ANOMALY DETECTION IN MIXED TIME-SERIES USING A CONVOLUTIONAL SPARSE REPRESENTATION WITH APPLICATION TO SPACECRAFT HEALTH MONITORING

Barbara Pilastre(1), *Gustavo Silva*(2), *Loïc Boussouf*(3), *Stéphane d'Escrivan*(4), *Paul Rodríguez*(2), *Jean-Yves Tourneret*(1),(5)

(1) TéSA, 7 Boulevard de la Gare, 31500 Toulouse, France
(2) Electrical Engineering Department, Pontifica Universidad Católica del Perú, Lima, Peru
(3) Airbus Defense and Space, 31 rue des Cosmonautes, 31400 Toulouse, France
(4) CNES, 18 Avenue Edouard Belin, 31400 Toulouse, France
(5) ENSEEIHT-IRIT, 2 Rue Camichel, 31071 Toulouse, France

ABSTRACT

This paper introduces a convolutional sparse model for anomaly detection in mixed continuous and discrete data. This model, referred to as C-ADDICT, builds upon the experiences of our previous ADDICT algorithm. It can handle discrete and continuous data jointly, is intrinsically shift-invariant, and crucially, it encodes each input signal (either continuous or discrete) from a joint activation and uniform combinations of filters, allowing the correlation across the input signals to be captured. The performance of C-ADDICT, is evaluated on a representative dataset composed of real spacecraft telemetries with an available ground-truth, providing promising results.

Index Terms— Anomaly detection, convolutional sparse representation, dictionary learning, shift-invariant.

1. INTRODUCTION

Anomaly detection (AD) is a wide area of research given its diverse applications [1, 2]. Motivated by the success of sparse coding (SC) and convolutional sparse coding (CSC) in many fields [3, 4, 5], these techniques have been applied to AD in images [6, 7, 8], videos [9, 10] or univariate and multivariate time-series [7, 11, 12, 13]. In this context, a univariate anomaly can be defined as an abnormal behaviour (never seen before) of a specific time-series, whereas a multivariate anomaly corresponds to a change in the relationships between several time-series. Note that a multivariate anomaly cannot be detected by the observation of a single time-series.

Most AD methods based on SC or CSC consider a semi-supervised learning and build the dictionary from training data composed of normal patterns (i.e., data without anomalies). New data can then be decomposed into the dictionary allowing potential anomalies to be detected by analyzing the sparse codes of this decomposition [8] or its residuals [7, 13]. In particular, for AD in mixed continuous and discrete time-series, several algorithms have been previously proposed (see Section 4 and [14, 15, 16, 17, 12]). For instance, we highlight that, in our previous work, the Anomaly Detection strategy based on a sparse decomposition on a DICTionary (ADDICT) provided state-of-the-art results to detect anomalies in mixed telemetry [13]. The ADDICT algorithm was able to capture univariate and multivariate anomalies by analyzing the residuals from the sparse coding phase and also benefited from analyzing shifted versions of the input data.

This paper introduces a new convolutional sparse model and its estimation algorithm referred to as convolutional ADDICT (C-ADDICT, see Section 3). The main motivation for C-ADDICT is to integrate in a single model all the ADDICT's goodness (see Section 3.1 for differences between the two models), resulting in an intrinsically shift-invariant algorithm that encodes each input signal (either continuous or discrete) from a joint activation and uniform combinations of filters, capturing correlations across input signals.

2. CSC AND DICTIONARY LEARNING

Convolutional sparse representations [18], also referred to as CSC, approximate a 1D or 2D signal $\mathbf{s}$ as a sum of convolutions between filters $\mathbf{d}_m$ and coefficient maps $\mathbf{x}_m$. The most common formulation of this convolutional model is an extension of the basis pursuit denoising (BPDN) problem

$$\underset{\{\mathbf{x}_m\}}{\arg\min}\ \frac{1}{2}\Big\|\sum_m \mathbf{d}_m * \mathbf{x}_m - \mathbf{s}\Big\|_2^2 + \lambda \sum_m \|\mathbf{x}_m\|_1 \tag{1}$$

where * denotes convolution, the regularization parameter λ controls the sparsity induced by the ℓ_1-norm and $\mathbf{d}_m$ is a set of pre-trained filters. The corresponding convolutional dictionary learning (CDL) problem is

$$\underset{\{\mathbf{x}_{m,k}\}\{\mathbf{d}_m\}}{\arg\min}\ \frac{1}{2}\sum_k \Big\|\sum_m \mathbf{d}_m * \mathbf{x}_{m,k} - \mathbf{s}_k\Big\|_2^2 + \lambda \sum_{k,m}\Big\|\mathbf{x}_{m,k}\Big\|_1 \quad \text{s.t.}\quad \|\mathbf{d}_m\|_2 \le 1 \quad \forall m \tag{2}$$

