Doctoral Program in Mechanical Engineering - 31st Cycle

Vibration Monitoring: Gearbox Identification and Faults Detection

A thesis submitted to the Politecnico di Torino for the degree of Doctor of Philosophy in Mechanical Engineering by

Alessandro Paolo Daga

Dynamic and Identification Research Group - DIRG
Department of Mechanical and Aerospace Engineering - DIMEAS
Politecnico di Torino

Supervisors:

Prof. Luigi Garibaldi, *Supervisor*
Prof. Alessandro Fasana, *Co-Supervisor*

Doctoral Examination Committee:

Prof. Keith Worden, Referee, University of Sheffield
Prof. Konstantinos Gryllias, Referee, KU Leuven

April 2019

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

…….....

Alessandro Paolo Daga

Turin, April 5, 2019

Acknowledgments

I would like to thank my reference professors Luigi Garibaldi, Alessandro Fasana and Stefano Marchesiello for their kindness, carefulness and always brilliant advices,

Prof. Keith Worden and Prof. Konstantinos Gryllias for their constructive revision of this work,

Prof. Francesco Castellani for involving me in the diagnostic of the Italian windfarm,

My parents Paolo and Paola for their constant support,

My girlfriend Federica for her admirable patience,

My mates for their natural joviality.

"As a multitude of laws often only hampers justice, so that a state is best governed when, with few laws, these are rigidly administered; in like manner, instead of the great number of precepts of which logic is composed, I believed that the four following would prove perfectly sufficient for me, provided I took the firm and unwavering resolution never in a single instance to fail in observing them.

> *The first was never to accept anything for true which I did not clearly know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in my judgment than what was presented to my mind so clearly and distinctly as to exclude all ground of doubt.*
>
> *The second, to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution.*
>
> *The third, to conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex; assigning in thought a certain order even to those objects which in their own nature do not stand in a relation of antecedence and sequence.*
>
> *And the last, in every case to make enumerations so complete, and reviews so general, that I might be assured that nothing was omitted."*
>
> René Descartes, A Discourse on Method, 1637

~

> *"Pure logical thinking cannot yield us any knowledge of the empirical world; all knowledge of reality starts from experience and end in it. Propositions arrived at by purely logical means are completely empty as regards reality. Because Galileo saw this, and particularly because he drummed it into the scientific world, he is the father of modern physics—indeed, of modern science altogether."*
>
> Ideas and opinions by Albert Einstein, 1954

~

> *"If you want to find the secrets of the universe, think in terms of energy, frequency and vibration."*
>
> attributed to Nikola Tesla, conversation with Ralph Bergstresser, 1942

Abstract

This thesis is devoted to the application of novel techniques to the identification and quantification of faults in gearboxes starting from vibration signals. In general, this kind of *abduction* can be considered a form of Non-Destructive Testing (NDT), whose scope is to increase the reliability of complex and expensive machines switching from programmed maintenance to preventive maintenance regimes based on such Vibration Monitoring (VM). The problem is then to perform a *data mining* on the available datasets so as to *recognize the patterns* and extract the useful *diagnostic information*.

Two parallel philosophies have been developed so as to comply with both *intermittent* and *continuous monitoring* of machines. The first allows the use of high-level signal processing techniques not only able to disclose the presence of a damage, but also the type, severity and location. The drawback is that the decision stage is usually not automated but left to a trained operator. In the second case, fast, real-time running statistical and machine learning algorithms can be used to trigger an alarm in case of *detection of damage*, leaving the quantification and localization of the damage for a further, more detailed analysis. Two methodologies are proposed by selecting from the literature the most suitable algorithms able to meet both the needs while ensuring model interpretability and satisfactory diagnostic results. These have been developed on theoretical modelled signals and on laboratory signals from a test rig at the DIRG lab (Dynamic & Identification Research Group test rig for high-speed aeronautical bearings) and later tested and compared on real signals from an aeronautical gearbox (SAFRAN aeronautical engine from the SAFRAN Contest, Conference Surveillance 8, October 20-21, 2015, Roanne, France) and from windmill gearboxes (Italian windfarm composed by six multi-megawatt wind turbines).

Keywords: Gearbox, Gear, Bearings, Vibration Monitoring, Damage Detection, Machine Diagnostics, Non-Destructive Testing, Data Mining, Signal Processing, Pattern Recognition, Machine Learning, Surveillance 8 SAFRAN aeronautical engine, Italian windfarm.

Nomenclature

Acronym	Full Name
AA NN	Auto-associative network
ADC	Analog-to-Digital Conversion
ALE	Adaptive Line Enhancement
ANC	Adaptive Noise Cancellation
ANN	Artificial Neural Networks
ANT	Adaptive neural tree
AR	Auto-Regressive
ART	Adaptive resonance theory network
BPFO/I	Ball Pass Frequency Outer/Inner race
BSF	Ball Spin Frequency
BSS	Blind Source Separation
CAA	Civil Aviation Authority (UK)
CART	Classification and Regression Tree
CBM	Condition-Based Maintenance
CEEMDAN	Complete EEMD with Adaptive Noise
CLINK	Complete Linkage Clustering
CM	Condition Monitoring
CNN	Convolutional Neural Network
DAG	Directed acyclic graph
DAQ	Data Acquisition System
DBN	Deep Belief Network
DRS	Discrete/Random Separation
EA	Envelope Analysis
EASA	European Aviation Safety Agency
EEMD	Ensemble EMD
EMD	Empirical Mode Decomposition
FA	False Alarms
FAA	Federal Aviation Authority (USA)
FK	Fast Kurtogram
FT (F/DFT)	(Fast/Discrete) Fourier Transform
FTF	Fundamental Train Frequency
GAN	Generative Adversarial Network
GAPF	Gear Assembly Phase Frequency
GLM	Generalized Linear Model
GMF	Gear Mesh Frequency
GMM	Gaussian Mixture Model
HMM	Hidden Markov Models
HTF	Hunting Tooth Frequency
HUMS	Health and Usage Monitoring Systems

IAS	Instantaneous Angular Speed
ICA	Independent Component Analysis
KDE	Kernel Density Estimators
k-NN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
LLE	Local Linear Embedding
LP	Linear Prediction (often LPC)
LSTM	Long/Short-Term Memory
MA	Missed Alarms
MA	Moving Average
MD	Mahalanobis Distance
MED	Minimum Entropy Deconvolution
ML	Machine Learning
MLP	Multi-Layer Perceptron
ND	Novelty Detection
NDT	Non-Destructive Testing
NI	Novelty Index
NN (ANN)	(Artificial) Neural Network
OT (COT)	(Computed) Order Tracking
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
RLS	Regularized Least Squares
RMS	Root-Mean Square
RNN	Recurrent Neural Network
SA (TSA)	(Time) Synchronous Average
SANC	Self-Adaptive Noise Cancellation
SHM	Structural Health Monitoring
SK	Spectral Kurtosis
SK*	Sliding filter estimate of SK
SLINK	Single-Linkage Clustering
SOM	Self-organizing maps
SSA	Singular Spectrum Analysis
SSD	Singular Spectrum Decomposition
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
SVR	Support Vector Regression
TBF	Time Between Failures
U/W PGMA	Unweighted or Weighted Pair Group Method with Arithmetic Mean
UPN	Unsupervised Pretrained Network
VM	Vibration Monitoring
WPT	Wavelet Packet Transform
WT (DWT)	(Discrete) Wavelet Transform

Table of contents

Chapter 1:

Motivation and scope

The work presented in this PhD thesis focuses on the application of vibration monitoring techniques to gearboxes, so as to identify the health condition of such mechanical components, detecting the eventual engendering of damages which could lead to a failure.

1. Motivation

Gearboxes are fundamental components of most industrial machines, as they are meant to transfer power from a motor (power source) to a user. In general, they are made of a metal casing within which a train of gears is sealed, in order to provide speed and torque conversion from a rotating power source connected to the input shaft to another device. Gearboxes are then critical components at different levels:

- Safety – whenever a power interruption may be harmful to people. For illustrative purposes, one can focus on the gearbox of a helicopter, or of any aeronautical means of transport in general. In this case, an interruption of the power implies an outage of the lift, which can lead to catastrophic results.
- Economic costs – whenever a failure results expensive in both repair cost and down-time. For example, one can consider an offshore windfarm. Such windmills are in general difficult to reach, and any intervention will be highly costly. Furthermore, a faulty wind turbine is unable to produce power, so that the downtime can cause a big loss of profit.

Additionally, one should consider that most of the machine faults are related to surface degradation which can be caused by corrosion or, primarily, by mechanical wear of the parts in contact (e.g. because of relative motion) such as rotating shafts, bearings and gears. Hence, machine faults are in the majority of cases imputable to gearboxes.

In order to improve the reliability of a mechanical device then, a diagnostic tool performing condition monitoring can be mounted on the machine, so as to gain in safety while having at the same time an economical advantage. Indeed, switching from a programmed maintenance regime to a predictive maintenance regime based on monitoring a parameter of condition, proves to be much more effective and efficient both from the technical and from the economical point of view. Because of this, condition monitoring gained increasing importance over the years, and is nowadays becoming an integral part of many maintenance regimes.

In particular, this thesis focuses on vibration monitoring (VM), a successful kind of condition monitoring based on accelerometric records. The high reliability of the used sensors (accelerometers), together with their small size and low weight, in fact, are the main advantages with respect to other technologies. In addition to the low impact of the sensors themselves, one should consider that the vibrations acquired outside a machine housing can convey information regarding the inner, hidden, mechanical parts, so that a diagnosis can be performed while the machine is in operation and without disassembling it.

2. Scope of the thesis

Despite the growing relevance of vibration monitoring, the main issue remains the interpretation of the vibration signal. Indeed, the problem is to process the large volumes of monitored data so as to extract some knowledge about the condition of the machine. The most reliable methods, in fact, date back to the 20th century (e.g. Time Synchronous Average, a fundamental technique of vibration monitoring, dates back to the 70s), moreover a univocal procedure of analysis is lacking, limiting the application in the industrial field.

The scope of this thesis is then to test a selection of promising and reliable methods on laboratory and real-life applications evaluating and comparing their performances. In the literature and in the scientific community one can find hundreds of methods and their variants. The scope is obviously not to review all of them, which would be impossible, but to identify some practical and representative ones and to organize them into a methodology which can cover the different needs of the monitoring of rotating machines, from the intermittent to the continuous monitoring. The main considerations used for the choice was the model interpretability, the degree to which the criteria for a decision can be understood by a human. The selected algorithms were developed according to the literature and, in some cases, improved so as to better match the particular needs of the practical real-life applications under analysis. Gearboxes, in fact, can equip many different machines and have then to work under very diverse conditions according to the particular application. In this thesis, three different applications are considered, two are related with aeronautical engines and one with windmills.

The first application regards a test rig built at the Dynamic & Identification Research Group (DIRG) laboratory and specifically conceived to test high-speed aeronautical bearings. The rig was already available, but the acquisitions were incomplete, and a long work of reorganization was needed to get the recordings ready for processing.

The second application refers to the accessory gearbox for the equipment (pumps, filters, alternators, starter etc.) of a SAFRAN civil aircraft engine with two damaged bearings. This dataset was part of the SAFRAN Contest of the Surveillance 8 conference held on October 20-21, 2015 in Roanne, France, to which the candidate participated gaining the second prize.

The third application concerns multimegawatt windmill gearboxes. The candidate was involved in the measurement campaign by an Italian windfarm in Molise region. The acquisitions were coordinated by prof. Francesco Castellani of the University of Perugia and were later analysed by the candidate with the proposed algorithms.

The three applications, despite being all centred on gearboxes (and in particular on bearing damages), are substantially different in terms of geometries, sizes, loads, rotational speeds, sensors kind, number and positioning and also size of tolerated damages.

To better compare the algorithms then, a simple synthetic signal of an ideal gearbox showing a damaged bearing was also produced by the candidate.

In general, the algorithms of interests are meant to extract the information hidden inside the accelerometric data (data mining), so as to produce a knowledge about the state of health of the machine. When this is done automatically by an Artificial Intelligence (a software running on a machine), it concerns the recognition of patterns in the available data. This Pattern Recognition problem is commonly solved using algorithms running on computers able to learn information from the data, both in the time domain and in the frequency domain. Therefore, researchers talk about Machine Learning.

The main steps of a Condition Monitoring system based on Machine Learning are:

- Acquire data for the purpose of training: measurements on the machine under analysis must be acquired at least in the healthy case as a reference; when possible, also acquisitions in known damaged conditions should be recorded.
- Pre-process the data to highlight the information: e.g. improve the signal of interest with respect to the background noise (Signal Processing).
- Extract the best features: a good feature is a quantity which summarizes the dataset and whose behaviour is correlated with the damage but, possibly, not with the operational and the environmental variables. Good features are then difficult to find e.g. defect frequencies, signal level, kurtosis, etc.
- Produce knowledge about damage: e.g. detect the presence of incipient damage, track the damage evolution in time so as to understand its severity, distinguish among damage location and types, prognose the remaining useful life.

In this work then, different algorithms for the different tasks will be tested to cover the entire analysis, establishing a methodology which can be used to extract the diagnostic knowledge from the data. The importance of a method, as theorized by Descartes [1], is central. Indeed, a methodology is an essential requirement for reaching knowledge and wisdom. The strength and the weakness of each method will be then objectively highlighted, so as to foster the development of diagnostic tools toward a maturity which can help vibration monitoring to keep spreading, guaranteeing for the future more accurate diagnosis and efficient maintenance regimes.

3. Outline

In the next chapters, the subject of vibration monitoring applied to gearbox identification and vibration monitoring will be treated. In particular:

- In chapter 2 the vibration monitoring motivations and philosophy are investigated. Two different ways of thinking the monitoring targeted on intermittent or continuous acquisitions are identified. The scope of diagnostics is declared, and the damage identification problem is defined in comparison to the well-known subject of pattern recognition.

- In chapter 3 the scientific literature is surveyed to produce a state-of-the-art review on vibration monitoring through signal processing and machine learning.

- In chapter 4 the selected algorithms from the literature review are illustrated and organized in a methodology for both the Intermittent Monitoring (based on Signa Processing algorithms) and for the Continuous Monitoring (based on Machine Learning).

- In chapter 5 the main signals used in this work are introduced. Firstly, the considerations in chapter 3 are applied to produce a synthetic simulated signal. This is used to make many preliminary tests on relevant algorithms. Anyway, the

performance assessment is mainly achieved on three experimental acquisitions. The first regards a test rig built at the Dynamic & Identification Research Group (DIRG) laboratory and specifically conceived to test high-speed aeronautical bearings. The other two on the contrary refers to real life applications such as a SAFRAN aeronautical engine and an Italian windfarm.

- In chapter 6 the proposed intermittent monitoring methodology is first tested on a synthetic dataset generated with different noise contamination levels. A novel algorithm for computing estimates of the spectral kurtosis (SK*) is introduced and compared to the reference algorithms selected from the literature review. Then, a second test of the methodology on real aeronautical engine acquisitions is performed, using the SAFRAN Contest dataset from Conference Surveillance 8 held in Roanne, France on October 20&21 2015. In detail, the Synchronous Average (SA) is compared to prediction-based algorithms such Linear Prediction (LP) and Discrete/Random Separation (DRS) for separating the gears deterministic signal from the non-deterministic contribution (bearings + noise). Finally, to highlight the bearing signature, the Fast Kurtogram (FK) taken as a reference is compared to the proposed novel spectral kurtosis estimator SK*, to the Empirical Mode Decomposition (EMD) and to the Stochastic Resonance (SR). The results of these interchangeable, alternative algorithms are compared mainly in terms of diagnostic performance and computational times.

- In chapter 7 Novelty Detection considerations are produced on the basis of Monte Carlo simulations. In particular, the thresholding operation for Mahalanobis Distance Novelty Detection is analysed and comparisons of the different criteria are made. In the 1-d case the Pierce's criterion is compared to Monte Carlo thresholding and to an Extreme Values Theory-based threshold, while for multivariate analyses, an EVT-threshold and the Wilks's critical values are compared against multivariate Monte Carlo thresholding. The curse of dimensionality is also tackled: the sample size for a proper MD-ND is investigated via Monte Carlo simulations, while considerations about the reliability of the covariance matrix estimation are made. Robust estimation procedures are compared on Monte Carlo repetition of signals from a simulated non-linear 1-D oscillator undergoing a white noise force.

- In chapter 8 the proposed methodology for processing the features extracted from the continuous monitoring signals is applied. The chapter starts analysing the DIRG test rig data using univariate statistics and hypothesis testing techniques such as the analysis of variance (ANOVA). Then, in order to "condensate" the information contained in the different features enhancing the effect of damage, multivariate analyses are applied. In particular, Fisher's Linear Discriminant Analysis (LDA), k-Nearest Neighbours (k-NN) classification, Principal Component Analysis (PCA) and Novelty Detection (ND) are applied. The issue of confounding influences is taken into account, and possible solutions are considered. As first, the Kernel Density Estimation (KDE) and the Gaussian Mixture Models (GMM) are used to improve the Novelty Detection by increasing the model complexity. Then, techniques for compensation of these confounding influences are tested (e.g. PCA orthogonal regression and whitening). The same procedure is repeated on a real-life dataset from the Italian Windfarm in Molise region.

- In chapter 9, in conclusion, the two methodologies are summarized and final considerations about the selected algorithms are given, together with the perspective for future improvements.

Bibliography

[1] René Descartes, *A Discourse on Method*, 1637

Chapter 2:

Vibration Monitoring: A Machine Learning Approach

1. Introduction

Vibration monitoring is a particular kind of condition monitoring in which vibration is used as a condition indicator. Hence, it can be considered an online Non-Destructive Testing (NDT) method. Vibration is a mechanical phenomenon whereby small oscillations about an equilibrium point occur, due to a continuous alternation between potential (strain) energy and kinetic energy. In general, it may be either desirable or not. This is for example essential for voice and music, as sound, or pressure waves, are generated by vibrating structures (e.g. vocal cords, etc.). On the contrary it is undesirable when it wastes energy creating unwanted annoying noise. It is the case of vibrating mechanical structures, which are usually carefully designed to minimize unwanted vibrations. In any case, even in good conditions, all the mechanical devices generate vibrations, but this unfavourable effect can be exploited to perform a monitoring of the health condition of a machine while the machine is in normal operation. The vibration measured on the housing of a machine, in fact, conveys information regarding its sources' conditions, usually hidden by noise. The monitoring can be then performed while the machine is operating, without stopping or disassembling it. Another advantage is that the vibration reacts very quickly to sudden changes, so that the moment in which a damage appears can be recognized soon after the generation. Furthermore, the accelerometers used in the monitoring are usually very small and light and can be added to instrument an existing machine without precluding the correct operation.

Obviously, both at an academic and at an industrial level, several different kinds of monitoring exist. Among all, it is important to remember:

- Acoustic: when a machine (or a component) is showing a defect or a damage, the sound produced during operation or via hammer testing can be different. Through microphones and pressure sensors then the airborne noise can be recorded and investigated to identify the sound sources.
- Acoustic Emission: it refers to solid-borne high-frequency acoustic signals generated by the developing of cracks or other permanent deformations. The damage detection in this case can be performed very soon, but the difficulty of application of the sensors limits its spreading.
- Oil Debris Analysis: when a damage occurs, metal chips or particles (usually called debris) may detach from the part. Using filters and magnetic chip detectors, a systematic analysis of the debris in the circulating lubricant system can be performed, analysing the quantity, type, shape size, etc. of such particles. The main limit is that the damage cannot be highlighted at an early stage, but just after some metal chip breaking off. When this occurs, a peak can be found in the recorded signal, but then no warning will be given until the next release.
- Performance Analysis: when the data regarding a machine functioning and performance are available, these can be used to monitor the health condition. For example, the power adsorbed by a motor in the usual work condition, or the

power produced by a windmill, are performance parameters which may be used as indicators for a possible damage. Unfortunately, such data are not always available.

- Thermography: by measuring the temperature of a part of a machine, it is possible to detect the presence of a damage. For example, in the last stages of life of a bearing, the rolling elements start to slide, leading to a substantial rise in the temperature due to friction. The main issue, then, is again the inability to detect incipient damage.

