Proc. Int. Conf. on Neural Information Processing (ICONIP 2015)

Robust Detection of Anomalies via Sparse Methods

Zoltán Á. Milacski[1], Marvin Ludersdorfer[3],
András Lőrincz[1], and Patrick van der Smagt[2,3]

[1] Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
[2] Department for Informatics, Technische Universität München, Munich, Germany
[3] fortiss, An-Institut Technische Universität München, Munich, Germany
srph25@gmail.com, ludersdorfer@fortiss.org, lorincz@inf.elte.hu

Abstract. The problem of *anomaly detection* is a critical topic across application domains and is the subject of extensive research. Applications include finding frauds and intrusions, warning on robot safety, and many others. Standard approaches in this field exploit simple or complex system models, created by experts using detailed domain knowledge.
In this paper, we put forth a statistics-based anomaly detector motivated by the fact that anomalies are sparse by their very nature. Powerful sparsity directed algorithms—namely Robust Principal Component Analysis and the Group Fused LASSO—form the basis of the methodology. Our novel unsupervised single-step solution imposes a convex optimisation task on the vector time series data of the monitored system by employing group-structured, switching and robust regularisation techniques.
We evaluated our method on data generated by using a Baxter robot arm that was disturbed randomly by a human operator. Our procedure was able to outperform two baseline schemes in terms of F_1 score. Generalisations to more complex dynamical scenarios are desired.

1 Introduction

The standard approach to describing system behaviour over time is by devising intricate dynamical models—typically in terms of first- or second-order differential equations—which are based on detailed knowledge about the system. Such models can then help to both control as well as predict the temporal evolution of the system and thus serve as a basis to detect faults.

We investigate the common case where models are either difficult to obtain or not rich enough to describe the dynamic behaviour of the nonlinear plant. This can happen when the plant has many degrees of freedom or high-dimensional sensors, and when it is embedded into a complex environment. Such is typically true for robotic systems, intelligent vehicles, or manufacturing sites; i.e., for a typical modern actor–sensor system that we depend on.

In such cases, the quality of fault detection deteriorates: too many false positives (i.e., false alarms) make the fault detection useless, while too many false negatives (i.e., unobserved faults) may harm the system or its environment.

Rather than fully trusting incomplete models, we put forth a methodology which creates a *probabilistic vector time series model* of the system from the recorded data and detects outliers with respect to this learned model. Following standard procedures [4], detecting such outliers will be called *anomaly detection*.

This type of detection is notoriously difficult as it is an ill-posed problem. First, the notion of anomaly strongly depends on the domain. Then, the boundary between "normal" and "anomalous" might not be precise and might evolve over time. Also, anomalies might appear normal or be obscured by noise. Finally, collecting anomalous data is very difficult, and labelling them is even more so [4]. Two observations are important to make: (i) anomalies are sparse by their very nature, and (ii) in a high-dimensional real-world scenario it will not be possible to rigorously define "normal" and "anomalous" regions of the data space. We therefore focus on an unsupervised approach: in a single step, a probabilistic vector time series model of the system's data is created, and (patterns of) samples that do not fit in the model are conjectured to be the sought anomalies.

In this paper we introduce a new two-factor convex optimisation problem using (i) group-structured, (ii) switching, and (iii) robust regularisation techniques; aiming to discover anomalies in stochastic dynamical systems. We assume that there is a *family of behaviours* between which the system switches randomly. We further assume that the time of the switching is stochastic, and that the system is coupled, i.e., both switching points and anomalies span across dimensions. Given that the general behaviour between the switches can be approximated by a random parameter set defining a *family of dynamics*, we are interested in detecting rare anomalies that may occur with respect to the "normal" behaviour (hence there are two factors). To the best of our knowledge, the combination of the techniques (i)–(iii) is novel.

To test our methods and demonstrate our results, we generated data with a Baxter robot. This system serves as a realistic place holder for a general system with complex dynamics in a high-dimensional space, in which the data cannot be easily mapped to a lower-dimensional plane, while the sensory data are not trivial. We generate realistic anomalies by having the robot perform predefined movements, with random physical disturbances from a human.

