Signal Processing 168 (2020) 107320

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

# Anomaly detection in mixed telemetry data using a sparse representation and dictionary learning

Barbara Pilastre[a,*], Loïc Boussouf[b], Stéphane D'Escrivan[d], Jean-Yves Tourneret[c,a]

[a] TéSA laboratory, 7 boulevard de la Gare, Toulouse, 31500, France
[b] Airbus Defense and Space, 31 rue des Cosmonautes, Toulouse 31400, France
[c] INP-ENSEEIHT/IRIT, 2 rue Charles Camichel, Toulouse, 31071, France
[d] Centre National d'Etudes Spatiales (CNES), 18 av Edouard Belin, Toulouse CEDEX 9 31401, France

ARTICLE INFO

*Article history:*
Received 26 April 2019
Revised 9 August 2019
Accepted 24 September 2019
Available online 24 September 2019

*Keywords:*
Spacecraft health monitoring
Anomaly detection
Sparse representation
Dictionary learning
Shift-Invariance

ABSTRACT

Spacecraft health monitoring and failure prevention are major issues in space operations. In recent years, machine learning techniques have received an increasing interest in many fields and have been applied to housekeeping telemetry data via semi-supervised learning. The idea is to use past telemetry describing normal spacecraft behaviour in order to learn a reference model to which can be compared most recent data in order to detect potential anomalies. This paper introduces a new machine learning method for anomaly detection in telemetry time series based on a sparse representation and dictionary learning. The main advantage of the proposed method is the possibility to handle multivariate telemetry time series described by mixed continuous and discrete parameters, taking into account the potential correlations between these parameters. The proposed method is evaluated on a representative anomaly dataset obtained from real satellite telemetry with an available ground-truth and compared to state-of-the-art algorithms.



## 1. Introduction

Spacecraft health monitoring and failure prevention are major issues in space operations. By monitoring housekeeping telemetry data, an anomaly affecting an equipment, a system or a sub-system can be detected from the abnormal behaviour of one or several telemetry parameters. A simple method for detecting anomalies in telemetry is the well-known out-of-limits (OOL) checking, which consists of defining an upper and a lower bound for each parameter and checking whether the values of this parameter exceed these bounds. This method is very simple and useful but has also some limits. Indeed, the determination of bounds for each parameter can be difficult and costly given the number of spacecraft sensors. Moreover, all anomalies are not detected by the OOL checking, e.g, when the parameter affected by an anomaly does not exceed the predefined bounds. An example of anomaly not detected by OOL checking is displayed in Fig. 1 (box #2).

Anomaly detection (AD) is a huge area of research given its diverse applications. Recent years have witnessed a growing interest for data-driven or machine learning (ML) techniques that have been used as effective tool for AD [1–5]. Motivated by this success, some ML methods have been applied to housekeeping telemetry after an appropriate preprocessing step [6–10]. These methods usually consider a semi-supervised learning that can be outlined in two steps: 1) learning from past telemetry describing only nominal spacecraft events and 2) detecting abnormal behaviour in the different parameters by an appropriate comparison to the model learned in step 1).

ML-based algorithms for AD in telemetry can be divided in two categories depending on their application to univariate or multivariate data. Univariate AD strategies process the different telemetry parameters independently, which is the most widely used approach. Popular ML methods that have been investigated in this framework include the one-class support vector machine [7], nearest neighbour techniques [8–10] or neural networks [11,12]. These solutions showed competitive results and improved significantly spacecraft heath monitoring. However, in order to improve AD in telemetry, it is important to formulate the problem in a multivariate framework and take into account possible correlations between the different parameters, allowing contextual anomalies to be detected. An example of contextual anomaly is shown in Fig. 1 (box #7). The detection of this kind of abnormal behaviour requires a multivariate detection rule. Some recent multivariate AD are based on feature extraction and dimensionality reduction [13] or on a probabilistic model for mixed discrete and continuous telemetry parameters [14].

This paper studies a new AD method based on a sparse data representation for spacecraft housekeeping telemetry. This method

* Corresponding author.
*E-mail address:* barbara.pilastre@tesa.prd.fr (B. Pilastre).

https://doi.org/10.1016/j.sigpro.2019.107320
0165-1684/

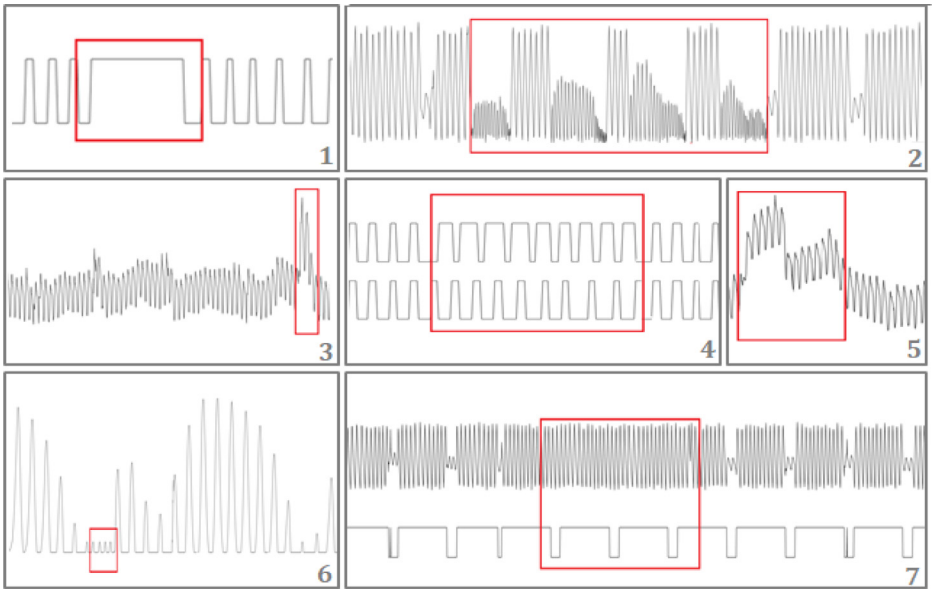


**Fig. 1.** Examples of univariate and multivariate anomalies (highlighted in red boxes) that are considered in this work. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is inspired by the works conducted in [15]. However, it has the advantage of handling mixed continuous and discrete data, and taking into account possible correlations between the different telemetry parameters thanks to an appropriate multivariate framework. The proposed algorithm requires to build a dictionary of normal patterns. New telemetry signals can then be decomposed into this dictionary using a sparse representation allowing potential anomalies to be detected by analyzing the residuals resulting from this sparse decomposition.

The paper is organized as follow. Section 2 introduces the context of AD in mixed telemetry data considered in this work. Section 3 briefly summarizes the theory of sparse representations and dictionary learning. Section 4 introduces the proposed AD method adapted to mixed continuous and discrete telemetry parameters. Section 5 evaluates the performance of the proposed method using a heterogeneous anomaly dataset with a controlled ground-truth. A comparison to other state-of-art techniques shows the potential of using a sparse representation on a dictionary of normal patterns for detecting abnormal behaviour in telemetry. Conclusions and future work are reported in Section 6.

## 2. Anomaly detection for telemetry

### 2.1. Characteristics of spacecraft telemetry

Spacecraft telemetry consists of hundred to thousand housekeeping parameters. All these parameters are quantized and take their values in a discrete set. However, it makes sense to make a distinction between parameters taking few distinct values (such as equipment operating modes or status), which can be considered as observations of discrete random variables, and parameters that can be considered as observations of continuous random variables (such as temperature, voltage, pressure etc.). Detecting anomalies in telemetry requires to consider these discrete and continuous random variables jointly leading to what we will call mixed vectors, in the sense that they contain discrete and continuous random variables.

In order to take into account relationships between the different parameters, it is necessary to learn their behaviour jointly, which requires to consider vectors belonging to a possibly high dimensional subspace. Considering high-dimensional data leads to major issues such as the well known curse of dimensionality [16,17]. Note also that this high-dimensionality has been considered in many recent works such as those devoted to big data [18,19].

Another important characteristics of telemetry data is that they are generally subjected to several preprocessings. As an example, it is quite classical to remove trivial outliers caused by errors in the data conversion and transmission using simple outlier detection methods [14]. Some telemetry parameters can have been resampled to account for the fact the data have been acquired at different sampling frequencies. In addition, some reconstruction methods may have been applied to compensate for missing data [14,20]. Finally, it is interesting to note that additional preprocessing is necessary for the learning phase to select telemetry which describes only usual normal behavior of the spacecraft. Indeed, behaviors representing rare operations, e.g., destocking or equipment calibration operations (abnormal in an other context) are not selected for learning.

### 2.2. Anomalies in telemetry

Anomalies that occur in housekeeping telemetry data can be divided in two categories that can be referred to as univariate and multivariate anomalies. Univariate anomalies correspond to an unusual individual behaviour (never seen before) affecting one specific parameter. Univariate anomalies can be classified in three main categories [1] summarized below

- Collective anomalies: a collection of consecutive data instances or time series considered as anomalous with respect to the entire signal. Two examples of collective anomalies are displayed in Fig. 1 (boxes #1 and #4).
- Point anomalies: an individual data instance considered as anomalous with respect to the rest of the data. A point anomaly is the easiest to detect because it corresponds to an excessive value of individual samples. It is not necessary to observe a collection of time samples to detect this kind of anomaly. Point anomalies can be generally detected by simple thresholding, e.g., using the OOL AD method. Two examples of consecutive point anomalies are displayed in Fig. 1 (boxes #3 and #5).

- Univariate contextual anomalies: an individual data instance or a time series considered as anomalous in a specific context, but not otherwise. Fig. 1 displays examples of contextual anomalies for consecutive data instances (box #2) or time series (box #6).

Note that collective anomalies and some individual contextual anomalies may not be detected if data instances are processed independently. The detection of these anomalies requires to consider collections of data instances or time series.

