ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

*Dottorato di Ricerca in*

Ingegneria Elettronica, Telecomunicazioni e Tecnologie dell'Informazione

Ciclo 37

*Settore Concorsuale*: 09/E3 — Elettronica

*Settore Scientifico Disciplinare*: ING-INF/01 — Elettronica

# Anomaly Detection Challenges in Monitoring Applications

*Presentata da:*
Andriy Enttsel

*Coordinatore Dottorato:*
Davide Dardari

*Supervisore:*
Riccardo Rovatti

Esame finale anno 2025

*“What makes us the most normal,” said Reiko, “is knowing that we’re not normal”.*

Haruki Murakami, Norwegian Wood

ALMA MATER STUDIORUM — UNIVERSITY OF BOLOGNA

# *Abstract*

School of Engineering
Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" (DEI)

Ph.D. Programme in Electronics, Telecommunications and Information Technologies Engineering

**Anomaly Detection Challenges in Monitoring Applications**

by Andriy ENTTSEL

Monitoring physical systems has become pervasive, particularly in critical applications where ensuring operational integrity is paramount. A fundamental task in this context is identifying anomalous behaviors, commonly referred to as anomaly detection. Over the past years, significant attention has been devoted to anomaly detection, leading to the development of more accurate and robust algorithms capable of processing complex data. However, several fundamental challenges remain. Firstly, anomaly detection is inherently unsupervised, making it difficult to exploit prior knowledge about possible anomalies during the design phase. Secondly, despite advances in related fields, anomaly detection has not been thoroughly analyzed from an information-theoretic perspective. This lack of exploration complicates its integration with other well-established tasks, such as signal compression. This dissertation aims to address these challenges by presenting both practical and information-theoretic frameworks for anomaly detection.

In the first part, we design a tool designed to mitigate the challenge of evaluating anomaly detection performance in the absence of real anomalies. To this end, we develop robust mathematical models that emulate possible anomalies in time series data and propose a procedure for generating synthetic anomalies. Furthermore, we establish a theoretical framework for performance assessment based on a novel concept of distinguishability.

In the second part, we employ these assessment tools to study the interaction between compression and anomaly detection from an information-theoretic standpoint. We demonstrate that common lossy compression algorithms can compromise the effectiveness of anomaly detection performed on compressed data. We then study how these tasks can be jointly optimized and offer insights for developing practical systems that integrate both functions. In a similar spirit, we design an autoencoder-based compression scheme that not only minimizes distortion but also preserves information critical for anomaly detection.

# *Acknowledgments*

When you come to the end of the doctorate journey, it feels natural to reflect on the academic achievements and professional growth of the past three years. While this dissertation should highlight those aspects, I want to use this short space to focus on human connections and relationships that have shaped this experience and appreciate it as a whole.

Throughout these three years, I lived just five minutes from the lab, and even when I moved back in with my parents, I had never thought of skipping a day. Of course, I enjoyed what I have been working on and, of course, I was exploring fascinating topics. Yet, I truly believe that the main driving force to go to the lab each day was my "fascinating colleagues". In the lab I was not just writing papers, I was building new and consolidating old friendships. All the stupid things we brought to light, all the memes we created, made these years much less of an agony and can be considered, without a doubt, our major contributions to society.

Coming home after work was an instant relief from the inevitable stresses of the day, thanks to the warm atmosphere created by the house and my "flatmates". All the meals we have shared, all the nights we have spent together, and all the parties we have thrown in our "ballroom" made this domestic partnership a domestic friendship.

I must also thank my long-standing friends, who always encouraged me and tolerated my absence with the ever-present excuse, "I have to study".

Of course, no acknowledgment would be complete without recognizing the mentorship. I am deeply grateful to my seniors for the invaluable things they taught me and for their constant guidance in the intricate academic realm. I would also like to apologize for my stubbornness and the stressful discussion I made them go through.

Most importantly, I want to thank my family for supporting me and believing in my capabilities during this time. I am especially grateful to my mother, whose courage and love gave me the unique opportunity to live a better and more peaceful life.

Finally, a special thanks to the U.S. consular officer whose professionalism gave me time to focus on this dissertation instead of enjoying autumn in New York.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **ACC** | **Acc**elerometer |
| **AD** | **A**nomaly **D**etection |
| **AE** | **A**uto**E**ncoder |
| **AEC** | **A**uto**E**ncoder **C**ompression |
| **APP** | **App**lication |
| **AR** | **A**uto **R**egressive Model |
| **AUC** | **A**rea **U**nder the **C**urve |
| **DbD** | **D**istribution-**b**ased **D**etectors |
| **DCT** | **D**iscrete **C**osine **T**ransform |
| **DEC** | **DEC**oder |
| **DL** | **D**eep **L**earning |
| **DNNC** | **D**eep **N**eural **N**etwork **C**lassifier |
| **DPCC** | **D**isciplined **C**onvex-**C**oncave **P**rogramming |
| **DWT** | **D**iscrete **W**avelet **T**ransform |
| **ENC** | **ENC**oder |
| **GC** | **G**aussian **C**ompressed |
| **GD** | **G**aussian **D**istribution |
| **GNN** | **G**aussian **N**arrowband **N**oise |
| **GWN** | **G**aussian **W**hite **N**oise |
| **IF** | **I**solation **F**orest |
| **IoT** | **I**nternet **o**f **T**hings |
| **KLT** | **K**arhunen-**L**oève **T**ransform |
| **LbD** | **L**earning-**b**ased **D**etectors |
| **LD** | **L**ikelihood **D**etector |
| **LOF** | **L**ocal **O**utlier **F**actor |
| **MD** | **M**ahalanobi **D**istance |
| **ML** | **M**achine **L**earning |
| **MSE** | **M**ean **S**quared **E**rror |
| **NGC** | **N**on-**G**aussian **C**ompressed |
| **NPD** | **N**eyman-**P**earson **D**etector |
| **OCSVM** | **O**ne **C**lass **S**upport **V**ector |
| **PCA** | **P**rincipal **C**omponent **A**nalysis |
| **PCC** | **P**rincipal **C**omponent **C**ompression |
| **PDF** | **P**robability **D**ensity **F**unction |
| **PS** | **P**rincipal-**S**ubspace |
| **PSNR** | **P**eak **S**ignal-to-**N**oise **R**atio |
| **RAE** | **R**enyi entropy-regularized **A**uto**E**ncoder |
| **RBF** | **R**adial **B**asis **F**unction |
| **RDC** | **R**ate-**D**istortion **C**ompression |
| **RDD** | **R**ate-**D**istortion-**D**istinguishability |
| **ROC** | **R**eceiver **O**perating **C**haracteristic |
| **RSNR** | **R**econstruction **S**ignal-to-**N**oise **R**atio |

| | |
|---|---|
| **SAE** | **S**hrink **A**uto**E**ncoder |
| **SPE** | **S**quared **P**rediction **E**rror |
| **SVDD** | **S**upport **V**ector **D**ata **D**escription |
| **TV** | **T**otal **V**ariation |
| **VQ** | **V**ector **Q**uantization |
| **WOMBATS** | **W**ide **O**pen **M**odel **B**ased **A**nomaly **T**est **S**uite |
| **ZC** | **Z**ero-**C**rossing |

# List of Symbols

| | |
|---|---|
| $\mathbb{R}$ | set of the real numbers |
| $\mathbf{v}$ | column vector |
| $v_j$ | $j$-th element of a vector $\mathbf{v}$ |
| $\mathbf{M}$ | matrix |
| $M_{i,j}$ | $(i,j)$-th element of a matrix $\mathbf{M}$ |
| $\mathrm{tr}$ | trace of a matrix |
| $.^{\top}$ | transposition operator |
| $\mathrm{diag}(\cdot)$ | diagonal matrix with a vector argument as diagonal |
| $\vert\cdot\vert$ | absolute value |
| $\Vert\cdot\Vert$ | norm |
| $\mathbf{I}_n$ | identity matrix with dimension $n \times n$ |
| $\boldsymbol{\Sigma}$ | covariance matrix |
| $\mathbf{x}$ | input signal |
| $n$ | input signal dimension |
| $\mathbf{y}$ | transformed signal |
| $k$ | transformed signal dimension |
| $\mathbf{z}$ | compressed signal |
| $\hat{\mathbf{x}}$ | reconstructed signal |
| $\mathcal{L}_{\times}$ | Localization of the signal $\times$ |
| $f_{\times}$ | probability density function of $\times$ |
| $F_{\times}$ | cumulative distribution function of $\times$ |
| $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ | Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ |
| $\mathcal{U}\left(\cdot\right)$ | uniform distribution on the set defined by the argument |
| $\mathrm{Prob}\{\cdot\}$ | probability |
| $\mathbf{E}[\cdot]$ | expectation |
| $\mathrm{ok}$ | normal behavior |
| $\mathrm{ko}$ | anomaly |

*Моїй мамі.*

# Introduction

Physical reality is constantly monitored, analyzed, and evaluated to guide decisions or actions to modify it. This is achieved by integrating sensors into natural environments, engineered systems, or even the human body, creating a cyber-physical system [69, 66, 110]. These sensors convert physical quantities of interest into electrical signals. Once acquired, the signals are digitized for processing by digital systems, effectively capturing the physical dynamics as data streams. These data streams are frequently transmitted to cloud facilities for further analysis, where algorithms are used to extract valuable information for various applications.

A critical task in monitoring scenarios is to assess whether the system's behavior is typical. When behavior deviates from the norm, this is reflected in the signals and should be flagged by an *anomaly detection* algorithm [2]. Rapid and accurate identification of abnormal behaviors is vital for counteracting fraud, repairing malfunctions, or addressing health issues before they lead to financial loss or harm to human well-being.

Despite yielding a simple binary outcome—normal or anomalous—anomaly detection is not trivial and faces several challenges. High dimensionality, the non-stationary and heterogeneous nature of the signals, and noise contamination must all be addressed to ensure accurate decision-making [125]. While the deployment of more advanced algorithms can mitigate some of these issues, other problems remain unresolved. A fundamental challenge lies in the unsupervised nature of anomaly detection: anomalies are rarely known in advance, requiring algorithms to rely solely on signals that represent normal behavior. This unavailability of anomalies also complicates the process of assessing which algorithm will perform best for identifying specific behaviors, presenting a general *model selection* problem in signal processing [40, 51].

As sensors collect more signals, efficient management of computational, memory, and energy resources becomes imperative. A common approach is to exploit signal properties, such as redundancy, structure, and sparsity to create a more compact representation while retaining most of the original information [94, 87]. This process is known as compression, and when it results in some loss of information, it is referred to as lossy compression [54].

Although compression helps reduce energy consumption and transmission costs, its impact on anomaly detection performance remains an open question. Lossy compression has been extensively explored in signal processing and Information Theory, particularly through the *rate-distortion* curve, which illustrates the trade-off between the number of bits required to represent a signal and the loss of fidelity due to compression [54, 34]. However, the information-theoretic analysis of anomaly detection remains largely underexplored in the literature, complicating the understanding of how compression and detection performance interact [125]. In parallel, with recent advances in computer vision, supervised tasks such as classification and segmentation have been successfully integrated with lossy compression, optimizing it not only to maintain signal fidelity but also to preserve critical information for these tasks [147, 90, 88].

The objective of this dissertation is two-fold: *i*) to address the performance assessment problem of the anomaly detection task, and *ii*) to analyze, using information-theoretic methods, the synergy between compression and anomaly detection. This work is organized as follows:

Chapters 1 and 2 introduce the task of anomaly detection and the mechanisms of lossy compression. In particular, we provide the mathematical models and outline the algorithms that will be adopted throughout the dissertation.

**Part I** addresses the model selection and evaluation of the performance of the anomaly detection task. In Chapter 3, we define a framework for benchmarking anomaly detection algorithms that work with signals modeled as time series. The core of this framework is an extensive set of anomaly models designed to capture a wide range of effects that real-world anomalies have on the signal representing normal behavior. These models facilitate a synthetic injection of corruptions into normal signals, enabling realistic simulations of real-world anomalies. The effectiveness of this approach is validated in two monitoring scenarios, in which multiple anomaly detection techniques are evaluated. In Chapter 4, we focus on a more theoretical perspective. To measure anomaly detection performance, we introduce the concept of *distinguishability* between normal and anomalous sources and propose two information-theoretic metrics: one assumes prior knowledge of the anomalies, while the other is completely agnostic. Additionally, we establish a framework based on Gaussian signals to facilitate the derivation of theoretical results related to anomaly detection.

**Part II** explores the interplay between compression and anomaly detection. In Chapter 5, we examine how lossy compression techniques, governed by the rate-distortion curve, impact detection performance. To achieve this, we adapt the evaluation tools from Part I to work with compressed signals, allowing us to derive theoretical insights and validate them through practical applications. Next, in Chapter 6, we extend the rate-distortion framework by introducing and analyzing the *rate-distortion-distinguishability* trade-off. Considering Gaussian signals, we formulate and solve two optimization problems, one assuming prior knowledge of anomalies and the other operating without such knowledge, revealing how distinguishability is influenced by both compression rate and distortion. The observed theoretical trends are then confirmed through empirical tests in realistic scenarios. Finally, Chapter 7 focuses specifically on autoencoder-based compression. Here, we leverage the concept of distinguishability to design a novel loss function that includes a regularization term, enabling the compressor to optimize for both reconstruction and anomaly detection tasks. The effectiveness of this approach is validated in two different use cases and for various types of anomalies.

# Chapter 1

# Anomaly Detection

Anomaly detection (AD) plays a fundamental role in various monitoring applications, where the objective is to identify deviations from normal system behavior [2]. These deviations, commonly referred to as anomalies[1], may indicate important events such as equipment malfunctions, fraud in financial transactions, structural damage in buildings, or system failures in industrial processes. Prompt detection of such anomalies can prevent catastrophic outcomes, making AD a crucial tool in ensuring the reliability and safety of complex systems.

In this chapter, we first present the mathematical model used to formalize the AD problem. Section 1.2 provides an overview of the most common detection methods, while the final section discusses a metric commonly adopted to evaluate detector performance.

## 1.1 Mathematical model

In a monitoring application, the system is tracked by acquiring at a specific time instant $t$ samples of a generic quantity $\mathbf{x}[t]$ representative of the system. In the AD context, $\mathbf{x}[t]$ has to be considered as a realization from one of two distinct sources: one representing normal behavior, $\mathbf{x}^{\mathrm{ok}}$, and the other representing an anomaly, $\mathbf{x}^{\mathrm{ko}}$. We model these two sources as discrete-time, stationary, $n$-dimensional stochastic processes, each generating independent and identically distributed (i.i.d.) vectors $\mathbf{x}^{\mathrm{ok}} \in \mathbb{R}^n$ and $\mathbf{x}^{\mathrm{ko}} \in \mathbb{R}^n$, with different probability density functions (PDFs) $f^{\mathrm{ok}}_{\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}^+$ and $f^{\mathrm{ko}}_{\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}^+$. As a result, at any given time $t$, the observed process is either $\mathbf{x}[t] = \mathbf{x}^{\mathrm{ok}}[t]$ or $\mathbf{x}[t] = \mathbf{x}^{\mathrm{ko}}[t]$. Since the vectors are i.i.d., the time index can be dropped.

The goal of an anomaly detector is to distinguish between the normal instance $\mathbf{x}^{\mathrm{ok}} \sim f^{\mathrm{ok}}_{\mathbf{x}}$ and the anomalous $\mathbf{x}^{\mathrm{ko}} \sim f^{\mathrm{ko}}_{\mathbf{x}}$. When a detector processes an input $\mathbf{x}$, whether normal $(\mathrm{ok})$ or anomalous $(\mathrm{ko})$, its output can be interpreted as a function $z(\mathbf{x})$ that assigns a *score* to the instance. The higher the scores, the more abnormal the behavior [2]. Typically a detector is designed only based on $f^{\mathrm{ok}}_{\mathbf{x}}$, as the anomalous source $f^{\mathrm{ko}}_{\mathbf{x}}$ is usually unknown. We refer to this case as *anomaly-agnostic*, which in the literature is commonly considered unsupervised AD. In contrast, when both $f^{\mathrm{ok}}_{\mathbf{x}}$ and $f^{\mathrm{ko}}_{\mathbf{x}}$ are available to the detector, we call this scenario *anomaly-aware*, which covers semi-supervised or supervised AD approaches [142, 108, 9, 127, 125]. The anomaly-aware scenario is less typical than the anomaly-agnostic AD, but it will be an interesting reference case.

## 1.2 Anomaly detectors

In the anomaly-agnostic scenario, a typical detector assigns a score to each instance $\mathbf{x}$ based on the negative log-likelihood, specifically $z(\mathbf{x}) = -\log f^{\mathrm{ok}}_{\mathbf{x}}(\mathbf{x})$. This score increases as the processed instance deviates from the distribution $f^{\mathrm{ok}}_{\mathbf{x}}$ representing the

[1] Depending on the context, anomalies may also be called outliers, or novelties.

normality, effectively measuring how unlikely the instance is to be generated by $f_{\mathbf{x}}^{\mathrm{ok}}$. This type of detector is referred to as a likelihood-based detector (LD).

In contrast, when the detector has access to both the normal and anomalous distributions, it computes the log-likelihood ratio, assigning the score $z(\mathbf{x}) = -\log f_{\mathbf{x}}^{\mathrm{ok}}(\mathbf{x})/f_{\mathbf{x}}^{\mathrm{ko}}(\mathbf{x}) = \log f_{\mathbf{x}}^{\mathrm{ko}}(\mathbf{x}) - \log f_{\mathbf{x}}^{\mathrm{ok}}(\mathbf{x})$. This ratio compares the likelihood of the instance under the anomalous distribution $f_{\mathbf{x}}^{\mathrm{ko}}$ relative to the distribution of normality $f_{\mathbf{x}}^{\mathrm{ok}}$, offering a more robust detection method. This approach comes from the Neyman-Pearson lemma, which provides an optimal framework for hypothesis testing [34, Theorem 12.7.1], [74, Theorem 3.1]. We refer to this detector as a Neyman-Pearson detector (NPD).

Both LD and NPD detection methods, are indeed valuable theoretical tools. However, in practical applications, not only is $f_{\mathbf{x}}^{\mathrm{ko}}$ unknown, making the anomaly-aware case less typical, but also $f_{\mathbf{x}}^{\mathrm{ok}}$ is usually not explicitly available, precluding the use of likelihood-based scores. Instead, what is typically accessible is data representing normal behavior. In the literature different techniques have been proposed to generate the score $z(\mathbf{x})$ from this data [29, 125], with each method trying to capture certain statistical characteristics of normal data. In the remainder of this section, we will provide a brief overview of common unsupervised detectors[2], highlighting their main attributes and hyperparameters.

### 1.2.1 Principal component analysis (PCA)-based

Principal Component Analysis (PCA) [70] aims at representing a signal in a reduced $k < n$-dimensional subspace, where the majority of the signal's energy is concentrated. When this subspace is built using normal data, anomalies can be detected by evaluating how well each data point fits the subspace. To this end, the first step involves the covariance matrix $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{E}\left[\mathbf{x}^{\mathrm{ok}}\mathbf{x}^{\mathrm{ok}\top}\right]$[3] estimation from the data. Next, a spectral decomposition of the covariance matrix is performed, yielding $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{U}^{\mathrm{ok}}\mathbf{\Lambda}^{\mathrm{ok}}\mathbf{U}^{\mathrm{ok}\top}$, where $\mathbf{U}^{\mathrm{ok}}$ is an orthonormal matrix $n \times n$ that contains the eigenvectors of $\mathbf{\Sigma}^{\mathrm{ok}}$, and $\mathbf{\Lambda}^{\mathrm{ok}}$ is a diagonal matrix of size $n \times n$ with the corresponding eigenvalues $\lambda_0^{\mathrm{ok}} \geq \lambda_1^{\mathrm{ok}} \geq \cdots \geq \lambda_{n-1}^{\mathrm{ok}} \geq 0$ on the diagonal. This allows us to identify the basis of the signal space $\mathbf{U}^{\mathrm{ok}}$ and the energy distribution in that space $\mathbf{\Lambda}^{\mathrm{ok}}$. The $k$ largest eigenvalues correspond to the most significant (principal) components and will identify the subspace defined by the matrix $\mathbf{U}_k^{\mathrm{ok}}$ containing the first $k$ columns of $\mathbf{U}^{\mathrm{ok}}$.

With this information, two families of detectors can be derived [160]:

- Squared prediction error (SPE$_k$): This detector analyzes the residual space defined by the eigenvectors that are not part of the principal components. The score is computed as

$$z(\mathbf{x}) = \left\|\mathbf{x} - \mathbf{U}_k^{\mathrm{ok}}\mathbf{U}_k^{\mathrm{ok}\top}\mathbf{x}\right\|^2. \tag{1.1}$$

- Hotelling's T-squared test (T$_k^2$): This detector aims at identifying anomalies within the principal subspace. It computes the score as

$$z(\mathbf{x}) = \left\|\left(\mathbf{\Lambda}_k^{\mathrm{ok}}\right)^{-1/2}\mathbf{U}_k^{\mathrm{ok}\top}\mathbf{x}\right\|^2, \tag{1.2}$$

  where $\mathbf{\Lambda}_k^{\mathrm{ok}}$ is the $k \times k$ upper-left submatrix of $\mathbf{\Lambda}^{\mathrm{ok}}$.

[2] In this dissertation, the anomaly-aware scenario will primarily be explored from a theoretical perspective, with a focus on NPD. In the few instances where practical supervised detectors are considered, they will be based on binary classifiers [142].

[3] Without loss of generality, we assume the mean vector of $\mathbf{x}^{\mathrm{ok}}$ to be null, i.e., $\mathbf{E}\left[\mathbf{x}^{\mathrm{ok}}\right] = \mathbf{0}$. If the mean is not zero, the data can be centered by subtracting the mean.

### 1.2.2 Gaussian distribution (GD)-based

These methods assume that normal data follows a multivariate Gaussian distribution. We recall two different detectors:

- Mahalanobis distance (MD) [29]: This approach calculates the score as the Mahalanobis distance, which measures how far the vector $\mathbf{x}$ is from the center of the estimated Gaussian distribution, with each component weighted by its variance. Formally, the score is:
$$z(\mathbf{x}) = \left\| (\mathbf{\Lambda}^{\mathrm{ok}})^{-1/2}\, {\mathbf{U}^{\mathrm{ok}}}^{\top} \mathbf{x} \right\|. \tag{1.3}$$

- Autoregressive model (AR$_p$) [2]: This method fits a linear regression model to predict future samples based on past samples. During testing, the score is computed from the residuals of these predictions. In particular, for a model of order $p \in [1, n)$, the score is derived as:
$$z(\mathbf{x}) = \frac{1}{n-p} \sum_{i=0}^{n-p-1} (x_{p+i} - \tilde{x}_i)^2 \tag{1.4}$$
where $\tilde{\mathbf{x}}$ contains the predictions for the last $n-p$ samples of $\mathbf{x}$, based on the first $p$ samples of the vector.

While MD and AR are simple and low-complexity detectors effective in simpler scenarios, they may struggle in more complex, real-world cases where the data deviates from the assumed Gaussian distribution.

### 1.2.3 Machine learning (ML)-based

- Local Outlier Factor (LOF$_h$) [22]: This detector relies on nearest neighbors to calculate the anomaly score for each data point. The score is computed as the ratio between the average local density around the $h$ nearest neighbors and the local density of the point itself.

- Isolation Forest (IF$_q$) [89]: This detection method is based on the principle that anomalies can be more easily isolated from the rest of the data. It constructs $q$ tree-like structures by recursively partitioning the dataset. The anomaly score depends on the average path length (or depth) required to isolate a given data point across multiple trees.

- One-Class Support Vector Machine (OCSVM$_{\mathrm{kernel},\,\nu}$) [130]: This technique finds a non-linear transformation that maps the input data into a higher-dimensional space where a simple hyperplane can separate normal and anomalous points. The score assigned to each point is the distance from the origin in this new space. The main hyperparameters are the *kernel* of the non-linear mapping and $\nu$, which is the lower bound of the fraction of vectors belonging to the training set used as support vectors. In the case of specific kernels, such as the radial basis function (RBF), OCSVM is equivalent to Support Vector Data Description (SVDD) [126].

### 1.2.4 Deep learning (DL)-based

Most of the techniques mentioned above have been adapted to neural networks [48, 76, 154, 126, 125, 156] to better capture non-linear data patterns. Some of the most notable examples include:

- Autoencoder (AE) [154]: This technique extends the concept of PCA by replacing the linear projection onto a subspace with a non-linear transformation onto a *manifold*. Transformation is performed by an encoder neural network $(\mathrm{ENC})$, while reconstruction is handled by a decoder neural network $(\mathrm{DEC})$. Both networks are trained simultaneously to best reconstruct normal data and then the reconstruction error is used as an anomaly score:

  $$z(\mathbf{x}) = \|\mathbf{x} - \mathrm{DEC}(\mathrm{ENC}(\mathbf{x}))\|^2 . \tag{1.5}$$

  Anomalous instances typically result in a high reconstruction error, as they deviate significantly from the learned distribution of the normal data.

- Deep Support Vector Data Description (Deep SVDD) [126]: This method generalizes the traditional SVDD framework [143] to a deep learning context. It uses a deep neural network $(\mathrm{ENC})$ to project the data into a latent space. $\mathrm{ENC}$ is trained to enclose the projected normal data within a minimum-volume hypersphere and anomalies are identified by evaluating the distance of data points from the center $\mathbf{c}$ of this hypersphere:

  $$z(\mathbf{x}) = \|\mathbf{c} - \mathrm{ENC}(\mathbf{x})\| . \tag{1.6}$$

  Points lying far from the center are classified as anomalous.

  Other techniques to mention are *i*) Variational AE [76] that extends upon the classical autoencoder to learn the distribution of the normal data $f_{\mathbf{x}}^{\mathrm{ok}}$, and uses the approximation to the negative log-likelihood as anomaly score; *ii*) in [156] the authors present the Deep Isolation Forest, while *iii*) the autoregressive model has been adapted to Recurrent Neural Networks [48].

## 1.3 Metrics

To make a binary decision, the detector's score $z(\mathbf{x})$ is compared against a threshold, which must be carefully chosen based on the specific requirements of the application. However, to assess detector performance independently of the threshold, a common metric is the Area Under the Curve $(\mathrm{AUC})$ of the Receiver Operating Characteristic (ROC), defined as:

$$\mathrm{AUC} = \mathrm{Prob}\{z(\mathbf{x}^{\mathrm{ko}}) > z(\mathbf{x}^{\mathrm{ok}})\} \tag{1.7}$$

The $\mathrm{AUC} \in [0, 1]$ represents the probability that, given a random normal instance and a random anomalous instance, the anomalous one will be scored higher. In practice, the $\mathrm{AUC}$ does not typically have a closed-form solution and is usually estimated from the detector's scores [45]. A significant challenge with the $\mathrm{AUC}$ is that it requires a sufficient number of anomalous examples to be reliably estimated. Typically, algorithms are evaluated on standard datasets containing normal and anomalous data [5, 41, 16, 125]. However, this approach limits evaluation to specific scenarios and does not fully account for real-world monitoring applications, where different normal data describes the system and anomalous examples are scarce. In Chapter 3, we will introduce an alternative method that addresses this issue by synthetically generating anomalous data based on the available normal data.

# Chapter 2

# Compression

The increasing growth of sensor networks and Internet of Things (IoT) devices highlights the need for efficient data acquisition systems [69, 66, 110]. These systems are crucial in various fields, including infrastructure and industrial process monitoring, environmental observation, and scientific research. By collecting and storing large volumes of data, they provide a detailed overview of the monitored environment, facilitating real-time analysis and informed decision-making. However, managing the vast amounts of generated data presents significant challenges in terms of storage, transmission, and processing.

Modern large-scale acquisition systems can typically be modeled as numerous sensing units, each acquiring an unknown physical quantity and transforming it into samples of random processes for transmission over a network. To reduce transmission bitrate, these sensor readings are often compressed aiming to retain useful information while reducing the data size [54, 59]. This compression can be either lossless or lossy. While lossless algorithms completely preserve information, lossy algorithms are often preferred due to significantly higher reduction factors. However, lossy techniques introduce a trade-off between the bitrate and the amount of information loss, which is generally addressed in Information Theory through the rate-distortion curve [34], a theoretical limit that defines the minimum required bitrate for a given maximum level of distortion.

In practical scenarios, a compression mechanism corresponds to a specific rate-distortion curve, and the compression level determined by the application identifies a point on that curve. This is the case of a wide range of scenarios that require the extraction and monitoring of features relevant to the system under monitoring, such as structural health systems [24, 165], industrial plant sensorization [115, 157, 30], and biomedical signal processing [98, 21, 97]. Lower signal distortion generally increases the chance of meeting the requirements of the main task.

Compression schemes are often asymmetric. In the specific case of sensor data, encoding is lightweight and designed for low-complexity devices, while decoding is more resource-intensive and typically performed in the cloud. This contrasts with video codecs, where the asymmetry is reversed: encoding is more complex, involving computationally intensive tasks to achieve high compression efficiency, whereas decoding is optimized to be lightweight, enabling real-time playback across a wide range of devices.

This chapter presents the main tools of lossy compression that will be utilized throughout the dissertation. Specifically, Section 2.1 recalls the rate-distortion theory, with a particular focus on its application to Gaussian sources. In Section 2.2, a brief review of some of the most well-known lossy compression schemes is presented.

## 2.1 Rate-Distortion Theory

Signals are compressed by encoding their information into symbols, which are then transmitted over a communication channel with a limited capacity, measured by the maximum number of symbols per second. When the available rate is insufficient, the compression process discards part of the information to comply with the constraints of the channel. This loss of information causes the receiver to observe a distorted version of the original signal. Intuitively, as the channel capacity decreases, the distortion increases. This relationship is thoroughly studied in the rate-distortion theory [34, Chapter 13].

In this context, we consider a system whose main function is to transmit the information content of a signal source $\mathbf{x}$ to a receiver over a channel with a constraint in rate. At any given time $t$, a signal instance $\mathbf{x}[t]$ is fed to an encoder generating a compressed version $\mathbf{z}[t]$, which can then be decompressed into an approximation $\hat{\mathbf{x}}[t] \subset \mathbb{R}^n$.

The constraint in rate is such that lossy compression must be adopted, which means that the encoding process is not injective and introduces some degree of distortion. The encoder is specifically designed for the source $\mathbf{x}$, which is modeled as an independent, discrete-time, stochastic process of dimension $n$.

Distortion can be defined as:

$$D = \mathbf{E}\left[\|\mathbf{x}[t] - \hat{\mathbf{x}}[t]\|^2\right] \tag{2.1}$$

where $\mathbf{E}[\cdot]$ denotes the expectation. The minimum rate $\rho$ required to achieve a maximum allowed distortion level $\delta$ is given by the rate-distortion optimization problem (RD) [34, Theorem 13.2.1]:

$$\text{(RD)} \quad \rho(\delta) = \begin{array}{l} \inf\limits_{f_{\hat{\mathbf{x}}|\mathbf{x}}} \mathcal{I}\left(\hat{\mathbf{x}};\mathbf{x}\right) \\ \text{s.t.} \quad D \leq \delta \end{array} \tag{2.2}$$

where $\mathcal{I}\left(\hat{\mathbf{x}};\mathbf{x}\right)$ is the mutual information between $\hat{\mathbf{x}}$ and $\mathbf{x}$ [34, Chapter 8], and $f_{\hat{\mathbf{x}}|\mathbf{x}}$ is the conditional probability density function (PDF) modeling the possibly stochastic relationship between the encoder and decoder. Although [34, Theorem 13.2.1] originally defines the rate-distortion function for discrete sources, it can also be derived for well-behaved continuous sources [34, Chapter 13], which is the case we will focus on.

For a memoryless source (which allows us to neglect the time index $t$) that generates vectors of independent zero-mean Gaussian variables, i.e., $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_0, \dots, \lambda_{n-1})$ with $\lambda_0 \geq \lambda_1 \geq \dots \geq \lambda_{n-1} \geq 0$, the solution to (2.2) is:

$$\rho = \frac{1}{2}\sum_{j=0}^{n-1} \log_2 \frac{\lambda_j}{\min\left\{\theta, \lambda_j\right\}} = -\frac{1}{2}\sum_{j=0}^{n-1} \log_2 \tau_j \tag{2.3}$$

$$\delta = \sum_{j=0}^{n-1} \min\left\{\theta, \lambda_j\right\} = \sum_{j=0}^{n-1} \lambda_j \tau_j \tag{2.4}$$

where $\theta \in [0, \lambda_0]$ is the so-called reverse *water-filling* parameter [34, Theorem 13.3.3], and $\tau_j = \min\left\{1, \theta/\lambda_j\right\}$ represents the fraction of energy lost to distortion along the $j$-th component.

The coding theorems behind this classical framework suggest that the optimal rate-distortion trade-off, as given by (2.2), is asymptotically achievable by encoding an increasing number of consecutive source symbols into a single block. This block is then decoded into a sequence of distorted symbols.

## 2.2 Compressors

Lossy compression techniques aim to reduce the size of the source data while tolerating some loss in precision or fidelity. The objective is to efficiently encode the data while

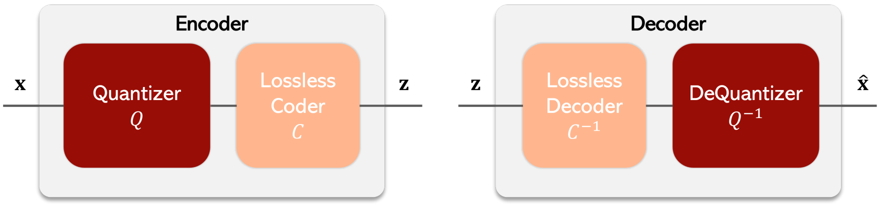


FIGURE 2.1: Main blocks of a lossy compressor. Encoder compresses the source, while the decoder recovers the signal from its compressed version by performing inverse operations of the encoder.

minimizing the reconstruction error. A practical lossy compressor can be represented as in Figure 2.1. The encoder $(\mathrm{ENC})$ compresses the source signal $\mathbf{x}$ into $\mathbf{z}$ by first discretizing it with a quantization stage [54] and then generates a bitstream by encoding each symbol with lossless compression techniques, such as Huffman coding or arithmetic coding [65, 122]. Lossless coders tend to assign shorter codes to more probable quantized values, which helps to reduce the number of bits required to represent the quantized data.

The decoder $(\mathrm{DEC})$ reconstructs the original signal from $\mathbf{z}$ by performing the inverse steps: lossless decoding followed by dequantization.

The rest of this section provides a concise overview of several widely used lossy compression methods.

### 2.2.1 Vector quantization-based compression

As rate-distortion theory suggests, it is more convenient to simultaneously compress multiple source symbols. The higher the number of symbols $n$ processed by the compressor, the better the high dimensionality of the signal space can be leveraged when discretizing this space [54]. This idea is closely related to the *sphere packing* phenomenon, where efficiently packing spheres in higher-dimensional spaces minimizes overlap and maximizes space utilization, leading to better compression performance. A natural scheme that arises from this consideration is vector quantization (VQ) [53]. In VQ, a quantizer partitions the $n$-dimensional space into cells by approximating the vectors using a finite set $\mathcal{C}$ of representative vectors called centroids. Formally, a vector quantizer is a mapping $Q : \mathbb{R}^n \to \mathcal{C} = \{\mathbf{c}_i, \dots, \mathbf{c}_{M-1}\} \subset \mathbb{R}^n$ that, given an input vector of dimension $n$, maps $\mathbf{x}$ to the nearest centroid $\mathbf{c}_i \in \mathcal{C}$ based on a distance metric (typically the Euclidean distance). The objective of VQ is to minimize the total distortion $D$ which in this case becomes:

$$D = \mathbf{E}\left[\|\mathbf{x} - \mathbf{c}_i\|_2^2\right]. \tag{2.5}$$

In practice, the design of an optimal set $\mathcal{C}$ is typically performed starting from a training dataset $\mathcal{X} = \{\mathbf{x}_0, \dots, \mathbf{x}_{N-1}\}$ via Lloyd's algorithm, a variant of k-means clustering, which iteratively refines the centroids to minimize the distortion. However, the complexity of designing an optimal VQ system increases exponentially with the input vector dimension $n$, the size of the dataset $N$, and the number of centroids $M$. At inference time, the search for the nearest centroid can also become computationally expensive, making this approach challenging for resource-constrained devices. Generally, other techniques are preferred to VQ.

### 2.2.2 Transform coding-based compression

Transform coding is a fundamental technique in lossy compression [52]. In this technique, data is first transformed from its original domain (often the time or spatial domain) into a new domain where it can be more efficiently represented. The key idea is that, in the transform domain, the coefficients are often *decorrelated*, which largely simplifies the compression process.

Once the signal components are decorrelated, simple scalar quantization (equivalent to VQ with $M = 1$) can be applied independently to each component. Transform coding represents a computationally efficient alternative to VQ: transformation is usually linear, implemented either through an $n \times n$ matrix multiplication or more efficient $O(n \log_2 n)$ algorithms. This approach largely simplifies quantization by allowing it to be performed separately for each transformed coefficient. Despite this simplification, transform coding offers good rate-distortion performance, as $n$ signal components are still compressed together.

The most well-known application of transform coding is in image compression algorithms such as JPEG and JPEG2000 [152, 139]. JPEG uses the Discrete Cosine Transform (DCT) [7] to transform $8 \times 8$ pixel blocks into the frequency domain, where most of the high-frequency components are discarded. JPEG2000 uses the Discrete Wavelet Transform (DWT) [120, 10] for multi-resolution image compression. In formats like MP3 and AAC, transform coding is applied to audio signals using a modified version of the DCT [23]. The coefficients corresponding to frequencies outside the range of human hearing are discarded to achieve lossy compression without noticeable quality degradation.

### Mathematical model

Given an $n$-dimensional signal $\mathbf{x}$, the goal of transform coding is to apply a linear transformation $\mathbf{T}$ to map $\mathbf{x}$ to a new set of coefficients $\mathbf{y}$ in a transformed domain. This transformation is represented as:

$$\mathbf{y} = \mathbf{T}\mathbf{x} \tag{2.6}$$

In most cases, the $n \times n$ transformation matrix $\mathbf{T}$ is orthogonal (i.e., $\mathbf{T}^\top \mathbf{T} = \mathbf{I}_n$), which implies that the inverse transformation can be easily performed to reconstruct the original data, i.e.,

$$\mathbf{x} = \mathbf{T}^{-1}\mathbf{y} = \mathbf{T}^\top \mathbf{y} \tag{2.7}$$

In reality, after transformation, each of the coefficients is quantized $z_j = Q(y_j)$, $j = 0, \ldots, n-1$ forming a new vector $\mathbf{z}$. The simplest form of quantization is uniform scalar quantization, where each coefficient is divided by a quantization width $q$ and rounded down to the nearest integer:

$$z_j = \left\lfloor \frac{y_j}{q} \right\rfloor, \tag{2.8}$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

After quantization, each component of $\mathbf{z}$ is then encoded using entropy coding techniques, such as Huffman or arithmetic coding to produce a bitstream. At the receiver, the quantized coefficients are decoded, and the inverse transformation $\mathbf{T}^\top$ is applied to $\mathbf{z}$ to reconstruct an approximation of the original signal $\hat{\mathbf{x}}$.

### Common transformations

Several transformations are commonly used in transform coding, each with specific properties that make it suitable for different types of data. Some of the most popular transforms are described below.

**Karhunen-Loève transform (KLT)** The Karhunen-Loève Transform (KLT) is a transform derived from the statistical properties of the data and is closely related to the

Principal Component Analysis (PCA) [8, 70]. The KLT is considered optimal in the sense that the transform diagonalizes the signal covariance matrix, thus minimizing redundancy among transformed components.

To compute the KLT, the covariance matrix of the signal, $\mathbf{\Sigma} = \mathbf{E}[\mathbf{x}\mathbf{x}^\top]$, is first estimated. The spectral decomposition of the covariance matrix is then performed, yielding

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{2.9}$$

where $\mathbf{U}$ is an $n \times n$ orthonormal matrix that contains the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ is a diagonal matrix of size $n \times n$ with the corresponding eigenvalues $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq 0$. The eigenvectors in $\mathbf{U}$ represent the directions along which the data is most spread out, while the eigenvalues in $\mathbf{\Lambda}$ quantify the amount of variance along each direction.

The KLT is applied by projecting the original signal $\mathbf{x}$ onto the eigenvector space:

$$\mathbf{y} = \mathbf{T}\mathbf{x} = \mathbf{U}^\top\mathbf{x} \tag{2.10}$$

This transformation decorrelates the components of $\mathbf{y}$, as shown by:

$$\mathbf{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \mathbf{\Lambda} = \operatorname{diag}\left(\lambda_0, \dots \lambda_{n-1}\right), \tag{2.11}$$

which means that the transformed coefficients in $\mathbf{y}$ are uncorrelated, with variances given by the corresponding eigenvalues.

Although the KLT is theoretically optimal, it is computationally expensive due to the need to compute (and store) the eigenvectors of the covariance matrix. This requires solving an eigenvalue problem at design time and performing $O(n^2)$ operations to project the signal onto the eigenvector space during deployment. As a result, the KLT is rarely used directly in practice but serves as a theoretical benchmark for comparing the performance of more computationally efficient transforms, such as the DCT and DWT. These transforms approximate the properties of the KLT while being simpler to compute.

**DCT** The DCT [7] is one of the most widely used transforms, with applications extending beyond image, audio, and video compression [6]. It works by representing a signal as a sum of cosines that oscillate at different frequencies, helping to concentrate the signal energy into a small number of low-frequency coefficients. The 1D DCT of a signal $\mathbf{x}$ is defined as:

$$y_j = \alpha_j \sum_{i=0}^{n-1} x_i \cos\left[\frac{\pi(2i+1)j}{2n}\right], \quad j = 0, 1, \dots, n-1 \tag{2.12}$$

with $\alpha_0 = \frac{1}{\sqrt{n}}$ and $\alpha_j = \frac{2}{\sqrt{n}}$ for $j = 1, \dots, n-1$.

The Inverse DCT (IDCT) allows the reconstruction of the original signal:

$$x_i = \sum_{j=0}^{n-1} \alpha_j y_j \cos\left[\frac{\pi(2i+1)j}{2n}\right], \quad i = 0, 1, \dots, n-1 \tag{2.13}$$

In 2D (e.g., for image compression), the DCT is applied separately to each row and column of an image matrix, resulting in the transformation of the entire image into the frequency domain. The DCT's ability to concentrate the energy of natural signals in a few low-frequency components makes it ideal for compression, as high-frequency components (which often correspond to fine details or noise) can be discarded with minimal perceptual loss. This principle is at the basis of the JPEG image compression algorithm.

**DWT** DWT [120, 10] represents a signal as a composition of wavelet functions that are localized in both the time and frequency domains. Unlike DCT, which uses globally defined cosine functions that span the entire signal, wavelets are compact and can efficiently capture both high- and low-frequency components with localized detail.

The DWT is obtained by recursively decomposing the signal into approximation (low-frequency) and detail (high-frequency) components. Formally, the 1D DWT of a signal $\mathbf{x}$ can be expressed as

$$\mathbf{y} = [\mathbf{y}_{\text{low}}, \mathbf{y}_{\text{high}}] \tag{2.14}$$

where $\mathbf{y}_{\text{low}}$ contains low-frequency components, and $\mathbf{y}_{\text{high}}$ embeds the high-frequency components. This process can be repeated on the low-frequency coefficients to achieve multi-level decomposition, providing a hierarchical representation of the signal structure.

In image compression, such as in JPEG2000, the 2D DWT is applied to images, resulting in a multi-resolution analysis of the image. This hierarchical decomposition captures details at different scales, allowing efficient compression by prioritizing the more significant low-frequency components while selectively neglecting less important high-frequency details.

### 2.2.3 Dimensionality reduction-based compression

Dimensionality reduction, a concept closely related to lossy compression, performs signal approximation by mapping high-dimensional data into a lower-dimensional space, retaining only the most significant information [151]. While dimensionality reduction is not a proper compression technique since the data is not quantized, it still leads to information loss. As such, it can be used as a preprocessing stage in compression pipelines, selecting the most informative components prior to quantization. This process results in a more efficient compression strategy in which the essential components are retained, while the redundant or less informative components are discarded.

In practice, dimensionality reduction and transform coding often complement each other. Many transforms have a property of coefficients *concentration*, meaning that a large part of the signal's energy is contained within just a few transform coefficients. This enables elimination of insignificant coefficients in the transformed domain, leading to a reduced bitstream size and computational complexity at the cost of some additional distortion.

#### Common dimensionality reduction methods

**Principal component analysis** PCA is one of the most widely used methods for dimensionality reduction [68, 24, 100]. It is based on KLT, as this transform is also optimal in the sense that it concentrates the maximum possible energy of a signal $\mathbf{x}$ into the fewest number of coefficients. In fact, the eigenvectors in $\mathbf{U}$, corresponding to the largest eigenvalues, identify the directions in the signal space that capture the most variance, allowing for an effective reduction in the dimensionality of the signal.

Given a target dimensionality $k$ of the subspace, the PCA dimensionality reduction is performed by projecting the signal $\mathbf{x}$ onto the reduced subspace:

$$\mathbf{y} = {\mathbf{U}_k}^\top \mathbf{x} \tag{2.15}$$

where $\mathbf{U}_k = [\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}]$ contains the first $k$ columns of $\mathbf{U}$.

The original data is reconstructed as:

$$\hat{\mathbf{x}} = \mathbf{U}_k \mathbf{y}. \tag{2.16}$$

By selecting the top $k$ eigenvectors (or *principal components*), the reconstruction error measured with $D$ in (2.1) is minimized, ensuring that the most variance is retained in the reduced subspace.

**Autoencoder (AE)** It is a particular neural network that is used for dimensionality reduction [78]. An AE consists of an encoder $(\mathrm{ENC})$ and a decoder $(\mathrm{DEC})$, where the encoder maps the input $\mathbf{x}$ to a *latent* representation $\mathbf{y}$, and the decoder reconstructs the input from $\mathbf{y}$, i.e.,

$$\mathbf{y} = \mathrm{ENC}(\mathbf{x}), \quad \hat{\mathbf{x}} = \mathrm{DEC}(\mathbf{y}). \tag{2.17}$$

The whole network is trained to minimize the reconstruction loss that, up to a normalization constant $n$, is measured with distortion $D$ and estimated with Mean Squared Error (MSE):

$$\mathcal{L}(\mathbf{x}) = \frac{1}{n}\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{nN}\sum_{i=0}^{N-1} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{2.18}$$

where $N$ is the number of training examples of $\mathbf{x}$.

Autoencoders are well-suited for capturing complex, non-linear relationships within data that often reside on lower-dimensional *manifolds* embedded in higher-dimensional spaces. A manifold represents a lower-dimensional structure that signals naturally form within a larger space. Unlike PCA, which assumes linear subspaces, autoencoders can learn non-linear mappings, reducing the dimensionality of signals by identifying and preserving the manifold structure. This is especially useful for minimizing reconstruction error when the signal's underlying geometry is non-linear, as seen in cases such as images or audio.

### 2.2.4 Neural network-based compression

AEs play a key role in many of the so-called end-to-end optimized image compression frameworks. Optimizing the components in the scheme shown in Figure 2.1 is not trivial. Improving one module may not necessarily translate to enhanced overall rate-distortion performance. With the rapid advancement of deep learning, several studies [145, 103, 63, 44] have investigated the potential of neural networks to simultaneously optimize all components of the compression pipeline, allowing for better adaptability among the modules. However, to function effectively as compression tools, AEs must be equipped with two fundamental components: a quantizer and an *entropy model*.

The problem with quantization is its non-differentiability, which complicates training with gradient-based optimization techniques. To address this, several methods have been proposed, such as adding uniform noise during training to approximate the effects of quantization [13] or using the straight-through estimator [145], which treats quantization as the identity function during backpropagation. Some approaches also involve learning quantization centers during training, as shown in [103], where smooth kernels are used to relax the quantization during the backward pass.

In addition to quantization, entropy models play an important role in predicting the probability distribution of quantized latent variables, allowing efficient encoding using entropy coding techniques, such as arithmetic coding. Advanced entropy models, such as those that incorporate *context models* [103], predict the probability of a latent variable based on its context (i.e., previously decoded values), further improving compression efficiency by capturing dependencies in the data.

The interplay between the quantizer and the entropy model is essential for achieving high compression ratios, and both components are typically trained together with the AE in an end-to-end manner, with the objective of optimizing rate-distortion performance.

A comprehensive survey and benchmark of learned image compression techniques can be found in [63], along with a survey on non-linear transform coding in [13].

# Part I

# Performance Assessment of Anomaly Detectors

# Chapter 3

# Synthetic Anomalies for Performance Assessment in Time Series

In sectors such as healthcare, industry, and structural engineering, the use of monitoring systems is a common practice aimed at tracking, controlling, and optimizing physical environments. Such systems, equipped with multiple sensors, generate a large amount of time-series data that are then processed to identify system malfunctions, data chain failures, or unauthorized intrusions. As anticipated in Chapter 1, in this scenario, the essential processing task becomes anomaly detection (AD). The literature offers numerous AD algorithms [29] and the main challenge often lies in which AD algorithm to select for the application at hand.