2 Theoretical Background

2.1 LASSO and Group LASSO

LASSO [9] is an ℓ_1-regularised least-squares problem defined as follows. Let $D < N$ and let us denote an input vector by $\mathbf{x} \in \mathbb{R}^D$, an overcomplete vector system by $\mathbf{D} \in \mathbb{R}^{D \times N}$, a representation of $\mathbf{x}$ in system $\mathbf{D}$ by $\mathbf{a} \in \mathbb{R}^N$, a tradeoff of penalties by λ. Then LASSO tries to find

$$\min_{\mathbf{a}} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1, \tag{1}$$

which—for a sufficiently large value of λ—will result in a gross-but-sparse representation for vector $\mathbf{a}$: only a small subset of components $\mathbf{a}_i$ will be non-zero

(but large) while the corresponding columns $\mathbf{D}_{.,i}$ will still span $\mathbf{x}$ closely. Model complexity (sparsity) is implicitly controlled by λ. LASSO is the best convex approximation of the NP-hard ℓ_0 version of the same problem, since convexity ensures a unique global optimum value and polynomial-time algorithms grant one to find a global solution [5, 3].

A useful extension to LASSO is the so-called Group LASSO [11]: instead of selecting individual components, groups of variables are chosen. This is done by defining a disjoint group structure on $\mathbf{a}$ (or equivalently on the columns of $\mathbf{D}$):

$$G_i \subseteq \{1, \ldots, N\}, \quad |G_i| = N_i, \quad i = 1, \ldots, L, \tag{2}$$

$$G_i \cap G_j = \emptyset, \quad \forall i \neq j, \tag{3}$$

$$\bigcup_{i=1}^{L} G_i = \{1, \ldots, N\}. \tag{4}$$

Then one can impose a mixed ℓ_1/ℓ_2-regularisation task:

$$\min_{\mathbf{a}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \sum_{i=1}^{L} \|\mathbf{a}_{G_i}\|_2, \tag{5}$$

where $\mathbf{a}_{G_i}$ denotes the subset of components in $\mathbf{a}$ with indices contained by set G_i. The last term is often referred to as the $\ell_{1,2}$ norm of $\mathbf{a}$. Thus the elements from the same group are either forced to vanish together or form a dense representation within the selected groups, resulting in so-called *group sparsity*.

2.2 Robust Principal Component Analysis

Similarly to the LASSO problem, one can define an ℓ_1 norm-based optimisation for compressing the data into a small subspace in the presence of a few outliers. Let $\|\cdot\|_*$ and $\|\mathrm{vec}(\cdot)\|_1$ denote the singular valuewise ℓ_1 norm (nuclear norm) and the elementwise ℓ_1 norm of a matrix, respectively. Then the Robust Principal Component Analysis (RPCA) [2] task for given $\mathbf{X} \in \mathbb{R}^{D\times T}$ and unknowns $\mathbf{U}, \mathbf{S} \in \mathbb{R}^{D\times T}$ is as follows:

$$\min_{\mathbf{U},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{U} - \mathbf{S}\|_F^2 + \lambda\|\mathbf{U}\|_* + \mu\|\mathrm{vec}(\mathbf{S})\|_1, \tag{6}$$

where $\mathbf{U}$ is a robust low-rank approximation to $\mathbf{X}$, gross-but-sparse (non-Gaussian) anomalous errors are collected in $\mathbf{S}$, and $\|\cdot\|_F$ denotes the Frobenius norm. Then as a post-processing step, singular value decomposition can be performed on the outlier-free component $\mathbf{U} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^T$, resulting in a robust estimate of rank $k \in \mathbb{N}$. Note that the classical Principal Component Analysis (PCA) of the input matrix $\mathbf{X}$ would be vulnerable to outliers.

2.3 Fused LASSO and Group Fused LASSO

LASSO can also be modified to impose sparsity for linearly transformed ($\mathbf{Q} \in \mathbb{R}^{T\times M}$ for arbitrary M) components of $\mathbf{v} \in \mathbb{R}^T$:

$$\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{y} - \mathbf{v}\|_2^2 + \lambda\|\mathbf{Q}^T\mathbf{v}\|_1. \tag{7}$$

In the special case when $\mathbf{y} \in \mathbb{R}^T$ is a time series and $\mathbf{Q}$ is a finite differencing operator of order p, this technique is called Fused LASSO [10] and yields a piecewise polynomial approximation of $\mathbf{v}$ of degree $(p-1)$. One can also consider the version when $\mathbf{y}$ is replaced with a multivariate time series $\mathbf{X}, \mathbf{V} \in \mathbb{R}^{D \times T}$:

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2 + \lambda \|\mathrm{vec}(\mathbf{V}\mathbf{Q})\|_1. \tag{8}$$

Change points can be localised in the different components and by assuming a coupled system, change points may be co-localised in time. The formulation that allows for such group-sparsity is the $\ell_{1,2}$ norm of Group LASSO:

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2 + \lambda \sum_{t=1}^{T-p} \|\mathbf{V}\mathbf{Q}_{\cdot,t}\|_2 \tag{9}$$

for order of differentiation p (with $M = T - p$) [1].

