6. In order to check discriminative abilities of the classifier, run again (many times) the Simulation process I, but this with the curves that are "far" from the class "0". For example, select a "false" $\bar{\theta}^0$, which is far from the original one. This time, interpret the result in n_{imp} as correct decisions.

6.2 Testing – correlated descriptors

Testing for correlated descriptors goes along similar lines as above with several changes only. The main one is in the presence of the correlation learning phase that is organized as follows.

Simulating the correlation learning phase

Simulating historical data Select $K \times K$ non-singular matrix B and use it for generating descriptors[7] as follows for $n = -1, -2, \ldots, -N_H$

$$\bar{\theta}_n = \bar{\theta}^0 + B\bar{\varepsilon}_n + \bar{\xi}_n, \tag{47}$$

where $\bar{\varepsilon}_n$'s are zero mean random vectors with the unit covariance matrix. The third summand in (47) is $K \times 1$ random vector with zero mean and uncorrelated components. These components have the same variances that are equal to the right hand side of (28). Their role is to approximately[8] incorporate errors introduced by observations corrupted by ζ_i's into the simulation process.

Simulated learning of Σ Denote by $\hat{\Sigma}(n)$ an estimate of Σ obtained from n historical observations. Set $\hat{\Sigma}(0)$ to be $K \times K$ matrix of zeros. For $n = 1, 2, \ldots, N_H$ run learning as follows

$$\hat{\Sigma}(n) = \nu_n \hat{\Sigma}(n-1) + \frac{1}{n}(\bar{\theta}_{-n} - \bar{\theta}^0)(\bar{\theta}_{-n} - \bar{\theta}^0)^T, \tag{48}$$

where $\nu_n \stackrel{def}{=} \frac{n-1}{n}$.

Checking Set $\hat{\Sigma} = \hat{\Sigma}(N_H)$ and check non-singularity[9] of this matrix. If not, increase N_H in (47). For diagnostic purposes one may verify how far is $\hat{\Sigma}$ from BB^T, e.g., in the Frobenius norm.

Forming a decision function $\hat{\phi}$ Compute the inverse of $\hat{\Sigma}$ and prepare

$$\hat{\phi}(\bar{\theta}) = (\bar{\theta} - \bar{\theta}^0)^T \hat{\Sigma}^{-1} (\bar{\theta} - \bar{\theta}^0). \tag{49}$$

In (48) and in (49) one may use $\hat{\bar{\theta}}^0$ estimated from historical data, but then one has to reduce the number of degrees of freedom when establishing the distribution of $\hat{\phi}(\bar{\theta})$.

Simulation process II

After finishing the above learning phase, one can run simulations aiming at testing the Meta-algorithm for its ability to correctly classify newly incoming $\bar{\theta}$'s (see also path: E), D), F) in Fig. 1). To this end it suffices to run the Simulation process I with the following changes.

1. Replace Step b) by:
   Generate $\bar{\theta} = \bar{\theta}^0 + B\bar{\varepsilon}$ and extend it to $m \times 1$ vector by padding $m - K$ zeros after $\bar{\theta}$. Store the result in $\bar{\theta}_{ext}$ vector.
2. In Step d) replace condition (46) by $\hat{\phi}(\bar{\theta}) > \rho$, where $\hat{\phi}(\bar{\theta})$ is as defined in (49).

The results of testing are summarized in the next subsection.

6.3 Summary of the test results

Firstly, the testing of synthetic, uncorrelated data was performed. The simulation parameters are presented in Table 1.

Firstly, the synthetic curves based on $\sqrt{t}$ were created. The results for different amounts of disturbance in θ are presented in Figure 2. Clearly, the results are correct (close to 0) until a certain level of disturbance is achieved. Only then the number of wrong classifications – both tested curves belong to the same class – increases. At this point the disturbed curve is very different then.

[7] When variability of descriptors is generated as in (47), we know that the covariance matrix of $\bar{\theta}_n$'s is $\Sigma = BB^T$, but we shall behave like a person with a split personality, i.e., B is given for generating data, but later we forget it and consider Σ as unknown.

