This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Anomaly Detection Based on Compressed Data: An Information Theoretic Characterization

Alex Marchioni, *Member, IEEE*, Andriy Enttsel, *Graduate Student Member, IEEE*, Mauro Mangia, *Member, IEEE*, Riccardo Rovatti, *Fellow, IEEE*, and Gianluca Setti, *Fellow, IEEE*

***Abstract*—Large monitoring systems produce data that is often compressed to be transmitted over the network. For latency or security reasons, compressed data may be processed at the edge, i.e., along the path from sensors to the cloud, for some purposes such as anomaly detection. However, the performance of a detector distinguishing between normal and anomalous behavior may be affected by the loss of information due to compression. We here analyze how lossy compression affects the performance of a generic anomaly detector. This relationship is formalized in terms of information-theoretic quantities. Within such a framework we leverage a Gaussian assumption to derive analytical results regarding the importance of white noise as a representative of both the average and asymptotic anomalies. Moreover, in an anomaly-agnostic scenario, we also show the existence of a level of compression for which an anomaly is undetectable though compression is not completely destructive. Numerical evidence confirms that the proposed information-theoretic quantities anticipate the performance of practical compressors and detectors in the case of Gaussian and non-Gaussian signals allowing an assessment of the tradeoff between compression and detection.**

***Index Terms*—Anomaly detection, edge computing, Internet of Things, lossy compression, rate-distortion theory.**

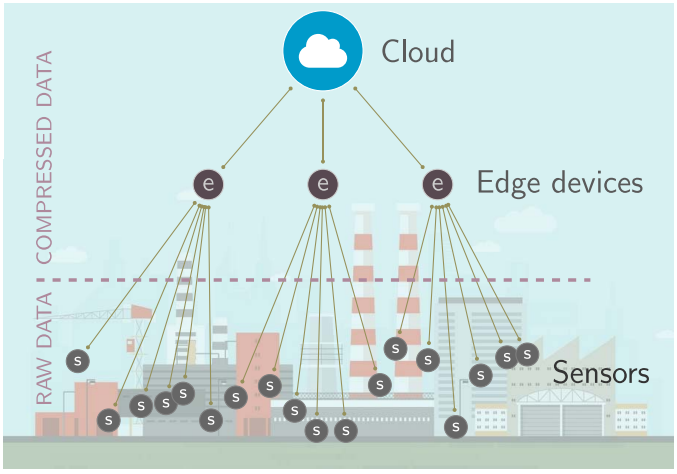

Fig. 1. Sensorized plant whose compressed acquisitions are aggregated at the edge before being sent to the cloud.

I. Introduction

A TYPICAL scenario for nowadays massive acquisition systems can be modeled as a large number of sensing units, each transforming some unknown physical quantity into samples of random processes that are then transmitted over a network. To reduce transmission bitrate, the sensor readings are often compressed by a lossy mechanism that aims to preserve the useful information [1]. Implicitly, this brings out a tradeoff between transmission bitrate and the amount of information loss which is addressed in general terms by the Information Theory with the optimal rate-distortion curve [2], i.e., the theoretical lower bound for the rate given a maximum distortion level.

In any practical application, a compression mechanism corresponds to a proper rate-distortion curve while the compression level imposed by the final application identifies a point on that curve. This is the case of a large variety of applications that include the extraction and the monitoring of features characterizing the object under observation as in structural health monitoring [3], [4], sensorization of industrial plants [5], [6], [7], or the processing biomedical signals [8], [9], [10] (e.g., Fig. 1 for an example). In general, the lower the signal distortion, the more likely to meet the requirements of a main task.

A common practice is to adopt a remote unit to process and store compressed data [1], [6], [11]. Before reaching a possible cloud facility, the corresponding bitstreams may traverse several levels of hierarchical aggregation and intermediate devices that are often indicated as the *edge* of the cloud [12].

Manuscript received 22 May 2023; accepted 13 July 2023. This work was supported in part by PNRR-M4C2-Investimento 1.3, Partenariato Esteso—"Future Artificial Intelligence Research (FAIR)"—Spoke 8 "Pervasive AI," funded by the European Commission through the NextGeneration EU Programme under Grant PE00000013. This article was recommended by Associate Editor M. Dotoli. *(Corresponding author: Alex Marchioni.)*

Alex Marchioni, Andriy Enttsel, and Mauro Mangia are with the Department of Electrical, Electronic, and Information Engineering, University of Bologna, 40136 Bologna, Italy, and also with the Advanced Research Center on Electronic Systems, University of Bologna, 40125 Bologna, Italy (e-mail: alex.marchioni@unibo.it; andriy.enttsel@unibo.it; mauro.mangia@unibo.it).

Riccardo Rovatti is with the Department of Electrical, Electronic, and Information Engineering, University of Bologna, 40136 Bologna, Italy, also with the Advanced Research Center on Electronic Systems, University of Bologna, 40125 Bologna, Italy, and also with the Alma Mater Research Institute for Human-Centered AI, University of Bologna, 40015 Bologna, Italy (e-mail: riccardo.rovatti@unibo.it).

Gianluca Setti is with CEMSE, King Abdullah University of Science and Technology, Jeddah 23955, Saudi Arabia, on leave from the Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy (e-mail: gianluca.setti@kaust.edu.sa).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2023.3299169.

Digital Object Identifier 10.1109/TSMC.2023.3299169

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

For latency, privacy, or cybersecurity reasons, some computational tasks may benefit from their deployment at the edge. One of those tasks is the detection of anomalies.[1]

Usually, compression schemes applied to sensor data are asymmetric and entail a lightweight encoding performed on very-low complexity devices paired with a possibly expensive decoding stage running on the cloud. In these conditions, it is sensible that anomaly detectors working on the edge access compressed data and not the raw signal. Yet, lossy compression bases its effectiveness on neglecting some of the signal details. This translates into a distortion between the original and the compressed signal but also in a loss of features that, in principle, could have been used to tell normal behaviors from anomalous ones.

In general, acquisition systems must obey a distortion constraint so that they are designed to best address the tradeoff between compression and distortion. However, such a tradeoff goes in parallel to the one between distortion and the ability to determine whether the signal is normal or anomalous. Here, we analyze the latter with the same information theoretic machinery used in the well-known rate-distortion analysis and show that the two tradeoffs are different.

A. Related Works

How compression affects the ability of a detector to distinguish between two sources of information is a topic that has been investigated in the literature. In [13], the problem of hypothesis testing is discussed for a single source under a rate constraint. Such a basis has been extended to information-theoretic problems of statistical inference in the case of multiterminal data compression in [14]. With respect to the proposed framework, the authors do not include a constraint on distortion since original signals are not required to be reconstructed. On the contrary, we assume that compression is designed to guarantee the quality of service needed by the processing tasks that receive the reconstructed data.

In a sense, the framework we identify is partially related to the information-bottleneck scheme [15], [16]. In that scheme, the main tradeoff between rate and distortion is replaced by a very general criterion that identifies which part of the information content of the original signal must be preserved during compression, i.e., the compressor limits the rate by maintaining some features. In detail, the preserved features consist of the information contained in the original signal about a second suitably introduced signal matching the target application. For example, compression can also be adapted to anomaly detection. However, in our context, compression is independent of anomaly detection as it is designed to maximally preserve the signal information for further analysis (see Fig. 2). Hence, we augment the classical rate-distortion framework to add a quantity representing the distinguishability between normal and anomalous signals. Moreover, we also consider cases in which we completely ignore the statistic of the anomaly thus spoiling the applicability of the information-bottleneck scheme that can model anomaly detection only if we can identify anomalies with a second source of information.

The information-bottleneck principle has also been employed for unsupervised tasks. To tackle one-class classification, in [17] and [18] an optimization problem is considered in terms of rate-distortion tradeoff and it is solved by applying the information-bottleneck principle. However, as in previously reported contributions, this tradeoff is used for anomaly detection with no consideration about signal distortion affecting further applications, i.e., without any constraint on the error committed in the reconstruction of the original signal.

Another way to relate unsupervised anomaly detection to data compression is described in [19]. Here, the level of abnormality of a data point with respect to the entire data set, called *coding cost*, is given by its ability to be efficiently compressed in a Huffman coding fashion.

The same reason that differentiates our work from the information-bottleneck principle makes the analysis we propose different from other modifications of classical rate-distortion theory that substitute energy-based distortion with perceptive criteria [20], [21].

Though not overlapping with the problem we address, it is also worthwhile to mention [22], [23], in which it is assumed that the original signal is characterized by some parameters (e.g., the mean) and the authors clarify how the estimation of such parameters is affected by lossy compression.

Note also that other applications exist in which rate and distortion are paired with additional merit figures taking into account relevant features of the system. As an example [24] adds computational effort considerations to the analysis of rate-distortion of wavelet-based video coding.

Finally, even without emphasis on compression, the relationship between the analysis of suitably defined subcomponents of a signal to detect possible outlier behaviors is a classic theme that is still under investigation [25], [26], [27].

B. Our Contribution

In this article, we propose an analysis of the performance of a generic anomaly detector working on a signal distorted by compression mechanisms addressing the rate-distortion tradeoff. In particular, we consider ideal and real detectors working on signals compressed by ideal and real encoders designed to preserve the original information. To anticipate the performance of a detector trying to distinguish anomalous from common signals, we lay down a theoretical framework and define two information-theoretic measures of distinguishability to model anomaly-agnostic and anomaly-aware scenarios.

Specializing our analysis to the case of Gaussian sources.

1) The distinguishability metrics in case of white anomalies are representative of the average detector's performance over many different anomalies.
2) As the signal dimension increases, white anomalies tend to be typical.

[1]Depending on the context, an anomaly may also be referred to as outlier, novelty, intrusion, attack, etc.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

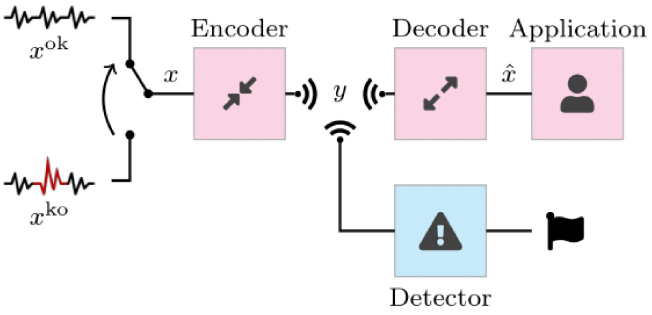

Fig. 2. Signal chain is tuned on the normal signal x^{ok} to best address the rate-distortion tradeoff, guaranteeing a certain quality of service to a given application. An anomalous signal x^{ko} may occur and a detector working on the compressed signal y should be able to detect it.

3) In the case of an anomaly-agnostic detector, the relation between distinguishability and distortion needs not to be monotonic and there may be at least one nondisruptive distortion level that makes the white anomaly undetectable.

We then present numerical evidence showing the effectiveness of the theoretical framework when Gaussian signals (GSs), the distortion-optimal compressor, and optimal detectors are considered.

Finally, we provide further numerical evidence when the assumptions under which the theory was developed are progressively relaxed, showing the effectiveness of the latter in modeling the behavior.

1) Real-world detectors monitoring GSs compressed by the encoder optimal in the rate-distortion sense.
2) Real-world detectors coping with GSs compressed by practical suboptimal encoders that adapt the compression mechanism to the statistical properties of the normal signal source.
3) Real-world detectors working on electrocardiogram (ECG) and accelerometer signals compressed by the same real-world suboptimal encoders.

Both theoretical and empirical results reveal how compression tuned to maximize the information transfer does not necessarily address at best the compromise with distinguishability.

This article is organized as follows. Section II reviews the classical rate-distortion theory to define the expression of the optimal compressor, while Section III defines the model for normal and anomalous signals together with the metrics to measure the distinguishability between them. Section IV lays down some theoretical results that are then confirmed with numerical evidence in Sections V and VI. The conclusion is finally drawn. Proofs of the theorems and lemmas are reported in the Appendix.

II. Rate Versus Distortion

Signals are compressed by encoding the information content into symbols which can be transmitted over a channel whose capacity is limited to a maximum rate of symbols per second. When the rate is not sufficient, the compression mechanism discards some information to fit the channel. This loss of information results in a receiver observing a signal that is distorted with respect to the original. Intuitively, the lower the channel capacity, the higher the distortion. This tradeoff is extensively studied in rate-distortion theory [2, Ch. 13].

We consider the context in which a system has the main task of transferring the information content of a signal source x to a receiver through a communication channel that has a constraint on rate. At any time instant t, an instance $x[t]$ is passed to an encoding stage producing a compressed version $y[t]$ that may then be decompressed into $\hat{x}[t] \in \widehat{\mathbb{R}^n}$, where $\widehat{\mathbb{R}^n} \subset \mathbb{R}^n$.

The constraint on the rate is such that it implies a lossy compression mechanism. The encoding stage is therefore not injective and introduces some distortion. The encoder is tuned on the source x, which is modeled as an independent discrete-time, n-dimensional stochastic process.

Distortion may be defined as follows:

$$D = \mathbf{E}\Big[\big\|x[t] - \hat{x}[t]\big\|^2\Big] \tag{1}$$

where $\mathbf{E}[\cdot]$ stands for expectation, and the minimum achievable rate ρ can be expressed as a function of the maximal accepted distortion δ as follows [2, Th. 13.2.1]:

$$\rho(\delta) = \inf_{f_{\hat{x}|x}} I\big(\hat{x}; x\big) \quad \text{s.t. } D \le \delta \tag{2}$$

where $I(\hat{x}; x)$ is the mutual information between $\hat{x}$ and x [2, Ch. 8], and $f_{\hat{x}|x}$ is a conditional probability density function (PDF) modeling the possibly stochastic mapping characterizing the encoder–decoder pair. Although [2, Th. 13.2.1] defines the rate-distortion function in the discrete case, it can also be proved for well-behaved continuous sources [2, Ch. 13] as considered in this work.

