Unsupervised anomaly detection using optimal transport for predictive maintenance

Amina Alaoui-Belghiti, Sylvain Chevallier, Eric Monacelli

Unsupervised anomaly detection using optimal transport for predictive maintenance

Amina Alaoui-Belghiti[1,2], Sylvain Chevallier[2,3], and Eric Monacelli[2,3]

[1] Nexeya, Massy, France
amina.alaouibelghiti@nexeya.com
[2] UVSQ, Universite Paris-Saclay
(sylvain.chevallier,eric.monacelli)@uvsq.fr
[3] LISV, Universite Paris-Saclay

Abstract. Anomaly detection is of crucial importance in industrial environment, especially in the context of predictive maintenance. As it is very costly to add an extra monitoring layer on production machines, non-invasive solutions are favored to watch for precursory clue indicating the possible need for a maintenance operation. Those clues are to be detected in evolving and highly variable working environment, calling for online and unsupervised methods. This contribution proposes a framework grounded in optimal transport, for the specific characterization of a system and the automatic detection of abnormal events. This method is evaluated on acoustic dataset and demonstrate the superiority of metrics derived from optimal transport on the Euclidean ones. The proposed method is shown to outperform one-class SVM on real datasets, which is the state-of-the-art method for anomaly detection.

Keywords: Predictive maintenance · Optimal transport · Anomaly detection · Unsupervised learning.

1 Introduction

In the industrial field, the equipment operational readiness is a central and unavoidable challenge. This is a key parameter in the production budgeting or in the solution development projects. The concept of operational readiness mainly covers the preventive maintenance which corresponds to systematic or periodic operations defined for each type of equipment, and/or the corrective maintenance which concerns operations to remedy the damage that has occurred during operation. But with these two kinds of maintenance, some unnecessary actions are performed or the equipment breaks down unexpectedly, moreover, these two kinds of maintenance do not take into account the specific characteristics of real operating conditions [5].

The predictive maintenance anticipates the needs and remedies its weaknesses by reducing prevention costs and avoiding unplanned downtime [15]. It becomes increasingly important in recent years, and is linked to terms such as the Internet of Things(IOT), big data and industry 4.0.

Predictive maintenance requires continuous monitoring of the components directly on the machine or on the installation, such as for wind turbines [8], or for aircraft engines [1], power plants, solar fields, etc. This is achieved by sensors that report information about the periphery and environmental characteristics such as temperature, vibration, noise or humidity, but also by looking at the historical and conceptual data of the equipment. These data are then recorded and evaluated by correlations, compared with models through well-targeted algorithms. In particular, anomaly detection aims to identify failing components and replace them before the occurrence of a failure.

A crucial step in predictive maintenance is the anomaly detection. It is a major challenge in industrial applications, especially for the identification of manufacturing defects or component failure. Whichever method is used, an anomaly is always defined, either implicitly or explicitly with respect to a model. The choice of the method and of the model depends entirely on the context, the objective, the available data, and their properties.

Theoretically, an observation is considered as abnormal or atypical compared to a given model, this model can be parametric, assuming a specific statistical law distribution, or non-parametric defined directly from the data. Without any assumption on the law governing the data, the model can be chosen in relation with the presence of a target variable to explain, to model, or to predict by regression or discrimination. In the opposite case, it can be related to presence of the probability density or multidimensional distribution of variables.

In the parametric cases, the law of the explanatory variables, or that of the residuals, is usually supposed to be Gaussian multidimensional. Relative to a target variable model, when considering a Gaussian linear model, the atypical or poorly adjusted observations are those diverging from the estimated variance [2]. Nevertheless, the detection of atypical observations is more similar to that of influential observations, for example using Cook's distance, well known in simple linear regression [3].

Without target variable to model, the abnormality at a multidimensional density can be characterized by the Mahalanobis distance, as used in [1] for the detection of anomalies in an aircraft engine. This distance is estimated from the inverse of the empirical covariance of the distribution.

In the non-parametric cases, there is no assumption about the multidimensional distribution of variables, it is estimated locally in different ways. Relative to a target variable model, Random forest [17] includes an original solution adapted to the taking into account of mixed variables.