where the ℓ_2-norm constraint on the filters $\mathbf{d}_m$ avoids scaling ambiguities between the filters and the coefficient maps. The CDL problem (2) is non-convex jointly with respect to $(\mathbf{d}_m, \mathbf{x}_{m,k})$. However, it can be recast as an alternating optimization of two convex problems. Due to the high complexity associated with the convolutional operators in the data fidelity terms of (1)-(2), many algorithms have been proposed to solve the problem in the frequency domain [19, 20, 21, 22, 23]. For instance, these algorithms use the alternating direction methods of multipliers (ADMM) [24] and an accelerated gradient descent (APG) [25] as main frameworks.

3. DETECTING ANOMALIES USING CSC

This section describes the proposed C-ADDICT algorithm, which is a multivariate AD method for mixed data, based on a convolutional sparse representation extension to the multiple measurement vector (MMV) [26] problem. MMV is a generalization of sparse signal representation techniques, including BPDN, in which a collection of K signals[1] are simultaneously represented by the same dictionary.

On what follows, we first highlight the differences between ADDICT and C-ADDICT (Section 3.1), to then proceed to describe the detection and learning stages of the latter (Sections 3.2 and 3.3).

3.1. Key differences between ADDICT and C-ADDICT

On what follows we describe a *toy example* in order to highlight the key differences between ADDICT [13] and C-ADDICT.

[1] The collection of signals can be composed by K independent signals or K windows of fixed size obtained from a single signal.

ADDICT, as well as C-ADDICT, use a set of two (learn) dictionaries which are used to encode a set of continuous $\mathbf{s}_1$ and discrete $\mathbf{s}_2$ signals respectively. Let $\{\Phi_C,\, \Phi_D\}$ and $\{\hat{\Phi}_C,\, \hat{\Phi}_D\}$ respresent such matrix and *convolutional* dictionaries respectively.

In a simplistic fashion, ADDICT will encode the input signal as

$$\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix} = \begin{bmatrix}\Phi_C & \\ & \Phi_D\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} = \begin{bmatrix}\Phi_C\mathbf{x}_1\\ \Phi_D\mathbf{x}_2\end{bmatrix}. \tag{3}$$

In order to capture univariate and multivariate anomalies, ADDICT includes a pre-processing stage (shifted versions of $\{\mathbf{s}_1,\, \mathbf{s}_2\}$) as well as a post-processing stage (mainly, residual analysis). For an explicit description of such stages, see [13, Section 4]).

In contrast, due to the use of convolutional dictionaries, C-ADDICT is intrinsically shift-invariant. Moreover, it encodes the input signal as

$$\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix} = \begin{bmatrix}\hat{\Phi}_C\\ \hat{\Phi}_D\end{bmatrix}\mathbf{x} \quad \left(\text{equivalent to } \begin{bmatrix}\hat{\Phi}_C\mathbf{x}_1\\ \hat{\Phi}_D\mathbf{x}_2\end{bmatrix} \text{ s.t. } \mathbf{x}_1 = \mathbf{x}_2\right). \tag{4}$$

Here we highlight that $\{\mathbf{s}_1,\, \mathbf{s}_2\}$ are encoded via a joint activation of dictionary atoms (i.e. uniform combinations of filters) and thus it captures correlations across the input signals. Furthermore, the alternative formulation (text in parenthesis) of (4) allows us to use a consensus approach (fully desccribed in Section 3.2) to ease the additional computational cost associated with C-ADDICT.

3.2. Proposed anomaly detection strategy

Consider P 1D-signals denoted as $\mathbf{s}_p$, $p = 1, ..., P$ that can be affected by univariate and multivariate anomalies. Each signal $\mathbf{s}_p$ is segmented into K windows of fixed size w denoted as $\mathbf{s}_{k,p} \in \mathbb{R}^w$, $k = 1, ..., K$. Inspired by [13, 7], each window $\mathbf{s}_{k,p}$ is decomposed onto M dictionaries (of size L) $\mathbf{d}_{m,p} \in \mathbb{R}^L$, $m = 1, ..., M$ as

$$\mathbf{s}_{k,p} = \sum_m \mathbf{d}_{m,p} * \mathbf{x}_{m,k,p} + \mathbf{e}_{k,p} + \mathbf{b}_{k,p} \tag{5}$$