If the source is memoryless (thus allowing us to drop the time index t) and generates vectors of independent and zero-mean Gaussian variables, i.e., when $x \sim \mathcal{G}(0, \Sigma)$ where Σ is a diagonal covariance matrix such that $\Sigma = \operatorname{diag}(\lambda_0, \ldots, \lambda_{n-1})$ with $\lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_{n-1} \ge 0$, then the solution of (2) is

$$\rho = \frac{1}{2}\sum_{j=0}^{n-1} \log_2 \frac{\lambda_j}{\min\{\theta, \lambda_j\}} = -\frac{1}{2}\sum_{j=0}^{n-1} \log_2 \tau_j \tag{3}$$

$$\delta = \sum_{j=0}^{n-1} \min\{\theta, \lambda_j\} = \sum_{j=0}^{n-1} \lambda_j \tau_j \tag{4}$$

where $\theta \in [0, \lambda_0]$ is the so-called reverse water-filling parameter [2, Th. 13.3.3], and $\tau_j = \min\{1, \theta/\lambda_j\}$ accounts for the fraction of energy canceled by distortion along the jth component.

The coding theorems behind such a classical development imply that the optimal tradeoff (2) between rate and distortion is asymptotically obtained by simultaneously encoding an increasing number of subsequent source symbols into a single block that can be then reverted to a sequence of distorted symbols. Hence, in principle, the intermediate symbols y feeding the anomaly detector in Fig. 2 cause it to work simultaneously on multiple instances of the signals.

Though this is not incoherent with what happens in real detectors that observe more than one suspect instance before

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

declaring an anomaly, we here instead consider a *per-use* analysis which is typical and scales the key merit figures (rate, distortion, and, in our case, distinguishability—see Section III) by the number of source symbols aggregated to obtain them.

This allows us to pursue the classical approach defining a test channel whose single user has the same expected behavior as the average of infinite uses and, in the case of Gaussian sources, has a particularly simple expression that we derive and exploit to imagine that a source instance x is encoded into a compressed symbol y from which $\hat{x}$ can be recovered [2, Ch. 13], [28].

In the same Gaussian framework, it is also possible to derive the PDF of the distorted signal $\hat{x}$ and the conditional PDF $f_{\hat{x}|x}$ that stochastically maps an input $x \sim \mathcal{G}(0, \Sigma)$ to $\hat{x}$. If we accept to identify a zero-variance Gaussian with Dirac's delta and define $S_\theta = I_n - T_\theta$ with $T_\theta = \mathrm{diag}(\tau_0, \ldots, \tau_{n-1})$ to account for the fraction of energy that survives distortion along each component, then we can derive the following lemma whose proof is in the Appendix.

Lemma 1: If $x \sim \mathcal{G}(0, \Sigma)$ is a memory-less source and we constraint the distortion $D \leq \delta$, the optimally distorted signal has distribution

$$\hat{x} \sim \mathcal{G}(0, \Sigma S_\theta) \tag{5}$$

and the optimal encoding mapping is

$$f_{\hat{x}|x}(\alpha, \beta) = G_{S_\theta \beta, \Sigma S_\theta T_\theta}(\alpha) \tag{6}$$

where $G_{m,K}(\cdot)$ represents the PDF of a Gaussian variable with mean vector m and covariance matrix K.

Although in general it is not explicitly reported, the expression of $f_{\hat{x}|x}$ is important when the compression mechanism is employed to encode a signal different from the one for which it was designed. This is the case of an unexpected anomalous source that replaces the normal signal.

III. Anomalies and Their Distinguishability

Once introduced the rate-distortion tradeoff that rules many lossy compression mechanisms, we analyze what happens when a compressed signal is processed for anomaly detection. As compression introduces distortion, it also reduces the information available for a detector to distinguish whether the transmitted signal differs from what is usually observed. Hence, in the case of anomaly detection on compressed signals, the tradeoff is threefold as, together with rate and distortion, also distinguishability comes into play.

To include this aspect in our model, each observable instance $x[t]$ has to be considered as a realization of two different sources: one modeling the normal behavior x^{ok} and one representing an anomaly x^{ko}. These two sources are modeled as two discrete-time, stationary, n-dimensional stochastic processes each generating independent and identically distributed (i.i.d.) vectors $x^{\mathrm{ok}} \in \mathbb{R}^n$ and $x^{\mathrm{ko}} \in \mathbb{R}^n$ with different PDFs $f^{\mathrm{ok}} : \mathbb{R}^n \to \mathbb{R}^+$ and $f^{\mathrm{ko}} : \mathbb{R}^n \to \mathbb{R}^+$. As a result, at any time t the observable process is either $x[t] = x^{\mathrm{ok}}[t]$ or $x[t] = x^{\mathrm{ko}}[t]$ (visually represented in Fig. 2). Since we assume the generated vectors as i.i.d., from now on we may drop the time indication.

Coherently to the previous section, we here consider the case in which both sources are Gaussian. In particular, we focus on signals with zero-mean and covariance matrices $\Sigma^{\mathrm{ok}}, \Sigma^{\mathrm{ko}} \in \mathbb{R}^{n \times n}$. In general, $\Sigma^{\mathrm{ok}} \neq \Sigma^{\mathrm{ko}}$, but we will assume $\mathrm{tr}(\Sigma^{\mathrm{ok}}) = \mathrm{tr}(\Sigma^{\mathrm{ko}}) = n$, where $\mathrm{tr}(\cdot)$ stands for matrix trace, meaning that, on the average, each sample in the vector contributes with unit energy. With the assumption of signals to be zero-mean and of equal energy, we can focus our analysis on one of the possible effects of anomalies, i.e., the distribution of energy over the signal subspace. Moreover, with no loss of generality, we assume $\Sigma^{\mathrm{ok}} = \mathrm{diag}(\lambda_0^{\mathrm{ok}}, \ldots, \lambda_{n-1}^{\mathrm{ok}})$ with $\lambda_0^{\mathrm{ok}} \geq \lambda_1^{\mathrm{ok}} \geq \cdots \geq \lambda_{n-1}^{\mathrm{ok}} \geq 0$.

Typically, a compressor aims to guarantee the best tradeoff between rate and distortion for the typical condition in which the observed signal behaves normally, i.e., $x = x^{\mathrm{ok}}$. The optimal encoder for Gaussian sources is the one indicated in (6) for which $\Sigma = \Sigma^{\mathrm{ok}}$ and $f_{\hat{x}|x} = f_{\hat{x}|x}^{\mathrm{ok}}$ have to be considered. However, once deployed, the same encoder works on anomalies, i.e., $x = x^{\mathrm{ko}}$, for which performance is suboptimal.

The encoded signal y is then observed for anomaly detection. Since we assume the decoding stage to be injective, y brings the same information of the reconstructed signal $\hat{x}$ so that, in abstract terms, processing y is equivalent to working on $\hat{x}$. As a result, the detector seeks to distinguish between normal $\hat{x}^{\mathrm{ok}} \sim f_{\hat{x}}^{\mathrm{ok}}$ and anomalous $\hat{x}^{\mathrm{ko}} \sim f_{\hat{x}}^{\mathrm{ko}}$ reconstructed signals.

Since the encoder is tuned on the normal signal, x^{ok} is optimally compressed in the rate-distortion sense into $\hat{x}^{\mathrm{ok}}$ whose distribution is reported in (5) with $\Sigma = \Sigma^{\mathrm{ok}}$, while the anomalous signal x^{ko} is suboptimally transformed into $\hat{x}^{\mathrm{ko}}$ for which $f_{\hat{x}}^{\mathrm{ko}}$ is given by the following lemma, whose proof is in the Appendix.

Lemma 2: If an anomalous source $x^{\mathrm{ko}} \sim \mathcal{G}\left(0, \Sigma^{\mathrm{ko}}\right)$ is encoded with the compression scheme $f_{\hat{x}|x}^{\mathrm{ok}}$ of Lemma 1, then

$$\hat{x}^{\mathrm{ko}} \sim \mathcal{G}\Big(0, S_\theta \Sigma^{\mathrm{ko}} S_\theta + \theta S_\theta\Big). \tag{7}$$

Such a result has two noteworthy corner cases.

1) If $\theta \to 0^+$ there is no distortion. In fact, since $S_\theta = I_n$, Lemma 2 gives $\hat{x}^{\mathrm{ko}} \sim x^{\mathrm{ko}}$.
2) If $x^{\mathrm{ko}} \sim x^{\mathrm{ok}}$ there is no anomaly, $\Sigma^{\mathrm{ok}} = \Sigma^{\mathrm{ko}}$, and

$$S_\theta \Sigma^{\mathrm{ko}} S_\theta + \theta S_\theta = \left[S_\theta + \theta\left(\Sigma^{\mathrm{ok}}\right)^{-1}\right]\Sigma^{\mathrm{ok}} S_\theta = \Sigma^{\mathrm{ok}} S_\theta$$

where the last equality holds since $S_\theta = \max\{0, I_n - \theta(\Sigma^{\mathrm{ok}})^{-1}\}$, the possible disagreements between $S_\theta + \theta(\Sigma^{\mathrm{ok}})^{-1}$ and I_n correspond to components multiplied by zero by the last S_θ factor. Hence, Lemma 2 can be compared with Lemma 1 to confirm that $\hat{x}^{\mathrm{ko}} \sim \hat{x}^{\mathrm{ok}}$.

Lemmas 1 and 2 imply that when the normal and anomalous signals are Gaussian before compression, the performance of anomaly detectors depends on how much we are capable of distinguishing between the two distributions in (5) and (7). We quantify the difference between them with two kinds of information-theoretic measures, which model two distinct scenarios, one in which the detector knows both $f_{\hat{x}}^{\mathrm{ok}}$ and $f_{\hat{x}}^{\mathrm{ko}}$ and one in which it knows only $f_{\hat{x}}^{\mathrm{ok}}$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

To proceed further it is convenient to define the functional

$$L(x';x'') = -\int_{\mathbb{R}^n} f_{x'}(\alpha)\log_2\left[f_{x''}(\alpha)\right]\mathrm{d}\alpha \tag{8}$$

that is the average coding rate, measured in bits per symbol, of a source characterized by the PDF $f_{x'}$ with a code optimized for a source with PDF $f_{x''}$, so that $L(x;x)$ is equal to the differential entropy of x [2, Ch. 8]. As an alternative statistical point of view, if $f_{x'}$ is the PDF of the symbols generated by a source x', $f_{x''}$ is the PDF of the symbols generated by a source x'' and $\ell_x(\alpha) = -\log_2 f_x(\alpha)$ is the negative log-likelihood that the symbol α has been generated by the sources x, then $L(x';x'') = \mathbf{E}[\ell_{x''}(\alpha)|x']$, i.e., the average negative likelihood that an instance is generated by the source x'' when it is actually generated by the source x'.

Within the Gaussian assumption, we can derive the analytical expression for L in the following lemma whose derivation is in the Appendix.

Lemma 3: If $x' \sim \mathcal{G}(0, \Sigma')$ and $x'' \sim \mathcal{G}(0, \Sigma'')$ then

$$L(x';x'') = \frac{1}{2\ln 2}\left\{\ln\left[(2\pi)^n\left|\Sigma''\right|\right] + \mathrm{tr}\left[\left(\Sigma''\right)^{-1}\Sigma'\right]\right\} \tag{9}$$

where $|\cdot|$ indicates the determinant of its matrix argument.

A. Distinguishability in Anomaly-Agnostic Detection

When $f^{\mathrm{ko}}_{\hat{x}}$ is unknown and only $f^{\mathrm{ok}}_{\hat{x}}$ is given, we can only consider the average coding rates referring to code optimized for $\hat{x}^{\mathrm{ok}}$, i.e., $L(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ok}})$ and $L(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ok}})$. One may quantify the difference between a normal behavior and an anomalous one by measuring the increase or decrease in the average coding rate with respect to the expected case $L(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ok}})$ as follows:

$$\zeta = L\left(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ok}}\right) - L\left(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ok}}\right) \tag{10}$$

$$= \int_{\mathbb{R}^n}\left[f^{\mathrm{ok}}_{\hat{x}}(\alpha) - f^{\mathrm{ko}}_{\hat{x}}(\alpha)\right]\log_2 f^{\mathrm{ok}}_{\hat{x}}(\alpha)\mathrm{d}\alpha. \tag{11}$$

Since there may be anomalies whose encoding yields a lower rate with respect to normal signals, ζ is not always positive. As a result, a distinguishability measure is given by considering its magnitude, i.e., $|\zeta|$.

From a statistical perspective, ζ corresponds to the difference in the expectation of the negative log-likelihood for α to be normal given either α is actually an instance of $\hat{x}^{\mathrm{ok}}$ or $\hat{x}^{\mathrm{ko}}$

$$\zeta = \mathbf{E}\left[\ell^{\mathrm{ok}}_{\hat{x}}(\alpha)|\hat{x}^{\mathrm{ko}}\right] - \mathbf{E}\left[\ell^{\mathrm{ok}}_{\hat{x}}(\alpha)|\hat{x}^{\mathrm{ok}}\right].$$

The use of the quantity $\ell^{\mathrm{ok}}_{\hat{x}}(\alpha) = -\log_2 f^{\mathrm{ok}}_{\hat{x}}(\alpha)$ can be found in other works related to anomaly detection, e.g., in [19] where it is referred as a *coding cost* of α.