Without target variable to model, this is the case with the most extensive literature with many proposed methods. A comprehensive review is proposed in [10], most notable methods are Local Outlier Factor [4] and Isolation Forest [11]. Nonetheless, these methods are applicable in situations where abnormal observations are frequent and where all data are corrupted with a low density noise. Here, the situation is different as the monitored components are first assumed to operate correctly. The abnormal events could be identified as a novelty, a sensible variation from the clean reference data. The most know algorithm in novelty detection is the One-Class SVM [13, 12], which is the current state of the art.

The shift in the observed distribution could be quantified, calling for metrics and topological spaces for handling probability measures [16]. These metrics follow from the Kantorovitch problem, that define a transport plan (couplings) which is the solution of transportation problem defined by a cost matrix. Common applications are image processing, shape interpolation, similar terms in text documents [14]. The computation of this transport plan often requires a large amount of computational resources and is not feasible in high dimensions.

Recent advances yield new algorithms for fast computation of transportation distance, such as the Sinkhorn distance [6]. Adding entropic regularization to transportation distance, the Sinkhorn distance has several interesting properties as it is a scale free, non-Euclidean formulation which is less subject to the curse of dimensionality. Also, clever implementation that makes use of parallelization and GPU computation are available [9].

In this contribution, we propose to formulate a straightforward algorithm for anomaly detection in a predictive maintenance context. We consider the case of a system to monitor with non-invasive sensors, where a large amount of data from a normal behavior is available. This algorithm should be able to identify novel behavior or abnormal samples, that will be eventually process to determine if a predictive maintenance is required. Drawing on optimal transport framework, our contributions are the following:

- A novel unsupervised method for anomaly detection
- Evaluation with Euclidean methods on toy data set
- Application to real data, comparison with state-of-the-art
- Possible extension for online implementation

The rest of this paper is organized as follows. Section 2 describes the proposed model and provides a formal description of the system. In Sect. 3, the model is evaluated on a toy data set and on real acoustic data set and compared with Euclidean methods and one-class SVM. The Sect. 4 concludes this paper.

2 Optimal transport for anomaly detection

This section first provides some insight on transport algorithms, with details on the Sinkhorn algorithm. The rest of the section goes through the explanation on the proposed algorithm.

2.1 Optimal transport with entropic regularization

We consider two metric spaces $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{M}(X)$ is the set of Radon probability measures on $\mathcal{X}$. Whereas we focus on discrete measures, the proposed approach could be extended to the continuous case without any specific problem. With discrete measure α with weight a defined on $\mathcal{X}$ in n locations and β

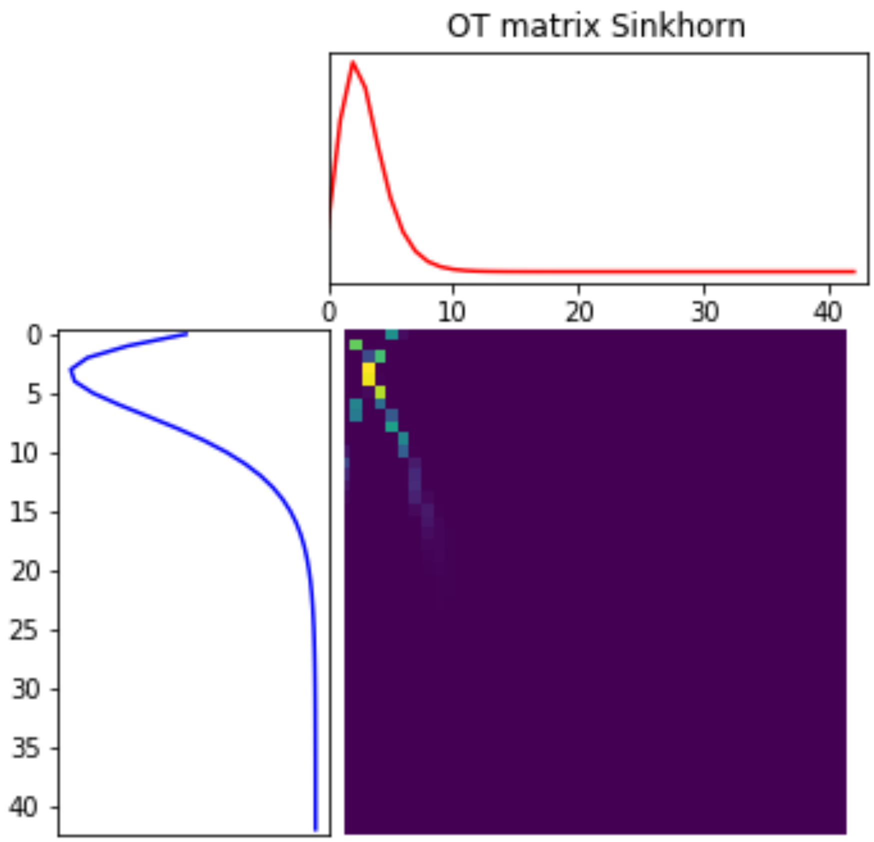

Fig. 1. Illustration of a coupling P solving the optimal transport from measures a on left to measures b on top.