where $\mathbf{e}_{k,p}$ is an anomaly vector (possibly equal to $\mathbf{0}$) for $\mathbf{s}_{k,p}$ and $\mathbf{b}_{k,p}$ is an additive noise. In order to identify both univariate and multivariate anomalies, we propose to build a multivariate dictionary $\mathbf{d}_m = [\mathbf{d}_{m,1}^T, ..., \mathbf{d}_{m,P}^T]^T$ composed of the M filters $\mathbf{d}_{m,p}$ describing the normal behaviours of the different signals $\mathbf{s}_p$.
The proposed AD problem (see also Section 3.1) is defined as

$$\begin{aligned}\underset{\substack{\{\mathbf{x}_{m,k,p}\},\\ \{\mathbf{e}_{k,p}\}}}{\arg\min} \quad & \frac{1}{2}\sum_{p,k}\Big\|\sum_m \mathbf{d}_{m,p} * \mathbf{x}_{m,k,p} + \mathbf{e}_{k,p} - \mathbf{s}_{k,p}\Big\|_2^2 \\ & + \lambda \sum_{p,k,m}\|\mathbf{x}_{m,k,p}\|_1 + \beta\sum_{p,k}\|\mathbf{e}_{k,p}\|_2 \\ & \text{s.t.} \quad \mathbf{x}_{m,k,1} = \mathbf{x}_{m,k,2} = \cdots = \mathbf{x}_{m,k,P}\end{aligned} \tag{6}$$

where λ and β are parameters that control the level of sparsity of $\mathbf{x}_{m,k,p}$ and $\mathbf{e}_{k,p}$. This formulation reflects the fact that nominal signals can be well approximated by a sum of convolution between *few* filters and coefficient maps and that anomalies are *rare*. Given the equality constraint on the coefficient maps, the different signals are approximated using the same combination of multivariate filters allowing relationships between signals to be preserved and multivariate anomalies to be detected.

For convenience of notation, we introduce a Toeplitz matrix $\mathbf{D}_{m,p}$ such that $\mathbf{D}_{m,p}\mathbf{x}_{m,k,p} = \mathbf{d}_{m,p} * \mathbf{x}_{m,k,p}$ and the two matrices

$$\mathbf{D}_p = (\mathbf{D}_{0,p},\, \mathbf{D}_{1,p},\, \ldots)\,,\ \mathbf{X}_{k.p} = \left(\mathbf{x}_{0,k,p}^T,\, \mathbf{x}_{1,k,p}^T,\, \cdots\right)^T \tag{7}$$

allowing (6) to be rewritten using the following simplified form

$$\begin{aligned}\underset{\substack{\{\mathbf{X}_{k,p}\},\\ \{\mathbf{e}_{k,p}\}}}{\arg\min}\ & \frac{1}{2}\sum_{p,k}\|\mathbf{D}_p\mathbf{X}_{k,p} + \mathbf{e}_{k,p} - \mathbf{s}_{k,p}\|_2^2 + \lambda\sum_k\|\mathbf{X}_{k,p}\|_1 \\ & + \beta\sum_{p,k}\|\mathbf{e}_{k,p}\|_2 \quad \text{s.t.} \quad \mathbf{X}_{k,1} = \cdots = \mathbf{X}_{k,P}\ .\end{aligned} \tag{8}$$

By adding an auxiliary variable $\mathbf{Y}_k$ that is constrained to be equal to each pth primary variable $\mathbf{X}_{k,p}$, (8) is a global consensus problem (9), which can be solved using a consensus ADMM approach, i.e.,

$$\begin{aligned}\underset{\substack{\{\mathbf{X}_{k,p}\},\{\mathbf{Y}_k\},\\ \{\mathbf{e}_{k,p}\}}}{\arg\min}\ & \frac{1}{2}\sum_{p,k}\|\mathbf{D}_p\mathbf{X}_{k,p} + \mathbf{e}_{k,p} - \mathbf{s}_{k,p}\|_2^2 + \lambda\sum_k\|\mathbf{Y}_k\|_1 \\ & + \beta\sum_{p,k}\|\mathbf{e}_{k,p}\|_2 \quad \text{s.t.} \quad \mathbf{X}_{k,p} = \mathbf{Y}_k \quad \forall p\ .\end{aligned} \tag{9}$$