With the assumption of Gaussian sources, the optimal encoder (in the rate-distortion sense) lets survive only the components j for which $\lambda^{\mathrm{ok}}_j > \theta$. Hence, $f^{\mathrm{ok}}_{\hat{x}}$ and $f^{\mathrm{ko}}_{\hat{x}}$ given in (5) and (7) have only the first n_θ components non-null with $n_\theta = \arg\max_j\{\lambda^{\mathrm{ok}}_j > \theta\}$. The other $n - n_\theta$ components are set to 0 and thus cannot be used to tell anomalous from normal cases. We therefore focus on the first n_θ components of $\hat{x}^{\mathrm{ok}}$ and $\hat{x}^{\mathrm{ko}}$ which are Gaussian with covariance matrices $\hat{\Sigma}^{\mathrm{ok}}_\theta$ and $\hat{\Sigma}^{\mathrm{ko}}_\theta$ corresponding to the $n_\theta \times n_\theta$ upper-left submatrix of $\Sigma^{\mathrm{ok}}S_\theta$ in (5) and of $S_\theta\Sigma^{\mathrm{ok}}S_\theta + \theta S_\theta$ in (7), respectively.

By properly combining the definition of ζ in (10) with the expression of L within the Gaussian assumption in (9), we obtain

$$\zeta = \frac{1}{2\ln 2}\mathrm{tr}\left[\tilde{\Sigma}_\theta - I_{n_\theta}\right] \tag{12}$$

where $\tilde{\Sigma}_\theta = (\hat{\Sigma}^{\mathrm{ok}}_\theta)^{-1}\hat{\Sigma}^{\mathrm{ko}}_\theta$ which corresponds to the $n_\theta \times n_\theta$ upper-left submatrix of $(\Sigma^{\mathrm{ok}})^{-1}\Sigma^{\mathrm{ko}}S_\theta + T_\theta$. Note that, since $\tilde{\Sigma}_\theta$ is linear with respect to $\hat{\Sigma}^{\mathrm{ko}}_\theta$, so is ζ. In addition, ζ vanishes when $\hat{\Sigma}^{\mathrm{ok}}_\theta = \hat{\Sigma}^{\mathrm{ko}}_\theta$.

As a noteworthy particular case, when the normal signal is white, i.e., when $\Sigma^{\mathrm{ok}} = I_n$, we have that $\theta \in [0, 1]$ and that for any $\theta < 1$, $T_\theta = \theta I_n$ and $n_\theta = n$. Hence, $\tilde{\Sigma}_\theta = (1-\theta)\Sigma^{\mathrm{ko}} + \theta I_n$ that leads to $\zeta = 0$. This result is not surprising since the distinguishability modeled by $|\zeta|$ depends only on the statistics of x^{ok} that has no exploitable structure.

B. Distinguishability in Anomaly Aware Detection

When both $f^{\mathrm{ok}}_{\hat{x}}$ and $f^{\mathrm{ko}}_{\hat{x}}$ are known, the anomaly detection task reduces to a binary classification problem for which we may resort to the Neyman–Pearson Lemma [2, Th. 12.7.1], [29, Th. 3.1]. This lemma can be understood in the sense that the cardinal quantity to observe is

$$r(\alpha) = \log_2\left[\frac{f^{\mathrm{ko}}_{\hat{x}}(\alpha)}{f^{\mathrm{ok}}_{\hat{x}}(\alpha)}\right]$$

which can be interpreted as a measure of abnormality of α, i.e., a score that the detector employs to distinguish whether the single α behaves normally or not. Consequently, one may measure the distinguishability between the distributions $f^{\mathrm{ok}}_{\hat{x}}$ and $f^{\mathrm{ko}}_{\hat{x}}$ as the difference between the score observed in average when $\hat{x} = \hat{x}^{\mathrm{ok}}$ and the score obtained in average when $\hat{x} = \hat{x}^{\mathrm{ko}}$

$$\kappa = \mathbf{E}\left[r(\alpha)|\hat{x}^{\mathrm{ko}}\right] - \mathbf{E}\left[r(\alpha)|\hat{x}^{\mathrm{ok}}\right] \tag{13}$$

$$= \int_{\mathbb{R}^n} f^{\mathrm{ko}}_{\hat{x}}(\alpha)\log_2\left[\frac{f^{\mathrm{ko}}_{\hat{x}}(\alpha)}{f^{\mathrm{ok}}_{\hat{x}}(\alpha)}\right]\mathrm{d}\alpha + \int_{\mathbb{R}^n} f^{\mathrm{ok}}_{\hat{x}}(\alpha)\log_2\left[\frac{f^{\mathrm{ok}}_{\hat{x}}(\alpha)}{f^{\mathrm{ko}}_{\hat{x}}(\alpha)}\right]\mathrm{d}\alpha \tag{14}$$

$$= L\left(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ok}}\right) - L\left(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ko}}\right) + L\left(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ko}}\right) - L\left(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ok}}\right) \tag{15}$$

$$= \mathcal{D}_{\mathrm{KL}}\left(f^{\mathrm{ko}}_{\hat{x}}\|f^{\mathrm{ok}}_{\hat{x}}\right) + \mathcal{D}_{\mathrm{KL}}\left(f^{\mathrm{ok}}_{\hat{x}}\|f^{\mathrm{ko}}_{\hat{x}}\right) \tag{16}$$

where given distributions f' and f'', $\mathcal{D}_{\mathrm{KL}}(f'\|f'')$ refers to the Kullback–Leibler divergence [2, Ch. 2], of which κ results to be the symmetrized version. Note that, κ is similar to the *divergence* defined in [30] and [31] for binary classification problems.

The measure κ models a detector that knows the distributions of both normal and anomalous sources such that their optimal codes are also known. From (15), it is evident that κ may be interpreted as the sum of the differences in the average coding rate for both distorted sources with a code optimized for the normal source $L(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ok}}) - L(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ok}})$ and optimized for the anomalous source $L(\hat{x}^{\mathrm{ok}};\hat{x}^{\mathrm{ko}}) - L(\hat{x}^{\mathrm{ko}};\hat{x}^{\mathrm{ko}})$. Since the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

average coding rate is expected to be shorter when employed to code a source for which it is optimized, these differences are expected to be greater when the difference between two distributions $f_{\hat{x}}^{\text{ok}}$ and $f_{\hat{x}}^{\text{ko}}$ increases. As a result, large κ values correspond to system configurations with high detection capability.

Differently from ζ, κ is a quantity that is always positive and can be directly used as a distinguishability measure.

Within the Gaussian assumption, the distinguishability measure κ becomes

$$\kappa = \frac{1}{2\ln 2}\text{tr}\left[\tilde{\Sigma}_\theta + \tilde{\Sigma}_\theta^{-1} - 2I_{n_\theta}\right] \tag{17}$$

from which it is evident that κ is convex with respect to $\hat{\Sigma}_\theta^{\text{ko}}$ and, as for ζ, κ vanishes for $\hat{\Sigma}_\theta^{\text{ok}} = \hat{\Sigma}_\theta^{\text{ko}}$.

As a final remark, coherently with the typical *per use* analysis, distinguishability measures implicitly consider detectors that scrutiny an increasing number of subsequent source instances and scale their performance by such a number. Hence, as rate and distortion coming from (2) are best-case bounds that can be approximated by increasing the complexity of the system, the distinguishability measures indicate how fast a detector accumulates information to declare an anomaly. The higher such a figure, the lower the number of subsequent symbols needed to arrive at a conclusion or, alternatively, the higher the confidence in a conclusion drawn after analyzing a single instance.

IV. Average and Large-Window Distinguishability

Once framework and metrics are defined, it is interesting to analyze points of view from which white noise appears to be important. First, it is the average over the set of all possible anomalies. Second, it is the asymptotic behavior of anomalies when the signal dimension increases.

A. Average on the Set of Possible Anomalies

Anomalies modeled as zero-mean Gaussian vectors with fixed energy are completely defined by their covariance matrix Σ^{ko} where $\text{tr}(\Sigma^{\text{ko}}) = n$. We decompose $\Sigma^{\text{ko}} = U^{\text{ko}}\Lambda^{\text{ko}}U^{\text{ko}\top}$ with $\Lambda^{\text{ko}} = \text{diag}(\lambda_0^{\text{ko}}, \ldots, \lambda_{n-1}^{\text{ko}})$ and U^{ko} orthonormal.

The set of all possible $\lambda^{\text{ko}} = (\lambda_0^{\text{ko}}, \ldots, \lambda_{n-1}^{\text{ko}})^\top$ is

$$\mathbb{S}^n = \left\{\lambda \in \mathbb{R}^{+n} | \sum_{j=0}^{n-1}\lambda_j = n\right\}$$

while the set of all possible U^{ko} is that of orthonormal matrices

$$\mathbb{O}^n = \left\{U \in \mathbb{R}^{n\times n} | U^\top U = I_n\right\}.$$

By indicating with $\mathcal{U}(\cdot)$ the uniform distribution in the argument domain, we will assume that when λ^{ko} is not known then $\lambda^{\text{ko}} \sim \mathcal{U}(\mathbb{S}^n)$ and, similarly, when U^{ko} is not known then $U^{\text{ko}} \sim \mathcal{U}(\mathbb{O}^n)$, independently of λ^{ko}.

Note now that $\mathbb{S}^n$ is invariant with respect to any permutation of the λ_j. Since $\lambda^{\text{ko}} \sim \mathcal{U}(\mathbb{S}^n)$, also $\mathbf{E}[\lambda^{\text{ko}}]$ must be invariant with respect to the same permutations so that $\mathbf{E}[\lambda_j^{\text{ko}}] = \mathbf{E}[\lambda_k^{\text{ko}}]$ for any j, k. Since λ^{ok} has a constrained sum and is the diagonal of Λ^{ko} we have $\mathbf{E}[\Lambda^{\text{ko}}] = I_n$. This implies

$$\begin{aligned}\mathbf{E}\left[\Sigma^{\text{ko}}\right] &= \mathbf{E}\left[U^{\text{ko}}\Lambda^{\text{ko}}U^{\text{ko}\top}\right] \\ &= \mathbf{E}\left[U^{\text{ko}}\mathbf{E}\left[\Lambda^{\text{ko}}\right]U^{\text{ko}\top}\right] \\ &= \mathbf{E}\left[U^{\text{ko}}U^{\text{ko}\top}\right] = I_n. \end{aligned}\tag{18}$$

Hence, in our setting, the average anomaly is white and we may compute the corresponding distinguishability measures ζ_I and κ_I, i.e., ζ and κ when $\Sigma^{\text{ko}} = I_n$. Note that, in this case, $\tilde{\Sigma}_\theta$ is the $n_\theta \times n_\theta$ upper-left submatrix of $(\Sigma^{\text{ok}})^{-1}S_\theta + T_\theta$, which is a diagonal matrix whose diagonal elements are

$$u_{\theta,j} = \frac{1}{\lambda_j^{\text{ok}}}\left(1 - \frac{\theta}{\lambda_j^{\text{ok}}}\right) + \frac{\theta}{\lambda_j^{\text{ok}}}.$$

With these quantities, the expressions of the distinguishability measures become

$$\zeta_I = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta-1}\left(u_{\theta,j} - 1\right) \tag{19}$$

$$\kappa_I = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta-1}\left(u_{\theta,j} + \frac{1}{u_{\theta,j}} - 2\right). \tag{20}$$

Note that due to Jensen's inequality, the linearity of ζ and the convexity of κ, we have $\zeta_I = \mathbf{E}[\zeta]$ and $\kappa_I \leq \mathbf{E}[\kappa]$.

Moreover, the very simple structure of ζ_I allows the derivation of the following theorem whose proof is in the Appendix.

Theorem 1: If $\bar{k} = \arg\max_k\{\lambda_k^{\text{ok}} \geq \lambda_k^{\text{ko}} = 1\}$, then $\zeta_I = 0$ for at least one point $0 < \theta < \lambda_{\bar{k}}^{\text{ok}}$.

Considering a white anomaly, the intuition behind this theorem is the following. When distortion is null (no compression), since $\hat{x}^{\text{ko}}$ and $\hat{x}^{\text{ok}}$ have the same average energy and the coding is tuned on $\hat{x}^{\text{ok}}$, $L(\hat{x}^{\text{ko}}, \hat{x}^{\text{ok}}) > L(\hat{x}^{\text{ok}}, \hat{x}^{\text{ok}})$ such that ζ_I is positive. On the other hand, when distortion is so high that only the first component of x^{ok} survives, i.e., $\hat{\Sigma}_\theta^{\text{ok}} = \lambda_0^{\text{ok}} - \theta$, a single component also survives in $\hat{x}^{\text{ko}}$. In this setting, ζ_I depends on the difference between the two scalar quantities $\hat{\Sigma}_\theta^{\text{ok}}$ and $\hat{\Sigma}_\theta^{\text{ko}}$. With few numerical manipulations, it is possible to prove that $\hat{\Sigma}_\theta^{\text{ok}} > \hat{\Sigma}_\theta^{\text{ko}}$ thus ζ_I results to be negative. Since ζ_I is continuous in θ, it must pass through zero at least once. Therefore, at least one critical level of distortion exists that makes ineffective detectors not using information on the anomaly.

B. Asymptotic Distinguishability

White signals are not only the average anomalies but are also *typical* anomalies in a sense specified by the following theorem whose proof is in the Appendix.