A multivariate or contextual anomaly results from a parameter whose behaviour has never been observed jointly with the behaviour of one or several other parameters recorded at the same time. Fig. 1 (boxes #4 and #7) shows examples of contextual anomalies. Note that the anomaly of Fig. 1 in box #7 is a multivariate contextual anomaly that affects a set of two related discrete and continuous parameters. Note also that the top signal is supposed to evolve differently depending on the status of an equipment (that can be ON or OFF) associated with the binary bottom parameter. In this example, the expected behaviour of the continuous parameter is not observed in the red box, which corresponds to a multivariate contextual anomaly. The detection of this kind of anomaly requires to work in a multivariate framework in order to learn the behaviour of multiple parameters. The objective of this work is to propose a flexible AD method able to detect univariate as well as multivariate anomalies affecting telemetry.

## 3. Sparse representations and dictionary learning

Sparse representations have received an increasing attention in many signal and image processing applications. These applications include denoising [21–23], classification [24–26] or pattern recognition [27–29]. The use of sparse representations for AD is more original and has been considered in less applications such as hyperspectral imaging [30], detection of abnormal motions in videos [31], irregular heartbeat detection in electrocardiograms (ECG) or specular reflectance and shadow removal in natural images [15]. The next part of this section recalls some basic elements about sparse representations and dictionary learning.

### 3.1. Sparse representations

Building a sparse representation (also referred to as *sparse coding*) consists in approximating a signal $\boldsymbol{y} \in \mathbb{R}^N$ as $\boldsymbol{y} \approx \boldsymbol{\Phi x}$, where $\boldsymbol{\Phi} \in \mathbb{R}^{N \times L}$ is an overcomplete dictionary composed of $L$ columns called *atoms*, and $\boldsymbol{x} \in \mathbb{R}^L$ is a sparse coefficient vector. In others words, the signal $\boldsymbol{y}$ is expressed as a sparse linear combination of few atoms of the dictionary $\boldsymbol{\Phi}$. Once the dictionary $\boldsymbol{\Phi}$ has been determined, the sparse representation problem reduces to estimate the sparse coefficient vector $\boldsymbol{x}$ by solving the following problem

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{x}\|_0 \leq T \tag{1}$$

where $\|.\|_0$ is the $\ell_0$ pseudo-norm which counts the number of non-zero entries of $\boldsymbol{x}$, $\|.\|_2$ is the $\ell_2$ norm, $T$ is the allowed number of non-zeros entries of $\boldsymbol{x}$ and "s.t." means "subject to".

Problem (1) is NP-hard and can be solved by greedy algorithms such as matching pursuit (MP) [32], orthogonal matching pursuit (OMP) [33], or by convex relaxation. Convex relaxation replaces the $\ell_0$ pseudo-norm by the $\ell_1$ norm defined by $\|\boldsymbol{x}\|_1 = \sum_{l=1}^{L} |x_l|$, leading to a convex problem whose solution can be computed using algorithms such as the least absolute shrinkage and selection operator (LASSO) [34].

### 3.2. Dictionary learning

The quality of the approximation $\boldsymbol{y} \approx \boldsymbol{\Phi x}$ strongly relies on the choice of the dictionary $\boldsymbol{\Phi}$. Dictionaries can be divided in two main classes corresponding to parametric and data-driven dictionaries. Parametric dictionaries are composed of fixed atoms such as wavelets, curvelets, contourlets or short-time Fourier transforms. Data-driven dictionaries learn the dictionary from the data, which has shown to be interesting in many practical applications [35]. This paper focuses on a semi-supervised framework in which the dictionary is learned from clean data which do not contain any anomaly. A classic way to learn a dictionary from the data is to use data analysis methods such as the well known principal component analysis (PCA). However, more efficient data-driven methods for dictionary learning (DL), often referred to as DL methods, have attracted many attention in recent years. These methods learn dictionaries tailored for sparse representations by solving the following problem

$$\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{x},\boldsymbol{\Phi}} \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_F^2 \ \text{ s.t. } \ \|\boldsymbol{x}\|_0 \leq T. \tag{2}$$

Classical DL algorithms alternate between the estimation of $\boldsymbol{x}$ (in a first step of sparse coding) and the estimation of $\boldsymbol{\Phi}$ (in a second step of dictionary update). Many efficient DL algorithms have been proposed in the literature including K-SVD [36] or online DL (ODL) [37]. In K-SVD, the sparse coding step is done using a greedy algorithm. In the second step, the dictionary and the sparse vector are estimated using a singular value decomposition (SVD), allowing the columns of $\boldsymbol{\Phi}$ as well as the associated coefficients of $\boldsymbol{x}$ to be updated. The ODL algorithm has been designed to learn dictionaries from large and dynamic datasets, using a sparse coding step performed by the LARS-LASSO algorithm [38,39] and a dictionary update using block-coordinate descent with warm restarts.

Unfortunately K-SVD and ODL algorithms have not been designed for mixed discrete and continuous data and thus cannot be used for telemetry data. Learning a dictionary with discrete and continuous atoms is an interesting and challenging problem. However, the dictionary can also be built from representative training signals that are not affected by anomalies. In this work, the dictionary has been built from "normal" vectors (that are not affected by anomalies) belonging to a training database. Extending standard DL methods (such as K-SVD and ODL) to mixed data will be considered in future work.

## 4. Proposed anomaly detection method for mixed telemetry

This section describes the proposed Anomaly Detection using DICTionary (ADDICT) algorithm which is an AD method for mixed data. We focus here on the detection step and assume that the dictionary has been learned in a previous step using telemetry signals associated with a normal spacecraft behaviour.

### 4.1. Preprocessing

Telemetry times series acquired at the same time instant and considered as part of the same context are first segmented into overlapping windows of fixed size $w$ with a shift $\delta$ (with an overlapping area equal to $w - \delta$) as illustrated in Fig. 2. The resulting matrices are then concatenated into vectors yielding mixed vectors whose components are discrete or continuous depending on the considered parameter. One concatenated vector thus represents a specific context containing information from both continuous and discrete signals on a duration $w$. Given this preprocessing, input data for AD are mixed signals composed of telemetry time series formed by the different parameters, i.e., $\boldsymbol{y} = [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T]^T$ with $\boldsymbol{y}_k \in \mathbb{R}^w, k = 1, \ldots, K$, where $K$ is the number of telemetry parameters and $w$ is the size of the time window. To simplify notations, the $N_D$ first components of the mixed signal $\boldsymbol{y} \in \mathbb{R}^N$ are composed of the discrete parameters whereas the last $N_C$ components are associated with the continuous times series (with $N = N_D + N_C$). In

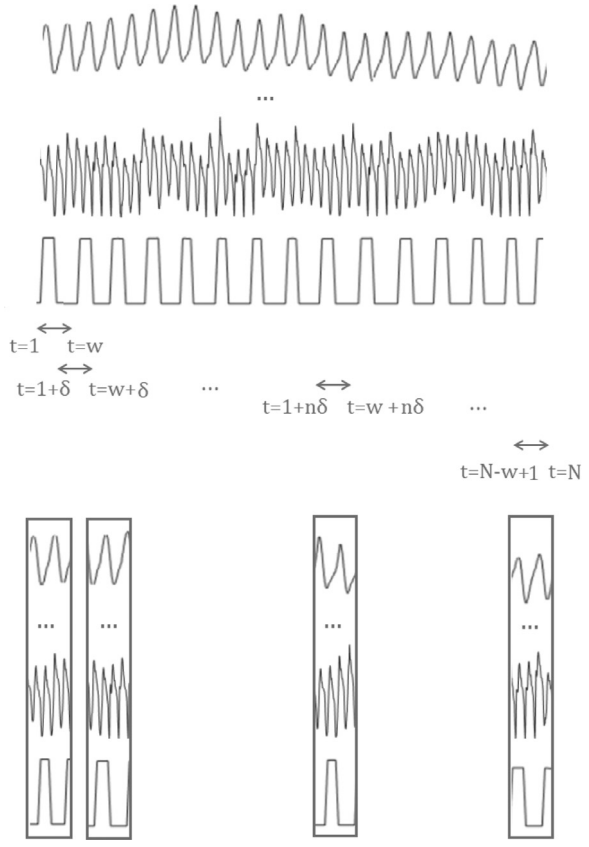


**Fig. 2.** Segmentation of telemetry into overlapping windows.

other words, the mixed signal is partitioned into discrete and continuous counterparts denoted as $\boldsymbol{y}_{\mathrm{D}} = [\boldsymbol{y}(1), \ldots, \boldsymbol{y}(N_{\mathrm{D}})]^T$ and $\boldsymbol{y}_{\mathrm{C}} = [\boldsymbol{y}(N_{\mathrm{D}}+1), \ldots, \boldsymbol{y}(N_{\mathrm{D}}+N_{\mathrm{C}})]^T$ such that $\boldsymbol{y} = (\boldsymbol{y}_{\mathrm{D}}^T, \boldsymbol{y}_{\mathrm{C}}^T)^T$.

### 4.2. Anomaly detection using a sparse representation

A mixed dictionary $\boldsymbol{\Phi} \in \mathbb{R}^{N \times 2L}$ composed of discrete and continuous atoms is defined as

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_{\mathrm{D}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{\mathrm{C}} \end{bmatrix}$$

where $\boldsymbol{\Phi}_{\mathrm{D}}$ and $\boldsymbol{\Phi}_{\mathrm{C}}$ contain the discrete and continuous dictionary atoms, respectively. The two dictionaries $\boldsymbol{\Phi}_{\mathrm{D}} \in \mathbb{R}^{N_{\mathrm{D}} \times L}$ and $\boldsymbol{\Phi}_{\mathrm{C}} \in \mathbb{R}^{N_{\mathrm{C}} \times L}$ have been extracted from a dictionary composed of $L$ mixed atoms taking into account possible correlation between the different parameters, especially between discrete and continuous ones. In other words, the $l$th discrete atom of the discrete dictionary $\boldsymbol{\Phi}_{\mathrm{D}}$ and the $l$th continuous atom of the continuous dictionary $\boldsymbol{\Phi}_{\mathrm{C}}$ are composed of discrete and continuous behaviours observed in the same mixed atom, which are potentially correlated The proposed AD strategy decomposes the mixed signal as follows

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e} + \boldsymbol{b} \tag{3}$$

where $\boldsymbol{\Phi}\boldsymbol{x}$ is the nominal part of $\boldsymbol{y}$, $\boldsymbol{e} = [\boldsymbol{e}_{\mathrm{D}}^T, \boldsymbol{e}_{\mathrm{C}}^T]^T$ is a possible anomaly signal ($\boldsymbol{e} = \mathbf{0}$ in absence of anomaly), $\boldsymbol{x} = [\boldsymbol{x}_{\mathrm{D}}^T, \boldsymbol{x}_{\mathrm{C}}^T]^T$ is a sparse vector and $\boldsymbol{b} \in \mathbb{R}^N$ is an additive noise. The proposed algorithm applies two distinct strategies to estimate the nominal part of $\boldsymbol{y}$ processing $\boldsymbol{x}_{\mathrm{D}}$ and $\boldsymbol{x}_{\mathrm{C}}$ differently.The anomaly signal $\boldsymbol{e}$ is estimated by analyzing residuals resulting from this sparse decomposition. The proposed detector assumes that the anomalies affecting telemetry data are additive, which is generally the case. Note that the proposed model (3) provides a specific structure of the residue $\boldsymbol{e}$, which allows its non zero values to be identified. These non-zero values correspond to the parameters affected by the anomalies.