3 Methods

3.1 Problem Formulation

We assume that there is a piecewise polynomial trajectory in a multi-dimensional space, like the motion of a robotic arm in configuration space. We also assume that the robot executes certain actions one-by-one, e.g., it is displacing objects giving rise to sharp changes in the trajectory. However, we are not aware of the plan and we have no additional information and thus, we do not know the points in time or space when the trajectory would switch. These points will be called *switching points* or *change points*. Yet we know that such change points of different configuration components are co-localised in time (i.e., the plan spans across dimensions).

We also assume that at random times an anomaly disturbs the motion, but the system compensates, reverts back to the trajectory, and executes the task.

We propose to use the following $\ell_{1,2}, \ell_{1,2}$ convex optimisation problem for the detection of the above kind of anomalies. With known $\mathbf{X} \in \mathbb{R}^{D \times T}$ and variables $\mathbf{V}, \mathbf{S} \in \mathbb{R}^{D \times T}$, solve

$$\min_{\mathbf{V},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{V} - \mathbf{S}\|_F^2 + \lambda \sum_{t=1}^{T-p} \|\mathbf{V}\mathbf{Q}_{\cdot,t}\|_2 + \mu \sum_{t=1}^{T} \|\mathbf{S}_{\cdot,t}\|_2, \tag{10}$$

where $\mathbf{Q}$ is a finite differencing operator. For the sake of simplicity, we assume that order of $\mathbf{Q}$ is $p = 2$:

$$\mathbf{Q}_{d,t} = \begin{cases} 1, & \text{if } d = t \text{ or } d = t + 2, \\ -2, & \text{if } d = t + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This is a combination of the Robust PCA and the Group Fused LASSO tasks: it tries to find a robust piecewise linear approximation $\mathbf{V}$ and additive error term $\mathbf{S}$ for $\mathbf{X}$ with group-sparse switches *and* group-sparse anomalies that contaminate the unknown plan. The second term says that we are searching for a fit, which has group-sparse second-order finite differences. The third term ($\mathbf{S}$) corresponds to gross-but-group-sparse *additive* anomalies not represented by the switching model.

We also use the corresponding ℓ_1, ℓ_1 variant as a comparison:

$$\min_{\mathbf{V},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{V} - \mathbf{S}\|_F^2 + \lambda\|\mathrm{vec}(\mathbf{VQ})\|_1 + \mu\|\mathrm{vec}(\mathbf{S})\|_1, \tag{12}$$

which allows change points and anomalies to occur independently in the components (via ordinary sparsity instead of group-sparsity).

We used Matlab R2014b with CVX 3.0 beta [6, 7] for minimisations, capable of transforming cost functions to equivalent forms that suit fast solvers, e.g., the Splitting Conic Solvers (SCS) 1.0 [8] and used the sparse direct linear system option.

Dependencies on the (λ, μ) tuple were searched on the whole data set within the domain of $\{2^{-9}, 2^{-7}, \ldots, 2^{15}\} \times \{2^{-16}, 2^{-14}, \ldots, 2^{8}\}$. The best values were selected according to the F_1 score, the harmonic mean of precision and sensitivity: $F_1 = \frac{2\,\mathrm{TP}}{(2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN})}$, where TP, FP, FN are the number of true positives, false positives and false negatives for anomalous segments, respectively. Predicted momentary values (i.e., norms of the gross-but-group-sparse additive error term: $\|\mathbf{S}_{\cdot,t}\|_2$, $t = 1, \ldots, T$) were thresholded into momentary binary labels with respect to 0.01. Note that these labels have many spikes and gaps between them, thus a morphological closing operator was used to fill-in such gaps up to 1 s length during post-processing, resulting in the segmented labels of the F_1 calculations. A baseline algorithm using internal torque information—not available for our methods—together with some differencing heuristics and parameter optimisation was also added for a rough orientation about performance.