[8] Simulations of exact errors would be computationally demanding. To this end, one has to repeat steps c) and d) in the Simulation process I for each $\bar{\theta}^0 + B\bar{\varepsilon}_n$, $n = -1, -2, \ldots, -N_H$.

[9] It is desirable to perform this checking by verifying whether the smallest eigenvalue of $\hat{\Sigma}$ is sufficiently larger than zero.

Table 1. Parameters for numerical tests

K	16
τ	1.
m	32
α	0.05
N_t	1000
σ_k^2	0.58
$Var\,\zeta$	0.8

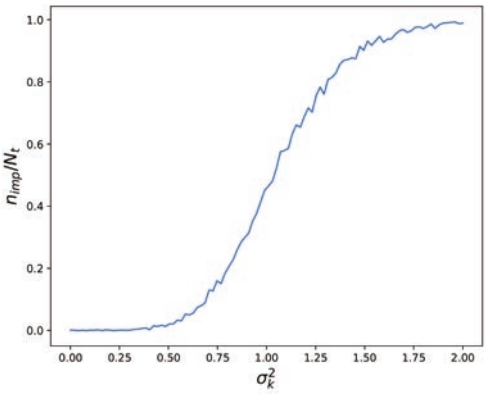

Figure 2. Uncorrelated descriptors, simulation process

When correlated descriptors are used, we can not visualise the change of B easily. Therefore, in Figure 3, on the horizontal axis, we can see the variance of ζ instead.

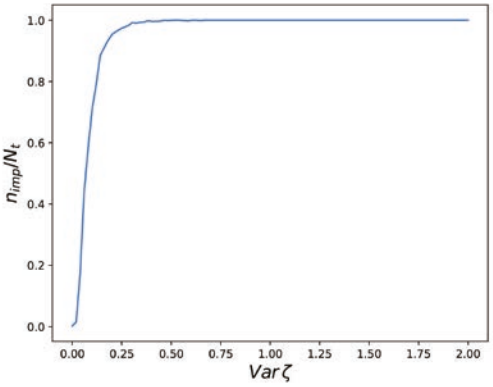

Figure 3. Correlated descriptors, simulation process

Similarly we can check how the algorithm can discriminate between two similar curves. First of them is already used

$$\sqrt{t}, \quad t \in [0,1]. \tag{50}$$

As a second one, the

$$1+\log_{10} t, \quad t \in [0,1], \tag{51}$$

was used. The comparison can be seen in Figure 4.

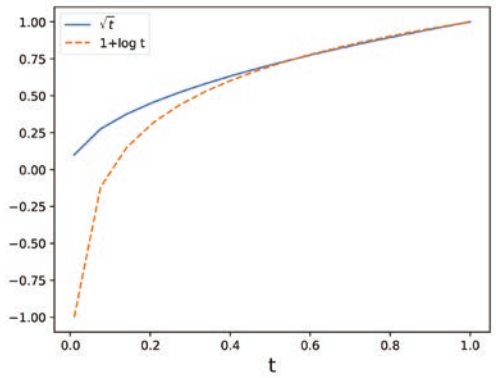

Figure 4. The comparison of valuse of function $\sqrt{t}$ and $1+\log t$

Obviously, when comparing dissimilar curves, the correct result would be the high number of counts in n_{imp}.

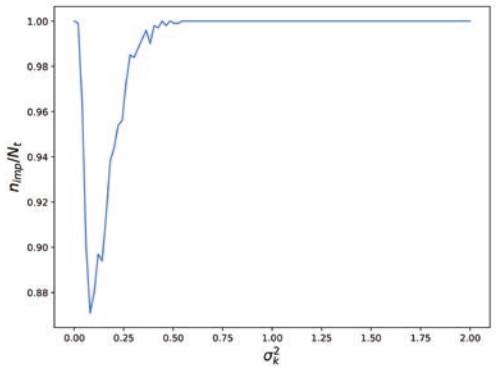

Figure 5. Uncorrelated descriptors, simulation process

Indeed that happens. In both cases (correlated, see Figure 6 and uncorrelated in Figure 5) we can clearly see that, when randomness is added, the number of correct classifications drops slightly (note scale on the vertical axis). Afterward, when the amount of disturbance is very high, the two curves are dissimilar enough to be classified as different.