The nominal component of $\boldsymbol{y}$ is approximated by linear combinations of atoms describing only nominal behaviours of the different parameters, which can be written as

$$\boldsymbol{\Phi}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\Phi}_{\mathrm{D}}\boldsymbol{x}_{\mathrm{D}} \\ \boldsymbol{\Phi}_{\mathrm{C}}\boldsymbol{x}_{\mathrm{C}} \end{bmatrix}.$$

with $\boldsymbol{x}_{\mathrm{D}} \in \mathbb{R}^{N^D}$ and $\boldsymbol{x}_{\mathrm{C}} \in \mathbb{R}^{N_{\mathrm{C}}}$ The discrete and continuous counterparts of the test signals will be approximated by two distinct strategies. However, it is important to preserve existing relationships between the signal parameters to allow for the detection of contextual anomalies. To this end, we propose to estimate the discrete approximation $\boldsymbol{\Phi}_{\mathrm{D}}\boldsymbol{x}_{\mathrm{D}}$ and the anomaly signal $\boldsymbol{e}_{\mathrm{D}}$ in a first step (leading to estimators denoted as $\widehat{\mathbf{e}}_{\mathrm{D}}$ and $\widehat{\mathbf{x}}_{\mathrm{D}}$) and the continuous approximation $\boldsymbol{\Phi}_{\mathrm{C}}\boldsymbol{x}_{\mathrm{C}}$ and the anomaly signal $\boldsymbol{e}_{\mathrm{C}}$ in a second step based on $\widehat{\mathbf{e}}_{\mathrm{D}}$ and $\widehat{\mathbf{x}}_{\mathrm{D}}$. Given the proposed preprocessing, the signals $\boldsymbol{e}_{\mathrm{D}}$ and $\boldsymbol{e}_{\mathrm{C}}$ are divided into $K_{\mathrm{D}}$ and $K_{\mathrm{C}}$ discrete and continuous parameters, i.e., $\boldsymbol{e}_{\mathrm{D}} = [\mathbf{e}_{\mathrm{D},1}^T, \ldots, \mathbf{e}_{\mathrm{D},\mathrm{K_D}}^T]^T$ and $\boldsymbol{e}_{\mathrm{C}} = [\mathbf{e}_{\mathrm{C},1}^T, \ldots, \mathbf{e}_{\mathrm{C},\mathrm{K_C}}^T]^T$.

#### 4.2.1. Sparse coding for discrete atoms

In order to solve the sparse coding for discrete atoms, we propose to solve the following problem

$$\arg \min_{\boldsymbol{x}_{\mathrm{D}} \in \mathcal{B}, \boldsymbol{e}_{\mathrm{D}} \in \mathbb{R}^{N_{\mathrm{D}}}} \| \boldsymbol{y}_{\mathrm{D}} - \boldsymbol{\Phi}_{\mathrm{D}}\boldsymbol{x}_{\mathrm{D}} - \boldsymbol{e}_{\mathrm{D}} \|_2^2 + b_{\mathrm{D}} \sum_{k=1}^{K_{\mathrm{D}}} \|\mathbf{e}_{\mathrm{D,k}}\|_2 \tag{4}$$

where $\|\mathbf{e}_{\mathrm{D},k}\|_2, k = 1, \ldots, K_{\mathrm{D}}$ is the Euclidean norm, $\mathbf{e}_{\mathrm{D,k}}$ corresponds to the $k$th time-series of $\mathbf{e}_{\mathrm{D}}$ associated with the $k$th parameter and $b_{\mathrm{D}}$ is a regularization parameter that controls the level of sparsity of $\boldsymbol{e}_{\mathrm{D}}$. The sparsity constraint for the anomaly signal reflects the fact that anomalies are rare and affect few parameters at the same time. Note that the discrete vector $\boldsymbol{x}_{\mathrm{D}}$ is constrained to belong to $\mathcal{B}$, where $\mathcal{B}$ is the canonical or natural basis of $\mathbb{R}^L$, i.e., $\mathcal{B} = \{\boldsymbol{\epsilon}_l, l = 1, \cdots, L\}$, where $\boldsymbol{\epsilon}_l$ is a vector whose $l$th component equals 1 and whose other components equal 0. In other words, only one atom of the discrete dictionary $\boldsymbol{\Phi}_{\mathrm{D}}$ is chosen to represent the discrete signal, this amounts to looking for the nearest neighbour of $\boldsymbol{y_D}$ in the dictionary. This strategy has proved to be an effective method to reconstruct discrete signals (compared to a representation using a linear combination of atoms), which explains this choice. Since $\boldsymbol{x}_{\mathrm{D}}$ belongs to a finite set, its estimation is combinatorial and can be solved for each atom $\boldsymbol{\phi}_{\mathrm{D},l}$ (where $\boldsymbol{\phi}_{\mathrm{D},l}$ is the $l$th column of $\boldsymbol{\Phi}_{\mathrm{D}}$) as follows

$$\widehat{\mathbf{e}}_{\mathrm{D,l}} = \arg \min_{\mathbf{e}_{\mathrm{D},l}} \|\boldsymbol{y}_{\mathrm{D}} - \boldsymbol{\phi}_{\mathrm{D},l} - \boldsymbol{e}_{\mathrm{D},l}\|_2^2 + b_{\mathrm{D}} \sum_{k=1}^{K_{\mathrm{D}}} \|\mathbf{e}_{\mathrm{D,k}}\|_2. \tag{5}$$

The solution of the optimization problem (5) is classically obtained using a shrinkage operator $\widehat{\mathbf{e}}_{\mathrm{D,l}} = T_{b_{\mathrm{D}}}(\mathbf{h})$, with $\mathbf{h} = \boldsymbol{y}_{\mathrm{D}} - \boldsymbol{\phi}_{\mathrm{D},l}$

$$\left[T_{b_{\mathrm{D}}}(\mathbf{h})\right]_{\mathrm{k}} = \begin{cases} \dfrac{\|\mathbf{h}_{\mathrm{k}}\|_2 - b_{\mathrm{D}}}{\|\mathbf{h}_{\mathrm{k}}\|_2}\mathbf{h}_{\mathrm{k}} & \text{if } \|\mathbf{h}_{\mathrm{k}}\|_2 > b_{\mathrm{D}} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{h}_k$ is the $k$th part of $\mathbf{h}$ associated with the $k$th parameter for $\mathrm{k} = 1, \ldots, K_{\mathrm{D}}$. All the atoms $\boldsymbol{\phi}_{\mathrm{D},l}$ yielding an anomaly signal equal to zero are selected and the corresponding values of $l$ are stored in a subset $\mathcal{M}$ defined as

$$\mathcal{M} = \{l \in \{1, \cdots, L\} |\ \|\widehat{\mathbf{e}}_{\mathrm{D},l}\|_2 = 0\}. \tag{7}$$

Note that $\mathcal{M}$ contains the values of $l$ associated with the discrete atoms of $\boldsymbol{\Phi}_{\mathrm{D}}$ that are the closest to $\boldsymbol{y}_{\mathrm{D}}$. The regularization parameter $b_{\mathrm{D}}$ plays an important role in the atom selection step because it fixes the level of authorized deviation from a discrete parameter and an atom of the dictionary. The lower the value of $b_{\mathrm{D}}$, the lower the number of selected atoms that will be used to estimate

the continuous nominal signal $\boldsymbol{y}_{\mathrm{C}}$. Conversely, the higher the value of $b_{\mathrm{D}}$, the better the nominal estimation of $\boldsymbol{y}_{\mathrm{C}}$, with a higher risk to break the links between discrete and continuous parameters by selecting non-representative atoms and miss multivariate contextual anomalies.