Selecting the most suitable algorithm for building an anomaly detector requires knowing the most common anomalies within the application and having a statistically significant number of examples of such anomalies to evaluate and compare the algorithms' performance. However, such data are not available in many practical cases, especially during the design phase.

This problem can be categorized under the broader task of *model selection*, which consists of choosing an appropriate statistical model based on available data [40, 51]. For AD on time series, this issue has only recently been systematically addressed [51, 125, 129, 155], and the research suggests that generating synthetic anomalies can provide a promising solution. However, to implement this approach, designers must: *i*) identify common anomalies relevant to the application; *ii*) model and characterize such anomalies; and *iii*) develop a procedure for the generation of synthetic anomalies that resemble them. The literature still lacks a comprehensive framework that includes and automates all these steps.

This approach has previously been explored in the context of image data. In [60], the authors propose a framework to generate common image corruptions at varying intensities to assess the robustness of classifiers. In [125] the authors employ this framework as one of the reference benchmarks in image-based AD.

In the case of time series, most previous studies have focused on particular types of abnormalities [131, 109, 132, 67], with only a few offering systematic procedures for the generation of anomalies [83, 141, 56, 113, 51].

Some works have investigated anomalies affecting a monitoring system itself. For example, [131, 109, 132] focus on common sensor faults such as spikes, stuck-at values, and low battery, categorizing them into three types: *short*, *noise*, and *constant*. Similarly, [67] models a limited number of faults to evaluate the detection capabilities of a One-Class Support Vector Machine (OCSVM) detector [130].

More recent works [83, 56, 141] detach from the source of the anomaly and instead classify anomalies based on their impact on the signal. For example, [83] models normal

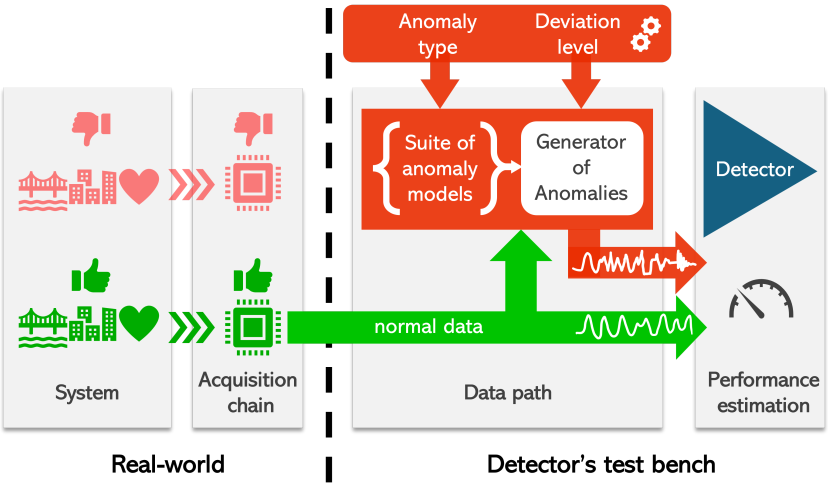


FIGURE 3.1: WOMBATS allows to assess a detector targeting unknown anomalies coming from potentially multiple sources.

signals as a combination of *seasonality* and *trend* and generates anomalies by perturbing these components. The works in [141, 56] present a more general model for normal signals, but focus their analysis only on a few types of anomalies, such as *global*, *local*, and *dependency* anomalies.

Finally, [113, 51] propose a broader set of anomalies, covering several effects that anomalous signals may have on normal data in the real world. However, while these contributions are important, the anomaly sets remain incomplete and there is still a lack of a common parameter to all anomalies to quantify their severity, an essential requirement for comparing detector performance across different types of anomalies.

In this chapter, we present a framework named WOMBATS (Wide Open Model-Based Anomaly Test Suite)[1] and depicted in Figure 3.1 for a systematic evaluation of anomaly detectors based on synthetic anomalies that designers can tailor to any context using historical data and domain-specific information. Importantly, the framework only requires normal data, as various anomalies are generated synthetically by controlling a single parameter. We focus on the unsupervised (anomaly-agnostic) scenario, as supervised detectors require prior knowledge of anomalies, which is not available by assumption[2].

In detail, in Section 3.1 we model both normal and anomalous time series signals. For anomalies, we first perform a taxonomy of the most common anomalies in time series and then define a general model that accounts for multiple effects anomalies have on the normal signal. This model is flexible enough to cover the analyzed and potentially other anomalies.

In Section 3.2 the full framework for anomaly detector selection is described. We start by categorizing anomalies according to their effect on the signal's power — whether it increases, remains invariant, or decreases. The first group includes superimposed disturbances, the second captures deviations that leave the power unchanged but affect information content, and the third comprises nonlinear distortions that reduce power. To ensure consistency among different anomalies, each perturbation is modeled using a parameter that controls the amount of deviation from the normal signal. After detailing a generation procedure for each anomaly, we define a metric to evaluate the detector starting from normal and anomalous examples.

[1] The source code can be found at `https://github.com/SSIGPRO/wombats`

[2] Exploring the use of synthetic anomalies for training and testing supervised detectors is an interesting direction but requires further consideration beyond the scope of this chapter.

Finally, in Section 3.3 we present and analyze the results of the experiments. The framework was applied to two real-world scenarios: human health monitoring (using electrocardiogram data) and structural health monitoring (using acceleration data). In both cases, signals were synthetically corrupted to create datasets for testing and evaluating various detectors. The performance of the detectors was compared on these synthetic anomalies, demonstrating the framework's effectiveness in assessing detector performance across different contexts. Additionally, we proved its ability to predict how detectors would behave when faced with real-world anomalies.

## 3.1 Models of normal and anomalous signals

In this section, we describe the model and assumptions used to represent normal observations and introduce a general model for defining anomalous behaviors.

### 3.1.1 From time series to time instances

Time series data consist of continuously collected and recorded measurements over a given time. Formally, a time series can be defined as a set $S$ of pairs, where each pair contains a vector of $m$ simultaneously monitored variables $\mathbf{s}_i = (s^{(0)}, \ldots, s^{(m-1)})$ observed at a specific timestamp $t_i$: $S = \{(\mathbf{s}_i, t_i) | i \in \mathbb{N}, t_p < t_q \text{ if } p < q\}$. If only a single variable is tracked ($m = 1$), the time series is called *univariate*, while if $m > 1$, $S$ is a *multivariate* time series.

In most cases, the time intervals between successive measurements remain constant, allowing us to omit the explicit reference to time. For simplicity, in this work, we focus on univariate time series, which can be seen as an ordered sequence of measurements $\mathsf{s} = (s_0, s_1, \ldots, s_p, \ldots)$. From now on, we treat $\mathsf{s}$ as a signal sampled at regular intervals, with a sampling period $T$.

In a realistic scenario where the signal $\mathsf{s}$ must be processed in real-time, it is impractical to treat $\mathsf{s}$ as a single entity. Instead, we model $\mathsf{s}$ as a sequence of non-overlapping signal instances, also referred to as windows, each consisting of $n$ consecutive samples: $\mathsf{s} = (\mathbf{x}_0^\top, \mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_p^\top, \ldots)$, where $\mathbf{x}_j = (s_{jn}, \ldots, s_{j(n+1)-1})^\top$.

We assume that the length of each instance $n$ is sufficiently large to ensure that each $\mathbf{x}_j$ captures the main statistical properties of the complete signal $\mathsf{s}$. Furthermore, we assume that the dependency between different segments, $\mathbf{x}_j$ and $\mathbf{x}_p$ (for $j \neq p$), is negligible. This allows us to analyze one instance $\mathbf{x} = (x_0, \ldots, x_{n-1})^\top$ at a time.

### 3.1.2 Model of normal signals

Given these assumptions, the acquired signal samples are represented as a vector $\mathbf{x} \in \mathbb{R}^n$, which can be an example of a source of either normal readings $\mathbf{x}^{\mathrm{ok}} \in \mathbb{R}^n$ or anomalous readings $\mathbf{x}^{\mathrm{ko}} \in \mathbb{R}^n$. The two types of sources, normal and anomalous, are assumed to have different statistical properties, which should help to distinguish between them. Without loss of generality, we assume that the normal readings $\mathbf{x}^{\mathrm{ok}}$ have zero mean $\mathbf{E}\left[\mathbf{x}^{\mathrm{ok}}\right] = 0$ and covariance matrix $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{E}\left[\mathbf{x}^{\mathrm{ok}}\mathbf{x}^{\mathrm{ok}\top}\right]$. We also assume that $\mathbf{x}^{\mathrm{ok}}$ is normalized to have unit power, so that $\frac{1}{n}\mathbf{E}\left[\|\mathbf{x}^{\mathrm{ok}}\|^2\right] = 1$, where $\|\cdot\|$ denotes the $\ell_2$ norm.

In real-world signals, power is typically unevenly distributed in the signal space. This is evident from the spectral decomposition of the covariance matrix $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{U}^{\mathrm{ok}}\mathbf{\Lambda}^{\mathrm{ok}}\mathbf{U}^{\mathrm{ok}\top}$, where $\mathbf{U}^{\mathrm{ok}}$ is an $n \times n$ orthonormal matrix that contains the eigenvectors of $\mathbf{\Sigma}^{\mathrm{ok}}$, and $\mathbf{\Lambda}^{\mathrm{ok}}$ is a diagonal $n \times n$ matrix with the corresponding eigenvalues $\lambda_0^{\mathrm{ok}} \geq \lambda_1^{\mathrm{ok}} \geq \cdots \geq \lambda_{n-1}^{\mathrm{ok}} \geq$

TABLE 3.1: Examples of sensor faults and alterations of the system

**Sensor faults**

| **Type** | **Description** |
|---|---|
| Noise faults | Occur due to hardware failure, environmental conditions, or battery supply issues, resulting in unexpected high variance or noise in sensor values [131]. |
| Short fault/spike | Manifest as an instantaneous increase in the rate of change of sensor values and is caused by hardware or connection failures [67, 161]. |
| Clipping or saturation | Result in a clipping of the extreme values in sensor readings, either due to physical limitations of certain sensor types or calibration issues [37, 64]. |
| Stuck-at faults | Refer to situations when the output shows low variance with samples concentrated around a constant value [118]. |
| Constant | Due to improper calibration or drift of calibration parameters over time, sensor readings may manifest a consistent deviation by a constant value [106]. |
| Narrow-band interference | Considers spurious narrow-band signals contaminating the band of the signal of interest [153]. |
| Dead-zone | Is a non-linearity that manifests when a sensor or actuator fails to provide a non-null output value [86]. |

**System alterations**

| **Type** | **Description** |
|---|---|
| Aging | Occurs due to the natural and unavoidable variation of the physical properties of the system over time [119, 92]. |
| Wearing | Alterations caused by degradation of the system's components due to repeated environmental or mechanical stimuli, e.g., erosion, friction [72]. |
| Abrupt changes | Sudden variations in the system's behavior. Examples of such alterations include damages in a civil structure (tendon/strand breakages [24] and earthquakes [101]); irregular heart rate, irregular rhythm, and ectopic rhythm during heart activity [85]. |

$0$. The eigenvalues are monotonically non-increasing and we can identify an integer $k$ such that the fraction of power contained in the first $k$ components is given by:

$$\gamma = \frac{1}{n}\sum_{j=0}^{k-1} \lambda_j^{\mathrm{ok}} \gg \frac{k}{n} \tag{3.1}$$

This value of $k$ estimates the number of *principal components* of the signal. The components beyond the principal ones are usually dominated by noise and provide minimal information about the features of the normal signal.

### 3.1.3 Model of anomalous signals

Anomalies are often described according to either their cause or their effect on the system within a specific application. Table 3.1 provides examples of sensor failures and anomalies that affect the monitored system in various scenarios. Although understanding domain-specific causes and effects is important, creating a general framework for anomaly detection requires abstracting from these specifics and focusing on the alterations that anomalies introduce to acquired signals.

To achieve this, we propose a dictionary of anomalies based on a common mathematical model specialized to highlight changes in signal features that are typically affected by real-world anomalies. The key idea is to represent anomalies as *variations* of a normal signal in the sense that anomalous waveforms can be derived from normal waveforms as

$$\mathbf{x}^{\text{ko}} = c\left(\mathbf{x}^{\text{ok}}\right) + \mathbf{d} \tag{3.2}$$

where $c : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a potentially nonlinear function that captures how the anomaly alters the normal signal $\mathbf{x}^{\text{ok}}$, with the assumption that $\mathbf{E}\left[c\left(\mathbf{x}^{\text{ok}}\right)\right] = 0$. The vector $\mathbf{d} \in \mathbb{R}^n$ represents an independent disturbance, characterized by a power level $\frac{1}{n}\mathbf{E}\left[\|\mathbf{d}\|^2\right] = a^2$. The independence between $\mathbf{d}$ and $\mathbf{x}^{\text{ok}}$ implies that the power contributions of normal signal and disturbance are separable, i.e., $\mathbf{E}\left[{\mathbf{x}^{\text{ok}}}^\top \mathbf{d}\right] = \mathbf{E}\left[\mathbf{d}^\top c\left(\mathbf{x}^{\text{ok}}\right)\right] = 0$. The function $c$ can represent a range of transformations, including simple scaling, linear mappings (using an $n \times n$ matrix), or more complex nonlinear functions.

To effectively explore the anomaly space, we apply different variations to the normal signal. When $a^2 > 0$, the anomaly is described by a signal scaling. If $a^2 = 0$, the anomaly is obtained by either redistributing signal power through a linear transformation or by introducing a non-linearity to the signal via $c$.

To quantify the difference between normal and anomalous signals, we introduce the concept of *deviation*, defined as:

$$\Delta = \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{x}^{\text{ok}} - \mathbf{x}^{\text{ko}}\right\|^2\right] = 1 + \frac{1}{n}\text{tr}\left[\mathbf{\Sigma}^{\text{ko}}\right] - \frac{2}{n}\mathbf{E}\left[{\mathbf{x}^{\text{ok}}}^\top c\left(\mathbf{x}^{\text{ok}}\right)\right] \tag{3.3}$$

where $\mathbf{\Sigma}^{\text{ko}} = \mathbf{E}\left[\mathbf{x}^{\text{ko}}{\mathbf{x}^{\text{ko}}}^\top\right]$ is the covariance matrix of the anomalous signal. This expression leverages the facts that $\frac{1}{n}\text{tr}\left[\mathbf{\Sigma}^{\text{ok}}\right] = 1$ and that $\mathbf{x}^{\text{ok}}$ and $\mathbf{d}$ are uncorrelated.

Deviation acts as a key parameter for controlling the difficulty of the AD task, allowing detectors to be tested against anomalies of varying levels of challenge.

In cases where $a^2 = 0$ and $c\left(\mathbf{x}^{\text{ok}}\right) = \mathbf{C}\mathbf{x}^{\text{ok}}$, for some matrix $\mathbf{C}$, we may set

$$\mathbf{C} = \mathbf{U}^{\text{ko}}\sqrt{\mathbf{\Lambda}^{\text{ko}}}\sqrt{\mathbf{\Lambda}^{\text{ok}}}^{-1}{\mathbf{U}^{\text{ok}}}^\top \tag{3.4}$$

where $\mathbf{U}^{\text{ko}}$ and $\mathbf{\Lambda}^{\text{ko}}$ represent the eigenvectors and eigenvalues from the spectral decomposition of the anomalous signal's covariance matrix $\mathbf{\Sigma}^{\text{ko}} = \mathbf{U}^{\text{ko}}\mathbf{\Lambda}^{\text{ko}}{\mathbf{U}^{\text{ko}}}^\top$ and where we have used the straightforward definition of the square root of a diagonal matrix. Leveraging the linearity of this mapping, the expectation in (3.3) simplifies to:

$$\mathbf{E}\left[{\mathbf{x}^{\text{ok}}}^\top c\left(\mathbf{x}^{\text{ok}}\right)\right] = \text{tr}\left[\mathbf{C}\mathbf{\Sigma}^{\text{ok}}\right] = \text{tr}\left[\mathbf{U}^{\text{ko}}\sqrt{\mathbf{\Lambda}^{\text{ko}}}\sqrt{\mathbf{\Lambda}^{\text{ok}}}{\mathbf{U}^{\text{ok}}}^\top\right]. \tag{3.5}$$

## 3.2 WOMBATS: a framework for detector selection

In this section, we describe the proposed framework called WOMBATS for detector selection, as illustrated in Figure 3.1. After training the target detectors on normal data, the designer can generate an anomalous version of the test set using a predefined suite of anomalies. The intensity of these anomalies is set by the deviation parameter. The detectors are then evaluated by testing them on both normal and anomalous data, allowing for an assessment of their ability to discriminate between the two.

To elaborate further on the procedure, we: *i*) introduce the anomaly suite, categorized by the effect of the anomalies on the signal's power; *ii*) explain the implementation of these anomalies; *iii*) derive the deviation expressions for each type of anomaly; and *iv*) define a metric to assess the detectors' performance.

### 3.2.1 Anomalies that increase signal power

If we assume $c$ is the identity and $a^2 > 0$, the anomalous signal has power $\frac{1}{n}\mathbf{E}\left[\|\mathbf{x}^{\mathrm{ko}}\|^2\right] = 1 + a^2$. We will consider several possible disturbances.

#### Constant

The disturbance is constant, with $d_j := \pm a$ for $j = 0, \dots, n-1$.

#### Step

We define

$$d_j := \pm a \begin{cases} r\sqrt{n/l} & j = 0, \dots, l-1 \\ (1-r)\sqrt{n/(n-l)} & j = l, \dots, n-1 \end{cases} \tag{3.6}$$

here $r \in \{-1, 1\}$ determines whether the step is rising or falling, and $l \in \{0, \dots, n-1\}$ sets the step's position within the time window.

#### Impulse

$$d_j := \pm a \begin{cases} \sqrt{n} & \text{if } j = l \\ 0 & \text{otherwise} \end{cases}, \quad 0 \le j < n \tag{3.7}$$

where $l \in \{0, \dots, n-1\}$ sets the impulse's position.

#### Gaussian white noise (GWN)

We define $\mathbf{d}$ as a simple White Gaussian noise, i.e., $d_j \sim \mathcal{N}(0, 1)$.

#### Gaussian narrowband noise (GNN)

The disturbance is a Gaussian signal with a frequency band of $[f_0 - B/2, f_0 + B/2]$ where $0 \le f_0 \le 1/2$ is the center frequency and $0 \le B \le \min\{f_0, 1/2 - f_0\}$ is the bandwidth.

### 3.2.2 Anomalies that do not change signal power

In this case, anomalies modify the normal signal by redistributing its power within the signal space. This redistribution can either occur between the normal signal and an independent disturbance or within the components of the normal signal itself.

#### Mixing with a disturbance

When a disturbance with power $a^2 \le 1$ is present, we model the mixing by defining $c\left(\mathbf{x}^{\mathrm{ok}}\right) = \sqrt{1-a^2}\mathbf{x}^{\mathrm{ok}}$. In this case, we focus on two types of disturbances: constant and GWN.

### Intra-signal mixing

When power redistribution occurs solely among the components of the normal signal, we assume this happens via a linear transformation, represented by a matrix $\mathbf{C}$, which preserves power, ensuring that $\frac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}^{\mathrm{ko}}\right] = 1$.

**Time warping** Time warping [20] produces local decelerations in the normal signal. This can be modeled by assuming that the entries in $\mathbf{x}^{\mathrm{ok}}$ are discrete-time samples at rate $f_s$, taken from a continuous-time waveform $\mathbf{x}^{\mathrm{ok}}(t)$. The time-warping transformation $w(t)$ creates an anomalous waveform $\mathbf{x}^{\mathrm{ko}}(t) = \mathbf{x}^{\mathrm{ok}}(w(t))$, which is then sampled into the vector $\mathbf{x}^{\mathrm{ko}}$. Considering the sampling times $t_j = \frac{j}{f_s}$ with $j = 0, \ldots, n-1$, we have:

$$x_j^{\mathrm{ko}} = \mathbf{x}^{\mathrm{ko}}(t_j) = \mathbf{x}^{\mathrm{ok}}(w(t_j)) = \sum_{q=0}^{n-1} x_q^{\mathrm{ok}} \eta(w(t_j)f_s - q)\,\mathrm{d}w_q \tag{3.8}$$

where $\eta(\cdot)$ is the interpolation function allowing the reconstruction of the continuous waveform from its samples, and $\mathrm{d}w$ accounts for local time warping, guaranteeing power preservation. In this case, the linear transformation matrix is $C_{j,q} = \eta(w(t_j)f_s - q)\,\mathrm{d}w_q$.

Further to redistribution along time, we also take into account power redistribution in the spectrum. To this end, based on the spectral decomposition of the covariance matrix $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{U}^{\mathrm{ok}}\mathbf{\Lambda}^{\mathrm{ok}}\mathbf{U}^{\mathrm{ok}\top}$, where the first $k$ components represent the principal ones and characterize normality, we consider two types of spectral alterations.

**Spectral alteration** This anomaly modifies how power is distributed across the principal components by altering the eigenvalues in $\mathbf{\Lambda}^{\mathrm{ok}}$, resulting in a new diagonal matrix $\mathbf{\Lambda}^{\mathrm{ko}}$. The transformation matrix $\mathbf{C}$ can be derived using the expression in (3.4).

**Principal subspace alteration** This anomaly alters the principal components themselves by modifying the columns of $\mathbf{U}^{\mathrm{ok}}$, resulting in a new orthonormal matrix $\mathbf{U}^{\mathrm{ko}}$. The transformation matrix $\mathbf{C}$ is again obtained relying on (3.4).

## 3.2.3 Anomalies that decrease signal power

### Saturation

Saturation is modeled by clipping the largest samples in a window to a maximum value $x_{\mathrm{SAT}}$ as in

$$x_j^{\mathrm{ko}} = \begin{cases} x_{\mathrm{SAT}} \ \text{sign}\left(x_j^{\mathrm{ok}}\right) & \text{if } \left|x_j^{\mathrm{ok}}\right| > x_{\mathrm{SAT}} \\ x_j^{\mathrm{ok}} & \text{otherwise} \end{cases} \tag{3.9}$$

This ensures that any sample that exceeds the threshold is set to $x_{\mathrm{SAT}}$, preserving its original sign.

### Dead-zone

In contrast to saturation, the dead-zone anomaly sets to $0$ the samples smaller than $x_{\mathrm{DZ}}$ as in

$$x_j^{\mathrm{ko}} = \begin{cases} 0 & \text{if } \left|x_j^{\mathrm{ok}}\right| < x_{\mathrm{DZ}} \\ x_j^{\mathrm{ok}} & \text{otherwise} \end{cases} \tag{3.10}$$

### 3.2.4 Anomaly implementation

Considering the taxonomy of anomalies outlined above, we define how the random components of the mathematical models translate into actual instances:

- The symbol $\pm$ is randomly assigned to either a $+$ or $-$ with equal probability of $\frac{1}{2}$.
- The rising parameter $r$ is uniformly selected from $\{0, 1\}$, while the position $l$ is set to the midpoint of the window $\lceil n/2 \rceil$.
- When $\mathbf{d}$ is a Gaussian noise within the frequency band $[f_0 - \frac{B}{2}, f_0 + \frac{B}{2}]$, it is drawn as a Gaussian random vector with zero mean and covariance $\boldsymbol{\Sigma}$, defined as [133]:

$$\Sigma_{j,q} = a^2 \cos\left(2\pi(j-q)f_0\right) \operatorname{sinc}\left((j-q)B\right) \tag{3.11}$$

  for $j, q = 0, \dots, n-1$. If white noise is required, the covariance is simply $\boldsymbol{\Sigma} = a^2 \mathbf{I}_n$, with $\mathbf{I}_n$ the identity matrix.

- The interpolation function $\eta$ is implemented using $3$-rd order cardinal splines described in [150].
- The warping function is modeled as $w(t) = (1-\alpha)\,t$, with $\alpha \in [0, 1]$ controlling the degree of local time deceleration.
- Spectral alterations are applied only to the principal components, while the noise subspace remains untouched. Specifically, the eigenvalues $\lambda_j^{\mathrm{ko}}$ differ from $\lambda_j^{\mathrm{ok}}$ for $j = 0, \dots, k-1$, but remain equal for $j = k, \dots, n-1$. The number of principal components $k$ is determined using the criterion proposed in [47].

  The first $k$ eigenvalues of $\boldsymbol{\Lambda}^{\mathrm{ok}}$ are altered by defining vectors

  $\ell^{\mathrm{ko}} = \frac{1}{\sqrt{n\gamma}} \left(\sqrt{\lambda_0^{\mathrm{ko}}}, \dots, \sqrt{\lambda_{k-1}^{\mathrm{ko}}}\right)^\top$ and $\ell^{\mathrm{ok}} = \frac{1}{\sqrt{n\gamma}} \left(\sqrt{\lambda_0^{\mathrm{ok}}}, \dots, \sqrt{\lambda_{k-1}^{\mathrm{ok}}}\right)^\top$.

  Since power invariance requires $\left\|\ell^{\mathrm{ok}}\right\| = \left\|\ell^{\mathrm{ko}}\right\| = 1$, we set $\ell^{\mathrm{ko}} = \mathbf{R}_{k,\theta}\ell^{\mathrm{ok}}$, where $\mathbf{R}_{k,\theta}$ is a random $k \times k$ rotation matrix parameterized by an angle $\theta$. The modified eigenvalues are then recovered by squaring and scaling the entries of $\ell^{\mathrm{ko}}$.

- For principal subspace alterations, the new matrix $\mathbf{U}^{\mathrm{ko}}$ is generated by applying a random $n \times n$ rotation matrix $\mathbf{R}_{n,\theta}$ to the columns of $\mathbf{U}^{\mathrm{ok}}$.
- Random rotation matrices are constructed assuming that both $n$ and $k$ are even. The rotation matrix $\mathbf{R}_\theta$ is given by:

$$\mathbf{R}_\theta = \mathbf{Q} \begin{pmatrix} \mathbf{r}_\theta & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{r}_\theta \end{pmatrix} \mathbf{Q}^\top, \quad \mathbf{r}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{3.12}$$

  where $\mathbf{r}_\theta$ represents a 2D rotation matrix, and $\mathbf{Q}$ is derived by orthonormalizing a sample of the Ginibre ensemble [49].

- In saturation and dead-zone anomalies, the thresholds are dynamically adapted in each window to match the desired deviation $\Delta$ specified by equations (3.18) and (3.19).
- Multiple anomalies can be combined to simulate complex real-world anomalies involving superimposed effects. Further details can be found in the Appendix A.

### 3.2.5 Deviations

Table 3.2 reports the relationship between the parameters used in the definition of each type of anomaly and the resulting deviation $\Delta$. These formulas enable the application of alterations that impose a uniform predefined level of challenge on the detector under test.

TABLE 3.2: Expressions of deviation $\Delta$ for each type of anomaly

| **Anomalies** | $\Delta$ | |
|---|---|---|
| Increasing | $a^2$ | (3.13) |
| Mixing | $2\left[1-\sqrt{1-a^2}\right]$ | (3.14) |
| Time warping | $2\left[1-\frac{1}{n}\mathrm{tr}\left[\mathbf{C}\boldsymbol{\Sigma}^{\mathrm{ok}}\right]\right]$ | (3.15) |
| Spectral alt. | $2\gamma\left[1-\cos(\theta)\right]$ | (3.16) |
| Principal subspace alt. | $2\left[1-\cos(\theta)\right]$ | (3.17) |
| Saturation | $\sum_{\lvert x_j^{\mathrm{ok}}\rvert > x_{\mathrm{SAT}}} \left[x_j^{\mathrm{ok}} - x_{\mathrm{SAT}}\text{sign}\left(x_j^{\mathrm{ok}}\right)\right]^2$ | (3.18) |
| Dead-zone | $\sum_{\lvert x_j^{\mathrm{ok}}\rvert \leq x_{\mathrm{DZ}}} \left(x_j^{\mathrm{ok}}\right)^2$ | (3.19) |

In (3.15), the deviation is governed by the parameter $\alpha$ inherent to $\mathbf{C}$, while in (3.16), the deviation arises from the alterations applied by $\mathbf{R}_{k,\theta}$. Since the principal subspace alterations induced by $\mathbf{R}_{n,\theta}$ are equivalent to rotating each $\mathbf{x}^{\mathrm{ok}}$ into $\mathbf{x}^{\mathrm{ko}}$ by the same angle, equation (3.17) is derived using equations (3.4) and (3.16). Finally, equations (3.18) and (3.19) are used to determine the threshold values $x_{\mathrm{SAT}}$ and $x_{\mathrm{DZ}}$, ensuring that the average deviation across the dataset corresponds to the specified $\Delta$.

### 3.2.6 Metrics

As anticipated in Chapter 1, when a detector processes an input $\mathbf{x}$, whether normal $(\mathrm{ok})$ or anomalous $(\mathrm{ko})$, its output can be interpreted as a function $z(\mathbf{x})$ that assigns a score to the instance. Once normal and anomalous scores have been generated, it is possible to use the Area Under the Curve $(\mathrm{AUC})$ of the Receiver Operating Characteristic $(\mathrm{ROC})$ [45] to measure the overall performance of the detector. We propose to adopt a variation of $\mathrm{AUC}$, reflecting the *probability of correct detection*:

$$P_D = \begin{cases} \mathrm{AUC} & \text{if } \mathrm{AUC} \geq 0.5 \\ 1-\mathrm{AUC} & \text{if } \mathrm{AUC} < 0.5. \end{cases} \tag{3.15}$$

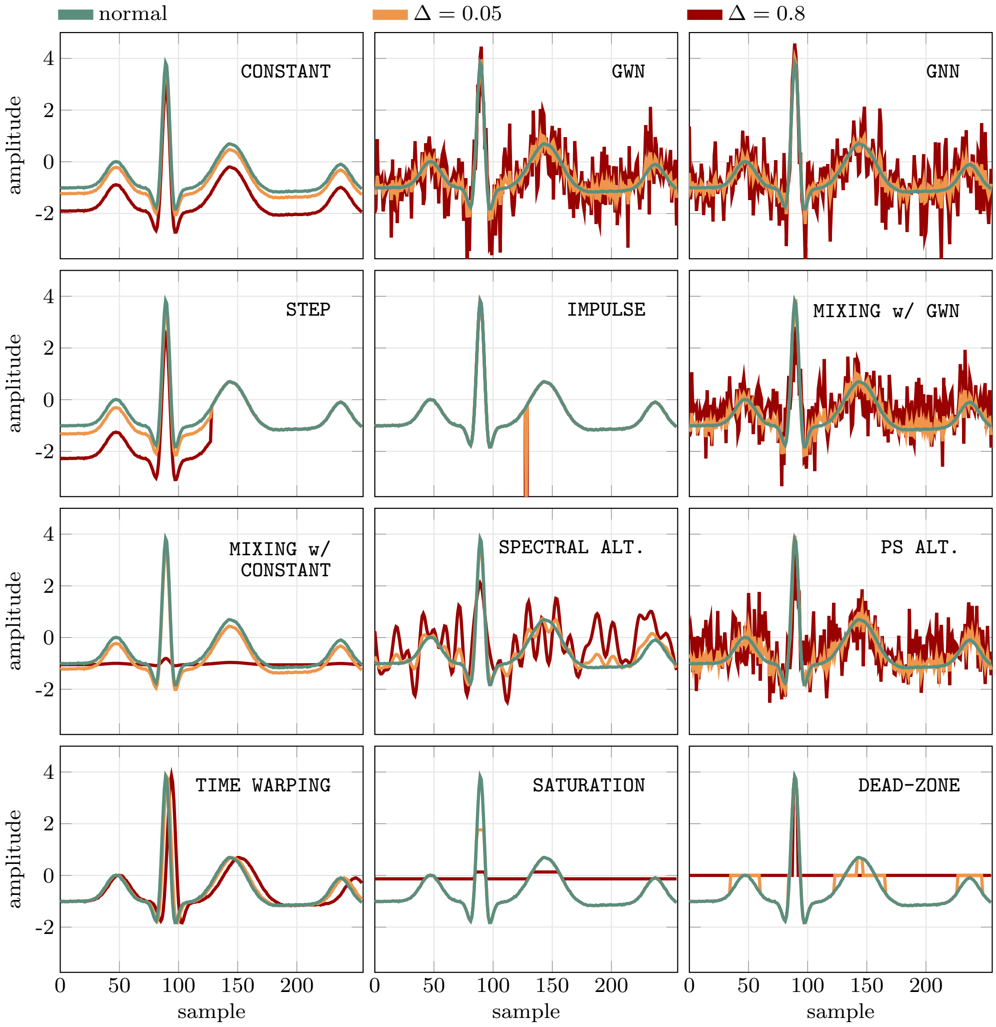


FIGURE 3.2: Examples of ECG anomalies for two different values of deviation $\Delta = \{0.05, 0.8\}$.

The definition of $P_D$ accounts for cases where a detector consistently assigns higher scores to normal instances, meaning $\text{AUC} < 0.5$. In such cases, the detector can still be useful if its output is interpreted in a reversed manner.

## 3.3 Numerical examples

To prove the effectiveness of the proposed anomaly models, we perform a numerical evaluation of detector performance using two distinct datasets: Electrocardiogram (ECG) signals from a health monitoring application and accelerometer (ACC) waveforms from a structural health monitoring system.

### 3.3.1 Numerical setup

#### Detectors

We validate our approach by deploying a range of detectors, each utilizing different techniques that can be found in the literature [29]. Specifically, we focus on the detectors

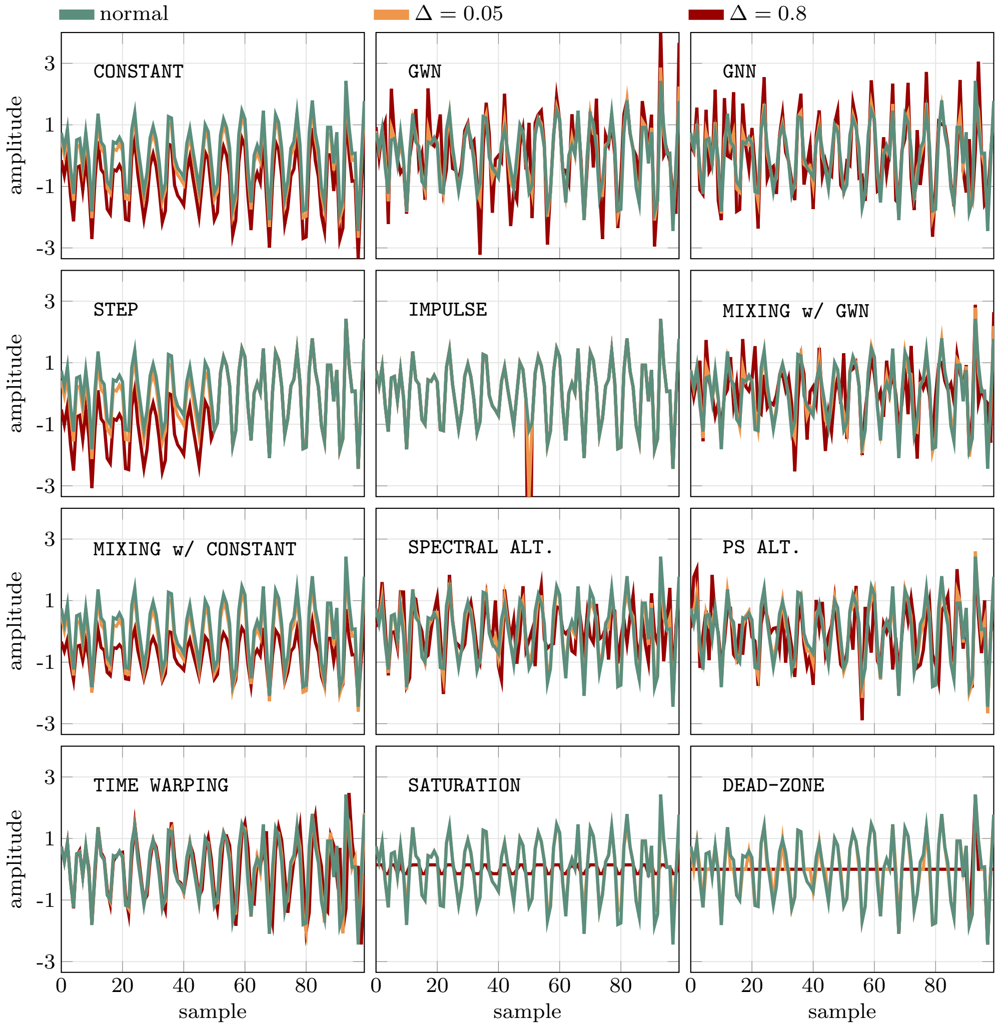


FIGURE 3.3: Examples of ACC anomalies for two different values of deviation $\Delta = \{0.05, 0.8\}$.

TABLE 3.3: Significant signal features and the corresponding score $z(\mathbf{x})$

| **Feature** | $z(\mathbf{x})$ | **Description** |
|---|---|---|
| pk-pk | $\max \mathbf{x} - \min \mathbf{x}$ | Peak-to-peak value |
| energy | $\|\mathbf{x}\|^2$ | Energy of the vector $\mathbf{x}$ |
| TV | $\sum_{i=0}^{n-1} \|x_{i+1} - x_i\|$ | Total variation [124] |
| ZC | $\frac{1}{2}\sum_{i=0}^{n-2} [1 - \mathrm{sign}(x_i x_{i+1})]$ | Number of zero-crossings |

TABLE 3.4: Performance of different detectors (columns) in terms of $P_D$, working on ECG anomalies (rows) with a fixed deviation $\Delta = 0.05$

| | | PCA-BASED | | | | GD-BASED | | | ML-BASED | | | FEATURE-BASED | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ANOMALY** | SPE$_{52}$ | SPE$_{59}$ | T$^2_{52}$ | T$^2_{59}$ | AR$_4$ | AR$_{16}$ | MD | OC$_{\text{poly, 0.01}}$ | LOF$_5$ | IF$_{250}$ | energy | TV | ZC | pk-pk |
| INCREASING | GWN | 1.00 | 1.00 | 0.55 | 0.63 | 1.00 | 1.00 | 1.00 | 0.50 | 0.98 | 0.59 | 0.56 | 1.00 | 1.00 | 0.70 |
| | IMPULSE | 1.00 | 1.00 | 0.54 | 0.62 | 1.00 | 1.00 | 1.00 | 0.50 | 0.98 | 0.54 | 0.56 | 0.76 | 0.66 | 0.76 |
| | STEP | 0.54 | 0.60 | 0.65 | 0.63 | 0.96 | 0.99 | 0.99 | 0.50 | 0.89 | 0.58 | 0.56 | 0.51 | 0.53 | 0.59 |
| | CONSTANT | 0.50 | 0.50 | 0.52 | 0.51 | 0.50 | 0.51 | 0.51 | 0.50 | 0.63 | 0.56 | 0.55 | 0.50 | 0.54 | 0.50 |
| | GNN | 0.93 | 0.92 | 0.55 | 0.59 | 0.93 | 0.93 | 0.93 | 0.50 | 0.97 | 0.59 | 0.56 | 0.95 | 0.94 | 0.67 |
| INVARIANT | MIXING w/ GWN | 1.00 | 1.00 | 0.50 | 0.59 | 1.00 | 1.00 | 1.00 | 0.59 | 0.98 | 0.55 | 0.50 | 1.00 | 1.00 | 0.62 |
| | MIXING w/ CONSTANT | 0.52 | 0.51 | 0.53 | 0.53 | 0.56 | 0.56 | 0.57 | 0.59 | 0.62 | 0.52 | 0.50 | 0.54 | 0.54 | 0.60 |
| | SPECTRAL ALT. | 0.50 | 0.50 | 0.62 | 0.60 | 0.59 | 0.55 | 0.55 | 0.59 | 0.97 | 0.54 | 0.50 | 0.73 | 0.80 | 0.51 |
| | PRINCIPAL SUBSPACE ALT. | 1.00 | 1.00 | 0.50 | 0.58 | 1.00 | 1.00 | 1.00 | 0.59 | 0.98 | 0.55 | 0.50 | 1.00 | 1.00 | 0.61 |
| | TIME WARPING | 0.52 | 0.52 | 0.51 | 0.51 | 0.70 | 0.73 | 0.72 | 0.50 | 0.52 | 0.50 | 0.50 | 0.51 | 0.51 | 0.50 |
| DEC. | SATURATION | 0.86 | 0.99 | 0.79 | 0.81 | 1.00 | 1.00 | 1.00 | 0.83 | 0.81 | 0.61 | 0.70 | 0.69 | 0.50 | 0.97 |
| | DEAD-ZONE | 0.92 | 0.98 | 0.50 | 0.55 | 1.00 | 1.00 | 1.00 | 0.73 | 0.93 | 0.53 | 0.56 | 0.50 | 0.86 | 0.50 |

described in Chapter 1. However, in our analysis, we exclude Deep Learning-based methods, as the focus is to validate the effectiveness of the evaluation framework rather than testing individual detectors.

**Feature-based** Along with the most common detectors, we also track some of the signal's representative features that can be efficiently computed. Each of the features listed in Table 3.3 corresponds to a different score and its detection capabilities can highlight the nature of the modeled anomalies.

### Datasets

**ECG signals** The reference ECG signals were generated using a realistic ECG signal generator as described in [102], following the setup adopted in [96]. Specifically, the sampling rate was set at $256\,\mathrm{sps}$, and heart rates were drawn uniformly in the range $60$-$100\,\mathrm{bpm}$. In total $7.7 \times 10^4$ segments of $2\,\mathrm{s}$ each were generated and randomly divided into non-overlapping windows, each of length $n = 256$, corresponding to vectors $\mathbf{x} \in \mathbb{R}^n$. To emulate realistic conditions, Gaussian white noise was added, guaranteeing a signal-to-noise ratio (SNR) of $40\,\mathrm{dB}$. The number of principal components resulted in $k = 52$. Finally, the dataset was split, with the $10^4$ vectors reserved for testing performance and the remaining vectors used for training.

TABLE 3.5: Performance of different detectors (columns) in terms of $P_D$, working on ECG anomalies (rows) with a fixed deviation $\Delta = 0.8$

| | | PCA-BASED | | | | GD-BASED | | | ML-BASED | | | FEATURE-BASED | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ANOMALY | SPE$_{52}$ | SPE$_{59}$ | $T^2_{52}$ | $T^2_{59}$ | AR$_4$ | AR$_{16}$ | MD | OC$_{\text{poly, 0.01}}$ | LOF$_5$ | IF$_{250}$ | energy | TV | ZC | pk-pk |
| INCREASING | GWN | 1.00 | 1.00 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 | 0.51 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 0.99 |
| | IMPULSE | 1.00 | 1.00 | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 | 0.53 | 1.00 | 0.55 | 0.99 | 1.00 | 0.66 | 1.00 |
| | STEP | 0.90 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.52 | 1.00 | 0.98 | 0.96 | 0.56 | 0.82 | 0.81 |
| | CONSTANT | 0.50 | 0.50 | 0.75 | 0.71 | 0.55 | 0.66 | 0.59 | 0.51 | 0.99 | 0.96 | 0.94 | 0.50 | 0.89 | 0.50 |
| | GNN | 0.94 | 0.93 | 0.70 | 0.74 | 0.96 | 0.96 | 0.97 | 0.50 | 1.00 | 0.99 | 0.97 | 0.99 | 0.97 | 0.96 |
| INVARIANT | MIXING w/ GWN | 1.00 | 1.00 | 0.52 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.85 | 0.51 | 1.00 | 1.00 | 0.52 |
| | MIXING w/ CONSTANT | 0.79 | 0.77 | 0.91 | 0.91 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.77 | 0.50 | 0.94 | 0.97 | 1.00 |
| | SPECTRAL ALT. | 0.52 | 0.51 | 0.96 | 0.94 | 0.91 | 0.81 | 0.85 | 0.99 | 1.00 | 0.73 | 0.53 | 1.00 | 1.00 | 0.72 |
| | PRINCIPAL SUBSPACE ALT. | 1.00 | 1.00 | 0.50 | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.80 | 0.50 | 1.00 | 1.00 | 0.50 |
| | TIME WARPING | 0.59 | 0.58 | 0.54 | 0.55 | 0.77 | 0.79 | 0.78 | 0.50 | 0.57 | 0.50 | 0.50 | 0.54 | 0.56 | 0.50 |
| DEC. | SATURATION | 0.86 | 0.58 | 1.00 | 1.00 | 0.72 | 0.87 | 0.87 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.50 | 1.00 |
| | DEAD-ZONE | 1.00 | 1.00 | 0.99 | 0.61 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.99 |

TABLE 3.6: Performance of different detectors (columns) in terms of $P_D$, working on ACC anomalies (rows) with a fixed deviation $\Delta = 0.05$

| | | PCA-BASED | | | | GD-BASED | | | ML-BASED | | | FEATURE-BASED | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ANOMALY | SPE$_{16}$ | SPE$_{85}$ | $T^2_{16}$ | $T^2_{85}$ | AR$_4$ | AR$_{16}$ | MD | OC$_{\text{RBF, 0.01}}$ | LOF$_5$ | IF$_{500}$ | energy | TV | ZC | pk-pk |
| INCREASING | GWN | 0.80 | 0.98 | 0.55 | 0.87 | 0.76 | 0.93 | 1.00 | 0.64 | 0.85 | 0.59 | 0.58 | 0.59 | 0.57 | 0.60 |
| | IMPULSE | 0.80 | 0.99 | 0.55 | 0.88 | 0.76 | 0.94 | 1.00 | 0.64 | 0.86 | 0.55 | 0.58 | 0.54 | 0.51 | 0.73 |
| | STEP | 0.82 | 0.97 | 0.50 | 0.91 | 0.85 | 0.90 | 0.93 | 0.64 | 0.86 | 0.59 | 0.58 | 0.50 | 0.70 | 0.54 |
| | CONSTANT | 0.81 | 0.51 | 0.50 | 0.84 | 0.85 | 0.90 | 0.83 | 0.64 | 0.84 | 0.59 | 0.58 | 0.50 | 0.72 | 0.50 |
| | GNN | 0.78 | 0.80 | 0.54 | 0.83 | 0.71 | 0.85 | 0.88 | 0.63 | 0.84 | 0.58 | 0.58 | 0.58 | 0.55 | 0.59 |
| INVARIANT | MIXING w/ GWN | 0.79 | 0.98 | 0.54 | 0.87 | 0.75 | 0.93 | 1.00 | 0.63 | 0.85 | 0.58 | 0.58 | 0.58 | 0.57 | 0.60 |
| | MIXING w/ CONSTANT | 0.81 | 0.51 | 0.51 | 0.84 | 0.85 | 0.90 | 0.83 | 0.64 | 0.84 | 0.58 | 0.58 | 0.51 | 0.72 | 0.51 |
| | SPECTRAL ALT. | 0.50 | 0.50 | 0.51 | 0.51 | 0.50 | 0.51 | 0.51 | 0.50 | 0.53 | 0.50 | 0.50 | 0.50 | 0.52 | 0.50 |
| | PRINCIPAL SUBSPACE ALT. | 0.58 | 0.83 | 0.50 | 0.63 | 0.55 | 0.72 | 0.94 | 0.51 | 0.60 | 0.50 | 0.50 | 0.50 | 0.51 | 0.51 |
| | TIME WARPING | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.51 | 0.51 | 0.50 | 0.51 | 0.50 | 0.50 | 0.50 | 0.51 | 0.50 |
| DEC. | SATURATION | 0.54 | 0.69 | 0.56 | 0.53 | 0.55 | 0.59 | 0.90 | 0.56 | 0.54 | 0.55 | 0.55 | 0.55 | 0.50 | 0.68 |
| | DEAD-ZONE | 0.58 | 0.81 | 0.50 | 0.62 | 0.55 | 0.72 | 0.93 | 0.50 | 0.61 | 0.51 | 0.51 | 0.51 | 0.82 | 0.50 |

**ACC signals** The acceleration signals used in this example were collected during the structural health monitoring of a viaduct located in Italy [101, 24]. The dataset consists of measurements from 90 three-axis accelerometers, acquired at a sampling rate of $100\,\mathrm{sample/s}$. These signals capture the elastic response of the structure to vehicle transit and environmental stimuli. For this example, we considered only one axis of a selected sensor and divided the data into $3.77 \times 10^5$ windows, each of length $n = 100$, resulting in vectors $\mathbf{x} \in \mathbb{R}^n$. Due to the inherent noise in the dataset, the number of principal components was determined to be $k = 16$. Similarly to the ECG data, the $10^4$ vectors were allocated for performance assessment, while the remaining vectors were used for training.

### 3.3.2 Results

For each signal class and anomaly category, anomalous instances are obtained by modifying the original input vectors as illustrated in Section 3.2. Figures 3.2 and 3.3 show how different anomalies affect normal ECG and ACC signals, respectively. In these figures, each subfigure represents a different anomaly type, showing both the original and the altered signals for two levels of deviation $\Delta$. As expected, higher levels of deviation lead to more noticeable anomalies.