Using the scaled ADMM model, the associated updates are given by

$$\begin{aligned}\mathbf{X}_{k,p}^{(i+1)} = & \underset{\{\mathbf{X}_{k,p}\}}{\arg\min}\ \frac{1}{2}\sum_{p,k}\|\mathbf{D}_p\mathbf{X}_{k,p} + \mathbf{e}_{k,p}^{(i)} - \mathbf{s}_{k,p}\|_2^2 \\ & + \frac{\rho}{2}\sum_k\|\mathbf{X}_{k,p} - \mathbf{Y}_k^{(i)} + \mathbf{U}_{k,p}^{(i)}\|_2^2\end{aligned} \tag{10}$$

$$\mathbf{Y}_k^{(i+1)} = \underset{\{\mathbf{Y}_k\}}{\arg\min}\ \lambda\sum_k\|\mathbf{Y}_k\|_1 + \frac{\rho}{2}\sum_k\left\|\mathbf{X}_{k,p}^{(i+1)} - \mathbf{Y}_k + \mathbf{U}_{k,p}^{(i)}\right\|_2^2 \tag{11}$$

$$\mathbf{e}_{k,p}^{(i+1)} = \underset{\{\mathbf{e}_{k,p}\}}{\arg\min}\ \frac{1}{2}\sum_{p,k}\|\mathbf{D}_p\mathbf{X}_{k,p}^{(i+1)} + \mathbf{e}_{k,p} - \mathbf{s}_{k,p}\|_2^2 + \beta\sum_{p,k}\|\mathbf{e}_{k,p}\|_2 \tag{12}$$

$$\mathbf{U}_{k,p}^{(i+1)} = \mathbf{U}_{k,p}^{(i)} + \mathbf{X}_{k,p}^{(i+1)} - \mathbf{Y}_k^{(i+1)}\ . \tag{13}$$

The update of $\mathbf{Y}_k$ has a closed form expression defined using a soft-thresholding operator $\mathcal{S}_\gamma(x) = \text{sign}(x)\odot\max(0, |x| - \gamma)$, i.e.,

$$\mathbf{Y}_k^{(i+1)} = \mathcal{S}_{\lambda/\rho}\left[\frac{1}{P}\sum_p(\mathbf{X}_{k,p}^{(i+1)} + \mathbf{U}_{k,p}^{(i)})\right]. \tag{14}$$

The update equation of $\mathbf{e}_{k,p}$ is obtained using a shrinkage operator $\mathcal{T}_b$ on the reconstruction residue

$$\mathbf{e}_{k,p}^{(i+1)} = \mathcal{T}_\beta\Big[\mathbf{s}_{k,p} - \mathbf{D}_p\mathbf{X}_{k,p}^{(i+1)}\Big] \tag{15}$$

where $\mathcal{T}_\beta(x) = \begin{cases}\left(\frac{\|x\|_2 - \beta}{\|x\|_2}\right)x & \text{if } \|x\|_2 > \beta\\ 0 & \text{otherwise.}\end{cases}$.

As in [19], we propose to efficiently address the update equation of $\mathbf{X}_{k,p}$ in the frequency domain. However, the current formulation (10) that is based on independent windows, which can generate many boundary artefacts if the window size is small and comparable to the filter size. Considering that the CSC model is commonly optimized for entire signals (in which artefacts can be negligible), we assume that the set of coefficient maps $\mathbf{X}_p$ that encodes a complete signal s_p is approximately equal to the concatenation of coefficient maps $\mathbf{X}_{k,p}$ corresponding to the subsequent windows $s_{k,p}$, defined as $\mathbf{X}_P = [\mathbf{X}_{1,P} \;\; \mathbf{X}_{2,P} \;\; \cdots \;\; \mathbf{X}_{K,P}]$. Based on these comments, we can reformulate problem (10) in the frequency domain as

$$\underset{\{\hat{\mathbf{X}}_p\}}{\arg\min}\ \frac{1}{2}\sum_p\|\hat{\mathbf{D}}_p\hat{\mathbf{X}}_p - \hat{\mathbf{r}}_p\|_2^2 + \frac{\rho}{2}\|\hat{\mathbf{X}}_p - \hat{\mathbf{Z}}_p^{(i)}\|_2^2 \tag{16}$$

where $\hat{\mathbf{r}}_p^{(i)} = \hat{\mathbf{s}}_p - \hat{\mathbf{e}}_p^{(i)}$, $\hat{\mathbf{Z}}_p^{(i)} = \hat{\mathbf{Y}}^{(i)} - \hat{\mathbf{U}}_p^{(i)}$, and $\hat{\mathbf{D}}_p$, $\hat{\mathbf{X}}_p$, $\hat{\mathbf{r}}_p$ and $\hat{\mathbf{Z}}_p$ denote the frequency domain variables that are obtained after applying the discrete Fourier transform (DFT) to the variables $\mathbf{D}_p$, $\mathbf{X}_p$, $\mathbf{r}_p$ and $\mathbf{Z}_p$. The resulting linear system obtained from (16) after simple algebra has a similar structure as the one obtained in [19], which can be solved using the Sherman-Morrison formula.