*4.2.2. Sparse coding for continuous atoms*

The nominal continuous signal is approximated using a sparse linear combination of atoms contained in a dictionary denoted as $\boldsymbol{\Phi}_{\mathcal{M}}$, composed of the continuous atoms $\boldsymbol{\phi}_{\mathrm{C},l}, l = 1, \cdots, L$ whose discrete parts $\boldsymbol{\phi}_{\mathrm{D},l}$ have been selected in the discrete atom selection (i.e., $l \in \mathcal{M}$). More precisely, when $\mathcal{M} = \emptyset$, a discrete anomaly is detected and no sparse coding is performed for continuous atoms. When $\mathcal{M} \neq \emptyset$, the continuous atoms corresponding to the elements of $\mathcal{M} \neq \emptyset$ are selected and a continuous sparse decomposition is performed using the resulting representative continuous atoms (in view of the discrete test signal $\boldsymbol{y}_{\mathrm{D}}$). These continuous atoms are used to detect anomalies in multivariate correlated mixed data by preserving the relationships between discrete and continuous parameters. As a consequence, the sparse representation model used for the continuous parameters is defined as

$$\min_{\boldsymbol{x}_{\mathrm{C}}, \mathbf{e}_{\mathrm{C}}} \frac{1}{2}\|\boldsymbol{y}_{\mathrm{C}} - \boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{x}_{\mathrm{C}} - \mathbf{e}_{\mathrm{C}}\|_2^2 + a_{\mathrm{C}}\|\boldsymbol{x}_{\mathrm{C}}\|_1 + b_{\mathrm{C}}\sum_{k=1}^{K_{\mathrm{C}}}\|\mathbf{e}_{\mathrm{C},k}\|_2 \tag{8}$$

where $\|\boldsymbol{x}\|_1 = \sum_n |x_n|$ is the $\ell_1$ norm of $\boldsymbol{x}$, $\mathbf{e}_{\mathrm{C},k}$ corresponds to the $k$th time series of $\mathbf{e}_{\mathrm{C}}$ associated with the $k$th parameter with $k = 1, \ldots, K_{\mathrm{C}}$, $a_{\mathrm{C}}$ and $b_{\mathrm{C}}$ are regularization parameters that control the level of sparsity of the coefficient vector $\boldsymbol{x}_{\mathrm{C}}$ and the anomaly signal $\mathbf{e}_{\mathrm{C}}$, respectively. Note that (8) considers two distinct sparsity constraints for the coefficient vector $\boldsymbol{x}_{\mathrm{C}}$ and the anomaly signal $\mathbf{e}_{\mathrm{C}}$. This formulation reflects the fact that a nominal continuous signal can be well approximated by a linear combination of *few* atoms of the dictionary (sparsity of $\boldsymbol{x}_{\mathrm{C}}$) and that anomalies are *rare* and affect few parameters at the same time (sparsity of $\mathbf{e}_{\mathrm{C}}$).

Problem (8) can be solved with the alternating direction method of multipliers (ADMM) [40] by adding an auxiliary variable $\mathbf{z}$

$$\min_{\boldsymbol{x}_{\mathrm{C}}, \mathbf{e}_{\mathrm{C}}, \mathbf{z}} \frac{1}{2}\|\boldsymbol{y}_{\mathrm{C}} - \boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{x}_{\mathrm{C}} - \mathbf{e}_{\mathrm{C}}\|_2^2 + a_{\mathrm{C}}\|\mathbf{z}\|_1 + b_{\mathrm{C}}\sum_{k=1}^{K_{\mathrm{C}}}\|\mathbf{e}_{\mathrm{C},k}\|_2 \tag{9}$$

and the constraint $\mathbf{z} = \boldsymbol{x}_{\mathrm{C}}$. Note that, contrary to Problem (8), the first and second terms of (9) are decoupled, which allows an easier estimation of the vector $\boldsymbol{x}_{\mathrm{C}}$. The ADMM algorithm associated with (9) minimizes the following augmented Lagrangian

$$\mathcal{L}_A(\boldsymbol{x}_{\mathrm{C}}, \mathbf{z}, \mathbf{e}_{\mathrm{C}}, \mathbf{m}, \mu) = \frac{1}{2}\|\boldsymbol{y}_{\mathrm{C}} - \boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{x}_{\mathrm{C}} - \mathbf{e}_{\mathrm{C}}\|_2^2 + a_{\mathrm{C}}\|\mathbf{z}\|_1$$

$$+b_{\mathrm{C}}\sum_{k=1}^{K_{\mathrm{C}}}\|\mathbf{e}_{\mathrm{C},k}\|_2 + \mathbf{m}_{\mathrm{C}}^T(\mathbf{z} - \boldsymbol{x}_{\mathrm{C}}) + \frac{\mu_{\mathrm{C}}}{2}\|\mathbf{z} - \boldsymbol{x}_{\mathrm{C}}\|_F^2 \tag{10}$$

where $\mathbf{m}_{\mathrm{C}}$ is a Lagrange multiplier vector and $\mu_{\mathrm{C}}$ is a regularization parameter controlling the level of deviation between $\boldsymbol{z}$ and $\boldsymbol{x}_{\mathrm{C}}$. The ADMM algorithm is iterative and alternatively estimates $\boldsymbol{x}_{\mathrm{C}}$, $\mathbf{z}$, $\mathbf{e}_{\mathrm{C}}$ and $\mathbf{m}_{\mathrm{C}}$. More details about the update equations of the different variables at the $k$th iteration are provided below.

**Updating $\boldsymbol{x}_{\mathrm{C}}$**

$\boldsymbol{x}_{\mathrm{C}}$ is classically updated as follows

$$\widehat{\boldsymbol{x}}_{\mathrm{C}}^{k+1} = \arg\min_{\boldsymbol{x}_{\mathrm{C}}} \frac{1}{2}\|\boldsymbol{y}_{\mathrm{C}} - \boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{x}_{\mathrm{C}} - \mathbf{e}_{\mathrm{C}}^k\|_2^2 + \mathbf{m}_{\mathrm{C}}^k(\mathbf{z}^k - \boldsymbol{x}_{\mathrm{C}}) + \frac{\mu_{\mathrm{C}}^k}{2}\|\mathbf{z}^k - \boldsymbol{x}_{\mathrm{C}}\|_2^2. \tag{11}$$

Simple algebra leads to

$$\widehat{\boldsymbol{x}}_{\mathrm{C}}^{k+1} = (\boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{\Phi}_{\mathcal{M}}^T + \mu_{\mathrm{C}}^k I)^{-1}(\boldsymbol{\Phi}_{\mathcal{M}}^T\mathbf{r}_{\mathrm{C}}^k + \mathbf{m}_{\mathrm{C}}^k + \mu_{\mathrm{C}}^k\mathbf{z}^k) \tag{12}$$

where $\mathbf{r}_{\mathrm{C}}^k = \boldsymbol{y}_{\mathrm{C}} - \mathbf{e}_{\mathrm{C}}^k$.

**Updating z**

The update of $\mathbf{z}$ is defined as

$$\widehat{\mathbf{z}}^{k+1} = \arg\min_{\mathbf{z}} a_{\mathrm{C}}\|\mathbf{z}\|_1 + (\mathbf{m}_{\mathrm{C}}^k)^T(\mathbf{z} - \boldsymbol{x}_{\mathrm{C}}^{k+1}) + \frac{\mu_{\mathrm{C}}^k}{2}\|\mathbf{z} - \boldsymbol{x}_{\mathrm{C}}^{k+1}\|_2^2. \tag{13}$$

The solution of (13) is given by the element-wise soft thresholding operator

$$\widehat{\mathbf{z}}^{k+1} = S_{\gamma^k}\left[\boldsymbol{x}_{\mathrm{C}}^{k+1} - \frac{1}{\mu_{\mathrm{C}}^k}\mathbf{m}_{\mathrm{C}}^k\right]$$

with $\gamma^k = \frac{a_{\mathrm{C}}}{\mu_{\mathrm{C}}^k}$, where the thresholding operator $S_\gamma(\mathbf{u})$ is defined by

$$S_\gamma(\mathbf{u}) = \begin{cases} \mathbf{u}(n) - \gamma & \text{if } \mathbf{u}(n) > \gamma \\ 0 & \text{if } |\mathbf{u}(n)| \leq \gamma \\ \mathbf{u}(n) + \gamma & \text{if } \mathbf{u}(n) < -\gamma \end{cases} \tag{14}$$

where $\mathbf{u}(n)$ is the $n$th component of $\mathbf{u}$.

**Updating $\mathbf{e}_{\mathrm{C}}$**

The error vector $\boldsymbol{e}$ is also updated using the shrinkage operator already defined in (6) for the sparse coding of discrete atoms, i.e., as $\widehat{\boldsymbol{e}}_{\mathrm{C}} = T_{b_{\mathrm{C}}}[\boldsymbol{y}_{\mathrm{C}} - \boldsymbol{\Phi}_{\mathcal{M}}\boldsymbol{x}_{\mathrm{C}}]$ The ADMM resolution of (9) is detailed in [15] and summarized in Algorithm 1 (theoretical convergence properties are detailed in [41]).

**Algorithm 1** $\boldsymbol{x}, \mathbf{e}, \mathbf{z}, \mathbf{m} = \mathrm{ADMM}(\boldsymbol{y}, \boldsymbol{\Phi}, \mu, \rho, a, b)$.

**Initialisation:** k=1, $\mathbf{z}^0, \mathbf{e}^0, \mathbf{m}^0, \mu^0, \rho, \epsilon, a, b$
**repeat**
$\boldsymbol{x}^{k+1} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mu^k I)^{-1}[\boldsymbol{\Phi}^T(\boldsymbol{y} - \mathbf{e}^k) + \mathbf{m}^k + \mu^k\mathbf{z}^k]$
$\mathbf{z}^{k+1} = S_\gamma(\boldsymbol{x}^{k+1} - \frac{1}{\mu^k}\mathbf{m}^k), \gamma = \frac{a}{\mu}^k$
$\mathbf{e}^{k+1} = T_b(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x})$
$\mathbf{m}^{k+1} = \mathbf{m}^k + \mu^k(\mathbf{z}^{k+1} - \boldsymbol{x}^{k+1})$
$\mu^{k+1} = \rho\mu^k$
k = k+1
**until** stop criteria

*4.2.3. Proposed anomaly detection strategy*

The estimated anomaly signal $\widehat{\mathbf{e}}$ associated with the test signal $\boldsymbol{y}$ is built by the concatenation of its discrete and continuous counterparts $\widehat{\mathbf{e}} = (\widehat{\mathbf{e}}_{\mathrm{D}}^T, \widehat{\mathbf{e}}_{\mathrm{C}}^T)^T$. The proposed anomaly detection rule is described below:

A discrete anomaly is detected if $\|\mathbf{e}_{\mathrm{D}}\|_2 > 0$ (i.e $\mathcal{M} = \emptyset$). Moreover, a continuous or contextual anomaly is detected when $\|\mathbf{e}_{\mathrm{C}}\|_2 > S_{\mathrm{PFA}}$.

where $S_{\mathrm{PFA}}$ is a threshold depending on the probability of false alarm of the detector. This threshold can be adjusted by the user or determined using receiver operating characteristic (ROC) curves if a ground-truth is available (which will be the case in this paper). Note that the set $\mathcal{M}$ is used to detect anomalies in discrete data when it reduces to the empty set, i.e., when $\mathcal{M} = \emptyset$, and to extract the continuous atoms corresponding to the elements of $\mathcal{M}$ when $\mathcal{M} \neq \emptyset$. These continuous atoms are used to detect anomalies in multivariate correlated mixed data using the decision rule (8), which preserves the relationships between the discrete and continuous parameters. More precisely, the first step of the algorithm (detailed in Section 4.2.1) considers the discrete part of $\boldsymbol{y}$, namely $\boldsymbol{y}_{\mathrm{D}}$, and detects potential anomalies affecting the discrete parameters. In the second part of the algorithm (detailed in Section 4.2.2), (8) is used to detect univariate continuous anomalies and contextual discrete/continuous anomalies using the continuous part of

**Algorithm 2** Anomaly detection rule in mixed telemetry using a sparse representation $(\boldsymbol{y}, \boldsymbol{\Phi}^C_{\mathcal{M}}, \tau_{\max})$.