Thus, the vibration analysis has a number of advantages with respect to other methods. Moreover, it also has the ability to identify the actual faulty component, as different damaged parts show different "signatures" in the vibration.

In short, vibration analysis has an optimal cost/benefit ratio. Despite permanent transducers may have a high cost and should be built into the machine at design stage, the advantages outperform such weaknesses. Indeed, the minimum impact on the machine and the quick response to sudden faults are much more relevant to guarantee a reactive and reliable damage detection. The overall economic advantage, in fact, depends upon the possibility of improving the maintenance regime, switching from a preventive maintenance (i.e. time programmed) to a predictive maintenance (or Condition-Based Maintenance).

Preventive maintenance is a maintenance planned at regular time intervals, shorter than the expected *time between failures (tbf)*. This greatly reduces the possibility of catastrophic failure but has some big limitations. In particular, one should bear in mind that the *time between failures* is a statistical variable which can only be estimated from a population of machines. This means that the estimated *tbf* can show a large statistical variability (also depending on the machines population size). In order to guarantee a reliable maintenance regime, then, the *tbf* ensuring the absence of failure at a confidence of 95-99% should be used. This value can be many times larger than the expected *tbf*, as visualized in Figure 1 which refers to a practical example reported in Appendix 1.

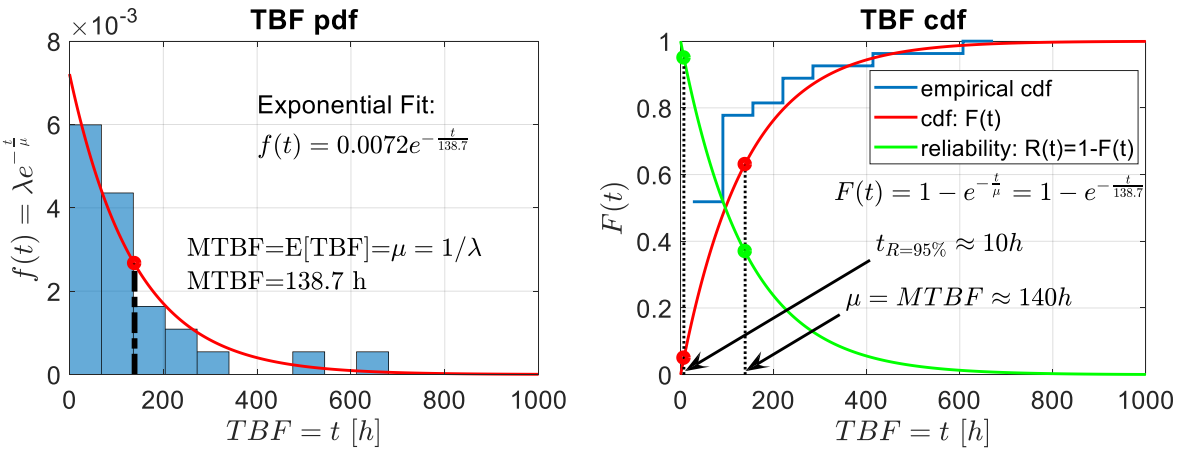


Figure 1: On the left, the fitted exponential PDF is compared to the empirical probability histogram obtained from the experimental data. On the right, the fitted CDF and the empirical CDF are shown, together with the reliability function. The MTBF and the 95% survival time $t_{R=95\%}$ are reported.

This means that in 1-5% of the cases, catastrophic failure can still occur, but, in most of the cases, healthy components are replaced, even if they could last for much

more time. This is obviously a waste, but it is the only way to limit the risk of catastrophic failure.

However, when reliable condition monitoring techniques are available, the potential breakdown of the machine can be predicted, and the maintenance can be carried out at the optimum time. This condition takes the name of preventive maintenance or Condition-Based Maintenance (CBM) [10] and it has become recognized as the best maintenance strategy in most cases, as it can strongly reduce the maintenance costs when appropriately implemented.

Diagnostics and prognostics are two important aspects in a CBM program. Diagnostics deals with fault detection, isolation and identification (i.e. it looks for the damage occurrence). Prognostics deals with fault prediction before its occurrence. The maintenance actions will be then based on the health information obtained by data processing of the acquired measurements of vibration. Obviously when catastrophic failure risk is high, permanent vibration transducers are used, so as to immediately react in case of damage establishment. In other cases, an intermittent monitoring can be implemented.

2. Permanent vs Intermittent Monitoring

In general, two types of Vibration Monitoring are possible. The selection of one instead of another basically depends on the risk of catastrophic failure and on the cost of the equipment itself compared to the cost of the production loss due to down times. Whenever the risk of catastrophic failure is very high, and a failure can cause costly damages to the machine, a permanent condition monitoring is desirable. In fact, it can ensure great reactivity to the sudden changes which may indicate a damage. Obviously, the cost of permanently mounted sensors is very high, and their employment should be evaluated at design stage, so that built-in sensors could be used. Furthermore, the data processing should be performed on-line, and needs to be very quick. Due to this, it is normally based on relatively simple parameters such as RMS, peak level, etc. Hence, the scope is limited to diagnose impending failure to give a warning in advance. And this is likely to be just a matter of hours or days, as opposed to the much longer periods of time which can be reached by other more advanced techniques. Of course, it is possible to add a second-level more detailed analysis in a non-continuous way. An intermittent monitoring can still be carried out in parallel and updated monthly, weekly or daily according to the need of the particular application.

In any case, for the vast majority of machines a permanent monitoring is not economically justified. The cost of the machine itself in fact can be outweighed by that of lost production. An intermittent monitoring is anyway desirable to update and optimize the maintenance plan. In this case less sensors will be used (typically one) together with a lighter data acquisition system usually made by a mere data logger. This simplicity will be then compensated by a more advanced data processing able to work out a highly detailed analysis giving long-term advance warning of developing faults.

3. Diagnostics via Vibration Monitoring: principles and definitions

In order to have a clear picture of a whole diagnostic system, an overall introduction about the motivation, the principles, the main physical components and the general procedure is fundamental. This paragraph is devoted to the materialization of the idea of vibration monitoring into a physical system performing diagnostics.

3.1. Damage vs fault

First of all, a definition of the used terminology is required to avoid misunderstandings, in particular when referring to the terms like defect, damage or failure, which can be vague and relative to a context.

From an engineering point of view, a defect is an appreciable deviation of the geometry of a component with respect to the theoretical design geometry. Obviously, imperfections exist in every structure and at any scale since all the manufacturing processes always show a degree of uncertainty (e.g. microscopic inclusions, roughness, dimensional and geometrical tolerances etc.). The term defect is then often used at a manufacturing level.

On the contrary, when a geometrical alteration appears while a machine is operating, it is usual to classify it as a damage. Incipient fractures or cracks and any alteration with respect to the healthy reference condition, belong to this type. Obviously cracks often originates from microscopical defects, so that the concept of damage and defect result linked.

In any case, nowadays, a damage tolerant design is common. To increase safety in fact, a component is usually engineered so as to be able to sustain defects while awaiting repair. Damage can be then accumulated until a critical level; the attainment of such a level is the so-called failure. The failure, in fact, can be considered as a damage growth (an increase in damage level) up to the limit for which the component can no more serve its intended purpose, and can potentially lead to catastrophic results.

The scope of a CBM regime is then to prevent failure. This can be done by diagnosing the presence of a damage, classifying its level and trying to get an estimate of the remaining useful life so as to improve the scheduling of the maintenance operations.

3.2. Diagnostics: principle and logic

The fact that damages cause changes in the dynamic characteristic of an item was a well-established knowledge since the first pottery was ever produced. The item struck by a finger or a hammer in fact, produces a sound which is different in case it is intact or in case it is showing a defect or a damage. The effects of a change in the dynamic characteristic of a component can be then measured and used for safety inspection and control in the very same way as a doctor can diagnose a heart malfunction by auscultating a patient's heartbeat.

Diagnostics is in fact a discipline which tries to get back to the cause of a malfunction starting from an observation. From a mathematical logic point of view, according to Peirce definition [1,2], it can be considered an abductive reasoning. Abduction is a form of logic inference which seeks to find the most likely explanation for an experimental observation. Knowing the rules and results, the premise can be then inferred at a degree of confidence.

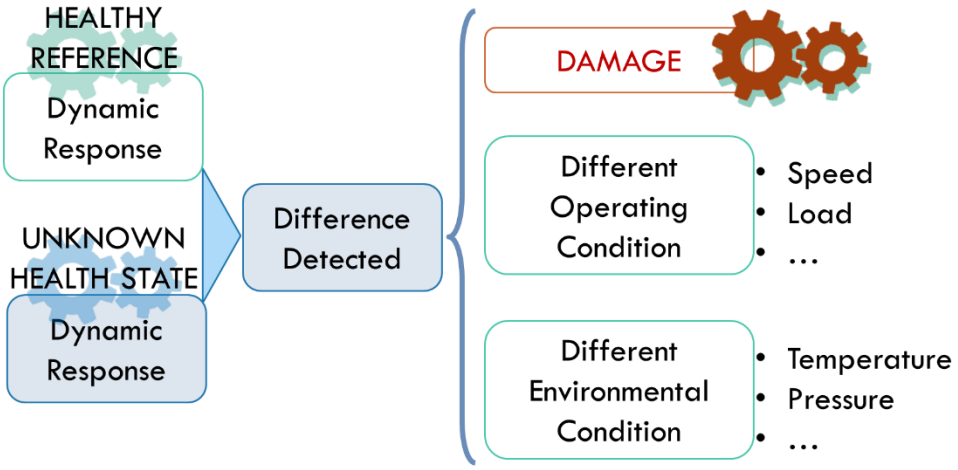


Figure 2: Estimation of the health condition from the data via abductive reasoning

Important considerations follow:

- No sensor can directly measure damage, only its effects can be detected.
- A change of the dynamic characteristic can be induced by the establishment of a damage, but also by other mechanisms.

Therefore, to minimize the remaining level of uncertainty of the inference, one should also exclude the other possible explanations. For example, environmental variables (e.g. temperature) can induce changes in the material properties, so that a different dynamic response is obtained. Such an effect must be considered in the diagnostic inferential process: either environmental variability must be excluded by controlling the environmental variables or, at least, it must be compensated with some particular method. Diagnostics is then a matter of damage identification from a measurement of a dynamic response. These considerations are summarized in the scheme in Figure 2.

3.3. Diagnostics: objectives and categorization

As widely described, the final objective is the optimization in terms of safety and cost-effectiveness of the maintenance regime of a machine (i.e. *wisdom*) thanks to the *knowledge* about its state of health. This knowledge is reached through some sort of *information* extracted by the experimental measurements of the dynamic response of a system (i.e. *data*) via diagnostic abduction.

The knowledge achieved from condition monitoring can range over different levels which were concretized by Rytter in 1993 [3]. His four-categories hierarchical definition of the damage identification problem was later modified by Worden and Dulieu-Barton in 2004 [4] to produce the following sequence:

- Level 1 – DETECTION: A qualitative indication that damage is present in the structure (N.B. possibly, at a given confidence).
- Level 2 – LOCALIZATION: Knowledge about the probable location of the damage (i.e. in which component?).
- Level 3 – CLASSIFICATION: Knowledge about the damage type (crack, spall etc.).
- Level 4 – ASSESSMENT (or Quantification): Information about the damage size.
- Level 5 – CONSEQUENCE (or Prognosis): Knowledge of the actual degree of safety. How far is the component from failure?

The first four levels are usually included in the definition of diagnostics, while the last one belongs to prognostics, a discipline focused on predicting the time at which a system or a component will fail, which corresponds to its Remaining Useful Life (RUL), the final and most important decision-making parameter. It is relevant to highlight that success at any level of the hierarchy depends upon having successfully achieved all the prior levels.

The process of reaching knowledge from the information contained in the experimental data follows in general two approaches, as a function of the a priori knowledge strength and of the amount of data available:

- *Model-based* – When the preconceptual knowledge of the system (e.g. physical laws) is high enough, a model can be built so that the damage identification can be seen as an inverse problem. New data is used to update the model parameters, whose variation will be put in relation with the presence of a damage.
- *Data-based (or data-driven)* – On the contrary, when the a priori knowledge is weak (e.g. for a complex machine like a gearbox), the damage identification problem will rely on the data, and it can be traced back to the field of pattern recognition. Basically, a data mining is performed to discover the regularities (i.e. the rules) which allows to classify the data into different categories.

This work will mainly focus on the data-driven methods, which is preferable for real, complex machines, whose true signals are commonly too difficult to be modelled. The use of models will be then limited to some signal separation algorithms or to the computation of relevant parameters such as the characteristic frequencies of the gears or of the bearings, which can be used as *features* (paragraph 4.3 and 5).

4. Damage identification as a Pattern Recognition problem

The process of building wisdom from data, often referred to as D2D (Data to Decision) process, was neatly summed up by Farrar and Doebling in 1999 [5] for Structural Health Monitoring (SHM), which is the analogous of Condition Based Monitoring to structures. Notice that CBM is often used with reference to rotating machinery, while in SHM “structural” has usually a civil engineering connotation.

Conceptually, even if CBM and SHM can be considered analogous, they are meant for different applications and have then to face different conditions and requirements. In particular, the size of the structures investigated by SHM (e.g. bridges, buildings, etc.) is usually relatively larger than the rotating machines object of CBM. Furthermore, even if some environmental parameters such as the temperature can affect CBM machines, these are often installed in controlled environments. This is not the case of SHM structures, which are commonly exposed to the open air and affected by seasonality. Additionally, unlike in SHM, failure detection and identification are usually more precise in CBM, as in many cases it is possible to find specific dynamic responses for specific fault classes [16].

In any case, the overall paradigm of SHM holds also for CBM, so that it is worth to start by studying it.

The fundamental steps which can be found in the literature [4,5,17] are summarized in a waterfall model which will be analysed in detail. This is summarized in Figure 3.

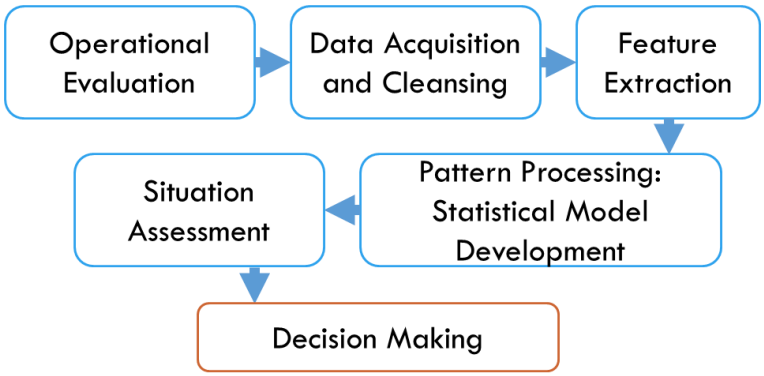


Figure 3: SHM and CBM flow chart [4,5]

4.1. Operational evaluation

In order to optimize both the costs and the performances of the damage identification, system-specific condition monitoring apparatus needs to be implemented. In this regard, the design limitations which can affect the diagnostic equipment may range from the operational and the environmental conditions undergone by the system, up to the available volume, weight, and powering system. This information can be summarized to define a budget for the requirements of the sensing unit and the data transmission which will be selected in the next steps. Also, information regarding the expected damage type and location can be relevant to decide whether a *local* monitoring, targeted on a particular component, can be preferred or added to a *global* monitoring investigating the whole machine.

4.2. Data acquisition and cleansing

The definition of the data acquisition hardware is not trivial. Every different stage of the measurement chain illustrated in Figure 4 needs to fit the budget while fulfilling the requirements imposed by the subsequent analysis.

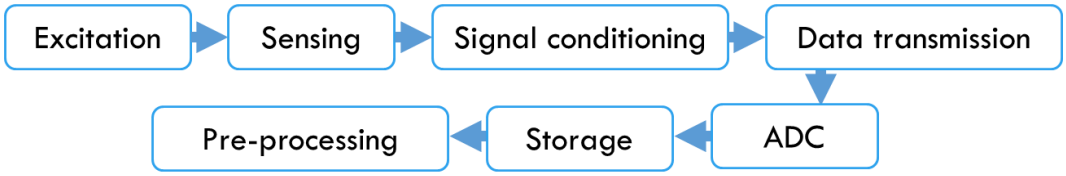


Figure 4: Data acquisition flow chart

4.2.1. Excitation

In order to measure the dynamic response of the system under analysis, this must be excited with a forcing term. Such an input signal can be either controlled or not. In the first case, an actuator can be used to generate a force which follows a given time signal (e.g. random or deterministic, stationary or non-stationary, etc.). Obviously, different kind of actuators exist with different performances in terms of force and frequency ranges: hydraulic, electro-dynamic, piezoelectric etc. When the force is controlled and recorded, the dynamic response of the system can be normalized over such input signal, so as to find a Frequency Response Function (in frequency domain, or an Impulse Response in time domain) which is a system invariant independent from the excitation level (N.B. when the assumption of system linearity holds). Nevertheless, in many other situations the excitation cannot be measured so that a response-only analysis is the only choice. It is the case of the ambient excitation (e.g. wind, etc) or the excitation due to internal

phenomena. Focusing on gearboxes for example, it is common to use the shafts rotation as excitation source.

4.2.2. Sensing and signal conditioning

The complete definition of the sensors to be employed in terms of location, number and type is typically performed according to some parameters like sensors cost, dimensions, mass, reliability, stability and required dynamic performances (e.g. bandwidth, frequency resolution, dynamic range, amplitude sensitivity, etc.).

In general, "sensors type" may refer to multiple categorizations. The physical principle on which the sensor relies (e.g. piezoelectric, piezoresistive, capacitive, etc.) or the sensors contact versus non-contact nature (e.g. accelerometers, acoustic emission, strain gauges, force transducers etc. vs laser sensors, eddy current sensors etc.) for example. In this work anyway, sensor type is used to define the physical quantity the transducer is sensitive to. Many types of transducers exist for measuring all three of the parameters in which lateral vibration can be expressed, namely displacement, velocity and acceleration. However, the only practical VM transducers are:

- Proximity probes for relative displacement: capacitive, inductive, eddy current sensors, magnetic, Hall effect – *non-contact.*
- Optical position measurement: laser interferometer, laser triangulation, laser time-of-flight, chromatic confocal sensors – *non-contact*.
- Velocity pickup: seismically suspended coil in the magnetic field of a permanent magnet attached to the housing of the sensor – *contact*.
- Accelerometers: piezoelectric elements sandwiched between a mass and the housing base; such crystals generate an electric charge proportional to strain – *contact*.

The signal is usually conditioned by the sensor to produce a voltage output which will be sent to the digital acquisition system. For example, the most common accelerometers exploit piezoelectric crystals to translate the acceleration information into a charge information. An operational amplifier-based circuit (a charge amplifier) is then integrated in the sensor to produce a voltage output proportional to the integrated value of the input current.

4.2.3. Data transmission

The physical connection of the sensors to the digital acquisition system is another issue to be considered. Fundamentally two kind of transmission can be taken into account: wired vs wireless. The difference is in the transmission medium as wired transmissions need copper wires or optical fiber cables to carry different forms of electrical signals from one end to the other, while wireless communication basically relies on electromagnetic waves which can travel not only through air and solid materials, but also through vacuum.

Traditional wired transmission is obviously more reliable and is unmatched in terms of speed of transmission and bandwidth of the signal. Nevertheless, the installation can be cumbersome and limits the mobility. In particular cases then, wireless transmission is the only solution, as it ensures more flexibility at lower maintenance costs (e.g. rotating components). Wireless transmissions are often digital, so that in most of cases the analog-to-digital conversion (ADC) is moved upstream.

4.2.4. Digital acquisition: conditioning, ADC and storage

Digital signals show many advantages with respect to analog. Digital data in fact, can be stored and retrieved very easily. No bulky recording medium (e.g. magnetic tape, rotating drums, etc.) is needed, with the advantage of being noise immune. Indeed, data stored in electronic supports (e.g. Hard Disk Drives, Solid-State Drives, flash memory etc.) are not prone to deterioration or noise contamination. Furthermore, signal compression is possible to reduce the data volumes without losing information. The possibility of performing time multiplexing is another great advantage which allows to send many digital signals at the same time (just a little delayed) on the same channel.