TABLE 3.7: Performance of different detectors (columns) in terms of $P_D$, working on ACC anomalies (rows) with a fixed deviation $\Delta = 0.8$

| | | PCA-BASED | | | | GD-BASED | | | ML-BASED | | | FEATURE-BASED | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ANOMALY** | $\text{SPE}_{16}$ | $\text{SPE}_{85}$ | $\text{T}^2_{16}$ | $\text{T}^2_{85}$ | $\text{AR}_4$ | $\text{AR}_{16}$ | MD | $\text{OC}_{\text{RBF, 0.01}}$ | $\text{LOF}_5$ | $\text{IF}_{500}$ | energy | TV | ZC | pk-pk |
| **INCREASING** | GWN | 0.96 | 1.00 | 0.76 | 0.99 | 0.97 | 1.00 | 1.00 | 0.89 | 0.99 | 0.81 | 0.80 | 0.82 | 0.67 | 0.83 |
| | IMPULSE | 0.96 | 1.00 | 0.77 | 1.00 | 0.98 | 1.00 | 1.00 | 0.89 | 0.99 | 0.58 | 0.80 | 0.64 | 0.51 | 0.95 |
| | STEP | 0.97 | 1.00 | 0.52 | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 0.99 | 0.82 | 0.80 | 0.51 | 0.92 | 0.69 |
| | CONSTANT | 0.97 | 0.69 | 0.51 | 0.98 | 1.00 | 1.00 | 0.98 | 0.90 | 0.99 | 0.83 | 0.80 | 0.50 | 0.93 | 0.50 |
| | GNN | 0.96 | 0.89 | 0.67 | 0.97 | 0.93 | 0.97 | 0.98 | 0.88 | 0.98 | 0.80 | 0.79 | 0.77 | 0.57 | 0.81 |
| **INVARIANT** | MIXING w/ GWN | 0.96 | 1.00 | 0.70 | 0.99 | 0.96 | 1.00 | 1.00 | 0.86 | 0.99 | 0.77 | 0.76 | 0.77 | 0.69 | 0.79 |
| | MIXING w/ CONSTANT | 0.96 | 0.58 | 0.63 | 0.97 | 1.00 | 0.99 | 0.97 | 0.87 | 1.00 | 0.79 | 0.76 | 0.64 | 0.96 | 0.65 |
| | SPECTRAL ALT. | 0.56 | 0.51 | 0.63 | 0.57 | 0.55 | 0.58 | 0.57 | 0.56 | 0.74 | 0.52 | 0.51 | 0.53 | 0.66 | 0.53 |
| | PRINCIPAL SUBSPACE ALT. | 0.78 | 0.98 | 0.54 | 0.87 | 0.73 | 0.93 | 0.99 | 0.59 | 0.93 | 0.51 | 0.50 | 0.52 | 0.65 | 0.53 |
| | TIME WARPING | 0.61 | 0.50 | 0.52 | 0.59 | 0.51 | 0.50 | 0.57 | 0.52 | 0.67 | 0.50 | 0.50 | 0.51 | 0.54 | 0.50 |
| **DEC.** | SATURATION | 0.89 | 0.69 | 0.94 | 0.88 | 0.90 | 0.74 | 0.65 | 0.93 | 0.83 | 0.92 | 0.92 | 0.94 | 0.50 | 0.98 |
| | DEAD-ZONE | 0.57 | 0.91 | 0.72 | 0.66 | 0.50 | 0.81 | 0.98 | 0.63 | 0.77 | 0.78 | 0.71 | 0.87 | 1.00 | 0.51 |

Both original and corrupted with two levels of deviation vectors are used to test multiple detectors. The detection results, measured by $P_D$, are presented in Tables 3.4, 3.5, 3.6, and 3.7, where each column characterizes a different detector. These results highlight which detector is most effective for each anomaly type, providing valuable feedback on the detector's performance during the design phase.

Tables 3.4 and 3.5 provide insight into the detectability of ECG signals for two deviation levels: $\Delta = 0.05$ and $\Delta = 0.8$, respectively. The results show how each anomaly affects the normal signal differently, leading to different detection performances. In particular, certain anomalies, such as constant or time-warping, are challenging for any detection method.

When comparing the results between Table 3.4 and 3.5, it becomes clear that increasing $\Delta$ improves detection performance, proving that $\Delta$ effectively controls the intensity of the anomalies[3].

For ACC signals, the detection results are summarized in Tables 3.6 and 3.7 for $\Delta = 0.05$ and $\Delta = 0.8$, respectively. As expected, the detection performance for ACC signals differs significantly from that obtained with ECG signals. For the same level of deviation, the values of $P_D$ are generally lower in the ACC case. Moreover, spectral alterations, time warping, and dead-zone anomalies were the most difficult to detect. Additionally, feature-based detectors tend to underperform compared to more sophisticated techniques.

We want to point out that, to understand how the framework assesses a detector's performance for a specific class of signals, the results in Tables 3.4, 3.5, 3.6, and 3.7 should be interpreted column-wise. Each column provides a summary of the expected detection performance of the respective detector.

Figure 3.4 further illustrates the tool's capability to predict detector performance with real-world anomalies. On the left-hand side, we present the scores of different detectors, which were trained on a reference synthetic ECG signal. These scores change based on the variations of a real-world ECG signal with anomalies [50] plotted at the top. While all detectors can identify the anomaly between $16$ and $22\,\text{s}$, some, e.g., $\text{T}^2_{52}$ and the energy-based detector, fail in detecting a noise increase between 23-26$\,\text{s}$ that does not affect the signal's power. This result, even in testing on a real ECG signal, aligns with the insights provided in Table 3.4 as it is able to anticipate the performance of $12$ out of $14$ considered detectors. In fact, for the exception of $\text{LOF}_5$ and ZC, the detectors failing with this real

[3] This is further demonstrated in the Appendix A, where $P_D$ is plotted against $\Delta$ for each detector and anomaly.

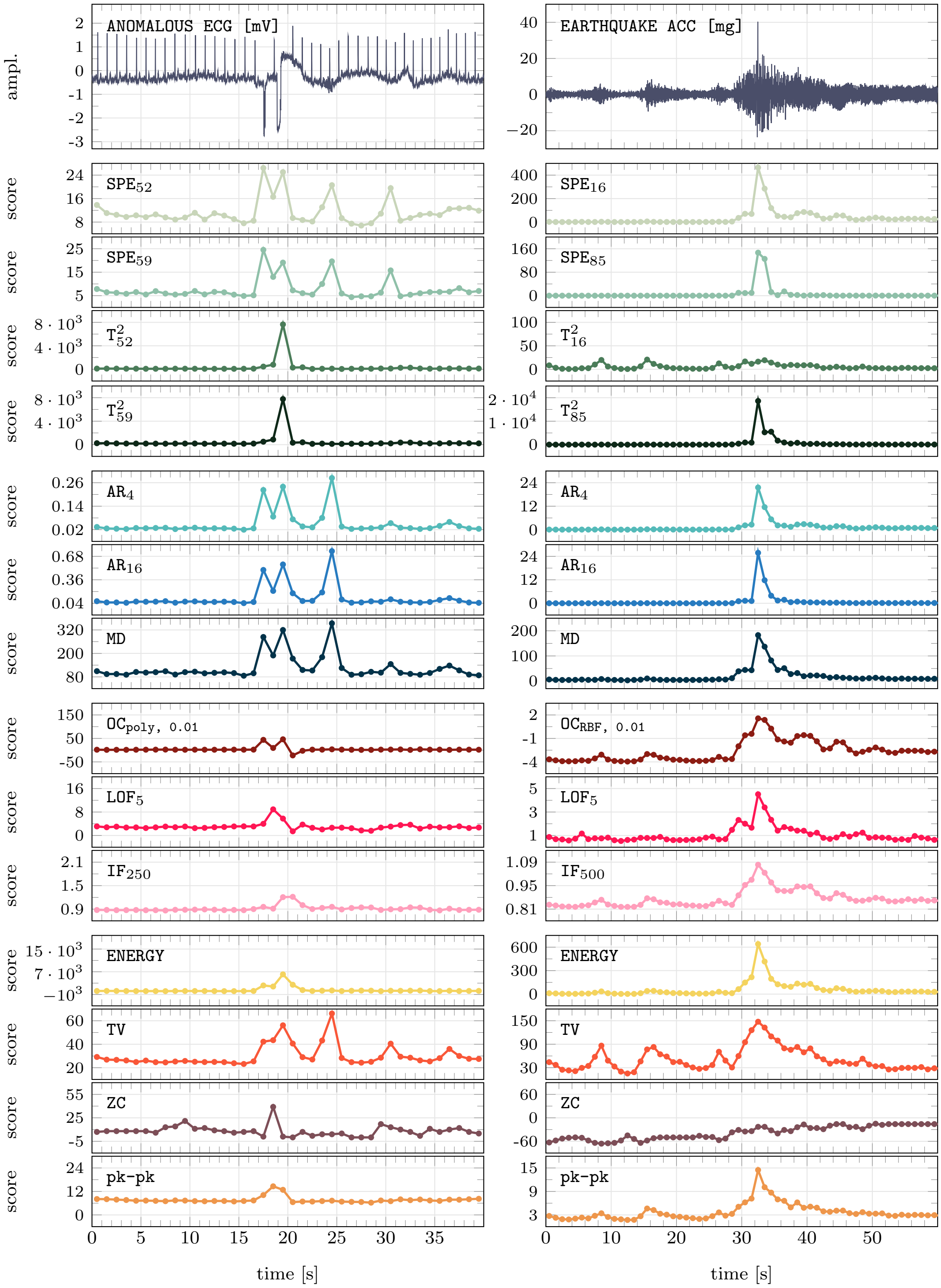


Figure 3.4: Scores trends (bottom) for four detectors working on anomalous real ECG time-series (right) and for four detectors applied on an ACC time series (left) containing an earthquake event.

anomaly, are the ones marked with a red cell in correspondence of the row MIXING w/ GWN in Table 3.4.

In addition, on the right side of Figure 3.4, the tool's ability to predict detector performance in identifying an earthquake occurring in the time interval of $28\text{-}38\,\mathrm{s}$ is demonstrated. Since an earthquake typically amplifies the energy of all signal components, it can be considered a GWN anomaly. Table 3.7 shows that SPE$_{85}$, AR$_{16}$ and MD are the most suitable for detecting this type of anomaly, while T$^2_{16}$ and ZC perform poorly. This is consistent with the score trends shown in the figure, where SPE$_{85}$, AR$_{16}$ and MD exhibit sharp variations during the earthquake, while T$^2_{16}$ and ZC show little deviation compared to their behavior during normal operation.

Interestingly, in both cases, the most effective detectors are not machine learning-based, but rather simpler and less computationally demanding solutions.

## 3.4 Conclusion

The assessment of anomaly detectors is often hampered by the limited availability of anomalous data. In this chapter, we described a systematic framework, we call WOMBATS, for assessing the performance of anomaly detectors on time series data. This framework is based on a dictionary of anomalies modeled as deviations from normal signal behavior, which capture abnormal conditions in the monitored system or acquisition process, such as aging, wear, sudden shifts, hardware malfunctions, or disturbances. Anomalies can be synthetically generated and controlled by a unique parameter that adjusts the severity of the deviation, allowing for control over the difficulty of anomaly detection.

The effectiveness of the WOMBATS framework has been demonstrated in two real-world applications: human health monitoring using ECG signals and structural health monitoring using acceleration data. By assessing various anomaly detectors, the framework highlights which detectors perform best and worst for specific types of anomalies. The numerical results show that the framework can accurately anticipate performance in real-world scenarios, such as identifying artifacts in ECG signals or detecting an earthquake during bridge monitoring.

In summary, this framework provides several important benefits: *i*) it is flexible and adaptable to any monitoring scenario; *ii*) it only requires access to normal data; *iii*) it offers a comprehensive coverage of the effects of real anomalies; *iv*) it ensures consistency with a unified anomaly model and a single parameter controlling anomaly severity.

# Chapter 4

# Theoretical Performance Assessment

As shown in Chapter 3, the probability of detection ($P_D$) is a reliable metric for evaluating anomaly detection (AD) performance. The problem with $P_D$ is that its analytical expression is never available and even for simple sources, e.g., Gaussian sources, it has to be estimated. Although Monte Carlo simulations can address this in practical applications [45], the derivation of theoretical results using $P_D$ remains challenging.

When dealing with complex performance measures, in statistics and Information Theory it is common to adopt criteria that are easier to evaluate and manipulate. For instance, distance measures, such as Jeffrey's divergence or the Bhattacharyya distance, have been studied as simpler substitutes for error probability in hypothesis testing [17, 68, 80, 71].

On this line, to overcome this limitation of $P_D$, in the first Section of this chapter, we introduce information-theoretic measures of *distinguishability* for anomaly-aware and anomaly-agnostic cases. In Section 4.2, we define a framework based on the assumption of Gaussian signals, i.e., both normal and anomalous sources are considered to follow a Gaussian distribution. This is a common assumption in Information Theory as we have seen for the rate-distortion problem in (2.3)-(2.4). In this setting, we highlight the importance of the white anomaly. In particular, we show that the white anomaly is the average anomaly over the set of possible anomalies. It also becomes *typical*, meaning that most of the anomalies resemble the white one, as the signal dimension increases. At the end of the Section, we also outline a procedure for signal generation.

## 4.1 Distinguishability

Given two sources, the normal one $\mathbf{x}^{\mathrm{ok}} \sim f_{\mathbf{x}}^{\mathrm{ok}}$ and anomalous one $\mathbf{x}^{\mathrm{ko}} \sim f_{\mathbf{x}}^{\mathrm{ko}}$, the performance of anomaly detectors depends on the degree to which these two distributions can be *distinguished*. We quantify this using two distinct information-theoretic measures, each modeling a different scenario: one where the detector knows both $f_{\mathbf{x}}^{\mathrm{ok}}$ and $f_{\mathbf{x}}^{\mathrm{ko}}$, and another where it only knows $f_{\mathbf{x}}^{\mathrm{ok}}$.

To formalize these measures, we rely on the differential cross-entropy functional defined as:

$$\mathcal{C}(\mathbf{x}';\mathbf{x}'') = -\int_{\mathbb{R}^n} f_{\mathbf{x}'}(\alpha)\log_2 f_{\mathbf{x}''}(\alpha)\mathrm{d}\alpha \tag{4.1}$$

The differential cross entropy represents the average coding rate, in bits per symbol, of a source with (PDF) $f_{\mathbf{x}'}$ when encoded using a code optimized for a source with PDF $f_{x''}$. In particular if one considers $\mathcal{C}(\mathbf{x};\mathbf{x})$, it equals the differential entropy of $\mathbf{x}$ [34, Chapter 8].

From a statistical point of view $\ell_{\mathbf{x}}(\alpha) = -\log_2 f_{\mathbf{x}}(\alpha)$ represents the negative log-likelihood of observing a symbol $\alpha$ from source $\mathbf{x}$, and $\mathcal{C}(\mathbf{x}';\mathbf{x}'') = \mathbf{E}\left[\ell_{\mathbf{x}''}(\alpha)|\mathbf{x}'\right]$ is the average negative likelihood that an instance is generated by source $\mathbf{x}''$, when it would be from source $\mathbf{x}'$.

### 4.1.1 Distinguishability in anomaly-agnostic detection

When only the distribution of normal signals $f_{\mathbf{x}}^{\mathrm{ok}}$ is known and the anomalous distribution $f_{\mathbf{x}}^{\mathrm{ko}}$ is unknown, we can only consider average coding rates based on a code optimized for $\mathbf{x}^{\mathrm{ok}}$, i.e., $\mathcal{C}(\mathbf{x}^{\mathrm{ko}};\mathbf{x}^{\mathrm{ok}})$ and $\mathcal{C}(\mathbf{x}^{\mathrm{ok}};\mathbf{x}^{\mathrm{ok}})$. To quantify the deviation between normal and anomalous behavior we may measure the increase or decrease of average coding rate compared to the expected case $\mathcal{C}(\mathbf{x}^{\mathrm{ok}};\mathbf{x}^{\mathrm{ok}})$. This can be expressed as

$$\zeta = \mathcal{C}\left(\mathbf{x}^{\mathrm{ko}};\mathbf{x}^{\mathrm{ok}}\right) - \mathcal{C}\left(\mathbf{x}^{\mathrm{ok}};\mathbf{x}^{\mathrm{ok}}\right) \tag{4.2}$$

or equivalently,

$$\zeta = \int_{\mathbb{R}^n} \left[f_{\mathbf{x}}^{\mathrm{ok}}(\alpha) - f_{\mathbf{x}}^{\mathrm{ko}}(\alpha)\right] \log_2 f_{\mathbf{x}}^{\mathrm{ok}}(\alpha)\, d\alpha. \tag{4.3}$$

Since some anomalies can lead to a lower coding rate than normal signals, $\zeta$ is not always positive. As a result, to measure the distinguishability, we use the absolute value, $\mathcal{Z} = |\zeta|$.

From a statistical perspective, $\zeta$ represents the difference in the expected negative log-likelihood for an instance $\alpha$ to be normal, given that $\alpha$ comes from either $\mathbf{x}^{\mathrm{ok}}$ or $\mathbf{x}^{\mathrm{ko}}$:

$$\zeta = \mathbf{E}\left[\ell(\alpha)|\mathbf{x}^{\mathrm{ko}}\right] - \mathbf{E}\left[\ell(\alpha)|\mathbf{x}^{\mathrm{ok}}\right]. \tag{4.4}$$

As anticipated in Chapter 1, the negative log-likelihood term $\ell(\alpha) = -\log_2 f_{\mathbf{x}}^{\mathrm{ok}}(\alpha)$ is a natural AD score, indicating whether the instance $\alpha$ deviates from the distribution representing the normal behavior.

### 4.1.2 Distinguishability in anomaly-aware detection

When both the distributions $f_{\mathbf{x}}^{\mathrm{ok}}$ and $f_{\mathbf{x}}^{\mathrm{ko}}$ are known, AD can be seen as a binary classification problem and tackled using the Neyman-Pearson Lemma [34, Theorem 12.7.1], [74, Theorem 3.1]. According to this lemma, the key quantity to track is

$$r(\alpha) = \log_2 \left[\frac{f_{\mathbf{x}}^{\mathrm{ko}}(\alpha)}{f_{\mathbf{x}}^{\mathrm{ok}}(\alpha)}\right] \tag{4.5}$$

and used as an abnormality score, indicating whether the instance $\alpha$ deviates from normal behavior. The distinguishability between $f_{\mathbf{x}}^{\mathrm{ok}}$ and $f_{\mathbf{x}}^{\mathrm{ko}}$ can be measured by comparing the average score for normal $\mathbf{x}^{\mathrm{ok}}$ and anomalous $\mathbf{x}^{\mathrm{ko}}$ signals, i.e.,

$$\mathcal{J} = \mathbf{E}\left[r(\alpha) \mid \mathbf{x}^{\mathrm{ko}}\right] - \mathbf{E}\left[r(\alpha) \mid \mathbf{x}^{\mathrm{ok}}\right]. \tag{4.6}$$

This expression can also be formulated as

$$\mathcal{J} = \int_{\mathbb{R}^n} f_{\mathbf{x}}^{\mathrm{ko}}(\alpha) \log_2 \left[\frac{f_{\mathbf{x}}^{\mathrm{ko}}(\alpha)}{f_{\mathbf{x}}^{\mathrm{ok}}(\alpha)}\right] \mathrm{d}\alpha + \int_{\mathbb{R}^n} f_{\mathbf{x}}^{\mathrm{ok}}(\alpha) \log_2 \left[\frac{f_{\mathbf{x}}^{\mathrm{ok}}(\alpha)}{f_{\mathbf{x}}^{\mathrm{ko}}(\alpha)}\right] \mathrm{d}\alpha, \tag{4.7}$$

or equivalently:

$$\mathcal{J} = \mathcal{C}\left(\mathbf{x}^{\mathrm{ko}};\mathbf{x}^{\mathrm{ok}}\right) - \mathcal{C}\left(\mathbf{x}^{\mathrm{ko}};\mathbf{x}^{\mathrm{ko}}\right) + \mathcal{C}\left(\mathbf{x}^{\mathrm{ok}};\mathbf{x}^{\mathrm{ko}}\right) - \mathcal{C}\left(\mathbf{x}^{\mathrm{ok}};\mathbf{x}^{\mathrm{ok}}\right), \tag{4.8}$$

Finally, we can write $\mathcal{J}$ as:

$$\mathcal{J} = \mathcal{D}_{\mathrm{KL}}\left(f_{\mathbf{x}}^{\mathrm{ko}} \| f_{\mathbf{x}}^{\mathrm{ok}}\right) + \mathcal{D}_{\mathrm{KL}}\left(f_{\mathbf{x}}^{\mathrm{ok}} \| f_{\mathbf{x}}^{\mathrm{ko}}\right), \tag{4.9}$$

where $\mathcal{D}_{\mathrm{KL}}\left(f' \| f''\right)$ represents the Kullback-Leibler divergence, for which $\mathcal{J}$ becomes a symmetrized version.

In fact, $\mathcal{J}$ is not exactly a novel quantity. The concept behind it has been extensively used in binary classification and pattern recognition problems under the name of *divergence* [68, 81, 148, 107].

In the AD context, measure $\mathcal{J}$ models a detector that is aware of both the normal and anomalous distributions and therefore has access to their optimal codes. Looking at (4.8), $\mathcal{J}$ can be interpreted as the sum of the differences in the average coding rate for both compressed sources with a code optimized for the normal source $\mathcal{C}\left(\mathbf{x}^{\mathrm{ko}}; \mathbf{x}^{\mathrm{ok}}\right) - \mathcal{C}\left(\mathbf{x}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ok}}\right)$ or for the anomalous source $\mathcal{C}\left(\mathbf{x}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ko}}\right) - \mathcal{C}\left(\mathbf{x}^{\mathrm{ko}}; \mathbf{x}^{\mathrm{ko}}\right)$. Since the average coding rate is expected to be shorter when adopted to code a source for which it has been optimized, these differences are expected to grow when the disparity between the distributions $f_{\mathbf{x}}^{\mathrm{ok}}$ and $f_{\mathbf{x}}^{\mathrm{ko}}$ increases. Consequently, higher values of $\mathcal{J}$ correspond to a stronger ability to detect anomalies.

Unlike $\zeta$, the quantity $\mathcal{J}$ is always positive and can be used directly as a measure of the distinguishability between normal and anomalous signals.

### 4.1.3 Discussion

The distinguishability metrics $\mathcal{Z}$ and $\mathcal{J}$ are introduced as simple yet effective measures for evaluating AD performance. Unlike the probability of detection, $P_D$, which quantifies detection in terms of probability, $\mathcal{Z}$ and $\mathcal{J}$ measure distinguishability in bits per symbol. In the case of $\mathcal{Z}$, as highlighted by equation (4.4), this metric quantifies the difference in average scores produced by the likelihood-based detector (LD), where $z(\alpha) = \ell(\alpha)$ for normal and anomalous cases. Similarly, equation (4.6) suggests that $\mathcal{J}$ measures the difference in the average scores generated by the Neyman-Pearson detector (NPD) for which $z(\alpha) = r(\alpha)$. In contrast, $P_D$ can offer a complete characterization of detectors (including LD and NPD) as it considers not only the first-order statistics, i.e., the average, of the scores but their entire distributions.

In future analyses, we will compare the trends of $P_D$ with those of $\mathcal{Z}$ and $\mathcal{J}$ to demonstrate how theoretical metrics align with practical detection outcomes. While this comparison will be largely qualitative, $\mathcal{Z}$ and $\mathcal{J}$ will effectively capture the main characteristics of $P_D$.

Furthermore, while $\mathcal{Z}$ and $\mathcal{J}$ map to the true distinguishability measure $P_D$, $\zeta$ corresponds to a commonly adopted performance metric $\mathrm{AUC}$. Therefore, the results derived in the next section will also involve this functional.

Finally, distinguishability measures implicitly assume that detectors examine an increasing number of signal instances, with performance improving as more data is processed. Similar to how rate and distortion from (2.2) represent best-case bounds that can be approached by increasing system complexity, distinguishability measures indicate how fast a detector gathers enough information to identify an anomaly. The higher the value of the measure, the fewer signal instances are required to confidently declare an anomaly, or, conversely, the greater the confidence in a decision made after analyzing only one instance.

## 4.2 Gaussian framework

### 4.2.1 Signal models

Specifically, we consider signals with $\boldsymbol{\mu}^{\mathrm{ok}} = \boldsymbol{\mu}^{\mathrm{ko}} = \mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}^{\mathrm{ok}}$ and $\boldsymbol{\Sigma}^{\mathrm{ko}} \in \mathbb{R}^{n\times n}$. In general, $\boldsymbol{\Sigma}^{\mathrm{ok}} \neq \boldsymbol{\Sigma}^{\mathrm{ko}}$, but we assume that $\mathrm{tr}(\boldsymbol{\Sigma}^{\mathrm{ok}}) = \mathrm{tr}(\boldsymbol{\Sigma}^{\mathrm{ko}}) = n$, where $\mathrm{tr}(\cdot)$ denotes the matrix trace. This implies that each vector sample contributes on average one unit of energy. With the assumption of zero-mean signals with equal energy, we focus on one possible effect of anomalies: energy redistribution across the signal's subspace. Furthermore, without loss of generality, we assume $\boldsymbol{\Sigma}^{\mathrm{ok}} = \mathrm{diag}\left(\lambda_0^{\mathrm{ok}}, \dots, \lambda_{n-1}^{\mathrm{ok}}\right)$ with $\lambda_0^{\mathrm{ok}} \geq \lambda_1^{\mathrm{ok}} \geq \dots \geq \lambda_{n-1}^{\mathrm{ok}} \geq 0$.

With the established metrics, it is worth exploring the perspectives that emphasize the importance of white noise within the Gaussian framework. One view suggests that white noise represents the average over the set of all possible anomalies. In addition, it reflects the asymptotic behavior of anomalies as the signal's dimensionality increases.

### 4.2.2 Average on the set of possible anomalies

The anomalies modeled as fixed-energy, zero-mean Gaussian vectors are completely characterized by their covariance matrix $\boldsymbol{\Sigma}^{\mathrm{ko}}$, with $\mathrm{tr}(\boldsymbol{\Sigma}^{\mathrm{ko}}) = n$. We can decompose the covariance matrix as $\boldsymbol{\Sigma}^{\mathrm{ko}} = \mathbf{U}^{\mathrm{ko}}\boldsymbol{\Lambda}^{\mathrm{ko}}\mathbf{U}^{\mathrm{ko}\top}$, where $\boldsymbol{\Lambda}^{\mathrm{ko}} = \mathrm{diag}(\lambda_0^{\mathrm{ko}}, \dots, \lambda_{n-1}^{\mathrm{ko}})$ is a diagonal matrix, and $\mathbf{U}^{\mathrm{ko}}$ is an orthonormal matrix.

The set of all possible eigenvalues $\boldsymbol{\lambda}^{\mathrm{ko}} = (\lambda_0^{\mathrm{ko}}, \dots, \lambda_{n-1}^{\mathrm{ko}})^\top$ is defined by:

$$\mathbb{S}^n = \left\{ \boldsymbol{\lambda} \in \mathbb{R}^{+n} \mid \sum_{j=0}^{n-1} \lambda_j = n \right\} \tag{4.10}$$

Meanwhile, the set of all possible orthonormal matrices $\mathbf{U}^{\mathrm{ko}}$ is:

$$\mathbb{O}^n = \left\{ \mathbf{U} \in \mathbb{R}^{n\times n} \mid \mathbf{U}^\top\mathbf{U} = \mathbf{I}_n \right\}. \tag{4.11}$$

By indicating with $\mathcal{U}(\cdot)$ the uniform distribution over the domain of the argument, we assume that $\boldsymbol{\lambda}^{\mathrm{ko}} \sim \mathcal{U}(\mathbb{S}^n)$ when $\boldsymbol{\lambda}^{\mathrm{ko}}$ is unknown and similarly $\mathbf{U}^{\mathrm{ko}} \sim \mathcal{U}(\mathbb{O}^n)$ when $\mathbf{U}^{\mathrm{ko}}$ is unknown, with the two distributions being independent.

Since $\mathbb{S}^n$ is invariant under any permutation of $\lambda_j$, the expected value $\mathbf{E}[\boldsymbol{\lambda}^{\mathrm{ko}}]$ must also be invariant under the same permutations. Therefore, $\mathbf{E}[\lambda_j^{\mathrm{ko}}] = \mathbf{E}[\lambda_k^{\mathrm{ko}}]$ for any $j, k$, and given the constraint on the sum of eigenvalues, we have $\mathbf{E}[\boldsymbol{\Lambda}^{\mathrm{ko}}] = \mathbf{I}_n$. This leads to:

$$\mathbf{E}[\boldsymbol{\Sigma}^{\mathrm{ko}}] = \mathbf{E}\left[\mathbf{U}^{\mathrm{ko}}\boldsymbol{\Lambda}^{\mathrm{ko}}\mathbf{U}^{\mathrm{ko}\top}\right] = \mathbf{E}_{\mathbf{U}}\left[\mathbf{U}^{\mathrm{ko}}\mathbf{E}_{\boldsymbol{\lambda}}[\boldsymbol{\Lambda}^{\mathrm{ko}}]\mathbf{U}^{\mathrm{ko}\top}\right] = \mathbf{E}\left[\mathbf{U}^{\mathrm{ko}}\mathbf{U}^{\mathrm{ko}\top}\right] = \mathbf{I}_n. \tag{4.12}$$

Hence, under this setting, the average anomaly behaves as white.

### 4.2.3 Asymptotic anomalies

White signals are not only the average anomalies but also represent the *typical* anomalies, as formalized by this theorem (proof provided in Appendix B):

**Theorem 4.1.** *If $\boldsymbol{\Sigma}^{\mathrm{ko}} = \mathbf{U}^{\mathrm{ko}}\mathrm{diag}(\boldsymbol{\lambda}^{\mathrm{ko}})\mathbf{U}^{\mathrm{ko}\top}$, where $\boldsymbol{\lambda}^{\mathrm{ko}} \sim \mathcal{U}(\mathbb{S}^n)$ and $\mathbf{U}^{\mathrm{ko}} \sim \mathcal{U}(\mathbb{O}^n)$, then for any $\beta > 1/2$, the quantity $\Delta_{\mathrm{F}} = n^{-\beta}\|\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n\|_F$, where $\|\cdot\|_F$ is the Frobenius norm, converges to 0 in probability as $n \to \infty$.*

This means that, as the dimension $n$ increases, most of potential anomalies exhibit behavior similar to white signals. From an AD perspective, when dealing with a signal of sufficiently large dimension, it is reasonable for the designer to consider the white anomaly as a reference.

### 4.2.4 Distinguishability in the Gaussian framework

First of all, under the Gaussian assumption, we can obtain an analytical expression for $\mathcal{C}$ as shown in the following Lemma, with the complete procedure provided in the Appendix B.

**Lemma 4.1.** *If* $\mathbf{x}' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}')$ *and* $\mathbf{x}'' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}'')$*, then*

$$\mathcal{C}(\mathbf{x}';\mathbf{x}'') = \frac{1}{2\ln 2}\left\{\ln\left[(2\pi)^n\left|\mathbf{\Sigma}''\right|\right] + \mathrm{tr}\left[(\mathbf{\Sigma}'')^{-1}\mathbf{\Sigma}'\right]\right\} \tag{4.13}$$

*where* $|\cdot|$ *denotes the determinant of the matrix.*

Under the assumption of Gaussian sources, by combining the definitions of $\zeta$ from (4.2) and $\mathcal{J}$ from (4.8) with the expression for $\mathcal{C}$ in the Gaussian case from (4.13), we obtain:

$$\zeta = \frac{1}{2\ln 2}\mathrm{tr}\left[\tilde{\mathbf{\Sigma}} - \mathbf{I}_n\right] \tag{4.14}$$

$$\mathcal{Z} = \frac{1}{2\ln 2}\left|\mathrm{tr}\left[\tilde{\mathbf{\Sigma}} - \mathbf{I}_n\right]\right| \tag{4.15}$$

$$\mathcal{J} = \frac{1}{2\ln 2}\mathrm{tr}\left[\tilde{\mathbf{\Sigma}} + \tilde{\mathbf{\Sigma}}^{-1} - 2\mathbf{I}_n\right], \tag{4.16}$$

where $\tilde{\mathbf{\Sigma}} = (\mathbf{\Sigma}^{\mathrm{ok}})^{-1}\mathbf{\Sigma}^{\mathrm{ko}}$ is an $n \times n$ matrix. Note that since $\tilde{\mathbf{\Sigma}}$ is linear with respect to $\mathbf{\Sigma}^{\mathrm{ko}}$, $\zeta$ also varies linearly, while it is evident that $\mathcal{Z}$ and $\mathcal{J}$ are convex with respect to $\mathbf{\Sigma}^{\mathrm{ko}}$. Additionally, both $\zeta$ and $\mathcal{J}$ vanish when $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{\Sigma}^{\mathrm{ko}}$.

A notable special case arises when the normal signal is white, i.e., when $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{I}_n$. This results in $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}^{\mathrm{ko}}$, leading to $\zeta = \mathcal{Z} = 0$. This outcome is not surprising, as the distinguishability modeled by $\mathcal{Z}$ relies on the statistics of $\mathbf{x}^{\mathrm{ok}}$, which in this scenario has no exploitable structure.

We can also compute the functional $\zeta_{\mathbf{I}}$, and the distinguishability measures $\mathcal{Z}_{\mathbf{I}}$ and $\mathcal{J}_{\mathbf{I}}$, representing $\zeta$ and $\mathcal{Z}$, $\mathcal{J}$ when $\mathbf{\Sigma}^{\mathrm{ko}} = \mathbf{I}_n$. In this case, $\tilde{\mathbf{\Sigma}} = (\mathbf{\Sigma}^{\mathrm{ok}})^{-1}$, and is a diagonal matrix whose elements on the diagonal are $u_j = \frac{1}{\lambda_j^{\mathrm{ok}}}$.

Using these quantities, the three measures can be expressed as:

$$\zeta_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta - 1}(u_j - 1) \tag{4.17}$$

$$\mathcal{Z}_{\mathbf{I}} = \frac{1}{2\ln 2}\left|\sum_{j=0}^{n_\theta - 1}(u_j - 1)\right| \tag{4.18}$$

$$\mathcal{J}_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{n_\theta - 1}\left(u_j + \frac{1}{u_j} - 2\right). \tag{4.19}$$

Note that, due to the linearity of $\zeta$, we have $\zeta_{\mathbf{I}} = \mathbf{E}[\zeta]$ and by Jensen's inequality, this leads to $\mathcal{Z}_{\mathbf{I}} \leq \mathbf{E}[\mathcal{Z}]$. Moreover, by Jensen's inequality and the convexity of $\mathcal{J}$ we also have $\mathcal{J}_{\mathbf{I}} \leq \mathbf{E}[\mathcal{J}]$.

### 4.2.5 Signal generation

Normal signals are modeled as $\mathbf{x}^{\mathrm{ok}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{ok}}\right)$, where $\boldsymbol{\Sigma}^{\mathrm{ok}}$ is the diagonal matrix resulting from the eigendecomposition of a given matrix $\mathbf{K}$, such that $\mathbf{K} = \mathbf{U}\boldsymbol{\Sigma}^{\mathrm{ok}}\mathbf{U}^{\top}$, with $\mathbf{U}$ orthonormal. The matrix $\mathbf{K}$ is an $n \times n$ positive semidefinite matrix, where the $(j,k)$-entry is assigned as $\omega^{|j-k|}$ for $j,k = 0,\dots,n-1$. The parameter $\omega$ is chosen to generate different degrees of non-whiteness, measured by the so-called *localization*, defined as:

$$\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}} = \frac{\mathrm{tr}({\boldsymbol{\Sigma}^{\mathrm{ok}}}^2)}{\mathrm{tr}^2(\boldsymbol{\Sigma}^{\mathrm{ok}})} - \frac{1}{n}. \tag{4.20}$$

The localization ranges from $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}} = 0$ for white signals to $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}} = 1 - \frac{1}{n}$ when the entire energy is concentrated along a single direction of the signal space (see [99] for more details). To illustrate the effects of realistic localizations [25], we consider the values of $\omega$ corresponding to $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}} \in \{0, 0.05, 0.2\}$.

Anomalous signals are generated as $\mathbf{x}^{\mathrm{ko}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{ko}}\right)$, where $\boldsymbol{\Sigma}^{\mathrm{ko}}$ is drawn randomly according to the uniform distribution defined in Section 4.2.2. Specifically, $\boldsymbol{\Sigma}^{\mathrm{ko}} = \mathbf{U}^{\mathrm{ko}}\boldsymbol{\Lambda}^{\mathrm{ko}}{\mathbf{U}^{\mathrm{ko}}}^{\top}$, with $\boldsymbol{\Lambda}^{\mathrm{ko}} = \mathrm{diag}(\boldsymbol{\lambda}^{\mathrm{ko}})$, where $\boldsymbol{\lambda}^{\mathrm{ko}} \sim \mathcal{U}\left(\mathbb{S}^n\right)$ and $\mathbf{U}^{\mathrm{ko}} \sim \mathcal{U}\left(\mathbb{O}^n\right)$ are generated independently. The term $\boldsymbol{\lambda}^{\mathrm{ko}}$ is sampled according to the procedure from [112]: we first draw $\xi_j \sim \mathcal{U}\left([0,1]\right)$ for $j = 0,\dots,n-1$ and then set

$$\lambda_j^{\mathrm{ko}} = \frac{n \log \xi_j}{\sum_{k=0}^{n-1} \log \xi_k}, \tag{4.21}$$

where $\lambda_j^{\mathrm{ko}} > 0$ and the entries of $\boldsymbol{\lambda}^{\mathrm{ko}}$ sum to $n$. To generate term $\mathbf{U}^{\mathrm{ko}} \sim \mathcal{U}\left(\mathbb{O}^n\right)$ we follow [104], where the matrix $\mathbf{A}$ is drawn from the Ginibre ensemble [49], which means its entries are independent and normally distributed, $A_{j,k} \sim \mathcal{N}\left(0,1\right)$ for $j,k = 0,\dots,n-1$. We assign to $\mathbf{U}^{\mathrm{ko}}$ the orthogonal factor in the $QR$ decomposition of $\mathbf{A}$.

### 4.2.6 Numerical evidence

This random sampling is first used to provide numerical support for Theorem 4.1. Figure 4.1 shows the vanishing trends of the average squared ($\Delta_{\mathrm{F}}$ with $\beta = 1$) and uniform (defined as $\Delta_{\max} = \|\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n\|_{\max} = \max_{j,k} |(\boldsymbol{\Sigma}^{\mathrm{ko}})_{j,k} - (\mathbf{I}_n)_{j,k}|$) deviations of uniformly distributed covariance matrices $\boldsymbol{\Sigma}^{\mathrm{ko}}$ from $\mathbf{I}_n$. The trend for $\Delta_{\mathrm{F}}$ confirms Theorem 4.1, while the trend for $\Delta_{\max}$ empirically extends this result to a stronger deviation measure. Thus, for sufficiently large $n$, white Gaussian noise can be considered a good candidate to represent anomalies that arise independently of the statistics of the normal signal.

In the next chapters, further numerical evidence will be provided that confirms the effectiveness of $\zeta$, $\mathcal{Z}$, and $\mathcal{J}$ in anticipating performance expressed in terms of $P_D$.

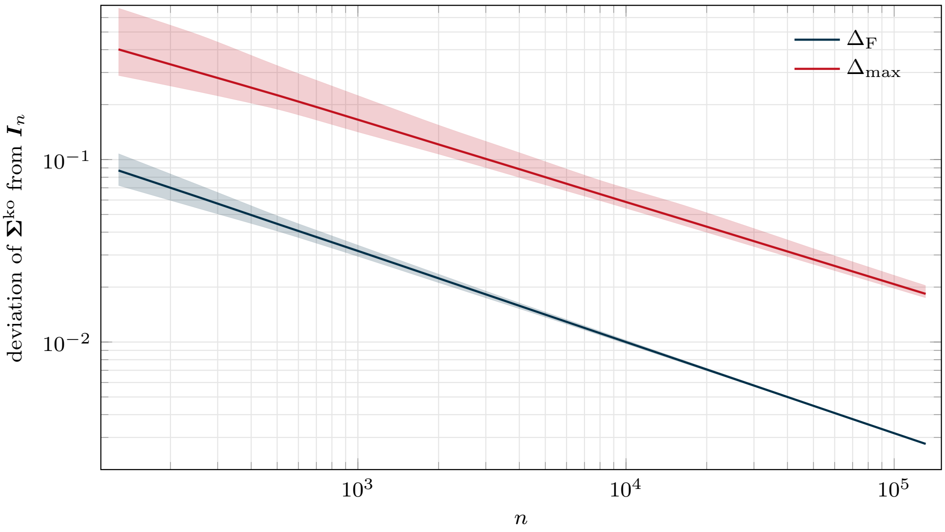


FIGURE 4.1: Trends of $\Delta_{\mathrm{F}}$ and $\Delta_{\max}$ for $n \in \{2^k\}_{k=7}^{17}$. Solid lines are mean values over $2\,000$ trials, while shaded areas represent $98\%$ of the population.

## 4.3 Conclusion

In this chapter, we defined a framework for the theoretical assessment of anomaly detection performance. We began by introducing two information-theoretic measures that quantify distinguishability between normal and anomalous signals, applicable in both anomaly-agnostic and anomaly-aware scenarios. These metrics are also complemented with a statistical interpretation. To derive closed-form expressions for these metrics and to model and subsequently generate possible anomalies, we described a set-up based on Gaussian signals. This setting allowed us to derive the following theoretical results:

- The white anomaly is the average anomaly over all possible anomalies.
- As the dimensionality of the signal increases, the white anomaly becomes typical.
- The distinguishability metrics for white anomalies are representative of a detector's average performance across various types of anomalies.

In this part of the dissertation, we have tackled the issue of performance assessment of the anomaly detection task. While in Chapter 3 the outlined framework offers a ready-to-use tool for practical AD applications, in this Chapter we defined a framework that can be adopted when theoretical derivations involving anomaly detection are required. In the next part, we will examine the interplay between signal compression and anomaly detection, where both practical and theoretical performance assessment tools will play a key role.

# Part II

# Compression and Anomaly Detection

# Chapter 5

# Rate-Distortion Theory and Distinguishability

As already anticipated in Chapter 2, in a monitoring system comprising numerous sensors, to reduce the transmission bitrate, sensor readings are often compressed using lossy techniques designed to retain useful information while reducing data size [94, 87]. A common practice is to store and process compressed data using a remote unit [59, 157, 58]. Before reaching cloud facilities, these compressed bitstreams often pass through hierarchical aggregation layers and intermediate devices, commonly referred to as the *edge* of the cloud [136]. The overall infrastructure is depicted in Figure 5.1. For latency, privacy, or security reasons, certain computational tasks, such as anomaly detection (AD), may benefit from being performed at the edge. However, lossy compression achieves efficiency by discarding some signal details, resulting in distortion between the original and compressed signals, and in a potential loss of features that could help anomaly detectors.

Typically, acquisition systems are subject to distortion constraints, designed to address the trade-off between compression and distortion in the best way possible. In the considered scenario, rate-distortion exists alongside another fundamental trade-off: the relationship between signal distortion and the ability to distinguish between normal and anomalous signals. In this Chapter, we analyze this trade-off using the same information-theoretic framework applied in rate-distortion analysis presented in Chapter 2, and the AD performance assessment framework defined in Chapter 4. We demonstrate that the rate-distortion and distinguishability-distortion trade-offs are fundamentally different.

The impact of compression on a detector's ability to distinguish between two information sources has been widely explored in the literature. In [4], the problem of hypothesis

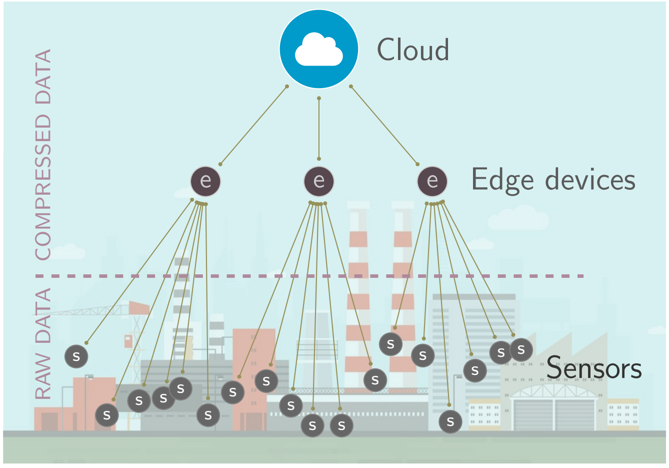


FIGURE 5.1: A sensor equipped plant whose compressed acquisitions are aggregated at the edge before being sent to the cloud.

testing under a rate constraint for a single source is examined. This work has been extended to multiterminal data compression scenarios, addressing statistical inference problems in [57]. However, unlike the analysis we perform in this chapter, these works do not impose a distortion constraint since reconstructing the original signals is not a requirement. In contrast, we assume that compression is designed to maintain the quality of service required by the processing tasks performed on the reconstructed data.

In some sense, our analysis is loosely related to the *information bottleneck* framework [146, 140]. That approach allows optimizing the trade-off between rate and a distortion measure more broadly defined with respect to (2.1). Unlike classical rate-distortion analysis that targets signal reconstruction, this framework selects which aspects of the original signal should be preserved during compression. The compressor reduces the rate while retaining the features relevant to a secondary signal that accounts for the specific needs of the target application. In the specific case of AD, this technique has been applied in works such as [35, 36]. For a comprehensive overview of the information bottleneck see [62]. Similarly to the information bottleneck principle, the authors of [138] proposed an end-to-end framework that jointly optimizes the rate constrained on a task-specific objective, evaluating its effectiveness in the context of classification tasks. In our case, compression is not adapted to AD but designed in a classical sense to preserve the signal's overall information for subsequent analysis (see Figure 5.2). We extend the traditional rate-distortion framework by adding a measure of distinguishability between normal and anomalous sources subject to classical compression.

The distinction between our approach and the information bottleneck method is the same that differentiates us from another variation of the classical rate-distortion theory that substitutes energy-based distortion with perceptual criteria [19].

Although not directly related to the problem considered in this chapter, it is worth mentioning studies such as [123, 137], which focus on how lossy compression affects the estimation of certain parameters (e.g., the mean) of the original signal.

Finally, there are other scenarios where rate and distortion are combined with additional performance metrics that account for specific characteristics of the system. For instance, [46] includes computational efficiency in the analysis of rate-distortion for wavelet-based video coding.

This chapter presents an analysis of the performance of a generic AD system receiving input signals distorted by compression mechanisms that comply with the rate-distortion trade-off. To this end, in Section 5.1 we revisit the traditional rate-distortion theory presented in Chapter 2 by explicating the optimal compression scheme and the distributions of compressed normal and anomalous Gaussian sources. Section 5.2 specializes the anomaly-aware and anomaly-agnostic distinguishability measures defined in Chapter 4 to the compressed domain, providing closed-form expressions for cases where the compressed signals follow a Gaussian distribution. The results on the interaction between distortion and distinguishability are supported by numerical evidence in Section 5.3. Additionally, we demonstrate that the findings hold even when certain assumptions are relaxed, in particular when analyzing signals from real-world applications and/or using practical compression methods.

In general, both theoretical and experimental results show that compression techniques optimized for information preservation may not always be the most effective in retaining distinguishability.

## 5.1 Rate-Distortion Compression

A typical compression system is designed to achieve the best trade-off between rate and distortion under normal operating conditions, i.e., when the observed signal follows the normal model $\mathbf{x} = \mathbf{x}^{\text{ok}}$. The trade-off between rate and distortion has been reviewed, in particular for the Gaussian sources, in Chapter 2. Here we also want to consider the situation when an anomalous source unexpectedly replaces the normal one and the encoder must process anomalous signals, i.e., $\mathbf{x} = \mathbf{x}^{\text{ko}}$. In this scenario, since the encoder is adapted to the normal signal, $\mathbf{x}^{\text{ok}}$ is optimally compressed in the rate-distortion sense into $\hat{\mathbf{x}}^{\text{ok}}$, and the anomalous signal $\mathbf{x}^{\text{ko}}$ is sub-optimally compressed into $\hat{\mathbf{x}}^{\text{ko}}$.

Within the Gaussian assumption, it is possible to derive the PDF of the distorted signal $\hat{\mathbf{x}}$ and the conditional PDF $f_{\hat{\mathbf{x}}|\mathbf{x}}$, which stochastically maps an input $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ to $\hat{\mathbf{x}}$. Although not often explicitly stated, the expression of $f_{\hat{\mathbf{x}}|\mathbf{x}}$ becomes important when the compression mechanism is used to encode a signal different from that for which it was originally designed, as is the case of our analysis.

Recall that the rate-distortion problem in (2.3)-(2.4) is governed by the reverse water-filling parameter $\theta \in [0, \lambda_0]$, such that $\tau_j = \min\{1, \theta/\lambda_j\}$ represents the fraction of energy lost to distortion along the $j$-th component. By treating a zero-variance Gaussian as a Dirac delta function we can define $\mathbf{S}_\theta = \mathbf{I}_n - \mathbf{T}_\theta$ with $\mathbf{T}_\theta = \text{diag}(\tau_0, \ldots, \tau_{n-1})$ to account for the fraction of energy that survives distortion along each component. With this setup, and leveraging the Gaussian framework introduced in Chapter 4, we can state the following lemmas, with the proofs provided in Appendix C.