*Discrete Model and Atom Selection*
**for** $l = 1$ to $L$ **do**
$\widehat{\mathbf{e}}_{\mathrm{D},l} = T_{b_\mathrm{D}}\left[\boldsymbol{y}_\mathrm{D} - \boldsymbol{\phi}_{\mathrm{D},l}\right]$
**end for**
$\mathcal{M} = \{l \in \{1, \ldots, L\} | \|\widehat{\mathbf{e}}_{\mathrm{D},l}\|_2 = 0\}$
Discrete Anomaly: if $\mathcal{M} = \emptyset$, a discrete anomaly is declared
If $\mathcal{M} \neq \emptyset$, the algorithm considers a continous model
*Continuous Model*
$\boldsymbol{\Phi}_{\mathcal{M}} = \{\boldsymbol{\Phi}_{\mathrm{C},l} \mid l \in \mathcal{M}\}$
*Anomaly Detection*
$\widehat{\mathbf{e}} = (\widehat{\mathbf{e}}_\mathrm{D}^T, \widehat{\mathbf{e}}_\mathrm{C}^T)^T$
Joint Anomaly: if $\mathcal{M} \neq \emptyset$ and $\|\widehat{\mathbf{e}}_\mathrm{C}\|_2 > S_{\mathrm{PFA}}$, a joint discrete/continuous anomaly is detected

the atoms selected in the first step (denoted as $\Phi_{\mathcal{M}}$). The two steps of the algorithms are summarized below and in Algorithm 2.

- **First step**: the discrete anomaly detection looks for the anomaly vectors $\widehat{\boldsymbol{e}}_{\mathrm{D},l}, l = 1, \ldots, L$ resulting from the discrete sparse decomposition that are equal to 0 and builds a subset $\mathcal{M}$ defined as

$$\mathcal{M} = \{l \in \{1, \cdots, L\} |\ \|\widehat{\mathbf{e}}_{\mathrm{D},l}\|_2 = 0\}. \tag{15}$$

  When the set $\mathcal{M}$ is empty, a discrete anomaly is detected.
- **Second step**: The set $\mathcal{M}$ is used to build a dictionary of continuous atoms (denoted as $\Phi_{\mathcal{M}} = \{\boldsymbol{\phi}_{\mathrm{C},l}, l = 1, \cdots, L\}$) associated with the discrete atoms selected in the first step, i.e., $\{\boldsymbol{\phi}_{\mathrm{D},l}, l = 1, \cdots, L\}$. This atom selection allows the continuous sparse decomposition to be performed using only representative continuous atoms (in view of the discrete test signal $\boldsymbol{y}_\mathrm{D}$). As a consequence, contextual anomalies between discrete and continuous parameters can be detected.

### 4.3. Shift-Invariant option

The proposed method has a shift-invariance (SI) optional step that can be activated for the discrete model and for continuous atom selection. This SI option allows a possible shift between the data of interest and the atoms of the discrete dictionary to be mitigated. Note that this option could also be applied to continuous data. However, since it increases the computational complexity significantly, it has only been considered for discrete data in this work. The SI option consists of building an overcomplete discrete dictionary by applying shifts to all the discrete atoms of the dictionary. In other words, each discrete atom $\boldsymbol{\phi}_{\mathrm{D},l}$ is shifted of $\tau$ lags to create a new discrete atom $\boldsymbol{\phi}_{\mathrm{D},l-\tau}$, with $\tau \in \{-\tau_{\max}, -(\tau_{\max} - 1), \ldots, -1, 0, 1, \ldots, \tau_{\max} - 1, \tau_{\max}\}$. Note that the maximum shift $\tau_{\max}$ has to be fixed by the user. By activating the SI option, the size of the discrete dictionary increases from $L$ to $2L\tau_{\max} + L$ atoms. This option is potentially interesting since it allows more representative atoms to be considered for the estimation of the nominal signal and for atom selection.

## 5. Experimental results

### 5.1. Overview

The first experiment considers a simple dataset composed of $K_\mathrm{D} = 3$ discrete and $K_\mathrm{C} = 7$ continuous parameters with an available ground-truth. The dictionary was constructed using two months of nominal telemetry (without anomalies), which represents approximately 30000 mixed training signals obtained after applying the preprocessing described in Section 4 with the parameters $\delta = 5$ and $w = 50$ (i.e., the signal length is $N = 500$). As explained before, the existing dictionary learning methods such as K-SVD or ODL have not been designed for mixed discrete and continuous data. In this work, we built the dictionary of mixed discrete and continuous parameters as follows: 1) the dictionary is initialized with $L = 2000$ training signals selected randomly in the training database (the choice of $L$ will be discussed later), 2) the proposed sparse coding algorithm is applied to the training data to determine the sparse representation $\boldsymbol{\Phi x}$ and select the $L$ training signals having the highest residuals $\|\boldsymbol{y} - \boldsymbol{\Phi x}\|$. This process is repeated 100 times and the $L$ signals most often selected among the iterations are selected as the columns of the mixed dictionary.

The performance of the different AD methods is evaluated using a test database associated with 18 days of telemetry, i.e., composed of 1000 signals including 90 affected by anomalies. Note that the 90 anomaly signals of the dataset are divided in 7 anomaly periods with various durations displayed in Fig. 1. Note also that a specific attention was devoted to the construction of a heterogeneous test database containing all kinds of anomalies, i.e., univariate discrete and continuous anomalies and two multivariate contextual anomalies. Finally, it is important to note that the majority of these anomalies are actual anomalies observed in operated satellites. This work investigates four AD methods whose principles are summarized below

- The one-class support vector machine (OC-SVM) method [42]: the OC-SVM algorithm was investigated in a multivariate framework by using input vectors composed of mixed continuous and discrete parameters. The input vectors were obtained using the preprocessing step described in Section 4. A. Denote as $\boldsymbol{y} \in \mathbb{R}^N$ one of these input vectors obtained by concatenating time series of the different telemetry parameters. The strategy adopted by OC-SVM is to map the training data in a higher-dimensional subspace $\mathcal{H}$ using a transformation $\varphi$, and to find a linear separator in this subspace, separating the training data (considered as mostly nominal) from the origin with the maximum margin. The separator is found by solving the following problem

$$\min_{\mathbf{w}, \rho, \mathcal{E}_i} \frac{1}{2} \|\ \mathbf{w}\ \|_2 + \frac{1}{\nu N} \sum_{i=1}^{N} \mathcal{E}_i - \rho$$
$$\text{s.t.} \quad \langle \mathbf{w}, \varphi(y) \rangle_{\mathcal{H}} \geq \rho - \mathcal{E}_i, \mathcal{E}_i \geq 0 \tag{16}$$

  where $\mathbf{w}$ is the normal vector to the linear separator, $< ., . >_{\mathcal{H}}$ is the Hilbert inner product for $\mathcal{H}$ where $\mathcal{H}$ is equiped with reproducing kernel $\rho$ is the so-called bias, $\mathcal{E}_i, i = 1, \ldots, N$ are slack variables (which are equal to 0 when $\boldsymbol{y}$ satisfies the constraint and are strictly positive when the constraint is not satisfied) and $\nu$ is a relaxation factor that can be interpreted as the fraction of training data allowed to be outside of the nominal class. Note that the parameter $\nu$ has to be fixed by the user. The kernel used in this work is the Gaussian kernel defined as

$$k(\boldsymbol{y}, \boldsymbol{y}') = \exp\left(-\gamma\ \|\ \boldsymbol{y} - \boldsymbol{y}'\ \|^2\right) \tag{17}$$

  where $\gamma$ is a parameter (also adjusted by the user) controlling the regularity of the separator. Once the separator has been found, determining whether a new data vector is nominal or abnormal is an easy task as it only consists of testing whether this vector falls inside or outside the separating curve. In other terms, the decision rule can be formulated as follows

$$f(\boldsymbol{y}) = \mathrm{sign}[k(\mathbf{w}, \mathbf{y}) - \rho] \tag{18}$$

  where sign is the function defined by

$$\mathrm{sign}(\boldsymbol{y}) = \begin{cases} -1 & \text{if } \boldsymbol{y} < 0 \\ 0 & \text{if } \boldsymbol{y} = 0 \\ 1 & \text{if } \boldsymbol{y} > 0 \end{cases}. \tag{19}$$

Note that an anomaly score can be defined as the distance between the test vector $\boldsymbol{y}$ and the separator with a positive score if $f(y)<0$ and a score equal to zero if $f(y)>0$. This anomaly score in the first case is defined by

$$a(\boldsymbol{y}) = \frac{\rho - k(\mathbf{w}, \mathbf{y})}{\|\mathbf{w}\|}. \tag{20}$$

Finally, we would like to mention that the parameters $\nu$ and $\gamma$ have been tuned by cross validation in this study.