The Analog to Digital conversion is then the fundamental step of a digital acquisition system. The analog output of the sensor must be discretized in both amplitude and time [Appendix 3]. This operation is mainly ruled by few parameters:

- *Sampling frequency* f_s: by means of a clock, the acquisition system is able to read amplitude values at equispaced time instants, so that just one sample every time increment $\Delta t = \frac{1}{f_s}$ is recorded. It is important to remember that, to avoid the phenomenon of aliasing, the analog signal is usually pre-processed with an analog anti-alias filter limiting the frequency content to f_c. Typically, setting $f_s = 2{,}56 f_c$ ensures that no frequency component exceeding the Nyquist limit $f_{ny} = \frac{f_s}{2}$ will be present to introduce aliases. Finally, a high f_s may seem a good choice regarding at time resolution, but it implies heavier recordings. A limit on the duration of the acquisition, can lead to a coarser resolution in frequency domain $df = \frac{1}{T} = \frac{f_s}{N}$.
- *Amplitude range* E and *number of bits* B: During the process of amplitude quantization, the analog signal amplitude is rounded to the nearest discrete value. The amplitude resolution ΔA, or quantization step, is imposed by the amplitude range and the number of bits of the DAC by the relation $\Delta A = \frac{E}{2^B - 1}$. The range E can be often selected to improve the amplitude resolution, while avoiding the possible *overloads* (signal amplitude exceeding the limits imposed by the range). DC vs AC coupling

should also be evaluated, as, when the mean value is not relevant, it can be filtered away (AC) at the advantage of a better use of the available range.

The digitalized signals are finally stored.

4.2.5. Data Pre-processing and acquisition system self-diagnostics

Before any further analysis of such stored signals, it is common to pre-treat the data to maximize the reliability and the efficiency of the system. In particular, four steps are usual:

Data cleansing

In order to highlight possible malfunctioning of the sensors, the cables or the acquisition system a consistency check is highly recommended. This self-diagnostic step, resulting in the rejection of corrupted data, is fundamental to increase the reliability of the whole machine-diagnostic system.

Data Normalization

Different signals often show different magnitude, but this may prevent an effective comparison. Furthermore, some trends related to environmental or operational variability may mask the presence of a damage. To foster the detection, then, a normalization is needed.

- Centring: Mean removal or trend removal are fundamental steps to isolate damage variability from certain “external” gross influences. For example, the vibration level can be related to the load cycle. When this effect is evident, it must be compensated, otherwise it will introduce a confounding effect to the detection.
- Re-scaling: Signals are commonly re-scaled to a notionally common nondimensional scale, so that different measurements can be compared meaningfully.

Data Fusion

Typically, many sensors could be used at the same time. Integrating multiple data sources then, may produce more consistent, accurate, and useful information. All data can be brought together into a single view in which a more complete picture of the system is created. This can enhance the damage detectability.

Data compression

The amount of data to be stored can be massive. Because of this, it may be reasonable to encode information using fewer bits than the original, so as to reduce the resources required to store and transmit data, at the cost of some computational resources consumed in the compression and decompression processes. Compression may be either lossy or lossless. In the diagnostics framework the second must be preferred, as no information should be lost at this step. The reduction should then identify and eliminate only statistical redundancy.

It is important to point out that the **redundancy** in the data which is advisable to remove at this stage (i.e. after the cleansing), corresponds to a redundancy in the sensors network which is usually added by purpose. If sensors output is correlated and consistent in fact, a sensor failure can be noticed, and actions can be taken (e.g. discard the data

from the anomalous sensor) abating the risk of missed alerts, which may have a severe cost and safety implications. Therefore, redundancy or diversity should be built in as an error detection and correction mean (self-monitoring), to increase the performance, the robustness and the reliability of the diagnostic system which will then feature a fail-safe design: the failure of a sensor causes no or minimal harm to other equipment, the environment or to people and does not prevent the system from working properly.

4.3.Signal processing: Feature selection, extraction and metric

The scope of the data mining, as largely understood, is to link a symptom appearing in the signal to the presence of a damage, and later to distinguish the damage type, location and severity, which corresponds to highlighting patterns in the data. Unfortunately, raw data is often a messy sum of different effects corrupted by noise. It is typical then to use a priori knowledge and engineering judgements to focus the attention on dominant traits which are known to be sensitive to damage. Such damage distinguishing characteristics are usually called *features*. The selection of the most promising features is then a critical step, which must be tackled focusing on the sensitivity to incipient damage while maximizing the accuracy and stability of the detection.

If the features, for example, show a high fluctuation in the measurement from normal, healthy condition, it will be harder to notice a deviation due to damage, unless the damage is severe. There is clearly a dependency between the resolution of the diagnosis and the noise rejection capabilities of the algorithm. Hence, features should be selected to limit as far as possible the fluctuations on the normal condition data.

It is interesting to point out that this procedure mimics what human beings and animals can easily do in an eye-blink. Colours, shapes, dimensions, proportions can be distinguished by the brain very easily and are used as features to classify every single entity entering the field of vision. For illustrative purposes, two Monna Lisa portraits are reported in Figure 5. Different features can be extracted, but only some will highlight the hidden patterns and enable a recognition. Hair and eyes colour for example are not exhaustive but focusing on proportions (e.g. head dimension with respect to the body) the human brain of the reader can with no doubt recognize the Leonardo da Vinci portrait from the Botero interpretation.

Figure 5: Monna Lisa from Leonardo da Vinci and Botero perspectives.

Features are then fundamental to enhance the peculiarities of the different patterns to be recognized. But for a reliable classification, quantitative features are preferable. A metric to quantify the level a feature in fact, is fundamental to the recognition process.

Suppose now that the raw data acquired is a time-series of accelerations from the outside of a gearbox casing. Most of the vibration will be directly linked to periodic events in the machine's operation, such as rotating shafts, meshing gear-teeth, etc. then in the signal spectrum, particular spectral lines will appear, like the meshing frequency and its harmonics. These are known to be sensitive to damage, as the engineering experience highlights the high correlation and dependence. Such spectral lines can be then used as features. This also enable to reduce the dimension of the data discarding the parts of the signal which does not contain useful diagnostic information. High level features can be then extracted using sophisticated signal processing algorithms, which aims to highlight the signal of interest with respect to the noise (increase the signal to noise ratio), to compensate for the transmission path from the source to the sensor and to isolate the different sources to enhance their contribution (Blind Source Separation). These algorithms are generally applied for intermittent monitoring, when a deeper analysis can be performed. On the contrary, a permanent monitoring usually requires faster responses for a real time implementation, so that lower level features (e.g. the common time domain RMS, skewness, kurtosis etc. – Chapter 5, Section 2.1) may result more appealing. The extraction of high-level features through established signal processing techniques will be examined in section 5.

4.4. Pattern processing: Statistical model development and validation

Once the features are selected and extracted, an intelligence should be used to univocally put in relation the statistically significant changes in the features to the presence of a damage. This can be performed by an expert looking at the features, or, more likely, by a statistical model. The last, in fact, in addition to the high effectiveness, can also give a quantitative information about the confidence of the estimated state of health, which depends on the natural fluctuation of the healthy features, but also on the amount of data used to train the algorithm. Indeed, a statistical model mimics the cognitive function of "learning" (machine learning) and applies it to the recognition of data corresponding to a healthy condition from the data produced by a damaged state, so that it must be trained for this purpose.

Four different kind of algorithms can be distinguished depending on the desired diagnosis (enriched Rytter levels, paragraph 3.3) and on the available data:

- *Classification*: a discrete, multi-class supervised problem. Level 1 to 4 diagnostic can be achieved with such algorithms if supported by the necessary training data (different labels for different damage types, locations and severities). Generally, the probability of a class membership can be easily retrieved from such algorithms.
- *Novelty detection*: a two-class supervised (or sometimes semi-supervised) problem. Level 1 diagnostic can be tackled just by looking for discordancy of the new data with respect to the training, healthy data. After having removed any other external influence in fact, damage is the only remaining cause of inconsistency and can be detected just by comparison against a limit threshold.

- *Regression*: a continuous-class supervised problem. This is often used for Level 2 or 4, in which the variable of interest is continuous e.g. the cartesian coordinate of the damage location or the damage size. Regression can be also extended to sortable classes using logistic functions able to map from a continuous variable to a categorical set, possibly enabling Level 1 and 3.
- *Clustering*: a discrete, multi-class unsupervised problem. It is an exploratory data mining trying to group a set of acquisitions in a way that the data in the same class (or cluster) are "more similar" in some sense (i.e. distance according to a selected metric) to each other than to those in other clusters.

In particular, if the machine learning algorithm is trained on labelled data (i.e. a training example is a *pair* of input-output information: the corresponding health state is known) it is said to be **supervised**. In the other case, when the learning algorithm must identify commonalities in non-labelled, non-classified or non-categorized data, it is called **unsupervised**. In any case, acquisitions in a damaged condition may be dangerous, so that in most of cases an engineer will rely only on healthy acquisitions. The possibility of training an algorithm with a machine believed healthy, but actually damaged, albeit improbable, must always be considered, as it could strongly affect the performance of the algorithm. An Outlier Analysis is then always recommended before the training, so as to prevent this eventuality.

Finally, an important part of the statistical model development process is the **test stage** on actual data collected in a so-defined validation set. The performance of the selected features in terms of effectiveness of the detection and damage sensitivity can be then established together with the prominent information about the so-called false alarm (FA) or missed alarms (MA). In particular, from a statistical point of view, Level 1 diagnostics can be seen as a test for the null hypothesis "H_0: $the\ machine\ is\ healthy$ ". In this regard, as also highlighted in Figure 6, two kind of errors are possible. The $type\ I\ error$, or rejection of a true H_0, corresponds to an indication of damage when none is present (FA) and can erode the confidence of the damage detection. The $type\ II\ error$, or failure in rejecting a false H_0, on the contrary, is a missed indication of damage although present (MA), which can be very detrimental as can bring serious economic and life-safety implications. In any case, depending on the application, minimum amounts of FA and MA are acceptable in a reliable monitoring system.

		True Health Condition:	
		Healthy (H_0)	*Damaged*
CBM Actions	*accept H_0: Healthy*	No Alarm	**Missed Alarm *type II error***
	reject H_0: Damaged	**False Alarm *type I error***	Alarm

Figure 6: Type I and II errors in hypothesis testing for CBM

4.5.Situation assessment and decision making

The final stage of the waterfall model is to take a decision. Hence, an intelligence has to use the knowledge about the health state of the machine produced by the pattern recognition algorithm to decide whether an action needs to be taken and what that action should be. In case of intermittent monitoring, this intelligence can be the human brain of an expert engineer, which looks at the features trends and evaluate, on the basis of its experience, if the machine can keep on working properly or if an intervention is needed. Obviously, this process is critical and shows some issues related to the unambiguousness of the interpretation, the number of parameters which can be considered at the same time and their correlations, and the decision times. Because of this, an Artificial Intelligence (AI) can be used to substitute the human contribution in the assessment of the state of health of the machine. In particular, engineers' knowledge can be captured to form a set of rules which emulates their experience. An "expert" system can then determine or recommend an action on the basis of the sensors data, at confidence and speed which are unmatched by the human brain. According to the European Aviation Safety Agency (EASA) the response of a diagnostic system is the triggering of an *alert,* namely an indication that requires further investigation or an *alarm,* an alert that needs a corrective maintenance action [6]. (N.B. note that the word "alarm" is often used in a figurative way, not to be confused with the alert vs alarm definition of this paragraph, which refers to real actions taken from the diagnostic system).
Different alerts can be produced according to a given colour code:

1. Red for alerts that require action prior to the next flight (Level 1 alert).
2. Yellow or Amber for Level 2 alerts that require maintenance personnel awareness, but do not preclude continued operation.
3. A third, distinct colour of the applicant's choice (e.g. Green), for "advisory" conditions that may influence future maintenance.

It is important to mention that, at the moment, the use of an AI in *safety-critical* applications is not allowed by regulatory bodies such as the EASA, the UK Civil Aviation Authority (CAA) or the US Federal Aviation Authority (FAA). In case of component failures leading to death or serious injury to people, to severe damage or disruption of equipment and eventually to environmental harms, in fact, operating at the margin of safety without inspections is not an option, even if the maximum advantage of a CBM would be in such cases. Regulatory bodies are anyway pushing for technological improvements to ameliorate the existing systems such as the Health and Usage Monitoring Systems (HUMS). This commercial system equips many helicopters often implementing *usage monitoring* (i.e. flight time recording and exceedance monitoring, in case some parameters as oil temperature and pressure or shaft torque etc. exceed the predefined levels), *oil debris detection* and *vibration monitoring* of the drive train (i.e. the gearbox) but many shortcomings remains. In particular:

- Very few maintenance indications are produced,
- The damage detection is sometimes below expectations,
- No accurate indication of "Go/No go" is given.

These drawbacks leave scope for a large number of possible improvements.

Bibliography

[1] Peirce C. S., "*Prolegomena to a science of reasoning: phaneroscopy, semeiotic, logic*", Peter Lang AG, 2015. ISBN: 9783631666029 3631666020.

[2] Fann, K.T. "*Peirce's Theory of Abduction",* The Hague: Martinus Nijhoff, 1970. DOI: 10.1007/978-94-010-3163-9

[3] Rytter A., "*Vibration based inspection of civil engineering structures"*, Ph.D. thesis. Department of Building Technology and Structural Engineering, University of Aalborg, Denmark, 1993.

[4] Worden K., Dulieu-Barton J.M., "*An overview of intelligent fault detection in systems and structures",* Structural Health Monitoring, 3 (1), 85-98, 2004. Doi: 10.1177/1475921704041866.

[5] Farrar C.R., Doebling S.W., "*Damage Detection and Evaluation II".* In: Silva J.M.M., Maia N.M.M. (eds) Modal Analysis and Testing. NATO Science Series (Series E: Applied Sciences), vol 363. Springer, Dordrecht, 1999. ISBN 978-0-7923-5894-7.

[6] EASA ED Decision 2018/007/R CS-29 Amendment 5, 2018, "*Certification Specifications and Acceptable Means of Compliance for Large Rotorcrafts"*.

[7] Antoniadou I., "*Accounting for nonstationarity in the condition monitoring of wind turbine gearboxes",* PhD thesis, University of Sheffield, 2013.

[8] Jardine A.K.S., Lin D., Banjevic D., "*A review of machinery diagnostics and prognostics implementing condition-based maintenance",* Mechanical Systems and Signal Processing, 20, 1483- 1510, 2006. Doi: 10.1016/j.ymssp.2005.09.012

[9] Carden E. P., Fanning P., "*Vibration Based Condition Monitoring: A Review"* Structural Health Monitoring, 3(4), 355-377, 2004. Doi: 10.1177/1475921704047500.

[10] Randall R.B., "*Vibration-based condition monitoring: Industrial, Aerospace and Automotive applications*", Wiley, 2011. ISBN: 978-0-470-74785-8

[11] Antoni J., Randall R. B., "*Rolling Element Bearing Diagnostics - A Tutorial",* Mechanical Systems and Signal Processing, 25(2), 485-520, 2011. Doi: 10.1016/j.ymssp.2010.07.017.

[12] Antoni J., Randall R. B., "*Unsupervised noise cancellation for vibration signals: part I—evaluation of adaptive algorithms"*, Mechanical Systems and Signal Processing 18(1), 89-101, 2004. Doi: 10.1016/S0888-3270(03)00012-8.

[13] Antoni J., Randall R. B., "*Unsupervised noise cancellation for vibration signals: part II—a novel frequency-domain algorithm"*, Mechanical Systems and Signal Processing 18(1), 103-117, 2004. Doi: 10.1016/S0888-3270(03)00013-X.

[14] Yiakopoulos C.T., Gryllias K.C., Antoniadis I.A., "*Rolling element bearing fault detection in industrial environments based on a K-means clustering approach*", Expert Systems with Applications, Volume 38, Issue 3, 2011. DOI: 10.1016/j.eswa.2010.08.083.

[15] Fassois S.D., Sakellariou J.S., "*Time-series methods for fault detection and identification in vibrating structures",* Phil. Trans. R. Soc. A (2007) 365, 411–448, 2007. Doi: 10.1098/rsta.2006.1929

[16] N. Dervilis, *"A machine learning approach to Structural Health Monitoring with a view towards wind turbines"*, Ph.D. thesis, University of Sheffield, 2013.

[17] C. R. Farrar, K. Worden, *"Structural Health Monitoring: A Machine Learning Perspective"*, John Wiley & Sons Inc, 2012. ISBN: 978-1-119-99433-6

Chapter 3:

Literature Survey: The state of the art

1. Vibration Signals from machines

Measured vibration signals are always combination of source effects and transmission path effects, so that a measurement is practically a sum of responses from the different sources, which can be modelled as the output of a multiple input system (MISO) [1,2].

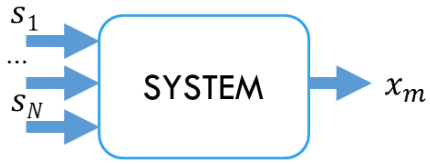


Figure 1: Measured response x of a MISO system due to multiple excitation sources.

In the time domain, the contribution of one source to the output corresponds to the convolution of the force signal with the impulse response function of the transmission path from the source to the measurement (IRF).

$$x_m = \sum_j s_j * h_{mj} \qquad X_m = \sum_j S_j \, H_{mj} \qquad (1$$

where x_m is the measured signal, s_j is the force signal and h_{mj} is the IRF, corresponding to the frequency response function (FRF) H_{mj} in the frequency domain. This can be for example a Mobility (vibration velocity output over force input) which "distorts" the measured output according to its own resonance peaks. In any case, the transmission path resonances not always appear in the measured spectrum: in case of a strong broadband noise excitation, resonance peaks will be present, but for the common discrete frequency sources typical in many machineries, they will not be highlighted unless directly excited by a forcing frequency.

In a simplified way, a pure sinusoidal forcing function generates then a sinusoidal component (the fundamental frequency) in the response signal together with additional harmonics induced by structural non-linearities. Anyway, different forcing functions can cause modulation effects giving rise to families of sidebands around the harmonics of the "carrier" frequency. The strength of the frequency analysis is then its ability to highlight such effects allowing a first-order visual separation of the different harmonics and enabling accurate measurements of the fundamental frequencies.

Indeed, the basic problem of damage detection is finally to decide whether some changes in the response signal are due to a change in a source or in the system, which is in general a blind source separation problem (BSS).

1.1. General signal classification

In order to understand which kind of source signals can be found, a general signal classification is schematized in Figure 2.

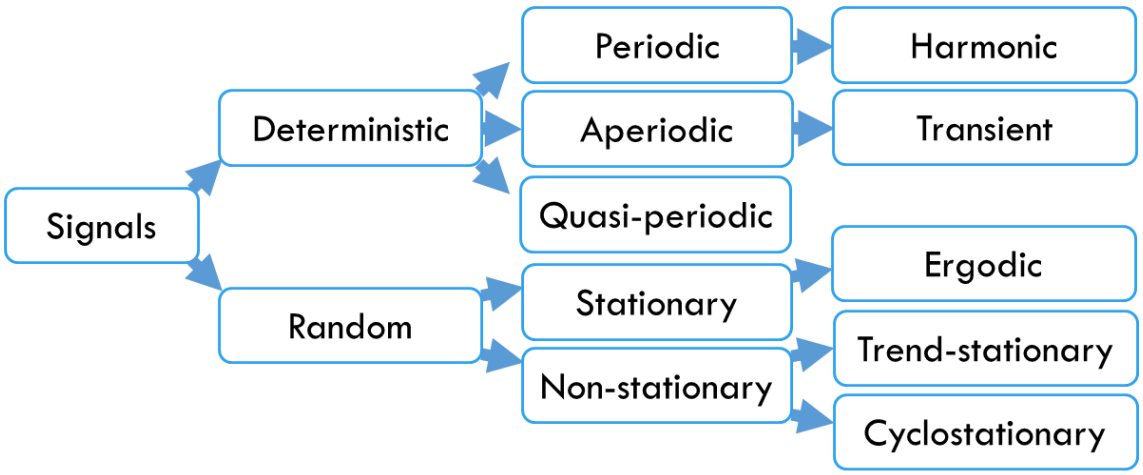


Figure 2: Signal classification [1,2]

In this classification, *deterministic* refers to a signal which can be predicted at any time in the future or past. On the contrary, *random* means that its values in time are unpredictable. Nevertheless, for the *stationary random* signals, whose statistical properties are not changing with time, a deterministic characterization is possible in the frequency domain, as their spectrum is proved to be deterministic [Appendix 3]. For example, the so-called **white noise** is a random signal having equal intensity at different frequencies, giving it a constant power spectral density. Individual random signals may be considered as realizations of a "random process", consisting of an ensemble of realizations. A random process is then *ergodic* if its time average is the same as its average over the different realizations.