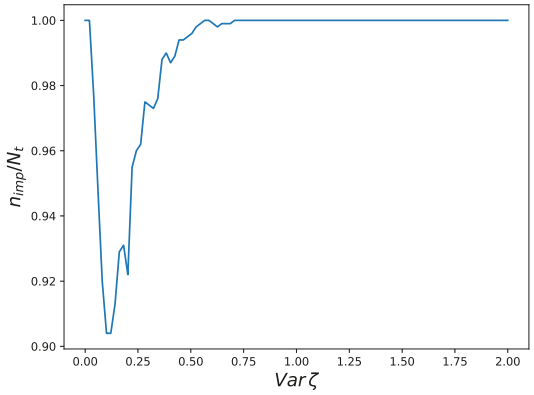

Figure 6. Correlated descriptors, simulation process

Overall, we can say that on the synthetic data the proposed algorithm works correctly.

7 Case study - COVID-19 growth rates

The COVID-19 is a global pandemic that nearly instantly changed the world. Its occurrence has also become a focus for many researchers trying to analyse and/or predict how the pandemic would unfold. General research had centered on predictions and how the current mitigation strategy would work like in [6].

In this section, we would like to use the novelty detection algorithm to investigate whether two so called waves of COVID are similar or different. Distinguish those spikes in COVID activity can help in deciding if a new mutation is present in considerable numbers and in assessing current mitigation strategy.

As a base data the number of COVID-19 cases in Croatia from January 1st 2020 to September 29th 2020 was chosen. The data were retrieved from ECDC COVID-19 dataset [19].

In Figure 7 the data in a relevant time period are shown. The first of the horizontal lines delimits the first wave of infections. The extent was determined by looking at the data. The dates in the said figure are for the beginning and the end of this observations. We can also distinguish the second possible wave and those observations are marked with second line and also annotated with relevant dates.

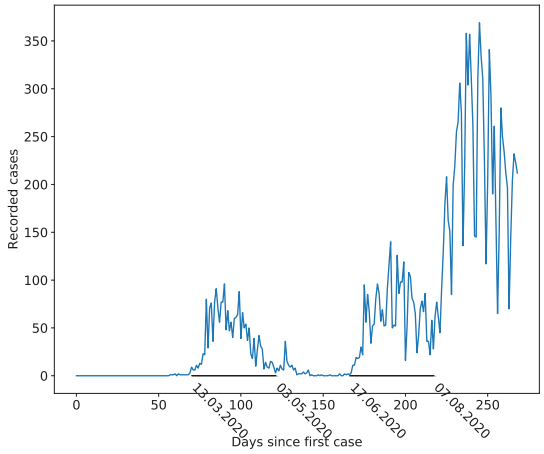

Figure 7. Data for COVID 19 cases in Croatia. Black lines indicates where samples were taken. The dates corresponds to the beginning and to the end of each sample respectively.

Even if we can see the shape as similar, the values are different. Since we are looking for similarities in shape, they have to be normalized to $[0, 1]$. The result for those observation groups is presented in Figure 8.

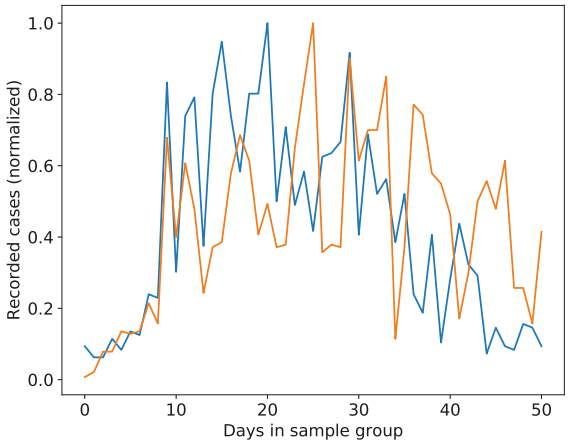

Figure 8. Normalized data in two stretches of time

The Algorithm was executed as in Section 5.1 using the first $m = 10$ terms. In order to verify the initial assessment of the second sample, the moving window was used. As a result, we obtain Figure 9. We can clearly see that the proposed second wave is where we have initially estimated. Please note that the smaller value around 70 is due to using the first wave as a reference and, obviously the data fits itself much better.