3.3. Correlated filters using convolutional dictionary learning

Convolutional dictionary learning can be used to estimate groups of filters $\mathbf{d}_{m,p}$ that accurately reconstruct the nominal signals $\mathbf{s}_{k,p}$ and can be plugged in the CSC-based anomaly detector investigated in the previous section. The proposed CDL extension of (6) is

$$\begin{aligned} \underset{\substack{\{\mathbf{x}_{m,k,p}\}\\ \{\mathbf{d}_{m,p}\}}}{\arg\min} \quad & \frac{1}{2}\sum_{p,k}\Big\|\sum_{m}\mathbf{d}_{m,p} * \mathbf{x}_{m,k,p} - \mathbf{s}_{k,p}\Big\|_2^2 + \lambda\sum_{k,m}\|\mathbf{x}_{m,k,p}\|_1 \\ \text{s.t.} \quad & \|\mathbf{d}_{m,p}\|_2 \leq 1 \quad \forall m,p \\ & \mathbf{x}_{m,k,1} = \mathbf{x}_{m,k,2} = \cdots = \mathbf{x}_{m,k,P}. \end{aligned} \tag{17}$$

As mentioned in Section 2, a non-convex CDL problem can be usually split in two convex sub-problems. For (17), we obtain

$$\begin{aligned} \underset{\{\mathbf{x}_{m,k,p}\}}{\arg\min} \quad & \frac{1}{2}\sum_{p,k}\Big\|\sum_{m}\mathbf{d}_{m,p} * \mathbf{x}_{m,k,p} - \mathbf{s}_{k,p}\Big\|_2^2 + \lambda\sum_{k,m}\|\mathbf{x}_{m,k,p}\|_1 \\ \text{s.t.} \quad & \mathbf{x}_{m,k,1} = \mathbf{x}_{m,k,2} = \cdots = \mathbf{x}_{m,k,P} \end{aligned} \tag{18}$$

$$\begin{aligned} \underset{\{\mathbf{d}_{m,p}\}}{\arg\min} \quad & \frac{1}{2}\sum_{p,k}\Big\|\sum_{m}\mathbf{d}_{m,p} * \mathbf{x}_{m,k,p} - \mathbf{s}_{k,p}\Big\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{d}_{m,p}\|_2 \leq 1 \quad \forall m,p\ . \end{aligned} \tag{19}$$

Problem (18) can be solved as (6) without considering the anomaly term, while problem (19) can be efficiently solved in the frequency domain using the accelerated proximal gradient approach previously proposed in [21, 27].

3.4. Remarks

Consider the P 1D-signals, previously denoted as s_p, can have distinct ranges of amplitudes between them, and the ℓ_2-norm corresponding to the $\ell_{1,2}$-norm penalty term is a restriction based on energy. In the problem (12) where $\beta\sum_{p,k}\|\mathbf{e}_{p,k}\|_2 = \beta\|\mathbf{e}\|_{1,2}$ and a single regularization parameter is used, it is easy to note that it would be necessary to choose different values of the regularization parameters for each signal to avoid suppressing low energy components (anomalies) associated to low amplitude signals. However, in order to avoid the selection of too many regularization parameters and avoid possible low energy suppression by using a single regularization parameter, the training and test signals are normalized using the maximum and minimum values of each corresponding the training signal composed by nominal data.

4. EXPERIMENTAL RESULTS

4.1. Simulation scenario

The proposed C-ADDICT algorithm has been applied to telemetry time-series for spacecraft health monitoring (see [12, 14, 15, 16, 17] for description of the application). Spacecraft telemetry data involves hundreds to thousands of parameters describing the evolution over time of physical quantities (such as temperature, pressure, voltage, ...) or of equipment operating modes (such as antenna position, status ON/OFF, ...). The training and test datasets consist of 7 continuous telemetry parameters and 3 discrete ones obtained from real satellite telemetry. For the training stage, the dictionary composed of 100 filters of length 100 per parameter was learnt from two months of telemetry describing normal behaviour of spacecraft. Furthermore, the proposed learning algorithm uses a regularization parameter λ = 0.01 and 5000 iterations. For the test stage, the proposed C-ADDICT algorithm was evaluated on 18 days of telemetry data with 7 gathered anomaly partitions displayed in Fig. 1. This algorithm used fixed windows of size $w = 200$ with regularization parameters λ and β which were adjusted by a grid-search. In order to properly compare performance of the proposed C-ADDICT algorithm with state-of-the-art methods, we use a detection rule (defined in Section 4.1), in which the estimated anomaly signals are divided into windows of size $w = 50$.