3.2 Data Set

We used 300 trials of 7-dimensional joint configurations of one arm of a Baxter research robot[4] as our data set. The configuration was characterised by 2 shoulder, 2 elbow and 3 wrist angles. The transitions between prescribed configurations were realised as an approximately piecewise linear time series with common change points (due to some minor irregularities of the position controller). Anomalies were generated by manual intrusion into the process: an assistant for the experiment kept hitting the robot arm when asked, as depicted in Fig. 1. The controller reverted back to the original trajectory as soon as possible. Timestamps of 822 collision commands were logged and served as ground-truth labels for the anomalies. The actual impacts were delayed by up to 5 s in the data because of human reaction time. We took this uncertain delay into consideration

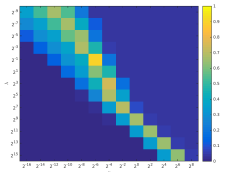
[4] `http://www.rethinkrobotics.com/baxter-research-robot/`

when computing the F_1 performance metric: we tried to pair detected anomalous segments with the timestamps provided with respect to the 5 s threshold. The data were recorded at 800 Hz frequency and interpolated uniformly to 50 Hz.

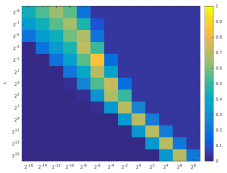
Fig. 1. Frame series of the operator hitting the robot arm.

4 Results

Parameter dependencies with respect to the mean F_1 score for the $\ell_{1,2}, \ell_{1,2}$ (10) and the ℓ_1, ℓ_1 (12) methods are shown in Fig. 2 (a) and (b), respectively. The best



(a)



(b)

$\ell_{1,2}, \ell_{1,2}$	0.90
ℓ_1, ℓ_1	0.86
Baseline	0.79

(c)

Fig. 2. Mean F_1 scores with different (λ, μ) parameters for (a): $\ell_{1,2}, \ell_{1,2}$ (10), (b): ℓ_1, ℓ_1 (12) algorithms. (c): Best achieved mean F_1 scores including baseline heuristics.

parameter combination for both schemes was $(\lambda, \mu) = (2^{-1}, 2^{-6})$. With these settings, the former algorithm achieved 709 true positive, 94 false positive and 113 false negative anomalous segments. Figure 2 (c) includes the best achieved mean F_1 scores for all approaches, including the heuristic baseline procedure.

An example highlights prediction scenarios for the $\ell_{1,2}, \ell_{1,2}$ method in Fig. 3: (a) shows the approximately piecewise linear input time series, with markers indicating common change points on each curve, as well as 5 anomalies pointed out by red rectangles; (b) shows the output gross-but-group-sparse norm values of our procedure, the 0.01 threshold level and the logged green ground-truth anomaly labels; while (c) provides close-up views of the actual and predicted anomalous segments. Anomalies 3 to 5 (around time points 700, 1160 and 1440) are true positives, as they are paired with ground-truth labels and are above the threshold. Anomaly 2 (around 420) is marked false negative, as it is improperly

thresholded (there are 71 examples for this in the entire data set). Anomaly 1 (around 80) is stigmatised false positive, as it lacks the ground-truth label, while an anomaly is certainly present in the segment due to controller imprecision (the total number of such cases is 17). Mean F_1 scores would be somewhat higher with more sophisticated thresholding and controller policies.