In this figure we can also see some other local minima. This had shown that we can use the pro-

posed algorithm also to find similarities in the data as well as to verify them.

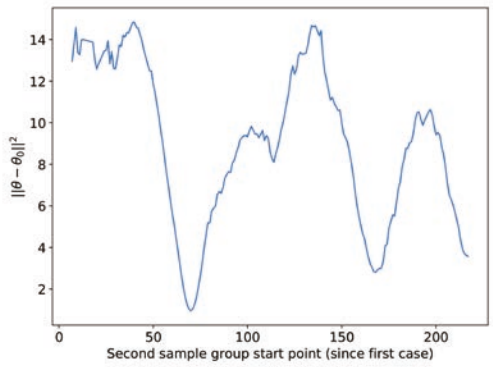

Figure 9. Results for moving windows

8 Concluding remarks

The aggregated learning algorithm for one-class classification of functional data, having the form of repetitively occurring random curves, has been proposed. It gathers off-line learning of the covariance matrix of curves' descriptors and pass-to-pass, on-line learning of descriptors of a current curve to be classified. The bounds for the probability of errors of the both learning processes have been derived. The descriptors are projections of a curve onto a finite dimensional subset of a selected orthogonal set of functions. Up to implementation details, the derivations are kept at the level of finite dimensional subsets of a Hilbert space and – therefore – can be formally generalized to, e.g., surfaces in two dimensions.

At the implementation level, curves can be provided as functions or represented as SRV, which is shape sensitive. A fast version of the on-line branch of the method is proposed using the fast version of the discrete cosine transform.

References

[1] C. Abraham, G. Biau, and B. Cadre, On the kernel rule for function classification, Annals of the Institute of Statistical Mathematics, 58(May 2005): 619–633, 2006.

[2] TW. Anderson, The Statistical Analysis of Time Series, Wiley Online Library, 1971.

[3] G. Aneiros, E. Bongiorno, R. Cao, P. Vieu, et al, Functional statistics and related fields. Springer, Cham 2017.

[4] G. Biau, F. Bunea, and M. Wegkamp, Functional classification in hilbert spaces. IEEE Transactions on Information Theory, 51(6): 2163–2172, 2005.

[5] P. Bickel and K. Doksum, Mathematical statistics: basic ideas and selected topics, volume I, volume 117. CRC Press,Boca Raton 2015.

[6] W. Bock, B. Adamik, M. Bawiec, V. Bezborodov, M. Bodych, J. Burgard, T. Goetz, T. Krueger, A. Migalska, B.a Pabjan, T. Ożański, E. Rafajłowicz, W. Rafajłowicz, E. Skubalska-Rafajłowicz, S. Ryfczyńska, E. Szczureki, and P. Szymański, Mitigation and herd immunity strategy for COVID-19 is likely to fail, medRxiv, 2020.

[7] V. Britanak, P. Yip, and K. Rao, Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations, Elsevier, Oxford, 2010.

[8] D. Clifton, S. Hugueny, and L. Tarassenko, A comparison of approaches to multivariate extreme value theory for novelty detection, In: IEEE Workshop on Statistical Signal Processing Proceedings, pages 13–16, 2009.

[9] A. Cuevas, A partial overview of the theory of statistics with functional data, Journal of Statistical Planning and Inference, 147: 1–23, 2014.

[10] A. Cuevas, M. Febrero, and R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Computational Statistics, 22(3): 481–496, 2007.