Theorem 2: If $\Sigma^{\text{ko}} = U^{\text{ko}}\text{diag}(\lambda_0^{\text{ko}}, \ldots, \lambda_{n-1}^{\text{ko}})U^{\text{ko}\top}$, where $\lambda^{\text{ko}} \sim \mathcal{U}(\mathbb{S}^n)$ and $U^{\text{ko}} \sim \mathcal{U}(\mathbb{O}^n)$, then $\forall\beta > 1/2$, $\Delta_{\text{F}} = n^{-\beta}\|\Sigma^{\text{ko}} - I_n\|_F$, with $\|\cdot\|_F$ the Frobenius matrix norm, tends to 0 in probability as $n \to \infty$.

Hence, when n increases, most of the possible anomalies behave as white signals. From an anomaly detection perspective, if the signal is characterized by a sufficiently large

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

dimension n, the designer may consider the white anomaly as a reference.

V. Simulation Setup

In this section, we propose a framework in which we define models for normal signal and anomalies, three different encoders, well-known detectors, and a practical figure of merit to be compared with the proposed distinguishability measures ζ and κ, i.e., a metric that can be computed in both anomaly-agnostic and anomaly-aware scenarios.

A. Normal Signal and Anomalies

Normal signals are assumed to be $x^{\text{ok}} \sim \mathcal{G}\left(0, \Sigma^{\text{ok}}\right)$ where Σ^{ok} is the diagonal matrix of the eigendecomposition of a given matrix K, such that $K = U\Sigma^{\text{ok}}U^\top$ where U is an orthonormal matrix. K is a semidefinite positive $n \times n$ square matrix where the j,k-entry is equal to $\omega^{|j-k|}$, with $j,k = 0,\dots,n-1$. The parameter ω is set to yield a different degree of nonwhiteness measured with the so-called *localization* defined as follows:

$$\mathcal{L}_{x^{\text{ok}}} = \frac{\text{tr}\left(\Sigma^{\text{ok}^2}\right)}{\text{tr}^2\left(\Sigma^{\text{ok}}\right)} - \frac{1}{n}. \tag{21}$$

The localization goes from $\mathcal{L}_{x^{\text{ok}}} = 0$ when the signal is white to $\mathcal{L}_{x^{\text{ok}}} = 1 - (1/n)$ when the whole energy is concentrated along a single direction of the signal space (see [32] for more details). To show the effect of realistic localization [33], we consider values of ω corresponding to $\mathcal{L}_{x^{\text{ok}}} \in \{0, 0.05, 0.2\}$. We refer to this class of signals as GSs. We also consider two alternative and more realistic settings in which the vectors x contain samples of an ECG waveform and accelerometers (ACC) signals coming from a structural health monitoring system (see Section VI-D). Since these classes of signals cannot be modeled as a stationary Gaussian process, we refer to them as non-GSs (NGSs).

Anomalous signals are generated as $x^{\text{ko}} \sim \mathcal{G}\left(0, \Sigma^{\text{ko}}\right)$ where Σ^{ko} is randomly drawn according to the uniform distribution defined in Section IV-A. In detail, $\Sigma^{\text{ko}} = U^{\text{ko}}\Lambda^{\text{ko}}U^{\text{ko}^\top}$ with $\Lambda^{\text{ko}} = \text{diag}(\lambda^{\text{ko}})$ where the terms $\lambda^{\text{ko}} \sim \mathcal{U}(\mathbb{S}^n)$ and $U^{\text{ko}} \sim \mathcal{U}(\mathbb{O}^n)$ are generated independently. The former term λ^{ko} is drawn according to [34], i.e., we first draw $\xi_j \sim \mathcal{U}([0,1])$ for $j = 0,\dots,n-1$ and then set

$$\lambda_j^{\text{ko}} = \frac{n \log \xi_j}{\sum_{k=0}^{n-1} \log \xi_k} \tag{22}$$

where $\lambda_j^{\text{ko}} > 0$ and the sum of the entries of λ^{ko} is n. For the latter term $U^{\text{ko}} \sim \mathcal{U}(\mathbb{O}^n)$, we follow [35]. First, the matrix A is generated within the Ginibre ensemble [36], i.e., with independent entries $A_{j,k} \sim \mathcal{G}(0,1)$ for $j,k = 0,\dots,n-1$. Then, U^{ko} is obtained as the orthonormal factor in the QR-decomposition of A.

The first use of this random sampling is to support Theorem 2 with some numerical evidence. Fig. 3 reports the vanishing trends of the average squared (Δ_{F} with $\beta = 1$) and uniform (intended as $\Delta_{\max} = \|\Sigma^{\text{ko}} - I_n\|_{\max} = \max_{j,k} |\Sigma^{\text{ko}}_{j,k} - (I_n)_{j,k}|$) deviations of uniformly distributed covariance matrices Σ^{ko} from I_n. The Δ_{F} trend confirms the result of Theorem 2 while the $\Delta_{\max}$ trend empirically generalizes it to a stronger deviation measure. With this, we may safely think that in the case of n sufficiently large, the white Gaussian noise is a good candidate to represent the class of anomalies coming independently of the statistics of the normal signal.

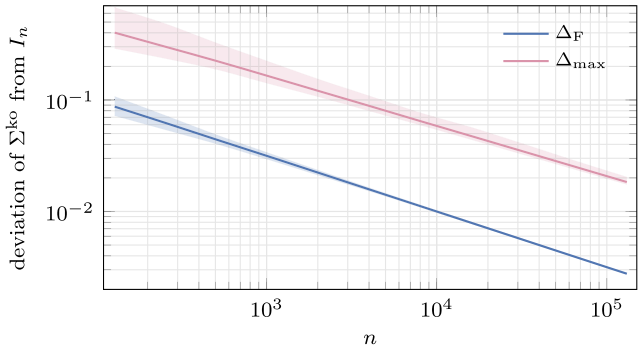

Fig. 3. Trends of Δ_{F} and $\Delta_{\max}$ for $n \in \{2^k\}_{k=7}^{17}$. Solid lines are mean trends over 2000 trials, while shaded areas contain 98% of the population.

B. Encoders

We identify three possible compression techniques tuned to the normal signal to be applied to normal and anomalous instances. More specifically, x is first compressed and then uncompressed to $\hat{x}$.

1) The rate-distortion compression (RDC), identified by the optimization problem in (2), which gives the lowest possible transmission rate for a given level of distortion.
2) The principal component compression (PCC) which consists of projecting x along the subspace spanned by the eigenvectors of Σ^{ok} with the largest eigenvalues. PCC is a linear compressor minimizing distortion when the n-dimensional vector x is represented in a lower dimension [37], [38].
3) Auto-encoder compression (AEC) based on suitably arranged deep neural structures [39, Ch. 14] that learn a latent nonlinear representation of the signal generalizing PCC [38]. The network is defined indicating with $p < n$ the dimension of the compressed representation, to deploy a chain of fully connected layers of dimensions n, $4n$, $2n$, and p, followed by a dual network whose layers have dimensions p, $2n$, $4n$, and n. These networks are trained to minimize distortion computed as in (1).[2]

These three schemes address in a different way the tradeoff between compression and distortion. Since we refer to a theoretical model based on continuous quantities and for which the rate is potentially infinite, the compressors have to be paired with a quantization stage ensuring a finite rate. In particular,

[2]To smooth performance degradation in AEC, we first train an autoencoder with $p = n-1$. Then, the node of the latent representation along which we measure the least average energy is dropped to produce a smaller network with an $(p-1)$-dimensional latent space. The obtained network is retrained using the previous weights as initialization. This process is repeated with decreasing p and thus considering larger distortion values. Trainings employ ADAM optimizer [40] with batch-size 128, and an initial learning rate of 0.01.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

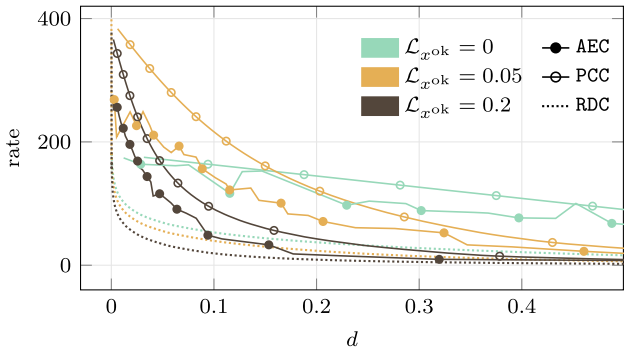

Fig. 4. Rate distortion curves for the three compression schemes we consider and for different values of the localization of the original signal.

we consider $n = 32$ and we encode each component of $\hat{x}$ with 16 bits so that rates are a maximum of $16n = 512$ bits per time step. We assume quantization fine enough to substantially preserve the Gaussian distribution of $\hat{x}$ and thus evaluate the mutual information between x and $\hat{x}$ as if they were jointly Gaussian with a covariance matrix estimated by the Monte Carlo simulation [41]. Such estimation yields the rate-distortion curves in Fig. 4 where on the x-axis is the normalized distortion $d = D/n$ in the range $d \in [0, 0.5]$ as larger values are usually beyond operative ranges. As expected, RDC yields the lowest rates confirming the role of the theoretical lower bound. Among the realistic compressors, PCC gives the largest rates while AEC, being a nonlinear generalization of PCC, better approaches the theoretical limit given by RDC. Note that, only the results of Fig. 4 refer to the additional quantization stage while in the remaining part of our analysis, we consider continuous sources.

From the signals point of view, both RDC and PCC preserve the Gaussian distribution so that when the input is Gaussian so is the compressed output. Conversely, AEC may alter the statistical distribution of a Gaussian source. We refer to the output of a compressor as Gaussian compressed signal (GC) when it is Gaussian, or non-GC (NGC), otherwise.

As a final remark, we recall that the result in Theorem 1 is only guaranteed for RDC in the case of Gaussian sources. However, we will show some numerical evidence that confirms Theorem 1 also for PCC and AEC, which are two compressors that adapt the encoder–decoder pair to the statistical characterization of the signal.

C. Detectors

The compressed version of the signal is then passed to a detector whose task is to compute a score such that high-score instances should be more likely to be anomalous. The final binary decision is taken by matching the score against a threshold. We first consider two detectors not relying on the information of the anomaly (anomaly-agnostic).

1) A likelihood detector (LD) whose score is the same considered for ζ, so that to each instance x we associate the score $\ell^{\text{ok}}_{\hat{x}}(\hat{x}) = -\log f^{\text{ok}}_{\hat{x}}(\hat{x})$.

TABLE I
ADOPTED MODELS FOR INPUT SIGNALS, COMPRESSED SIGNALS, AND DETECTOR FAMILIES

TAG	Description	Setting
	Input signal model	
GS	stationary Gaussian signals	$\mathcal{L}^{\text{ok}}$
NGS	realistic, non-stationary, and non-Gaussian signals	ECG, ACC
	Compressed signal model	
GC	GS compressed preserving the Gaussianity	$\mathcal{L}^{\text{ok}}$ with RDC or PCC
NGC	non-Gaussian compressed signal	$\mathcal{L}^{\text{ok}}$ with AEC: compressed ECG and ACC
	Detector family	
DbD	detector exploiting the PDF of the compressed signal	LD and NPD
LbD	data-driven detectors	OCSVM and DNN

2) A one-class support-vector machine (OCSVM) [42] with a Gaussian kernel,[3] trained on a set of instances of normal signals contaminated by 1% of unlabeled white instances to help the algorithm in finding the envelope of normal instances.

For what concerns the class of anomaly-aware methods, we consider two detectors that are able to leverage information on the anomaly.

1) A Neyman–Pearson detector (NPD), whose score is the same considered for κ, so that to instance x we associate the score $r(\hat{x}) = \log f^{\text{ko}}_{\hat{x}}(\hat{x}) - \log f^{\text{ok}}_{\hat{x}}(\hat{x})$.
2) A deep neural network (DNN) with three fully connected hidden layers with p, $2n$, and n neurons, ReLu activations, and a final sigmoid neuron producing the score. The network is trained[4] with a binary cross-entropy loss against a dataset containing labeled normal and anomalous instances.

LD and NPD detectors rely on the statistical characterization of the signals, hence they can only be employed with GSs compressed by RDC or PCC. Conversely, OCSVM and DNN are data-driven detectors that can also be applied to NGCs, i.e., non-Gaussian sources compressed with any detector or Gaussian sources compressed by AEC. We classify the former detectors as distribution-based detectors (DbDs) while the latter as learning-based detectors (LbDs). Table I summarizes the considered models for input and compressed signals along with the adopted families of detectors. The table also reports tags that will be used in future plots.

LbDs do not make assumptions on signals but require a training set that in our case contains 10^5 normal signal examples. When OCSVM is considered, this set is contaminated with a 1% of white noise instances while the training of

[3]The signal components are normalized by their variance and the scale parameter of the Gaussian kernel is fixed to $1/n_\theta$.