- Mixture of probabilistic principal component analyzers and categorical distributions (MPPCAD) [14]: this is a multivariate AD method based on probabilistic clustering and dimensionality reduction. The input data vector $\boldsymbol{y}$ is divided into two parts associated with continuous and discrete vectors denoted as $\boldsymbol{y}_C \in \mathbb{R}^{K_C}$ (containing one data instance of each continuous parameter) and $\boldsymbol{y}_D \in \mathbb{R}^{K_D}$ (containing one data instance of each discrete parameter) acquired at the same time instant. Each discrete parameter $\boldsymbol{y}_D(j)(j = 1, \ldots, K_D)$ takes its values in the set $\{1, \ldots, M_j\}$ containing $M_j$ different values. MPPCAD assumes that the vector of continuous variables is distributed according to a mixture of Gaussian distributions and that the vector of discrete variables is distributed according to a mixture of categorical distributions. This assumption leads to the following probability density distribution for the continuous data $\boldsymbol{y}_C$

$$p(\boldsymbol{y}_C|\Theta_C) = \sum_{g=1}^{G} \pi_g \mathcal{N}(\boldsymbol{y}_C|\mu_g, \mathbf{C}_g) \tag{21}$$

where the $g$th Gaussian distribution has mean vector $\mu_g$ and covariance matrix $\mathbf{C}_g = \mathbf{W}_g\mathbf{W}_g^T + \sigma_g^2\mathbf{I}_{K_C}$ with $\mathbf{W}_g$ the factor loading matrix, $\sigma_g^2$ the noise variance, $\mathbf{I}_{K_C}$ is the identity matrix of size $K_C \times K_C$ and $\pi_g$ the prior probability of the $g$th cluster. The distribution of the discrete data based on a mixture of categorical distributions is defined as

$$p(\boldsymbol{y}_D|\Theta_D) = \sum_{g=1}^{G} \pi_g \prod_{j=1}^{K_D} \mathrm{Cat}(\boldsymbol{y}_D(j)|\boldsymbol{\theta}_{g,j}) \tag{22}$$

where Cat(.) is the categorical distribution, $\boldsymbol{y}_D(j)$ is the $j$th component of $\boldsymbol{y}_D$ and $\boldsymbol{\theta}_{g,j} = [\theta_{g,j,1}, \ldots, \theta_{g,j,M_j}]$ denotes the parameter vectors of the categorical distributions, i.e., $P(\boldsymbol{y}_D(j) = l|g) = \theta_{g,j,l}$. Finally the joint distribution of the mixed data is obtained assuming independence between $\boldsymbol{y}_C$ and $\boldsymbol{y}_D$

$$p(\boldsymbol{y}_C, \boldsymbol{y}_D|\Theta) = p(\boldsymbol{y}_C|\Theta_C)p(\boldsymbol{y}_D|\Theta_D) \tag{23}$$

where $\Theta = \{\Theta_C^T, \Theta_D^T\}^T$, $\Theta_C = \{\pi_g, \mu_g, \mathbf{W}_g, \sigma_g^2, g = 1, \ldots, G\}$ and $\Theta_D = \{\theta_{g,j}, g = 1, \ldots, G, j = 1, \ldots, M_j\}$. The unknown parameter vector $\Theta$ can be classically estimated using the Expectation-Minimization (EM) algorithm [43] yielding an estimator denoted as $\widehat{\Theta}$. The EM algorithm was initialized using $k$-means clustering following Ding's method [44]. The authors of Yairi et al. [14] proposed to estimate the number of cluster $K$ and the dimensionality of the continuous latent space $L$ using heuristic rules. More precisely, the value of $L$ was tuned using the so-called "elbow-law" after applying a principal component analysis to the continuous data. The number of clusters $K$ was manually estimated based on the scatter plot of the principal component scores. Finally, it is interesting to note that an anomaly score $a(\boldsymbol{y}_C, \boldsymbol{y}_D|\widehat{\Theta})$ can be defined as the minus log likelihood of the mixed data

$$a(\boldsymbol{y}_C, \boldsymbol{y}_D|\widehat{\Theta}) = -\ln\ p(\boldsymbol{y}_C, \boldsymbol{y}_D|\widehat{\Theta}). \tag{24}$$

- New Operational SofTwaRe for Automatic Detection of Anomalies based on Machine-learning and Unsupervised feature Selection (NOSTRADAMUS): this is a univariate method developed by the french space agency CNES based on the OC-SVM method applied to each telemetry parameter individually [7]. The input data are vectors of features (mean, median, minimum, maximum, standard deviation...) computed on time windows resulting from a segmentation for these parameters on a fixed period of time. Different features are computed depending on the discrete or continuous nature of the parameter. The OC-SVM method requires to define an appropriate kernel, which was chosen as the Gaussian kernel in [7]. An anomaly score was also defined in order to quantify the "degree of abnormality" of any test vector. This degree of abnormality corresponds to the distance between this vector and the separator normalized to [0,1] in order to provide a probability of anomaly. Given the univariate framework of NOSTRADAMUS, a score is assigned to each parameter and is denoted as $a(\boldsymbol{y}_k)$ for the kth parameter. In order to compare with multivariate AD methods studied in this work, we define a multivariate score for NOSTRADAMUS corresponding to the sum of the univariate scores

$$a(\boldsymbol{y}) = \sum_{k=1}^{K} a(\boldsymbol{y}_k) \tag{25}$$

where K is the number of parameters.

- ADDICT: the proposed strategy is a multivariate AD method based on a sparse decomposition of any test vector $\boldsymbol{y}$ on a DICTionary (ADDICT) of normal patterns. The dictionary is learned from mixed training signals associated with a period of time where no anomaly was detected. The input data of this method are mixed vectors composed of telemetry parameters acquired during the same period of time. The preprocessing applied to the vector $\boldsymbol{y}$ and the AD algorithm were detailed in Section 4. An anomaly score can also be defined for this method

$$a(\boldsymbol{y}) = \begin{cases} -1 & \text{if } \|\widehat{\boldsymbol{e}}_D\|_2 > 0 \text{ (i.e. } \mathcal{M} = \emptyset) \\ \|\ \widehat{\boldsymbol{e}}_C\ \|_2 & \text{otherwise} \end{cases}. \tag{26}$$

All the regularization parameters $(a, b_C, b_D)$ were determined by cross validation in this study. At this point, it is worth mentioning that it might be interesting to consider other approaches such as Bayesian inference [45] to estimate these regularization parameters.

### 5.2. Performance evaluation

This section compares detection results obtained with the AD methods summarized in the previous section when they are applied on an anomaly dataset with available ground-truth. Fig. 3 shows the different anomaly scores with ground-truth marked by red backgrounds for OCSVM (a), MPPCAD (b), NOSTRADAMUS (c) and the proposed method ADDICT (d). The higher the score, the higher the probability of anomaly. For each method, the detection rule compares the anomaly score to a threshold and detects an anomaly if this score exceeds an appropriate threshold. In an operational context, the threshold can be set in order to obtain an acceptable probability of detection by constraining the probability of false alarm to be upper-bounded, since detecting too many false alarms is a problem for operational missions. In this paper, we determined the threshold associated with the value of the pair (probability of false alarm $P_{FA}$, probability of detection $P_D$) located the closest from the ideal point (0,1). Fig. 3 shows that point anomalies located in boxes #3 and #5 of Fig. 1) are well detected by all the methods. Indeed, the scores returned by all the methods during this anomaly period are significantly higher than the average score. The second anomaly (box #2 in Fig. 1 and indicated as 2 in Fig. 3) is a univariate anomaly that is also relatively well detected by all the methods. The first collective anomaly (box #1 in Fig. 1), which corresponds to an abnormal duration of a discrete

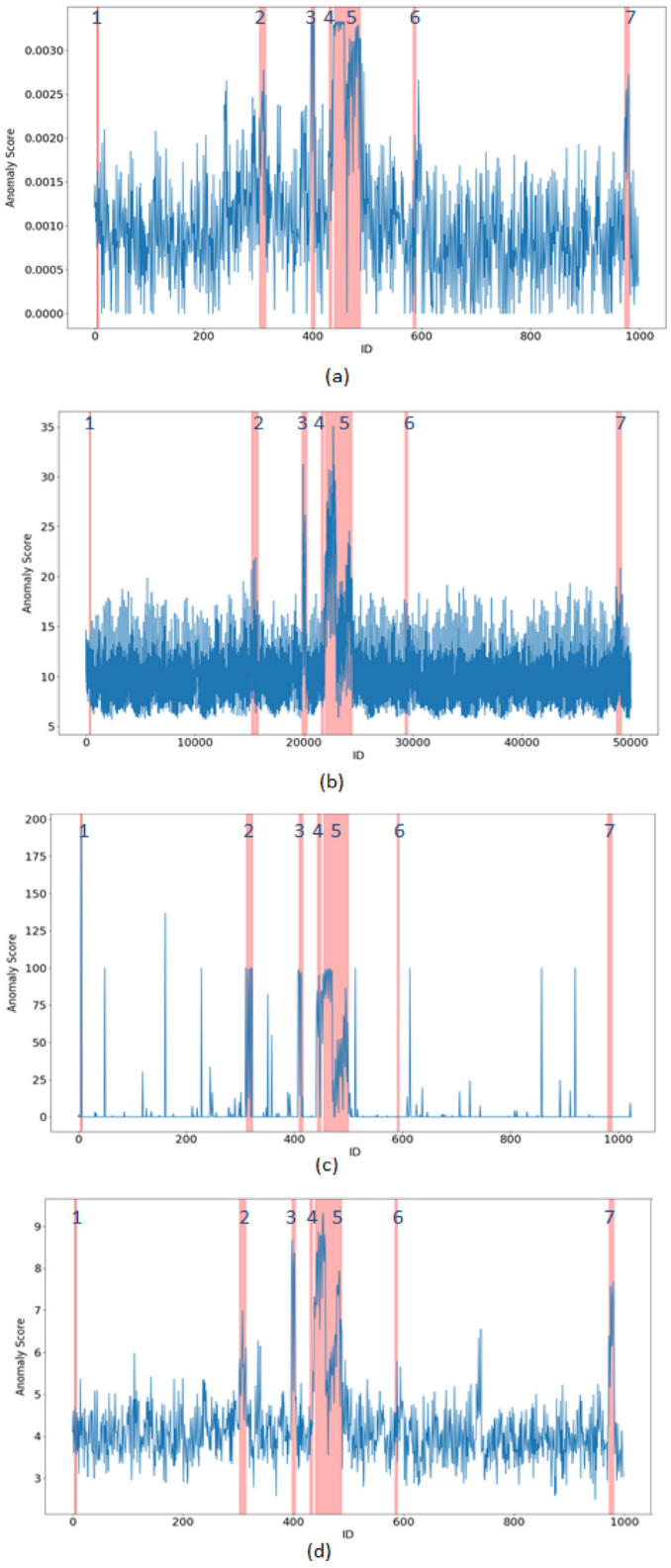


**Fig. 3.** Anomaly scores for the dataset with ground-truth marked by red background. OCSVM (a), MPPCAD (b), NOSTRADAMUS (c) and ADDICT (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