[11] L. Devroye, L. Gyorfi, and G. Lugosi, A probabilistic theory of pattern recognition, volume 31. Springer Science & Business Media, New York 2013.

[12] L. Devroye and G. Lugosi, Almost sure classification of densities, Journal of Nonparametric Statistics, 14(6): 675–698, 2002.

[13] P. Duda, K. Przybyszewski, and L. Wang, A novel drift detection algorithm based on features' importance analysis in a data streams environment, Journal of Artificial Intelligence and Soft Computing Research, 10(4): 287–298, 2020.

[14] P. Duda, L. Rutkowski, M. Jaworski, and D. Rutkowska, On the parzen kernel-based probability density function learning procedures over time-varying streaming data with applications to pattern classification, IEEE Transactions on Cybernetics, 50(4), 2018.

[15] P. Duda, L. Rutkowski, M. Jaworski, and D. Rutkowska, On the Parzen Kernel-Based Probability Density Function Learning Procedures Over Time-Varying Streaming Data With Applications to

Pattern Classification, IEEE Trans. on Cybernetics, 50(4): 1683–1696, 2020.

[16] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant, A general lower bound on the number of examples needed for learning, Information and Computation, 82: 247–261, 1989.

[17] F. Ferraty and P. Vieu, Nonparametric functional data analysis: theory and practice, Springer Science & Business Media, New York 2006.

[18] Ralph Foorthuis, On the nature and types of anomalies: A review, arXiv preprint arXiv: 2007.15634, 2020.

[19] European Centre for Disease Prevention and Control, Data on the geographic distribution of covid-19 cases worldwide.

[20] P. Galeano, J. Esdras, and R. Lillo, The mahalanobis distance for functional data with applications to classification, Technometrics, 57(2): 281–291, 2015.

[21] T. Gałkowski, A. Krzyżak, and Z. Filutowicz, A new approach to detection of changes in multidimensional patterns, Journal of Artificial Intelligence and Soft Computing Research, 10(2):125–136, 2020.

[22] T. Galkowski and L. Rutkowski, Nonparametric Fitting of Multivariate Functions, IEEE Transactions on Automatic Control, 31(8): 785–787, 1986.

[23] F. Gouin, C. Ancourt, and C. Guettier, Threewise: A local variance algorithm for GPU, Proceedings - 19th IEEE International Conference on Computational Science and Engineering, 14th IEEE International Conference on Embedded and Ubiquitous Computing and 15th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, CSE-EUC-DCABES 2016, pages 257–262, 2017.

[24] W. Greblicki, Pattern recognition procedures with nonparametric density estimates, IEEE Transactions on Systems, Man and Cybernetics, 8: 809–812, 1978.

[25] W. Greblicki and M. Pawlak, Classification using the Fourier series estimate of multivariate density functions, IEEE Transactions on Systems, Man and Cybernetics, 11: 726–730, 1981.

[26] W. Greblicki and M. Pawlak, Fourier and {H}ermite series estimates of regression functions, Ann. Inst. Stat. Math., 37: 443–454, 1985.

[27] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, New York, 2002.

[28] T. Harris, D. Tucker, B. Li, and L. Shand, Elastic depths for detecting shape anomalies in functional data, Technometrics, pages 1–11, 2020.

[29] D. Haussler, M. Kearns, H Sebastian Seung, and N. Tishby, Rigorous learning curve bounds from statistical mechanics, Machine Learning, 25(2-3): 195–236, 1996.

[30] W. Homenda, A. Jastrzębska, W. Pedrycz, and F. Yu, Combining classifiers for foreign pattern rejection, Journal of Artificial Intelligence and Soft Computing Research, 10(2): 75–94, 2020.

[31] L. Horváth and P. Kokoszka, Inference for functional data with applications, volume 200, Springer Science & Business Media, 2012.

[32] J. Jurečková and J. Kalina, Nonparametric multivariate rank tests and their unbiasedness, Bernoulli, 18(1): 229–251, 2012.