Among the deterministic, the **harmonic** are the main *periodic* signals of interest in this work. They are composed entirely of discrete frequency sinusoids whose frequency is an entire multiple of a fundamental frequency. When the frequencies are not entire multiples of the fundamental, *quasi-periodic* signals arises. Other aperiodic signals which can be found in machines are the *transient*. Transient signals practically exist only for a finite length of time (they theoretically decay to infinity, but they show measurable values just for a finite time) and are typically analysed as an entity. Finally, among the non-stationary, the most relevant are the *trend-stationary* signals, which can be transformed into a stationary signal by removing the underlying trend, and the **cyclostationary** signals, random signals which vary cyclically with time.

1.2. Gearbox signals

In a gearbox, the main sources of vibration are:

- Shaft unbalance, misalignment or cracks
- Meshing gears
- Rolling element bearings local damages
- Noise

while the transmission path often involves the structure and the casing of the machine. Such sources contribute, convolved with their transmission paths, to the measured vibration according to the MISO model introduced at the beginning of this chapter.

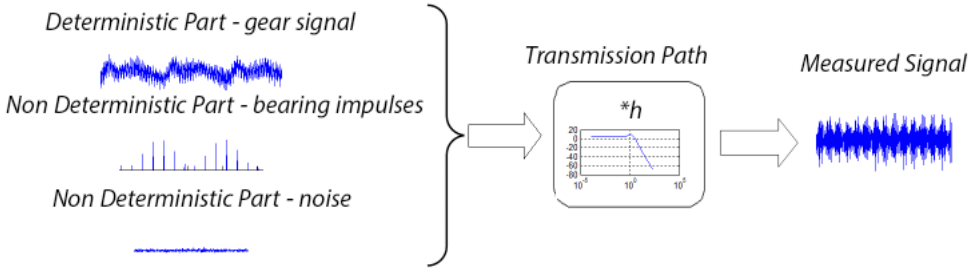


Figure 3: Measured signal from the MISO perspective.

To improve the diagnostic process, it is important to understand how the presence of a damage affects the machine vibration response. This analysis is usually done at constant environmental and operational conditions (speed and load), so as to isolate and highlight the effect of a damage. In this condition, a change in the health state of a component, results quite often in a change at a source. It is the case of an increase in a shaft unbalance, which only affects the shaft forcing function. On the other hand, some faults may primarily lead to a change in the structural response, like a crack in the machine casing. Moreover, the two effects can sometimes couple with each other, so that a change in the structural response gives a variation in the forcing function. For example, a tooth root crack affects both the force at the tooth-mesh and the local structural stiffness involved in the transmission path.

In any case, the damage induces a change in the resulting vibration signal which can be appreciated and recognized in the vibration spectrum. The characteristic spectral lines forming the so called “signature” can be then used as features to distinguish among the different damages in the different components. The results of Randall’s analysis [2] based on failure models and practical tests is reported in the following paragraphs for the main different sources of vibration.

1.2.1. Shaft signature

A number of faults related to the rotating shaft manifest themselves at a frequency corresponding to the speed of the shaft in question. In addition to such fundamental frequency, low harmonics and subharmonics can also be found. In particular, the signature is affected by

- **Rotor unbalance**: When the shaft mass is not perfectly distributed around its centre of mass, a centrifugal force rotating at Ω and proportional to $Me\Omega^2$ is generated, where M is the shaft mass, e is the radial displacement of the centre of mass of the rotor (eccentricity) and Ω is the shaft rotational speed. Causes of unbalance include design defects (e.g. poor design

tolerances leading to eccentricities, non-compensated addition of keys and keyways …), manufacturing defects (e.g. blow holes or sand traps in metal casts …), distortions (deformations induced by temperature or high stresses), corrosion, wear, deposits etc. In any case, unbalance gives vibration responses at shaft speed ($1 \times \Omega$) and its low harmonics, mainly in radial direction. To quantify the degree of unbalance of a rotor, the ISO 1940 norm for rotors balance quality requirements [3] establishes a classification of the balance quality requirements for typical machinery in terms of the balance quality grade $G = e\Omega_{MAX}$ in $\left[mm \cdot {}^{rad}/_{s}\right]$.

- **Misalignment**: When two shafts are coupled, two possible alignments errors can occur. A radial misalignment appears when the two shafts are mounted parallel, but a radial offset remains; when the two shafts' axes are incident on the contrary, an angular misalignment verifies.

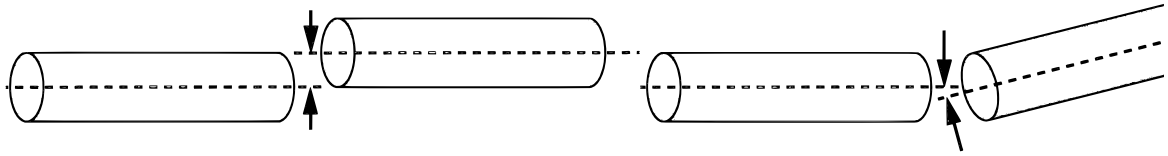


Figure 4: Radial and Angular misalignment

Such misalignments introduce a spatially-fixed bending deflection, which is then rotating with respect to the shaft, and generating fluctuating moments and forces, typically varying twice per revolution. Hence, misalignment gives vibrations in the axial direction at the low harmonics of shaft speed, with some preference for the even harmonics, in particular the second ($2 \times \Omega$).

- **Crack**: A crack in the shaft is a severe fault, involving the reduction of the load supporting section which can lead to failure. Unfortunately, even large cracks minimally affect the natural frequencies of the system, in particular when they are closed. In general, both the amplitude and the phase of the vibration at the low harmonics of the shaft speed may undergo small variations: sometimes the amplitude is first reduced while the crack enlarges because of opposing phases. Breathing cracks on the contrary often give greater changes at the third harmonic ($3 \times \Omega$).
- **Whirl**: A number of phenomena cause the centre of the shaft to whirl (namely to show a precessional orbit either forwards or backwards) at a frequency possibly different from the rotation speed. A synchronous whirling motion can be naturally experienced by flexible shafts supporting unbalanced discs (e.g. Jeffcott rotor). The support flexibility should be also taken into account at design stage, so as other whirl-generating phenomena. Whirl, in fact, may turn unstable and become destructive (*whip*), so that it must be controlled. Other forms of whirl induced by the journal bearings can be either a *dry friction* or an *oil* whirl. In any case it may show up under certain conditions (e.g. low load on the fluid film bearing) at a frequency close to one-half of rotor angular speed. Subharmonics will be visible in the spectra ($0.5 \times \Omega$).

1.2.2. Gear signal

Many different designs are possible for gears. Anyway, the simplest form is known as *spur gear* and is used to transfer power between two parallel shafts, usually running at different speeds. The transmission is performed through conjugate profiles designed to keep a constant transmission ratio τ during the whole meshing cycle:

$$\tau = \frac{\omega_o}{\omega_i} = \frac{n_o}{n_i} = \frac{f_o}{f_i} = \frac{\theta_o}{\theta_i} \equiv \frac{r_i}{r_o} = \frac{z_i}{z_o} \qquad (2$$

where ω is the shaft speed in rad/s,n is the shaft speed in rpm, f is the corresponding shaft frequency in Hz, θ is the rotation angle in rad, r is the radius of the wheel and z is its number of teeth, as highlighted in Table 1.

Table 1: Gear shafts kinematics

Shaft	***Rotational speed***	***Number of teeth***
Input	$\omega_i = 2\pi f_i$	z_i
Output	$\omega_o = 2\pi f_o$	z_o

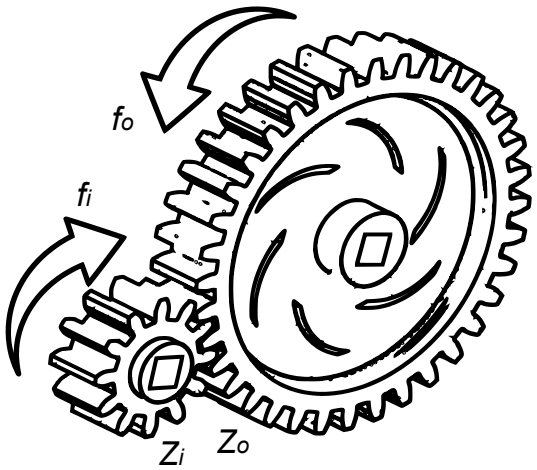


Figure 5: Mating spur gear wheels

The stability of the transmission ratio is ensured by the involute tooth profile, which implies a force-exchange occurring on a unique line of action identified by the tangent to the base circles of the two gear-wheels (the circle defining the base for the involute curve of teeth profile). In practice, the actual profile may show deviations from the ideal profile such as:

- Tip relief modifications, introduced by design
- Inaccuracy or systematic errors from the gear cutting machine
- Wear of the profile
- Tooth deflection, depending on the load

so that a transmission error (TE) occurs. This can be defined as an angular difference [4]:

$$TE_\theta = \theta_o - \theta_i \frac{r_i}{r_o} \qquad (3$$

that is, the difference between the perfect kinematic transmission and the actually achieved one. Dynamic TE involves the stiffness at the gear-mesh and is then commonly used as a condition monitoring parameter. Anyway, it must be remembered that the number of teeth in contact is usually non-constant during a meshing cycle. Indeed, the gear-characteristic contact ratio (CR) corresponding to the average number of teeth in contact is usually between 1 and 2, and the same transmitted force discharges either on one or two teeth, leading to a variable deflection which gives a strong parametric excitation at the so-called gear-mesh frequency (GMF). The GMF, together with the rotational speeds of the two gears ω_i and ω_o are then the main spectral lines characterizing the gear signature. Further relevant frequencies are the Gear Assembly Phase frequency (GAPF), which highlight the presence of wear, and the Hunting Tooth Frequency (HTF), established when a damage appears on a single tooth of both wheels.

Gear signature

The main spectral lines of the spectrum of a vibration signal induced by gear are:

- ***Input and output shaft frequency***
  Referring to a speed reducer, rotational frequencies expressed in Hz are given by:

$$f_i = \frac{n_i}{60} = \frac{\omega_i}{2\pi} \qquad f_o = \frac{n_o}{60} = \frac{\omega_o}{2\pi} = \tau f_i \qquad (4$$

- ***Gear mesh frequency***
  The GMF defines the rate at which gear-teeth mesh together. It is given by:

$$GMF = f_i z_i = f_o z_o \qquad (5$$

- ***Gear assembly phase frequency***

Depending on the number of teeth of a gear, different wear paths are possible, in accordance to the number of assembly phases. This number determines the distribution of wear between the teeth of the input and the output wheels. For example, considering a gearset of two wheels with 9 (z_i) and 15 (z_o) teeth, one tooth of the input wheel meshes with only 5 other teeth of the second wheel, leading to 3 possible paths: 1, 10, 4, 13, 7 or 2, 11, 5, 14, 8 or 3, 12, 6, 15, 9.

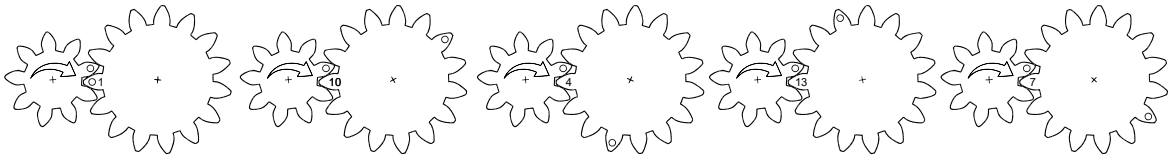


Figure 6: First wear path of the considered gear with $z_i = 9$ and $z_o = 15$

In general, the number of wear patterns (3 in the example in Figure 6) can be calculated as the *greatest common factor (gcf)* among the number of teeth z_i and z_o.

When wear occurs, the particular teeth combination offers a unique vibration characteristic and then the Gear Assembly Phase frequency appears in the vibration spectrum:

$$GAPF = \frac{GMF}{gcf} \qquad (6$$

- ***Hunting tooth frequency***

It is the frequency at which a damaged couple of teeth enters in contact.

Referring to previous example, considering the first wear path, tooth 1 of input wheel meshes only with teeth 1, 10, 4, 13 and 7 of the output wheel. If the first teeth of both wheels are considered, a contact occurs every 5 revolutions of the input shaft.

The calculation of the Hunting Tooth Frequency (HTF) depends on the number of teeth of each wheel and involves their gcf:

$$HTF = \frac{GMF \cdot gcf}{z_i z_o} = f_i \cdot \frac{gcf}{z_o} \qquad (7$$

Finally, in the overall spectrum, the gear signature can be easily highlighted. A qualitative spectrum as a function of orders of the input shaft is given in Figure 7. The orders correspond to a measure of relative frequency normalized on the input shaft frequency.

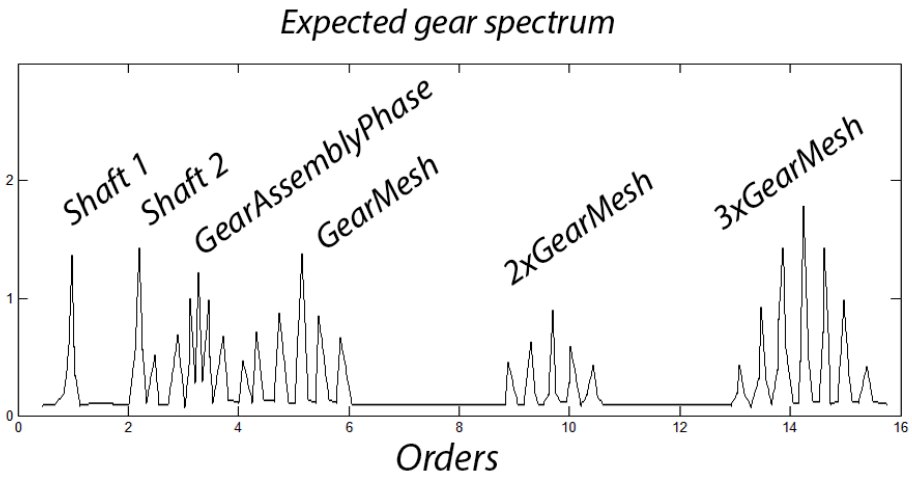


Figure 7: Expected spectrum: gear signature

As clearly depicted in Figure 7, important sidebands are always detectable also in the healthy condition, as even small mounting eccentricities or misalignments can lead to an amplitude modulation of the gear mesh signal at the rotational speed. The same but stronger modulation occurs in case of distributed faults (e.g. wear), enhancing the amplitude of the first sidebands which can be used as a damage indicator. Also, incipient wear causes an increase in the even GMF harmonics (in particular the second, $2 \times \mathrm{GMF}$). Then, as wear proceeds, the involute tooth-profile deteriorates leading to an enhancement of all the GMF harmonics and eventually to the appearance of the GAPF. Local faults such as root cracks or spalls on the contrary, bring wider distributions of harmonics and sidebands.

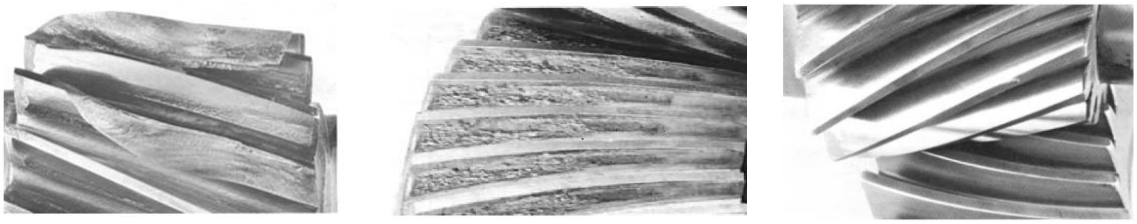
In the acquisition stage then, in order to ensure the gear signature to be well pictured in the spectrum ($f_{ny} \approx 3 \div 4\, GMF$), the sampling frequency must be set at least around $fs \cong 10\, GMF = 10\, z_i f_i$, so that a meshing cycle will be described by a minimum of 10 samples.

1.2.3. Gear fault models

Gear are critical components which may fail in various ways. The most common failure modes are:

- Breakage – when the stresses exceed fatigue limit, cracks originate and propagate until failure.
- Surface fatigue – Hertzian contact may induce surface fatigue phenomenon such as pitting, a progressive loss of material. When the involved stresses are high, larger and massive material pieces may detach, leading to spalling.
- Wear – degradation of the contact area due to a loss of material mainly affected by lubrication

Pictures of these different phenomena are reported in Figure 8.

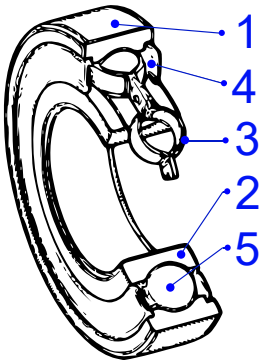


Crack *Fatigue (pitting or spalling)* *Wear*

Figure 8: Gear fault modes

1.2.4. Bearing signal

Rolling contact bearings are extensively used in all types of rotating machines, and their failure is one of the most frequent reasons of breakdown. They consist of a number of balls or rollers spaced by a metal cage and rolling within an inner and outer race.



Rolling element bearing components:

1. Outer Ring
2. Inner Ring
3. Cage
4. Packing and sealing
5. Rolling elements (balls)

Figure 9: Ball Bearing scheme

When a fault develops on a racetrack, a series of impacts is produced by the rolling elements passing on it. These shocks excite high frequency resonances of the whole structure between the bearing and the transducer.

The vibration signal induced by bearings presents peculiar characteristics:

- The strength of the impacts depends on the load supported by rolling elements and it is modulated by the rate at which the fault is passing through the load zone,
- Where the fault is moving with respect to the fixed position of the transducer, the transfer function of the transmission path varies.

Assuming to study an **ideal bearing** (ideal dimensions, no slip), it's easy to find the kinematic frequencies of occurrence of these impacts, and the typical modulation pattern in case of unidirectional (vertical) load, as shown in Figure 10.