[4]DNN is trained by backpropagation with an ADAM optimizer [40], a batch-size of 20 instances, and an initial learning rate of 0.01 that is scaled by 0.2 whenever the validation loss does not decrease for 5 epochs, where the validation set consists of the 10% of the instances initially devoted to training. This setting is the result of tuning.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

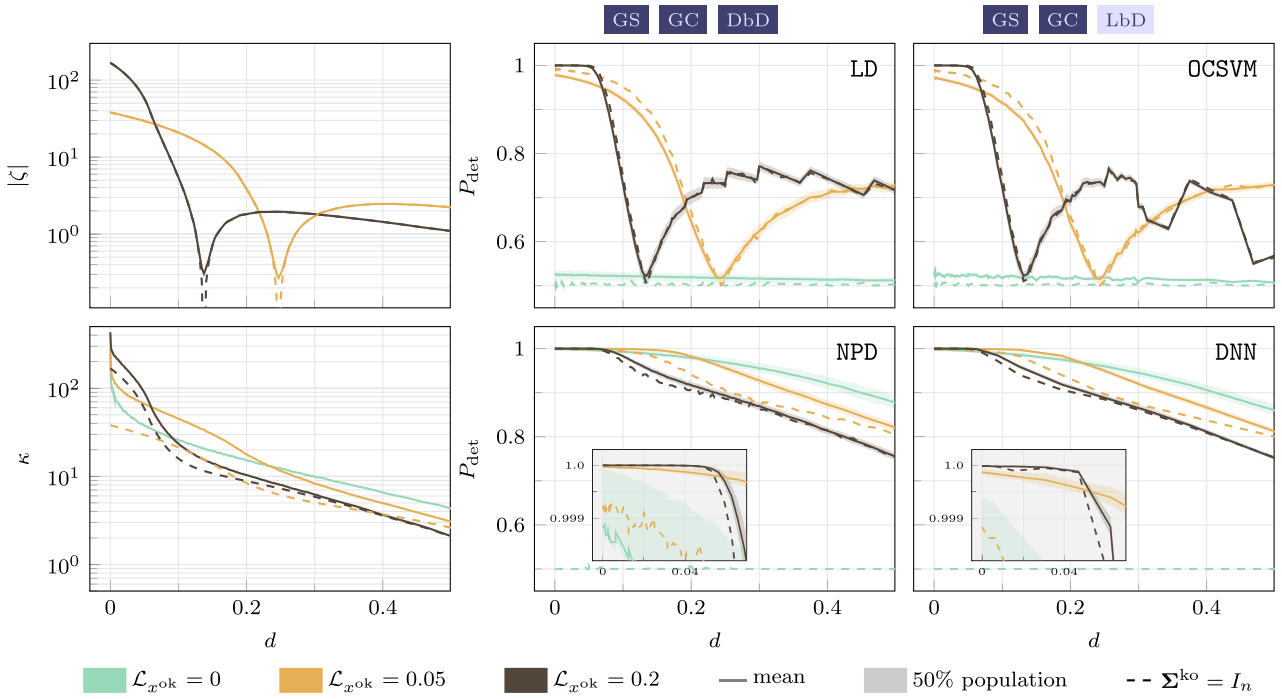

Fig. 5. Distinguishability measures ζ, κ, and P_{det} against normalized distortion d in case of RDC.

DNNs requires additional anomalous examples. In this case, we enlarge the training set with 10^5 anomalous examples where 50 different Σ^{ko} are generated such that for each one a different model with the same architecture is trained.

D. Metrics

All our detectors generate a score for each instance $\hat{x}$ that must be compared with a threshold to classify $\hat{x}$ as normal or anomalous. To be independent of thresholds, detectors' performance is assessed by the area-under-the-curve (AUC) methodology [43]. AUC $\in [0, 1]$ estimates the probability for a score of a random normal instance to be lower than the one computed for a random anomalous instance.

Clearly, detectors with AUC $= 0.5$ are no better than coin tossing. Yet, AUC < 0.5 represents the case in which the detector is more likely to score normal instances higher than anomalous ones. Such scores can be still used to distinguish normal from anomalous signals if they are interpreted in a reverse way. Hence, it is convenient to set our *probability of detection* defined as follows:

$$P_{\text{det}} = \begin{cases} \text{AUC}, & \text{if AUC} \geq 0.5 \\ 1 - \text{AUC}, & \text{if AUC} < 0.5. \end{cases} \tag{23}$$

Note that, if AUC must be estimated from samples, reversing values lower than 0.5 is not always possible. There are classes of estimators for which values less than 0.5 are not reliable [44], [45]. From now on, we report results referring to AUC estimated as in [43] for which reversing values lower than 0.5 is possible.

In the following, the trends of P_{det} are reported and matched with the trends of $|\zeta|$ and κ to show how theoretical properties reflect on real cases. Comparison is in part qualitative as ζ and κ quantify the distinguishability with bits per symbol while P_{det} comes from the probability of correct detection. Note also that ζ and κ refer to the difference between the average values of the score in the normal and anomalous cases, while P_{det} takes into account the entire distributions of these scores.

VI. Numerical Evidence

Considering the framework described above, here we match the theoretical derivations with the quantitative assessment of the performance of the described anomaly detectors for both anomaly-agnostic and anomaly-aware scenarios applied to signals compressed according to RDC, PCC, and AEC. Then, we discuss the case of ECG and ACC signals.

In the following plots, P_{det} is computed by considering 1000 normal signal and 1000 anomalous examples. The involved anomalies are the white noise plus 1000 different distributions, each characterized by a randomly generated Σ^{ko}. For DNN, we limit the analysis to 50 anomalies since the training process must be repeated for each of them.

A. RDC

Fig. 5 summarizes the results we have in this case with two rows of three plots each. The upper row of plots corresponds to detectors that do not exploit information on the anomaly, while the lower row of plots concerns detectors that may leverage information on the anomaly. Colors correspond to different $\mathcal{L}_{x^{\text{ok}}}$, dashed trends assume that the anomaly is the average one, i.e., white, and shaded areas show the span of 50% of the Monte Carlo population. The profiles of $|\zeta|$ and κ on the left shall be matched with the profiles on the right that correspond to the four detectors we consider. No $|\zeta|$ profile appears for $\mathcal{L}_x = 0$ as in that case $\zeta = 0$ (as anticipated in Section III-A). In Fig. 5 and in the successive

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

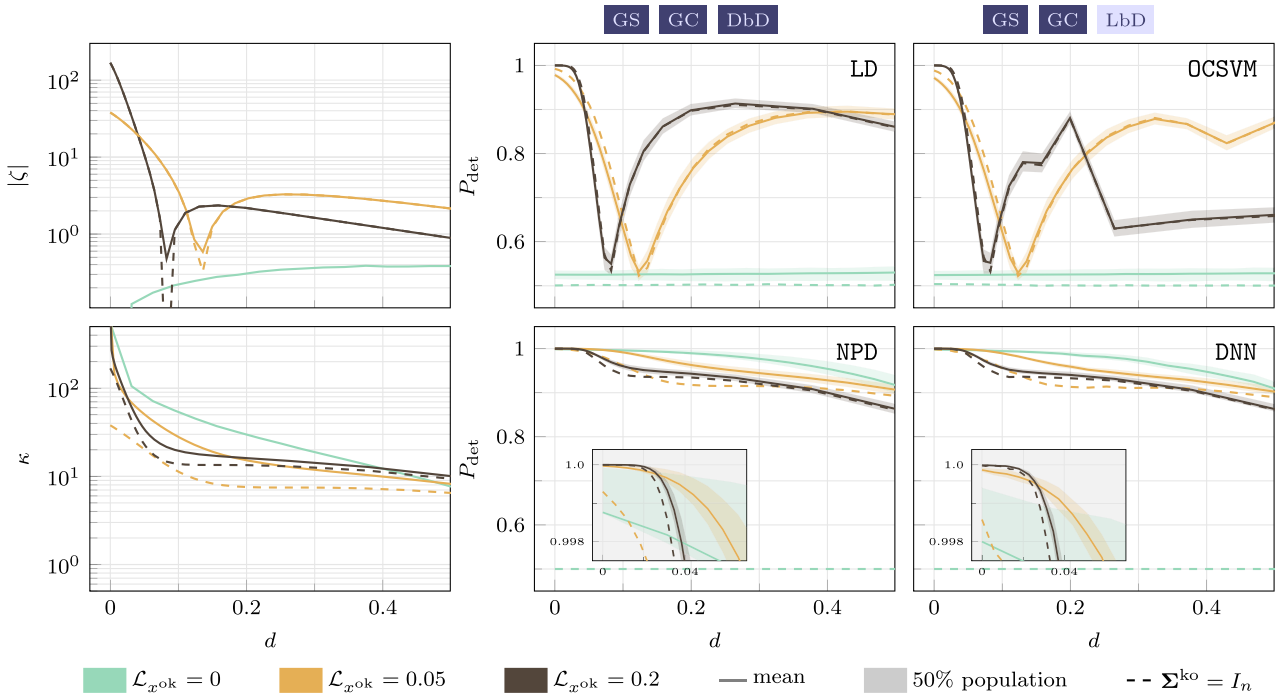

Fig. 6. Distinguishability measures ζ, κ, and P_{det} against normalized distortion d in case of PCC.

figures, we display above of each plot (or column of plots) a tag corresponding to the setting as defined in Table I.

Numerical results confirm that $\zeta_I = \mathbf{E}[\zeta]$, $\kappa_I \leq \mathbf{E}[\kappa]$ as discussed in Section IV-A. This corroborates the role of the white anomaly as a reference case since it represents the average behavior in the case of an anomaly-agnostic detector or a lower bound in anomaly-aware scenarios. The white anomaly is not only a reference case but also the case to which any possible anomaly tends when n increases as demonstrated in Theorem 2.

The theory also anticipates that, in the anomaly-agnostic scenario (upper row), the relation between detection performance and distortion is nonmonotonic and there exists a proper amount of distortion making distinguishability vanish and thus detectors fail. This critical point corresponds to the distortion level for which $|\zeta|$ crosses zero and the relation between normal and anomalous scores inverts. This behavior can be observed for both detectors, such as LD and OCSVM. The distortion level at which detectors fail depends also on $\mathcal{L}_{x^{\text{ok}}}$ as predicted by Theorem 1. Overall, theoretical measures $|\zeta|$ and κ anticipate that in the low-distortion region, more localized signals are more distinguishable from anomaly though they cause detector failures at smaller distortions with respect to less localized signals.

Detectors leveraging the knowledge of the anomaly (lower row) fail completely only at the maximum level of distortion as revealed by the abstract distinguishability measure κ. Also in this case, by comparing the trend of κ with the zoomed areas in the NPD and DNN plots, we see how theoretical measures anticipate that, in the low-distortion region, more localized signals tend to be more distinguishable from anomalies but cause more definite performance degradation of detectors when d increases.

B. PCC

From the point of view of the rate-distortion tradeoff PCC is largely suboptimal. Yet, due to its linear nature, x and $\hat{x}$ are still jointly Gaussian, so that, also in this case, we can compute the theoretical $|\zeta|$ and κ by means of (12) and (17).

Fig. 6 summarizes the results we have in this case with plots of the same kind of Fig. 5. The qualitative behaviors commented on in the previous section appear in the new plots and are anticipated by the trends of the theoretical quantities.

The distortion levels at which anomaly-agnostic detectors fail to change with respect to the RDC case but are still anticipated by the theoretical curves and Theorem 1.

In this case, the values of $|\zeta|$ beyond breakdown distortion levels increase slightly more than in the optimal compression scenario. Hence, by adopting a compression strategy that is suboptimal in the rate-distortion sense one may obtain a better distinguishability of the compressed normal signal from the compressed anomalies. This is, indeed, what happens in practice as highlighted by the LD and OCSVM plots in the first row of Fig. 6.

C. AEC

In this case, compression is nonlinear so that x and $\hat{x}$ may not be jointly Gaussian. This prevents us from computing the theoretical curves $|\zeta|$ and κ and from applying LD and NPD that rely on the knowledge of the distribution of the signals. For this reason, Fig. 7 reports only the performance of OCSVM and DNN detectors.

Notice how the qualitative trends of those performances still follow, though with a larger level of approximation, what is indicated by the theoretical curves for PCC.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

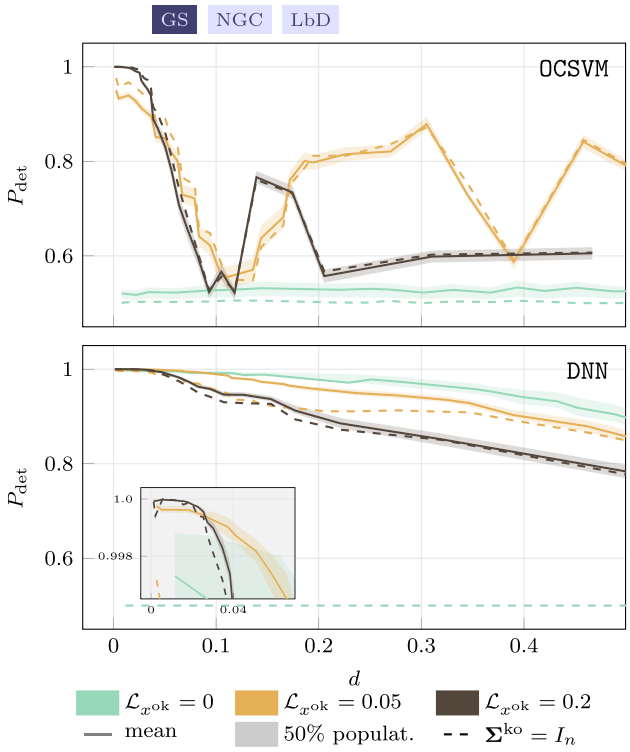

Fig. 7. Distinguishability measure P_{det} against normalized distortion d in case of AEC.