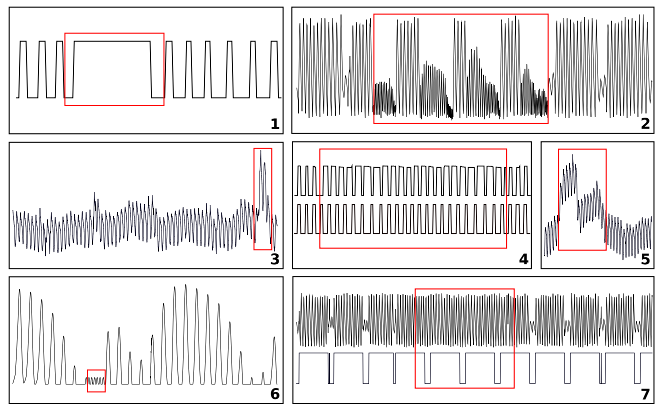


Fig. 1: Examples of univariate (1,2,3,5,6) and multivariate (4,7) anomalies in telemetry data (red boxes).

The C-ADDICT algorithm was compared to four state-of-the-art methods which were evaluated on the same dataset

- The ADDICT algorithm [13] which is a multivariate anomaly detection based on a sparse representation. In this experiment the shift-invariant option of the algorithm was activated with a maximum shift $\tau_{\max} = 5$.
- The W-ADDICT algorithm [28]. It is an extension of the ADDICT method which allows external information to be included via appropriate weights obtained from the correlation coefficient between the test signal and its decomposition into the dictionary.
- The one-class support vector machine (OC-SVM)[29] tested in a multivariate framework with the ADDICT preprocessing.
- The mixture of probabilistic principal component analysers and categorical distribution (MPPCAD) algorithm [10], which is a multivariate anomaly detection method based on probabilistic clustering and dimensionality reduction. The MPCCAD strategy approximates the joint distribution of continuous variables by a mixture of Gaussian distributions and the joint distribution of discrete variables by a mixture of categorical distributions.

4.2. Anomaly detection rule

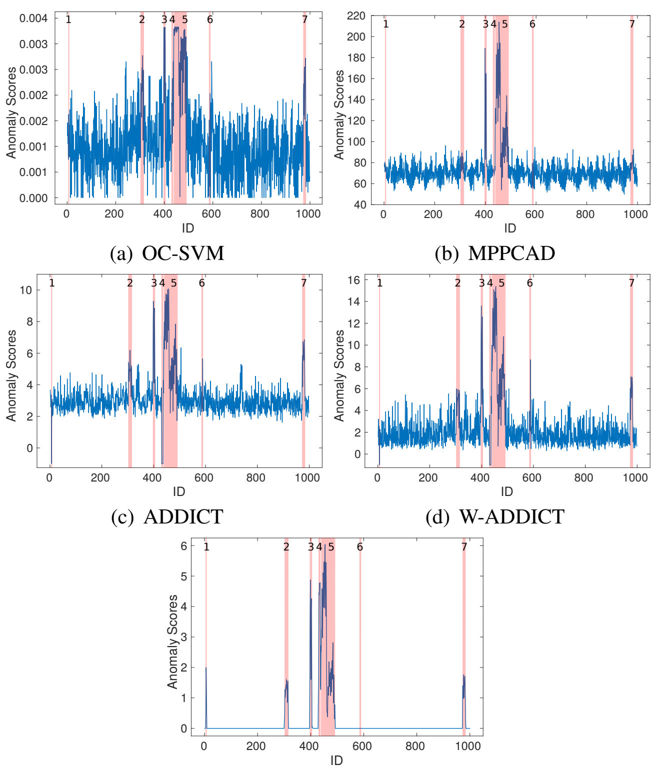
The proposed AD rule (also used in [13]) is based on the estimated anomaly signal $\mathbf{e}_k = [\mathbf{e}'_{1,k}, ..., \mathbf{e}'_{P,k}]'$, where P is the number of time-series acquired by the telemetry system. An anomaly score is defined as the norm of the anomaly signal, i.e., $a(s_k) = \|\mathbf{e}_k\|_2$. This anomaly score is compared to a threshold and an anomaly is detected if the score exceeds the threshold, i.e.,

$$\text{Anomaly detected if } a(s_k) > S_{\text{PFA}} \tag{20}$$

where S_{PFA} is a threshold depending on the probability of false-alarm of the anomaly detector. This threshold was tuned by cross validation from data with an available ground truth.