with weight b defined on $\mathcal{Y}$ in m locations, we could consider the set of coupling matrices defined by:

$$U(a,b) = \left\{P \in \mathbb{R}_+^{n\times m} : P\mathbf{1}_m = a \text{ and } P^T\mathbf{1}_n = b\right\} \tag{1}$$

where $\mathbf{1}_n$ is the vector of ones with n dimensions.

If we consider a cost matrix $C \in \mathbb{R}^{n\times m}$, the cost of the mapping from a to b using a transport matrix P is $\langle P, C\rangle$, with $\langle\cdot,\cdot\rangle$ is the Euclidean dot product.

$$d_C(a,b) = \min_{P\in U(a,b)} \langle P, C\rangle, \tag{2}$$

is called the optimal transport problem between a and b given cost C.

Defining the entropy of the coupling matrix as

$$H(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1), \tag{3}$$

it is possible to add an entropic regularization term to the optimal transport problem (2):

$$d_C^{\epsilon}(a,b) = \min_{P \in U(a,b)} \langle P, C \rangle - \epsilon H(P). \tag{4}$$

This problem admit a unique solution of the form $P_{i,j} = u_i K_{i,j} v_j$, with scaling variables u_i and v_j belonging respectively to $\mathbb{R}_+^n$ and $\mathbb{R}_+^m$. When conforming to mass conservation, the solution of (4) could be written as:

$$u * (Kv) = a \text{ and } v * (K^T u) = b \tag{5}$$

where $*$ is the element-wise vector multiplication. By alternatively solving the following update, an iteration l of the Sinkhorn algorithm is:

$$u^{(l+1)} = \frac{a}{Kv^{(l)}} \text{ and } v^{(l+1)} = \frac{b}{K^T v^{(l)}}, \tag{6}$$

starting with an arbitrary initialization.

An illustration of the obtained coupling when solving Problem 4 is shown on Fig. 1

2.2 Anomaly detection with optimal transport

The proposed detection algorithm is designed to model normal behaviors and identify abnormal behavior. We focus in this contribution on acoustic signals or vibration, as it is easy to embed such sensors in a non-invasion monitoring system. Nonetheless, the described approach could be adapted to various problems, as any sample distribution could act as an input feature. An abnormal signal can be defined by the distance between its own representation and the representation of another signal defined as a reference. Noisy signals show the highest distance from the reference representation.

Assuming a set of initial signals $\mathbf{X} = \{X_i\}_{i=1\ldots k}, X \in \mathbb{R}^t$, and a signal containing abnormal events, denoted $\tilde{X} = X + \eta N$, where N is the abnormal component and η is the component level. The signals are considered in the frequency domain, by estimating the power spectral density. For each signal, the power spectral density is evaluated with the Welch estimator $F(\cdot)$, that is the signals are split into several partially overlapping segments, a windowing function is applied (here Hamming) before computing their Fourier transform, and averaging the results for each segment. The resulting signals are $F(X) \in \mathbb{R}^n$.

The initial signals are then averaged, using a Euclidean mean, to obtain a barycenter acting as reference $F(\bar{\mathbf{X}}) = \frac{1}{k}\sum_k F(X_k)$. During preliminary tests, we have evaluated the opportunity to use Wasserstein barycenter [7] but it turned out that it yield poorer results. The distances from individual PSD

Algorithm 1 Anomaly detection with OT

Input: set of reference signals $\mathbf{X}$, signal to evaluate $\hat{X}$
Output: binary classification, 1 if normal signal, -1 if abnormal

procedure PREDICT ANOMALY($\mathbf{X}, \hat{X}$)
  $F(\bar{\mathbf{X}}) \leftarrow \frac{1}{k}\sum_k F(X_k)$
  for i in 1 .. k **do**
    $d_i \leftarrow d_C^{\epsilon}(F(\bar{\mathbf{X}}), F(X_i))$
  Set threshold ϑ from LogNormal fit on $\{d_i\}_{i=1 \ldots k}$
  $\hat{d} \leftarrow d_C^{\epsilon}(F(\bar{\mathbf{X}}), F(\hat{X}))$
  if $\hat{d} > \vartheta$ **then**
    return -1 ▷ Anomaly
  else
    return 1 ▷ Normal