**Lemma 5.1.** *If $\mathbf{x}^{\text{ok}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\text{ok}}\right)$ is a memoryless source and we constrain the distortion $D \leq \delta$, the optimally distorted signal has distribution*

$$\hat{\mathbf{x}}^{\text{ok}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\text{ok}}\mathbf{S}_\theta\right) \tag{5.1}$$

*and the optimal encoding mapping is*

$$f_{\hat{\mathbf{x}}|\mathbf{x}}(\alpha, \beta) = G_{\mathbf{S}_\theta\beta, \mathbf{\Sigma}^{\text{ok}}\mathbf{S}_\theta\mathbf{T}_\theta}(\alpha) \tag{5.2}$$

*where $G_{\mu,\mathbf{\Sigma}}(\cdot)$ denotes a Gaussian PDF with mean vector $\mu$ and covariance matrix $\mathbf{\Sigma}$.*

**Lemma 5.2.** *If an anomalous source $\mathbf{x}^{\text{ko}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\text{ko}}\right)$ is encoded with the compression scheme $f^{\text{ok}}_{\hat{\mathbf{x}}|\mathbf{x}}$ of Lemma 5.1, then*

$$\hat{\mathbf{x}}^{\text{ko}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{S}_\theta\mathbf{\Sigma}^{\text{ko}}\mathbf{S}_\theta + \theta\mathbf{S}_\theta\right). \tag{5.3}$$

This result has two important corner cases.

- If $\theta \to 0^+$ there is no distortion. In fact, since $\mathbf{S}_\theta = \mathbf{I}_n$, according to Lemma 5.2, $\hat{\mathbf{x}}^{\text{ko}} \sim \mathbf{x}^{\text{ko}}$.
- If $\mathbf{x}^{\text{ko}} \sim \mathbf{x}^{\text{ok}}$ no anomaly is observed, $\mathbf{\Sigma}^{\text{ok}} = \mathbf{\Sigma}^{\text{ko}}$, and

  $$\mathbf{S}_\theta\mathbf{\Sigma}^{\text{ko}}\mathbf{S}_\theta + \theta\mathbf{S}_\theta = [\mathbf{S}_\theta + \theta(\mathbf{\Sigma}^{\text{ok}})^{-1}]\mathbf{\Sigma}^{\text{ok}}\mathbf{S}_\theta = \mathbf{\Sigma}^{\text{ok}}\mathbf{S}_\theta$$

  where the last equality uses $\mathbf{S}_\theta = \max\left\{0, \mathbf{I}_n - \theta(\mathbf{\Sigma}^{\text{ok}})^{-1}\right\}$. The possible disagreements between $\mathbf{S}_\theta + \theta(\mathbf{\Sigma}^{\text{ok}})^{-1}$ and $\mathbf{I}_n$ correspond to components multiplied by zeros in the last $\mathbf{S}_\theta$ factor. Therefore, Lemma 5.2 and Lemma 5.1 can be compared to confirm that $\hat{\mathbf{x}}^{\text{ko}} \sim \hat{\mathbf{x}}^{\text{ok}}$.

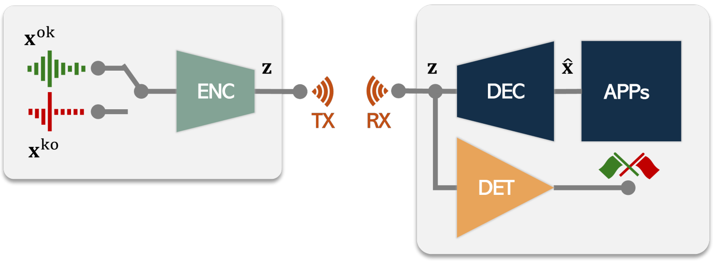


FIGURE 5.2: Signal chain is adapted to the normal signal $\mathbf{x}^{\text{ok}}$ to best address the rate-distortion trade-off, guaranteeing a certain quality of service to a given application (APP). An anomalous signal $\mathbf{x}^{\text{ko}}$ may appear and a detector working on the compressed signal $\mathbf{z}$ should be able to detect it.

Lemmas 5.1 and 5.2 imply that when both normal and anomalous signals follow Gaussian distributions prior to compression, the two resulting compressed sources also follow Gaussian distributions.

## 5.2 Distinguishability in the compressed domain

After introducing the rate-distortion trade-off that rules lossy compression methods, we now analyze the effects of processing a compressed signal on AD performance. Since compression introduces distortion, it also limits the information available to the detector, reducing its ability to distinguish whether the transmitted signal differs from the usual observations. Consequently, in the case of AD performed on compressed signals, the trade-off becomes three-dimensional, involving rate, distortion and distinguishability.

As illustrated in Figure 5.2, the compressed signal $\mathbf{z}$ is fed to the anomaly detector. Since we assume the decoding stage to be injective, $\mathbf{z}$ contains the same information as the reconstructed signal $\hat{\mathbf{x}}$, meaning that, in principle, analyzing $\mathbf{z}$ is equivalent to analyzing $\hat{\mathbf{x}}$. The goal of the detector is to distinguish between normal reconstructed signals $\hat{\mathbf{x}}^{\text{ok}} \sim f_{\hat{\mathbf{x}}}^{\text{ok}}$ and anomalous ones $\hat{\mathbf{x}}^{\text{ko}} \sim f_{\hat{\mathbf{x}}}^{\text{ko}}$. According to the framework developed in Chapter 4, we quantify this difference by specializing the $\mathcal{J}$ and $\mathcal{Z}$ measures, modeling different scenarios: $\mathcal{J}$ is applied when the detector knows both $f_{\hat{\mathbf{x}}}^{\text{ok}}$ and $f_{\hat{\mathbf{x}}}^{\text{ko}}$, and $\mathcal{Z}$ when it only knows $f_{\hat{\mathbf{x}}}^{\text{ok}}$. Additionally, when both normal and anomalous signals follow Gaussian distributions prior to compression, the performance of anomaly detectors depends on the degree to which the two resulting distributions in (5.1) and (5.3) can be distinguished.

### 5.2.1 Distinguishability in anomaly-agnostic detection

When only the distribution of normal signals $f_{\hat{\mathbf{x}}}^{\text{ok}}$ is known and the anomalous distribution $f_{\hat{\mathbf{x}}}^{\text{ko}}$ is unknown, we can consider $\zeta$ and $\mathcal{Z}$ specialized to the compressed domain:

$$\zeta = \int_{\mathbb{R}^n} \left[ f_{\hat{\mathbf{x}}}^{\text{ok}}(\alpha) - f_{\hat{\mathbf{x}}}^{\text{ko}}(\alpha) \right] \log_2 f_{\hat{\mathbf{x}}}^{\text{ok}}(\alpha)\, d\alpha \tag{5.4}$$

$$\mathcal{Z} = \left| \int_{\mathbb{R}^n} \left[ f_{\hat{\mathbf{x}}}^{\text{ok}}(\alpha) - f_{\hat{\mathbf{x}}}^{\text{ko}}(\alpha) \right] \log_2 f_{\hat{\mathbf{x}}}^{\text{ok}}(\alpha)\, d\alpha \right| \tag{5.5}$$

or equivalently

$$\zeta = \mathcal{C}\left(\hat{\mathbf{x}}^{\text{ko}}; \hat{\mathbf{x}}^{\text{ok}}\right) - \mathcal{C}\left(\hat{\mathbf{x}}^{\text{ok}}; \hat{\mathbf{x}}^{\text{ok}}\right) \tag{5.6}$$

$$\mathcal{Z} = \left|\mathcal{C}\left(\hat{\mathbf{x}}^{\text{ko}}; \hat{\mathbf{x}}^{\text{ok}}\right) - \mathcal{C}\left(\hat{\mathbf{x}}^{\text{ok}}; \hat{\mathbf{x}}^{\text{ok}}\right)\right| \tag{5.7}$$

Under the assumption of Gaussian sources, the optimal encoder (in rate-distortion terms) retains only those components $j$ for which $\lambda_j^{\mathrm{ok}} > \theta$. Hence, the distributions $f_{\hat{\mathbf{x}}}^{\mathrm{ok}}$ and $f_{\hat{\mathbf{x}}}^{\mathrm{ko}}$ given by (5.1) and (5.3) have non-zero values only for the first $k_\theta$ components, with $k_\theta = \arg\max_j\{\lambda_j^{\mathrm{ok}} > \theta\}$. The remaining $n - k_\theta$ components are set to zero, making them useless for distinguishing normal and anomalous cases. For this reason, we only focus on the first $k_\theta$ components of $\hat{\mathbf{x}}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ko}}$, which follow Gaussian distributions with covariance matrices $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}}$ and $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$ corresponding to the $k_\theta \times k_\theta$ upper-left submatrices of $\mathbf{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta$ in (5.1) and $\mathbf{S}_\theta\mathbf{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta + \theta\mathbf{S}_\theta$ in (5.3), respectively.

By defining $\tilde{\mathbf{\Sigma}}_\theta = (\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}})^{-1}\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$ as the $k_\theta \times k_\theta$ upper-left submatrix of $(\mathbf{\Sigma}^{\mathrm{ok}})^{-1}\mathbf{\Sigma}^{\mathrm{ko}}\mathbf{S}_\theta + \mathbf{T}_\theta$ and using the expression of $\zeta$ in the Gaussian case from (4.14), we obtain

$$\zeta = \frac{1}{2\ln 2}\mathrm{tr}\left[\tilde{\mathbf{\Sigma}}_\theta - \mathbf{I}_{k_\theta}\right]. \tag{5.8}$$

In this case, since $\tilde{\mathbf{\Sigma}}_\theta$ varies linearly with respect to $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$, $\zeta$ is also linear. Moreover, with the expression of $\zeta$ in (5.8), we can consider important corner cases:

- When $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}} = \hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$, $\zeta = 0$.
- When normal signal is white, i.e., $\mathbf{\Sigma}^{\mathrm{ok}} = \mathbf{I}_n$, and for any $\theta < 1$, we have $\mathbf{T}_\theta = \theta\mathbf{I}_n$ and $k_\theta = n$. This results in $\tilde{\mathbf{\Sigma}}_\theta = (1-\theta)\mathbf{\Sigma}^{\mathrm{ko}} + \theta\mathbf{I}_n$, leading to $\zeta = 0$.
- When the anomaly is white, i.e., $\mathbf{\Sigma}^{\mathrm{ko}} = \mathbf{I}_n$, the matrix $\tilde{\mathbf{\Sigma}}_\theta$ is diagonal with elements on the diagonal $u_{\theta,j} = \frac{1}{\lambda_j^{\mathrm{ok}}}\left(1 - \frac{\theta}{\lambda_j^{\mathrm{ok}}}\right) + \frac{\theta}{\lambda_j^{\mathrm{ok}}}$, which results in

$$\zeta_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{k_\theta - 1}(u_{\theta,j} - 1). \tag{5.9}$$

The simpler expression of $\zeta_{\mathbf{I}}$ in (5.9) allows us to derive the following theorem, whose proof is found in the Appendix C.

**Theorem 5.1.** *If* $\bar{k} = \arg\max_k \left\{\lambda_k^{\mathrm{ok}} \geq \lambda_k^{\mathrm{ko}} = 1\right\}$, *then* $\zeta_{\mathbf{I}} = 0$ *for at least one point* $0 < \theta < \lambda_{\bar{k}}^{\mathrm{ok}}$.

For the case of a white anomaly, the intuition behind this theorem is as follows. When there is no distortion, i.e., no compression, since both $\hat{\mathbf{x}}^{\mathrm{ko}}$ and $\hat{\mathbf{x}}^{\mathrm{ok}}$ have the same average energy, and the coding is adapted for $\hat{\mathbf{x}}^{\mathrm{ok}}$, $\mathcal{C}(\hat{\mathbf{x}}^{\mathrm{ko}}; \hat{\mathbf{x}}^{\mathrm{ok}}) > \mathcal{C}(\hat{\mathbf{x}}^{\mathrm{ok}}; \hat{\mathbf{x}}^{\mathrm{ok}})$, resulting in a positive $\zeta_{\mathbf{I}}$. On the other hand, when the distortion is so high that only the first component of $\mathbf{x}^{\mathrm{ok}}$ remains, i.e., $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}} = \lambda_0^{\mathrm{ok}} - \theta$, only one component survives in $\hat{\mathbf{x}}^{\mathrm{ko}}$. In this setting, $\zeta_{\mathbf{I}}$ depends on the difference between the scalar quantities $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}}$ and $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$. With some algebraic manipulation, it can be proven that $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}} > \hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$, resulting in a negative $\zeta_{\mathbf{I}}$.

Since $\zeta_{\mathbf{I}}$ is continuous with respect to $\theta$, it must cross zero at least once. Therefore, there exists at least one critical distortion level where detectors that do not use information on the anomaly become ineffective.

The previous properties of $\zeta$ together with Theorem 5.1 in the Gaussian case allow us to formulate the following corollary which proof arises naturally from considering that $\mathcal{Z} = |\zeta|$.

**Corollary 5.1.** *The anomaly-agnostic distiguishability measure vanishes, i.e.,* $\mathcal{Z} = 0$, *when at least one of the following conditions is satisfied:*

- $\hat{\mathbf{\Sigma}}_\theta^{\mathrm{ok}} = \hat{\mathbf{\Sigma}}_\theta^{\mathrm{ko}}$

- $\boldsymbol{\Sigma}^{\mathrm{ok}} = \mathbf{I}_n$
- $\boldsymbol{\Sigma}^{\mathrm{ko}} = \mathbf{I}_n$ *and* $\theta = \theta^*$ *with* $\theta^*$ *one of the critical distortion levels.*

### 5.2.2 Distinguishability in anomaly-aware detection

When both the distributions $f^{\mathrm{ok}}_{\hat{\mathbf{x}}}$ and $f^{\mathrm{ko}}_{\hat{\mathbf{x}}}$ are known, distinguishability can be measured with $\mathcal{J}$ specialized to the compressed domain such that:

$$\mathcal{J} = \int_{\mathbb{R}^n} f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\alpha) \log_2 \left[ \frac{f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\alpha)}{f^{\mathrm{ok}}_{\hat{\mathbf{x}}}(\alpha)} \right] \mathrm{d}\alpha + \int_{\mathbb{R}^n} f^{\mathrm{ok}}_{\hat{\mathbf{x}}}(\alpha) \log_2 \left[ \frac{f^{\mathrm{ok}}_{\hat{\mathbf{x}}}(\alpha)}{f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\alpha)} \right] \mathrm{d}\alpha, \tag{5.10}$$

or equivalently:

$$\mathcal{J} = \mathcal{C}\left(\hat{\mathbf{x}}^{\mathrm{ko}}; \hat{\mathbf{x}}^{\mathrm{ok}}\right) - \mathcal{C}\left(\hat{\mathbf{x}}^{\mathrm{ko}}; \hat{\mathbf{x}}^{\mathrm{ko}}\right) + \mathcal{C}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \hat{\mathbf{x}}^{\mathrm{ko}}\right) - \mathcal{C}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \hat{\mathbf{x}}^{\mathrm{ok}}\right), \tag{5.11}$$

Under the Gaussian assumption, the distinguishability measure $\mathcal{J}$ becomes:

$$\mathcal{J} = \frac{1}{2\ln 2} \mathrm{tr}\left[ \tilde{\boldsymbol{\Sigma}}_\theta + \tilde{\boldsymbol{\Sigma}}_\theta^{-1} - 2\mathbf{I}_{k_\theta} \right], \tag{5.12}$$

From this expression it is evident that $\mathcal{J}$ is convex with respect to $\hat{\boldsymbol{\Sigma}}^{\mathrm{ko}}_\theta$, and it vanishes when $\hat{\boldsymbol{\Sigma}}^{\mathrm{ok}}_\theta = \hat{\boldsymbol{\Sigma}}^{\mathrm{ko}}_\theta$. Additionally, the same rationale used to derive $\zeta_{\mathbf{I}}$ in (5.9) allows us to express:

$$\mathcal{J}_{\mathbf{I}} = \frac{1}{2\ln 2} \sum_{j=0}^{k_\theta - 1} \left( u_{\theta,j} + \frac{1}{u_{\theta,j}} - 2 \right). \tag{5.13}$$

## 5.3 Numerical evidence

### 5.3.1 Simulation setup

In this subsection, we introduce three different encoding schemes adopted to compress both normal signals and anomalies. We also describe well-established detectors and recall a practical performance metric. This metric can be computed in both anomaly-aware and anomaly-agnostic scenarios and will serve as a basis for comparing the distinguishability measures $\mathcal{Z}$ and $\mathcal{J}$.

**Normal signal and anomalies** The Gaussian signals $\mathbf{x}^{\mathrm{ok}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{ok}}\right)$ and $\mathbf{x}^{\mathrm{ko}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{ko}}\right)$ are generated following the procedure outlined in Chapter 4.

To further validate our theoretical framework, we consider two realistic applications involving non-stationary and non-Gaussian signals: ECG and ACC.

ECG signals are generated following the approach in [102][1], with a setup taken from [96]. Specifically, the heart rate is randomly drawn from the range of $60$-$100\,\mathrm{bpm}$, and the sampling rate is set to $256\,\mathrm{sps}$. We generate $10^5$ segments, each containing $512$ samples, from which vectors $\mathbf{x}^{\mathrm{ok}} \in \mathbb{R}^n$ with $n = 64$ are randomly selected.

ACC signals come from structural health monitoring of a viaduct along an Italian motorway [101, 105]. A total of 90 three-axis accelerometers have been deployed, each recording 100 samples per second along three different axes. These signals represent the

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

viaduct's elastic response to external stimuli like traffic or environmental conditions. In this study, however, we focus on readings relative to a single axis of a single sensor.

**Compressors** We consider three compression schemes adapted to the normal signal, which are applied to both normal and anomalous instances. Specifically, the signal $\mathbf{x}$ is compressed and then reconstructed as $\hat{\mathbf{x}}$ using the theoretically optimal compressor and the dimensionality reduction-based strategies recalled in Chapter 2:

1. Rate-distortion (RD): This compression method is defined by the optimization problem in (2.2), which achieves the minimum transmission rate possible for a given level of distortion.

2. Principal component analysis (PCA): This compression technique consists of projecting $\mathbf{x}$ onto the subspace spanned by the eigenvectors of $\mathbf{\Sigma}^{\mathrm{ok}}$ defined by the largest eigenvalues. PCA is a linear compression technique that minimizes distortion when the $n$-dimensional vector $\mathbf{x}$ is represented in a lower-dimensional space [101, 105].

3. Autoencoder (AE): This approach uses the autoencoder neural network to learn a latent, non-linear signal representation, essentially generalizing PCA. The autoencoder we consider consists of fully connected layers, with compression dimension $k < n$. The network architecture includes layers of size $n$, $4n$, $2n$, and $k$, followed by a mirrored structure of layers of size $k$, $2n$, $4n$, and $n$. These networks are trained to minimize distortion $D$ as defined in (2.1)[2].

The distortion introduced by the compressors is measured with the normalized distortion $d = D/n$, that by considering a memoryless source $\mathbf{x} = \mathbf{x}^{\mathrm{ok}}$ in (2.1) can be expressed as:

$$d = \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{x}^{\mathrm{ok}} - \hat{\mathbf{x}}^{\mathrm{ok}}\right\|^2\right]. \tag{5.14}$$

The three compression schemes handle the trade-off between compression and distortion in different ways. Since our model operates with continuous quantities, which can lead to potentially infinite rates, each compressor must be paired with a quantization stage to ensure a finite rate. Specifically, considering $n = 32$, we encode each component of $\hat{\mathbf{x}}$ using $16$ bits, limiting the maximum rate to $16n = 512$ bits per time step. The quantization is assumed to be fine enough to preserve the Gaussian characteristics of $\hat{\mathbf{x}}$. Hence, the mutual information between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is evaluated as if they are jointly Gaussian, with their covariance matrix estimated by Monte Carlo simulation [11].

This estimation yields the rate-distortion curves shown in Figure 5.3, where the x axis represents the normalized distortion, within the range $d \in [0, 0.5]$, as higher distortions typically extend beyond the operative ranges. As expected, the RD reaches the lowest rates, confirming its role as a theoretical lower bound. Among the practical compressors, PCA results in the highest rates, while AE, being a non-linear version of PCA, more closely approaches the theoretical limit set by RD. Note that, only the results in Figure 5.3 rely on the quantization stage, while the remainder of our analysis considers continuous sources.

From a signal perspective, both the RD and PCA preserve the Gaussian distribution of the input, meaning that if the input is Gaussian, the compressed output remains Gaussian

[2] To mitigate performance degradation in AE, the autoencoder is initially trained with $k = n - 1$. The latent node with the least average energy is then removed to reduce the network to a latent space of dimension $k - 1$, and the network is re-trained with the weights learned previously as initialization. This process is repeated, decreasing $k$ and progressively increasing the distortion. Training is performed using the ADAM optimizer [75] with a batch size of $128$ and an initial learning rate of $0.01$.

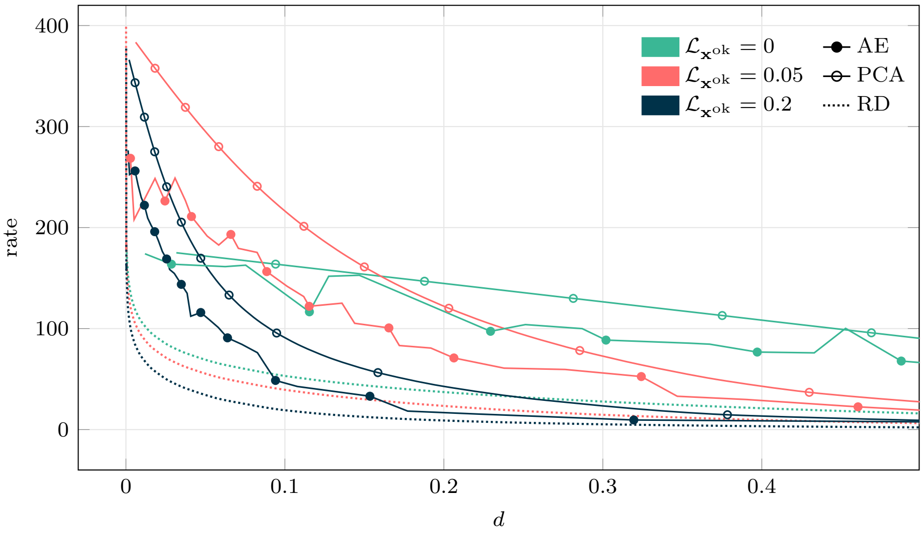


FIGURE 5.3: Rate distortion curves for the three compression techniques we consider and for different values of the localization of the normal signal.

as well. In contrast, AE can alter the statistical distribution of a Gaussian source. We refer to the output of a compressor as a Gaussian Compressed signal (GC) if the signal maintains its Gaussian properties, and as a Non-Gaussian Compressed signal (NGC) if it does not.

Lastly, it is important to note that the result of Theorem 5.1 is strictly guaranteed for RD when applied to Gaussian sources. Nevertheless, we will show numerical evidence that supports the validity of Theorem 5.1 for PCA and AE as well, demonstrating that these compressors, which adapt the encoder-decoder pair to the statistical properties of the normal signal, exhibit similar behavior.

**Detectors** The compressed signal is then processed by a detector, which assigns a score to each instance, with higher scores indicating a greater likelihood for the instance to be anomalous. A final binary decision is made by comparing the score to a predefined threshold. We first focus on two detectors that do not rely on information about anomalies (anomaly-agnostic):

- Likelihood detector (LD): This detector computes the same score used for the distinguishability measure $\mathscr{Z}$, assigning to each instance $\mathbf{x}$ the score $z(\hat{\mathbf{x}}) = \ell(\hat{\mathbf{x}}) = -\log f^{\text{ok}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$.
- One-Class Support-Vector Machine (OCSVM) [130]: This detector uses a Gaussian kernel and is trained on normal signal instances contaminated with $1\%$ of unlabeled white anomalies to help the algorithm in defining the boundary of normal instances[3].

We also consider two detectors that leverage the knowledge of the anomalies (anomaly-aware):

- Neyman-Pearson detector (NPD): This detector uses the same score as the distinguishability measure $\mathscr{J}$, assigning to each instance $\mathbf{x}$ the quantity $z(\hat{\mathbf{x}}) = r(\hat{\mathbf{x}}) = \log f^{\text{ko}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) - \log f^{\text{ok}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$.

[3] The signal components are normalized by their variance, and the kernel scale is set to $1/k_\theta$.

TABLE 5.1: Models we consider for input signals, compressed signals, and detector families.

| TAG | Description | Setting |
|---|---|---|
| | Input signal model | |
| GS | stationary Gaussian signals | $\mathcal{L}^{\mathrm{ok}}$ |
| NGS | realistic, non-stationary, and non-Gaussian signals | ECG, ACC |
| | Compressed signal model | |
| GC | GS compressed preserving the Gaussianity | $\mathcal{L}^{\mathrm{ok}}$ with RD or PCA |
| NGC | non-Gaussian compressed signal | $\mathcal{L}^{\mathrm{ok}}$ with AE; compressed ECG and ACC |
| | Detector family | |
| DbD | detector exploiting the PDF of the compressed signal | LD and NPD |
| LbD | data-driven detectors | OCSVM and DNNC |

- Deep neural network classifier (DNNC): This detector is a neural network classifier with three fully connected hidden layers (with $k$, $2n$, and $n$ neurons, respectively), ReLU activations, and a final sigmoid output neuron generating the score. The network is trained[4] with a binary cross-entropy loss on a dataset of labeled normal and anomalous instances.

As explained in Chapter 1, the LD and NPD detectors depend on the statistical characterization of the signals, which means that they are only suitable to be employed with Gaussian signals compressed by either RD or PCA. Conversely, OCSVM and DNNC are data-driven detectors that can be applied to non-Gaussian compressed signals, such as non-Gaussian sources processed by any compressor or Gaussian sources compressed with AE. We classify the first group as distribution-based detectors (DbD) and the latter as learning-based detectors (LbD). Table 5.1 summarizes the input signal models, compressed signal models, and the corresponding detector families, including the tags used in the upcoming figures.

LbD detectors do not require any assumptions about the signal distribution, but rely on a training dataset, which in this case consists of $10^5$ instances of normal signals. For OCSVM, this dataset is contaminated with $1\%$ of white noise instances, while DNNC training requires additional anomalous examples. To this end, the training set is augmented with $10^5$ anomalous samples, where $50$ different covariance matrices $\mathbf{\Sigma}^{\mathrm{ko}}$ are generated and a different model with the same architecture is trained for each one.

We evaluate the performance of each detector with probability of detection $P_D$ introduced in Chapter 3:

$$P_D = \begin{cases} \mathrm{AUC} & \text{if } \mathrm{AUC} \geq 0.5 \\ 1 - \mathrm{AUC} & \text{if } \mathrm{AUC} < 0.5. \end{cases} \tag{5.15}$$

[4] Training is done via backpropagation with the ADAM optimizer [75], using a batch size of $20$, an initial learning rate of $0.01$, which is scaled by $0.2$ whenever the validation loss reaches a plateau for $5$ epochs. The validation set contains $10\%$ of the instances devoted to training. This setup is the result of a tuning.

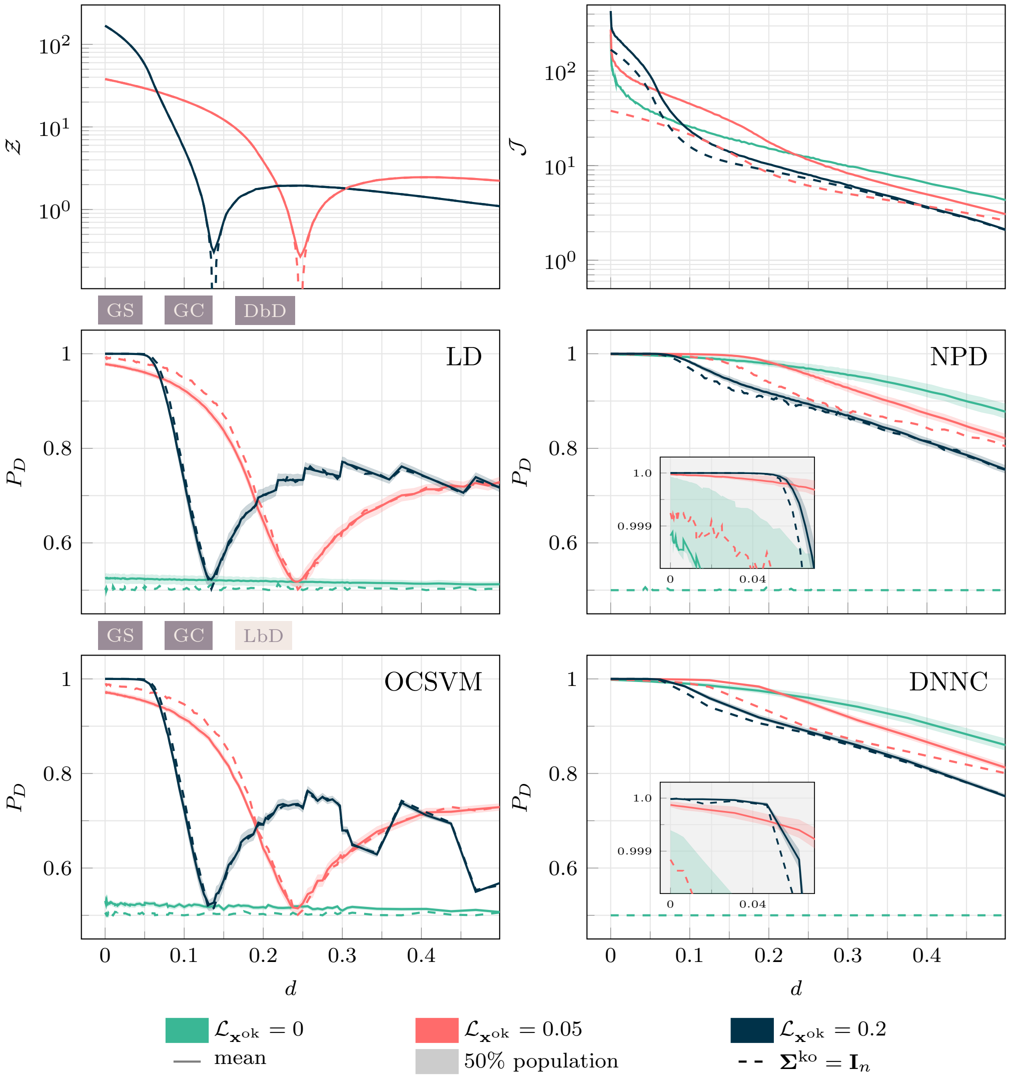


FIGURE 5.4: Distinguishability measures $\mathcal{Z}$, $\mathcal{J}$ and $P_D$ against normalized distortion $d$ in case of RD. The zoomed areas in the NPD and DNNC plots emphasize performance in the low-distortion region.

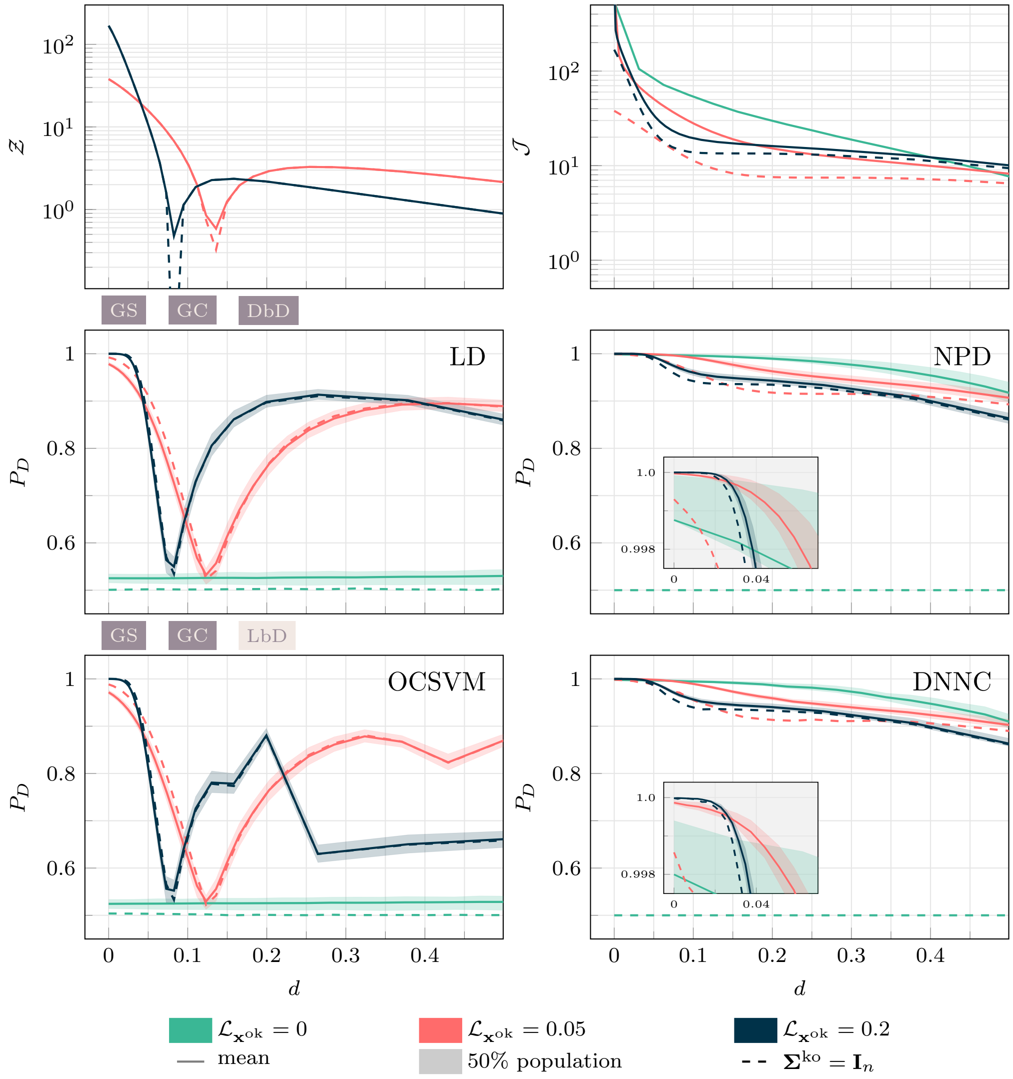


FIGURE 5.5: Distinguishability measures $\mathcal{Z}$, $\mathcal{J}$ and $P_D$ against normalized distortion $d$ in case of PCA. The zoomed areas in the NPD and DNNC plots emphasize performance in the low-distortion region.

In the forthcoming analysis, we present the trends of $P_D$ alongside those of $\mathcal{Z}$ and $\mathcal{J}$ to highlight how theoretical metrics translate to practical detection scenarios. As anticipated in Chapter 4, the comparison is somewhat qualitative — since $\mathcal{Z}$ and $\mathcal{J}$ measure distinguishability as the difference in average scores between normal and anomalous cases expressed in bits per symbol, while $P_D$ considers the entire distribution of these scores and represents the probability of correct detection — this approach highlights the relationship between theoretical and empirical performance.

### 5.3.2 Results

Starting from the established setup, we now match the theoretical derivations with the quantitative evaluation of anomaly detectors' performance in both anomaly-agnostic and anomaly-aware contexts. These evaluations are applied to signals compressed using the RD, PCA, and AE methods. Then we provide an analysis for the specific cases of ECG and ACC signals.

In the upcoming plots, $P_D$ is computed from $1,000$ examples of both normal and anomalous signals. The anomalies comprise white noise as well as $1,000$ different distributions, each defined by a randomly generated $\mathbf{\Sigma}^{\mathrm{ko}}$. For the DNNC detector, we restrict the analysis to $50$ anomalies due to the need to retrain the network for each anomaly.

**RD** Figure 5.4 illustrates the results organized in two rows with three plots each. The left-hand side shows detectors that have no information on the anomaly, while the right-hand side shows detectors that can exploit such information. Different colors correspond to various levels of $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}}$, dashed lines represent the average anomaly (white noise), and shaded areas contain the spread of $50\%$ of the Monte Carlo population. The plots on the top depict the $\mathcal{Z}$ and $\mathcal{J}$ profiles, which should be compared to the detector performances shown on the two lower rows. As expected, no $\mathcal{Z}$ profile appears for $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}} = 0$, since $\zeta = 0$ in this case (as mentioned in Section 5.2.1). Additionally, tags corresponding to the settings in Table 5.1 are added above each row of plots.

The numerical evidence confirms the theoretical results: $\zeta_{\mathbf{I}} = \mathbf{E}[\zeta]$, $\mathcal{Z}_{\mathbf{I}} \leq \mathbf{E}[\mathcal{Z}]$ and $\mathcal{J}_{\mathbf{I}} \leq \mathbf{E}[\mathcal{J}]$, as discussed in Section 4.2.2. This supports the idea that white noise can be adopted as a reference anomaly, allowing to compute average and lower bound behavior for anomaly-agnostic detectors or a lower bound in anomaly-aware scenarios. As demonstrated in Theorem 4.1, with increasing $n$, the white noise anomaly becomes the dominant anomaly, to which tends any possible anomaly.

In the anomaly-agnostic scenario (left-hand side), the theory anticipates that detection performance and distortion exhibit a non-monotonic relationship. There exists a point at which distortion nullifies distinguishability, causing detectors to fail. This critical distortion level corresponds to the point where $\mathcal{Z}$ crosses zero, and the relationship between normal and anomalous scores inverts. Both the LD and OCSVM detectors exhibit this behavior. The critical distortion point is also influenced by $\mathcal{L}_{\mathbf{x}^{\mathrm{ok}}}$, as predicted by Theorem 5.1. In general, abstract measures $\mathcal{Z}$ and $\mathcal{J}$ anticipate that in low-distortion regions, more localized signals are easier to distinguish from anomalies, but at the same time detector failure occurs at lower distortions with respect to less localized signals.

In the case of anomaly-aware detectors (right-hand side), complete failure only happens at the highest levels of distortion, as predicted by the distinguishability measure $\mathcal{J}$. By comparing the trend of $\mathcal{J}$ with the zoomed areas in the NPD and DNNC plots, we show that, in the low-distortion region, more localized signals are easier to distinguish from anomalies, although they also lead to more pronounced performance degradation as distortion increases.

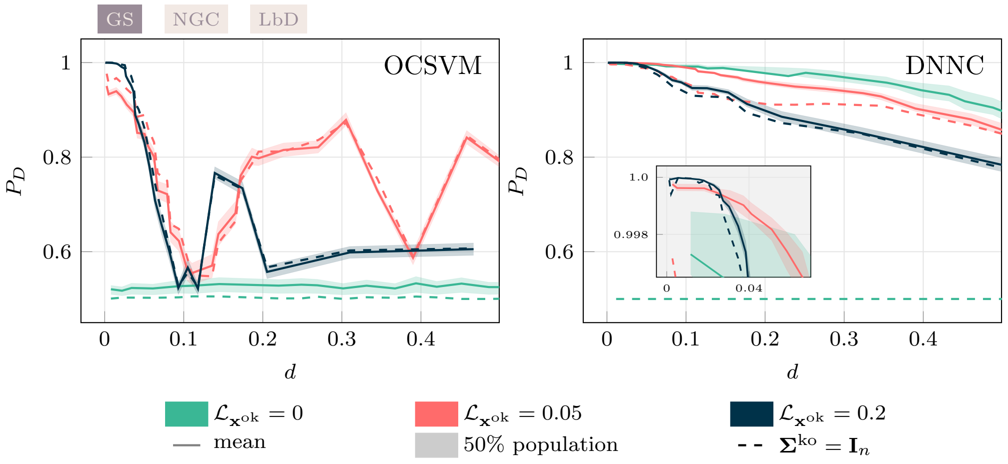


FIGURE 5.6: Distinguishability measure $P_D$ versus normalized distortion $d$ in case of AE. The zoomed area in the DNNC plot emphasizes performance in the low-distortion region.

**PCA** In terms of the rate-distortion trade-off, PCA is highly suboptimal. Yet, because it is linear, both $\mathbf{x}$ and $\hat{\mathbf{x}}$ remain jointly Gaussian, allowing us to compute the theoretical values of $\mathscr{Z}$ and $\mathscr{J}$ using (5.8) and (5.12).

Figure 5.5 summarizes the results for this case, representing plots similar to those in Figure 5.4. The qualitative behaviors discussed in the previous subsection are also present here, and they align with the trends predicted by the theoretical measures.

The distortion levels at which anomaly-agnostic detectors fail are different from those seen with RD, yet these points are still well-predicted by the theoretical profiles and Theorem 5.1.

In this case, beyond the breakdown distortion level, the values of $\mathscr{Z}$ increase slightly more than in the optimal compression scenario. This suggests that adopting a suboptimal compression method in terms of rate-distortion may enhance the distinguishability between compressed normal and anomalous signals. This consideration is also confirmed in practice, as shown by the improved performance of the LD and OCSVM detectors in the left column of Figure 5.5.

**AE** In this scenario, the compression process is non-linear, meaning that $\mathbf{x}$ and $\hat{\mathbf{x}}$ may no longer be jointly Gaussian. As a result, it is not possible to compute the theoretical measures $\mathscr{Z}$ and $\mathscr{J}$, nor can we apply the LD and NPD detectors, which rely on the knowledge of the signals. For this reason, Figure 5.6 only shows the performance of the OCSVM and DNNC detectors.

Despite this limitation, it is worth noting that the qualitative behavior of these detectors still roughly aligns, even if with more approximation, with the trends predicted by the theoretical curves plotted for PCA.

**Distinguishability in real applications** To further confirm our theoretical results, we also examine two practical applications involving non-stationary and non-Gaussian signals: ECG and ACC.

Both ECG and ACC signals are compressed via PCA, while OCSVM and DNNC are employed to distinguish between normal signals and instances of white anomalies $\mathbf{x}^{\text{ko}}$. Given that the input $\mathbf{x}$ is NGS, the compressed signal $\hat{\mathbf{x}}$ is also considered as NGC, regardless of the compression method used.

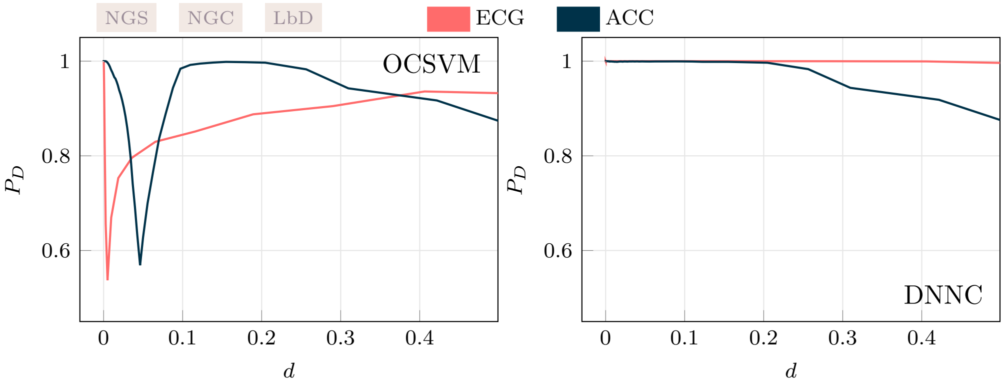


FIGURE 5.7: Distinguishability measure $P_D$ versus normalized distortion $d$ in case of PCA for ECG and ACC signals with windows of $n = 64$ samples.

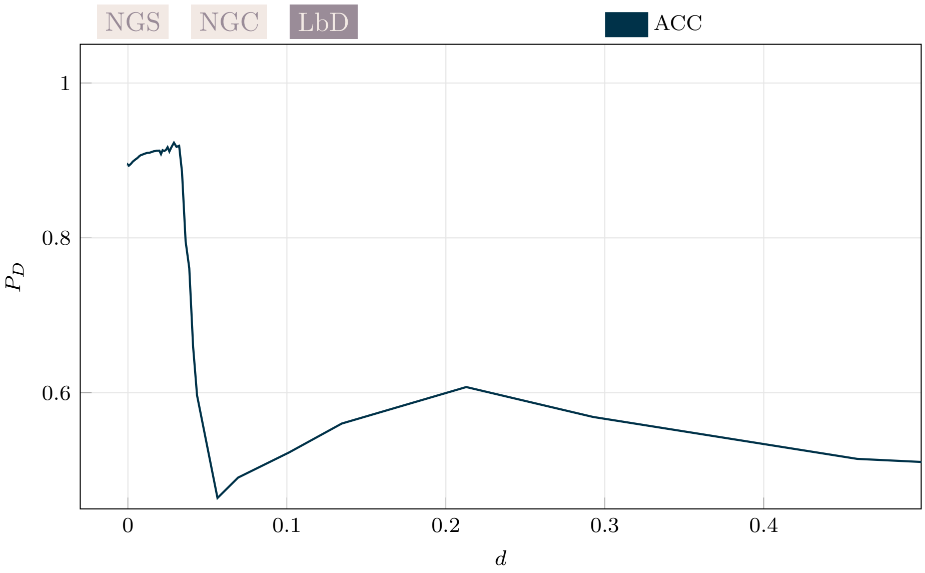


FIGURE 5.8: Distinguishability measure $P_D$ versus normalized distortion $d$ for a real anomaly in accelerometers (ACC) signals with windows of $n = 64$ samples.

The results in Figure 5.7 resemble the patterns observed in previous experiments. In the anomaly-agnostic scenario, OCSVM shows a critical but not disruptive distortion point at which detecting white anomalies becomes impossible. As with prior settings, different types of signals exhibit varying critical distortion values. In the anomaly-aware case, the performance of DNNC is consistently monotonic with respect to distortion $d$. Specifically, for ECG, $P_D$ remains close to $1$ across all $d$ values, while for ACC, detection performance gradually decreases as $d$ increases.

As a final case study, we examine a real-world anomaly altering the ACC signal. Specifically, this anomaly is a subtle, non-disruptive failure in the monitored civil structure, characterized by a slight shift in its modal frequencies [101]. With such characteristics, the anomaly cannot be described by a white noise distribution. The results, in terms of $P_D$ for the LD detector, are shown in Figure 5.8. The trend reported here matches with the presence of a critical, non-disruptive distortion value, consistent with the distortion level previously identified in Figure 5.7, where anomalies were emulated as white noise instances.

### 5.3.3 Discussion

To validate the distinguishability measures and the theoretical results based on them, we analyzed several numerical settings. First, we examined a case that matches the theoretical framework, where synthetic Gaussian signals are compressed using an optimal encoder in the rate-distortion sense. The results, illustrated in Figure 5.4, demonstrate how the two metrics can predict the performance of detectors that either rely on the signal's distribution or learn from data. Furthermore, in the anomaly-agnostic scenario, the results confirm the presence of a critical distortion level that makes anomalies undetectable, consistent with the prediction of Theorem 5.1.

Next, we incrementally relaxed the assumptions used to derive the theoretical results. In Figure 5.5, the optimal encoder was replaced with PCA, a more realistic encoder that still maintains Gaussianity in the compressed domain. In Figure 5.6, we considered the non-linear AE encoder, which alters the statistical properties of the compressed signals. In Figure 5.7, we substituted Gaussian signals with ECG and ACC, real-world non-Gaussian signals. Finally, Figure 5.8 considered a real anomaly that also exhibited non-Gaussian characteristics.

The numerical results provide strong evidence that the derived results offer valuable insights for practical applications. However, it is important to stress that this analysis specifically focused on cases where the encoder was adapted to the statistical properties of the normal signal. Thus, these results cannot be directly applied to other compression techniques that differ significantly from the conditions of Lemma 5.1 and Lemma 5.2.

## 5.4 Conclusion

Massive sensing systems often rely on lossy compression to reduce the bitrate of data transmission while preserving the fidelity of the signals. As these compressed sensor readings travel to centralized servers, they may be processed along the way for early anomaly detection. Therefore, this detection is performed on compressed data.

In a framework where both normal and anomalous signals are modeled as Gaussian sources, we extend the classical rate-distortion theory to describe the distributions of the distorted signals and their corresponding mappings. This allows us to specialize $\mathscr{Z}$ and $\mathscr{J}$ measures to quantify the distinguishability between normal and anomalous compressed signals, applicable in both anomaly-agnostic and anomaly-aware scenarios.

When the analysis considers Gaussian sources, it reveals that

- The distinguishability metrics anticipate the detection performance measured in terms of probability of detection.
- The distinguishability metrics for white anomalies are representative of a detector's average performance across various types of anomalies.
- For anomaly-agnostic detectors, the relationship between distinguishability and distortion may not be monotonic, and it can feature at least one non-disruptive distortion level where white anomalies become undetectable.

We also provide numerical evidence supporting the effectiveness of our theoretical framework in detecting Gaussian signals, using both rate-distortion-optimal compression schemes and ideal detectors.

In addition, we broaden our analysis by relaxing some of the initial assumptions and offering empirical evidence that demonstrates the robustness of the results in real-world scenarios. Specifically, the predictive capability extends to generic data-driven detectors, even when they do not process Gaussian signals and when compression alters the distribution of the original input signal. Moreover, the information-theoretic metrics remain effective when applied to realistic non-Gaussian signals.

# Chapter 6

# A Theoretical Framework for Rate-Distortion-Distinguishability

A typical scenario in large-scale acquisition systems is characterized by numerous sensing units that convert unknown physical phenomena into data samples, which are then transmitted over a network. To reduce the volume of transmitted data and retain only key information, these sensor readings are often compressed using lossy techniques [94, 87]. As shown in Chapter 2, this introduces a trade-off between the bit rate of transmission and the amount of lost information, quantified by distortion.