4.3. Performance evaluation

This section compares the detection performance of the proposed C-ADDICT algorithm to the four state-of-the-art methods recalled before. Fig. 2 shows the anomaly scores returned by OC-SVM (a), MPPCAD (b), ADDICT(c), W-ADDICT (d) and C-ADDICT (e) for the anomaly dataset with ground-truth marked by red backgrounds. The MPPCAD algorithm returns few false alarms and detects univariate anomalies corresponding to extreme values of the time-series (anomalies $\#3$ and $\#5$, see Fig. 1). However the other anomalies of the dataset are not detected by this method. The OC-SVM algorithm detects anomalies affecting continuous parameters but fails for discrete anomalies. The C-ADDICT algorithm detects the most serious anomalies on continuous as well as on discrete data and returns few false alarms compared to the other methods. Note that the 6th anomaly is not detected by C-ADDICT. This anomaly is very difficult to detect since it has a low amplitude and is observed on a limited time interval. The fact that this anomaly only exists on a small interval will not affect the global probability of detection (as seen in Table 1). Finally, not that the 6th anomaly is only detected by ADDICT and W-ADDICT, with a preference for W-ADDICT whose weights are adjusted to each time-series allowing AD to be improved.



(a) OC-SVM (b) MPPCAD (c) ADDICT (d) W-ADDICT (e) C-ADDICT

Fig. 2: Anomaly scores for test signals of the anomaly dataset with ground-truth marked by red backgrounds.

Fig. 3 displays the receiver operational characteristics (ROCs) of the five methods. Quantitative results in terms of probability of detection P_{D}, probability of false alarm P_{FA} and area under the curve (AUC) are also summarized in Table 1 (for a manually selected threshold satisfying the best compromise for spacecraft health monitoring). AD methods based on standard (ADDICT and W-ADDICT) or convolutional (C-ADDICT) sparse representations are more competitive than the other approaches for this dataset. Indeed, ADDICT and W-ADDICT provide high probabilities of detection ($P_{\text{D}} = 80.4\%$ for ADDICT and $P_{\text{D}} = 85.1\%$ for W-ADDICT) and low probabilities of false alarm ($P_{\text{FA}} = 4.43\%$ for ADDICT and $P_{\text{FA}} = 2.7\%$ for W-ADDICT). Despite one missed detection, C-ADDICT seems to be even more competitive with $P_{\text{D}} = 94.7\%$ and $P_{\text{FA}} = 1.7\%$. Furthermore, an interesting property of C-ADDICT is the possibility of setting the detection threshold to a small value thanks to a better sparsity of the anomaly signal $\mathbf{e}_k$. This represents a real advantage with respect to the other methods, for which it is necessary to adjust an appropriate threshold using a ground-truth (that is not always available in operating context).

Table 1: Values of P_{D}, P_{FA} and Area Under Curve (AUC) for OC-SVM, MPPCAD, ADDICT, W-ADDICT and C-ADDICT.

Method	Threshold	P_{D}	P_{FA}	AUC
OC-SVM	0.019	80.9%	7%	0.9413
MPPCAD	76	81.9%	25.9%	0.8779
ADDICT	4.1	81%	3%	0.937
W-ADDICT	4.5	85.1%	2.7%	0.9703
C-ADDICT	0	94.7%	1.7%	0.9706

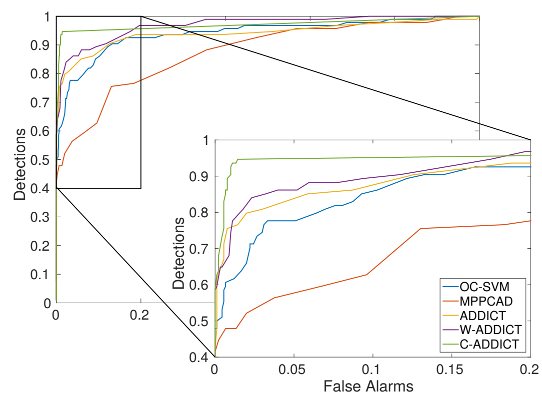


Fig. 3: Roc curves of ADDICT, W-ADDICT, G-ADDICT, W-G-ADDICT and C-ADDICT.

5. CONCLUSION

This paper introduced a new anomaly detection method based on convolutional sparse coding (referred to as C-ADDICT), generalizing an existing anomaly detection method, for mixed signals (discrete and continuous), based on a sparse representation in a dictionary of normal patterns (ADDICT). Due to the optimization problems associated with C-ADDICT, it is intrinsically shift-invariant and enforces correlation across the mixed signals. The performance of C-ADDICT was evaluated for anomaly detection in spacecraft telemetry data and showed its competitiveness with respect to the state-of-the-art. An interesting property of C-ADDICT is the simplicity of adjusting the detection threshold, which is interesting for practical applications. Future works include an evaluation of the method in an operational context including hundreds to thousands time-series. However, it will not affect the global probability of detection since it. Including weights into the proposed anomaly detection rule based on CSC would be also interesting.

6. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Survey*, vol. 43, no. 3, Jul. 2009.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection for discrete sequences: A survey,” *ACM Computing Survey*, vol. 24, no. 5, May 2012.