$F(X_k)$ and the reference barycenter $F(\bar{\mathbf{X}})$ is computed with the Sinkhorn distance $d_C^{\epsilon}(F(\bar{\mathbf{X}}), F(X_k))$, using a Chebyshev cost function. The same is done for signals with anomalies $d_C^{\epsilon}(F(\bar{\mathbf{X}}), F(\tilde{X})$.

A distribution following a Log-Normal distribution is then obtained from the histogram of distances. This distribution allowed us to set a distance threshold which was used to design a classifier in order to predict the tested signal as strict less than normal or abnormal. The different steps of the algorithm are shown in Algorithm 1.

The evaluation of the method by two performance metrics and its comparison with an Euclidean baseline and the one-class SVM is detailed in the following section.

3 Experimental analysis

3.1 Datasets description

This study is a part of a ongoing project aiming to implement a predictive maintenance solution based on the test bench monitoring of an electro-mechanical aircraft actuator. Unfortunately experiments related to this test bench are confidential and dataset from this experiment could not be shared.

To ensure the reproducibility of this study, we decided to apply our approach on public data. A first batch of data is chosen to demonstrate the robustness of the method. It also offers the possibility to compare with a Euclidean approach, to verify that the optimal transport distance has a real positive influence. We selected a sound recording of a working day in an open space[4] as reference data and different levels of pink noise are blended to simulate noisy data in a controlled way.

To further test our approach on realistic data, we also evaluated the method on a dataset similar to the private, industrial data. We selected a recording of

[4] The selected sound recordings are available upon request.

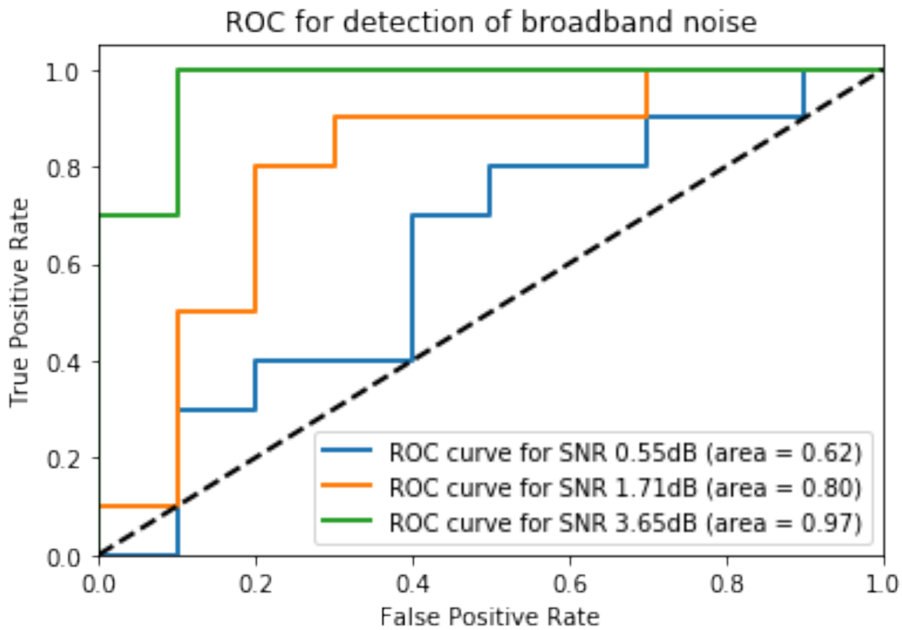

Fig. 2. ROC estimation for detecting peak noise

the rotating industrial machine sound for the reference acoustic signal, which is close to the one encountered in the industrial situation. We also selected some recordings of some abnormal event sounds to represent the anomalies that are also similar to those occurring on test bench[5]. The results of these experiments are presented in the following subsections.

3.2 Robustness and comparison with Euclidean metrics

To test the robustness of the anomaly detection method, it is evaluated on a monoraul track recording encoded at 44100 Hz of office open space sounds. The recording is 15 minutes long. The reference is computed on the first 7 minutes, the remaining audio is either left as it or corrupted with pink noise. Two different types of abnormal noise are considered: a large broadband noise is added to the signal or a sharp peak noise.