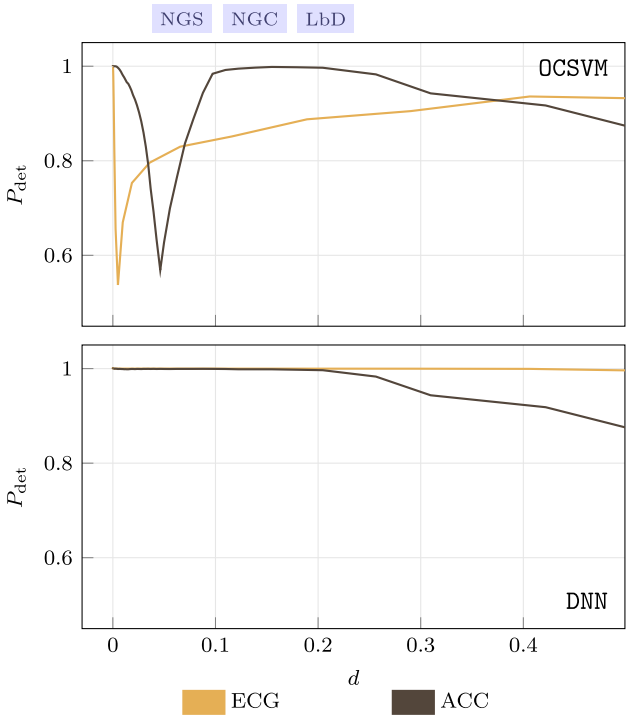

Fig. 8. Distinguishability measure P_{det} against normalized distortion d in case of PCC for ECG and accelerometers (ACC) signals considering windows of $n = 64$ samples.

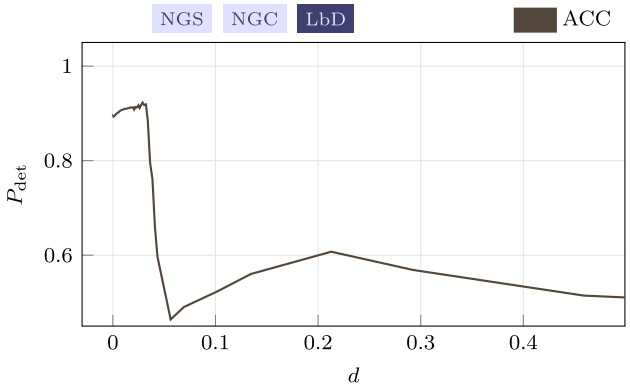

Fig. 9. Distinguishability measure P_{det} against normalized distortion d in case of a real anomaly in accelerometers (ACC) signals considering windows of $n = 64$ samples.

D. Distinguishability in Real Applications

To further test the effectiveness of our theoretical framework, we consider two realistic applications in which signals and nonstationary and non-Gaussian: 1) ECG and 2) ACC.

ECG signals are generated as in [46][5] with the same setup as in [47] in which the heart-beat rate is randomly drawn in the range 60–100 beats per minute, and the sampling rate is set to 256 samples per second. We generate 10^5 chunks containing 512 samples, from which we randomly pick vectors $x^{\text{ok}} \in \mathbb{R}^n$ with $n = 64$.

ACC signals come from a viaduct along an Italian motorway [37], [38], which is monitored by means of 90 three-axis accelerometers providing a stream of 100 samples per second for each axis. Such signals report the elastic response of the viaduct to external stimuli including car traffic or environmental factors. Here, we focus on the readings of a single sensor and a single axis.

Both ECG and ACC instances are then compressed by PCC, while OCSVM and DNN discriminate normal instances from instances of white anomalies x^{ko}. Since the input x is NGS then $\hat{x}$ is NGC, independently of the adopted compression mechanism.

Results in Fig. 8 confirm the trends seen in the previous settings. In the agnostic-anomaly scenario, the performance of OCSVM features a critical but not disruptive distortion for which the white anomaly is undetectable. As observed in the previous settings, different classes of signals exhibit different critical distortion levels. In the anomaly-aware scenario, the performance of DNN is monotonic in d. In particular, P_{det} is very close to 1 for all considered d values in the ECG case, while for ACC, it decreases with d.

As the last case study, we consider a real-world anomaly affecting the ACC signal. In particular, we refer to a nondisruptive failure affecting the monitored civil structure [37], which consists of a slight change in the modal frequencies. Hence, the anomaly statistical characterization is not white. Results in terms of P_{det} in the case of LD detectors are depicted in Fig. 9. The reported trend confirms the presence of a critical and nondisruptive value of d in correspondence with the distortion level anticipated in Fig. 8 where anomalies are instead generated as instances of white noise.

[5]MATLAB and C code available at the Physionet website https://physionet.org/content/ecgsyn/1.0.0/.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

E. Summary

Having as main focus the validation of both proposed measures of distinguishability, we analyzed some numerical evidence. First, a scenario matching the theoretical framework has been considered, where Gaussian synthetic signals are compressed using an optimal encoder in terms of rate-distortion. Fig. 5 illustrates how the two metrics can predict the performance of detectors that either leverage the knowledge of the signal distribution or learn from data. Moreover, in the anomaly-agnostic case, the results also confirm the existence of a critical level of distortion for which the anomaly is undetectable, as predicted in Theorem 1.

Subsequently, we relax the assumptions on which the theoretical results were derived, one by one. In Fig. 6, the optimal encoder is replaced with PCC, which is a realistic encoder for which the compressed signals can still be modeled as Gaussian. In Fig. 7, the encoder is AEC which alters the statistical properties of the signals. In Fig. 8, GSs representing the normal behavior are replaced with real signals ECG and ACC which are not Gaussian. Finally, in Fig. 9, also the anomaly is a real NGS.

This empirical evidence strongly supports how the presented framework offers valuable guidance for real-world scenarios. However, it is important to note that our analysis focused on cases where the encoder was tuned to the statistical characterization of the normal signal, and these results cannot be directly extrapolated to different compression mechanisms that may strongly differ from both Lemmas 1 and 2.

VII. Conclusion

Massive sensing systems may rely on lossy compression to reduce the bitrate needed to transmit acquisitions to the remote node while theoretically maintaining the relevant information. At some intermediate point along their path to centralized servers, compressed sensor readings may be processed for early detection of anomalies. Such detection is therefore carried out on compressed data.

In a framework approximating normal and anomalous signals with Gaussian sources, we revise the classical rate-distortion theory to report the distributions of the distorted signals and the mapping to obtain them (see Lemmas 1 and 2).

We define two information-theoretic metrics to measure the distinguishability between normal and anomalous sources in anomaly-agnostic and anomaly-aware scenarios. For these metrics, we provide a statistical interpretation and we compute their closed forms when GSs are involved.

We prove with Theorem 1 that, in the anomaly-agnostic scenario, there exists at least one critical level of distortion for which the detector is ineffective. We show that the white anomaly is a reference case that can be employed in the design of the system. Indeed, it provides information about the average and minimum performance in the anomaly-agnostic and anomaly-aware scenarios, respectively. In addition, we demonstrate with Theorem 2 that any possible anomaly tends to be white in the asymptotic case.

All these results are confirmed with numerical examples considering several settings. We first show that, in the case of Gaussian sources, the theoretical measures of distinguishability anticipate the performances of detectors well-aligned with the Gaussian framework. The same capability of anticipating the performance of a detector has been confirmed in the case of generic data-driven detectors, i.e., when the detection mechanism is not based on the hypothesis of Gaussian sources, and in the case of a compression mechanism altering the input signal distribution. Finally, we further confirm the effectiveness of the proposed metric in the case of a realistic NGS.

Appendix

Proof of Lemma 1: Distortion is tuned to the normal case that entails a memoryless source. Hence we may drop time indications and concentrate on a vector x with independent components $x_j \sim \mathcal{G}(0, \lambda_j)$ for $j = 0, \ldots, n-1$.

We know from [28] that for a given value of the parameter θ, each component x_j is transformed separately into $\hat{x}_j$. In particular

$$\hat{x}_j = \begin{cases} 0, & \text{if } \lambda_j \le \theta \\ x_j + \Delta_j, & \text{if } \lambda_j > \theta \end{cases} \tag{24}$$

where to achieve the Shannon lower bound, Δ_j must be an instance of a Gaussian random variable independent of $\hat{x}_j$. Hence, the three quantities $\hat{x}_j$, x_j and Δ_j must be such that $(\hat{x}_j, x, \Delta_j)^\top \sim \mathcal{G}\left(0, \Sigma_{\hat{x}_j, x_j, \Delta_j}\right)$ with

$$\Sigma_{\hat{x}_j, x_j, \Delta_j} = \begin{pmatrix} \lambda_j - \theta & \lambda_j - \theta & 0 \\ \lambda_j - \theta & \lambda_j & -\theta \\ 0 & -\theta & \theta \end{pmatrix}. \tag{25}$$

That explains in which sense $\hat{x}_j$ *encodes* x_j. In fact, the nondiagonal elements $\lambda_j - \theta$ are positive, and thus $\hat{x}_j$ and x_j are positively correlated.

From (25), if we agree to identify a Gaussian with 0 variance with a Dirac's delta we infer that $\hat{x}_j \sim \mathcal{G}\big(0, \max\{0, \lambda_j - \theta\}\big)$ and thus $\hat{x} \sim \mathcal{G}(0, \Sigma S_\theta)$.

Moreover, $(\hat{x}_j, x_j)^\top \sim \mathcal{G}\left(0, \Sigma_{\hat{x}_j, x_j}\right)$ with $\Sigma_{\hat{x}_j, x_j}$ the upper-left 2×2 submatrix of $\Sigma_{\hat{x}_j, x_j, \Delta_j}$ in (25). If we assume that $\theta < \lambda_j$, from the joint probability of x_j and $\hat{x}_j$, we may compute the action of $f_{\hat{x}|x}$ on the jth component of x_j as the PDF of $\hat{x}_j$ given x_j, i.e.,

$$\begin{aligned} f_{\hat{x}_j|x_j}(\alpha, \beta) &= \frac{f_{\hat{x}_j, x_j}(\alpha, \beta)}{f_{x_j}(\beta)} = \frac{G_{0, \Sigma_{\hat{x}_j, x_j}}\begin{pmatrix}\alpha \\ \beta\end{pmatrix}}{G_{0, \lambda_j}(\beta)} \\ &= \frac{1}{\sqrt{2\pi \lambda_j \tau_j s_j}} \exp\left(-\frac{1}{2} \frac{\left[\alpha - s_j \beta\right]^2}{\lambda_j \tau_j s_j}\right) \end{aligned}$$

where $\tau_j = \min\{1, \theta/\lambda_j\} \in [0, 1]$, and $s_j = 1 - \tau_j$. Note that, $f_{\hat{x}_j|x_j}(\alpha, \beta)$ becomes $\delta(\alpha)$ for $\tau_j \to 1$ (maximum distortion of this component implies that the corresponding output is set to 0) and $\delta(\alpha - \beta)$ for $\tau_j \to 0$ (no distortion of this component, the output is equal to the input).

We may collect the component-wise PDFs into a vector PDF by using the matrix $T_\theta = \operatorname{diag}(\tau_0, \ldots, \tau_{n-1}) = \min\{I_n, \theta(\Sigma^{\text{ok}})^{-1}\}$, and the matrix $S_\theta = I_n - T_\theta$ thus yielding the thesis. ∎

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Proof of Lemma 2: The PDF of $\hat{x}^{\mathrm{ko}}$ distorted by means of $f^{\mathrm{ok}}_{\hat{x}|x}$ can be computed as

$$f^{\mathrm{ko}}_{\hat{x}}(\alpha) = \int_{\mathbb{R}^n} f^{\mathrm{ko}}_{\hat{x},x}(\alpha,\beta)\mathrm{d}\beta = \int_{\mathbb{R}^n} f^{\mathrm{ok}}_{\hat{x}|x}(\alpha,\beta) f^{\mathrm{ko}}_{x}(\beta)\mathrm{d}\beta.$$

Assume first to be in the low-distortion condition $\theta < \lambda^{\mathrm{ok}}_{n-1}$ that implies $T_\theta = \theta(\Sigma^{\mathrm{ok}})^{-1}$, and write

$$\begin{aligned} f^{\mathrm{ko}}_{\hat{x}}(\alpha) &= \int_{\mathbb{R}^n} G_{S_\theta\beta,\Sigma^{\mathrm{ok}}S_\theta T_\theta}(\alpha) G_{0,\Sigma^{\mathrm{ko}}}(\beta)\mathrm{d}\beta \\ &= G_{0,\Sigma^{\mathrm{ok}}S_\theta T_\theta}(\alpha) \\ &\quad \times \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left[\beta^\top S_\theta (\Sigma^{\mathrm{ok}}S_\theta T_\theta)^{-1} S_\theta \beta - 2\alpha^\top (\Sigma^{\mathrm{ok}}S_\theta T_\theta)^{-1} S_\theta \beta\right]} \\ &\quad \times G_{0,\Sigma^{\mathrm{ko}}}(\beta)\mathrm{d}\beta \\ &= G_{0,\Sigma^{\mathrm{ok}}S_\theta T_\theta}(\alpha) \\ &\quad \times \frac{1}{\sqrt{(2\pi)^n \det \Sigma^{\mathrm{ko}}}} \underbrace{\int_{\mathbb{R}^n} e^{-\frac{1}{2}(\beta^\top Q\beta - 2q^\top\beta)}\mathrm{d}\beta}_{g(\alpha)} \end{aligned}$$

with $Q = S_\theta(\Sigma^{\mathrm{ok}}S_\theta T_\theta)^{-1}S_\theta + (\Sigma^{\mathrm{ko}})^{-1} = (\theta I_n)^{-1} - (\Sigma^{\mathrm{ok}})^{-1} + (\Sigma^{\mathrm{ko}})^{-1}$ and $q = (\Sigma^{\mathrm{ok}}S_\theta T_\theta)^{-1}S_\theta\alpha = \alpha/\theta$. To compute $g(\alpha)$ let $Q = \mathrm{U}\mathrm{D}\mathrm{U}^\top$ with D diagonal and U orthonormal, and set $\beta' = D^{1/2}U^\top\beta$ so that $\beta = UD^{-1/2}\beta'$ and $\mathrm{d}\beta = \mathrm{d}\beta'/\sqrt{\det Q}$. With this write

$$g(\alpha) = \frac{1}{\sqrt{\det Q}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(\beta'^\top\beta' - 2q^\top UD^{-1/2}\beta'\right)}\mathrm{d}\beta'$$

at the exponent of which one may add and subtract $q^\top Q^{-1}q = q^\top UD^{-1/2}D^{-1/2}U^\top q$ to yield

$$\begin{aligned} g(\alpha) &= \frac{1}{\sqrt{\det Q}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(\left\|\beta' - D^{-1/2}U^\top q\right\|^2 - q^\top Q^{-1}q\right)}\mathrm{d}\beta' \\ &= \sqrt{\frac{(2\pi)^n}{\det Q}} e^{\frac{1}{2}q^\top Q^{-1}q}. \end{aligned}$$