Compressed data are often sent to a remote location, such as a cloud facility, for further processing and storage. Before reaching these facilities, the compressed data may pass through the intermediate edge devices [136].

Whether performed on the edge of the cloud or the cloud itself, an important task in this context is the detection of anomalies, which implies determining whether a signal originates from a normal or anomalous source [24, 82, 33, 93, 32]. If the statistical characteristics of the anomalies are unknown, this task becomes unsupervised; otherwise, it can be treated as a supervised problem.

Nevertheless, compression can negatively affect the performance of anomaly detectors. Lossy compression is most effective when it discards information that is irrelevant to reconstructing typical signals, i.e., minimizing distortion. Yet, the discarded information might be essential to detect anomalies. Thus, designing a system that considers both compression and detection requires addressing a three-fold trade-off: *rate-distortion-distinguishability* (RDD).

In Chapter 5 we explored how compression schemes that balance rate and distortion influence the ability to distinguish signals, proving that the rate-distortion and distortion-distinguishability trade-offs are inherently different. This chapter extends this analysis to include the full RDD trade-off. We examine the relationship between compression, information loss, and the capacity to distinguish between normal and anomalous compressed signals in both supervised and unsupervised contexts. This extended analysis yields a Pareto surface that generalizes the rate-distortion curve [34] and optimally characterizes the rate-distortion-distinguishability interaction.

In existing literature, the joint trade-off among rate, distortion, and distinguishability has only been considered in the context of classification and other supervised tasks, and not yet for anomaly detection (AD).

Focusing on classification, the authors of [42] introduce a rate-distortion framework to assess the performance of various compression schemes combined with algorithms designed to estimate target orientation. Similarly, [163] presents a theoretical framework that investigates the interplay between rate, distortion, and classification accuracy, offering practical insights into compression techniques designed for the classification of compressed images. This trade-off is further explored with practical compression methods. Several studies [111, 14, 55, 84] focus on vector quantization and adapt standard vector quantizer design algorithms by integrating a classification-related term into the distortion metric to improve classification performance. In [95], the authors fine-tune the JPEG quantization tables to enhance either classification accuracy or image reconstruction quality, given a rate constraint. Meanwhile, [27] and [12] propose neural network-based frameworks that optimize compression, classification, and optionally signal decoding, all in a unified process.

The works discussed in [95, 27] fall under a broader paradigm that targets communication for both human and machine vision. These studies address the three-fold trade-off between rate, distortion, and a third metric reflecting performance in image or video tasks. For instance, [147, 90, 88] focus on tasks such as classification, semantic segmentation, object detection, foreground extraction, and depth estimation in images. On the other hand, works such as [43, 135, 159] consider video processing tasks, comprising action recognition, denoising, super-resolution, scene classification, semantic segmentation, and object classification.

It is also worth mentioning the studies [19, 18] that examine how image perception quality is affected by distortion and investigate the trade-off between rate, distortion, and a perception measure.

With respect to these related works, the analysis proposed in this chapter focuses on evaluating the trade-off between rate, distortion, and the performance of a detector in distinguishing between normal and anomalous signals in both supervised and unsupervised settings. Unlike typical binary classification, AD assumes that positive class occurrences (anomalies) are rare, causing the compression mechanism to be adapted primarily on the negative class (normal signals).

Specifically, in Section 6.1 we expand the traditional rate-distortion framework by introducing anomaly-aware and anomaly-agnostic distinguishability metrics into the optimization process, formulating two RDD optimization problems. To derive theoretical results, in Section 6.2, we adapt the framework assuming that both normal and anomalous sources, as well as their compressed counterparts, obey Gaussian distributions. This assumption allows us to solve the two RDD optimization problems numerically with convex optimization tools. Finally, Section 6.3 presents the numerical results for the Gaussian source and includes two Pareto surfaces, one for anomaly-aware and one for anomaly-agnostic scenarios. In this theoretical setting, we also highlight how the signal components are distorted when the compression mechanism is optimized according to the rate-distortion-distinguishability trade-off. We then validate the trends observed in the theoretical outcomes by applying them to real-world compression schemes and signals. Specifically, we examine the following approaches: *(i)* a compression method based on the Karhunen-Loève Transform (KLT) applied to Gaussian signals; *(ii)* JPEG compression, where quantization tables are adapted to enhance AD; *(iii)* a neural network-based lossy compression that incorporates a parameter for improving AD performance. In each scenario, we demonstrate how the performance of AD is influenced by both the rate and the distortion.

## 6.1 A joint look at rate, distortion, and distinghuishability

In this section, we first recall the concepts behind rate-distortion problem and distinguishability measures, introduced in Chapter 2 and Chapter 5, respectively. Then we blend these concepts by defining two rate-distortion-distinguishability problems.

Let $\mathbf{x}$ represent a generic stationary discrete-time source generating a random vector $\mathbf{x} \in \mathbb{R}^n$ at each time step $t$. This source typically produces typical signal instances but rarely generates anomalies. To account for normal and anomalous signals, we consider two independent sources $\mathbf{x}^{\mathrm{ok}}$ and $\mathbf{x}^{\mathrm{ko}}$, each following distinct probability density functions (PDFs), $f_{\mathbf{x}}^{\mathrm{ok}}$ and $f_{\mathbf{x}}^{\mathrm{ko}}$, respectively.

### 6.1.1 Rate and distortion

In the considered scenario, the observable signal $\mathbf{x}$ is compressed in a lossy manner into symbols suitable for transmission over a channel with a rate constraint. Specifically, the vector $\mathbf{x} \in \mathbb{R}^n$ is encoded into a symbol $\mathbf{z}$ with reduced information, then sent through the channel and later decoded into an approximation $\hat{\mathbf{x}}$. As a result, the receiver works with the distorted versions $\hat{\mathbf{x}}$ of $\mathbf{x}$. The design of the encoder that processes $\mathbf{x}$ is guided by the rate-distortion theory defined in [34, Chapter 13] and summarized in Chapter 2.

Since anomalies are rare, encoders are generally adapted to the normal signal, assuming $\mathbf{x} = \mathbf{x}^{\mathrm{ok}}$. Distortion defined in (2.1) in this context can be expressed as:

$$D = \mathbf{E}\left[\left\|\mathbf{x}^{\mathrm{ok}} - \hat{\mathbf{x}}^{\mathrm{ok}}\right\|^2\right], \tag{6.1}$$

with $\mathbf{E}[\cdot]$ the expected value.

The level of compression is measured by the transmission rate of $\mathbf{z}$, which maps into $\hat{\mathbf{x}}$. Given that the mapping from $\mathbf{z}$ to $\hat{\mathbf{x}}$ is injective and the encoder assumes $\mathbf{x} = \mathbf{x}^{\mathrm{ok}}$, this rate can be expressed by the mutual information $\mathcal{I}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ok}}\right)$ between $\mathbf{x}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ok}}$ [34, Chapter 8].

Using these definitions, the minimum achievable transmission rate $\rho$ and the maximum allowable distortion $\delta$ are driven by the relationship:

$$\text{(RD)} \quad \rho(\delta) = \begin{array}{l} \inf\limits_{f_{\hat{\mathbf{x}}|\mathbf{x}}} \mathcal{I}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ok}}\right) \\ \text{s.t.} \quad D \leq \delta \end{array} \tag{6.2}$$

as described in [34, Theorem 13.2.1]. Here, $f_{\hat{\mathbf{x}}|\mathbf{x}}$ represents the conditional PDF modeling the potentially stochastic mapping between the encoder and decoder, so that the PDF of $\hat{\mathbf{x}}$ is given by:

$$f_{\hat{\mathbf{x}}}(\alpha) = \int_{\mathbb{R}^n} f_{\hat{\mathbf{x}}|\mathbf{x}}(\alpha, \beta) f_{\mathbf{x}}(\alpha)\, d\alpha. \tag{6.3}$$

### 6.1.2 Distinguishability

The encoder $f_{\hat{\mathbf{x}}|\mathbf{x}}$, designed considering $\mathbf{x} = \mathbf{x}^{\mathrm{ok}}$, is also applied to both normal and anomalous signals. Consequently, the retrieved signal $\hat{\mathbf{x}}$ can be either $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{\mathrm{ok}}$ or $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{\mathrm{ko}}$, where these two vectors generally obey PDFs, $f_{\hat{\mathbf{x}}}^{\mathrm{ok}}$ and $f_{\hat{\mathbf{x}}}^{\mathrm{ko}}$, as described by (6.3) with $f_{\mathbf{x}} = f_{\mathbf{x}}^{\mathrm{ok}}$ for normal signals and $f_{\mathbf{x}} = f_{\mathbf{x}}^{\mathrm{ko}}$ for anomalous ones.

As compression increases and the rate decreases, the statistical properties of $\hat{\mathbf{x}}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ko}}$ tend to align. In the limit case of maximum compression, i.e., zero rate, $\hat{\mathbf{x}}$ becomes a constant, independent of the source. This constitutes a second trade-off: the one between compression and the ability to distinguish between $\hat{\mathbf{x}}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ko}}$. This trade-off directly conditions the effectiveness of detectors identifying anomalies given $\hat{\mathbf{x}}$.

Distinguishability is defined in information-theoretic terms in Chapter 4 and has been specialized to compressed sources in Chapter 5 for two different scenarios:

- *Anomaly-agnostic detector*: If the detector only knows the normal source, distinguishability is measured by:

$$\mathcal{Z} = \left|\int_{\mathbb{R}^n} \left[f_{\hat{\mathbf{x}}}^{\mathrm{ok}}(\alpha) - f_{\hat{\mathbf{x}}}^{\mathrm{ko}}(\alpha)\right] \log_2 f_{\hat{\mathbf{x}}}^{\mathrm{ok}}(\alpha)\, d\alpha\right|, \tag{6.4}$$

- *Anomaly-aware detector*: When the detector knows both normal and anomalous signal statistics, distinguishability is measured by:

$$\mathcal{J} = \int_{\mathbb{R}^n} \left[ f_{\hat{\mathbf{x}}}^{\mathrm{ok}}(\alpha) - f_{\hat{\mathbf{x}}}^{\mathrm{ko}}(\alpha) \right] \log_2 \left[ \frac{f_{\hat{\mathbf{x}}}^{\mathrm{ok}}(\alpha)}{f_{\hat{\mathbf{x}}}^{\mathrm{ko}}(\alpha)} \right] d\alpha. \tag{6.5}$$

Distinguishability reflects the detector's ability to recognize the source of the signal. However, in practice, the PDFs of signals are either rarely available or not easily estimable. Therefore, in practice, detector performance is typically measured with the probability of detection $P_D$, defined in Chapter 3 as:

$$P_D = \begin{cases} \mathrm{AUC} & \text{if } \mathrm{AUC} \geq 0.5, \\ 1 - \mathrm{AUC} & \text{if } \mathrm{AUC} < 0.5, \end{cases} \tag{6.6}$$

where $\mathrm{AUC}$ represents the Area Under the Curve of the Receiver Operating Characteristic.

### 6.1.3 Rate-Distortion-Distinguishability

The distinguishability metrics recalled in (6.4) and (6.5) are employed in Chapter 5, after solving (6.2), to connect each rate-distortion pair $(\rho, \delta)$ with a corresponding level of distinguishability. This approach still addresses the classical rate-distortion trade-off. In Chapter 5, we also show that, for Gaussian sources, these information-theoretic measures accurately predict the detector performance measured in terms of $P_D$.

We now extend the analysis to a more general trade-off between rate-distortion and distinguishability. To quantify distinguishability, we rely on the metrics in (6.4) and (6.5), formulating this triple trade-off through the following optimization problems:

(RDD $\mathcal{Z}$)

$$\begin{aligned} \rho(\delta, \omega) = \; & \inf_{f_{\hat{\mathbf{x}}|\mathbf{x}}} \; \mathcal{I}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ok}}\right) \\ & \text{s.t.} \quad D \leq \delta, \\ & \text{s.t.} \quad \mathbf{E}[\mathcal{Z}] \geq \omega. \end{aligned} \tag{6.7}$$

(RDD $\mathcal{J}$)

$$\begin{aligned} \rho(\delta, \omega) = \; & \inf_{f_{\hat{\mathbf{x}}|\mathbf{x}}} \; \mathcal{I}\left(\hat{\mathbf{x}}^{\mathrm{ok}}; \mathbf{x}^{\mathrm{ok}}\right) \\ & \text{s.t.} \quad D \leq \delta, \\ & \text{s.t.} \quad \mathcal{J} \geq \omega, \end{aligned} \tag{6.8}$$

where $\omega$ accounts for the minimum acceptable level of distinguishability. We refer to RDD $\mathcal{Z}$ and RDD $\mathcal{J}$ to differentiate between the anomaly-agnostic and the anomaly-aware scenarios, respectively.

In the (6.7), the distinguishability constraint $\mathbf{E}[\mathcal{Z}]$ represents the average distinguishability across a set of anomalies, rather than $\mathcal{Z}$ itself. This is because directly evaluating $\mathcal{Z}$ requires prior knowledge of $f_{\mathbf{x}}^{\mathrm{ko}}$, which is not available to anomaly-agnostic detectors.

The interplay between $\rho$, $\delta$, and $\omega$ forms a Pareto surface in the three-dimensional space of these parameters. This surface represents the locus of optimal solutions where improving one quantity necessarily makes degrade at least one of the others.

It is important to note that, not only in (6.2) but also in (6.7) and (6.8), the rate and distortion are considered for the normal source $\mathbf{x}^{\mathrm{ok}}$, since the primary objective is AD. AD can be viewed as a one-class classification problem, focusing on identifying deviations from normal behavior, assuming these events (anomalies) to be rare. This scenario is different from traditional binary classification, where both classes are equally relevant, and there would be no reason for tuning rate and distortion based on only one class.

## 6.2 The Gaussian case

The main distinction between problems (6.7) and (6.8) and the rate-distortion problem (6.2) is the additional constraint that allows the encoder to be optimized for larger distinguishability. Solving (6.7) and (6.8) for general sources is fundamentally challenging, certainly more so than solving (6.2). For this reason, we approach this problem by making some assumptions. Specifically, we rely on the Gaussian framework defined in Chapter 4 such that $\mathbf{x}^{\text{ok}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\text{ok}}\right)$ with $\mathbf{\Sigma}^{\text{ok}}$ a diagonal matrix and $\text{tr}(\mathbf{\Sigma}^{\text{ok}}) = n$. For the anomalous signal, we still consider $\mathbf{x}^{\text{ko}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\text{ko}}\right)$, but we relax the assumption on the average energy of the anomalous samples, allowing it to be a generic $\alpha > 0$, i.e., $\text{tr}(\mathbf{\Sigma}^{\text{ko}}) = \alpha n$.

Under these assumptions, we recall some properties derived in Chapter 5. First, the average anomaly is white, that is, $\mathbf{E}[\mathbf{\Sigma}^{\text{ko}}] = \alpha \mathbf{I}_n = \mathbf{W}$. In addition, the average distinguishability values $\mathbf{E}[\mathcal{Z}]$ and $\mathbf{E}[\mathcal{J}]$ are related to those calculated for the white anomaly, denoted as $\mathcal{Z}_{\mathbf{W}}$ and $\mathcal{J}_{\mathbf{W}}$. Specifically, according to Jensen's inequality, $\mathcal{Z}_{\mathbf{W}} \leq \mathbf{E}[\mathcal{Z}]$ and $\mathcal{J}_{\mathbf{W}} \leq \mathbf{E}[\mathcal{J}]$. Therefore, in (6.7), $\mathcal{Z}_{\mathbf{W}}$ can serve as a lower bound for $\mathbf{E}[\mathcal{Z}]$.

For the constraint on $\mathcal{J}$ in (6.8), the problem can be solved within the Gaussian framework by considering anomalies where $\mathbf{\Sigma}^{\text{ko}} = \text{diag}\left(\boldsymbol{\lambda}^{\text{ko}}\right)$, with $\boldsymbol{\lambda}^{\text{ko}} = \left(\lambda_0^{\text{ko}}, \dots, \lambda_{n-1}^{\text{ko}}\right)^\top$.

Although our approach will be generalized later, we initially assume that the encoding is *Gaussian-additive*. That is, $f_{\hat{\mathbf{x}}|\mathbf{x}}$ is such that $\mathbf{\Delta} = \hat{\mathbf{x}} - \mathbf{x}$ is a random vector that contains zero-mean components that are mutually independent but jointly Gaussian with those of $\mathbf{x}$ (and therefore $\hat{\mathbf{x}}$). The covariance matrices for the triads $\hat{x}_j^{\text{ok}}, x_j^{\text{ok}}, \Delta_j$ are given by

$$\mathbf{\Sigma}_{\left(\hat{x}_j^{\text{ok}}, x_j^{\text{ok}}, \Delta_j\right)} = \begin{pmatrix} \lambda_j^{\text{ok}} - \theta_j & \lambda_j^{\text{ok}} - \theta_j & 0 \\ \lambda_j^{\text{ok}} - \theta_j & \lambda_j^{\text{ok}} & -\theta_j \\ 0 & -\theta_j & \theta_j \end{pmatrix} \tag{6.9}$$

for some variances $0 \leq \theta_j \leq \lambda_j^{\text{ok}}$ for $j = 0, \dots, n-1$, which define the degrees of freedom for the encoding operation.

For Gaussian-additive encodings, we can state two lemmas whose proofs can be found in Appendix D.

**Lemma 6.1.** *If signal $\mathbf{x}^{\text{ok}}$ consists of independent Gaussian components with zero mean and variances $\lambda_j^{\text{ok}}$. When a Gaussian-additive encoding is applied with variances $\theta_j$ for $j = 0, \dots, n-1$, the resulting rate is given by*

$$\mathcal{I}\left(\hat{\mathbf{x}}^{\text{ok}}; \mathbf{x}^{\text{ok}}\right) = \frac{1}{2} \sum_{j=0}^{n-1} \log_2 \frac{\lambda_j^{\text{ok}}}{\theta_j} \tag{6.10}$$

*which represents the minimum achievable rate for any encoding that satisfies the condition $\mathbf{E}\left[\left(\hat{x}_j^{\text{ok}} - x_j^{\text{ok}}\right)^2\right] = \theta_j$ for each $j = 0, \dots, n-1$.*

**Lemma 6.2.** *For a Gaussian-additive encoder defined by parameters $\theta_j$ for $j = 0, \ldots, n-1$, the following hold:*

$$D = n - \sum_{j=0}^{n-1} \lambda_j^{\mathrm{ok}} \xi_j, \tag{6.11}$$

$$\mathcal{Z} = \frac{1}{2 \ln 2} \left| \sum_{j=0}^{n-1} r_j \xi_j \right|, \tag{6.12}$$

$$\mathcal{J} = \frac{1}{2 \ln 2} \sum_{j=0}^{n-1} \frac{r_j^2 \xi_j^2}{1 - r_j \xi_j}, \tag{6.13}$$

*with $r_j = 1 - \frac{\lambda_j^{\mathrm{ko}}}{\lambda_j^{\mathrm{ok}}}$ and $\theta_j = \lambda_j^{\mathrm{ok}}(1 - \xi_j)$, where $\xi_j \in [0, 1]$ represent normalized degrees of freedom for the encoder.*
*Both $\mathcal{Z}$ and $\mathcal{J}$ are convex functions of the $\xi_j$.*

Now, consider the case where both a distortion and a distinguishability constraint are imposed. According to Lemma 6.1, starting with a feasible encoding $f'_{\hat{\mathbf{x}}|\mathbf{x}}$, one can compute $\theta_j = \mathbf{E}\left[\left(\hat{x}_j^{\mathrm{ok}} - x_j^{\mathrm{ok}}\right)^2\right]$ and then define a Gaussian-additive encoder $f''_{\hat{\mathbf{x}}|\mathbf{x}}$ with the corresponding parameters. If $f''_{\hat{\mathbf{x}}|\mathbf{x}}$ satisfies both constraints, it provides a rate that is not higher than that of $f'_{\hat{\mathbf{x}}|\mathbf{x}}$.

However, since $D = \sum_{j=0}^{n-1} \theta_j$ is independent of the specific encoder, the only constraint that needs re-evaluation when transitioning from $f'_{\hat{\mathbf{x}}|\mathbf{x}}$ to $f''_{\hat{\mathbf{x}}|\mathbf{x}}$ is the distinguishability one.

To formalize this consideration, we rely on (6.12), (6.13), and (6.11) to define the following two subsets of $[0, n] \times \mathbb{R}^+$:

$$H_{\mathcal{Z}} = \{(\delta, \omega) \mid \exists \xi_0, \ldots, \xi_{n-1} \text{ s.t. } (6.11) = \delta \text{ and } (6.12) = \omega\}, \tag{6.14}$$

$$H_{\mathcal{J}} = \{(\delta, \omega) \mid \exists \xi_0, \ldots, \xi_{n-1} \text{ s.t. } (6.11) = \delta \text{ and } (6.13) = \omega\}. \tag{6.15}$$

Thanks to its special characteristics, when $(\delta, \omega) \in H_{\mathcal{Z}|\mathcal{J}}$, the rate-distortion-distinguishability (RDD) problems in (6.7) and (6.8) are effectively solved by using a Gaussian-additive encoder. Outside of these regions, however, this property cannot be derived directly and must instead be considered as an additional assumption, along with the Gaussian nature of both normal and anomalous sources.

Regardless of the generality of the following derivations, the RDD trade-off under investigation relies on solving (6.7) and (6.8), expressed in terms of the normalized degrees

TABLE 6.1: Assumptions on the signals and the encoder on which the problem in (6.16)-(6.20) relies.

| Signals<br>$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | | $\boldsymbol{\mu}$ | $\boldsymbol{\Sigma}$ | $\mathrm{tr}(\boldsymbol{\Sigma})$ |
|---|---|---|---|---|
| ok | | $\mathbf{0}$ | $\mathrm{diag}\left(\boldsymbol{\lambda}^{\mathrm{ok}}\right)$ | $n$ |
| ko | RDD $\mathcal{Z}$ | $\mathbf{0}$ | $\boldsymbol{\Sigma}^{\mathrm{ko}}$ | $\alpha n$ |
| | RDD $\mathcal{J}$ | $\mathbf{0}$ | $\mathrm{diag}\left(\boldsymbol{\lambda}^{\mathrm{ko}}\right)$ | $\alpha n$ |
| Encoder | Gaussian-additive if $(\delta, \omega) \notin H_{\mathcal{Z}\mid\mathcal{J}}$ | | | |

of freedom $\xi_0, \ldots, \xi_{n-1}$:

$$
\begin{aligned}
\rho = \min_{\xi_0,\ldots,\xi_{n-1}} \quad & -\frac{1}{2\ln 2}\sum_{j=0}^{n-1} \ln(1-\xi_j) && (6.16)\\
\text{s.t.} \quad & 0 \le \xi_j \le 1, && (6.17)\\
\text{s.t.} \quad & \sum_{j=0}^{n-1} \lambda_j^{\mathrm{ok}} \xi_j \ge n - \delta, && (6.18)\\
\text{either} \quad \text{s.t.} \quad & \left| \sum_{j=0}^{n-1} r_j^{\mathbf{W}} \xi_j \right| \ge 2\omega \ln 2, && (6.19)\\
\text{or} \quad \text{s.t.} \quad & \sum_{j=0}^{n-1} \frac{(r_j \xi_j)^2}{1 - r_j \xi_j} \ge 2\omega \ln 2, && (6.20)
\end{aligned}
$$

where only one of the constraints (6.19) or (6.20) should be considered at a time. In the case of (6.19), $r_j = 1 - \frac{\alpha}{\lambda_j^{\mathrm{ok}}}$. Moreover, it is important to note that the derivation of this constraint makes use of the inequality $\mathcal{Z}_{\mathbf{W}} \le \mathbf{E}[\mathcal{Z}]$.

Table 6.1 provides a summary of the assumptions under which the RDD problem defined by (6.16)–(6.20) holds. The assumption about the normal signal does not reduce generality. In the anomaly-agnostic scenario, we assume knowledge only of the scaling factor $\alpha$, which sets the average energy of the anomaly with respect to the normal signal. In the anomaly-aware case, anomalies are modeled with a diagonal covariance matrix.

### 6.2.1 Solution to the optimization problems

We can express the constraint (6.19) as

$$
\sum_{j=0}^{n-1} r_j^{\mathbf{W}} \xi_j \ge 2\omega \ln 2 \quad \cup \quad \sum_{j=0}^{n-1} r_j^{\mathbf{W}} \xi_j \le -2\omega \ln 2. \tag{6.21}
$$

This disjunctive form allows us to find the minimum rate solution by solving two separate maximization problems. In each case, the objective function is concave, the constraints are linear, and the feasible space is convex.

As already shown, the left-hand side of (6.20) is a convex function, which must be paired with a $\le$ relation to meet convex programming rules. However, in our case, the presence of the $\ge$ relation turns this into a reverse convex optimization problem, as described in [61]. We address this issue heuristically, relying on the algorithm proposed in [149, Algorithm 5.1], which reduces the problem to a series of sub-problems, each of

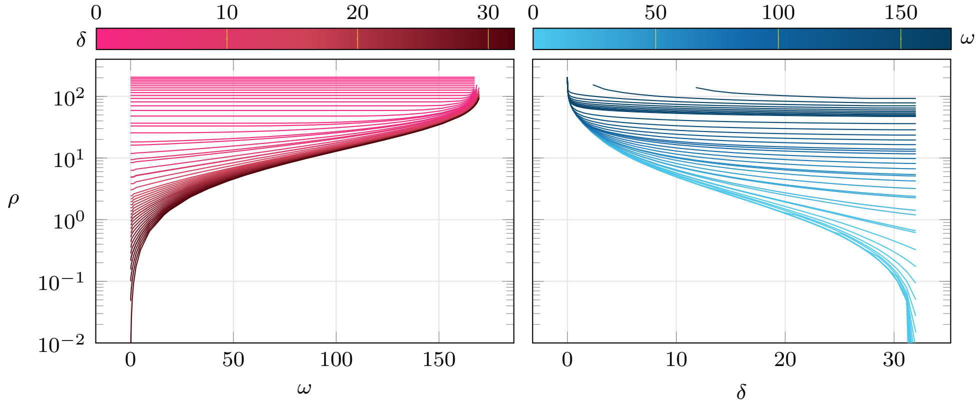


FIGURE 6.1: Trade-offs between rate-distinguishability and rate-distortion in the anomaly-agnostic case with $\alpha = 1$.

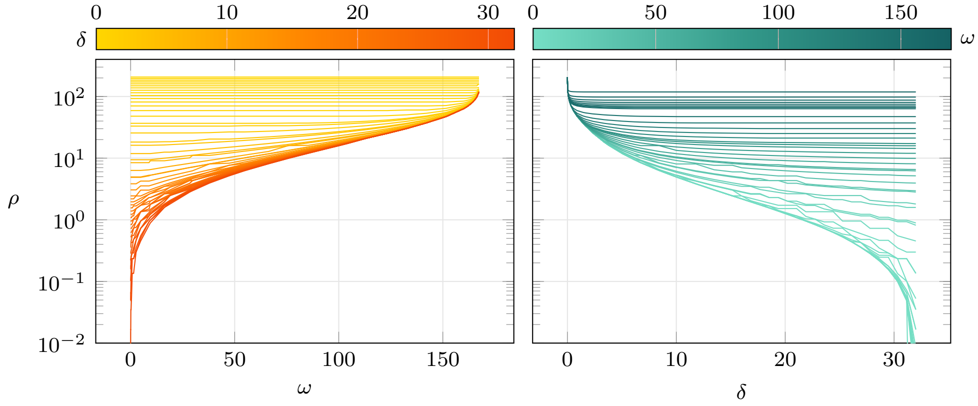


FIGURE 6.2: Trade-offs between rate-distinguishability and rate-distortion in the anomaly-aware case with $\mathbf{\Sigma}^{\text{ko}} = \mathbf{I}_n$.

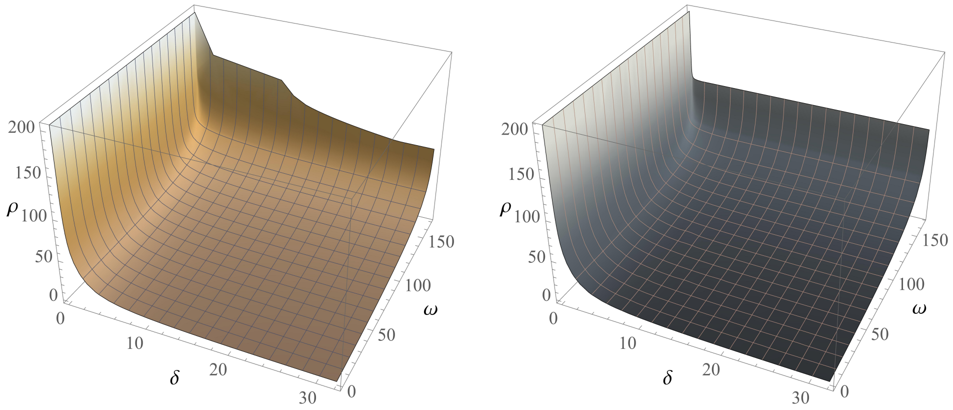
which involves maximizing a convex function. Then, we solve each sub-problem using the Disciplined Convex-Concave Programming (DCCP) framework [134].

## 6.3 Numerical evidence

The numerical results consist of four parts. We begin by analyzing the solution of the optimization problem defined in (6.16)-(6.20), which assumes Gaussian signals and Gaussian-additive encoding (as highlighted in Table 6.1). Next, we progressively relax these assumptions through three examples: *(i)* a compression scheme based on dimensionality reduction, allowing us to process Gaussian signals; *(ii)* a modified version of the JPEG compression standard, which lets us jointly optimize for distortion and detection performance relying on (6.16)-(6.20); *(iii)* a compression technique using an autoencoder, incorporating a regularization term to enhance detection.

### 6.3.1 Solution to the Pareto problems

We begin by illustrating the solutions to both RDD Pareto problems, using the setup for signal generation outlined in Chapter 4. Specifically, $n = 32$, and $\mathbf{\Sigma}^{\text{ok}}$ is a diagonal matrix such that $\mathcal{L}_{\mathbf{x}^{\text{ok}}} = 0.2$. The resulting eigenvalue profile $\boldsymbol{\lambda}^{\text{ok}}$ is shown in the top plot of Figure 6.6. The solution to the RDD minimization problem in (6.16)-(6.20), in the anomaly-agnostic scenario defined by (6.19), is obtained by setting $\alpha = 1$. For the



(A) Pareto surface for the anomaly-agnostic case with $\alpha = 1$.

(B) Pareto surface for the anomaly-aware case where $\mathbf{\Sigma}^{\mathrm{ko}} = \mathbf{I}_n$.

FIGURE 6.3: Rate-distortion-distinguishability Pareto surfaces.

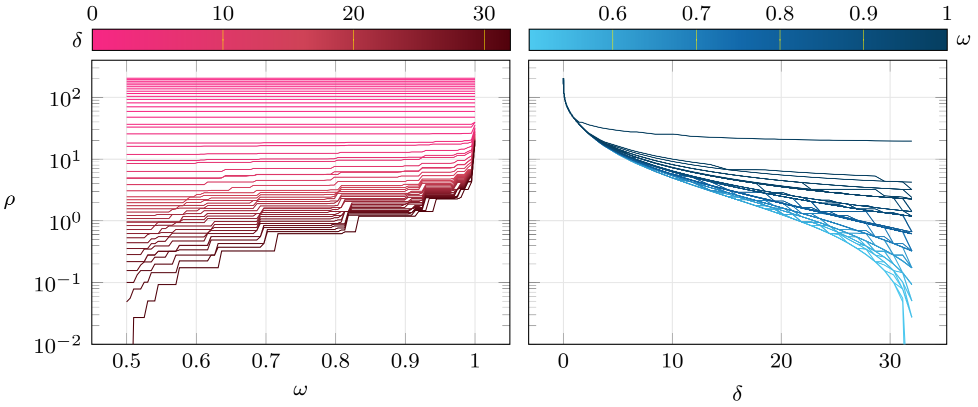


FIGURE 6.4: Trade-offs between rate-distinguishability and rate-distortion for the LD detector detecting the white anomaly.

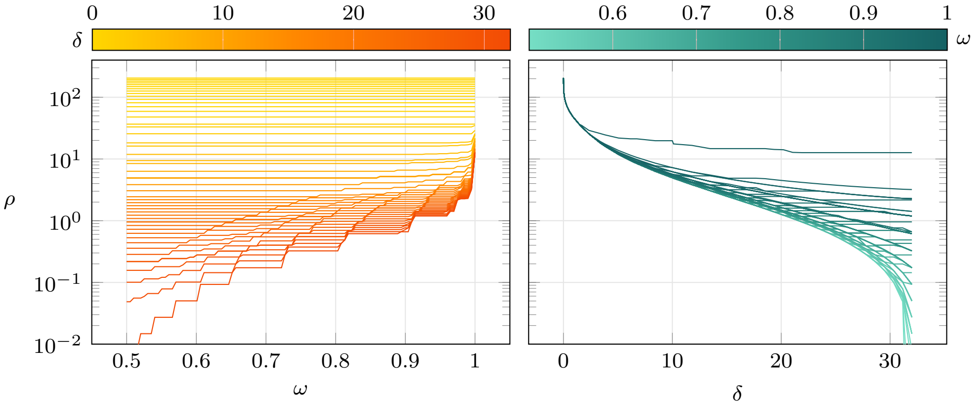


FIGURE 6.5: Trade-offs between rate-distinguishability and rate-distortion for the NPD detector detecting the white anomaly.

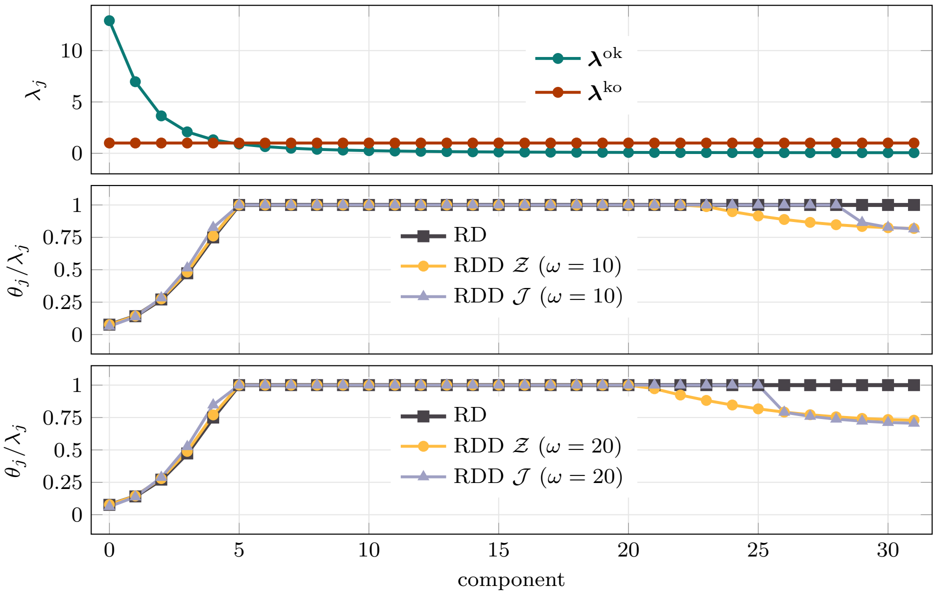


FIGURE 6.6: Distortion profiles of signal components for rate-distortion (RD) and rate-distortion-distinguishability (RDD) with $\delta = 10$.

anomaly-aware case, where detectability is set through (6.20), we assume $\boldsymbol{\Sigma}^{\mathrm{ko}} = \mathbf{I}_n$ (refer to Figure 6.6 for the $\boldsymbol{\lambda}^{\mathrm{ko}}$ profiles).

The solutions for the RDD Pareto problems are shown in Figure 6.1 and Figure 6.2, corresponding to the constraints from (6.19) and (6.20), respectively. These figures highlight the relationship between distinguishability and rate at different distortion levels, as well as the relationship between distortion and rate for different distinguishability levels. In both the anomaly-aware and anomaly-agnostic scenarios, the results demonstrate that for a fixed rate, increasing distortion improves distinguishability. Similarly, for a given distortion, the distinguishability is enhanced as the rate increases. To further confirm these trends, the Pareto surfaces are plotted in Figure 6.3.

Since $\mathcal{Z}$ and $\mathcal{J}$ are not directly comparable, we assess distinguishability using $P_D$ to compare the performance between anomaly-agnostic and anomaly-aware scenarios. To estimate $P_D$, we first generate $N = 10^4$ samples of both $\hat{\mathbf{x}}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ko}}$. For each sample, we then calculate the anomaly score. In the anomaly-agnostic case, we use a likelihood-based detector (LD) with a score $z(\hat{\mathbf{x}}) = -\log f^{\mathrm{ok}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$. In the anomaly-aware scenario, we use the Neyman-Pearson detector (NPD) with the score $z(\hat{\mathbf{x}}) = \log f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) - \log f^{\mathrm{ok}}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$. Finally, we estimate $\mathrm{AUC}$ [45] from the normal and anomalous scores.

The results, shown in Figure 6.4 and Figure 6.5, compare the agnostic and aware scenarios. It is evident that in the anomaly-aware case, the same level of distinguishability (measured by $P_D$) is achieved at a lower rate. This outcome is expected, as the detector in the anomaly-aware scenario can leverage additional information about the anomaly.

While Figure 6.4 and Figure 6.5 provide a performance comparison, Figure 6.6 offers insight into how the optimization problem in (6.16)-(6.20) affects the compression process. As already mentioned, the top plot shows the eigenvalue profiles of the normal ($\boldsymbol{\lambda}^{\mathrm{ok}}$) and anomalous ($\boldsymbol{\lambda}^{\mathrm{ko}}$) signals. In contrast, the two lower plots illustrate the relative distortion applied to the input components by the compressor. These distortions depend on the active constraints in the RDD problem for a distortion level of $\delta = 10$ and two different values of the distinguishability constraint $\omega$.

When only constraints (6.17) and (6.18) are active (i.e., $\omega = 0$), we refer to the problem as the classical rate-distortion (RD) case. This is equivalent to the Gaussian

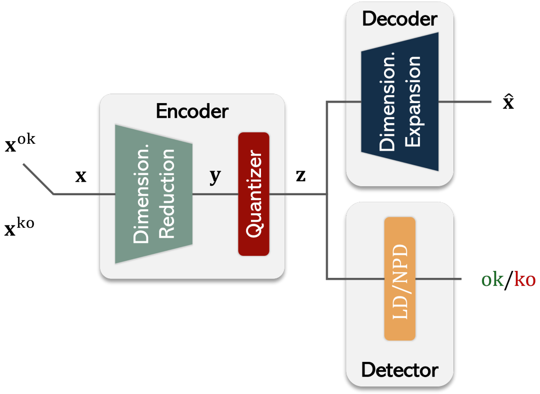


FIGURE 6.7: Scheme for the compression relying on Random Component Selection.

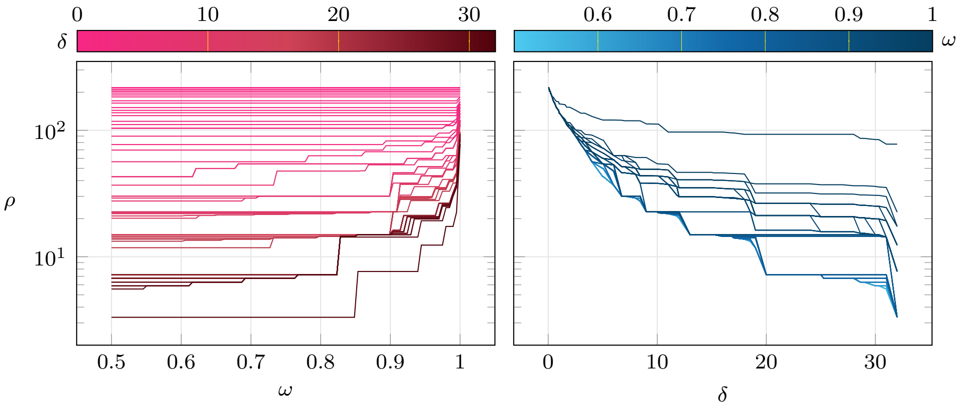


FIGURE 6.8: Rate-distinguishability and rate-distortion trade-offs administered by RCS compressor in case of LD detecting the white anomaly

source rate-distortion problem described in Chapter 2. However, when (6.19) or (6.20) is also active, the problem becomes the rate-distortion-distinguishability (RDD) case. We denote RDD $\mathcal{Z}$ and RDD $\mathcal{J}$ to identify the anomaly-agnostic and anomaly-aware scenarios, respectively.

In the RD case, the compressor applies most of the distortion to the components corresponding to the largest eigenvalues in $\boldsymbol{\lambda}^{\text{ok}}$ (the principal components), while it completely distorts the minor components. Essentially, the compressor neglects the lower-energy components to allocate more rate to those with higher energy content, leading to the increase of relative distortion $\theta_j/\lambda_j$ with $j$ until it reaches $1$ (complete distortion).

In the RDD cases, a similar behavior occurs, but the compressor reallocates some of the rate from the principal components to the lower-energy components. This effect is more evident with a stricter distinguishability constraint (higher $\omega$). The main difference between RDD $\mathcal{Z}$ and RDD $\mathcal{J}$ is that for the latter, the compressor can utilize information about the anomaly distribution to direct rate allocation where $\boldsymbol{\lambda}^{\text{ok}}$ and $\boldsymbol{\lambda}^{\text{ko}}$ differ most.

The observation that components with the lowest variance are the most informative for detection aligns with the theoretical results from [144], where it was demonstrated that, during dimensionality reduction, retaining the low-variance components is more advantageous for detection than keeping the high-variance ones. This fact has been empirically validated in the context of both image data [121] and time series analysis [101].

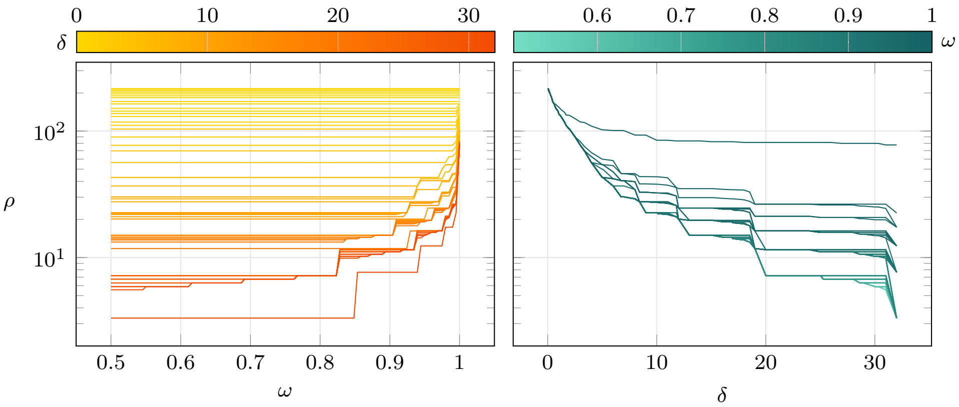


FIGURE 6.9: Rate-distinguishability and rate-distortion trade-offs administered by RCS compressor in case of NPD detecting the white anomaly

### 6.3.2 Random component selection

The RDD optimization problem works under certain assumptions about the signals and the encoder. In this case, we relax the assumption on the encoder and consider a dimensionality reduction based on the Karhunen-Loève Transform (KLT) followed by a quantization stage, as shown in Figure 6.7. Dimensionality reduction leads to distortion by projecting the input signal $\mathbf{x}$ onto a subspace spanned by $k < n$ selected components, which are the eigenvectors of $\mathbf{\Sigma}^{\text{ok}}$. The quantization stage, on the other hand, is used to limit the rate. When the principal components are selected, this approach is equivalent to the Principal Component Analysis (PCA)-based compression, described in Chapter 2, but here we also explore the option of selecting components other than the principal ones.

Since dimensionality reduction is a linear operation and the quantization can be modeled as a sum of Gaussian variables, the compressed signals retain their Gaussian nature.

The optimal dimensionality reduction for minimizing distortion relies on PCA, which picks the $k$ components corresponding to the largest eigenvalues of $\mathbf{\Sigma}^{\text{ok}}$. However, the addition of a distinguishability constraint makes the problem extremely hard to solve analytically. Therefore, we adopt a heuristic approach to jointly optimize distortion and distinguishability by randomly selecting $k$ components. Specifically, we represent this selection with a vector $\mathbf{k}$, whose elements are the indices of the $k$ randomly chosen components from $\mathbf{x}$, resulting in the signal $\mathbf{y} = (x_j | j \in \mathbf{k})$.

Given that the procedure is linear and the input vector $\mathbf{x}$ is Gaussian, the output signal $\mathbf{y}$ is also Gaussian. The covariance matrices of $\mathbf{y}$, depending on whether the input vector is $\mathbf{x}^{\text{ok}}$ or $\mathbf{x}^{\text{ko}}$, are given by:

$$\hat{\mathbf{\Sigma}}_{\theta}^{\text{ok}} = \text{diag}\left(\left(\lambda_j^{\text{ok}} | j \in \mathbf{k}\right)\right), \qquad \hat{\mathbf{\Sigma}}_{\theta}^{\text{ko}} = \left[\mathbf{\Sigma}^{\text{ko}}\right]_{\mathbf{k}}, \tag{6.22}$$

where $[\cdot]_{\mathbf{k}}$ denotes an operator that obtains a sub-matrix from a square matrix by selecting rows and columns indexed by $\mathbf{k}$. If $\mathbf{\Sigma}^{\text{ok}}$ is diagonal, then $\left[\mathbf{\Sigma}^{\text{ko}}\right]_{\mathbf{k}} = \text{diag}\left(\left(\lambda_j^{\text{ko}} | j \in \mathbf{k}\right)\right)$.

The subsequent quantization stage solely limits the rate, which would otherwise be infinite without it [15]. We consider a Gaussian source quantizer optimal from a rate-distortion perspective, described in Chapter 2, and design it so that the introduced distortion is minimal, fixed at $\epsilon \ll \lambda_{n-1}^{\text{ok}}$. This quantizer is Gaussian-additive, yielding a quantized vector $\mathbf{z} = \mathbf{y} + \mathbf{\Delta}$, where $\mathbf{\Delta}$ is a zero-mean Gaussian vector with independent components, each with variance $\theta_\epsilon = \epsilon/k$.

Finally, the decoder reconstructs the signal $\hat{\mathbf{x}}$ from the compressed vector $\mathbf{z}$ such that:

$$\hat{x}_j = \begin{cases} z_i \text{ where } i \text{ satisfies } m_i = j, & j \in \mathbf{k} \\ 0, & j \notin \mathbf{k}. \end{cases} \tag{6.23}$$

In the designed scheme, the total distortion $D$ is the sum of the dimensionality reduction and quantization contributions. The former is the sum of the eigenvalues corresponding to the discarded components, while the latter is represented by $\epsilon$, a degree of freedom that we set to be negligible. Thus, the resulting distortion is given by

$$D = \sum_{j \notin \mathbf{k}} \lambda_j^{\text{ok}} + \epsilon. \tag{6.24}$$

On the other hand, the quantization stage determines the rate. Specifically, the rate $\rho = \mathcal{I}(\hat{\mathbf{x}}^{\text{ok}}; \mathbf{x}^{\text{ok}})$ is equivalent to $\mathcal{I}(\mathbf{z}^{\text{ok}}; \mathbf{x}^{\text{ok}})$, as $\hat{\mathbf{x}}^{\text{ok}}$ and $\mathbf{z}^{\text{ok}}$ contain the same information. Since $\mathbf{z}^{\text{ok}}$ is derived from $\mathbf{y}^{\text{ok}}$, which is in turn obtained from $\mathbf{x}^{\text{ok}}$, the information shared between $\mathbf{y}^{\text{ok}}$ and $\mathbf{z}^{\text{ok}}$ is the same as that shared between $\mathbf{z}^{\text{ok}}$ and $\mathbf{x}^{\text{ok}}$. Hence, using (2.3), the rate can be expressed as:

$$\rho(\epsilon) = \mathcal{I}_\epsilon(\hat{\mathbf{x}}^{\text{ok}}; \mathbf{x}^{\text{ok}}) = \mathcal{I}_\epsilon(\mathbf{z}^{\text{ok}}; \mathbf{y}^{\text{ok}}) = \frac{1}{2} \sum_{j \in \mathbf{k}} \log_2 \frac{\lambda_j^{\text{ok}}}{\theta_\epsilon}, \tag{6.25}$$

which highlights the rate's dependence on $\epsilon$, a free parameter. It is important to note that the choice of $\epsilon$ is irrelevant as long as $\theta_\epsilon = \epsilon/k < \lambda_{n-1}^{\text{ok}}$, meaning none of the $\mathbf{y}^{\text{ok}}$ components are fully distorted. This can be proven using an alternative expression of (6.25) [77]:

$$\mathcal{I}_\epsilon(\mathbf{z}^{\text{ok}}; \mathbf{y}^{\text{ok}}) = \frac{k}{2} \log \frac{k}{2\epsilon^2 \pi e} + \mathcal{H}(\mathbf{y}^{\text{ok}}) + o(1). \tag{6.26}$$

It shows that the rate depends mainly on the subspace dimensionality $k$, while the differential entropy $\mathcal{H}(\mathbf{y}^{\text{ok}})$ represents the rate's variability due to different combinations of $k$ components. The parameter $\epsilon$ only determines the rate offset. In fact, $\mathcal{I}_{\epsilon''}(\hat{\mathbf{x}}^{\text{ok}}; \mathbf{x}^{\text{ok}}) = \mathcal{I}_{\epsilon'}(\hat{\mathbf{x}}^{\text{ok}}; \mathbf{x}^{\text{ok}}) - k\left(\log {}^1/_{\epsilon'} - \log {}^1/_{\epsilon''}\right)$.