The algorithm of Alg. 1 is evaluated with various thresholds to compute a ROC estimation. As it is shown on Fig. 2, when the power of the noise signal is strong (SNR 3.65 dB), the algorithm easily detect the anomaly in the signal. As expected, the detection is harder when the noise peak is weaker, with a SNR of 0.55, the Area Under Curve (AUC) is 0.62 which is above chance level (here 0.5) but of limited precision.

[5] All of these recordings are available as well upon request

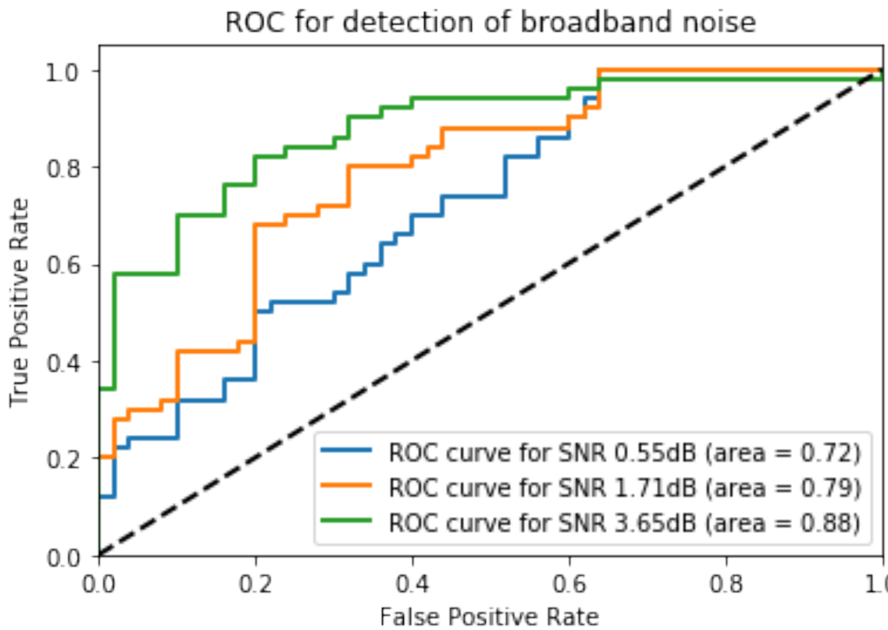

Fig. 3. Detection of broadband noise, SNR values are matching those of the peak noise experiment.

The second experiment rely on the same kind of experiment, but with a broadband noise. The Fig. 3 show that with equivalent SNR, the proposed algorithm is able to detect anomalous event with more easiliy than in the case of the peak noise. This reflects on the AUC values which are of 0.72 for a low SNR of 0.55 dB, 0.79 for SNR of 1.71 dB and 0.88 for 3.65 dB.

This experiment on a toy dataset demonstrates the feasibility of anomaly detection with optimal transport metrics. These results show that it is possible to detect a noise when the algorithm is calibrated with real sound. The proposed algorithm shows a limited sensitivity to anomalies concentrated on a narrow frequency peak but performed well with broadband change, even if the modifications are subtle.

3.3 Evaluation on real dataset

In these experiments, a reference signal similar to the noise of an industrial bench test is chosen. The signal is recorded in monaural, at 44100 Hz for 15 minutes long. Two qualitatively different kinds of faulty mechanical parts are considered: the sound of a light, high pitched whistling (dataset 1) and a cyclical low-pitched sound, similar to a faulty ball bearing (dataset 2). These two datasets are also 15 minutes long of 44100 Hz monaural recording.