[3] M. Elad and A. Aharon, “Image denoising via sparse and redundant representation over learned dictionaries,” *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 3736–3745, Dec. 2006.

[4] X. Mei and H. Ling, “Robust visual tracking and vehicle classification via sparse representation,” *IEEE Trans. Patt. Anal. Mach. intell.*, vol. 33, no. 11, pp. 2259–2272–18, Nov. 2011.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Patt. Anal. Mach. intell.*, vol. 31, no. 3, pp. 1–18, Feb. 2009.

[6] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, “Anomaly detection in hyperspectral images based on low-rank and sparse representation,” *IEEE Trans. on Geosci. and Remote Sensing*, vol. 54, no. 4, pp. 1990–2000, Apr. 2016.

[7] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, “Sparse coding with anomaly detection,” *J. Signal Process. Syst.*, vol. 79, no. 2, pp. 179–188, 2015.

[8] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg, “Detecting anomalous structures by convolutional sparse models,” in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015, pp. 1–8.

[9] S. Biswas and R. Venkatesh, “Sparse representation based anomaly detection with enhanced local dictionaries,” in *Proc. IEEE Int. Conf. Image Process.*, Paris, France, Oct. 2014.

[10] B. Zhao, L. Fei-Fei, and E. P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in *Proc. IEEE Int. Conf. Comp. Vision and Pttern. Recogn. (CVPR)*, Washington DC, USA, 2011, pp. 3312–3320.

[11] N. Takeishi and T. Yairi, “Anomaly detection from multivariate times-series with sparse representation,” in *Proc. IEEE Int. Conf. Syst. Man and Cybernetics*, San Siego, CA, USA, Oct. 2014.

[12] T. Yairi, N. Takeishi, T. Oda, Y. Nakajima, N. Nishimura, and N. Takata, “A data-driven health monitoring method for satellite housekeeping data based on probabilistic clustering and dimensionality reduction,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 3, pp. 1384–1401, Jun. 2017.

[13] B. Pilastre, L. Boussouf, S. D’Escrivan, and J-Y. Tourneret, “Anomaly detection in mixed telemetry data using a sparse representation and dictionary learning,” *Signal Process*, vol. 168, to appear, 2019.

[14] S. Fuertes, G. Picard, J.-Y. Tourneret, L. Chaari, A. Ferrari, and C. Richard, “Improving spacecraft health monitoring with automatic anomaly detection techniques,” in *Proc. Int. Conf. Space Operations (SpaceOps’2016)*, Daejeon, South Korea, May 2016.

[15] J.-A Martínez-Heras, A. Donati, M.G.F. Kirksch, and F. Schmidt, “New telemetry monitoring paradigm with novelty detection,” in *Proc. Int. Conf. Space Operations (SpaceOps’2012)*, Stockholm, Sweden, Jun. 2012.

[16] C. O’Meara, L. Schlag, L. Faltenbacher, and 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, May 2016.

[17] C. Barreyre, *Statistiques en grande dimension pour la détection d’anomalies dans les données fonctionnelles issues des satellites*, Ph.D. thesis, Université de Toulouse, Toulouse, France, May 2018.

[18] M. D Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *Proc. IEEE Int. Conf. Comp. Vision and Pttern. Recogn. (CVPR)*, San Francisco, CA, 2010, pp. 2528–2535.

[19] B. Wohlberg, “Efficient convolutional sparse coding,” in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7173–7177.

[20] B. Wohlberg, “Efficient algorithms for convolutional sparse representations,” *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[21] G. Silva and P. Rodrǵuez, “Efficient convolutional dictionary learning using partial update fast iterative shrinkage-thresholding algorithm,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018.

[22] G. Silva and P. Rodríguez, “Efficient algorithm for convolutional dictionary learning via accelerated proximal gradient consensus,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018.

[23] G. Silva and P. Rodríguez, “Generalized combinatorial approach using single filter basis for convolutional sparse modeling,” in *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Guadeloupe, West Indies, Dec. 2019.

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[25] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $o(1/k)^2$,” *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[26] Yonina C Eldar, *Sampling theory: Beyond bandlimited systems*, Cambridge University Press, 2015.

[27] C. Garcia-Cardona and B. Wohlberg, “Convolutional dictionary learning: A comparative review and new algorithms,” *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–381, Sep. 2018.

[28] B. Pilastre, L. Boussouf, S. D’Escrivan, and J-Y. Tourneret, “Représentation parcimonieuse pondérée pour la détection d’anomalies dans des signaux multivariés,” in *Proc. Groupe d’Etude du Traitement du Signal et des Images (GRETSI),*, Lille, France, Aug. 2019.

[29] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 3, no. 7, pp. 1443–1471, 2001.