For the carried numerical analysis, we set $\epsilon = \frac{\lambda_{n-1}^{\text{ok}}}{10^2}$ and randomly selected $K = \min\left\{\binom{n}{k}, 10^4\right\}$ vectors $\mathbf{k}$ for each $k \in \{1, \ldots, n-1\}$. For each encoder, we measured the rate and distortion using (6.25) and (6.24), respectively. As in the previous example, we modeled the anomaly with $\mathbf{\Sigma}^{\text{ko}} = \mathbf{I}_n$ and adopted $P_D$ to assess the performance of LD and NPD.

Figure 6.8 and Figure 6.9 show the resulting trade-offs. The profiles resemble the trends observed with the optimal Gaussian additive encoder in Figure 6.4 and Figure 6.5, where improving the performance of supervised and unsupervised detection performance comes at the cost of a higher rate or increased distortion. However, since this compression scheme is suboptimal, the rate required to achieve the same distortion-distinguishability performance is necessarily higher than that of the Gaussian-additive encoder optimized using (6.16).

### 6.3.3 JPEG compression and detection

In the previous section, we relaxed the assumption on the encoder. Here, we generalize the approach by also relaxing the assumption on the signals. Specifically, we apply the RDD framework to enhance the detection performance of a widely used compression algorithm. For this, we selected JPEG [152], a standard for lossy compression of both

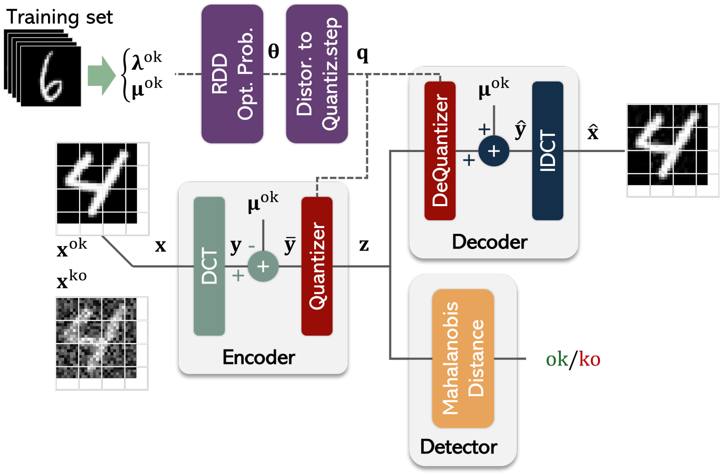


FIGURE 6.10: JPEG block diagram.

color and grayscale images. For simplicity, we only consider grayscale images from the MNIST dataset [39].

JPEG compression of grayscale images contains several stages. First, the image is divided into square blocks of $8 \times 8$ pixels, with each block independently processed by the Discrete Cosine Transform (DCT). The resulting DCT coefficients are then quantized based on a quantization table that reflects the required reconstruction quality. Finally, the quantized coefficients are encoded by entropy coding. Decoding involves reversing this procedure, starting with entropy decoding, followed by dequantization and inverse DCT (IDCT) to reconstruct the image blocks.

Since the quantization stage is the one responsible for information loss, it is also the main target of the proposed RDD framework. The core idea is to adapt the quantization table according to the distortion and distinguishability constraints derived from the RDD optimization process. For this example, we limit our analysis to the unsupervised scenario, where the compressor lacks knowledge of the anomaly, and we consider only RDD $\mathscr{Z}$ in (6.7).

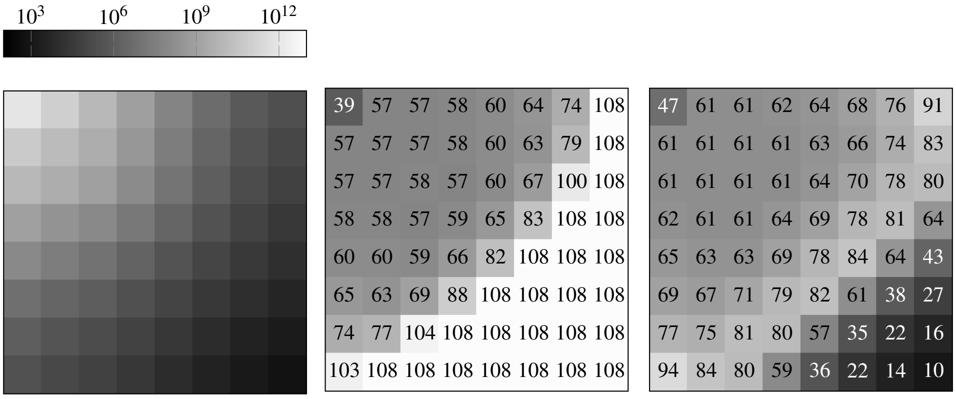
We assume that the DCT coefficients of each block are independent and obey a Gaussian distribution with mean $\mu_j^{\text{ok}}$ and variance $\lambda_j^{\text{ok}}$, where $j = 0, 1, \ldots, 63$ represents the index of the DCT coefficient. Let $\mathbf{x}$ be a block from a sample image in the training set and let $\mathbf{y} = \text{DCT}(\mathbf{x})$ represent its corresponding DCT coefficients. For the $j$-th DCT coefficient, the mean and variance are estimated as:

$$\mu_j^{\text{ok}} = \frac{1}{NB} \sum_{i=0}^{N-1} \sum_{l=0}^{B-1} y_j^{(i,l)}, \quad \lambda_j^{\text{ok}} = \frac{1}{NB} \sum_{i=0}^{N-1} \sum_{l=0}^{B-1} \left[ y_j^{(i,l)} - \mu_j^{\text{ok}} \right]^2, \tag{6.27}$$

where $y_j^{(i,l)}$ is the $j$-th DCT coefficient of the $l$-th block in the $i$-th image, $N$ is the number of images in the training set, and $B$ is the number of blocks per image.

To reformulate the optimization problem from (6.7) into (6.16)-(6.19), we define a zero-mean Gaussian signal $\bar{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}^{\text{ok}}$. The solution to this optimization problem generates a set of parameters $\xi = (\xi_0, \ldots, \xi_{63})$ that minimize the rate $\rho$ without violating the constraints on distortion $\delta$ and distinguishability $\omega$. From $\xi_j$, we compute the average distortion $\theta_j = \lambda_j^{\text{ok}}(1 - \xi_j)$ introduced by quantizing $\bar{y}_j$, to balance the rate and the constraints on distortion and distinguishability.

The relationship required to map the distortion into quantization steps is empirically estimated from the training set. Specifically, we map the distortion vector $\theta =$



(A) $\lambda^{\text{ok}}$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 39 | 57 | 57 | 58 | 60 | 64 | 74 | 108 |
| 57 | 57 | 57 | 58 | 60 | 63 | 79 | 108 |
| 57 | 57 | 58 | 57 | 60 | 67 | 100 | 108 |
| 58 | 58 | 57 | 59 | 65 | 83 | 108 | 108 |
| 60 | 60 | 59 | 66 | 82 | 108 | 108 | 108 |
| 65 | 63 | 69 | 88 | 108 | 108 | 108 | 108 |
| 74 | 77 | 104 | 108 | 108 | 108 | 108 | 108 |
| 103 | 108 | 108 | 108 | 108 | 108 | 108 | 108 |

(B) $\mathbf{q}$ with $\omega = 0$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 47 | 61 | 61 | 62 | 64 | 68 | 76 | 91 |
| 61 | 61 | 61 | 61 | 63 | 66 | 74 | 83 |
| 61 | 61 | 61 | 61 | 64 | 70 | 78 | 80 |
| 62 | 61 | 61 | 64 | 69 | 78 | 81 | 64 |
| 65 | 63 | 63 | 69 | 78 | 84 | 64 | 43 |
| 69 | 67 | 71 | 79 | 82 | 61 | 38 | 27 |
| 77 | 75 | 81 | 80 | 57 | 35 | 22 | 16 |
| 94 | 84 | 80 | 59 | 36 | 22 | 14 | 10 |

(C) $\mathbf{q}$ with $\omega = 10^3$

FIGURE 6.11: Variance of the DCT coefficients (A) and quantization tables as a result of two different values of the distinguishability constraint and same distortion $\delta = 0.3$ (B) and (C).

$(\theta_0, \ldots, \theta_{63})$ to a quantization table $\mathbf{q}$, where each $\theta_j$ corresponds to a quantization step $q_j$ applied to $\bar{y}_j$, resulting in the quantized version $z_j = \lfloor \bar{y}_j / q_j \rceil$, where $\lfloor \cdot \rceil$ represents the rounding to the nearest integer. Figure 6.10 shows the JPEG encoder and decoder block diagram, with the quantization table $\mathbf{q}$ derived from the RDD optimization in (6.16)-(6.19)[1].

In this scheme, the distortion is measured as the Mean Squared Error (MSE) between $\mathbf{x}^{\text{ok}}$ and $\hat{\mathbf{x}}^{\text{ok}}$, while the rate is calculated as the Shannon entropy $H$ of $\mathbf{z}^{\text{ok}}$, following [3]:

$$\rho = \mathcal{I}(\mathbf{x}^{\text{ok}}; \hat{\mathbf{x}}^{\text{ok}}) \simeq H(\mathbf{z}^{\text{ok}}) \simeq -n \sum_{l} \hat{p}_l \log \hat{p}_l \tag{6.28}$$

where $\hat{p}_l$ is the estimated probability that an entry in $\mathbf{z}$ is equal to the value $l$.

Considering the MNIST dataset, which contains $N = 60,000$ training samples of $28 \times 28$ images, each image is divided into $B = 16$ blocks. Figure 6.11-(A) illustrates the variance of the DCT coefficients $\boldsymbol{\lambda}^{\text{ok}}$ estimated from the training set. Figures 6.11-(B) and (C) present two examples of quantization tables generated with $\delta = 0.3$ and varying levels of the distinguishability constraint, $\omega = 0$ (no constraint) and $\omega = 10^3$, respectively. Like most natural images, the MNIST dataset exhibits low-pass characteristics, resulting in DCT coefficients associated with low frequencies (upper left corner of Figure 6.11-(A)) with higher magnitudes on average, while the high-frequency coefficients (lower right corner) have lower magnitudes.

In line with the observations from Figure 6.6, classical rate-distortion solutions apply uniform distortion across all components, meaning that low-frequency components experience less relative distortion compared to high-frequency ones. This is clear in Figure 6.11-(B), consistent with standard JPEG quantization tables [152]. However, as highlighted in Figure 6.11-(C), when the distinguishability constraint is added, high-frequency components undergo less distortion than mid-frequency ones, despite the latter being generally more informative for reconstruction. Both quantization tables in Figure 6.11-(B) and Figure 6.11-(C) adapt JPEG compression to the MNIST dataset, but the latter is optimized for AD, allowing the detection of anomalies affecting the input while maintaining the same distortion.

[1] Entropy encoding and decoding are omitted as they involve lossless compression, and jointly correspond to an identity mapping.

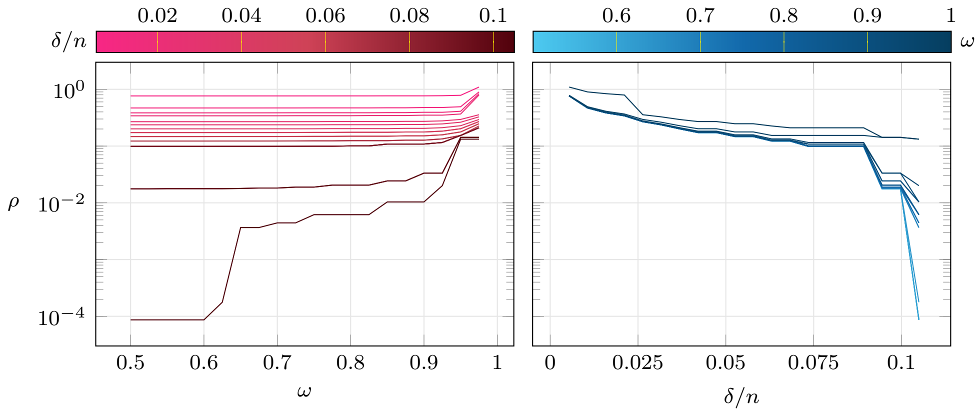


FIGURE 6.12: Rate-distortion-distinguishability trade-off with JPEG adapted to the MNIST dataset, using an anomaly modeled as uniform noise mixed with the signal.

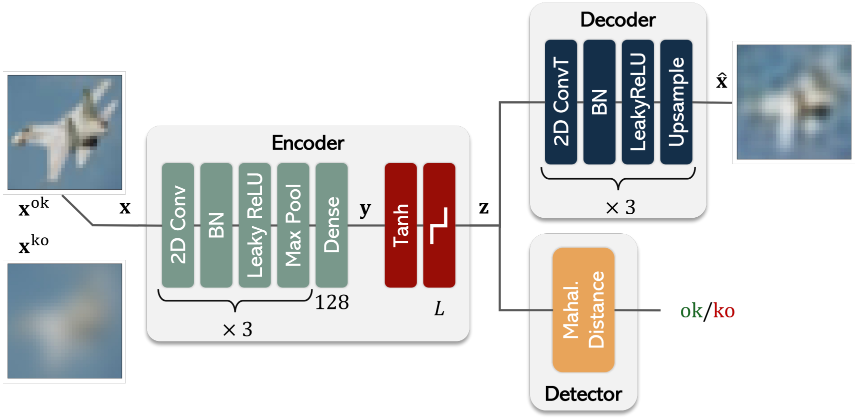


FIGURE 6.13: System for neural network-based CIFAR-10 image compression and detection.

To evaluate the performance of such quantization tables, we introduce anomalies in the form of uniform noise mixed with the MNIST images and use the Mahalanobis Distance (MD) [38] to detect them. Since JPEG performs block-wise compression, the detector analyzes the DCT coefficients of the image blocks to determine whether each block corresponds to a corrupted image. The mean and covariance of the DCT coefficients for each compressed block are estimated from the compressed images of the training set.

Figure 6.12 illustrates the Pareto curve for JPEG compression with the MD detector when uniform noise is mixed with the input images. The rate is computed as the Shannon entropy of the quantized DCT coefficients. The trends align with previous observations, underlying that improving detection performance or reducing distortion requires a higher rate. This result highlights the relevance of considering the rate-distortion-distinguishability trade-off, even when dealing with images as signals and using JPEG as a compressor.

### 6.3.4 End-to-end image compression and detection

As we have stressed in Chapter 2, autoencoder (AE)-based lossy compression methods have recently gained popularity as they enable the end-to-end optimization of all the components that form a compressor. In this context, we investigate the simultaneous training of the compressor and the detector.

While AEs perform dimensionality reduction, they need a quantization layer at the bottleneck to limit the rate for proper lossy compression. We apply the quantization method proposed in [3], which has been widely adopted in other compression techniques based on AE [103, 18]. The quantization process consists of scaling the latent vector within the range $[-1, 1]$ using a $\tanh$ function, followed by uniform quantization of each component with $L$ levels.

In this particular example, we focus on the CIFAR-10 dataset [79] and adopt the AE architecture from [126], adding a quantization layer. Specifically, the encoder reduces the input image of size $32\times32\times3$ (where $n = 3072$) to a latent vector $\mathbf{y}$ of dimension $k = 128$, which is then quantized. The components of the quantized signal $\mathbf{z}$ are computed as:

$$z_j = \arg\min_k |\tanh(y_j) - c_l| \tag{6.29}$$

with $c_l = -1 + 2l/(L-1)$ representing the $l$-th quantization level.

Given that this quantization step is non-differentiable, the backward pass is approximated using the soft quantization from [103]:

$$z_j = \sum_{l=0}^{L-1} \frac{e^{-|\tanh(y_j)-c_l|}}{\sum_{t=0}^{L-1} e^{-|\tanh(y_t)-c_t|}} c_l. \tag{6.30}$$

The quantized vector $\mathbf{z}$ is processed by the decoder, which reconstructs the input as $\hat{\mathbf{x}}$, while the detector outputs the anomaly score using the MD detector applied to $\mathbf{z}$. To promote detection, a regularization term acting on $\|\mathbf{y}\|^2$ is added, inspired by the Deep Support Vector Data Description [126] presented in Chapter 1, and Shrink AE [26] models. These models optimize the encoder for detection by minimizing the volume occupied by normal data $\mathbf{y}^{\text{ok}}$ in the latent space. The full system is illustrated in Figure 6.13 and is trained using the following loss function:

$$\mathcal{L}_\beta(\mathbf{x}^{\text{ok}}) = \frac{1}{n}\text{MSE}(\mathbf{x}^{\text{ok}}, \hat{\mathbf{x}}^{\text{ok}}) + \frac{\beta}{kN}\sum_{i=0}^{N-1} \|\mathbf{y}^{\text{ok}}\|^2 \tag{6.31}$$

where $\beta$ controls the regularization weight. This loss allows tuning the encoder for both reconstruction (through minimization of MSE) and AD (via minimizing the regularization term).

The distortion is estimated as MSE between $\mathbf{x}^{\text{ok}}$ and $\hat{\mathbf{x}}^{\text{ok}}$, while the rate is approximated by the Shannon entropy of $\mathbf{z}^{\text{ok}}$ as expressed in (6.28). In this case, the dimension of $\mathbf{z}^{\text{ok}}$ is $k$, and $\hat{p}_l$ is the estimated probability of $\mathbf{z}^{\text{ok}}$ to be equal to the $l$-th quantization level $c_l$.

As anomalies, we use the corruptions defined in [60], which has already been used as a benchmark for AD in [125]. Specifically, we focus on the Gaussian blur as a target corruption.

We train[2] several AEs, each characterized by a different combination of $L \in \{2, 2^2, \ldots, 2^9\}$, affecting the rate, and $\beta \in \{0, 10^{-6}, 10^{-5}, \ldots, 10^3\}$, acting on the trade-off between distortion and detection. The resulting three-fold trade-off between rate, distortion, and detection is presented in Figure 6.14. Despite making no assumptions about the signals or compressor, the results show that the relationship between distinguishability, rate, and distortion aligns with theoretical trends.

[2] Training is performed using the Adam optimizer, with an $l_2$ weight regularization coefficient of $10^{-6}$, a batch size of $200$, and an initial learning rate of $10^{-4}$, which is reduced when the loss reaches a plateau for $20$ epochs.

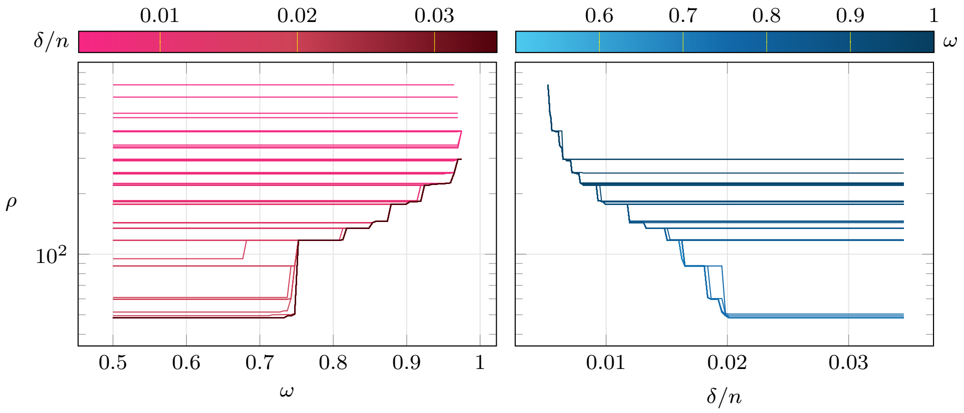


FIGURE 6.14: Rate-distortion-distinguishabilty trade-off for neural network-based system working with CIFAR-10 images in the Gaussian blur anomaly-agnostic case.

## 6.4 Conclusion

In applications where anomalies are detected from compressed signals, a trade-off arises between three quantities: rate, distortion, and distinguishability. To jointly address this trade-off, we extended the traditional rate-distortion theory by incorporating distinguishability constraints, resulting in the formulation of two optimization problems.

Assuming Gaussian signals and using a Gaussian-additive encoder, we solve these optimization problems that reveal and allow us to discuss the Pareto surface involving all three quantities in anomaly-agnostic and anomaly-aware scenarios. In the previous chapter, we have demonstrated that minimizing distortion under a fixed rate budget can impair the ability to distinguish normal and anomalous signals after compression. In this chapter, we showed that preserving distinguishability and enhancing detection performance requires increasing either the rate or the distortion. For instance, under a fixed distortion constraint, a compressor can maintain distinguishability by allocating part of the rate to components that may contribute less to reconstruction quality but are critical for anomaly detection.

To assess the generality of our framework, we evaluated it in three different scenarios: a compression scheme based on linear dimensionality reduction followed by quantization, a modified JPEG compressor incorporating distinguishability, and an autoencoder-based compression mechanism optimized for both reconstruction and detection. All three examples exhibited the theoretical trends we predicted, reinforcing the importance of jointly managing the trade-off between rate, distortion, and distinguishability in such applications.

# Chapter 7

# Anomaly Detection-Reconstruction Trade-off via Autoencoder

A lossy compression approach often adopted in acquisition systems is based on autoencoder-based dimensionality reduction (AE) [78]. As highlighted in Chapter 2, AEs are neural networks that extend Principal Component Analysis (PCA) [24] by learning a manifold, rather than just a subspace, that optimally preserves the signal's information content. The AE comprises two components: an encoder that compresses the input data by reducing its dimensionality, and a decoder that recovers the original data from the compressed version[1]. Both parts are trained simultaneously to minimize the reconstruction error, i.e., the deviation between the input and the output.

The impact of compression on reconstruction has been extensively studied since the development of rate-distortion theory [34, Chapter 8] and is summarized in Chapter 2, where the use of AE for compression has been highlighted.

Reconstruction may not be the main goal when working with compressed data. The idea of optimizing an AE-like architecture for tasks besides reconstruction has been explored in [27, 12], where Variational AE-based frameworks were introduced. These frameworks jointly optimize the encoder, decoder, and classifier for image data, improving classification accuracy without significantly compromising reconstruction quality.

As discussed in the previous chapters, anomaly detection (AD) is a fundamental analysis that goes along with reconstruction in fields such as structural health monitoring [101, 128], condition monitoring [73, 164], and healthcare monitoring [97, 1]. In this chapter, we aim to jointly optimize the AE for AD and reconstruction.

[1] While we have repeatedly emphasized that an AE is only a preprocessing step in the broader context of lossy compression, it is common for the terms dimensionality reduction and compression to be used interchangeably. This is often due to the assumption that the quantization is so fine that its effects can be neglected.

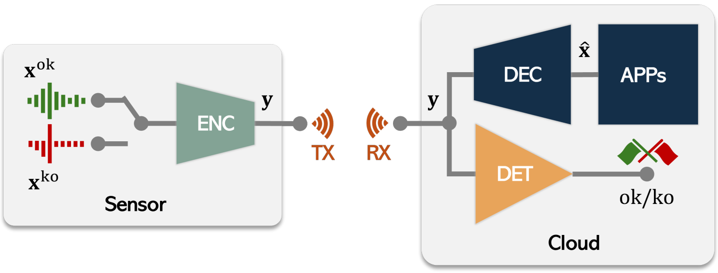


FIGURE 7.1: Block diagram of the system composed of an encoder (ENC) that compresses a signal and a decoder (DEC) that reconstructs it for further processing applications (APPs). The signal is either normal $\mathbf{x}^{\text{ok}}$ or anomalous $\mathbf{x}^{\text{ko}}$ and a detector (DET) discriminates between them by processing the information at the receiver side.

In case only AD is prioritized, AE-based dimensionality reduction remains a popular strategy for enhancing performance. For instance, in [126], the authors introduced Deep Support Vector Data Description (Deep SVDD), which first trains an AE to minimize reconstruction error, and then tunes the encoder to minimize the volume of the latent space for AD. On the other hand, the approach proposed in [26], known as Shrink AE (SAE), accounts for latent space volume minimization through a regularization term directly during the AE's training. This eliminates the need for subsequent encoder fine-tuning. However, since these methods are aimed at AD, they neglect reconstruction quality.

The relationship between compression and AD has been covered in the previous two chapters. Specifically, in Chapter 5 we have demonstrated that the detection performance degrades rapidly with increasing distortion when compression is performed by an AE optimized for only reconstruction. In Chapter 6 we have established a theoretical framework for the joint optimization of distinguishability and reconstruction. The observed theoretical trends were confirmed for several practical compression schemes, including one that exploits the AE structure.

In this chapter, we present a novel practical approach based on AE that targets both AD and reconstruction tasks. Specifically in Section 7.1, inspired by the information-theoretic guideline provided by $\zeta$ introduced in Chapter 4, we design a loss function for AE that allows not only to minimize the distortion but also to preserve useful information for AD. The latter is achieved by introducing into the loss function a regularization of the differential entropy in the latent space. We propose two different strategies, each based on a different differential entropy estimator. In one case the estimator assumes an isotropic Gaussian distribution of the latent vector and in the other, the estimator makes no assumptions about the distribution. In Section 7.2 we compare the performance of these two AEs in managing the trade-off between AD and reconstruction. We used both image data and time series ECG data as case studies, analyzing the trade-off across various types of anomalies commonly found in these datasets.

The results show that the entropy-regularization strategy, independently of the estimator, helps to retain features beneficial for AD, at a slight cost in reconstruction performance.

## 7.1 Mathematical models

We consider a system, illustrated in Figure 7.1, where a sensor acquires an $n$-dimensional signal $\mathbf{x}$, encodes it into a lower-dimensional $k$-dimensional vector $\mathbf{y} = \mathrm{ENC}(\mathbf{x})$, where $k < n$, and transmits it. At the receiver, the signal is reconstructed for further applications (APPs) using a decoder $\hat{\mathbf{x}} = \mathrm{DEC}(\mathbf{y})$. Simultaneously, a detector (DET) identifies whether the signal is normal $\mathbf{x}^{\mathrm{ok}} \sim f_{\mathbf{x}}^{\mathrm{ok}}$ or anomalous $\mathbf{x}^{\mathrm{ko}} \sim f_{\mathbf{x}}^{\mathrm{ko}}$. The detector operates on $\mathbf{y}$ as it contains information equivalent to $\hat{\mathbf{x}}$ due to the injectivity of $\mathrm{DEC}$.

In this setting, where an AE is utilized for compression, both $\mathrm{ENC}$ and $\mathrm{DEC}$ are neural networks. Since anomalies are rare and unknown, training is performed on normal signals, i.e., $\mathbf{x} = \mathbf{x}^{\mathrm{ok}}$. The reconstruction quality is assessed using distortion defined in (2.1) and which in this context can be expressed as:

$$D = \mathbf{E}\left[\left\|\mathbf{x}^{\mathrm{ok}} - \hat{\mathbf{x}}^{\mathrm{ok}}\right\|^2\right]. \tag{7.1}$$

$D$ is often estimated by Mean Squared Error (MSE):

$$\mathrm{MSE}(\mathbf{x}^{\mathrm{ok}}, \hat{\mathbf{x}}^{\mathrm{ok}}) = \frac{1}{N}\sum_{i=0}^{N-1} \left\| \mathbf{x}_i^{\mathrm{ok}} - \hat{\mathbf{x}}_i^{\mathrm{ok}} \right\|^2 \tag{7.2}$$

where $\mathbf{x}_i^{\mathrm{ok}}$ and $\hat{\mathbf{x}}_i^{\mathrm{ok}}$ represent individual instances of $\mathbf{x}^{\mathrm{ok}}$ and $\hat{\mathbf{x}}^{\mathrm{ok}}$ from a set of $N$ instances.

Using an MSE-based loss function makes AE focus solely on reconstruction, which, as highlighted in Chapter 5, can result in the elimination of the features of $\mathbf{x}$ crucial for AD. In this chapter, we want to ensure that the AE retains critical information for AD at the receiver's end.

In Chapter 4, we introduced a metric $\zeta$, whose absolute value measures the distinguishability between $\mathrm{ok}$ and $\mathrm{ko}$ sources. The definition of $\zeta$ in (4.2) is valid for any two sources and since the detector operates on the encoded signal $\mathbf{y}$, we specialize $\zeta$ within the latent space:

$$\zeta = \mathcal{C}\left(\mathbf{y}^{\mathrm{ko}}; \mathbf{y}^{\mathrm{ok}}\right) - \mathcal{C}\left(\mathbf{y}^{\mathrm{ok}}; \mathbf{y}^{\mathrm{ok}}\right) \tag{7.3}$$

where $\mathcal{C}(\mathbf{y}'; \mathbf{y}'') = -\int_{\mathbb{R}^n} f_{\mathbf{y}'}(\alpha) \log_2 f_{\mathbf{y}''}(\alpha) \mathrm{d}\alpha$ is the cross-entropy, which represents the average coding rate (in bits per symbol) of source $\mathbf{y}'$ when encoded using a code optimized for $\mathbf{y}''$. Hence, $\zeta$ captures the difference in average coding rate between anomalous and normal signals when encoded by a code optimized for the normal signal.

Additionally, using $\mathcal{H}(\cdot)$ to represent differential entropy and $\mathcal{D}_{\mathrm{KL}}\left(\cdot \| \cdot\right)$ for the Kullback-Leibler divergence [34, Chapter 2], we express:

$$\mathcal{C}(\mathbf{y}^{\mathrm{ok}}; \mathbf{y}^{\mathrm{ok}}) = \mathcal{H}(\mathbf{y}^{\mathrm{ok}}) \tag{7.4}$$

$$\mathcal{C}(\mathbf{y}^{\mathrm{ko}}; \mathbf{y}^{\mathrm{ok}}) = \mathcal{H}(\mathbf{y}^{\mathrm{ko}}) + \mathcal{D}_{\mathrm{KL}}\left(f_{\mathbf{y}}^{\mathrm{ko}} \| f_{\mathbf{y}}^{\mathrm{ok}}\right) \tag{7.5}$$

which allow us to write $\zeta$ as:

$$\zeta = \mathcal{H}(\mathbf{y}^{\mathrm{ko}}) - \mathcal{H}(\mathbf{y}^{\mathrm{ok}}) + \mathcal{D}_{\mathrm{KL}}\left(f_{\mathbf{y}}^{\mathrm{ko}} \| f_{\mathbf{y}}^{\mathrm{ok}}\right) \tag{7.6}$$

Thus, improving AD would involve increasing the entropy of $\mathbf{y}^{\mathrm{ko}}$, decreasing the entropy of $\mathbf{y}^{\mathrm{ok}}$, or increasing the divergence between $f_{\mathbf{y}}^{\mathrm{ko}}$ and $f_{\mathbf{y}}^{\mathrm{ok}}$. However, in an unsupervised setting where only $\mathbf{x}^{\mathrm{ok}}$ is available, the only term that can be optimized is $\mathcal{H}(\mathbf{y}^{\mathrm{ok}})$.

To make the AE retain the information necessary for AD at the receiver, we propose to reduce $\mathcal{H}(\mathbf{y}^{\mathrm{ok}})$, the differential entropy of the latent space representation $\mathbf{y}^{\mathrm{ok}}$. To achieve this, we design the following loss function:

$$\mathcal{L}(\mathbf{x}^{\mathrm{ok}}) = \frac{1}{n}\mathrm{MSE}(\mathbf{x}^{\mathrm{ok}}, \hat{\mathbf{x}}^{\mathrm{ok}}) + \frac{\beta}{k}\hat{\mathcal{H}}(\mathbf{y}^{\mathrm{ok}}) \tag{7.7}$$

where $\hat{\mathcal{H}}(\mathbf{y}^{\mathrm{ok}})$ is an estimator of $\mathcal{H}(\mathbf{y}^{\mathrm{ok}})$, and $\beta > 0$ is the entropy regularization weight.

### 7.1.1 Differential entropy estimation

We adopt two methods to minimize $\mathcal{H}(\mathbf{y}^{\mathrm{ok}})$. The first is based on the second-order Renyi entropy estimator, as described in [116], and is defined as:

$$\hat{\mathcal{H}}_{2,\sigma}(\mathbf{y}^{\mathrm{ok}}) = -\log_2\left(\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} G_{\sqrt{2}\sigma}(\mathbf{y}_j^{\mathrm{ok}} - \mathbf{y}_i^{\mathrm{ok}})\right) \tag{7.8}$$

with $G_\sigma(\cdot) = \exp\left(-\frac{\|\cdot\|^2}{2\sigma^2}\right)$ a Gaussian kernel having width $\sigma$. This is a non-parametric estimator that makes no assumptions about the distribution of $\mathbf{y}$.

The second method assumes that the latent space follows a zero-mean isotropic Gaussian distribution, such that the minimization of $\mathcal{H}(\mathbf{y}^{\mathrm{ok}})$ is equivalent to minimizing $\mathbf{E}\left[\|\mathbf{y}^{\mathrm{ok}}\|^2\right]$ [127].

The resulting loss functions are:

$$\mathcal{L}^{\mathsf{RAE}}_{k,\sigma,\beta}(\mathbf{x}^{\mathrm{ok}}) = \frac{1}{n}\mathrm{MSE}(\mathbf{x}^{\mathrm{ok}}, \hat{\mathbf{x}}^{\mathrm{ok}}) + \frac{\beta}{k}\hat{\mathcal{H}}_{2,\sigma}(\mathbf{y}^{\mathrm{ok}}) \tag{7.9}$$

$$\mathcal{L}^{\mathsf{SAE}}_{k,\beta}(\mathbf{x}^{\mathrm{ok}}) = \frac{1}{n}\mathrm{MSE}(\mathbf{x}^{\mathrm{ok}}, \hat{\mathbf{x}}^{\mathrm{ok}}) + \frac{\beta}{kN}\sum_{i=0}^{N-1}\|\mathbf{y}_i^{\mathrm{ok}}\|^2 \tag{7.10}$$

where RAE stands for Renyi entropy-regularized AE, and SAE refers to the Shrink AE introduced in [26].

Both approaches allow exploration of the trade-off between reconstruction and detection by acting on the regularization weight $\beta$. Although an increase in $\beta$ will result in a reduction in reconstruction performance, enhancing AD performance cannot be achieved by arbitrarily increasing $\beta$, rather the two terms should be properly weighted. In addition, (7.9) requires careful tuning of the kernel width $\sigma$.

### 7.1.2 Discussion

Since the literature on AD from an information point of view is limited [125], we want to stress the relevance of $\zeta$ and its role in some widely adopted detectors, some of which are detailed in Chapter 1. Considering (7.6), we only select the second term, i.e., normal signal differential entropy, to be minimized. As already stated, minimizing the differential entropy, under the Gaussian assumption, is equivalent to minimizing the Euclidean norm employed in SVDD [143] and its deep counterpart Deep SVDD. The authors in [144, 101] also embed the signal in components with lower variance, i.e., lower differential entropy, and show that it favors the detection of anomalies. The concept of differential entropy minimization aligns with the conclusions of the previous chapter, where components with the lowest variance proved to be the most informative for detection. A hint about the relevance of the first term of (7.6) can be found in [127] where the authors, besides minimizing normal signal differential entropy, maximize the differential entropy of the anomalous data to develop a semi-supervised version of Deep SVDD. Similarly, the third term of (7.6) plays a key role in OCSVM [130] in which the separation between normal and anomalous data samples is maximized.

The loss function in (7.9) is identical to the function in (6.31), utilized for the end-to-end image compression and detection of the previous chapter, as both come from the concept of SAE. In this chapter, however, the use of SAE is supported by the broader theoretical framework of differential entropy minimization, with SAE being a specific instance that leverages this principle.

## 7.2 Numerical evidence

We investigate how AE, RAE, and SAE administer the trade-off between reconstruction and detection performance in two distinct scenarios: ECG time series data and CIFAR-10 image data.

In particular, to perform AD in the latent representation $\mathbf{y}$, we employ two widely adopted unsupervised detectors: Mahalanobis Distance (MD) and OCSVM. The detection

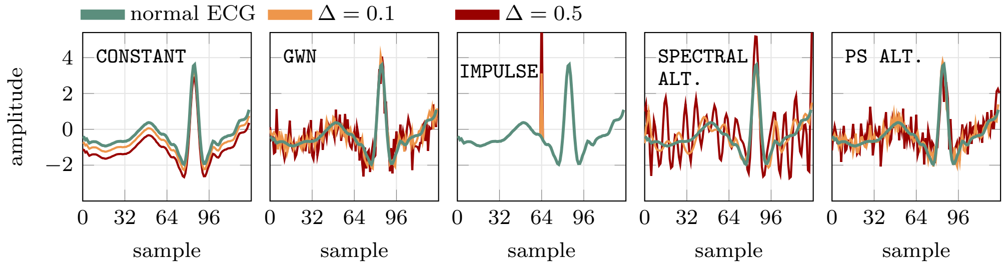
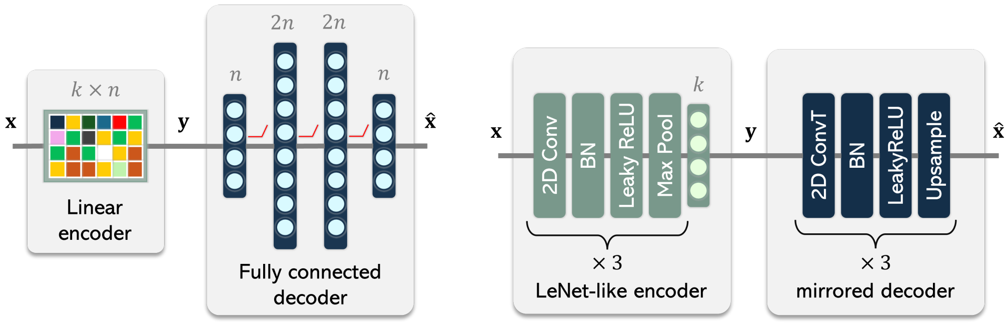


FIGURE 7.2: Examples of anomalies characterized by two different values of $\Delta$ injected into a window of the ECG signal.



(A) AE architecture for ECG signal.

(B) AE architecture for CIFAR-10 images.

FIGURE 7.3: AE architectures.

performance is measured directly from their anomaly scores, using the probability of correct detection ($P_D$), as defined in Chapter 3:

$$P_D = \begin{cases} \mathrm{AUC} & \text{if } \mathrm{AUC} \geq 0.5 \\ 1 - \mathrm{AUC} & \text{if } \mathrm{AUC} < 0.5, \end{cases} \tag{7.11}$$

where AUC represents the Area Under the Curve of the Receiver Operating Characteristic [45].

### 7.2.1 ECG

ECG signals were generated using a realistic synthetic generator[2] [102]. The setup, based on [96], includes heart rates uniformly distributed between 60-100 bpm, a sampling rate of 256 sps, and white noise was injected, guaranteeing a signal-to-noise ratio (SNR) of 40 dB. To train and validate the AEs, we generated $6.4 \times 10^5$ windows, each containing 128 samples. Two additional sets of $10^5$ and $10^4$ windows were created to train the detectors and evaluate the performance of both reconstruction and detection, respectively.

The reconstruction quality of the AEs is evaluated considering the Reconstruction Signal-to-Noise Ratio (RSNR) defined as:

$$\mathrm{RSNR} = \mathbf{E}\left[\frac{\|\mathbf{x}^{\mathrm{ok}}\|_2}{\|\mathbf{x}^{\mathrm{ok}} - \hat{\mathbf{x}}^{\mathrm{ok}}\|_2}\right]_{\mathrm{dB}}. \tag{7.12}$$

[2] MATLAB and C codes are freely available at the Physionet website https://physionet.org/content/ecgsyn/1.0.0/

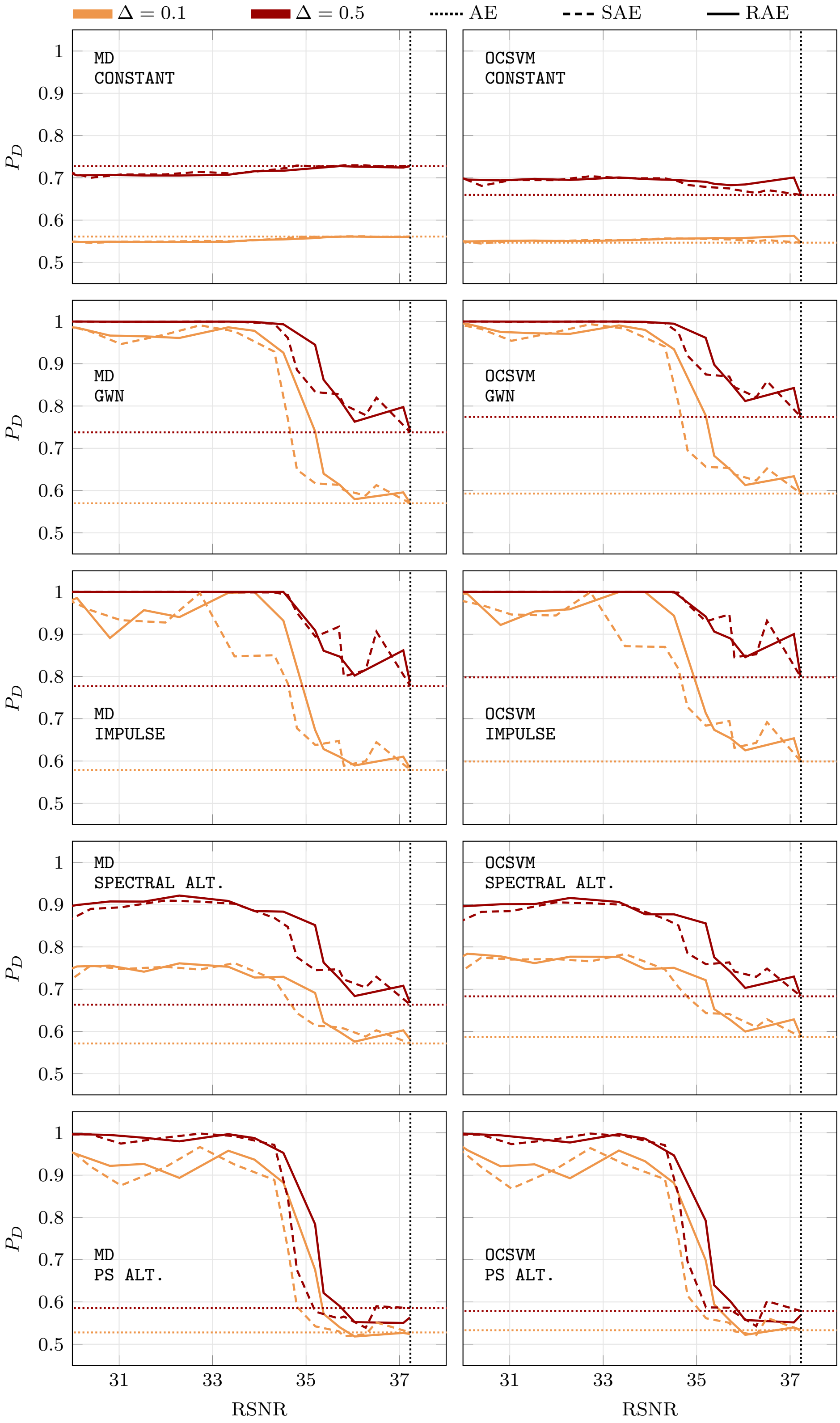


FIGURE 7.4: AD-reconstruction trade-off represented in terms of $P_D$-RSNR curves for ECG signal administered by AE, SAE and RAE for five anomalies with two levels of intensity $\Delta$.

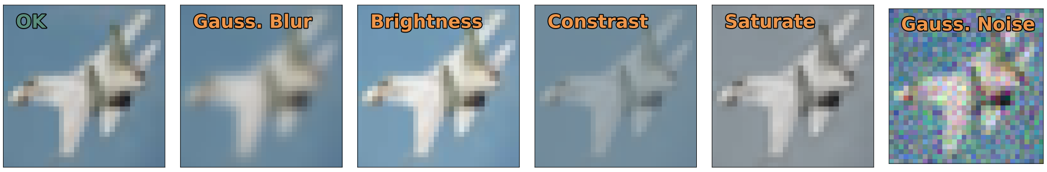


FIGURE 7.5: Examples of CIFAR-10 anomalies with severity $s = 1$.

As discussed in Chapter 3, injecting synthetic anomalies into signals representative of normal behavior can mitigate the lack of real anomaly data. From the perturbations introduced in Chapter 3, we selected three anomalies commonly associated with system faults: constant, impulse, and Gaussian white noise (GWN). Two additional anomalies were added to emulate changes in the sensed phenomenon: spectral alteration and principal subspace (PS) alteration. These anomalies were injected in each window of the normal test set, with intensity measured using the deviation metric from (3.3), recalled here for convenience:

$$\Delta = \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{x}^{\mathrm{ok}} - \mathbf{x}^{\mathrm{ko}}\right\|^2\right]. \tag{7.13}$$

Figure 7.2 illustrates examples of these anomalies, each sub-figure shows a different anomaly type, representing the original signal alongside the altered signals at two different deviation levels $\Delta$. As well established in Chapter 3, higher values of $\Delta$ result in more intense anomalies.

All AEs used an asymmetric neural network structure with a lightweight encoder $(\mathrm{ENC})$ and a more computationally expensive decoder $(\mathrm{DEC})$. Specifically, as illustrated in Figure 7.3-$(\mathrm{A})$, $\mathrm{ENC}$ applies an affine transformation to the $128$-dimensional input, generating a $20$-dimensional output. The decoder $\mathrm{DEC}$, processes the $20$-dimensional input through three hidden layers with $(128,\ 256,\ 256)$ units using ReLU activation functions, and ends with a linear output layer.

The training was performed using the Adam optimizer [75], with a batch size of $128$ and an initial learning rate of $0.001$, which was reduced each time the loss plateaued for $20$ epochs. For the RAE, the kernel width was set to $\sigma = 0.5$, selected after tuning on a validation set.

In Figure 7.4 we present the detection-reconstruction trade-offs for ECG signals. Each row corresponds to a different type of anomaly, while each column refers to a different detector. The anomaly intensities are indicated using different colors. Within each plot, we show the $P_D$-$\mathrm{RSNR}$ curves for SAE (dashed line) and RAE (solid line), generated by acting on the weight parameter $\beta$. The standard AE performance (dotted line) is also included for comparison.

As the results suggest, both RAE and SAE enhance detection performance compared to AE in nearly all configurations. This improvement is observed with both the OCSVM and MD detectors, but it comes at the cost of reduced reconstruction performance: the higher the $P_D$ improvement, the greater the loss in $\mathrm{RSNR}$. The nature of the detector or regularization method has minimal impact since the dashed and solid lines almost overlap. Across all types of anomalies, the curves across the columns (detectors) are quite similar. Significant detection improvements are achieved for anomalies such as GWN, impulse, spectral alteration, and PS alteration. In contrast, performance for the constant anomaly presents limited or no improvement. In most cases, the best $P_D$ gains are achieved with a reduction in $\mathrm{RSNR}$ that does not exceed $2\,\mathrm{dB}$. For example, in the case of PS alteration, both OCSVM and MD detectors achieve $P_D$ of $53\%$ (for $\Delta = 0.1$) and $59\%$ (for $\Delta = 0.5$) using the standard AE, but this increases by $40\%$ when one between SAE or RAE is used.

### 7.2.2 CIFAR-10

For the image data, we focus on the CIFAR-10 dataset [79], which contains $6 \times 10^5$ color images of size $32 \times 32$, evenly distributed across $10$ classes. To train the AEs and detectors, we select $5 \times 10^5$ images, while the remaining $10^4$ images are dedicated to performance evaluation. We chose five common image corruptions as anomalies: Gaussian blur (GB), brightness, contrast, saturation, and Gaussian noise (GN), with severity levels $s = \{1, 3\}$ as described in [60]. These corruptions have previously been used as benchmarks for AD performance [125]. Figure 7.5 shows examples of a normal image and its anomalous versions with severity $s = 1$.

For the AEs, we adopt a symmetric convolutional encoder-decoder architecture from [126], shown in Figure 7.3-$(\mathrm{B})$. The encoder $\mathrm{ENC}$ maps a $32 \times 32 \times 3$ image into a $128$-dimensional latent vector through three blocks, each comprising a sequence of 2D convolution, batch normalization, leaky ReLU activation, and max-pooling. These convolution layers contain $32$, $64$, and $128$ filters, respectively, all having stride $1$ and kernel size $5$. The decoder $\mathrm{DEC}$ mirrors the encoder, replacing convolutions with transposed convolutions and adopting upsampling instead of max-pooling.

Training is performed using the Adam optimizer with the $\ell_2$ weight regularization set to $10^{-6}$, a batch size of $200$, and an initial learning rate of $10^{-4}$, which is reduced when the loss reaches a plateau for the $20$ epochs. For RAE, the kernel width $\sigma$ is set to $1$.