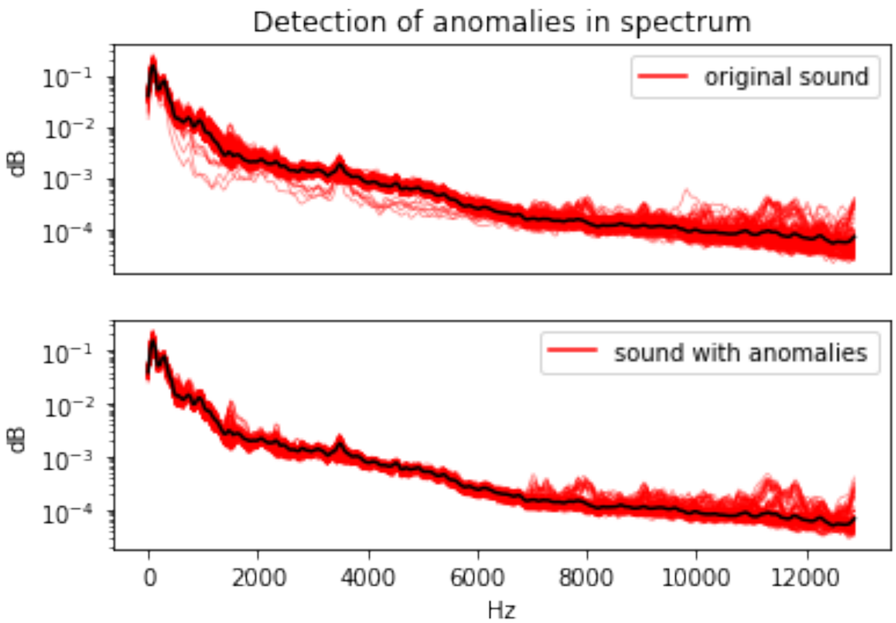

Fig. 4. Spectrum of the reference dataset is on the top plot. The PSD of the combined reference and anomalous sound is shown on the bottom.

The algorithm described in Alg. 1 is compared to a Euclidean baseline. In this baseline, this is the same algorithm but the Sinkhorn metrics are replaced with a Euclidean one on lines 4 and 6.

The state-of-the-art algorithm, one-class SVM, for novelty detection is benchmarked against the proposed algorithm. The one-class SVM relies on a Radial Basis Function (RBF) kernel and the ν value is set to reflect the ratio of anomalies.

The datasets are separated in training (500 samples) and test data (500 samples) using a repeated k-fold split, the training data are used to train the SVM and the compute the reference signal $F(\bar{\mathbf{X}})$. The algorithms (SVM, OT and Euclidean) are then evaluated on test data: half of the test data are mixed with the faulty mechanisms sound of dataset 1 and 2.

To illustrate the input features of the algorithm, the Fig. 4 shows the PSD of the 500 training signal on the top plot. The reference computed on these training signal is shown as a thick black line. On the bottom plot, one could see the mix of the 250 test signal unmodified and of the 250 test signal mixed with the anomalous sound of dataset 1. The reference indicated in black is the same as the one on the top plot.

The results are shown on Fig. 5, for the datasets 1 with light, high pitched whistling anomalies. As the results are qualitatively similar for dataset 2, the results for the dataset 1 and 2 are summarized in the Table 1. The left part

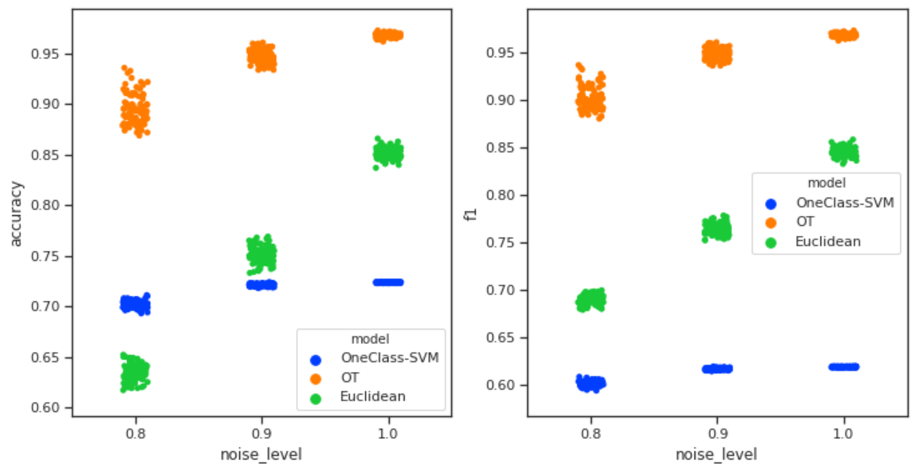

Fig. 5. Accuracy and F1-measure estimated on the first dataset for different noise levels. One-class SVM, Optimal Transport (OT) and Euclidean baseline are evaluated on this dataset.

of Fig. 5 shows the accuracy for 3 different noise levels. The one-class SVM (OCSVM) achieves good results but misses several anomalies lowering its score around 70-75 %. The Euclidean method demonstrates lower performance than OCSVM for low SNR but outperforms OCSVM for high SNR. The optimal transport algorithm yields the higher results, with 90-98 % accuracy.