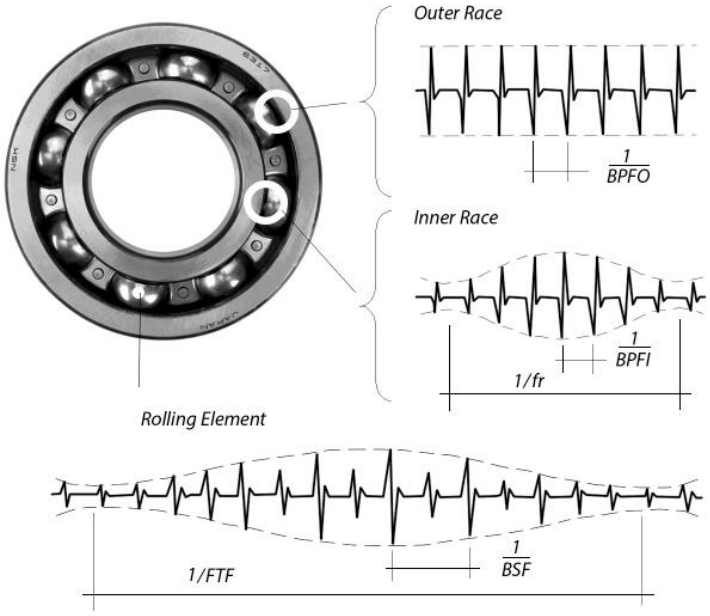


Figure 10: Typical signals induced by local faults in rolling element bearings

The ideal kinematic frequencies result:

Ball-pass frequency, outer race (BPFO)	$BPFO = \frac{nf_r}{2}\left\{1 - \frac{d}{D}cos\emptyset\right\}$	*(8*
Ball-pass frequency, inner race (BPFI)	$BPFI = \frac{nf_r}{2}\left\{1 + \frac{d}{D}cos\emptyset\right\}$	*(9*
Fundamental train frequency (cage speed, FTF)	$FTF = \frac{f_r}{2}\left\{1 - \frac{d}{D}cos\emptyset\right\}$	*(10*
Ball spin frequency (BSF)	$BSF = f_r\frac{D}{2d}\left\{1 - \left(\frac{d}{D}cos\emptyset\right)^2\right\}$	*(11*

Where:

- f_r is the shaft speed (relative speed among inner and outer ring),
- $\emptyset$ is the angle of the load from the radial plane
- n is the number of rolling elements,
- d is the inner ring diameter,
- D is the outer ring diameter.

Unfortunately, in case of actual bearings, slips are likely to occur: the size of the rolling elements in fact is given at a tolerance degree, so that different diameters imply distinct speeds, which will be "uniformed" by the cage causing random slips. The actual bearing frequencies can show variations of the order of 1-2% both as deviation from the calculated value and as random variation around the mean frequency, so that the resulting vibration signal is no more periodic but stochastic and is usually modeled as cyclostationary.

Cyclostationarity

The cyclostationarity is the property of a signal which, although not necessarily periodic, is produced by a hidden periodic mechanism. Most of the signals generated by rotating and reciprocating machines are then of this kind. Strictly speaking an *n-th order cyclostationary* signal is a signal whose n-th order statistics are periodic. For example, a fist order cyclostationary signal (**CS1**) has a **periodic mean value** (e.g. realizations of a purely periodic signal to which a noise is added – for details about random processes refer to Appendix 5), while a second order cyclostationary signal (**CS2**) shows a **periodic autocorrelation function** (e.g. realizations of a white noise modulated by a periodic amplitude, featuring then a periodic variance).

1.2.5. Bearings faults models

Bearing faults can be of various kind, but the most frequent are surely the pits and the spalls. A pit is a surface defect caused by impurities in oil at very high pressures which leads to small surface cracks. A spall on the contrary is a surface defect caused by the fatigue stress problems related to the motion of the rolling elements. A crack below the surface is then generated, leading to the detachment of a skin part.

a) Pit and Spall comparison b) Pitting c) Spalling (flanking)

Figure 11: a) Typical Pit and Spall scheme with dimensions, b) Tapered rolling bearing pitting on outer race, c) Well developed fatigue spall on a bearing inner race.

However, in order to study their effect on vibration signals, the fault models are generally diversified on the basis of the crack dimension, distinguishing between localized faults (typically smaller pits) or extended faults (typically large spalls).

Localized faults

In general, at an early stage, the most common bearing surface fault is a small crack (pit or spall) resulting in **sharp impacts**.

Two models are available for the vibration signals generated by localized faults:

- Pure 2nd order Cyclostationarity (CS2): In this case the period of the signal is modeled as $T_i = iT + \delta T_i$, where δT_i is the random variable;
- Pseudo-Cyclostationarity: The period of the signal is in this case modeled as a random variable $\Delta T_i = T_{i+1} - T_i$, so that the uncertainty of occurrence is increasing with the number of future periods (the system has no memory).

Although CS2 model can give good results, and in terms of envelope spectra the difference is minimum, it has been proved to be incorrect, so that Pseudo-Cyclostationarity should be taken into account [5].

Extended faults

As time passes, the small localized cracks are enlarged but smoothed by the balls rolling on them. The generated impacts are no longer sharp, and their detection becomes harder. Luckily, when bearings are coupled with gears, it is possible to exploit their vibration signal to evidence the presence of these faults. Indeed, the gear-induced vibration signal undergoes a modulation due to the extended fault crossing the loaded zone of the bearing, as it is highlighted in Figure 12.

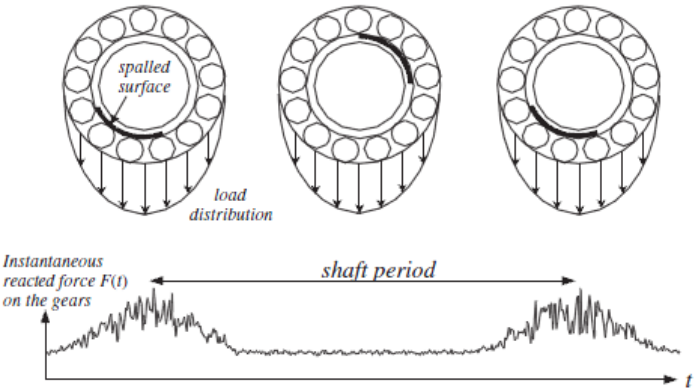


Figure 12: Typical modulating signal from the effect of an extended inner race fault on a gear signal [5]

1.2.6. Noise

Generally speaking, a noise is an unwanted sound judged to be unpleasant, loud or disruptive to hearing. In contrast, in physics noise usually refers to the unwanted part of an acquisition produced by a stochastic process which disrupts the signal of interest. The power spectrum (PSD) of a noise is usually called "colour". This practice started after the definition of the *white noise* as a signal whose PSD has equal power within any equal frequency bands (i.e. flat power spectrum). By analogy, the different colours are defined:

- *pink noise* PSD decreases of 3 dB per octave (density proportional to $1/f$),
- *red/brown noise (or Brownian)* PSD decreases of 6 dB per octave,
- *blue noise* PSD increases of 3 dB per octave,
- *violet noise* PSD increases of 6 dB per octave,
- *grey noise* PSD is modulated in accordance to the psychoacoustic equal loudness curve so that it is perceived by human hears as equally loud.

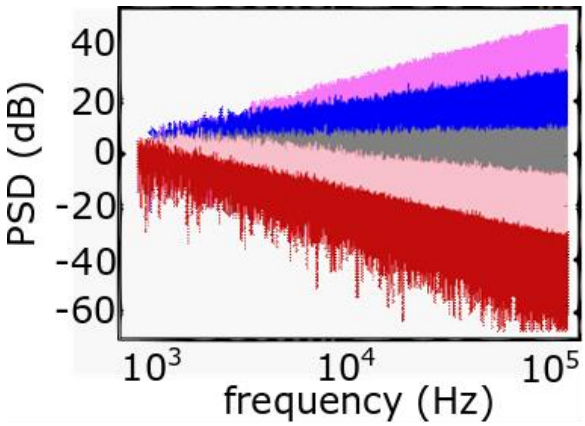


Figure 13: The colors of Noise

2. Commonly used features

Focusing on the D2D process, the decision is taken on the base of the knowledge of the state of health coming from an abductive reasoning. The core part is then the recognition of patterns in the dataset, in which the "symptoms" of a damage are sought. Such symptoms are usually abnormal variations of some selected features. According to the different components of a gearbox, different features can be taken into account [6].

2.1. Global Features

Historically, the most common features are related to the signal amplitude level and can be computed directly from the time measurements. In stationary conditions, in fact, higher acceleration values are ascribable to the presence of a malfunctioning. They can be referred to as *global* meaning that they give an overall indication of damage in the gearbox, without identifying the particular component which is failing.

Different time-features can be used as level indicators:

Peak:

The peak value of the signal is defined as half the difference between the maximum and minimum.

$$peak\big(s(n)\big) = \frac{1}{2}[\max(s(n)) - \min(s(n))] \qquad (12$$

Measurement noise can largely affect this level indicator, reducing its reliability.

Root Mean Square:

The Root Mean Square (RMS) of the signal is the normalized second statistical moment of the signal (standard deviation), namely the square root of the mean of the squared original signal:

$$RMS(s(n)) = \sqrt{E[s^2(n)]} \qquad (13$$

This indicator is more robust to noise, and quite reliable in stationary conditions.

Crest Factor:

The crest factor is defined as the ratio of the peak value to the RMS of the signal:

$$Crest\ Factor(s(n)) = \frac{Peak\big(s(n)\big)}{RMS\big(s(n)\big)} \qquad (14$$

In presence of significant impulsiveness, it is a very reliable indicator.

Based on the time measurements, different features can be extracted also by fitting a Linear Time Invariant model. In general, the simple autoregressive model is widely used to effectively characterize a signal:

AR coefficients:

A discrete time signal can be represented by an autoregressive model of the kind:

$$s(n) = -\sum_{l=0}^{L} a_l s(n-l) + e(n) \qquad (15$$

This formula models the signal $s(n)$ up to a residual error $e(n)$ using the previous L samples multiplied by corresponding a_l coefficients. L is the model order, which should be wisely selected. High order AR models, in fact, have a minimum in-sample error (error at the samples used to evaluate the model coefficients) but are very likely to overfit the signal, also modelling the acquisition noise.

In addition to the level indicators and LTI model coefficients, other statistical moments related to the shape of the sample distribution can be used as statistical features (see Appendix 3):

Mean value:

It represents the location of the distribution. For acceleration signals it is usually null.

$$Mean\big(s(n)\big) = \mu_s = E[s(n)] \qquad (16$$

Variance:

It corresponds the dispersion of the distribution. When the mean value is null, it is the square of the RMS.

$$Variance\ \big(s(n)\big) = \sigma_s^2 = E[(s(n) - \mu_s)^2] \qquad (17$$

For a stationery process with null average, the variance equals the RMS and represents the average power of the signal.

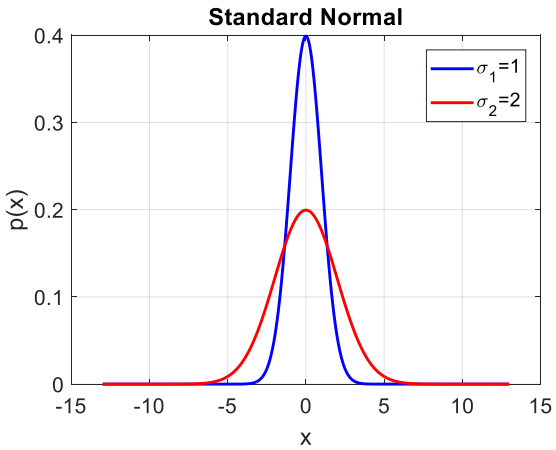


Figure 14: Dispersion of two different normal distributions with same mean value but different variance.

Skewness:

It indicates the degree of symmetry of the distribution; it is null for a symmetric distribution, while it increases (positively or negatively) when the distribution moves away from symmetry (the mean moves right-side or left-side with reference to the mode, respectively).

$$Skewness(s(n)) = E\left[\left(\frac{s(n)-\mu_y}{\sigma_y}\right)^3\right] \quad (18$$

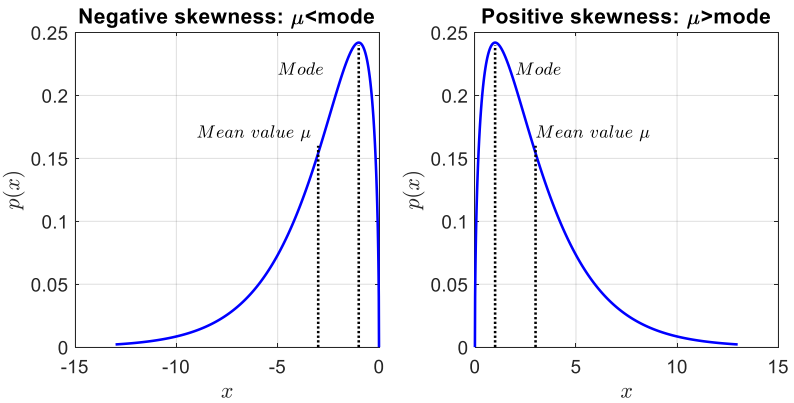


Figure 15: Skewness

Excess Kurtosis:

It is related to the *tailedness* of the distribution. When the distribution is normal, the excess kurtosis is made null (Mesokurtic) by removing a factor 3 from the standardized fourth order moment (usual definition of Kurtosis). When the tail area is larger ("fat tails"), with respect to the normal, the distribution is said Leptokurtic, and the excess kurtosis increases. On the contrary, a distribution is said Platykurtic when the tail area is reduced, and the excess kurtosis turns out to be negative.

$$Kurtosis(s(n)) = E\left[\left(\frac{s(n)-\mu_y}{\sigma_y}\right)^4\right] - 3 \quad (19$$

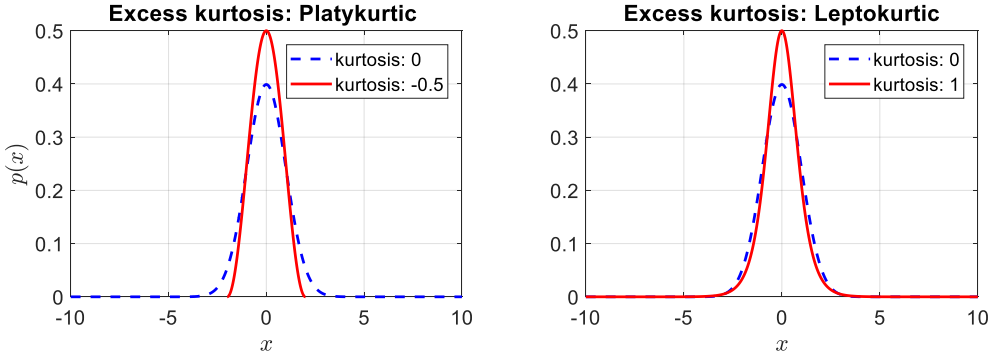


Figure 16: Platykurtic and Leptokurtic distributions

Frequency domain features are also widely used. They are based on the Power Spectral Density (PSD) of the signal, which can be easily computed through the Periodogram, a procedure based on Fourier transform (see Appendix 3). This gives a frequency domain representation characterized by a base frequency $df = {f_s}/{N}$ where N is the window length, and f_s the sampling frequency. In principle, all the spectral lines could be used as features, but clearly the frequencies related to damage phenomena or corresponding to some structural resonance will be much more of interest for diagnostic purposes. In this regard, gears and bearing features (the characteristic spectral lines) can be easily distinguished. However, as already introduced, it is preferable to separate the two contributions before exploring the spectrum to avoid possible cross participations.

Starting from the early 70's in fact, the Time Synchronous Averaging (TSA) is used on stationary, synchronous measurements to extract the deterministic part, completely determined as functions of time, leaving a residual signal which contains all the non-deterministic contributions, noise included. Besides TSA, different methods based on predictions from autoregressive models are able to isolate the deterministic component enabling the separation of the gear signal, periodic with respect to the shaft angle, from the bearing contribution, which is cyclostationary and will be contained together with noise in the residual signal.

2.2. Gear-targeted features

As stated in section 1.2.2, the main spectral features for gears are related to the spectrum amplitude at the gear mesh frequency (GMF) and its harmonics. The amplitude of the first order sidebands can also be used as a damage indicator. Indeed, in general, local faults such as root cracks or spalls entail wider distributions of harmonics and sidebands. Incipient wear on the contrary causes an increase in the even GMF harmonics (in particular the second, $2 \times \mathrm{GMF}$). Then, as wear proceeds, the involute tooth-profile deteriorates leading to an enhancement of all the GMF harmonics and eventually to the appearance of the Gear Assembly Phase frequency (GAPF). When damage on a single tooth of both wheels occurs, the Hunting Tooth Frequency (HTF) can also be found.

In addition to the plain spectral lines, further features can be extracted exploiting the discrete/random decomposition. The most common are FM0, NA4, NA4*, FM4, M6A, M8A, NB4 etc. [7] Below some of the most widespread are reported:

FM0

The zero-order figure of merit is a robust indicator of faults in a gear. It detects major changes in the meshing pattern by comparing the maximum peak-to-peak amplitude of the signal to the energy of the mesh frequency and its harmonics. Basically, it assumes that a damaged tooth in a gear produces a vibration signal with a significantly increased peak to peak value, while the overall meshing energy is not much affected. It is then similar to the crest factor but targeted on gears alone.

$$FM0 = \frac{PP(s(n))}{\sum_{h=0}^{H} P_h} \qquad (20$$

where $PP(s(n))$ is the maximum peak-to-peak amplitude of the signal $s(n)$ and P_h is the power spectrum of the h-th GMF.

FM4

Complementary to FM0 it aims to detect faults concerning a limited number of teeth. It measures whether the amplitude distribution of the difference signal is peaked or flat. It is then a computation of the normalized kurtosis of the difference signal $d(n)$, namely the SA residual signal (non-deterministic part) filtered to remove the gear mesh frequencies, the 1st order side bands, and the shaft revolution signal. Therefore, this parameter works under the assumption that a gearbox in good condition has a difference signal with a Gaussian amplitude distribution, whereas a gearbox with a defective tooth produces a difference signal with a series of major peaks, deforming the distribution. However, if too many teeth are defective, the data distribution turns flatter and the kurtosis value decreases, weakening the FM4 detection ability.

$$FM4 = \frac{N\sum_{i=1}^{N}\left(d(i) - \bar{d}\,\right)^4}{\left(\sum_{i=1}^{N}\left(d(i) - \bar{d}\right)^2\right)^2} \qquad (21$$

where $\bar{d}$ is the mean value of the difference signal, and N is the total number of observations. FM4 is non-dimensional and designed to have a nominal value of 3 if d is purely Gaussian.

NB4

NB4 is an indicator of localized gear tooth damage. When few teeth are deteriorating, transient load fluctuations arise, changing the envelope of the signal. NB4 exploit this information using the quasi-normalized kurtosis performed on the envelope of the SA residual signal, band-pass filtered about shaft and the mesh frequencies (differently from NA4 which uses the difference signal). The envelope signal is computed using the Hilbert transform (see Appendix 2).

2.3. Rolling element bearing-targeted features

Section 1.2.4 introduces the typical modulation pattern which can be found in ideal rolling element bearing signals in case of unidirectional (vertical) load. For actual bearings, because of possible slips occurring at the rolling elements, the real bearing characteristic frequencies can show variations of the order of 1-2% from the theoretical values given by the four ideal spectral features:

- Ball-pass frequency, outer race (BPFO)
- Ball-pass frequency, inner race (BPFI)
- Fundamental train frequency (cage speed, FTF)
- Ball spin frequency (BSF)

Unfortunately, these spectral features are often so weak with respect to the background noise, that they can rarely be detected before catastrophic failure by means of power spectra of raw signals from gearboxes. Also, the contribution separation performed by the SA, introduced in early 70's by Weichbrodt and Smith [9] and consolidated by Braun in 1975 [10], is not effective in highlighting the bearing signature. Because of this, the first bearing diagnostic studies (actually based on acoustic emissions) conducted by Balderston [8] in the late sixties, were based on the demodulation of high frequency resonant responses (tens of kHz). This developed in the late 70's into a number of techniques such as the ''Shock Pulse Meter'' (SPM) by Beercheck in 1976 [11] and later marketed for some time by the SKF bearing company, and the ''Spike Energy'' (SE) method marketed by IRD [18], where the demodulated frequency was the resonance of the transducer itself. In particular, the SPM used an accelerometer precisely tuned to a given frequency of 32 kHz, while the SE gave more tolerance for the transducer resonance, which was bounded in the range 5-50kHz. In both cases, the bandpass filtered signal was converted into a train of shock pulses with an amplitude proportional to the energy of the shocks produced by the damaged bearing. Particular figures of merit were finally used to detect incipient failure of bearings. An example is the IRD "**g-SE™**", or acceleration units of spike energy™, corresponding to a product of impact amplitude, pulse rate and high-frequency random vibration energy [14].