Putting this back into $f^{\mathrm{ok}}_{\hat{x}}$ we get

$$f^{\mathrm{ko}}_{\hat{x}}(\alpha) = G_{0,\left[(\theta I_n)^{-1} - (\Sigma^{\mathrm{ok}})^{-1} + (\Sigma^{\mathrm{ko}})^{-1}\right]\Sigma^{\mathrm{ko}}\Sigma^{\mathrm{ok}}S_\theta T_\theta}(\alpha).$$

A straightforward expansion of the definitions under the low-distortion assumption finally rearranges the covariance matrix into

$$\begin{aligned} &\left[(\theta I)^{-1} - \left(\Sigma^{\mathrm{ok}}\right)^{-1} + \left(\Sigma^{\mathrm{ko}}\right)^{-1}\right]\Sigma^{\mathrm{ko}}\Sigma^{\mathrm{ok}}S_\theta T_\theta \\ &= \left[(\theta I_n)^{-1} - \left(\Sigma^{\mathrm{ok}}\right)^{-1} + \left(\Sigma^{\mathrm{ko}}\right)^{-1}\right]\Sigma^{\mathrm{ko}}\Sigma^{\mathrm{ok}}\theta\left(\Sigma^{\mathrm{ok}}\right)^{-1}S_\theta \\ &= \left[\Sigma^{\mathrm{ko}} - \theta\left(\Sigma^{\mathrm{ok}}\right)^{-1}\Sigma^{\mathrm{ko}} + \theta I_n\right]S_\theta \\ &= \left[I_n - \theta\left(\Sigma^{\mathrm{ok}}\right)^{-1}\right]\Sigma^{\mathrm{ko}}S_\theta + \theta S_\theta \\ &= S_\theta\Sigma^{\mathrm{ko}}S_\theta + \theta S_\theta \end{aligned} \tag{26}$$

as in the statement of the lemma.

To address the case in which θ exceeds $\lambda^{\mathrm{ok}}_{n-1}$ note that for $\theta \to (\lambda^{\mathrm{ok}}_{n-1})^-$, the last diagonal entry of S_θ tends to 0 and thus by (26) the covariance tends to have zeros in its last row and column. Since a Gaussian with vanishing variance can be considered Dirac's delta, this model the fact that the last component of both x and x^{ko} is fully distorted and set to 0. With this, (26) is valid also for $\lambda^{\mathrm{ok}}_{n-1} < \theta < \lambda^{\mathrm{ok}}_{n-2}$. Yet, analogous considerations can be carried out for $\theta \to (\lambda^{\mathrm{ok}}_j)^-$ and $j = n-2, n-3, \ldots, 0$ so that (26) is valid for any value of θ. ∎

Proof of Lemma 3: Starting from the definition in (8), and assuming $x' \sim \mathcal{G}(0, \sum')$ and $x'' \sim \mathcal{G}(0, \sum'')$ then

$$\begin{aligned} L(x'; x'') &= -\int_{\mathbb{R}^n} G_{0,\Sigma'}(\alpha)\log_2\left[G_{0,\Sigma''}(\alpha)\right]\mathrm{d}\alpha \\ &= \frac{1}{2}\log_2\left[(2\pi)^n\left|\Sigma''\right|\right]\int_{\mathbb{R}^n} G_{0,\Sigma'}(\alpha)\mathrm{d}\alpha \\ &\quad + \frac{1}{2\ln 2}\int_{\mathbb{R}^n} \alpha^\top\left(\Sigma''\right)^{-1}\alpha\ G_{0,\Sigma'}(\alpha)\mathrm{d}\alpha \\ &= \frac{1}{2}\log_2\left[(2\pi)^n\left|\Sigma''\right|\right] + \frac{1}{2\ln 2}\mathrm{tr}\left[\left(\Sigma''\right)^{-1}\Sigma'\right] \end{aligned}$$

where the last summand has been computed as the expectation of a quadratic form in a Gaussian multivariate for which in [48, Ch. 3 and Corollary 3.2b.1] gives a formula. ∎

Proof of Theorem 1: From (19) we have that

$$\zeta_I = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta - 1} A_j(\theta)$$

with

$$A_j(\theta) = \frac{1}{\lambda^{\mathrm{ok}}_j}\left(1 - \frac{\theta}{\lambda^{\mathrm{ok}}_j}\right) + \frac{\theta}{\lambda^{\mathrm{ok}}_j} - 1.$$

Note that $A_j(\theta)$ is continuous and its derivative is $(\partial/\partial\theta)A_j = (1 - {}^1/_{\lambda^{\mathrm{ok}}_j})/\lambda^{\mathrm{ok}}_j$.

For simplicity's sake assume $\lambda^{\mathrm{ok}}_0 > \lambda^{\mathrm{ok}}_1 > \cdots > \lambda^{\mathrm{ok}}_{n-1} > 0$, set $\lambda^{\mathrm{ok}}_n = 0$, and define $\Theta_j =]\lambda^{\mathrm{ok}}_{j+1}, \lambda^{\mathrm{ok}}_j[$ for $j = 0, \ldots, n-1$ so that if $\theta \in \Theta_j$ then $n_\theta = j + 1$.

As a function of θ, ζ_I is continuous. In fact, it is trivially continuous in each Θ_j. Yet, it is continuous also at any chosen $\lambda^{\mathrm{ok}}_{\bar{\jmath}}$ with $\bar{\jmath} = 0, \ldots, n-1$. To see why, note that

$$\begin{aligned} \lim_{\theta\to\lambda^{\mathrm{ok}-}_{\bar{\jmath}}} \zeta_I &= \frac{1}{2\ln 2}\lim_{\theta\to\lambda^{\mathrm{ok}-}_{\bar{\jmath}}}\sum_{j=0}^{\bar{\jmath}} A_j(\theta) \\ &= \frac{1}{2\ln 2}\lim_{\theta\to\lambda^{\mathrm{ok}-}_{\bar{\jmath}}} A_{\bar{\jmath}}(\theta) + \sum_{j=0}^{\bar{\jmath}-1} A_j(\theta) \\ &= \frac{1}{2\ln 2}\lim_{\theta\to\lambda^{\mathrm{ok}+}_{\bar{\jmath}}}\sum_{j=0}^{\bar{\jmath}-1} A_j(\theta) = \lim_{\theta\to\lambda^{\mathrm{ok}+}_{\bar{\jmath}}} \zeta_I \end{aligned}$$

where we have exploited that the $A_j(\theta)$ are continuous and thus their left and right limits coincide, and that $A_{\bar{\jmath}}(\lambda^{\mathrm{ok}}_{\bar{\jmath}}) = 0$.

On the left-hand side of its domain, when $\theta = \lambda^{\mathrm{ok}}_n = 0$ (no distortion), we have $n_\theta = n$ and thus

$$\zeta_I = \frac{1}{2\ln 2}\sum_{j=0}^{n-1}\left(\frac{1}{\lambda^{\mathrm{ok}}_j} - 1\right) \geq 0$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

where the last inequality follows from the fact that $\sum_{j=0}^{n} \lambda_j^{\text{ok}} = n$ and thus $\sum_{j=0}^{n} 1/\lambda_j^{\text{ok}} \geq n$.

On the right-hand side of its domain, when $\theta = \lambda_0^{\text{ok}}$ (maximum distortion), we have $n_\theta = 0$ and thus $\zeta_I = 0$. Yet, we also have that

$$\frac{\partial}{\partial\theta}\zeta_I = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta-1}\frac{1}{\lambda_j^{\text{ok}}}\left(1-\frac{1}{\lambda_j^{\text{ok}}}\right)$$

in which the summands are positive if $\lambda_j^{\text{ok}} > 1$. Hence, if $\bar{k} = \arg\max_k\{\lambda_k^{\text{ok}} \geq 1\}$, for $\theta \geq \lambda_{\bar{k}}^{\text{ok}}$, all the summands in the above expression are positive and thus $(\partial/\partial\theta)\zeta_I > 0$ for $\lambda_{\bar{k}}^{\text{ok}} < \theta \leq \lambda_0^{\text{ok}}$. Given that $\zeta_I = 0$ at the end of that interval, it must be negative in its interior.

Since we know that ζ_I is positive for $\theta = \lambda_n^{\text{ok}} = 0$ and it is continuous for $\theta \in [\lambda_n^{\text{ok}}, \lambda_0^{\text{ok}}]$, it must pass through zero at least once whenever it is not negative, i.e., for $0 < \theta < \lambda_{\bar{k}}^{\text{ok}}$. ∎

Proof of Theorem 2: We will use the following lemma whose proof follows this one.

Lemma 4: If $\lambda^{\text{ko}} \sim \mathcal{U}(\mathbb{S}^n)$, then for any integrable function $f : \mathbb{R} \to \mathbb{R}$ and any $j = 0, \ldots, n-1$

$$\mathbf{E}\Big[f\Big(\lambda_j^{\text{ko}}\Big)\Big] = \frac{n-1}{n^{n-1}}\int_0^n f(p)(n-p)^{n-2}\mathrm{d}p.$$

To prove our thesis we start by writing $\Delta_{\text{F}} = n^{-\beta}\|\Sigma^{\text{ko}} - I_n\|_F$ as follows:

$$\Delta_{\text{F}} = \frac{1}{n^\beta}\sqrt{\text{tr}\Big[\big(\Sigma^{\text{ko}} - I_n\big)^\top\big(\Sigma^{\text{ko}} - I_n\big)\Big]}.$$

By noting that U^{ko} is orthonormal and thus that $\Sigma^{\text{ko}} - I_n = U^{\text{ko}}(\Lambda^{\text{ko}} - I_n)U^{\text{ko}\top}$ we get

$$\begin{aligned}\Delta_{\text{F}} &= \frac{1}{n^\beta}\sqrt{\text{tr}\Big[U^{\text{ko}}\big(\Lambda^{\text{ko}} - I_n\big)^2 U^{\text{ko}\top}\Big]}\\ &= \frac{1}{n^\beta}\sqrt{\text{tr}\Big[\big(\Lambda^{\text{ko}} - I_n\big)^2 U^{\text{ko}\top}U^{\text{ko}}\Big]}\\ &= \frac{1}{n^\beta}\sqrt{\sum_{k=0}^{n-1}\big(\lambda_k^{\text{ko}} - 1\big)^2}.\end{aligned}$$

Starting from the above expression one has

$$\begin{aligned}\mathbf{E}\Big[\Delta_{\text{F}}^2\Big] &= \frac{1}{n^{2\beta}}\sum_{k=0}^{n-1}\mathbf{E}\Big[\Big(\lambda_k^{\text{ko}} - 1\Big)^2\Big]\\ &= \frac{1}{n^{2\beta}}\sum_{k=0}^{n-1}\frac{n-1}{n+1} = \frac{1}{n^{2\beta-1}}\frac{n-1}{n+1}\end{aligned}$$

where we have used Lemma 4 to compute the expectation.

Hence, when $\beta > 1/2$

$$\mathbf{E}\Big[\Delta_{\text{F}}^2\Big] \xrightarrow[n\to\infty]{} 0 \tag{27}$$

that can be plugged into the Markov inequality to obtain

$$\Pr\Big(\Delta_{\text{F}}^2 \geq \bar{\Delta}\Big) \leq \frac{\mathbf{E}\big[\Delta_{\text{F}}^2\big]}{\bar{\Delta}} \quad \forall\bar{\Delta} > 0$$

and thus that Δ_{F} converges to 0 in probability with increasing n. ∎

Proof of Lemma 4: For any function $f : \mathbb{R} \mapsto \mathbb{R}$ we have

$$\begin{aligned}\mathcal{I}\big[f(p)\big] &= \int_{\mathbb{S}^n} f(p_0)\mathrm{d}p_0\ldots\mathrm{d}p_{n-1}\\ &= \int_0^n f(p_0)\int_0^{n-p_0}\int_0^{n-p_0-p_1}\cdots\\ &\qquad\int_0^{n-p_0-p_1-\cdots-p_{n-3}}\mathrm{d}p_0\ldots\mathrm{d}p_{n-2}\\ &= \int_0^n f(p_0)\frac{(n-p_0)^{n-2}}{(n-2)!}\mathrm{d}p_0.\end{aligned}$$

Since λ^{ko} is uniformly distributed over $\mathbb{S}^n$ the probability density is the constant $1/\mathcal{I}[1] = n^{-(n-1)}(n-1)!$ and the expectation of f is

$$\begin{aligned}\mathbf{E}\Big[f\Big(\lambda_j^{\text{ko}}\Big)\Big] &= n^{-(n-1)}(n-1)!\mathcal{I}\big[f(p)\big]\\ &= \frac{(n-1)!}{n^{n-1}}\int_0^n f(p)\frac{(n-p)^{n-2}}{(n-2)!}\mathrm{d}p\\ &= \frac{n-1}{n^{n-1}}\int_0^n f(p)(n-p)^{n-2}\mathrm{d}p.\end{aligned}$$

∎

ACKNOWLEDGMENT

This article reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] H. He and Y. Tan, "Unsupervised classification of multivariate time series using VPCA and fuzzy clustering with spatial weighted matrix distance," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1096–1105, Mar. 2020.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[3] A. Burrello, A. Marchioni, D. Brunelli, S. Benatti, M. Mangia, and L. Benini, "Embedded streaming principal components analysis for network load reduction in structural health monitoring," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4433–4447, Mar. 2021.