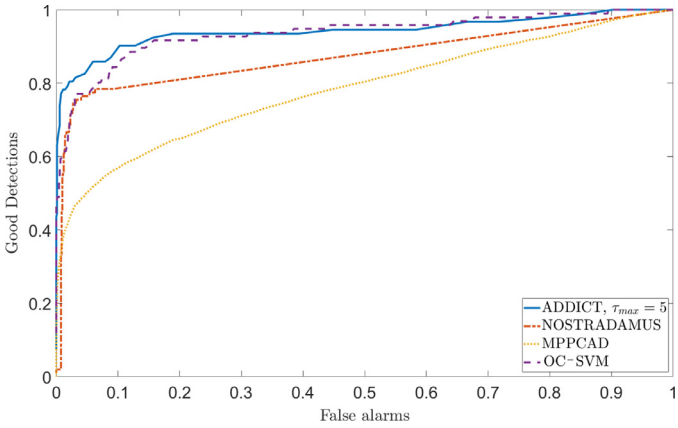


**Fig. 4.** ROC curves of OC-SVM, NOSTRADAMUS, MPPCAD and ADDICT for the anomaly dataset.

**Table 1**
Values of $P_D$ and $P_{FA}$ for OCSVM, MPPCAD, NOSTRADAMUS and ADDICT.

| Method | Threshold | $P_D$ | $P_{FA}$ |
|---|---|---|---|
| OC-SVM | 0.0016 | 89% | 12.3% |
| MPPCAD | 12 | 67% | 25% |
| NOSTRADAMUS | 29 | 77.26% | 6% |
| ADDICT ($\tau_{max} = 0$) | 3.8 | 84.6% | 9.8% |
| ADDICT ($\tau_{max} = 5$) | 3.7 | 89% | 10.2% |

parameter, is only detected by NOSTRADAMUS. This non detection by MPPCAD can be partly explained by the fact that this approach processes time windows of length $w = 1$ whereas this kind of anomaly clearly requires to consider longer time windows. The fourth anomaly (box #4 in Fig. 1) can be classified as a collective anomaly if we consider its abnormal duration, or as a multivariate contextual anomaly if we consider the abnormal joint behaviour of the two discrete parameters. This anomaly is only detected by NOSTRADAMUS. This result is due to the fact that anomalies affecting discrete data are poorly managed by the multivariate AD methods. Note that in this first experiment, the SI option of the proposed method which aims at solving this problem was not active. On the other hand, the sixth anomaly (box #6 in Fig 1 and referred to as 6 in Fig. 3), corresponding to a univariate contextual anomaly that occurs on continuous data, is detected by the proposed method but not by the others. The non detection of this anomaly by MPPCAD can be explained by the same arguments used for the collective anomaly. Moreover, this anomaly is not detected by NOSTRADAMUS since it does not affect significantly features that form the input vector of this algorithm. Finally, the last anomaly corresponding to a multivariate contextual anomaly for a continuous parameter (labelled 7 in Fig. 3 and located in box #7 of Fig. 1), is perfectly detected by OCSVM and ADDICT. However, it is less significant for MPPCAD and is not detected by NOSTRADAMUS, which is not able to handle anomalies due to correlations between the different parameters.

Quantitative results in terms of probability of detection and probability of false alarm are given in Fig. 4 which displays ROC curves of the four methods for the anomaly dataset. ROC curves were built using ground-truth. The performances corresponding to the pair (probability of false alarm $P_{FA}$, probability of detection $P_D$) located the closest from the ideal point (0,1) are reported in Table 1. Our comments are summarized below

- Few false alarms are generated by the MPPCAD algorithm but an important proportion of anomalies from the database is not detected.

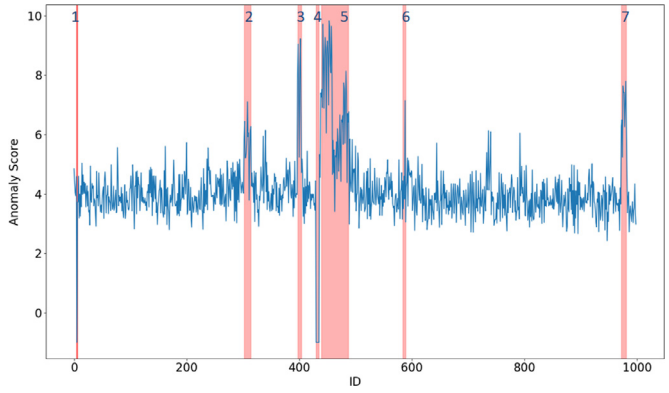


**Fig. 5.** Anomaly scores obtained with ADDICT for the dataset with ground-truth marked by red background and the shift-invariance option enabled with $\tau_{\max} = 5$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

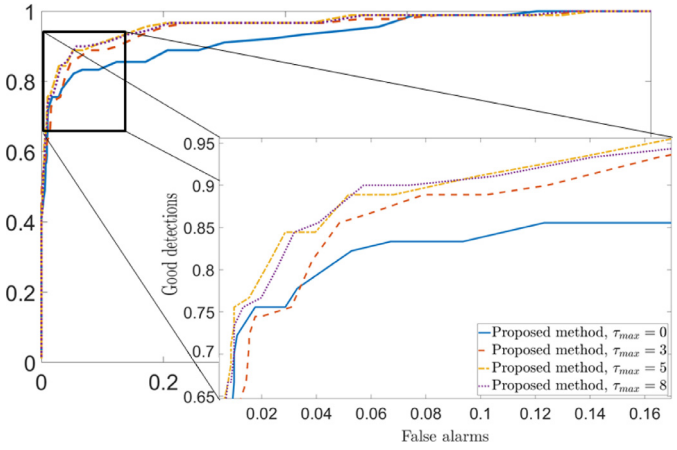


**Fig. 6.** ROC curves for ADDICT with different values of $\tau_{\max} \in \{0, 1, 3, 5, 8\}$.

- The OC-SVM method detects the most serious anomalies affecting continuous parameters but is not able to detect anomalies associated with discrete parameters.
- NOSTRADAMUS is able to detect the majority of the univariate anomalies but fails for the multivariate ones. It would be interesting to adapt NOSTRADAMUS to detect multivariate anomalies in mixed data.
- The results obtained with ADDICT are very encouraging with a high probability of detection $P_D = 0.846$ and a small probability of false alarm $P_{FA} = 0.098$. These results are improved with the SI option leading to $P_D = 0.89$ and $P_{FA} = 0.102$, which corresponds to the best performance for this dataset.

### 5.3. Shift invariant option

This section investigates the usefulness of the SI option for the proposed method. Fig. 5 displays the anomaly scores of the proposed method with a maximum allowed shift $\tau_{\max} = 5$. By comparing these results with those in Fig. 3(d) obtained without using the SI option (i.e., with $\tau_{\max} = 0$), we observe that the SI option allows anomalies affecting discrete parameters to be detected (anomalies #1 and #4 in Fig. 1). In addition, the activation of the SI option decreases scores of nominal signals, which is an interesting property. More precisely, 70% of nominal signals yield a lower anomaly score when the SI option is enabled. Reducing this score allows the number of false alarms to be decreased, improving the performance of the proposed method.

The maximum allowed shift $\tau_{\max}$ was determined using ROCs that express the probability of detection $P_D$ as a function of the probability of false alarm $P_{FA}$. Fig. 6 shows ROCs for different values of $\tau_{\max}$ showing that $\tau_{\max} = 5$ leads to a good compromise in terms of performance and computational complexity (the higher $\tau_{\max}$ the higher the execution time). These results confirm the importance of the SI option for the proposed method.

### 5.4. Selecting the number of atoms in the dictionary

This section explains how the proposed method ADDICT selects the number of atoms in the dictionary. Intuitively, the more atoms in the dictionary, the better the sparse representation of nominal signals and the lower the probability of false alarms. Our experiments have shown that the anomalies are also better approximated when the number of atoms in the dictionary increases.

Fig. 7 shows the values of $P_D$ and $P_{FA}$ returned by the proposed method ADDICT versus the number of atoms in the dictionary. The performance starts by improving when the number of atoms increases. For instance, moving from 100 to 2000 atoms allows $P_D$ to increase from 77.78% to 88.89% and $P_{FA}$ to decrease from 26% to 5.72%, which is a significant improvement. Beyond 2000 atoms, the detection performance does not improve, which explains the choice $L = 2000$ in our experiments. This analysis emphasizes that choosing the number of atoms in the dictionary is important for AD using ADDICT.