Starting from the mid 80's, the focus moved on the spectrum of the envelope of a band-pass filtered acceleration signal, leading to the "High Frequency Resonance Technique" (McFadden and Smith [13]), later evolved into the well-established "Envelope Analysis". Long discussions were held to choose the optimum bandwidth for the demodulation associated with envelope analysis. This dispute ended with the development of the spectral kurtosis (SK), first introduced by Dwyer in 1983 [12,15].

Spectral Kurtosis based algorithm such as the Fast Kurtogram [16], in fact, proved to be effective in many practical cases in finding the optimal frequency band for envelope analysis to improve the signal/noise ratio of bearing signal with respect to background noise, becoming the reference algorithm for bearing diagnostics.

Alternative ways to highlight the bearing spectral features introduced at the beginning of this chapter can be anyway found in the literature. Minimum Entropy Deconvolution (MED), for example, initially proposed by Wiggins in 1978 [18] for seismic analysis, can be used to generate a filter that counteracts the effect of the transmission path under the assumption that the original excitation was impulsive, and thus having high kurtosis. With a similar scope, the Stochastic Resonance can be used to enhance the impulsive content of a vibration signal [19]. Wavelet Denoising, on the contrary, can enhance the bearing characteristic frequencies by reducing the background noise.

A deeper insight of the here selected algorithms will be given in the next Chapters, also investigating their performances on the datasets presented in Chapter 4.

3. Signal processing: State of the art

As extensively described in Chapter 2, the data recorded by the acquisition system need to be processed (and pre-processed) so as to highlight the diagnostic information, which is otherwise just a messy superimposition of the different contributions (i.e. many sources) covered by noise and distorted by the transmission path from source to sensor. This operation, usually called *features extraction*, is then fundamental to enhance the health information hidden inside the signals.

Many different algorithms are available in the literature, so that a review to select the more established and promising is fundamental. In first instance, such algorithms can be classified according to the type of data acquired [20,21]. At least two categories can be highlighted in case of vibration monitoring:

- *Waveform type*: Time series of vibration acquisitions (typically accelerations). They are basically discrete functions of time.
- *Value type*: Single value recorded at a specific time for a particular variable (e.g. temperature, pressure, humidity, etc.). Even the low-level features which can be extracted from time series (e.g. RMS, etc) can fit this category as they summarize with a single value an entire period of time.

A second categorization may refer to the main context of application, in reference to the different components of a gearbox. In that case, the types will be mainly:

- *Global*
- *Gear-targeted*
- *Rolling element bearing-targeted*

3.1. Waveform data analysis

In the vibration monitoring context, waveform data refers to the time series recorded from the employed acquisition system. In case of accelerometers, the waveform is a discrete function of time describing the signal's acceleration amplitude over the duration of the recording. This time-signal can be decomposed via Fourier analysis in a set of fundamental waveforms, the harmonic functions, which form a basis for the so-called frequency domain.

The algorithms will be then further categorized according to the domain of reference.

3.1.1. Time domain analysis

Time domain analysis is based on the time waveform itself. Traditionally, such analysis is related to the extraction of so-called time-domain features which can describe the general condition of a machine (*global*). For example, it is a common knowledge that the overall vibration level can rapidly rise as a result of the aggravation of a damage. Indeed, descriptive statistics like peak amplitude, Root Mean Square (RMS) or crest factor (peak/RMS), in addition to higher order moments as variance, skewness and kurtosis, can be used to summarize the waveform with a value type data (appendix 5). These features are very fast to compute but are usually quite sensitive to the operational and environmental conditions. This give raise to necessary considerations about their clear dependence on damage.

Advanced approaches apply parametric time series models to the waveform. The parameters resulting from the fitting will be then used as features to characterize the system. The popular models used in the literature are the autoregressive (AR) models or

the autoregressive moving average (ARMA) models. In practice, for complex systems, the application of AR or ARMA models is not straightforward and difficulties rise in the determination of the more appropriate model order.

Model fitting can be directly applied with diagnostic purposes. It can be noticed that a damage can deteriorate the goodness of the time-signal prediction, so that it is possible to find a dependence of the variance of the residual signal (or error) from the damage.

Taking a further step, model fitting can also be used with signal separation purposes. According to Wold's theorem [22] in fact, a stationary time series can always be written as a unique sum of a deterministic part (singular) and a stochastic part (regular), but only the first can be predicted by its past values using an AR model, so that a separation is possible. The optimal AR coefficients for this task take the name of Wiener filter. A number of algorithms are available based on the AR model, like the Linear Prediction (LP), also known as Adaptive Line Enhancement (ALE) in acoustics, the Adaptive Noise Cancellation (ANC) and the Self-Adaptive Noise Cancellation (SANC) [22] but the most popular techniques remains the Synchronous Averaging (SA), proposed since the early 70's in order to separate the deterministic content from the residual part. With this procedure in fact, the complex time-domain vibration signal from a whole gearbox can be reduced to deterministic estimates of the vibration for individual shafts and their associated gears. When the signal is stationary, a simple Time Synchronous Averaging (TSA) [23] can be applied, but in most of applications this does not hold. *Angular-domain* data is then required. Unfortunately, commercial Data Acquisition Systems (DAQ) are not so efficient in sampling at constant angular increments, so that resampling algorithms are needed, based on the additional record of the angular position of a reference shaft. This procedure takes the name of Computed Order Tracking (COT). Once the separation is performed, the analysis usually switches to the frequency domain, to extract features of interest from the spectra.

Other techniques are based on the amplitude-phase demodulation of the time signal. This is usually performed using the analytic signal produced via Hilbert transform, from which the envelope and the instantaneous angular frequency (or phase) can be extracted [Appendix 2]. Envelope Analysis (EA) is shown to enhance both gear fatigue crack detection and, in particular, rolling element bearings damages [24]. The selection of the demodulation band for EA is usually performed via spectral kurtosis-based algorithms such as the Fast Kurtogram (FK).

Finally, Minimum Entropy Deconvolution (MED), initially proposed by Wiggins in 1978 for seismic analysis, can be used to generate a filter that counteracts the effect of the transmission path from the source to the sensor, with particular applications for bearings, whose impulses can be separated and highlighted [23].

3.1.2. Frequency domain analysis

In the field of vibration monitoring, it is common to expand the time series in a sum of harmonic functions using the popular and efficient Fast Fourier Transform FFT. The advantage of analyzing the signal spectrum is that certain frequencies of interest can be easily identified and isolated. These form the "signature" of the machine, which is unique and highly responsive to fault. The main idea is then to look at the whole spectrum and to focus on certain spectral lines which can be used as features (see Chapter 3). The most commonly used tool for spectral analysis is the power spectrum, which can be estimated in two different ways:

- Non-parametric algorithms are usually preferred as they do not need any assumption about the process structure. They estimate the power spectrum through the Fourier transform of the estimated autocorrelation of the time sequence [Appendix 3].

- Parametric approaches on the contrary assume that the underlying stationary stochastic process has a simpler structure which can be represented by a parametric model (e.g. AR, ARMA etc.) from which the power spectrum can be obtained.

Despite the lower spread, other kind of spectra can be defined. The Cepstrum for example, which is the inverse Fourier transform of the logarithmic power spectrum, is commonly used to highlight the harmonics and the sideband patterns of the gears signature [24]. Higher order spectra like bi-spectrum, tri-spectrum are sometimes applied [21].

Other frequency domain methods based on parametric modelling can substitute equivalent time-domain, predictive algorithms. It is the case of the Discrete/Random Separation (DRS) [25,26], which outperforms ANC, SANC and LP in terms of efficiency and stability. Once the deterministic part is separated from the residual, the spectral lines of interest, magnified by the separation, can be extracted as features.

3.1.3. Time-Frequency analysis

The spectral analysis is a valuable tool but is unable to reflect the time varying nature of the signal, whose frequency content may change in time. It is then particularly inadequate for processing non-stationary waveform signals, which are common in machineries. In this case, a two-dimensional spectrum is sought, as a function of both frequency and time. The simplest method is the Short Time Fourier Transform (STFT) which produces the so-called *spectrogram* (the power of the STFT). It is based on the segmentation of the original time series which is then cut, windowed and Fourier-transformed. This obviously limits the frequency resolution. Because of this, bilinear time-frequency distributions (e.g. Wigner-Ville, Choi–Williams, etc) are sometimes used, even if they suffer the interferences due to cross terms.

Another established 2D representation is the *scalogram* produced via wavelet transform (WT). It differs from the STFT because of the use of different base functions. An orthonormal basis is generated by shifting and scaling a wavelet i.e. a function centered around zero and featuring finite energy. The advantage of the scalogram is a non-uniform resolution: it has a high frequency resolution at low frequencies and a high time resolution at high frequency. More recent developments of WT are the Discrete WT (DWT) and the Wavelet Packet Transform (WPT).

Finally, also algorithms for signal decomposition are often used to produce data-driven time-frequency representations. The Empirical Mode Decomposition (EMD) for example, is able to decompose the signal in a number of base functions which are the possibly nonlinear and nonstationary simple oscillatory modes. Ensemble EMD (EEMD), Complete EEMD with Adaptive Noise (CEEMDAN) are two evolution of EMD trying to improve the mode-mixing issue. When statistics and geometry elements are added to the time series analysis, further decompositions can be found. One simple method relies on the application of the well-known Principal Component Analysis and takes the name of Singular Spectrum Analysis (SSA), recently improved in the Singular Spectrum Decomposition (SSD). The name “singular spectrum” relates to the spectrum of eigenvalues, often used as a base for the reconstruction of the components which are believed to have a meaningful interpretation.

3.2. Value type data analysis

Value type data include both raw data from acquisition or extracted. Such kind of data is much simpler than waveforms, but the complexity lies in the correlation structure among the different variables, whose number may be large. Because of this, a multivariate analysis needs to be conducted.

One of the simplest but fundamental multivariate analysis is the Principal Component Analysis (PCA), which uses an eigenvalue decomposition of the covariance matrix estimated from the data to define a new orthogonal coordinate system. The final principal components are the uncorrelated linear combinations of the original variables. PCA is usually applied to reduce the dimensionality of large datasets, producing 2D or 3D projections which explain most of the data variability and can be easily visualized.

Another linear but supervised dimensionality reduction can be obtained via Linear Discriminant Analysis (LDA), which aims to find the direction that best discriminate some data classes. It is in fact a linear classifier algorithm such as Logistic Regression or Naive Bayes classifier.

When linearity proves to be a limit, Non-linear PCA, kernel PCA (k-PCA) or Local Linear Embedding (LLE) are commonly used, together with applications of Artificial Neural Networks (ANN) featuring bottleneck layers. Also, non-linear classifiers such as Support Vector Machine (SVM) or kernel classifiers (e.g. k-nearest neighbours k-NN) are widely used.

In some cases, it may be useful to find a transform whose basis is composed by independent components, although non-orthogonal. It is the case of the Independent Component Analysis (ICA), which can be used to perform a blind source separation of the data.

Finally, for explorative unsupervised analyses (i.e. the value type data is not labelled) clustering can substitute classification for discovering similarities and dividing the data-points into groups (cluster) [27].

3.3. Large Rotorcrafts Vibration Health Monitoring: acknowledged indicators and signal processing techniques

At the end of section 4.5 of Chapter 2, the Health and Usage Monitoring Systems (HUMS) of helicopters were considered as a commercial example of implementation of usage monitoring. Regulatory bodies such as the European Aviation Safety Agency already acknowledges such systems and also gives indications regarding the simplest but thrusted procedures for vibration health monitoring. Reference [28], at chapter CS 29.1465(a), reports the typical vibration monitoring indicators and signal processing techniques summarized in Table 2.

Table 2: Typical Vibration Health Monitoring Indicators & Signal Processing Techniques [28]

Main Gearbox components:	**Indicators:**
Shafts	Fundamental shaft order and its harmonics.
Gears	Gear-mesh frequency and its harmonics, modulation of meshing waveform, impulse detection and energy measurement, non-mesh-related energy content.
Bearings	High frequency energy content (band-passed signal energy measurements), impulse detection, signal envelope modulation patterns and energies correlated with bearing defect frequencies.
Rotor:	Fundamental shaft order and harmonics up to blade pass frequency, plus multiples of this.

The diagnostic relevance of the vibration data processing via extraction of the previously introduced indicators, is particularly highlighted in the document. The typical signal processing techniques reported include:

I. Asynchronous Power Spectrum, which does not require phase information or frequency tracking.
II. Synchronous Spectrum, which needs phase information or frequency tracking.
III. Band-pass filtered signal Envelope Power Spectrum Analysis (a recommended technique for gearbox bearings).
IV. Synchronous Averaging for time and frequency domain signal analysis (a recommended technique for gearbox gears).
V. Band-pass filtering and the measurement of filtered signal statistics, including crest factor (can be used for bearings not within engines or gearboxes).

3.4. Vocabulary and definitions of Signal Processing

As a final part of the state of the art, a short vocabulary is added to ensure a unique interpretation and definition of some relevant operations which can be performed on a measured time signal $x(t)$ [29].

***Detection**:* Deducing (or better "abducing") the presence of an unknown but peculiar signal $s(t)$ in a noisy, measured waveform $x(t)$ at a pre-set confidence level. The damage detection is then the binary decision-making problem pertaining to the presence or not of a fault in the considered system.

Identification: Deducing the presence of an unknown but peculiar $s(t)$ in a noisy, measured $x(t)$. In particular, damage identification is the multiple decision-making problem pertaining to the localization or classification of an incurred fault thorough the damage-characteristic signal $s(t)$.

***Estimation**:* Estimating the "amount" a of a known signal $s(t)$ in a noisy, measured $x(t)$, where $x(t) = a \cdot s(t) + n(t)$. It is then related to the estimation of the exact magnitude of a detected damage (i.e. damage level).

Filtering: Using a signal processing model and the past measurements to enhance the current measurement. It can be performed as an on-line analysis.

Smoothing: Using a signal processing model and all measurements (past and future) to enhance each measurement. It is not suitable for on-line analysis, but only for batch (or at line) analysis.

Predicting: Using a signal processing model and the past measurements to predict future measurements. It can be done both on-line and in batch mode.

4. Machine Learning State of the art

Machine Learning is the subject of building intelligent machines able to learn information from the data without the intervention of human specialists. This learning consists of highlighting an underlying set of patterns useful to understand relationships in the data which can be used to generalize some sort of inference [30,31,62].

A taxonomy of ML models is reported in Figure 17, where two main types of learning are shown. In particular, the learning is said supervised when labelled input data is available. Furthermore, when a numeric variable has to be predicted, we talk about regression, while classification is used when a categorical value (e.g. a class or a group) is pursued.

On the contrary, the unsupervised learning works without exploiting information about labels. In this case, two families of problems are commonly solved with unsupervised learning methods: dimensionality reduction and clustering.

In between, when labelled data from the healthy condition alone is used, we speak about semi-supervised learning. This can be tackled as a one-class classification problem which takes the name of novelty detection.

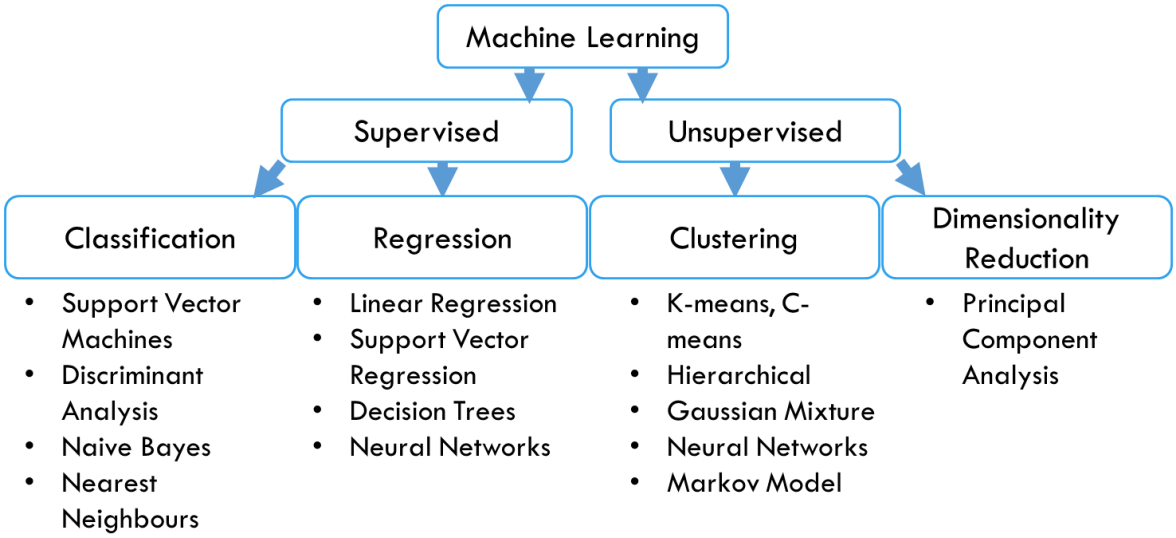


Figure 17: Taxonomy of the most common Machine Learning algorithms.

4.1. Regression

Regression is the statistical processes of estimating relationships among variables so as to predict numerical values of the dependent variable from one or more independent variables in a data set.

Regression can be either parametric or not. Non-parametric methods are commonly performing smoothing using a kernel function. Kernel regression estimates the continuous dependent variable from a limited set of data points by convolving the observations with a kernel function, a window that specifies how to "blur" the influence of the data points so that their values can be used to predict the value for nearby locations. Also, decision tree learning algorithms such as the Classification And Regression Tree (CART) [32] can be applied for predicting a dependent variable.

Regarding parametric regression on the contrary, the simplest model is the linear. In linear regression, the relationships are modelled with linear predictor functions whose

unknown parameters are usually estimated from the data by Ordinary Least Squares, which minimizes the sum of the squares if the differences between the observed dependent variable and the its predicted value from the linear function.

Linear regression is good for regressing data that follows a linear pattern and adhere to certain assumptions such as normally distributed error for the dependent variable.

If also the independent variables are affected by errors then, other error-in-variable models should be preferred, such as Orthogonal Regression (or Principal Component Regression).

On the contrary, when the limit is the linearity of the pattern, a simple way to improve the linear model is with a Generalized Linear Model (GLM) able to generalize linear regression to allow for fitting nonlinear models using a link function that transforms the data appropriately. For example, when the dependent variable is restricted to binary, a sigmoid function can be used to take any real input to a value bounded between zero and one that can be interpreted as a probability. In this case we speak about Logistic Regression.

The advantage of linear regression lies in its simplicity and easy interpretability. Nevertheless, Machine Learning can offer way more sophisticated extensions of GLMs based on Neural Networks which are more flexible as they can automatically select any nonlinearity and apply it locally to model very complex functions. In this sense NNs are very flexible and effective, so that the only drawback remains the model interpretability. Understanding the "logic" behind, in fact, can be critical in some applications as in diagnostics, where a human is often required to confirm or not the diagnosis.

In any case, nowadays trading-off ease of interpretability for predictive power is commonly accepted at scientific level, so that the use of models with high level of abstraction led to the development of an entire branch of ML called Deep Learning.

Deep Learning methods evolved from the Neural Network, which is inspired by the structure and function of biological neurons. The parameters to be optimized are now weights on the Neural Network connections.

The success of NNs in all the fields of science is strongly related to their supervised training called backpropagation, which uses gradient descent with respect to the weights to minimize output error.

The behaviour of a NN is dependent on its architecture: number of neurons, number of layers and their connections. Choosing the right architecture depending on the problem to model is obviously fundamental [33]. Ongoing research are producing every day novel architectures, so that is would be impossible to consider all of them. In any case, the most important ones are here reported.

The first basic feed-forward architecture is the Multi-Layer Perceptron (MLP [34]), made of at least three layers of nodes, an output layer, at least a hidden layer and an input layer. Except the input nodes, each node is a neuron that uses a nonlinear activation function.

Another well-known architecture is that of Recurrent Neural Networks (RNNs [35]) which is used to model time series as a recurrent function of sequential data where an output at time t depends on the input at time t and on the previous output at time $t-1$.