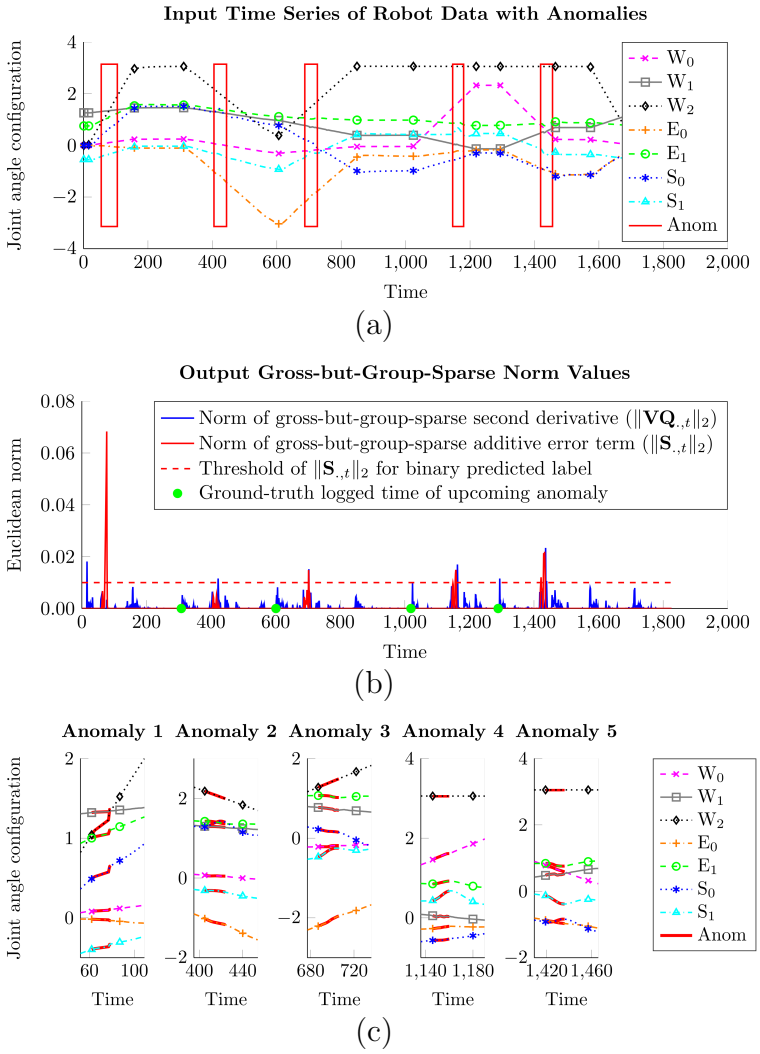


Fig. 3. Prediction scenarios for the $\ell_{1,2}, \ell_{1,2}$ method. (a): Input time series of joint angles. W: wrist, E: elbow, S: shoulder; Red rectangles: anomalies. (b): Blue (red): estimated change points (anomalies); red dashed line: optimal threshold level; green dots: ground-truth anomaly labels. (c): Close-ups for actual and detected (red) anomalous segments: Anomaly 1/2: false positive/negative, 3/4/5: true positives.

5 Conclusion

We showed that sparse methods can find anomalous disturbances affecting a dynamical system characterised by stochastic switches within a parametrised family of behaviours. We presented a novel method capable of locating anomalous events without supervisory information. The problem was formulated as a convex optimisation task. The performance of the algorithm was evaluated in a robotic arm experiment. Although the introduced anomalies were similar to the switching between behaviours, the procedure could still identify the disturbances with high F_1 values, outperforming two baseline approaches.

The piecewise polynomial approximation may be generalised further to more complex, e.g., autoregressive behavioural families and may be extended with more advanced post-processing techniques.

Our approach is motivated by the following: anomalies are sparse and in turn, sparsity based methods are natural choices for the discovery of not yet modelled events or processes; and that the early discovery of anomalous behaviour is of high importance in complex engineered systems.

Acknowledgments. Supported by the European Union and co-financed by the European Social Fund (TÁMOP 4.2.1./B-09/1/KMR-2010-0003) and by the EIT Digital grant on *CPS for Smart Factories*.

References

1. Bleakley, K., Vert, J.P.: The group fused Lasso for multiple change-point detection. arXiv:1106.4199 (2011)
2. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J of ACM 58(3), 11 (2011)
3. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Tr. on Inf. Theo. 51(12), 4203–4215 (2005)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection. ACM Comp. Surv. 41, 15:1–58 (2009)
5. Donoho, D.L.: Compressed sensing. IEEE Tr. on Inf. Theo. 52(4), 1289–1306 (2006)
6. Grant, M., Boyd, S., Ye, Y.: CVX. Recent Adv. Learn. Cont. pp. 95–110 (2012)
7. Grant, M.C., Boyd, S.P.: Graph implementations for nonsmooth convex programs. In: Recent advances in learning and control, pp. 95–110. Springer (2008)
8. O'Donoghue, B., Chu, E., Parikh, N., Boyd, S.: Operator splitting for conic optimization via homogeneous self-dual embedding. arXiv:1312.3039 (2013)
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B pp. 267–288 (1996)
10. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Roy. Stat. Soc. B 67(1), 91–108 (2005)
11. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B 68(1), 49–67 (2006)