To evaluate the reconstruction quality of the AEs we consider the Peak Signal-to-Noise Ratio (PSNR) defined as:

$$\mathrm{PSNR} = \mathbf{E}\left[\frac{\max(\mathbf{x}^{\mathrm{ok}})}{\|\mathbf{x}^{\mathrm{ok}} - \hat{\mathbf{x}}^{\mathrm{ok}}\|_2}\right]_{\mathrm{dB}} \tag{7.14}$$

The trade-off between the detection and reconstruction performance observed with the ECG data is validated for CIFAR-10 in Figure 7.6. However, in contrast to the ECG case, here the choice of the regularization method has a greater impact. SAE performs better for anomalies like brightness, saturation, and Gaussian noise. For the latter, specifically with the MD detector, SAE leads to a $P_D$ improvement of $24\%$ for $s = 1$ and $43\%$ for $s = 3$. On the other hand, RAE outperforms SAE in detecting Gaussian blur and contrast anomalies, in particular when paired with the OCSVM detector. In these cases, RAE yields a $15\%$ up to $30\%$ boost in $P_D$ while reducing $\mathrm{PSNR}$ by less than $0.5\,\mathrm{dB}$. However, for Gaussian blur and contrast anomalies, the MD detector used with AE already performs well, and regularization fails to improve it further.

SAE shows more consistent improvements across all configurations as $\beta$ increases. Finally, the relatively small difference between RAE and SAE performances with OCSVM and MD detectors indicates that choosing the detector is more critical when working with the standard AE than with the regularized versions. This suggests that carefully designing the compression phase could reduce the need for extensive detector tuning.

## 7.3 Conclusion

In this chapter, we considered monitoring systems where an autoencoder compresses sensor data and the compressed data is either reconstructed or inspected for anomaly detection. Although autoencoders are optimized for reconstruction, they require regularization to improve anomaly detection. Starting from the information-theoretic metric

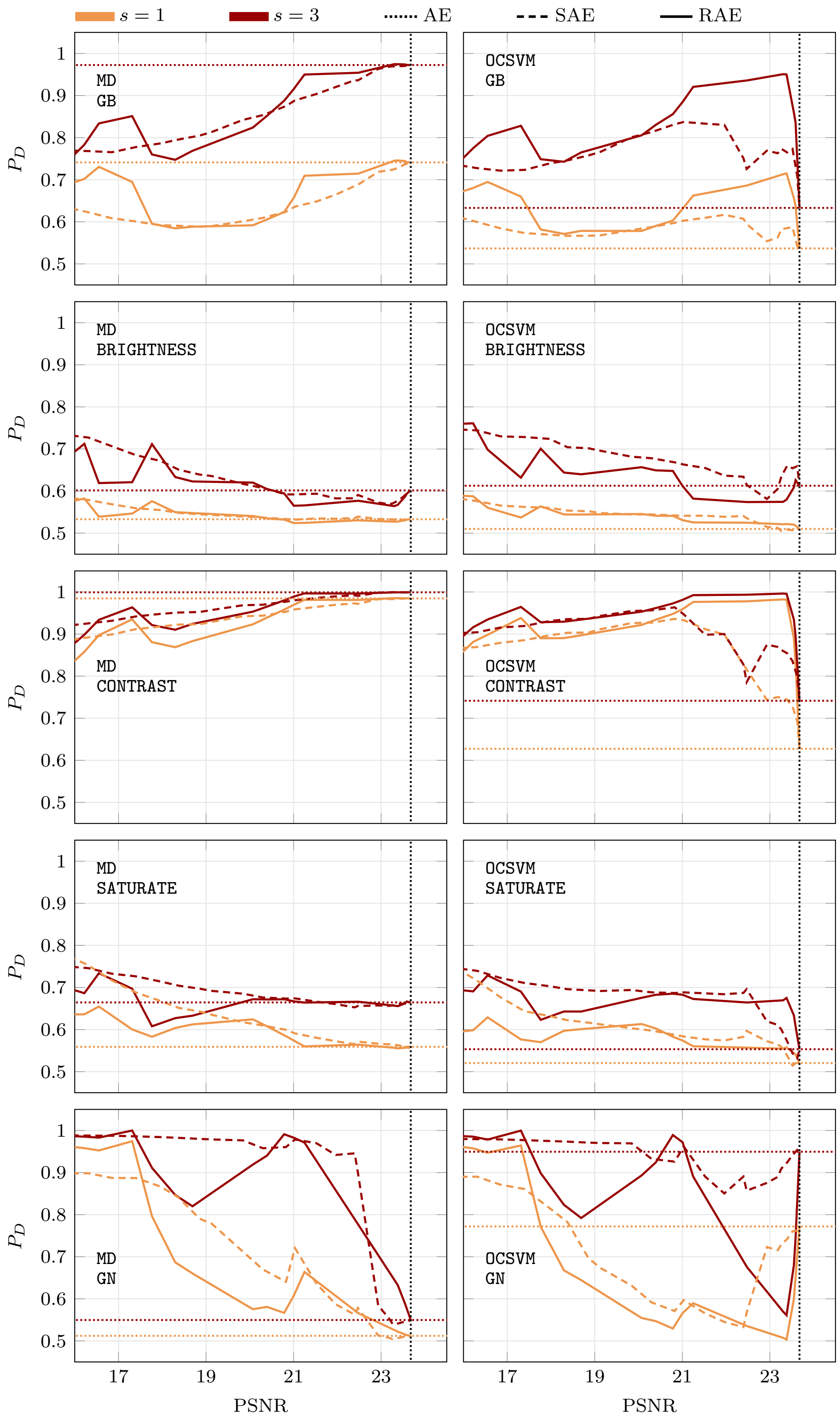


FIGURE 7.6: AD-reconstruction trade-off represented in terms of $P_D$-PSNR curves for CIFAR-10 images administered by AE, SAE, and RAE for five anomalies with two levels of intensity $s$.

$\zeta$, we show that effective regularization should aim to reduce the differential entropy of the compressed representation of the normal signal. We then propose two differential entropy-based regularization methods, SAE and RAE, and analyze how they address the trade-off between detection accuracy and reconstruction quality.

Our investigation considered ECG time series and CIFAR-10 images, adopting two detectors to characterize five common anomalies at different intensities. The experiments showed that AE regularization brings significant improvements in anomaly detection at the cost of an acceptable degradation in reconstruction. Specifically, for the ECG data, both SAE and RAE achieved a $20\%$ improvement in the detection probability $(P_D)$, with a corresponding $2\,\mathrm{dB}$ drop in reconstruction quality $(\mathrm{RSNR})$ across four out of five anomalies. RAE showed a slight advantage by achieving the same $P_D$ improvement with a smaller decrease in $\mathrm{RSNR}$. In the case of CIFAR-10 images, SAE proved more effective, outperforming RAE in detecting three out of five anomalies.

Overall, SAE offers consistent detection improvements (up to $43\%$ in $P_D$) across various scenarios. RAE, while potentially more effective in specific cases, requires a more complex parameter tuning. Additionally, using either regularization method, the choice of the detector becomes less important.

# Conclusion

This dissertation addressed two critical challenges in anomaly detection: the performance assessment of anomaly detectors and the impact of lossy compression on anomaly detection performance. First, we provided a framework for both practical and theoretical performance assessment of anomaly detectors. Second, we designed a framework for the joint analysis of rate, distortion, and distinguishability, offering insights into optimizing compressors for downstream anomaly detection tasks.

For practical assessment, we developed a framework named WOMBATS, which integrates all essential modules for detector evaluation. It includes a generator of synthetic anomalies with multiple signatures and controlled intensity levels, providing a systematic, flexible, consistent, and comprehensive tool for evaluating anomaly detectors in real-world monitoring scenarios.

From a theoretical perspective, we presented a framework based on novel information-theoretic measures, termed distinguishability, applicable in both anomaly-agnostic and anomaly-aware settings. These measures represent manageable quantities in theoretical analysis and effectively approximate practical performance metrics. Moreover, we introduced a Gaussian-based anomaly model, highlighting the white anomaly—a reference anomaly that represents both the average and typical behavior across all possible anomalies.

These assessment frameworks were leveraged to analyze how compression affects anomaly detection performance. Our theoretical and numerical results show that compressors optimized solely for the rate-distortion trade-off can compromise the effectiveness of anomaly-agnostic detectors. To address this issue, we develop the rate-distortion-distinguishability (RDD) framework that optimizes three key quantities: rate, distortion, and distinguishability. Specifically, we formulated two optimization problems that provide insights into designing compressors that balance these three factors, ensuring effective anomaly detection while maintaining efficient compression.

Finally, we designed an autoencoder-based dimensionality reduction scheme that preserves information not only for reconstruction but also for anomaly detection from the compressed representation. By incorporating differential entropy-based regularization techniques, we showed that anomaly detection performance can be significantly enhanced while maintaining an acceptable trade-off with reconstruction quality.

### Limitations and future developments

The proposed methodologies tackle both practical and theoretical challenges in anomaly detection performance assessment and compression design. While effective, they still present several limitations that open opportunities for future research.

A key limitation of WOMBATS is that it currently operates on univariate time series. Extending this framework to multivariate time series would enable the evaluation of detectors exploiting dependencies across multiple signals. Additionally, WOMBATS primarily assesses detectors that operate on windows of signal samples. Future work could explore approaches to evaluate detectors that process one time series sample at a time.

While the designed synthetic anomalies have proven useful for detector benchmarking at testing time, they could also be leveraged for self-supervised learning during training. This approach could potentially provide detectors that outperform fully unsupervised approaches. Similar techniques have been explored in image-based anomaly detection, where synthetic anomalies were used to design state-of-the-art detection methods [114, 158, 162].

Supervision could also help mitigate the performance degradation of anomaly-agnostic detectors when working with compressed data. As observed in our analysis, anomaly-aware detectors do not suffer from the same issue. This suggests the need for compression schemes that explicitly support supervised discriminators on the receiver side, ensuring that detection remains reliable despite compression.

Our RDD framework is currently specialized for Gaussian sources but can be extended to non-Gaussian ones. For instance, aligning with rate-distortion-perception theory [18], and adapting it to Bernoulli processes could provide valuable insights for image data. While the Gaussian assumption limits RDD's applicability, prior research has shown that deep feature extraction networks often produce representations that approximately follow Gaussian distributions [121]. This suggests that a hybrid approach, where a neural network extracts Gaussian-like features before applying our RDD framework for quantization, could bridge the gap between theory and real-world applications.

Most importantly, further research is needed to explore real-world scenarios where the interplay between rate, distortion, and distinguishability can be studied in depth. This would help refine our framework and guide the development of more practical solutions for anomaly detection in monitoring applications. Moreover, the insights provided by the RDD framework should be leveraged to design new compression pipelines or enhance existing ones, aligning anomaly detection-oriented compression with state-of-the-art methods that target reconstruction quality [28, 31, 91].

# Appendix A

# Synthetic Anomalies for Performance Assessment in Time Series

## A.1 $P_D$ vs. $\Delta$

This appendix section presents a set of figures that provide complementary results, further reinforcing the effectiveness of the proposed approach for systematically assessing anomaly detectors introduced in Chapter 3. These results are examined for the two considered use cases: Electrocardiogram (ECG) data and accelerometer telemetry from a structural health monitoring system (ACC).

Figures A.1-A.6 illustrate the performance of the considered anomaly detectors, including Principal Component Analysis (PCA)-based, Machine Learning (ML)-based, Gaussian distribution (GD)-based, and feature-based methods, by evaluating $P_D$ w.r.t the deviation parameter $\Delta$. These figures represent an extension to the results presented in Tables 3.4-3.7, where $P_D$ values were reported for two specific deviation levels, $\Delta = \{0.05, 0.8\}$. Each figure focuses on a different anomaly class and on a different category of the normal signal. Figures A.1-A.3 show the detector performance for power-increasing, power-invariant, and power-decreasing anomalies in ECG signals, whereas Figures A.4-A.6 depict the results for ACC signals.

Figure A.1 presents the detection performance, measured in terms of $P_D$, for power-increasing anomalies in ECG signals as a function of deviation $\Delta$. Each row corresponds to a different anomaly, while each column focuses on a different family of detectors. As expected, detection performance improves with increasing $\Delta$ across all cases.

Among the anomalies, the constant anomaly is the most difficult to detect, particularly for low values of $\Delta$, whereas Gaussian White Noise (GWN) is the easiest to identify. No single detector consistently outperforms the others across all anomalies and deviation levels. For instance, in the first column, SPE$_{59}$ achieves the best performance for GWN, Gaussian Narrowband Noise (GNN), and impulse while performing at an average level for constant and step anomalies. Conversely, T$^2_{52}$ is the most effective for constant and step but performs worst for GWN, GNN, and impulse.

On average, detectors such as SPE$_{59}$, LOF$_5$, and MD tend to outperform other detectors within their respective families. Feature-based detectors exhibit high variability, as their performance is strongly influenced by the specific type of anomaly.

Figure A.2 presents the detection performance, measured in terms of $P_D$, for power-invariant anomalies in ECG signals w.r.t deviation $\Delta$. Each row refers to a different anomaly, while each column corresponds to a different class of detectors. As expected, detection performance generally improves with increasing $\Delta$. However, monotonicity is not strictly maintained in certain cases—specifically, when GD-based detectors encounter

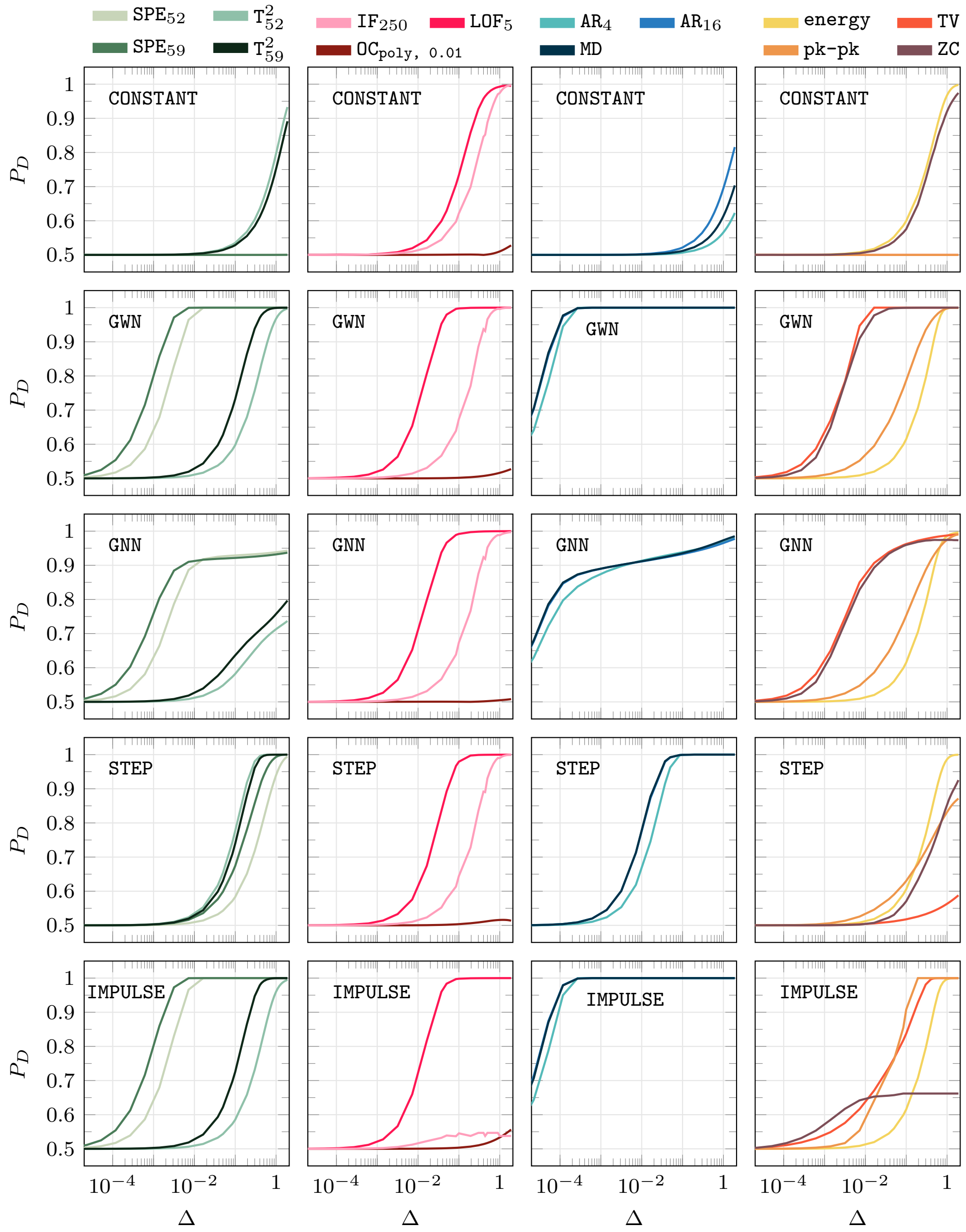


FIGURE A.1: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ECG power-increasing anomalies.

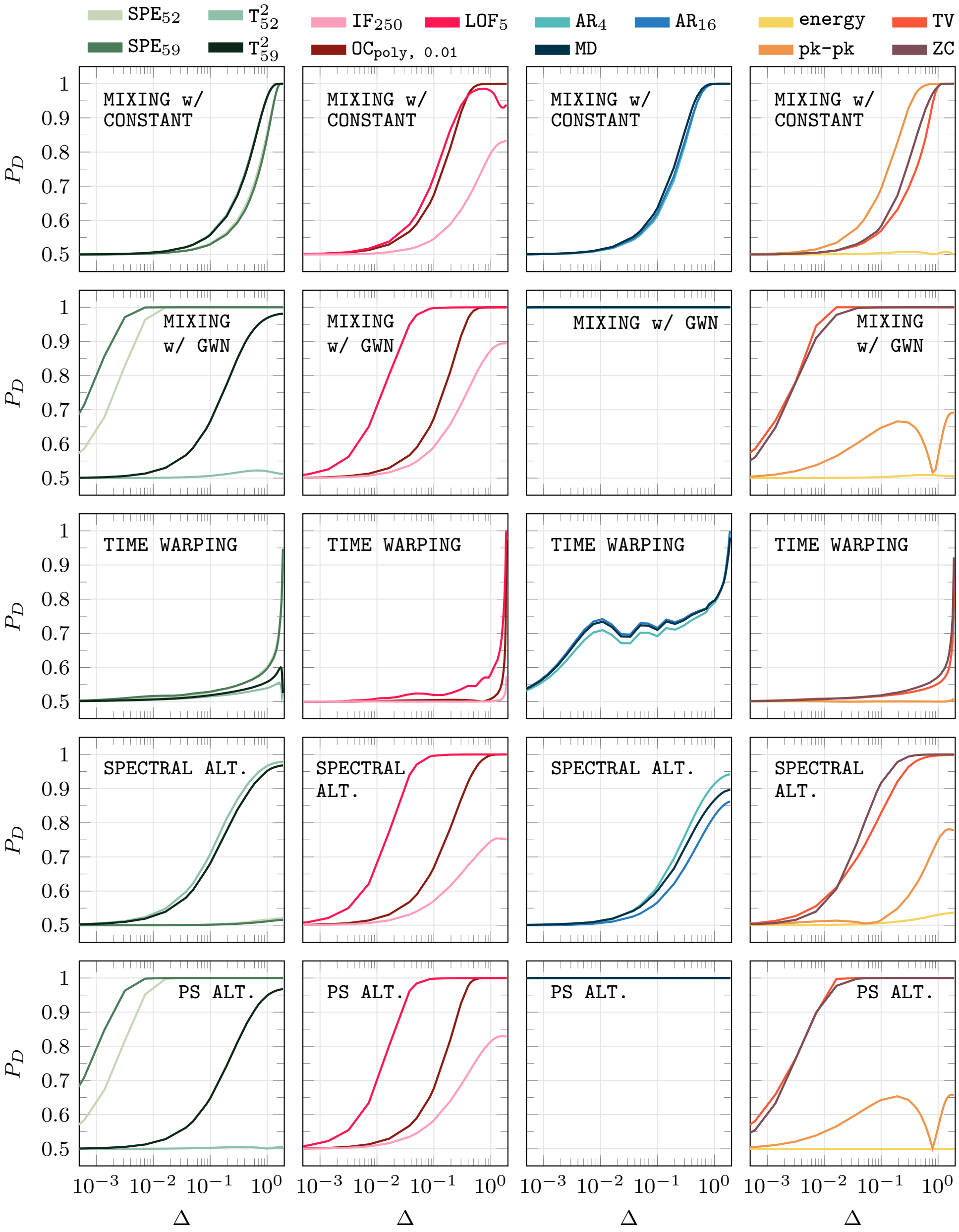


FIGURE A.2: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ECG power-invariant anomalies.
]

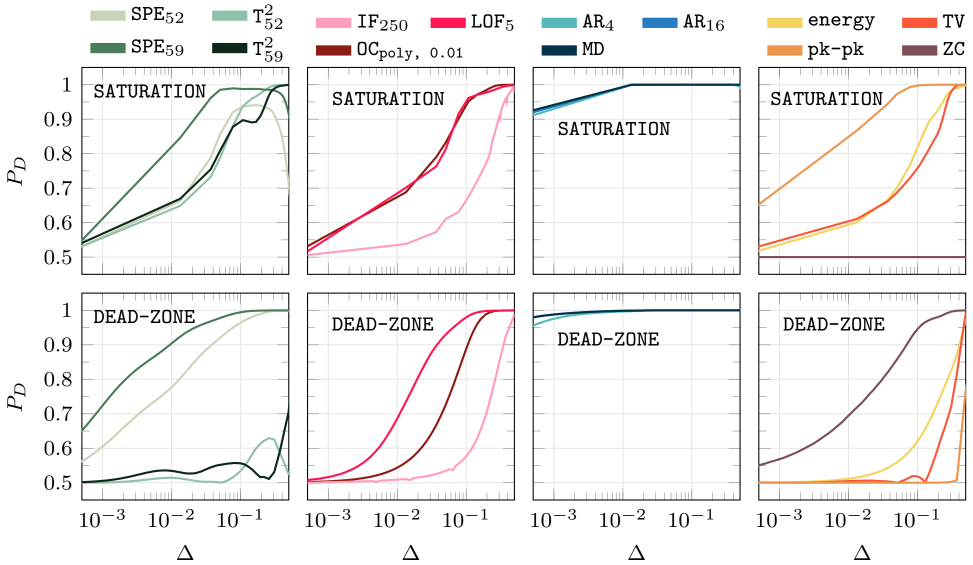


FIGURE A.3: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ECG power-decreasing anomalies.

a time warping anomaly and when the pk-pk detector faces the principal subspace (PS) alteration and mixing with GWN anomalies. As already noted, these two anomalies pose similar challenges for most detectors.

Among the anomalies, mixing with GWN and PS alteration are the easiest to detect, particularly for GD-based detectors, whereas time warping remains the most problematic, especially at low $\Delta$ values. As in the previous case, no single detector consistently outperforms the others across all anomalies and deviation levels. For instance, T$^2_{52}$ completely fails at detecting PS alteration, time warping, and mixing with GWN but surpasses other PCA-based detectors in identifying the spectral alteration.

On average, LOF$_5$ and MD outperform other detectors within their respective families. Regarding feature-based detectors, as expected, the energy-based detector struggles with power-invariant anomalies, while ZC tends to achieve the best performance across this category.

Figure A.3 illustrates the detection performance, measured in terms of $P_D$, for power-decreasing anomalies in ECG signals as a function of deviation $\Delta$. Each row focuses on a different anomaly, while each column corresponds to a different family of detectors. Once again, detection performance generally improves with increasing $\Delta$, though some PCA-based detectors do not exhibit strict monotonicity.

All GD-based detectors perform similarly and consistently outperform other detector families for both anomalies. Among feature-based detectors, the ZC-based detector completely fails at detecting saturation anomalies but achieves the best performance for the dead-zone anomaly. Conversely, the pk-pk detector struggles with the dead-zone anomaly but is the most effective among feature-based detectors for saturation anomalies.

Figure A.4 shows the detection performance, measured in terms of $P_D$, for power-increasing anomalies in ACC signals w.r.t deviation $\Delta$. Each row corresponds to a different anomaly, while each column illustrates the performance for different family of detectors. As expected, detection performance improves with increasing $\Delta$ across all cases.

Unlike the ECG results shown in Figure A.1, the constant and step anomalies exhibit a similar level of difficulty for the detectors compared to the other anomalies. Both T$^2_{16}$

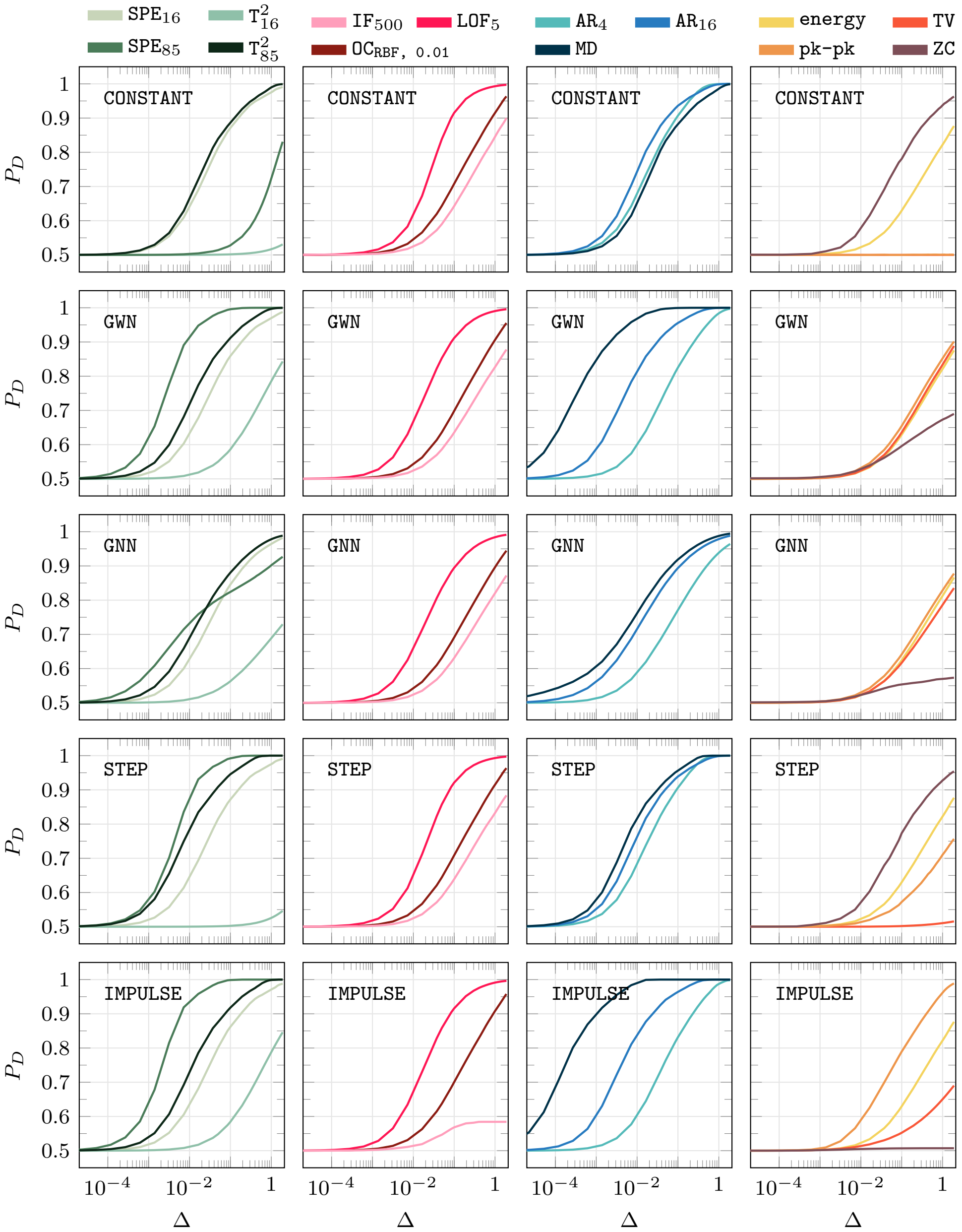


FIGURE A.4: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ACC power-increasing anomalies.

and TV detectors completely fail at detecting these two anomalies. While the ZC detector performs well for constant and step, it fails for the impulse anomaly.

Figure A.5 presents the detection performance, measured in terms of $P_D$, for power-invariant anomalies in ACC signals as a function of deviation $\Delta$. Each row corresponds to a different anomaly, while each column focuses on a different family of detectors. As in previous cases, detection performance generally improves with increasing $\Delta$. However, the time warping and spectral alteration anomalies remain particularly challenging to detect, even at high $\Delta$ values. Feature-based detectors perform relatively well only for the mixing with constant anomaly but struggle with the others. Note that, unlike the ECG case in Figure A.2, the mixing with GWN and PS alteration anomalies exhibit different levels of difficulty for the detectors.

Figure A.6 presents the detection performance, measured in terms of $P_D$, for power-decreasing anomalies in ACC signals as a function of deviation $\Delta$. The two rows consider the two types of power-decreasing anomalies, while each column represents a different family of detectors. The ZC-based detector fails at detecting saturation anomalies but performs best among feature-based detectors for the dead-zone anomaly. Conversely, the pk-pk detector struggles with dead-zone anomalies but it outperforms other feature-based detector for saturation anomalies.

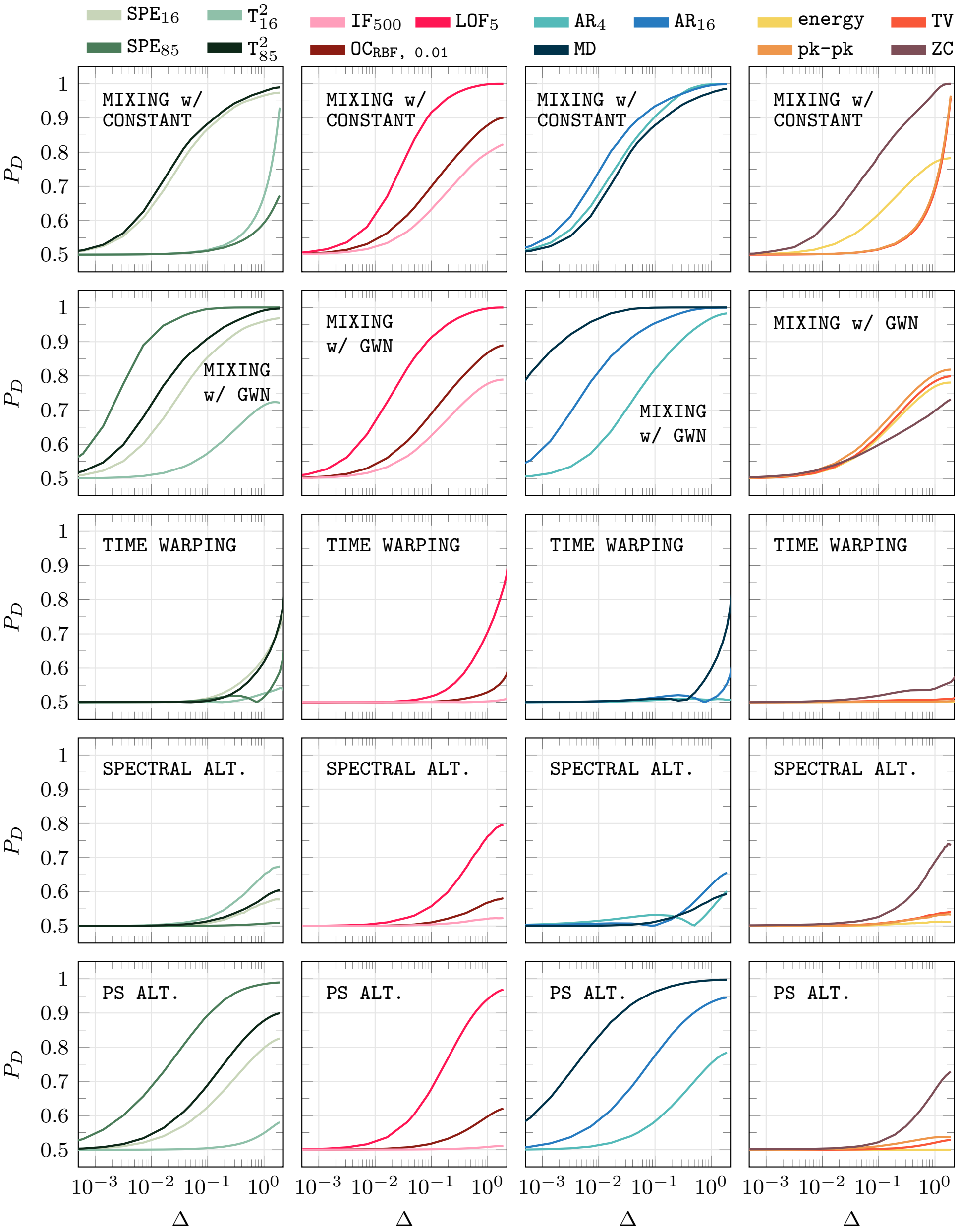


FIGURE A.5: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ACC power-invariant anomalies.

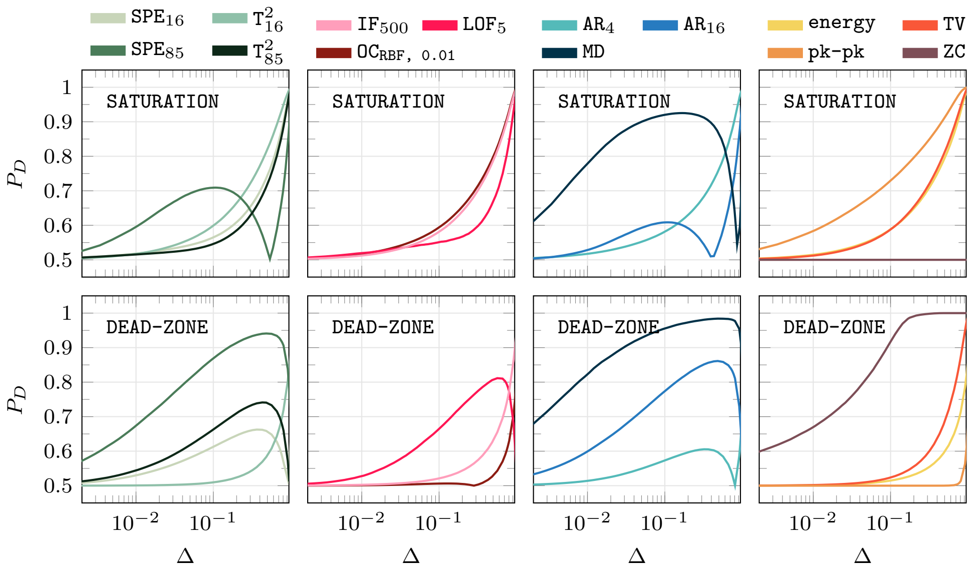


FIGURE A.6: Performance in terms of $P_D$ of four families of detectors (PCA-based, ML-based, GD-based, feature-based) against deviation $\Delta$ in case of ACC power-decreasing anomalies.

## A.2 Anomalies Mixtures

Different types of anomalies can be combined to simulate real-world anomalies that manifest through multiple superimposed effects. Referring to the general anomaly model introduced in equation (3.2) of Chapter 3:

$$\mathbf{x}^{\mathrm{ko}} = c\left(\mathbf{x}^{\mathrm{ok}}\right) + \mathbf{d}, \tag{A.1}$$

as previously discussed, anomalies can arise either from an alteration of the normal signal through $c(\cdot)$ or by adding an independent disturbance signal $\mathbf{d}$ to the normal data.

When combining two different types of anomalies, we can have three cases: intra-category combinations, where either two $\mathbf{d}$-like anomalies or two $c(\cdot)$-like anomalies are superimposed, and an inter-category combination of one $c(\cdot)$ and one $\mathbf{d}$-like anomalies. While intra-category combinations can be handled similarly, inter-category cases require special attention. For illustration, we focus on two examples: *i*) the mixture of constant (power-increasing) and saturation (power-decreasing) anomalies, and *ii*) the combination of spectral alteration and PS subspace alteration power-invariant anomalies.

### A.2.1 Constant + saturation

Based on the model in (A.1), to combine constant and saturation anomalies, the normal signal window $\mathbf{x}^{\mathrm{ok}}$ should first be altered by applying a saturation anomaly, followed by the application of a constant anomaly. To achieve this, it is sufficient to set:

$$c\left(x_j^{\mathrm{ok}}\right) = \begin{cases} x_{\mathrm{SAT}}\mathsf{sign}\left(x_j^{\mathrm{ok}}\right) & \text{if } \left|x_j^{\mathrm{ok}}\right| > x_{\mathrm{SAT}} \\ x_j^{\mathrm{ok}} & \text{otherwise} \end{cases} \tag{A.2}$$

$$d_j = \pm a \quad \text{for} \quad j = 0, \dots, n-1. \tag{A.3}$$

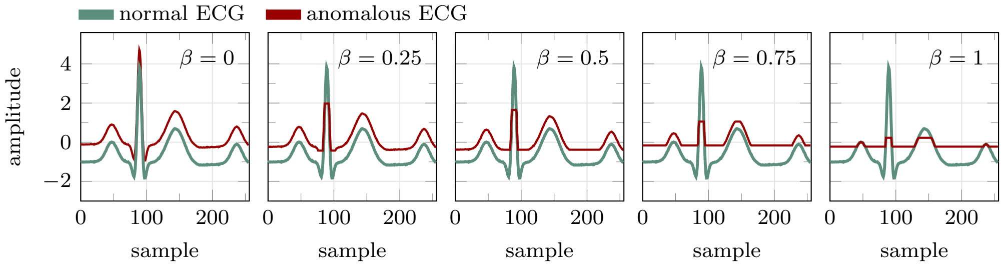


FIGURE A.7: Examples of saturation-constant mixture anomaly in case of ECG signal for five different values of weight $\beta = \{0, 0.25, 0.5, 0.75, 1\}$ increasing from left to right.

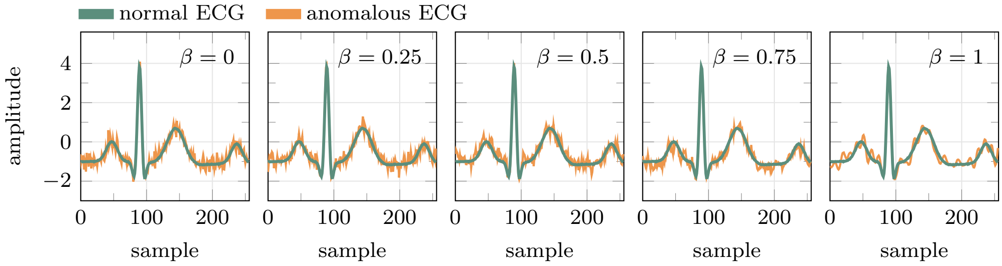


FIGURE A.8: Examples of spectral alteration-PS alteration mixture anomaly in case of ECG signal for five different values of weight $\beta = \{0, 0.25, 0.5, 0.75, 1\}$ increasing from left to right.

Parameters $x_{\text{SAT}}$ and $a$ are selected to satisfy a specified level of deviation

$$\Delta = \beta\Delta + (1-\beta)\,\Delta = \Delta_{\text{SAT}} + \Delta_{\text{CONST}} \tag{A.4}$$

where $\beta$ is a weighting parameter that controls the relative intensity of the two anomalies and

$$\Delta_{\text{SAT}} = \beta\Delta = \sum_{|x_j^{\text{ok}}|>x_{\text{SAT}}} \left[x_j^{\text{ok}} - x_{\text{SAT}}\text{sign}\left(x_j^{\text{ok}}\right)\right]^2 \tag{A.5}$$

$$\Delta_{\text{CONST}} = (1-\beta)\,\Delta = a^2 \tag{A.6}$$

The equation (A.4) holds since saturation and constant anomalies are independent.

We provide examples of the saturation-constant anomaly mixture applied to an ECG signal, with $\Delta = 0.8$ and five different values of $\beta$, in Figure A.7. The transition from a pure constant anomaly ($\beta = 0$) to a pure saturation anomaly ($\beta = 1$) is clearly visible. For intermediate $\beta$ values, the anomalies blend, resulting in a weighted combination of the two single anomaly types.

It is important to note that this example can be generalized to any other mixture where the first anomaly distorts $\mathbf{x}^{\text{ok}}$ through $c(\cdot)$, e.g., spectral alteration, dead-zone, and the second anomaly is caused by a disturbance $\mathbf{d}$, e.g., GWN, GNN; or, alternatively, to a mixture where both anomalies arise from two different disturbances $\mathbf{d}$.

### A.2.2 Sprectral alteration + PS alteration

To create a mixture of spectral alteration and PS alteration anomalies, starting from (A.1), it is enough to set $c(\cdot)$ as a composition of spectral alteration and PS alteration anomalies and $\mathbf{d} = \mathbf{0}$:

$$c\left(\mathbf{x}^{\mathrm{ok}}\right) = \mathbf{C}\mathbf{x}^{\mathrm{ok}} = \mathbf{U}^{\mathrm{ko}}\sqrt{\mathbf{\Lambda}^{\mathrm{ko}}}\sqrt{\mathbf{\Lambda}^{\mathrm{ok}}}^{-1}{\mathbf{U}^{\mathrm{ok}}}^{\top}\mathbf{x}^{\mathrm{ok}} \tag{A.7}$$

$$d_j = 0 \quad \text{for} \quad j = 0, \ldots, n-1. \tag{A.8}$$

where $\mathbf{U}^{\mathrm{ko}} = \mathbf{R}_{n,\theta_{\mathrm{PSA}}}\mathbf{U}^{\mathrm{ok}}$ and $\sqrt{\mathbf{\Lambda}^{\mathrm{ko}}}$ is built such that

$$\sqrt{\lambda_j^{\mathrm{ko}}} \quad \text{for} \quad j = 0, \ldots, k \quad \text{are s.t.} \quad \ell^{\mathrm{ko}} = \mathbf{R}_{k,\theta_{\mathrm{SA}}}\ell^{\mathrm{ok}} \tag{A.9}$$

$$\sqrt{\lambda_j^{\mathrm{ko}}} = \sqrt{\lambda_j^{\mathrm{ok}}} \quad \text{for} \quad j = k, \ldots, n-1 \tag{A.10}$$

with $\ell^{\mathrm{ok}} = \frac{1}{\sqrt{n\gamma}}\left(\sqrt{\lambda_0^{\mathrm{ok}}}, \ldots, \sqrt{\lambda_{k-1}^{\mathrm{ok}}}\right)^{\top}$ and $\ell^{\mathrm{ko}} = \frac{1}{\sqrt{n\gamma}}\left(\sqrt{\lambda_0^{\mathrm{ko}}}, \ldots, \sqrt{\lambda_{k-1}^{\mathrm{ko}}}\right)^{\top}$. The angles $\theta_{\mathrm{SA}}$ and $\theta_{\mathrm{PSA}}$ are set to satisfy

$$\Delta = g_\beta\left[\Delta'\right] = g_\beta\left[\beta\Delta' + (1-\beta)\,\Delta'\right] = g_\beta\left[\Delta_{\mathrm{SA}} + \Delta_{\mathrm{PSA}}\right] \tag{A.11}$$

$$\Delta_{\mathrm{SA}} = \beta\Delta' = 2\gamma\left[1 - \cos(\theta_{\mathrm{SA}})\right] \tag{A.12}$$

$$\Delta_{\mathrm{PSA}} = (1-\beta)\,\Delta' = 2\left[1 - \cos(\theta_{\mathrm{PSA}})\right] \tag{A.13}$$

with $g_\beta = 2\left[1 - \frac{1}{n}\mathrm{tr}\mathbf{C}_{\beta,\Delta'}\right]$ the function that links the effective deviation $\Delta$ with the imposed one $\Delta'$.

Examples of the spectral alteration-PS alteration anomaly mixture applied to an ECG signal, with $\Delta = 0.05$ and five different values of $\beta$ are shown in Figure A.8. By acting on $\beta$, it is noticeable how the mixture transitions from a pure PS alteration anomaly ($\beta = 0$) to a pure spectral alteration anomaly ($\beta = 1$). For intermediate $\beta$ values, the resulting anomaly appears as a weighted combination of the two.

Note that, the anomaly resulting from spectral alteration-PS alteration mixture can be further combined with disturbances following a procedure similar to the one illustrated in the previous example.

# Appendix B

# Theoretical Performance Assessment

## B.1 Proof of Theorem 4.1

*Proof of Theorem 4.1.* We will rely on the following Lemma, the proof of which is provided immediately after this one.

**Lemma B.1.** *If $\boldsymbol{\lambda}^{\mathrm{ko}} \sim \mathcal{U}\left(\mathbb{S}^n\right)$, then for any integrable function $f : \mathbb{R} \to \mathbb{R}$ and any $j = 0, \dots, n-1$*

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

To prove our thesis we start by writing $\Delta_{\mathrm{F}} = n^{-\beta} \|\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n\|_F$ as

$$\Delta_{\mathrm{F}} = \frac{1}{n^\beta} \sqrt{\mathrm{tr}\left[\left(\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n\right)^\top \left(\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n\right)\right]}.$$

By noting that $\mathbf{U}^{\mathrm{ko}}$ is orthonormal and thus that $\boldsymbol{\Sigma}^{\mathrm{ko}} - \mathbf{I}_n = \mathbf{U}^{\mathrm{ko}} \left(\boldsymbol{\Lambda}^{\mathrm{ko}} - \mathbf{I}_n\right) \mathbf{U}^{\mathrm{ko}\top}$ we get

$$\begin{aligned}
\Delta_{\mathrm{F}} &= \frac{1}{n^\beta} \sqrt{\mathrm{tr}\left[\mathbf{U}^{\mathrm{ko}} \left(\boldsymbol{\Lambda}^{\mathrm{ko}} - \mathbf{I}_n\right)^2 \mathbf{U}^{\mathrm{ko}\top}\right]} \\
&= \frac{1}{n^\beta} \sqrt{\mathrm{tr}\left[\left(\boldsymbol{\Lambda}^{\mathrm{ko}} - \mathbf{I}_n\right)^2 \mathbf{U}^{\mathrm{ko}\top} \mathbf{U}^{\mathrm{ko}}\right]} \\
&= \frac{1}{n^\beta} \sqrt{\sum_{k=0}^{n-1} \left(\lambda_k^{\mathrm{ko}} - 1\right)^2}.
\end{aligned}$$

Starting from the above expression, we obtain

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

where we have applied Lemma B.1 to compute the expectation.

Hence, when $\beta > 1/2$,

$$\mathbf{E}\left[\Delta_{\mathrm{F}}^2\right] \xrightarrow[n \to \infty]{} 0 \tag{B.1}$$

that can be substituted into the Markov inequality to yield

$$\text{Pr}\left(\Delta_{\mathrm{F}}^2 \geq \bar{\Delta}\right) \leq \frac{\mathbf{E}\left[\Delta_{\mathrm{F}}^2\right]}{\bar{\Delta}} \qquad\qquad \forall\bar{\Delta} > 0$$

and thus that $\Delta_{\mathrm{F}}$ converges to $0$ in probability with increasing $n$. □

*Proof of Lemma B.1.* For any function $f : \mathbb{R} \mapsto \mathbb{R}$ we have

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

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

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

□

## B.2 Proof of Lemma 4.1

*Proof of Lemma 4.1.*

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

where the last summand has been computed as the expectation of a quadratic form in a Gaussian multivariate for which Corollary 3.2b.1 in [117, chapter 3] gives a formula.

□

# Appendix C

# Rate-Distortion Theory and Distinguishability

## C.1 Proof of Lemma 5.1

*Proof of Lemma 5.1.* The distortion is tuned to the normal case, which assumes a memoryless source. Therefore, we can omit time indices and focus on a vector $\mathbf{x}$ with independent components $x_j \sim \mathcal{N}(0, \lambda_j)$ for $j = 0, \ldots, n-1$.

We know from [77] that for a given value of the parameter $\theta$, each component $x_j$ is independently transformed into $\hat{x}_j$. Specifically,

$$\hat{x}_j = \begin{cases} 0 & \text{if } \lambda_j \le \theta \\ x_j + \Delta_j & \text{if } \lambda_j > \theta \end{cases} \tag{C.1}$$

where, to achieve the Shannon lower bound, $\Delta_j$ must be a Gaussian random variable independent of $\hat{x}_j$. Consequently, the three quantities $\hat{x}_j$, $x_j$ and $\Delta_j$ must be such that $(\hat{x}_j, x_j, \Delta_j)^\top \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\hat{x}_j, x_j, \Delta_j}\right)$ with

$$\mathbf{\Sigma}_{\hat{x}_j, x_j, \Delta_j} = \begin{pmatrix} \lambda_j - \theta & \lambda_j - \theta & 0 \\ \lambda_j - \theta & \lambda_j & -\theta \\ 0 & -\theta & \theta \end{pmatrix}. \tag{C.2}$$

That illustrates in which sense $\hat{x}_j$ *encodes* $x_j$. In fact, the non-diagonal elements $\lambda_j - \theta$ are positive, indicating that $\hat{x}_j$ and $x_j$ are positively correlated.

From (C.2), if we accept to identify a Gaussian with $0$ variance with a Dirac's delta we conclude that $\hat{x}_j \sim \mathcal{N}(0, \max\{0, \lambda_j - \theta\})$ and thus $\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}\mathbf{S}_\theta)$.