The RNN is then a special feed-forward Neural Network that can send information over time-steps by creating loops in the network that feed forward to the same neuron at the next time step. This works similarly to Markov models but is superior for very large data sets because the RNN can capture long-range time dependencies.

The architecture of a Long/Short-Term Memory (LSTM [36]) network was designed explicitly to overcome the problems of long-range time dependencies by using three gates:

- a forget gate to propagate information preserved from the previous time step,
- an input gate to introduce new information at the time step, and
- an update gate that determines what will be preserved from the time step.

In this way, the gates can carry information efficiently across any number of time steps. The gates weights are finally optimized using another application of backpropagation through time steps.

Another well-known architecture is that of Convolutional Neural Networks (CNN), representing the mathematical convolution operation on the inputs [63,64]. A CNN consists of a number of hidden layers where convolution filters of various shape and size are implemented as sliding windows. CNNs are organized so as to learn filters to produce strong activation to spatially local input patterns as well as globally [33].

Weights in a CNN are optimized again with backpropagation, however, to limit the possible overfitting, the optimization is performed using regularized methods such as the Dropout, which randomly drop units (along with their connections) from the Neural Network during training to prevent co-adaptation and overfitting.

In general, regularized regression methods add information penalizing large regression coefficients and avoid then overfitting. Ridge and Lasso regression are two examples of regularized least squares (RLS).

Regression methods predicting a dependent variable (i.e. the label) from one or more independent variables (i.e. the features) are sometimes used for classification. For example, setting a threshold on the probability outcome of a Logistic regression, it easy to turn the regression output (practically a soft classification information) into a hard classification information (e.g. binary, 0 or 1).

On the contrary, classification algorithms such as Support Vector Machines (SVM) can be modified to perform regression (Support Vector Regression, SVR [37]). Instead of generating safety boundaries from a hyperplane which separate two classes, the optimization seeks to find a hyperplane as close as possible to the observations. The SVR is a non-parametric regression tool which can exploit a non-linear kernel to finally end up with a non-linear regression.

4.2. Classification

Classification corresponds to the problem of identifying the right category (class or group) to which a new observation belongs, on the basis of a training set of data containing labelled observations whose category membership is known. Labels can be either numerical or categorical. Classification is then the prediction of a label value. At its simplest, classification produces a binary output with labels of two classes: 0 and 1 (hard binary classification). Some models such as decision trees, directly create a binary output. Other models calculate a probability (value between 0 and 1) and separate the values according to a threshold value with a discriminant function (soft classification).

The first works on classification dates back to Fisher [38], who introduced Fisher's linear discriminant function assuming a multivariate normal distribution for each of the two groups. Improvements able to give group membership probabilities are based on Bayesian approaches which are considered more informative than simply attributing a single group-label to each new observation.

In general, a large number of algorithms for classification are based on a linear function that assigns a score to each possible category by combining the feature vector of a new observation with a vector of weights, using a dot product. The predicted category is the one with the highest score. This type of score function is known as a linear predictor function, and is comprehensive of Logistic regression, Support Vector Machines (SVM) and also the single Perceptron (building block of the Multi-Layer Perceptron) falls under this category.

Despite being a linear classifier, SVM can exploit the kernel trick to implicitly map the inputs into high-dimensional observation spaces in which the categories are separable by a clear gap. SVMs then can efficiently perform a non-linear classification.

A non-linear classification can be obtained also by adding multiple layers of Perceptrons with non-linear activation functions, which leads to a feed-forward artificial neural network called Multi-Layer Perceptron (MLP).

Quadratic classifiers are also possible, such as the Quadratic Discriminant Analysis (QDA), which allows for conic sections separating the classes (e.g. a line, a circle or ellipse, a parabola or a hyperbola).

Moving to non-parametric methods, the simplest method is probably the k-nearest neighbours (k-NN). In this case, an observation is simply classified by a plurality vote of its neighbours. The observation is assigned to the class most common among its k nearest neighbours.

Other non-parametric classifiers are based on decision trees. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf form the classification rules. Decision threes are commonly based on algorithms able to optimize the choice at each node, but this does not ensure to get to the global optimum configuration. Furthermore, particularly deep trees are prone to overfitting the dataset, so that trees are commonly implemented in Random Forests, an efficient type of Ensemble Learning. Random Forest models in fact, implement a level of differentiation by splitting the different features and using just some of them in different threes, whose results are later aggregated.

A different Ensemble Method is the so called BAGGing, or Bootstrap AGGregation. In this case, given a dataset, multiple bootstrapped subsamples are drawn. A Decision tree is then formed on each of the subsamples and the results are later aggregated. Decision trees are anyway proved to lead to better results.

In general, anyway, an objective comparison of the different classifiers' performance is difficult to set up. A classifier performance in fact is strongly related to the dataset under analysis, so that, as stated by the "no-free-lunch theorem", a single classifier that works best on all given problems does not exists.

Classification is considered an instance of supervised learning. The corresponding unsupervised procedure is known as clustering and involves grouping data into categories based on some measure of distance or similarity.

4.3. Clustering

Clustering is the general task of exploratory data mining whose target is the grouping a set of observations in such a way that the observations in the same group (which takes the name of cluster) are more similar to each other than to those in other groups.

Similarity can be evaluated in terms of distances between cluster members, density areas of the data-space or intervals.

Clustering is then a multi-objective optimization that involves trial and error. The iterative process seeks to find the best variable settings including the number of expected clusters, the thresholds or the intervals, which are the main parameters as well as the distance metric to use, which is usually selected previously.

The most famous clustering algorithm is the k-means, which belongs to the family of centroid-based algorithms. In brief, clusters are based on the distance of the observations from a number k of pre-determined centroids whose positions are iteratively optimized.

Different approaches are categorized as Hierarchical clustering. They agglomerate or divide observations in clusters on the basis of a measure of similarity, commonly distance. For example, agglomerative algorithms then “connect” near observations to form clusters. The most established algorithm of this kind, also known as Connectivity-based clustering, are single-linkage clustering (SLINK [39]), complete linkage clustering (CLINK [40]), and average linkage clustering such as UPGMA or WPGMA [41] (Unweighted or Weighted Pair Group Method with Arithmetic Mean).

Distribution-based clustering can also be found. Clusters in fact can be interpreted as observations belonging to the same distribution. Fitting for example a Gaussian mixture model (GMM) optimized by the expectation-maximization algorithm both hard clustering (binary information about observations belonging or not to a cluster) and soft clustering (also known as fuzzy clustering, it outputs the likelihood of belonging to a cluster) are possible.

Finally, also density-based clustering, can be found in the literature. In this case, clusters are defined as areas of higher density with respect to the remainder of the data set. The most popular method is DBSCAN [42] which connects points within certain distance thresholds only when they satisfy a density criterion. A generalization of this method takes the name of OPTICS [43], which removes the selection of the threshold and produces a hierarchical result related to that of linkage clustering.

4.4. Dimensionality reduction

Dimensionality Reduction is the process of reducing the number of variables under consideration. Approaches can be divided into feature selection and feature extraction methods. Feature extraction creates new features as functions of the original features, whereas feature selection returns a subset of the features. The main idea when using a feature selection technique is that the data contains some features that are either redundant or irrelevant and can thus be removed without incurring much loss of information. Subset selection algorithms can be broken up into wrappers, filters, and embedded methods. In general, anyway, they are based on a search algorithm which optimizes the scoring metric grading the selected subset of features and comparing it to all the other scorses from each possible subset.

The purpose of feature extraction dimensionality reduction on the contrary is to increase data efficiency by saving on computational expense while strengthening the signal-

to-noise ratio by eliminating dimensions high in noise or low in variance in data. Furthermore, the first principal features can be used to effectively visualize the dataset.

Some of the established linear dimensionality reduction methods are Principal component analysis (PCA), Singular value decomposition (SVD), Linear discriminant analysis (LDA), Factor analysis (FA) and Independent component analysis (ICA). Nevertheless, nonlinear mappings can be found by introducing a Support Vector Machine formulation to the problem of PCA, leading to the so-called Kernel PCA.

Other prominent nonlinear techniques which can falls under the name of Graph-based kernel PCA include manifold learning techniques such as Isomap, Locally Linear Embedding (LLE [44]), Hessian LLE, Laplacian eigenmaps, and methods based on tangent space analysis [45]. These techniques construct a low-dimensional data representation using a cost function that retains local properties of the data and can be viewed as defining a graph-based kernel for Kernel PCA.

The SVM kernel trick can be used also for extending LDA to a Generalized discriminant analysis (GDA) which provides a mapping of the input vectors into high-dimensional feature space to finally reduce the dimensionality by maximizing the ratio of between-class scatter to within-class scatter.

Finally, Neural Networks can also be used to learn non-linear dimension reduction functions. In particular, Unsupervised Pretrained Networks (UPNs) are very good at reproducing input data or generating outputs that share a likeness with inputs. Three main architectures can be found in the literature, Autoencoders, Deep Belief Networks (DBNs) and Generative Adversarial Networks (GANs).

Autoencoders are used as powerful dimensionality-reducers. In their simplest form they are made of a single layer of perceptrons (as in MLP) where the output layer has the same number of nodes as the input layer, but the hidden layer is restricted in a way that forces the autoencoder to reconstruct the input only approximately, prioritizing the most relevant aspects of the input data. This allows an autoencoder to construct a compressed representation of the input [56]

DBNs consist of layers of Restricted Boltzmann Machines (RBMs), a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs [33]. An RBM approximates the input using a series of stepped sigmoid units, or binary layers of equal weight but progressively stronger negative bias [46]. Layering RBMs allows a DBN to progressively learn more complex features automatically and, with adequately deep layering, generate convincing reproductions of inputs.

GANs are systems of two neural networks competing in a zero-sum (i.e. balanced gain or loss) game and consists of a generative model (typically a deconvolutional neural network) that estimates data distribution of a training set and a discriminative model (a convolutional neural network) that estimates the probability that a sample came from the training set rather than the generative model [47]. The generative network's training objective is to increase the error rate of the discriminative network (i.e. to "fool" the discriminator by producing novel candidates so similar to the input that the discriminator is not able to distinguish them). They are mainly used for image-to-image translation but can also be used for denoising or dimensionality reduction.

4.5. Novelty Detection

Novelty detection is the task of classifying test data that differ in some respect from the data that are available during training. This may be seen as “one-class classification”, in which a model is constructed to describe “normal” training data.

The simplest approach to ND derives directly from statistics. Based on the probability distribution of the training data in fact, appropriate thresholds on probability or distance can be found.

If a statistical model is used to fit the training data, we speak about statistical, parametric approaches. The most common distribution model is the multivariate normal, so that when the assumption of normality holds, it is easy to find the iso-probability ellipsoid (multivariate extension of an interval of critical values) that contains a given amount of data (e.g. the 97% of the dataset). This can be done by fixing a threshold on the Mahalanobis distance, a metric which accounts for the covariance structure of the normal dataset. The selection of a proper threshold is not an easy task also in this simple case because of the effect of high dimensional spaces and small training sets. In this cases Extreme Value Theory (EVT) can be used to optimize the threshold.

When the assumption of normality is too restrictive, it is possible to increase the complexity of the model by considering a k-components Gaussian Mixture Model (GMM). A GMM is a is a linear combination of k-Gaussian probability density functions and models more generic distributions estimating the density by using fewer gaussian kernels than number of patterns in the data.

In general, the problem of ND can be tackled by statistical tests such as the simple two-Sample t-Test for Equal Means. If t-test shows significant difference between the two sets of measurements (normal profile and test profile), then second set is considered to contain novel patterns. This can be extended for testing multiple-samples tests using for example the Analysis of Variance (ANOVA), and for multivariate multiple-samples tests with the Multivariate ANOVA (MANOVA).

Hidden Markov Models (HMM) are also suitable for ND. HMMs are used to model sequential data as a number of states together with probability of moving between pairs of states (transition probabilities). Novelty is assessed if the probability of observing a sequence to be tested is below a threshold.

Parametric approaches often need extensive a prior knowledge of the problem to be effective. Hence, when no assumption about the statistical distribution of the data can be verified, non-parametric approach can be more suitable.

Among them, the simplest are the Histograms, a graphical display of tabulated frequencies. If the histogram of the data to be tested is dissimilar from the training one, the test data is considered novel. The Kullback–Leibler divergence can be used as a similarity metric in this case.

Histograms are very sensitive to bin size, which is the free parameter. A more stable alternative is to sum rectangular windows of the same size as the bin centred on each sample. This leads to a family of methods called Kernel Density Estimators (KDE), or also Parzen density estimation, whose free parameters are the window shape (which must not be necessarily rectangular) and the bandwidth (the size of the window).

Histograms can be extended also to multivariate datasets, but for dimensions higher than two they usually become unfeasible. KDEs also have problems due to the curse of dimensionality as in high dimensional spaces all the samples become approximately equally far away from each other, so that KDEs lead to almost uniform results.

The idea of density is always related to that of distance; hence ND methods are available, based on the assumption that normal points have close neighbours while outliers lies far away. In k-Nearest Neighbours (kNN) method, for example, an observation is classified by a majority vote of its neighbours. Finally, to evaluate novelty, it is possible to use either distance (e.g. observation is novel if distance to the k-nearest neighbours exceeds a threshold) or density (e.g. using a local outlier factor LOF). [49,50,51]

Ultimately, as Novelty Detection is practically a one-class classification, most of the classification algorithms seen in the previous section can work or can be adapted to work in this framework [52].

For example, traditional Support Vector Machines can be used assuming existence of labelled data [53,54]. When this is not possible, SVM can be modified considering that normal points belong a to high density data regions which can be encompassed by an hypersphere. Separation of non-spherical distributed data is done in a high dimensional space into which observations are mapped using non-linear mapping (kernel).

Most of the recent developments are anyway related to Neural Networks, which can effectively detect novelty by analysing the response of the trained neural network (trained on the normal data) to a test input. The great generalization ability of NNs in fact can lead to very accurate results. The main architectures are reported hereinafter.

Multilayer Perceptron (MLP), a feedforward artificial neural network whose weights can be efficiently optimized by backpropagation to output values close to a target when the input is normal.

Auto-associative networks (AA NNs) consisting of an input, an output and at least a middle bottleneck compression layer are commonly used in dimensionality reduction as a variant of Principal Component Analysis (PCA). Similarly to the already introduced Autoencoders, these can be used for ND by computing the error in the reconstruction made by a network trained on normal data. If this exceeds a threshold data can be considered novel.

Auto-associative networks perform a functional mapping and should not be confused with associative memories (e.g. Hopfield network) which are classifiers, recalling a stored exemplar that most closely resembles a partial or corrupted input pattern [55].

Hopfield network is composed of binary threshold units with recurrent connections between them. These nets have a scalar value associated with each state of the network, referred to as the network energy; when a normal observation is given to a trained network energy remains close to a minimum, while it increases when novel observations are sent as input.

Other artificial neural network producing a dimensionality reduction are the Self-organizing maps (SOMs). SOMs produce a low-dimensional (typically two-dimensional), discretized representation of the input space called a map. This map can be used for ND as after the training, only some part of the map activates when normal data is presented. If regions of the map further than the normal activates, novelty is detected [56].

In a similar fashion, Radial Basis Functions (RBF) networks create a real net in the original feature space. Neurons are centred in this space and activates if an observation falls into their range of action. This way, an RBF network outputs an approximation of the density of the training set, as Gaussian Mixture Models (GMMs) do. Observations in low density regions are then considered novel.

The previously introduced decision trees, which can learn hierarchies over pre-specified features with data-driven architecture, can be merged to neural networks to form

neural trees. In their simplest form, a neural tree is a set of perceptrons functionally organized in a binary tree which leads to a hierarchical quantization of the training data into equiprobable regions. A tree built on training data can be compared to a tree computed on novel data by Kullbach-Leibler Divergence, which can be used to discover novelty. [57]

Adaptive neural trees (ANTs) are also emerging, to adaptively grows the architecture from primitive modules as an improvement to the original binary tree. [58]

Other tree-shaped networks are the Bayesian. A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables and outputs the probability of the variable represented by the node. In other words, it reconstructs the probability of every possible event as defined by the combination of the values of all the variables (i.e. the joint probability distribution). By aggregating results with a single root node, novelty detection can be performed. Similar ideas may be applied to undirected, and possibly cyclic, graphs such as Markov networks.

Combining classical network models with real-life inspired characteristics, different approaches to Neural Networks Novelty Detection can be found in the literature.

For example, Habituation based networks are based on “habituation”, a decrement in the response of a repeated stimulus. This way a network can be trained to non-respond to normal inputs but only to novel [49,51].

Oscillatory networks also combine a classical network with dynamics. Oscillating networks are stable systems trained to quickly reach their equilibrium point after when normal data are given as input. When novel observations are shown to the network then, longer times are necessary to the net to get back to equilibrium, so that novelty can be assessed [51,59].

Adaptive resonance theory (ART) networks on the contrary are inspired to the brain information processing and implement lateral inhibition, the ability of an excited neuron to reduce the activity of its neighbours. In the net then, an input is taken from the comparison field and transferred it to its best match in the recognition field, which is the only neuron whose set of weights most closely matches the input vector. This feature of the net can be easily used for classification purposes [60].

Finally, information theory-based methods can be also used in novelty detection [49,50,51]. The basic idea is that outliers significantly alter the information content in a dataset, so that the measure of information contained in novel data will differ from the information level evaluated in the training. Measures of information are the Kolomogorov complexity, the entropy and other diversity indices such as the Shannon index, the Rényi entropy, the Simpson index, the Gini–Simpson index, etc.

Bibliography

[1] Fasana A., & Marchesiello S., "*Meccanica delle vibrazioni*". Torino: Clut, 2006. ISBN: 88-7992-217-3

[2] Randall R.B., "*Vibration-based condition monitoring: Industrial, Aerospace and Automotive applications*", Wiley, 2011. ISBN: 978-0-470-74785-8

[3] ISO 1940-1, "*Mechanical vibration — Balance quality requirements for rotors in a constant (rigid) state — Part 1: Specification and verification of balance tolerances*", Second edition, 2003.

[4] Palermo A., Britte L., Janssens K., Mundo D., Desmet W., "*The measurement of Gear Transmission Error as an NVH indicator: Theoretical discussion and industrial application via low-cost digital encoders to an all-electric vehicle gearbox*", Mechanical Systems and Signal Processing, Volume 110, 2018. DOI: 10.1016/j.ymssp.2018.03.005.

[5] Antoni J., Randall R. B., "*Rolling Element Bearing Diagnostics - A Tutorial",* Mechanical Systems and Signal Processing, 25(2), 485-520, 2011. DOI: 10.1016/j.ymssp.2010.07.017.

[6] W. Caesarendra, T. Tjahjowidodo, "A Review of Feature Extraction Methods in Vibration-Based Condition Monitoring and Its Application for Degradation Trend Estimation of Low-Speed Slew Bearing", Machines, 2017. DOI: 10.3390/machines5040021.

[7] P. Večeř, M. Kreidl, R. Šmíd, "Condition Indicators for Gearbox Condition Monitoring" Systems Acta Polytechnica Vol. 45 No. 6/2005. https://ojs.cvut.cz/ojs/index.php/ap/article/view/782

[8] H. L. Balderston, "The detection of incipient failure in bearings," Materials Evaluation, vol. 27, no. 6. pp. 121–128, 1969.

[9] Weichbrodt B and Smith KA. Signature analysis–nonintrusive techniques for incipient failure identification. General Electric Technical Information Series paper 70-C-364, 1970.