[33] W C M Kallenberg, T Ledwina, and E Rafajlowicz, Testing bivariate independence and normality, Sankhya: The Indian Journal of Statistics, Series A, 59(1): 42–59, 1997.

[34] M. Kemmler, E. Rodner, E. Wacker, and J. Denzler, One-class classification with Gaussian processes, Pattern Recognition, 46(12): 3507–3518, 2013.

[35] J. T. Kwok, I. W. Tsang, and J. M Zurada, A class of single-class minimax probability machines for novelty detection, IEEE Transactions on Neural Networks, 18(3): 778–785, 2007.

[36] N. Ling and P. Vieu, Nonparametric modelling for functional data: selected survey and tracks for future, Statistics, 52(4): 934–949, 2018.

[37] M. Markou and S. Singh, Novelty detection: A review - Part 2:: Neural network based approaches, Signal Processing, 83(12): 2499–2521, 2003.

[38] M. Markou and S. Singh, Novelty detection: a review—part 1: statistical approaches, Signal Processing, 83(12): 2481–2497, 2003.

[39] J. Marron, J. Ramsay, L. Sangalli, and A. Srivastava, Functional data analysis of amplitude and phase variation, Statistical Science, 30(4): 468–484, 2015.

[40] J. S. Marron, J. Ramsay, L. Sangalli, and A. Srivastava, Functional data analysis of amplitude and phase variation, Statistical Science, 30(4): 468–484, 2015.

[41] D. Montgomery, Introduction to statistical quality control, John Wiley & Sons New York, 2009.

[42] H-G Mueller et al, Peter Hall, functional data analysis and random objects, The Annals of Statistics, 44(5): 1867–1887, 2016.

[43] K. Patan, M. Witczak, and J. Korbicz, Towards robustness in neural network based fault diagnosis, International Journal of Applied Mathematics and Computer Science, 18(4): 443–454, 2008.

[44] S. Perera and J. Liu, Complexity reduction, self/completely recursive, radix-2 dct i/iv algorithms, Journal of Computational and Applied Mathematics, 379: 112936, 2020.

[45] E. Rafajłowicz and Schwabe R, Halton and Hammersley sequences in multivariate nonparametric regression, Statistics and Probability Letters, 76(8): 803–812, 2006.

[46] E. Rafajłowicz and R. Schwabe, Equidistributed designs in nonparametric regression, Statistica Sinica, 13(1), 2003.

[47] E. Rafajłowicz and E. Skubalska-Rafajłowicz, FFT in calculating nonparametric regression estimate based on trigonometric series, Journal of Applied Mathematics and Computer and Computer Science, 3(4): 713–720, 1993.

[48] E. Rafajłowicz and A. Steland, A binary control chart to detect small jumps, Statistics, 43(3): 295–311, 2009.

[49] E. Rafajłowicz and A. Steland, The Hotelling—Like T2 Control Chart Modified for Detecting Changes in Images having the Matrix Normal Distribution, In Springer Proceedings in Mathematics and Statistics, volume 294, pages 193–206, 2019.

[50] E. Rafajłowicz, Nonparametric orthogonal series estimators of regression: a class attaining the optimal convergence rate in L2, Statistics and Probability Letters, 5: 219–224, 1987.

[51] J. Ramsay and B. Silverman, Applied functional data analysis: methods and case studies, Springer, 2007.

[52] D. Rutkowska and L. Rutkowski, On the Hermite series-based generalized regression neural networks for stream data mining, In: International Conference on Neural Information Processing, pages 437–448. Springer, 2019.

[53] L. Rutkowski, A general approach for nonparametric fitting of functions and their derivatives with applications to linear circuits identification, IEEE Transactions on Circuits and Systems, 33(8): 812–818, 1986.

[54] L. Rutkowski, M. Jaworski, and P. Duda, Stream data mining: algorithms and their probabilistic properties, Springer, Cham, 2020.

[55] L. Rutkowski and E. Rafajłowicz, On optimal global rate of convergence of some nonparametric identification procedures, IEEE Trans. Automatic Control, AC-34: 1089–1091, 1989.