We also evaluated the F1-score to take into account the precision and the recall for the anomaly. These scores are shown on the right part and one could see that the optimal transport outperforms all the methods.

4 Discussion and conclusion

Anomaly detection is a complex problem with no uniformly better solution because the choice of which method to use depends largely on the context, the properties of the variables and data observed, and also on the objective pursued. This paper focuses on a new method of anomaly detection in acoustic signals by comparing them with reference signals through the calculation of Sinkhorn distances in optimal transport. The choice fell on an unsupervised detection method, because there are no anomalies a priori defined, common in industrial equipment operating mode. This allows us an online implementation in order to an industrial deployment.

In the experimental analysis section, the robustness of the OT method has been demonstrated by observing the ROC estimation. It distinguished itself from the One Class-SVM and Euclidean methods by showing a high level of accuracy for different datasets through its evaluation by two performance metrics, accuracy and F1 measures.

		Accuracy			F1		
	Noise level	OT	OCSVM	Baseline	OT	OCSVM	Baseline
Dataset 1	0.8	**0.89**	0.70	0.63	**0.90**	0.60	0.69
	0.9	**0.95**	0.72	0.75	**0.95**	0.62	0.76
	1.0	**0.97**	0.72	0.85	**0.97**	0.62	0.84
Dataset 2	1.0	**0.54**	0.5	0.5	**0.90**	0.60	0.62
	1.5	**0.64**	0.5	0.5	**0.94**	0.61	0.62
	2.0	**0.84**	0.5	0.5	**0.97**	0.62	0.62

Table 1. Accuracy and F1-measure for two datasets of acoustic recording, corrupted by faulty mechanical sounds.

As already discussed before, this study is part of a larger PhD position project that is the predictive and automated maintenance of an industrial equipment. The overall architecture of the project is divided into several parts starting with the monitoring of the equipment until the alarms propagation and the proposal of the maintenance actions to be carried out, while passing through the anomaly detection and the fault identification.

The contribution of this paper has been devoted mainly to acoustic data, but we want to broaden the study of anomaly detection in future work on other perception mechanisms set up for maintenance, such as vibratory and thermographic measurements with well-defined specifications, while integrating the confidence interval questions for each type of measurement.

References

1. Abdel-Sayed, M., Duclos, D., Faÿ, G., Lacaille, J., Mougeot, M.: Dictionary comparison for anomaly detection on aircraft engine spectrograms. In: Machine Learning and Data Mining in Pattern Recognition, pp. 362–376. Springer (2016)
2. Aguinis, H., Gottfredson, R.K., Joo, H.: Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods **16**(2), 270–301 (2013)
3. Atkinson, A., Riani, M.: Robust diagnostic regression analysis. Springer Science & Business Media (2012)
4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM sigmod record. vol. 29, pp. 93–104. ACM (2000)
5. Camci, F.: System maintenance scheduling with prognostics information using genetic algorithm. IEEE Transactions on reliability **58**(3), 539–552 (2009)
6. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in neural information processing systems. pp. 2292–2300 (2013)
7. Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. In: International Conference on Machine Learning. pp. 685–693 (2014)
8. Garcia, M.C., Sanz-Bobi, M.A., del Pico, J.: Simap: Intelligent system for predictive maintenance: Application to the health condition monitoring of a windturbine gearbox. Computers in Industry **57**(6), 552–568 (2006)
9. Genevay, A., Cuturi, M., Peyré, G., Bach, F.: Stochastic optimization for large-scale optimal transport. In: Advances in Neural Information Processing Systems. pp. 3440–3448 (2016)

10. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one **11**(4), e0152173 (2016)
11. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422. IEEE (2008)
12. Ma, J., Perkins, S.: Time-series novelty detection using one-class support vector machines. In: Neural Networks, 2003. Proceedings of the International Joint Conference on. vol. 3, pp. 1741–1745. IEEE (2003)
13. Manevitz, L.M., Yousef, M.: One-class svms for document classification. Journal of machine Learning research **2**(Dec), 139–154 (2001)
14. Peyré, G., Cuturi, M., et al.: Computational optimal transport. Tech. rep. (2017)
15. Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A.: Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics **11**(3), 812–820 (2015)
16. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)
17. Zhang, J., Zulkernine, M., Haque, A.: Random-forests-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **38**(5), 649–659 (2008)