[4] F. Zonzini, M. Zauli, M. Mangia, N. Testoni, and L. De Marchi, "Model-assisted compressed sensing for vibration-based structural health monitoring," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7338–7347, Nov. 2021.

[5] F. Pilati, G. Lelli, A. Regattieri, and E. Ferrari, "Assembly line balancing and activity scheduling for customised products manufacturing," *Int. J. Adv. Manuf. Technol.*, vol. 120, pp. 3925–3946, May 2022.

[6] K. Yan, Z. Ji, H. Lu, J. Huang, W. Shen, and Y. Xue, "Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1349–1356, Jul. 2019.

[7] H. Chen, H. Yi, B. Jiang, K. Zhang, and Z. Chen, "Data-driven detection of hot spots in photovoltaic energy systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 8, pp. 1731–1738, Aug. 2019.

[8] M. Mangia, J. Haboba, R. Rovatti, and G. Setti, "Rakeness-based approach to compressed sensing of ECGs," in *Proc. IEEE Biomed. Circuits Systems Conf. (BioCAS)*, 2011, pp. 424–427.

[9] D. Bortolotti, M. Mangia, A. Bartolini, R. Rovatti, G. Setti, and L. Benini, "An ultra-low power dual-mode ECG monitor for healthcare and wellness," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, 2015, pp. 1611–1616.

[10] M. Mangia, L. Prono, A. Marchioni, F. Pareschi, R. Rovatti, and G. Setti, "Deep neural oracles for short-window optimized compressed sensing of biosignals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 3, pp. 545–557, Jun. 2020.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[11] G. He, Y. Pan, X. Xia, J. He, R. Peng, and N. N. Xiong, "A fast semi-supervised clustering framework for large-scale time series data," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 7, pp. 4201–4216, Jul. 2021.
[12] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
[13] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.
[14] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.
[15] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37-th Annu. Allerton Conf. Commun. Control Comput.*, 1999, pp. 368–377.
[16] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.*, 1999, pp. 617–623.
[17] K. Crammer and G. Chechik, "A needle in a haystack: Local one-class optimization," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 26.
[18] K. Crammer, P. P. Talukdar, and F. Pereira, "A rate-distortion one-class model and its applications to clustering," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 184–191.
[19] C. Böhm, K. Haegler, N. S. Müller, and C. Plant, "CoCo: Coding cost for parameter-free outlier detection," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2009, pp. 149–158.
[20] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6228–6237.
[21] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 675–685.
[22] M. R. D. Rodrigues, N. Deligiannis, L. Lai, and Y. C. Eldar, "Rate-distortion trade-offs in acquisition of signal parameters," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 6105–6109.
[23] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Hardware-limited task-based quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, Oct. 2019.
[24] B. Foo, Y. Andreopoulos, and M. van der Schaar, "Analytical complexity modeling of wavelet-based video coders," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 3, 2007, pp. 789–792.
[25] L. Vu, V. L. Cao, Q. U. Nguyen, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "Learning latent representation for iot anomaly detection," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3769–3782, May 2022.
[26] C. Huang et al., "Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13834–13847, Dec. 2022.
[27] C. Yin, S. Zhang, J. Wang, and N. N. Xiong, "Anomaly detection based on convolutional recurrent autoencoder for IoT time series," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 1, pp. 112–122, Jan. 2022.
[28] A. Kolmogorov, "On the shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inf. Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.
[29] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Hoboken, NJ, USA: Prentice Hall, 1988.
[30] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Royal Soc. London. Series A. Math. Phys. Sci.*, vol. 186, no. 1007, pp. 453–461, 1946.
[31] K. Solomon, *Information Theory and Statistics*. Hoboken, NJ, USA: Wiley, 1959.
[32] M. Mangia, F. Pareschi, V. Cambareri, R. Rovatti, and G. Setti, "Rakeness-based design of low-complexity compressed sensing," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 5, pp. 1201–1213, May 2017.
[33] V. Cambareri, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "A rakeness-based design flow for analog-to-information conversion by compressive sensing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2013, pp. 1360–1363.
[34] S. Onn and I. Weissman, "Generating uniform random vectors over a simplex with implications to the volume of a certain polytope and to multivariate extremes," *Ann. Oper. Res.*, vol. 189, no. 1, pp. 331–342, 2011.
[35] F. Mezzadri, "How to generate random matrices from the classical compact groups," *NoticesTrans. Amer. Math. Soc.*, vol. 54, no. 5, pp. 592–604, 2006.
[36] J. Ginibre, "Statistical ensembles of complex, quaternion, and real matrices," *J. Math. Phys.*, vol. 6, no. 3, pp. 440–449, 1965.
[37] A. Marchioni, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "Subspace energy monitoring for anomaly detection @sensor or @edge," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7575–7589, Aug. 2020.
[38] A. Moallemi, A. Burrello, D. Brunelli, and L. Benini, "Exploring scalable, distributed real-time anomaly detection for bridge health monitoring," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17660–17674, Sep. 2022.
[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980
[41] R. B. Arellano-Valle, J. E. Contreras-Reyes, and M. G. Genton, "Shannon entropy and mutual information for multivariate skew-elliptical distributions," *Scandinavian J. Statist.*, vol. 40, no. 1, pp. 42–62, 2013.
[42] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 1999, pp. 582–588.
[43] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[44] H. Jamalabadi, S. Alizadeh, M. Schönauer, C. Leibold, and S. Gais, "Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers," *Human Brain Map.*, vol. 37, no. 5, pp. 1842–1855, May 2016.
[45] L. Snoek, S. Miletić, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *NeuroImage*, vol. 184, pp. 741–760, Jan. 2019.
[46] P. McSharry, G. Clifford, L. Tarassenko, and L. Smith, "A dynamical model for generating synthetic electrocardiogram signals," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 3, pp. 289–294, Mar. 2003.
[47] M. Mangia, R. Rovatti, and G. Setti, "Rakeness in the design of analog-to-information conversion of sparse and localized signals," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 5, pp. 1001–1014, May 2012.
[48] S. B. Provost and A. Mathai, *Quadratic Forms in Random Variables: Theory and Applications/ A.M. Mathai, Serge B. Provost* (Statistics: Textbooks and Monographs). New York, NY, USA: Marcel Dekker, 1992.

Alex Marchioni (Member, IEEE) received the B.Sc. and M.Sc. degrees (with Hons.) in electronic engineering and the Ph.D. degree in electronic, telecommunication, and information technology from the University of Bologna, Bologna, Italy, in 20011, 2015, and 2022, respectively.

He is currently a Research Fellow with the Department of Electrical, Electronic, and Information Engineering, University of Bologna, where he is also a member of the Statistical Signal Processing Group. His research interests are in signal processing, machine learning, anomaly detection, compressed sensing, Internet of Things, and big data analytics.

Andriy Enttsel (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees (with Hons.) in electronic engineering from the University of Bologna, Bologna, Italy, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electronics, telecommunications, and information technologies engineering with Statistical Signal Processing Group.

His primary research focuses on signal processing, anomaly detection, and information theory.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Mauro Mangia (Member, IEEE) received the B.Sc. and M.Sc. degrees in electronic engineering and the Ph.D. degree in information technology from the University of Bologna, Bologna, Italy, in 2005, 2009, and 2013, respectively.

He was a visiting Ph.D. student with the École Polytechnique Federale de Lausanne, Lausanne, Switzerland, in 2009 and 2012. He is currently an Assistant Professor with the Department of Electrical, Electronic, and Information Engineering, University of Bologna within the Statistical Signal Processing Group. He is also a member of both the Advance Research Center for Electronic Systems ARCES, University of Bologna, and the Alma Mater Research Institute for Human-Centered Artificial Intelligence, University of Bologna. His research interests are in nonlinear systems, explainable artificial intelligence, machine learning and AI, anomaly detection, Internet of Things, big data analytics, and optimization.

Dr. Mangia was a recipient of the 2013 IEEE CAS Society Guillemin–Cauer Award and of the 2019 IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS Best Paper Award. He received the Best Student Paper Award at ISCAS2011. He was the Web and Social Media Chair for ISCAS2018.

Riccardo Rovatti (Fellow, IEEE) received the M.S. degree in electronic engineering and the Ph.D. degree in electronics, computer science, and telecommunications from the University of Bologna, Bologna, Italy, in 1992 and 1996, respectively.

He is currently a Full Professor of Electronics with the University of Bologna. He has authored more than 300 technical contributions to international conferences and journals and two volumes. His research focuses on mathematical and applicative aspects of statistical signal processing, on machine learning for signal processing, and on the application of statistics to nonlinear dynamical systems.

Prof. Rovatti was a recipient of the 2004 IEEE CAS Society Darlington Award, the 2013 IEEE CAS Society Guillemin–Cauer Award, and the 2019 IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS Best Paper Award. He received the Best Paper Award at ECCTD 2005 and the Best Student Paper Award at the EMC Zurich 2005 and ISCAS 2011. He is an IEEE Fellow for his contribution to nonlinear and statistical signal processing applied to electronic systems. He was a Distinguished Lecturer of the IEEE CAS Society for the years 2017–2018.

Gianluca Setti (Fellow, IEEE) received the Dr.Eng. (Hons.) and Ph.D. degrees in electronic engineering from the University of Bologna, Bologna, Italy, in 1992 and 1997, respectively.

From 1997 to 2017, he was with the Department of Engineering, University of Ferrara, Ferrara, Italy, as an Assistant Professor from 1998 to 2000, an Associate Professor from 2001 to 2008, and as a Professor from 2009 to 2017 of Circuit Theory and Analog Electronics. From 2017 to 2022, he was a Professor of Electronics, Signal, and Data Processing with the Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy. Since November 2022, he has been the Dean of the Computer, Electrical, Mathematical Sciences, and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. He held several positions as a Visiting Professor/a Scientist with EPFL, Lausanne, Switzerland, in 2002 and 2005; the University of California at San Diego, CA, USA, in 2004; IBM, Armonk, New York, USA, in 2004 and 2007; and the University of Washington, Seattle, WA, USA, in 2008 and 2010; and he is also a permanent (in-kind) Faculty Member of ARCES, University of Bologna. He was the Co-Editor of *Chaotic Electronics in Telecommunications* (CRC Press, Boca Raton, 2000), *Circuits and Systems for Future Generations of Wireless Communications* (Springer, 2009), and *Design and Analysis of Biomolecular Circuits* (Springer, 2011), coauthored *Adapted Compressed Sensing for Effective Hardware Implementations* (2018) as well as one of the guest editors of the May 2002 special issue of the IEEE Proceedings on "Applications of Non-linear Dynamics to Electronic and Information Engineering." His research interests include nonlinear circuits, recurrent neural networks, electromagnetic compatibility, compressive sensing and statistical signal processing, biomedical circuits and systems, power electronics, design and implementation of IoT nodes, circuits and systems for machine learning, and applications of AI techniques for anomaly detection and predictive maintenance.

Dr. Setti received the 2013 IEEE CAS Society Meritorious Service Award, he is co-recipient of the 2004 IEEE CAS Society Darlington Award, the 2013 IEEE CAS Society Guillemin–Cauer Award, the 2019 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS Best Paper Award, the Best Paper Award at ECCTD2005, and the Best Student Paper Award at EMCZurich2005 and at ISCAS2011. He also received the 1998 Caianiello Prize for the Best Italian Ph.D. thesis on Neural Networks. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS from 1999 to 2002 and from 2002 to 2004, and for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS from 2004 to 2007, as the Deputy-Editor-in-Chief, for the *IEEE Circuits and Systems Magazine* from 2004 to 2007, as well as the Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS from 2006 to 2007, and of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS from 2008 to 2009. He also served in the Editorial Board of IEEE ACCESS from 2013 to 2015 and of the PROCEEDINGS OF THE IEEE from 2015 to 2018. Since 2019, he has been serving as the first non-U.S. Editor-in-Chief of the PROCEEDINGS OF THE IEEE, the flagship journal of the Institute. In 2012, he was the Chair of the IEEE Strategic Planning Committee of the Publication Services and Products Board and from 2013 to 2014, he was the first non-North-American Vice President of the IEEE for Publication Services and Products. He served in the program committee of many conferences and was, in particular, the Special Sessions Co-Chair of ISCAS2005 (Kobe) and ISCAS2006 (Kos), the Technical Program Co-Chair of NDES2000 (Catania), ISCAS2007 (New Orleans), ISCAS2008 (Seattle), ICECS2012 (Seville), BioCAS2013 (Rotterdam), and MWSCAS2023 (Phoenix), as well as the General Co-Chair of NOLTA2006 (Bologna) and ISCAS2018 (Florence). He was a Distinguished Lecturer of the IEEE CAS Society from 2004 to 2005 and from 2013 to 2014 of IEEE CAS Society, a member of the CASS Board of Governors from 2005 to 2008, and served as the 2010 CAS Society President.

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI CARE Agreement