Moreover, $(\hat{x}_j, x_j)^\top \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\hat{x}_j, x_j}\right)$ where $\mathbf{\Sigma}_{\hat{x}_j, x_j}$ represents the upper-left $2 \times 2$ submatrix of $\mathbf{\Sigma}_{\hat{x}_j, x_j, \Delta_j}$ in (C.2). Assuming that $\theta < \lambda_j$, from the joint probability of $x_j$ and $\hat{x}_j$, we can compute the action of $f_{\hat{\mathbf{x}}|\mathbf{x}}$ on the $j$-th component of $x_j$ as the probability density function (PDF) of $\hat{x}_j$ given $x_j$, i.e.,

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

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 assemble the component-wise PDFs into a vector PDF by employing the matrix $\mathbf{T}_\theta = \mathrm{diag}\,(\tau_0, \dots, \tau_{n-1}) = \min\{\mathbf{I}_n, \theta(\mathbf{\Sigma}^{\mathrm{ok}})^{-1}\}$, and the matrix $\mathbf{S}_\theta = \mathbf{I}_n - \mathbf{T}_\theta$, which ultimately leads us to the conclusion of the thesis. □

## C.2 Proof of Lemma 5.2

*Proof of Lemma 5.2.* The PDF of $\hat{\mathbf{x}}^{\mathrm{ko}}$ distorted through $f^{\mathrm{ok}}_{\hat{\mathbf{x}}|\mathbf{x}}$ can be expressed as

$$f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\alpha) = \int_{\mathbb{R}^n} f^{\mathrm{ko}}_{\hat{\mathbf{x}},\mathbf{x}}(\alpha,\beta)\mathrm{d}\beta = \int_{\mathbb{R}^n} f^{\mathrm{ok}}_{\hat{\mathbf{x}}|\mathbf{x}}(\alpha,\beta) f^{\mathrm{ko}}_{\mathbf{x}}(\beta)\mathrm{d}\beta \tag{C.3}$$

First, consider the low-distortion regime where $\theta < \lambda^{\mathrm{ok}}_{n-1}$ that implies $\mathbf{T}_\theta = \theta(\mathbf{\Sigma}^{\mathrm{ok}})^{-1}$, and write

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

with $\mathbf{Q} = \mathbf{S}_\theta(\mathbf{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta\mathbf{T}_\theta)^{-1}\mathbf{S}_\theta + (\mathbf{\Sigma}^{\mathrm{ko}})^{-1} = (\theta\mathbf{I}_n)^{-1} - (\mathbf{\Sigma}^{\mathrm{ok}})^{-1} + (\mathbf{\Sigma}^{\mathrm{ko}})^{-1}$ and $\mathbf{q} = (\mathbf{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta\mathbf{T}_\theta)^{-1}\mathbf{S}_\theta\alpha = \alpha/\theta$. To compute $g(\alpha)$ let $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ with $\mathbf{D}$ diagonal and $\mathbf{U}$ orthonormal, and set $\beta' = \mathbf{D}^{1/2}\mathbf{U}^\top\beta$ so that $\beta = \mathbf{U}\mathbf{D}^{-1/2}\beta'$ and $\mathrm{d}\beta = {}^{\mathrm{d}\beta'}/_{\sqrt{\det\mathbf{Q}}}$. With this write

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

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

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

Substituting this back into $f^{\mathrm{ok}}_{\hat{\mathbf{x}}}$ we obtain

$$f^{\mathrm{ko}}_{\hat{\mathbf{x}}}(\alpha) = G_{\mathbf{0},[(\theta\mathbf{I}_n)^{-1} - (\mathbf{\Sigma}^{\mathrm{ok}})^{-1} + (\mathbf{\Sigma}^{\mathrm{ko}})^{-1}]\mathbf{\Sigma}^{\mathrm{ko}}\mathbf{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta\mathbf{T}_\theta}\left(\alpha\right).$$

Under the low-distortion assumption, a straightforward expansion of the definitions allows us to simplify and rearrange the covariance matrix into

$$\begin{aligned}
&\left[(\theta\mathbf{I}_n)^{-1} - (\boldsymbol{\Sigma}^{\mathrm{ok}})^{-1} + (\boldsymbol{\Sigma}^{\mathrm{ko}})^{-1}\right] \boldsymbol{\Sigma}^{\mathrm{ko}}\boldsymbol{\Sigma}^{\mathrm{ok}}\mathbf{S}_\theta\mathbf{T}_\theta = \\
&= \left[(\theta\mathbf{I}_n)^{-1} - (\boldsymbol{\Sigma}^{\mathrm{ok}})^{-1} + (\boldsymbol{\Sigma}^{\mathrm{ko}})^{-1}\right] \boldsymbol{\Sigma}^{\mathrm{ko}}\boldsymbol{\Sigma}^{\mathrm{ok}}\theta(\boldsymbol{\Sigma}^{\mathrm{ok}})^{-1}\mathbf{S}_\theta \\
&= \left[\boldsymbol{\Sigma}^{\mathrm{ko}} - \theta(\boldsymbol{\Sigma}^{\mathrm{ok}})^{-1}\boldsymbol{\Sigma}^{\mathrm{ko}} + \theta\mathbf{I}_n\right]\mathbf{S}_\theta \\
&= \left[\mathbf{I}_n - \theta(\boldsymbol{\Sigma}^{\mathrm{ok}})^{-1}\right]\boldsymbol{\Sigma}^{\mathrm{ko}}\mathbf{S}_\theta + \theta\mathbf{S}_\theta \\
&= \mathbf{S}_\theta\boldsymbol{\Sigma}^{\mathrm{ko}}\mathbf{S}_\theta + \theta\mathbf{S}_\theta && \text{(C.4)}
\end{aligned}$$

as in the statement of the lemma.

To handle the case where $\theta$ exceeds $\lambda_{n-1}^{\mathrm{ok}}$, note that as $\theta \to (\lambda_{n-1}^{\mathrm{ok}})^-$, the last diagonal entry of $\mathbf{S}_\theta$ tends to $0$. Consequently, by (C.4), the covariance tends to have zeros in its last row and column. Since a Gaussian with vanishing-variance can be considered a Dirac delta, this model the fact that the last component of both $\mathbf{x}^{\mathrm{ok}}$ and $\mathbf{x}^{\mathrm{ko}}$ is fully distorted and set to $0$. Thus, (C.4) remains valid also for $\lambda_{n-1}^{\mathrm{ok}} < \theta < \lambda_{n-2}^{\mathrm{ok}}$. Yet, analogous considerations can be extended for $\theta \to (\lambda_j^{\mathrm{ok}})^-$ and $j = n-2, n-3, \ldots, 0$ so that (C.4) is valid for any value of $\theta$. □

## C.3 Proof of Theorem 5.1

*Proof of Theorem 5.1.* From (4.17) we have that

$$\zeta_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{k_\theta - 1} A_j(\theta)$$

with

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

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

For the sake of simplicity assume $\lambda_0^{\mathrm{ok}} > \lambda_1^{\mathrm{ok}} > \cdots > \lambda_{n-1}^{\mathrm{ok}} > 0$, set $\lambda_n^{\mathrm{ok}} = 0$, and define $\Theta_j = ]\lambda_{j+1}^{\mathrm{ok}}, \lambda_j^{\mathrm{ok}}[$ for $j = 0, \ldots, n-1$ so that if $\theta \in \Theta_j$ then $k_\theta = j+1$.

As a function of $\theta$, $\zeta_{\mathbf{I}}$ is continuous. In fact, it is trivially continuous in each $\Theta_j$. Yet, it is continuous also at any chosen $\lambda_{\bar{j}}^{\mathrm{ok}}$ with $\bar{j} = 0, \ldots, n-1$. To understand why, note that

$$\begin{aligned}
\lim_{\theta\to\lambda_{\bar{j}}^{\mathrm{ok}-}} \zeta_{\mathbf{I}} &= \frac{1}{2\ln 2}\lim_{\theta\to\lambda_{\bar{j}}^{\mathrm{ok}-}}\sum_{j=0}^{\bar{j}} A_j(\theta) \\
&= \frac{1}{2\ln 2}\lim_{\theta\to\lambda_{\bar{j}}^{\mathrm{ok}-}} A_{\bar{j}}(\theta) + \sum_{j=0}^{\bar{j}-1} A_j(\theta) \\
&= \frac{1}{2\ln 2}\lim_{\theta\to\lambda_{\bar{j}}^{\mathrm{ok}+}}\sum_{j=0}^{\bar{j}-1} A_j(\theta) = \lim_{\theta\to\lambda_{\bar{j}}^{\mathrm{ok}+}}\zeta_{\mathbf{I}}
\end{aligned}$$

where we have utilized the fact that the $A_j(\theta)$ are continuous and thus their left and right limits are equal, and that $A_{\bar{j}}(\lambda_{\bar{j}}^{\mathrm{ok}}) = 0$.

At the leftmost point of the domain, when $\theta = \lambda_n^{\mathrm{ok}} = 0$ (no distortion), we have $k_\theta = n$ and thus

$$\zeta_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{n-1}\left(\frac{1}{\lambda_j^{\mathrm{ok}}} - 1\right) \geq 0$$

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

At the rightmost end of its domain, when $\theta = \lambda_0^{\mathrm{ok}}$ (maximum distortion), we have $k_\theta = 0$ and thus $\zeta_{\mathbf{I}} = 0$. Yet, we also have that

$$\frac{\partial}{\partial\theta}\zeta_{\mathbf{I}} = \frac{1}{2\ln 2}\sum_{j=0}^{k_\theta - 1}\frac{1}{\lambda_j^{\mathrm{ok}}}\left(1 - \frac{1}{\lambda_j^{\mathrm{ok}}}\right)$$

in which the summands are positive when $\lambda_j^{\mathrm{ok}} > 1$. Therefore, if $\bar{k} = \arg\max_k\{\lambda_k^{\mathrm{ok}} \geq 1\}$, for $\theta \geq \lambda_{\bar{k}}^{\mathrm{ok}}$, all the summands in the above expression are positive. This implies that $\frac{\partial}{\partial\theta}\zeta_{\mathbf{I}} > 0$ for $\lambda_{\bar{k}}^{\mathrm{ok}} < \theta \leq \lambda_0^{\mathrm{ok}}$. Given that $\zeta_{\mathbf{I}} = 0$ at the end of that interval, it must be negative in its interior.

We also know that $\zeta_{\mathbf{I}}$ is positive for $\theta = \lambda_n^{\mathrm{ok}} = 0$ and is continuous for $\theta \in ]\lambda_n^{\mathrm{ok}}, \lambda_0^{\mathrm{ok}}[$. Hence, it must pass through zero at least once whenever it is not negative, i.e., for $0 < \theta < \lambda_{\bar{k}}^{\mathrm{ok}}$. □

# Appendix D

# A Theoretical Framework for Rate-Distortion-Distinguishability

## D.1 Proof of Lemma 6.1

*Proof of Lemma 6.1.* Let $\mathcal{H}(\cdot)$ denote the differential entropy of a random vector and $\mathcal{H}(\cdot|\cdot)$ represent the conditional differential entropy between two random vectors. By applying the chain rule, we can express the mutual information as follows:

$$\begin{aligned} \mathcal{I}(\hat{\mathbf{x}}^{\text{ok}};\mathbf{x}^{\text{ok}}) &= \mathcal{H}\left(\mathbf{x}^{\text{ok}}\right) - \mathcal{H}\left(\mathbf{x}^{\text{ok}}|\hat{\mathbf{x}}^{\text{ok}}\right) && \text{(D.1)} \\ &= \mathcal{H}\left(\mathbf{x}^{\text{ok}}\right) - \mathcal{H}\left(\mathbf{x}^{\text{ok}} - \hat{\mathbf{x}}^{\text{ok}}|\hat{\mathbf{x}}^{\text{ok}}\right) && \text{(D.2)} \\ &\geq \mathcal{H}\left(\mathbf{x}^{\text{ok}}\right) - \mathcal{H}\left(\mathbf{x}^{\text{ok}} - \hat{\mathbf{x}}^{\text{ok}}\right) && \text{(D.3)} \end{aligned}$$

where the inequality is given by the property that conditioning reduces entropy.

Let us define $\mathbf{\Delta} = \hat{\mathbf{x}}^{\text{ok}} - \mathbf{x}^{\text{ok}}$. To further relax the bound, we observe that for the covariance matrix $\mathbf{\Sigma}^{\mathbf{\Delta}}$ of $\mathbf{\Delta}$, the entropy $\mathcal{H}\left(-\mathbf{\Delta}\right)$ is maximized when $\mathbf{\Delta}$ is a zero-mean Gaussian vector. This leads us to the following lower bound:

$$\mathcal{I}\left(\hat{\mathbf{x}}^{\text{ok}};\mathbf{x}^{\text{ok}}\right) \geq \mathcal{H}\left(\mathbf{x}^{\text{ok}}\right) - \frac{n}{2}\log_2(2\pi e) - \frac{1}{2}\log_2\left|\mathbf{\Sigma}^{\mathbf{\Delta}}\right| \tag{D.4}$$

Next, we note that the variances $\theta_j$ of the components of $\mathbf{\Delta}$ are specified, meaning the diagonal elements of $\mathbf{\Sigma}^{\mathbf{\Delta}}$ are fixed. According to Hadamard's inequality, the determinant $\left|\mathbf{\Sigma}^{\Delta}\right|$ is maximized when all non-diagonal elements of $\mathbf{\Sigma}^{\mathbf{\Delta}}$ are zero, implying that the components of $\mathbf{\Delta}$ are independent.

All these conditions minimize the lower bound on the rate, which can be achieved through a Gaussian additive encoding with variances $\theta_j$. Therefore, such an encoding yields to the minimum possible rate:

$$\mathcal{H}\left(\mathbf{x}^{\text{ok}}\right) - \mathcal{H}\left(-\mathbf{\Delta}\right) = \frac{1}{2}\sum_{j=0}^{n-1}\log_2\frac{\lambda_j^{\text{ok}}}{\theta_j} \tag{D.5}$$

Since such encoding satisfies the condition $\mathbf{E}\left[\left(\hat{x}_j^{\text{ok}} - \hat{x}_j^{\text{ok}}\right)^2\right] = \theta_j$ for each $j = 0,\ldots,n-1$, the total distortion can be expressed as:

$$D = \sum_{j=0}^{n-1}\theta_j \tag{D.6}$$

□

## D.2 Proof of Lemma 6.2

*Proof of Lemma 6.2.* Assuming a Gaussian-additive encoding, we can express $\hat{\mathbf{x}}^{\mathrm{ok}}$ as $\mathbf{x}^{\mathrm{ok}} + \boldsymbol{\Delta}$ having independent zero-mean Gaussian components, and each element has variance $\sigma^2_{\hat{\mathbf{x}}^{\mathrm{ok}}_j} = \lambda^{\mathrm{ok}}_j - \theta_j$ for $j = 0, \ldots, n-1$. The results of Chapter 5 can be directly applied in this slightly generalized scenario, allowing us to utilize the encoding $f_{\hat{\mathbf{x}}|\mathbf{x}}$ (Lemma 5.1) and its application to the components of $\mathbf{x}^{\mathrm{ko}}$ (Lemma 5.2).

This results in $\hat{\mathbf{x}}^{\mathrm{ko}}$ consisting of independent, zero-mean, Gaussian components with variances

$$\sigma^2_{\hat{x}^{\mathrm{ko}}_j} = \left(1 - \frac{\theta_j}{\lambda^{\mathrm{ok}}_j}\right)\left[\lambda^{\mathrm{ko}}_j \left(1 - \frac{\theta_j}{\lambda^{\mathrm{ok}}_j}\right) + \theta_j\right] \tag{D.7}$$

for each $j = 0, \ldots, n-1$.

Finally, by applying Lemma 4.1 of Chapter 4, we can express both $\mathcal{Z}$ and $\mathcal{J}$ as functions of the ratios

$$u_j = \frac{\sigma^2_{\hat{x}^{\mathrm{ko}}_j}}{\sigma^2_{\hat{x}^{\mathrm{ok}}_j}} = 1 - r_j \xi_j. \tag{D.8}$$

In particular, using Lemma 4.1, together with the definitions of $\mathcal{Z}$ and $\mathcal{J}$, leads to equations (6.12) and (6.13).

Both $\mathcal{Z}(\boldsymbol{\xi})$ and $\mathcal{J}(\boldsymbol{\xi})$ are convex functions of $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_{n-1})$. Specifically, $\mathcal{Z}$ is linear for $\sum_{j=0}^{n-1} r_j \xi_j \geq 0$ and for $\sum_{j=0}^{n-1} r_j \xi_j < 0$. Furthermore, since $r_j \xi_j < 1$, the function $\frac{(r_j \xi_j)^2}{1 - r_j \xi_j}$ is convex for $j = 0, \ldots, n-1$. Consequently, $\mathcal{J}$, being a sum of convex functions, is also convex.

For the new degrees of freedom, the distortion can be written as:

$$D = \sum_{j=0}^{n-1} \theta_j = \sum_{j=0}^{n-1} \lambda^{\mathrm{ok}}_j (1 - \xi_j) = n - \sum_{j=0}^{n-1} \lambda^{\mathrm{ok}}_j \xi_j. \tag{D.9}$$

□

# Bibliography

[1] M. Abououf et al. “Explainable AI for Event and Anomaly Detection and Classification in Healthcare Monitoring Systems”. In: *IEEE Internet of Things Journal* 11.2 (2024), pp. 3446–3457. DOI: 10.1109/JIOT.2023.3296809.

[2] C. C. Aggarwal. *Outlier Analysis*. Springer Cham, 2017. ISBN: 978-3-319-47577-6. DOI: https://doi.org/10.1007/978-3-319-47578-3.

[3] E. Agustsson et al. “Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, 1141–1151. ISBN: 9781510860964.

[4] R. Ahlswede and I. Csiszar. “Hypothesis testing with communication constraints”. In: *IEEE Transactions on Information Theory* 32.4 (1986), pp. 533–542. DOI: 10.1109/TIT.1986.1057194.

[5] S. Ahmad et al. “Unsupervised real-time anomaly detection for streaming data”. In: *Neurocomputing* 262 (2017). Online Real-Time Learning Strategies for Data Streams, pp. 134–147. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.04.070.

[6] N. Ahmed, P. J. Milne, and S. G. Harris. “Electrocardiographic Data Compression Via Orthogonal Transforms”. In: *IEEE Transactions on Biomedical Engineering* BME-22.6 (1975), pp. 484–487. DOI: 10.1109/TBME.1975.324469.

[7] N. Ahmed, T. Natarajan, and K. R. Rao. “Discrete Cosine Transform”. In: *IEEE Transactions on Computers* C-23.1 (1974), pp. 90–93. DOI: 10.1109/T-C.1974.223784.

[8] N. Ahmed and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing*. 1st ed. Springer-Verlag Berlin, Heidelberg, 1975. ISBN: 978-3-642-45450-9. DOI: 10.1007/978-3-642-45450-9.

[9] E. Anthi et al. “A Supervised Intrusion Detection System for Smart Home IoT Devices”. In: *IEEE Internet of Things Journal* 6.5 (2019), pp. 9042–9053. DOI: 10.1109/JIOT.2019.2926365.

[10] M. Antonini et al. “Image coding using wavelet transform”. In: *IEEE Transactions on Image Processing* 1.2 (1992), pp. 205–220. DOI: 10.1109/83.136597.

[11] R. B. Arellano-Valle, J. E. Contreras-Reyes, and M. G. Genton. “Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions”. In: *Scandinavian Journal of Statistics* 40.1 (2013), pp. 42–62. DOI: https://doi.org/10.1111/j.1467-9469.2011.00774.x.

[12] Y. Bai et al. “Towards End-to-End Image Compression and Analysis with Transformers”. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. Vol. 36. 2022, 104 – 112.

[13] J. Ballé et al. “Nonlinear Transform Coding”. In: *IEEE Journal of Selected Topics in Signal Processing* 15.2 (2021), pp. 339–353. DOI: 10.1109/JSTSP.2020.3034501.

[14] J.S. Baras and S. Dey. “Combined compression and classification with learning vector quantization”. In: *IEEE Transactions on Information Theory* 45.6 (1999), pp. 1911–1920. DOI: 10.1109/18.782112.

[15] A. J. Bell and T. J. Sejnowski. “An Information-Maximization Approach to Blind Separation and Blind Deconvolution”. In: *Neural Computation* 7.6 (Nov. 1995), pp. 1129–1159. ISSN: 0899-7667. DOI: 10.1162/neco.1995.7.6.1129.

[16] P. Bergmann et al. “The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection”. In: *International Journal of Computer Vision* 129.4 (2021), pp. 1038–1059. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01400-4.

[17] A. Bhattacharyya. “On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions”. In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–109.

[18] Y. Blau and T. Michaeli. “Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 675–685.

[19] Y. Blau and T. Michaeli. “The Perception-Distortion Tradeoff”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6228–6237. DOI: 10.1109/CVPR.2018.00652.

[20] J. Bonnel et al. “Nonlinear time-warping made simple: A step-by-step tutorial on underwater acoustic modal separation with a single hydrophone”. In: *The Journal of the Acoustical Society of America* 147.3 (2020), pp. 1897–1926. DOI: 10.1121/10.0000937.

[21] D. Bortolotti et al. “An Ultra-Low Power Dual-Mode ECG Monitor for Healthcare and Wellness”. In: *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2015, pp. 1611–1616. DOI: 10.7873/DATE.2015.0784.

[22] M. M. Breunig et al. “LOF: Identifying Density-Based Local Outliers”. In: *SIGMOD Record* 29.2 (2000), pp. 93–104. DOI: 10.1145/335191.335388.

[23] V. Britanak. “A Survey of Efficient MDCT Implementations in MP3 Audio Coding Standard: Retrospective and State-of-the-Art”. In: *Signal Processing* 91.4 (2011), pp. 624–672. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2010.09.009.

[24] A. Burrello et al. “Embedded Streaming Principal Components Analysis for Network Load Reduction in Structural Health Monitoring”. In: *IEEE Internet of Things Journal* 8.6 (2021), pp. 4433–4447. DOI: 10.1109/JIOT.2020.3027102.

[25] V. Cambareri et al. “A Rakeness-based Design Flow for Analog-to-Information Conversion by Compressive Sensing”. In: *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2013, pp. 1360–1363. DOI: 10.1109/ISCAS.2013.6572107.

[26] V.L. Cao, M. Nicolau, and J. McDermott. “Learning Neural Representations for Network Anomaly Detection”. In: *IEEE Transactions on Cybernetics* 49.8 (2019), pp. 3074–3087. DOI: 10.1109/TCYB.2018.2838668.

[27] L.D. Chamain, S. Qi, and Z. Ding. “End-to-End Image Classification and Compression With Variational Autoencoders”. In: *IEEE Internet of Things Journal* 9.21 (2022), pp. 21916–21931. DOI: 10.1109/JIOT.2022.3182313.

[28] S. Chandak et al. “LFZip: Lossy Compression of Multivariate Floating-Point Time Series Data via Improved Prediction”. In: *2020 Data Compression Conference (DCC)*. 2020, pp. 342–351. DOI: 10.1109/DCC47342.2020.00042.

[29] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly Detection: A Survey”. In: *ACM Computing Surveys* 41.3 (2009), p. 15. DOI: 10.1145/1541880.1541882.

[30] H. Chen et al. “Data-Driven Detection of Hot Spots in Photovoltaic Energy Systems”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49.8 (2019), pp. 1731–1738. DOI: 10.1109/TSMC.2019.2896922.

[31] G. Chiarot and C. Silvestri. “Time Series Compression Survey”. In: *ACM Comput. Surv.* 55.10 (Feb. 2023), p. 198. DOI: 10.1145/3560814.

[32] C. Ciancarelli et al. “Innovative ML-based Methods for Automated On-board Spacecraft Anomaly Detection”. In: *Studies in Computational Intelligence* 1088 (2023), 213 – 228. DOI: 10.1007/978-3-031-25755-1_14.

[33] C. Ciancarelli et al. “New Concepts Of Automated Anomaly Detection In Space Operations Through ML-Based Techniques”. In: vol. 2022-September. 2022.

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991. ISBN: 0471062596.

[35] K. Crammer and G. Chechik. “A Needle in a Haystack: Local One-Class Optimization”. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 26. ISBN: 1581138385. DOI: 10.1145/1015330.1015399.

[36] K. Crammer, P. P. Talukdar, and F. Pereira. “A Rate-Distortion One-Class Model and Its Applications to Clustering”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. Helsinki, Finland: Association for Computing Machinery, 2008, 184–191. ISBN: 9781605582054. DOI: 10.1145/1390156.1390180.

[37] Q. K. Dang and Y. S. Suh. “Sensor Saturation Compensated Smoothing Algorithm for Inertial Sensor-Based Motion Tracking”. In: *Sensors* 14.5 (2014), pp. 8167–8188. DOI: 10.3390/s140508167.

[38] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. “The Mahalanobis distance”. In: *Chemometrics and Intelligent Laboratory Systems* 50.1 (2000), pp. 1–18. ISSN: 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(99)00047-7.

[39] L. Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[40] J. Ding, V. Tarokh, and Y. Yang. “Model Selection Techniques: An Overview”. In: *IEEE Signal Processing Magazine* 35.6 (2018), pp. 16–34. DOI: 10.1109/MSP.2018.2867638.

[41] R. Domingues et al. “A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses”. In: *Pattern Recognition* 74 (2018), pp. 406–421. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.09.037.

[42] Y. Dong, S. Chang, and L. Carin. “Rate-distortion bound for joint compression and classification with application to multiaspect scattering”. In: *IEEE Sensors Journal* 5.3 (2005), pp. 481–492. DOI: 10.1109/JSEN.2005.844338.

[43] L. Duan et al. “Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8680–8695. DOI: 10.1109/TIP.2020.3016485.

[44] Z. Duan et al. "QARV: Quantization-Aware ResNet VAE for Lossy Image Compression". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.1 (2024), pp. 436–450. DOI: 10.1109/TPAMI.2023.3322904.

[45] T. Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.10.010.

[46] B. Foo, Y. Andreopoulos, and M. van der Schaar. "Analytical Complexity Modeling of Wavelet-based Video Coders". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 3. 2007, pp. III–789–III–792. DOI: 10.1109/ICASSP.2007.366798.

[47] M. Gavish and D.L. Donoho. "The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$". In: *IEEE Transactions on Information Theory* 60.8 (2014), pp. 5040–5053. DOI: 10.1109/TIT.2014.2323359.

[48] C. L. Giles, G. M. Kuhn, and R. J. Williams. "Dynamic recurrent neural networks: Theory and applications". In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 153–156. DOI: 10.1109/TNN.1994.8753425.

[49] J. Ginibre. "Statistical Ensembles of Complex, Quaternion, and Real Matrices". In: *Journal of Mathematical Physics* 6.3 (1965), pp. 440–449. DOI: 10.1063/1.1704292.

[50] A.L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: (2000). DOI: 10.1161/01.cir.101.23.e215.

[51] M. Goswami et al. "Unsupervised Model Selection for Time Series Anomaly Detection". In: *The Eleventh International Conference on Learning Representations*. 2023.

[52] V. K. Goyal. "Theoretical foundations of transform coding". In: *IEEE Signal Processing Magazine* 18.5 (2001), pp. 9–21. DOI: 10.1109/79.952802.

[53] R. Gray. "Vector quantization". In: *IEEE ASSP Magazine* 1.2 (1984), pp. 4–29. DOI: 10.1109/MASSP.1984.1162229.

[54] R.M. Gray and D.L. Neuhoff. "Quantization". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2325–2383. DOI: 10.1109/18.720541.

[55] R. Gupta and A.O. Hero. "High-rate vector quantization for detection". In: *IEEE Transactions on Information Theory* 49.8 (2003), pp. 1951–1969. DOI: 10.1109/TIT.2003.814482.

[56] S. Han et al. "ADBench: Anomaly Detection Benchmark". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022.

[57] T. S. Han and S. Amari. "Statistical inference under multiterminal data compression". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2300–2324. DOI: 10.1109/18.720540.

[58] G. He et al. "A Fast Semi-Supervised Clustering Framework for Large-Scale Time Series Data". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51.7 (2021), pp. 4201–4216. DOI: 10.1109/TSMC.2019.2931731.

[59] H. He and Y. Tan. "Unsupervised Classification of Multivariate Time Series Using VPCA and Fuzzy Clustering With Spatial Weighted Matrix Distance". In: *IEEE Transactions on Cybernetics* 50.3 (2020), pp. 1096–1105. DOI: 10.1109/TCYB.2018.2883388.

[60] D. Hendrycks and T. Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *Proceedings of the International Conference on Learning Representations* (2019).

[61] R.J. Hillestad and S.E. Jacobsen. “Reverse convex programming”. In: *Applied Mathematics and Optimization* (1980), pp. 63–68. DOI: 10.1007/BF01442883.

[62] S. Hu et al. “A Survey on Information Bottleneck”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (2024), pp. 5325–5344. DOI: 10.1109/TPAMI.2024.3366349.

[63] Y. Hu et al. “Learning End-to-End Lossy Image Compression: A Benchmark”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8 (2022), pp. 4194–4211. DOI: 10.1109/TPAMI.2021.3065339.

[64] C. Huang et al. “Event-Triggering State and Fault Estimation for a Class of Nonlinear Systems Subject to Sensor Saturations”. In: *Sensors* 21.4 (2021). DOI: 10.3390/s21041242.

[65] D. A. Huffman. “A Method for the Construction of Minimum-Redundancy Codes”. In: *Proceedings of the IRE* 40.9 (1952), pp. 1098–1101. DOI: 10.1109/JRPROC.1952.273898.

[66] S. M. R. Islam et al. “The Internet of Things for Health Care: A Comprehensive Survey”. In: *IEEE Access* 3 (2015), pp. 678–708. DOI: 10.1109/ACCESS.2015.2437951.

[67] S. U. Jan et al. “Sensor Fault Classification Based on Support Vector Machine and Statistical Time-Domain Features”. In: *IEEE Access* 5 (2017), pp. 8682–8690. DOI: 10.1109/ACCESS.2017.2705644.

[68] H. Jeffreys. “An invariant form for the prior probability in estimation problems”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007 (1946), pp. 453–461. DOI: 10.1098/rspa.1946.0056.

[69] J. Jin et al. “An Information Framework for Creating a Smart City Through Internet of Things”. In: *IEEE Internet of Things Journal* 1.2 (2014), pp. 112–121. DOI: 10.1109/JIOT.2013.2296516.

[70] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002. ISBN: 978-0-387-22440-4. DOI: https://doi.org/10.1007/b98835.

[71] T. Kailath. “The Divergence and Bhattacharyya Distance Measures in Signal Selection”. In: *IEEE Transactions on Communication Technology* 15.1 (1967), pp. 52–60. DOI: 10.1109/TCOM.1967.1089532.

[72] E. Karakose et al. “A New Experimental Approach Using Image Processing-Based Tracking for an Efficient Fault Diagnosis in Pantograph–Catenary Systems”. In: *IEEE Transactions on Industrial Informatics* 13.2 (2017), pp. 635–643. DOI: 10.1109/TII.2016.2628042.

[73] Y. Kawaguchi. “Anomaly Detection Based on Feature Reconstruction from Subsampled Audio Signals”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. 2018, pp. 2524–2528. DOI: 10.23919/EUSIPCO.2018.8553480.

[74] S. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall PTR, 1988. ISBN: 013504135x.

[75] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.

[76] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. 2014. DOI: arXiv:1312.6114.

[77] A. Kolmogorov. "On the Shannon theory of information transmission in the case of continuous signals". In: *IRE Transactions on Information Theory* 2.4 (1956), pp. 102–108. DOI: 10.1109/TIT.1956.1056823.

[78] M.A. Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *Aiche Journal* 37 (1991), pp. 233–243.

[79] A. Krizhevsky and G. Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. 0. Toronto, Ontario: University of Toronto, 2009.

[80] S. Kullback. *Information Theory and Statistics*. John Wiley & Sons, 1959. ISBN: 0844656259.

[81] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79 –86. DOI: 10.1214/aoms/1177729694.

[82] M.N. Kurt, Y. Yılmaz, and X. Wang. "Real-Time Nonparametric Anomaly Detection in High-Dimensional Settings". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.7 (2021), pp. 2463–2479. DOI: 10.1109/TPAMI.2020.2970410.

[83] K. H. Lai et al. "Revisiting Time Series Outlier Detection: Definitions and Benchmarks". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.

[84] M.A. Lexa. "Quantization via Empirical Divergence Maximization". In: *IEEE Transactions on Signal Processing* 60.12 (2012), pp. 6408–6420. DOI: 10.1109/TSP.2012.2217136.

[85] H. Li and P. Boulanger. "A Survey of Heart Anomaly Detection Using Ambulatory Electrocardiogram (ECG)". In: *Sensors* 20.5 (2020). DOI: 10.3390/s20051461.

[86] D. Liang et al. "Fuzzy-Sliding Mode Control for Humanoid Arm Robots Actuated by Pneumatic Artificial Muscles With Unidirectional Inputs, Saturations, and Dead Zones". In: *IEEE Transactions on Industrial Informatics* 18.5 (2022), pp. 3011–3021. DOI: 10.1109/TII.2021.3111655.

[87] B. Liu et al. "A Multitone Model-Based Seismic Data Compression". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.2 (2022), pp. 1030–1040. DOI: 10.1109/TSMC.2021.3077490.

[88] F. Liu et al. "Rate-Adaptive Multitask-Oriented Semantic Communication: An Extended Rate–Distortion Theory-Based Scheme". In: *IEEE Internet of Things Journal* 11.9 (2024), pp. 15557–15570. DOI: 10.1109/JIOT.2024.3350656.

[89] F. T. Liu, K. M. Ting, and Z.-H. Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining (ICDM)*. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[90] J. Liu, H. Sun, and J. Katto. "Improving Multiple Machine Vision Tasks in the Compressed Domain". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022, pp. 331–337. DOI: 10.1109/ICPR56361.2022.9956532.

[91] J. Liu et al. "Deep Dict: Deep Learning-Based Lossy Time Series Compressor for IoT Data". In: *ICC 2024 - IEEE International Conference on Communications*. 2024, pp. 4245–4250. DOI: 10.1109/ICC51166.2024.10622275.

[92] K. Liu et al. "A Data-Driven Approach With Uncertainty Quantification for Predicting Future Capacities and Remaining Useful Life of Lithium-ion Battery". In: *IEEE Transactions on Industrial Electronics* 68.4 (2021), pp. 3170–3180. DOI: 10.1109/TIE.2020.2973876.

[93] S.-Y. Lo, P. Oza, and V.M. Patel. "Adversarially Robust One-Class Novelty Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023), pp. 4167–4179. DOI: 10.1109/TPAMI.2022.3189638.

[94] V. Loia, S. Tomasiello, and A. Vaccaro. "Fuzzy Transform Based Compression of Electric Signal Waveforms for Smart Grids". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47.1 (2017), pp. 121–132. DOI: 10.1109/TSMC.2016.2578641.

[95] X. Luo et al. "The Rate-Distortion-Accuracy Tradeoff: JPEG Case Study". In: *2021 Data Compression Conference (DCC)*. 2021, pp. 354–354. DOI: 10.1109/DCC50243.2021.00049.

[96] M. Mangia, R. Rovatti, and G. Setti. "Rakeness in the Design of Analog-to-Information Conversion of Sparse and Localized Signals". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 59.5 (2012), pp. 1001–1014. DOI: 10.1109/TCSI.2012.2191312.

[97] M. Mangia et al. "Deep Neural Oracles for Short-Window Optimized Compressed Sensing of Biosignals". In: *IEEE Transactions on Biomedical Circuits and Systems* 14.3 (2020), pp. 545–557. DOI: 10.1109/TBCAS.2020.2982824.

[98] M. Mangia et al. "Rakeness-based Approach to Compressed Sensing of ECGs". In: *2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 2011, pp. 424–427. DOI: 10.1109/BioCAS.2011.6107818.

[99] M. Mangia et al. "Rakeness-Based Design of Low-Complexity Compressed Sensing". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.5 (2017), pp. 1201–1213. DOI: 10.1109/TCSI.2017.2649572.

[100] A. Marchioni et al. "Streaming Algorithms for Subspace Analysis: Comparative Review and Implementation on IoT Devices". In: *IEEE Internet of Things Journal* 10.14 (2023), pp. 12798–12810. DOI: 10.1109/JIOT.2023.3256529.

[101] A. Marchioni et al. "Subspace Energy Monitoring for Anomaly Detection @Sensor or @Edge". In: *IEEE Internet of Things Journal* 7.8 (2020), pp. 7575–7589. DOI: 10.1109/JIOT.2020.2985912.

[102] P.E. McSharry et al. "A dynamical model for generating synthetic electrocardiogram signals". In: *IEEE Transactions on Biomedical Engineering* 50.3 (2003), pp. 289–294. DOI: 10.1109/TBME.2003.808805.

[103] F. Mentzer et al. "Conditional Probability Models for Deep Image Compression". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4394–4402. DOI: 10.1109/CVPR.2018.00462.

[104] F. Mezzadri. "How to generate random matrices from the classical compact groups". In: *Notices of the American Mathematical Society* 54.5 (2006), pp. 592–604.

[105] A. Moallemi et al. "Exploring Scalable, Distributed Real-Time Anomaly Detection for Bridge Health Monitoring". In: *IEEE Internet of Things Journal* 9.18 (2022), pp. 17660–17674. DOI: 10.1109/JIOT.2022.3157532.

[106] S. R. Moreno et al. "Wind turbines anomaly detection based on power curves and ensemble learning". In: *IET Renewable Power Generation* 14.19 (2020), pp. 4086–4093. DOI: https://doi.org/10.1049/iet-rpg.2020.0224.

[107] S. D. Morgera and L. Datta. "Toward a Fundamental Theory of Optimal Feature Selection: Part I". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.5 (1984), pp. 601–616. DOI: 10.1109/TPAMI.1984.4767573.

[108] J. Mũnoz-Marí et al. "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 48.8 (2010), pp. 3188–3197. DOI: 10.1109/TGRS.2010.2045764.

[109] K. Ni et al. "Sensor Network Data Fault Types". In: *ACM Trans. Sen. Netw.* 5.3 (2009). ISSN: 1550-4859. DOI: 10.1145/1525856.1525863.

[110] A. B. Noel et al. "Structural Health Monitoring Using Wireless Sensor Networks: A Comprehensive Survey". In: *IEEE Communications Surveys and Tutorials* 19.3 (2017), pp. 1403–1423. DOI: 10.1109/COMST.2017.2691551.

[111] K.L. Oehler and R.M. Gray. "Combining image compression and classification using vector quantization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.5 (1995), pp. 461–473. DOI: 10.1109/34.391396.

[112] 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". In: *Annals of Operations Research* 189.1 (2011), pp. 331–342. DOI: 10.1007/s10479-009-0567-7.

[113] J. Paparrizos et al. "TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection". In: *Proc. VLDB Endow.* 15.8 (2022), 1697–1711. ISSN: 2150-8097. DOI: 10.14778/3529337.3529354.

[114] M. Pei et al. "Self-Supervised Learning for Industrial Image Anomaly Detection by Simulating Anomalous Samples". In: *International Journal of Computational Intelligence Systems* 16.1 (2023), p. 152. ISSN: 1875-6883. DOI: 10.1007/s44196-023-00328-0.

[115] F. Pilati et al. "Assembly Line Balancing and Activity Scheduling for Customised Products Manufacturing". In: *The International Journal of Advanced Manufacturing Technology* 120 (2022), 3925–3946. DOI: 10.1007/s00170-022-08953-3.

[116] J. C. Principe. *Information Theoretic Learning*. Springer New York, NY, 2010. ISBN: 978-1-4419-1569-6.

[117] S. B. Provost and A. M. Mathai. *Quadratic Forms in Random Variables: Theory and Applications*. Statistics: Textbooks and Monographs. Marcel Dekker, 1992.

[118] N. Ramanathan et al. "The Final Frontier: Embedding Networked Sensors in the Soil". In: *eScholarship*. 2006.

[119] L. Ren et al. "A Data-Driven Auto-CNN-LSTM Prediction Model for Lithium-Ion Battery Remaining Useful Life". In: *IEEE Transactions on Industrial Informatics* 17.5 (2021), pp. 3478–3487. DOI: 10.1109/TII.2020.3008223.

[120] O. Rioul and M. Vetterli. "Wavelets and signal processing". In: *IEEE Signal Processing Magazine* 8.4 (1991), pp. 14–38. DOI: 10.1109/79.91217.

[121] O. Rippel et al. "Gaussian Anomaly Detection by Modeling the Distribution of Normal Data in Pretrained Deep Features". In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–13. DOI: 10.1109/TIM.2021.3098381.

[122] J. Rissanen and G. G. Langdon. "Arithmetic Coding". In: *IBM Journal of Research and Development* 23.2 (1979), pp. 149–162. DOI: 10.1147/rd.232.0149.

[123] M. R. D. Rodrigues et al. "Rate-distortion trade-offs in acquisition of signal parameters". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 6105–6109. DOI: 10.1109/ICASSP.2017.7953329.

[124] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268. ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(92)90242-F.

[125] L. Ruff et al. "A Unifying Review of Deep and Shallow Anomaly Detection". In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795. DOI: 10.1109/JPROC.2021.3052449.

[126] L. Ruff et al. "Deep One-Class Classification". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 4393–4402.

[127] L. Ruff et al. "Deep Semi-Supervised Anomaly Detection". In: *International Conference on Learning Representations*. 2020.

[128] A. Sabato et al. "Noncontact Sensing Techniques for AI-Aided Structural Health Monitoring: A Systematic Review". In: *IEEE Sensors Journal* 23.5 (2023), pp. 4672–4684. DOI: 10.1109/JSEN.2023.3240092.

[129] S. Schmidl, P. Wenig, and T. Papenbrock. "Anomaly Detection in Time Series: A Comprehensive Evaluation". In: *Proc. VLDB Endow.* 15.9 (2022), 1779–1797. ISSN: 2150-8097. DOI: 10.14778/3538598.3538602.

[130] B. Schölkopf et al. "Support Vector Method for Novelty Detection". In: *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. Denver, CO: MIT Press, 1999, pp. 582–588.

[131] A. Sharma, L. Golubchik, and R. Govindan. "On the Prevalence of Sensor Faults in Real-World Deployments". In: *2007 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*. 2007, pp. 213–222. DOI: 10.1109/SAHCN.2007.4292833.

[132] A. B. Sharma, L. Golubchik, and R. Govindan. "Sensor Faults: Detection Methods and Prevalence in Real-World Datasets". In: *ACM Trans. Sen. Netw.* 6.3 (2010). ISSN: 1550-4859. DOI: 10.1145/1754414.1754419.

[133] R.K. Sharma and J.W. Wallace. "Improved Spectrum Sensing by Utilizing Signal Autocorrelation". In: *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*. 2009, pp. 1–5. DOI: 10.1109/VETECS.2009.5073595.

[134] X. Shen et al. "Disciplined convex-concave programming". In: *2016 IEEE 55th Conference on Decision and Control (CDC)* (2016), pp. 1009–1014.

[135] X. Sheng et al. "VNVC: A Versatile Neural Video Coding Framework for Efficient Human-Machine Vision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.7 (2024), pp. 4579–4596. DOI: 10.1109/TPAMI.2024.3356548.

[136] W. Shi et al. "Edge Computing: Vision and Challenges". In: *IEEE Internet of Things Journal* 3.5 (2016), pp. 637–646. DOI: 10.1109/JIOT.2016.2579198.

[137] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues. "Hardware-Limited Task-Based Quantization". In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5223–5238. DOI: 10.1109/TSP.2019.2935864.

[138] S. Singh et al. “End-to-End Learning of Compressible Features”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, pp. 3349–3353. DOI: 10.1109/ICIP40778.2020.9190860.

[139] A. Skodras, C. Christopoulos, and T. Ebrahimi. “The JPEG 2000 still image compression standard”. In: *IEEE Signal Processing Magazine* 18.5 (2001), pp. 36–58. DOI: 10.1109/79.952804.

[140] N. Slonim and N. Tishby. “Agglomerative Information Bottleneck”. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*. NIPS’99. MIT Press, 1999, 617–623. DOI: 10.5555/3009657.3009745.

[141] G. Steinbuss and K. Böhm. “Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data”. In: *ACM Trans. Knowl. Discov. Data* 15.4 (2021), p. 65. ISSN: 1556-4681. DOI: 10.1145/3441453.

[142] I. Steinwart, D. Hush, and C. Scovel. “A Classification Framework for Anomaly Detection”. In: *J. Mach. Learn. Res.* 6 (Dec. 2005), 211–232. ISSN: 1532-4435.

[143] D. M. J. Tax and R. P. W. Duin. “Support Vector Data Description”. In: *Machine Learning* 54 (2004). DOI: 10.1023/B:MACH.0000008084.60811.49.

[144] D. M. J. Tax and K.-R. Müller. “Feature Extraction for One-Class Classification”. In: *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003*. Ed. by O. Kaynak et al. Springer Berlin Heidelberg, 2003, pp. 342–349. ISBN: 978-3-540-44989-8.

[145] L. Theis et al. “Lossy Image Compression with Compressive Autoencoders”. In: *International Conference on Learning Representations*. 2017.

[146] N. Tishby, F. C. Pereira, and W. Bialek. “The information bottleneck method”. In: *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*. 1999, 368–377.

[147] R. Torfason et al. “Towards Image Understanding from Deep Compression Without Decoding”. In: *International Conference on Learning Representations*. 2018.

[148] G. T. Toussaint. “Probability of Error, Expected Divergence, and the Affinity of Several Distributions”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 8.6 (1978), pp. 482–485. DOI: 10.1109/TSMC.1978.4310001.

[149] H. Tuy. “Convex programs with an additional reverse convex constraint”. In: *Journal of Optimization Theory and Applications* 52.3 (Mar. 1987), pp. 463–486. DOI: 10.1007/BF00938217.

[150] M. Unser. “Splines: a perfect fit for signal and image processing”. In: *IEEE Signal Processing Magazine* 16.6 (1999), pp. 22–38. DOI: 10.1109/79.799930.

[151] M. Vetterli. “Wavelets, approximation, and compression”. In: *IEEE Signal Processing Magazine* 18.5 (2001), pp. 59–73. DOI: 10.1109/79.952805.

[152] G.K. Wallace. “The JPEG still picture compression standard”. In: *IEEE Transactions on Consumer Electronics* 38.1 (1992), pp. xviii–xxxiv. DOI: 10.1109/30.125072.

[153] S. J. Wenndt and A. J. Noga. “Narrow-Band Interference Cancellation for Enhanced Speaker Identification”. In: *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’99)*. 1999, pp. 123–126. DOI: 10.1109/ASPAA.1999.810865.

[154] G. Williams et al. “A comparative study of RNN for outlier detection in data mining”. In: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* 2002, pp. 709–712. DOI: 10.1109/ICDM.2002.1184035.

[155] R. Wu and E. J. Keogh. “Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.3 (2023), pp. 2421–2429. DOI: 10.1109/TKDE.2021.3112126.

[156] H. Xu et al. “Deep Isolation Forest for Anomaly Detection”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.12 (2023), pp. 12591–12604. DOI: 10.1109/TKDE.2023.3270293.

[157] K. Yan et al. “Fast and Accurate Classification of Time Series Data Using Extended ELM: Application in Fault Diagnosis of Air Handling Units”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49.7 (2019), pp. 1349–1356. DOI: 10.1109/TSMC.2017.2691774.

[158] M. Yang et al. “SLSG: Industrial image anomaly detection with improved feature embeddings and one-class classification”. In: *Pattern Recognition* 156 (2024), p. 110862. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2024.110862.

[159] W. Yang et al. “Video Coding for Machines: Compact Visual Representation Compression for Intelligent Collaborative Analytics”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.7 (2024), pp. 5174–5191. DOI: 10.1109/TPAMI.2024.3367293.

[160] S. Yin et al. “A Review on Basic Data-Driven Approaches for Industrial Process Monitoring”. In: *IEEE Transactions on Industrial Electronics* 61.11 (2014), pp. 6418–6428. DOI: 10.1109/TIE.2014.2301773.

[161] H. Q. Zhang and Y. Yan. “A Wavelet-Based Approach to Abrupt Fault Detection and Diagnosis of Sensors”. In: *IEEE Transactions on Instrumentation and Measurement* 50.5 (2001), pp. 1389–1396. DOI: 10.1109/19.963215.

[162] X. Zhang, M. Xu, and X. Zhou. “RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 16699–16708.

[163] Y. Zhang. “A Rate-Distortion-Classification approach for lossy image compression”. In: *Digital Signal Processing* 141 (2023), p. 104163. ISSN: 1051-2004. DOI: https://doi.org/10.1016/j.dsp.2023.104163.

[164] L. Zhao, Y. Liu, and L. Wang. “Image compression and reconstruction of transmission line monitoring images using compressed sensing”. In: *2017 8th International Conference on Mechanical and Aerospace Engineering (ICMAE)*. 2017, pp. 371–375. DOI: 10.1109/ICMAE.2017.8038674.

[165] F. Zonzini et al. “Model-Assisted Compressed Sensing for Vibration-Based Structural Health Monitoring”. In: *IEEE Transactions on Industrial Informatics* 17.11 (2021), pp. 7338–7347. DOI: 10.1109/TII.2021.3050146.