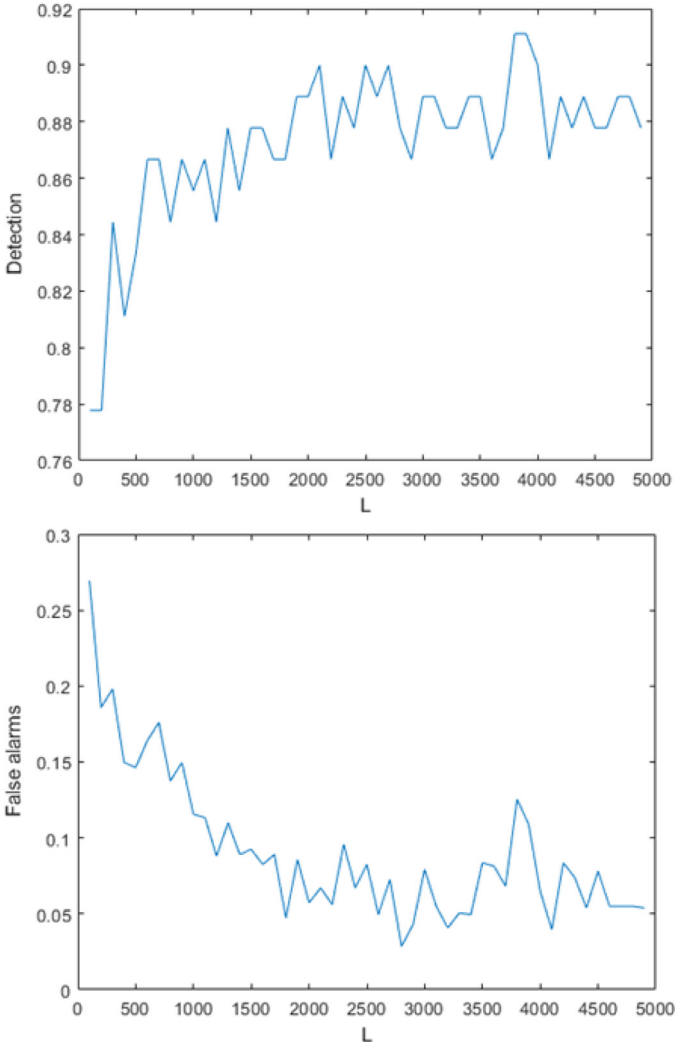


**Fig. 7.** Values of $P_D$ (top) and $P_{FA}$ (bottom) versus the number of dictionary atoms $L$.

## 6. Conclusion

This paper investigated a new data-driven method for anomaly detection in mixed housekeeping telemetry data based on a sparse representation and dictionary learning. The proposed method can handle mixed discrete and continuous parameters that are processed jointly allowing possible correlations between these parameters to be captured. The approach was evaluated on a heterogeneous anomaly dataset with available ground-truth. The first results demonstrated the competitiveness of this approach with respect to the state-of-the-art. Our experiments showed the usefulness of a shift invariant option for the detection of anomalies affecting discrete parameters, leading to a significant reduction in the probability of false alarm.

For future work, different issues might be investigated. The most challenging task is the dictionary learning step, which should be adapted to mixed discrete and continuous data. Another research prospect is the potential use of sparse codes or anomaly signals to identify the causes of the anomaly. Finally, we think that integrating the feedback of users in the algorithm might improve the future detection of anomalies, e.g., by reducing the number of false alarms. This opens the way for many works related to online or sequential anomaly detection, which could be useful for spacecraft health monitoring.

## Declaration of Competing Interest

None.

## Acknowledgements

The authors would like to thank Pierre-Baptiste Lambert from CNES and Clémentine Barreyre from Airbus Defence and Space for fruitful discussions about anomaly detection in spacecraft telemetry. This work was supported by CNES and Airbus Defence and Space.

## References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. 43 (3) (2009).
[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection for discrete sequences: a survey, ACM Comput. Surv. 24 (5) (2012).
[3] M. Pimentel, D. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.
[4] M. Markou, S. Singh, Novelty detection: a review - part 2: neural network based approaches, Signal Process. 83 (12) (2003) 2499–2521.
[5] M. Markou, S. Singh, Novelty detection: a review - part 1: statistical approaches, Signal Process. 83 (12) (2003) 2481–2497.
[6] C. Barreyre, B. Laurent, J.-M. Loubes, B. Cabon, L. Boussouf, Statistical methods for outlier detection in space telemetries, in: Proc. Int. Conf. Space Operations (SpaceOps'2018), Marseille, France, 2018.
[7] S. Fuertes, G. Picard, J.-Y. Tourneret, L. Chaari, A. Ferrari, C. Richard, Improving spacecraft health monitoring with automatic anomaly detection techniques, in: Proc. Int. Conf. Space Operations (SpaceOps'2016), Daejeon, South Korea, 2016.
[8] J.-A. Martínez-Heras, A. Donati, M. Kirksch, F. Schmidt, New telemetry monitoring paradigm with novelty detection, in: Proc. Int. Conf. Space Operations (SpaceOps'2012), Stockholm, Sweden, 2012.
[9] I. Verzola, A. Donati, J.-A. M. Heras, M. Schubert, L. Somodi, Project sybil: a novelty detection system for human spaceflight operations, in: Proc. Int. Conf. Space Operations (SpaceOps'2016), Daejeon, South Korea, 2016.
[10] C. O'Meara, L. Schlag, L. Faltenbacher, M. Wickler, ATHMoS: automated telemetry health monitoring system at GSOC using outlier detection and supervised machine learning, in: Proc. Int. Conf. Space Operations (SpaceOps'2016), Daejeon, South Korea, 2016.
[11] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, T. Soderstrom, Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding, in: Proc. Int. Conf. Knowledge Data Mining (KDD'2018), London, United Kingdom, 2018, pp. 387–395.
[12] C. O'Meara, L. Schlag, M. Wickler, Applications of deep learning neural networks to satellite telemetry monitoring, in: Proc. Int. Conf. Space Operations (SpaceOps'2018), Marseille, France, 2018.
[13] N. Takeishi, T. Yairi, Anomaly detection from multivariate times-series with sparse representation, in: Proc. IEEE Int. Conf. Syst. Man and Cybernetics, San Siego, CA, USA, 2014.
[14] T. Yairi, N. Takeishi, T. Oda, Y. Nakajima, N. Nishimura, N. Takata, A data-driven health monitoring method for satellite housekeeping data based on pobabilistic clustering and dimensionality reduction, IEEE Trans. Aerosp. Electron. Syst. 53 (3) (2017) 1384–1401.
[15] A. Adler, M. Elad, Y. Hel-Or, E. Rivlin, Sparse coding with anomaly detection, J. Signal Process. Syst. 79 (2) (2015) 179–188.
[16] R.O. Duda, Pattern Classification and Scene Analysis, Wiley, New-York, 1973.
[17] D. Dohono, High-dimensional data analysis: the curses and blessings of dimensionality, AMS Math Challenges Lecture, 2000.
[18] A. Tajer, V. Veeravalli, H.V.Poor, Outlying sequence detection in large data sets: a data-driven approach, IEEE Signal Process. Mag. 31 (5) (2014) 44–56.
[19] A. Gilbert, P. Indyk, M. Iwen, L. Schmidt, Recent developments in the sparse fourier transform: a compressed fourier transform for big data, IEEE Signal Process. Mag. 31 (5) (2014) 91–100.
[20] N. Li, Y. Yang, Robust fault detection with missing data via sparse decomposition, in: Proc. Int. Fed. Automatic Control (IFAC'2013), Shanghai, China, 2013.
[21] M. Elad, A. Aharon, Image denoising via sparse and redundant representation over learned dictionaries, IEEE Trans. Image Process. 14 (12) (2006) 3736–3745.
[22] W. Dong, B. Li, Sparsity-based image denoising via dictionary learning and structural clustering, in: Proc. IEEE conf. Comput. Vis. Pattern Recogni. (CVPR'2010), San Siego, CA, USA, 2010.
[23] S. Xu, X. Yang, S. Jiang, A fast nonlocally centralized sparse representation algorithm for image denoising, Signal Process. 131 (2017) 99–112.
[24] K. Huang, S. Aviyente, Sparse representation for signal classification, in: Proc. Neur. Inf. Process. Syst. (NIPS'2006), Whistler, B C, Canada, 2006.
[25] X. Mei, H. Ling, Robust visual tracking and vehicle classification via sparse representation, IEEE Trans. Patt. Anal. Mach. intell. 33 (11) (2011) 2259–2272–18.
[26] Y. Oktar, M. Turkan, A review of sparsity-based clustering methods, Signal Process. 148 (2018) 20–30.
[27] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Patt. Anal. Mach. intell. 31 (3) (2009) 1–18.
[28] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: Proc. IEEE conf. Comput. Vis. Pattern Recogni., San Francisco, CA, USA, 2010.
[29] H. Cheng, Z. Liu, L. Yang, X. Chen, Sparse representation and learning in visual recognition: theory and applications, Signal Process. 93 (6) (2013) 1408–1425.
[30] Y. Xu, Z. Wu, J. Li, A. Plaza, Z. Wei, Anomaly detection in hyperspectral images based on low-rank and sparse representation, IEEE Trans. Geosci. Remote Sens. 54 (4) (2016) 1990–2000.
[31] S. Biswas, R. Venkatesh, Sparse representation based anomaly detection with enhanced local dictionaries, in: Proc. IEEE Int. Conf. Image Process., Paris, France, 2014.
[32] S.G. Mallat, Z. Zhang, Matching pursuits with times-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397–3415.
[33] P.K.Y. Pati, R. Rezaiifar, Orthogonal matching pursuits: Recursive function approximation with application to wavelet decomposition, in: Proc. Asilomar Conf. Signals, Syst. Comput. (ACSSC'1993), Pacific Grove, CA, USA, 1993.
[34] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. Roy. Statist. Soc. Ser. B (Methodological) 58 (1) (1996) 267–288.
[35] I. Tosic, P. Frossard, Dictionary learning, IEEE Signal Process. Mag. 28 (2011) 27–38.
[36] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse decomposition, IEEE Trans. Signal. Process. 54 (11) (2006) 4311–4322.
[37] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proc. Int. Conf. Mach. Learn. (ICML'2009), Montreal, QC, Canada, 2009, pp. 689–696.
[38] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.
[39] M. Osborne, B. Presnell, B. Turlach, A new approach to variable selection in least squares problems, IMA J. Num. Anal. 20 (3) (2000) 389–403.
[40] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2010) 1–222.
[41] J. Eckstein, D. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Math. Program. (Ser. A B) 55 (3) (1992) 293–318.
[42] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 3 (7) (2001) 1443–1471.
[43] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. Ser. B (Methodological) 39 (1) (1997) 1–38.
[44] C. Ding, X. He, K-means clustering via principal component analysis, in: Int. Conf. Machine Learning (ICML'2004), Banff, Alberta, Canada, 2004.
[45] M.E. Tipping, Bayesian Inference: an Introduction to Principles and Practice in Machine Learning, Springer, 2004.