[56] S. Sameer and M. Markou, An approach to novelty detection applied to the classification of image regions, IEEE Transactions on Knowledge and Data Engineering, 16(4): 396–407, 2004.

[57] R. Serfling. Approximation theorems of mathematical statistics, volume 162. John Wiley & Sons, New York 2009.

[58] E. Skubalska-Rafajłowicz, One-dimensional Kohonen's Lvq nets for multidimensional patterns recognition, International Journal of Applied Mathematics and Computer Science, 10(4): 767–778, 2000.

[59] E. Skubalska-Rafajlowicz, Pattern recognition algorithms based on space-filling curves and orthogonal expansions, IEEE Transactions on Information Theory, 47(5): 1915–1927, 2001.

[60] E. Skubalska-Rafajłowicz, Random projection RBF nets for multidimensional density estimation, International Journal of Applied Mathematics and Computer Science, 18(4): 455–464, 2008.

[61] E. Skubalska-Rafajlowicz and A. Krzyzak, Fast k-NN classification rule using metric on space-filling curves, In Proceedings of 13th International Conference on Pattern Recognition, volume 2, pages 121–125. IEEE, 1996.

[62] A. Srivastava and E. Klassen, Functional and shape data analysis, volume 1, Springer, Cham, 2016.

[63] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, Shape Analysis of Elastic Curves in Euclidean Spaces, IEEE Journal on Selected Areas in Communications, 10(2): 391–400, 1992.

[64] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, Shape analysis of elastic curves in Euclidean spaces, IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(7): 1415–1428, 2010.

[65] A. Steland and R. von Sachs, Asymptotics for high-dimensional covariance matrices and quadratic forms with applications to the trace functional and shrinkage, Stochastic Process. Appl., 128(8): 2816–2855, 2018.

[66] L. Tarassenko, A. Nairac, N. Townsend, I. Buxton, and P. Cowley, Novelty detection for the identification of abnormalities, International Journal of Systems Science, 31(11): 1427–1439, 2000.

[67] B. Trawiński, M. Smętek, Z. Telec, and T. Lasota, Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms, International Journal of Applied Mathematics and Computer Science, 22(4): 867–881, 2012.

[68] M. Vidyasagar, A theory of learning and generalization, Springer-Verlag, Berlin, 2002.

[69] G. Vinue and I. Epifanio, Robust archetypoids for anomaly detection in big functional data, Advances in Data Analysis and Classification, pages 1–26, 2020.

[70] J. Wang, J. Chiou, and H-G Mueller, Review of Functional Data Analysis, pages 1–41, 2015.

[71] W. Xie, O. Chkrebtii, and S. Kurtek, Visualization and Outlier Detection for Multivariate Elastic Curve Data, IEEE Transactions on Visualization and Computer Graphics, 26(11): 3353–3364, 2020.

[72] Y. Yang and T. Mathew, The simultaneous assessment of normality and homoscedasticity in one-way random effects models, Statistics and Applications (ISSN 2452-7395(online)), 18(2): 97–119, 2020.

[73] M. Yao and H. Wang, One-Class Support Vector Machine for Functional Data Novelty Detection, In 2012 Third Global Congress on Intelligent Systems, pages 172–175. IEEE, 2012.

[74] M. Yao and H. Wang, One-class support vector machine for functional data novelty detection, In: Proceedings - 2012 3rd Global Congress on Intelligent Systems, GCIS 2012, number 1, pages 172–175. IEEE, 2012.

Wojciech Rafajłowicz received the M.Sc. degree from Wroclaw University of Science and Technology in 2011 and the Ph.D. degree with distinctions from the University of Zielona Góra in 2016. Currently is an assistant professor at the Department of Control Systems and Mechatronics, Wroclaw University of Science and Technology. His research concentrates mostly on computational aspects of control systems and decision sequences for different types of processes. His additional interests are in the application of image processing and embedded systems. He has published 13 research papers and was an author or editor in 4 books and an author or co-author in 36 book chapters or papers in conference proceedings.