[10] S. Braun, "The Extraction of Periodic Waveforms by Time Domain Averaging", Acta Acustica united with Acustica, Volume 32, Number 2, pp. 69-77(9), 1975

[11] Beercheck, RC. "Listening for the sounds of bearing trouble", Machine Design,48,82-6, 1976

[12] R.F. Dwyer, "Detection of non-Gaussian signals by frequency domain kurtosis estimation", International Conference on Acoustic, Speech, and Signal Processing, Boston, pp. 607–610, 1983. DOI: 10.1109/ICASSP.1983.1172264

[13] McFadden, P. D. and Smith, J.D. “Vibration monitoring of rolling element bearings by the high frequency resonance technique-a review”, Tribology International, 17, 3-10, 1984. DOI: 10.1016/0301-679X(84)90076-8

[14] A. Davies, “Handbook of Condition Monitoring: Techniques and Methodology”, Springer Science & Business Media, 1998, DOI: 10.1007/978-94-011-4924-2_12

[15] J. Antoni, “The spectral kurtosis: a useful tool for characterizing non-stationary signals”, Mechanical Systems and Signal Processing 20, 282–307, 2006, DOI: 10.1016/j.ymssp.2010.07.017

[16] J. Antoni, “*Fast computation of the kurtogram for the detection of transient faults*”, Mech. Syst. Signal Process. 21,108–124, 2007. DOI: 10.1016/j.ymssp.2005.12.002.

[17] R.B. Randall, J. Antoni, “Rolling Element Bearing Diagnostics - A Tutorial”, Mechanical Systems and Signal Processing 25(2):485-520, 2011, DOI: 10.1016/j.ymssp.2010.07.017

[18] R.A. Wiggins, “Minimum Entropy Deconvolution”, Geoexploration, vol. 16, Elsevier Scientific Publishing, Amsterdam, pp. 21–35,1978. DOI: 10.1016/0016-7142(78)90005-4

[19] J. Li, X. Chen, Z. He, “Adaptive stochastic resonance method for impact signal detection based on sliding window”, Mechanical Systems and Signal Processing, Volume 36, Pages 240-255, 2013. DOI: 10.1016/j.ymssp.2012.12.004.

[20] Antoniadou I., “*Accounting for nonstationarity in the condition monitoring of wind turbine gearboxes”,* PhD thesis, University of Sheffield, 2013.

[21] Jardine A.K.S., Lin D., Banjevic D., “*A review of machinery diagnostics and prognostics implementing condition-based maintenance”,* Mechanical Systems and Signal Processing, 20, 1483- 1510, 2006. Doi: 10.1016/j.ymssp.2005.09.012

[22] Antoni J., Randall R. B., “*Rolling Element Bearing Diagnostics - A Tutorial”,* Mechanical Systems and Signal Processing, 25(2), 485-520, 2011. Doi: 10.1016/j.ymssp.2010.07.017.

[23] Carden E. P., Fanning P., “*Vibration Based Condition Monitoring: A Review”* Structural Health Monitoring, 3(4), 355-377, 2004. Doi: 10.1177/1475921704047500.

[24] Randall R.B., “*Vibration-based condition monitoring: Industrial, Aerospace and Automotive applications*”, Wiley, 2011. ISBN: 978-0-470-74785-8

[25] Antoni J., Randall R. B., “*Unsupervised noise cancellation for vibration signals: part I—evaluation of adaptive algorithms”*, Mechanical Systems and Signal Processing 18(1), 89-101, 2004. Doi: 10.1016/S0888-3270(03)00012-8.

[26] Antoni J., Randall R. B., “*Unsupervised noise cancellation for vibration signals: part II—a novel frequency-domain algorithm”*, Mechanical Systems and Signal Processing 18(1), 103-117, 2004. Doi: 10.1016/S0888-3270(03)00013-X.

[27] Yiakopoulos C.T., Gryllias K.C., Antoniadis I.A., "*Rolling element bearing fault detection in industrial environments based on a K-means clustering approach*", Expert Systems with Applications, Volume 38, Issue 3, 2011. DOI: 10.1016/j.eswa.2010.08.083.

[28] EASA ED Decision 2018/007/R CS-29 Amendment 5, 2018, "*Certification Specifications and Acceptable Means of Compliance for Large Rotorcrafts"*.

[29] Fassois S.D., Sakellariou J.S., "*Time-series methods for fault detection and identification in vibrating structures",* Phil. Trans. R. Soc. A (2007) 365, 411–448, 2007. Doi: 10.1098/rsta.2006.1929

[30] Christopher M. Bishop, Pattern Recognition And Machine Learning, 2006, Information Science and Statistics, Springer Verlag, ISBN-13: 978-0387310732

[31] A Gentle Introduction to Machine Learning, Mark Dahl, P.Geo. 2018, Recorder, vol 43 n 01. Official publication of the Canadian society of exploration geophysicists, CSEG. https://csegrecorder.com/articles/view/a-gentle-introduction-to-machine-learning

[32] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[33] A. Gibson and J. Patterson. Major Architectures of Deep Networks. In: Deep Learning. O'Reilly Media, 2017. ISBN: 9781491924570 [Online] Available from: https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/ch04.html [Accessed 27 Nov 2017]

[34] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[35] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in Advances in Neural Information Processing Systems, 2013, pp. 190–198

[36] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," ArXiv e-prints, Feb. 2014.

[37] S. F. Crone, S. Lessmann, and S. Pietsch, "Forecasting with Computational Intelligence - An Evaluation of Support Vector Regression and Artificial Neural Networks for Time Series Prediction," presented at IEEE World Congress on Computational Intelligence, Vancouver (Canada), 2006.

[38] Fisher R.A. (1936) " The use of multiple measurements in taxonomic problems", Annals of Eugenics, 7, 179–188

[39] R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" (PDF). The Computer Journal. British Computer Society. 16 (1): 30–34. doi:10.1093/comjnl/16.1.30

[40] D. Defays (1977). "An efficient algorithm for a complete-link method". The Computer Journal. British Computer Society. 20 (4): 364–366. doi:10.1093/comjnl/20.4.364

[41] Sokal R, Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin. 38: 1409–1438.

[42] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). ISBN 1-57735-004-9.

[43] Ankerst, Mihael; Breunig, Markus M.; Kriegel, Hans-Peter; Sander, Jörg (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60. CiteSeerX 10.1.1.129.6542

[44] Zhang, Zhenyue; Zha, Hongyuan (2004). "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment". SIAM Journal on Scientific Computing. 26 (1): 313–338. doi:10.1137/s1064827502419154.

[45] Bengio, Yoshua; Monperrus, Martin; Larochelle, Hugo (2006). "Nonlocal Estimation of Manifold Structure". Neural Computation. 18 (10): 2509–2528. CiteSeerX 10.1.1.116.4230. doi:10.1162/neco.2006.18.10.2509. PMID 16907635.

[46] A. Gibson and J. Patterson. Fundamentals of Deep Networks. In: Deep Learning. O'Reilly Media, 2017. ISBN: 9781491924570 [Online] Available from: https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/ch03.html [Accessed 27 Nov 2017]

[47] I. Goodfellow, et al. Generative Adversarial Networks. arXiv:1406.2661v1 [stat.ML], 2014. [Online] Available from: https://arxiv.org/abs/1406.2661 [Accessed 23 Nov 2017]

[48] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, Lionel Tarassenko, A review of novelty detection, Signal Processing, Volume 99, 2014, Pages 215-249, ISSN 0165-1684, https://doi.org/10.1016/j.sigpro.2013.12.026.

[49] Miljković, Dubravko. (2010). Review of novelty detection methods. 593-598.

[50] A. Banerjee, V. Chandola, V. Kumar and A. Lazarević, "Anomaly Detection: A Tutorial", Proc. of SIAM Data Mining Conference, Atlanta, GA, April 2008,

[51] V. Chandola, A. Banerjee and V. Kumar, "Outlier Detection: A Survey", Univ. of Minnesota TR 07-017, August, 2007

[52] M.Markou and S.Singh, "Novelty Detection: A Review Part I: statistical approaches",Sig. Proc.,Vol. 83,No.12,Dec. 2003

[53] M.Markou and S.Singh,"Novelty Detection:A Review Part II: neural network approaches",Sig.Proc,Vol.83,No.12,Dec.2003

[54] B.Schölkopf, R.Williamson, A.Smolax, J.Shawe-Taylory and J. Platt, "Support Vector Method for Novelty Detection", in Advances in Neural Information Processing Systems 12, The MIT Press, June 2000

[55] M.A. Kramer, Autoassociative neural networks, Computers & Chemical Engineering, Volume 16, Issue 4, 1992, Pages 313-328, ISSN 0098-1354, DOI: 10.1016/0098-1354(92)80051-A.

[56] T.Kohonen, "Self-Organizing Maps",2nd.Ed.,Springer, 1997

[57] J A Sirat & J-P Nadal (1990) Neural trees: a new tool for classification, Network: Computation in Neural Systems, 1:4, 423-438, DOI: 10.1088/0954-898X_1_4_003

[58] Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A. & Nori, A.. (2019). Adaptive Neural Trees. Proceedings of the 36th International Conference on Machine Learning, in PMLR 97:6166-6175

[59] R. Borisyuk, M. Denham, F. Hoppensteadt, Y. Kazanovich and Olga Vinogradova, "Oscillatory Model of Novelty Detection", Network Computation in Neural Systems, Vol. 12, No 1, pages 1 - 20, Jan. 2001

[60] T.Tanaka and A.Weitzenfeld,"Adaptive Resonance Theory", in Neural Simulation Language, The MIT Press, 2002

[62] Adrian Stetco, Fateme Dinmohammadi, Xingyu Zhao, Valentin Robu, David Flynn, Mike Barnes, John Keane, Goran Nenadic, Machine learning methods for wind turbine condition monitoring: A review, Renewable Energy, Volume 133, 2019, Pages 620-635, ISSN 0960-1481, DOI: 10.1016/j.renene.2018.10.047.

[63] Chen Z., Gryllias K., Li W., Mechanical fault diagnosis using Convolutional Neural Networks and Extreme Learning Machine. MECHANICAL SYSTEMS AND SIGNAL PROCESSING, 2019. DOI: 10.1016/j.ymssp.2019.106272

[64] Chen Z., Gryllias K., Li W., Intelligent Fault Diagnosis for Rotary Machinery Using Transferable Convolutional Neural Network. IEEE Transactions on Industrial Informatics, 2019. DOI: 10.1109/TII.2019.2917233

Chapter 4:

The selected algorithms

1. Introduction

According to chapter 2, the gearbox identification through vibration monitoring can be considered a data to decision (D2D) process. Indeed, the vibration measurements collected in datasets are analysed in order to produce a decision about the possibility of triggering an alert to communicate the need of an eventual corrective maintenance action. In this regard, according to costs and safety considerations, two monitoring philosophies are possible, leading to two different kinds of methodologies. In case of very expensive machines, for example, the cost of a permanent, continuous monitoring system is easily justifiable. Very quick and human-independent algorithms are then required to work on-line. These algorithms rely on simple features and their scope is bounded to diagnose impending failure and trigger an alert with some limited advance. Unfortunately, it is far more common in the industrial field to find relatively cheap machines with important down-time costs. In such cases an intermittent monitoring is preferable, so that less sensors (typically one) are used, connected to a light data acquisition system. In this case a more advanced data processing is preferable to produce a detailed diagnostic with long-term advance warning of developing faults. In this work, a selection of the algorithms found in the literature review both for Signal Processing and from Machine Learning environments are described and organized in the procedural scheme reported in Figure 1 so as to create two methodologies suitable for either a permanent or an intermittent monitoring. The selection is obviously just a small subset with respect to what can be found in the state of the art (Chapter 3). The main criteria used for the choice was the model interpretability, the degree to which the criteria for a decision can be understood by a human [67]. The algorithms effectiveness and efficiency were then assessed, and the complexity gradually increased to reach optimal diagnostics results at the minimum cost.

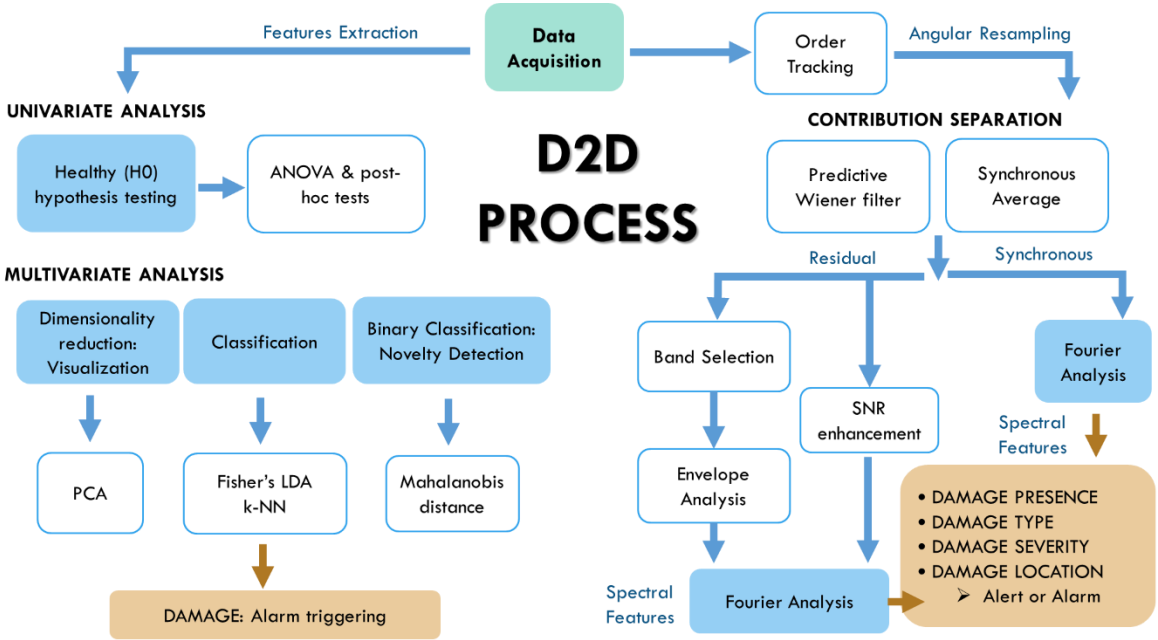


Figure 1: Proposed procedural scheme for Condition Based Monitoring

In particular, on the left half of the scheme (Figure 1), the methodology for a permanent, continuous monitoring derived from Machine Learning is given. In order to cope with the need of a quick, on-line data processing, it is designed using data-based techniques applied on simple features extracted directly from the raw time signals. These features can be treated independently, with multiple univariate analyses, or altogether, considering the complex correlation structure in a so-called multivariate analysis. Depending on the desired diagnostic level (enriched Rytter levels, chapter 2), different algorithms are proposed. In any case, the target of the analysis is mainly to detect the presence of a damage (Level 1 - Detection), triggering an alarm in case of danger.

On the contrary, on the right side of Figure 1, the methodology for an intermittent monitoring derived from Signal Processing can be found. This is based on the evaluation of frequency domain features for the different sources (basically gears and bearings). The raw, **asynchronously** sampled (i.e. with a constant sampling frequency in Hertz) signal is resampled at constant angular increments of a reference shaft and the different contributions are then separated so as to produce synchronous spectra, from which the damage-characteristic frequencies are identifiable (see chapter 3 for gearbox signatures). An additional tachometer is required for the Computed Order Tracking [14] to transform the measured signal from time domain to angular (or order) domain. The separation on the contrary is performed exploiting the different characteristics of gears and bearings signals. According to Wold's theorem [16] in fact, a stationary time series can always be written as a unique sum of a deterministic part (singular) and a stochastic part (regular). This enables to separate the gear signal, which is periodic with respect to the shaft angle (and then deterministic), from the bearing contribution, which is cyclostationary, as introduced in chapter 3. Several different algorithms can be used to perform the separation. One of the most reliable is the Synchronous Average (SA), extension of the Time Synchronous Averaging to angular domain signals [2]. A number of algorithms based on prediction is also available. Linear Prediction (LP), Adaptive Noise Cancellation (ANC), Self-Adaptive Noise Cancellation (SANC) or Discrete/Random Separation (DRS) belong to this category [16,17].

Despite the separation, the spectrum of the residual signal (the raw signal minus the deterministic signal) often contains little diagnostic information about bearing faults, whose characteristic signal is typically covered by stronger background noise. Anyway, over many years, Envelope Analysis established as the benchmark method for highlighting the bearing diagnostic information [15]. First, the signal is bandpass filtered in a high frequency band corresponding to some structural resonance, that can be excited by the bearing fault impulsivities. Then, it is amplitude demodulated by extracting its envelope, whose spectrum contains the desired diagnostic features.

2. Selected Signal Processing techniques for Intermittent monitoring

With a particular focus on bearing damages, whose detection is known to be very hard, the following methodology is proposed to perform an intermittent monitoring. As stated in chapter 2, intermittent monitoring usually has to rely on few sensors (typically one) and light data acquisition systems but at the same time, incipient damages need to be detected because long time can pass between successive acquisitions. In these regards, very advanced and complex data processing techniques are needed to work out a highly detailed analysis giving long-term advance warning of developing faults.

The proposed methodology is summarized in the procedural scheme reported in Figure 2.

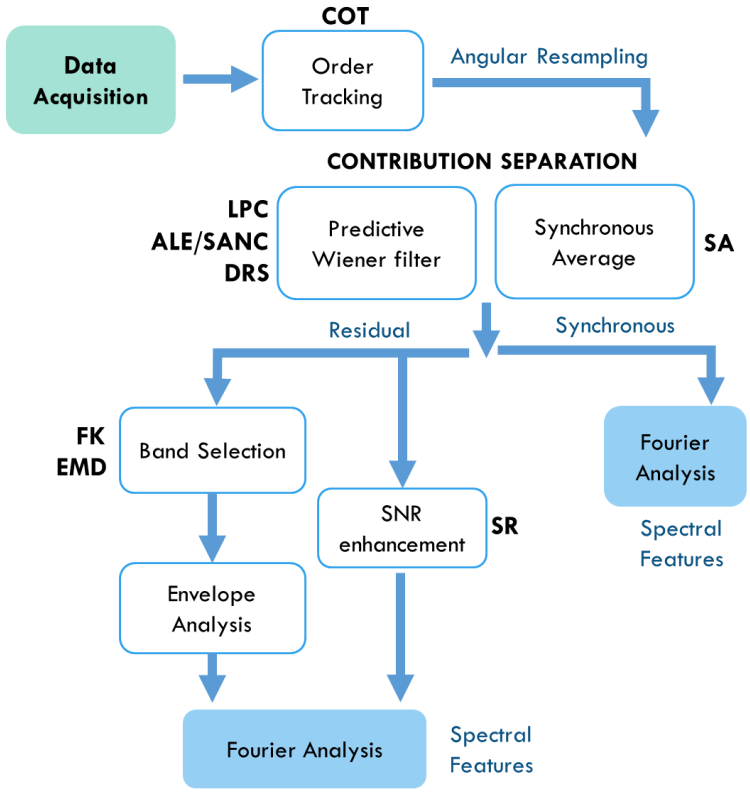


Figure 2: The proposed intermittent monitoring methodology.

The first issue to be faced when monitoring gearboxes is related to the fact that industrial machines usually work at variable speed. Despite a simple algorithm such as the Short Time Fourier Transform can highlight such speed variations (this is described in subsection 2.1 of this chapter), when a tachometer is available it is always wise to switch from the time domain to the angular domain. By resampling a time series at constant angular increments in fact, it is possible to “rephase” asynchronously sampled signals, so that the non-stationarity due to the variable speed can be compensated. This subject is tackled in subsection 2.2, describing the Computed Order Tracking algorithm, together with the Synchronous Average, a simple but effective algorithm able to extract the gear deterministic signal by averaging the resampled signal over time for the different angular positions.

Alternative algorithms for deterministic/non-deterministic separation based on prediction are treated in subsection 2.3. In particular Linear Prediction, Self-Adaptive

Noise Cancellation (known in acoustics as Adaptive Line Enhancer) and Discrete-Random Separation are considered. Once the residual containing the bearing signal and noise is obtained, Envelope Analysis can be used to demodulate the high frequency noise which becomes a carrier for the bearing diagnostic information. The reference method for selecting the optimal demodulation band is the Fast Kurtogram, described in subsection 2.4. Nevertheless, an alternative algorithm, the Empirical Mode Decomposition is proposed in subsection 2.5. Finally, in subsection 2.6 Stochastic Resonance is used to increase the signal to noise ratio of the residual to highlight the otherwise hidden bearing characteristic spectral lines in the spectrum.

2.1. Time-frequency STFT and asynchronous sampling issues

Rotating machines signals are known to be synchronous with the main shaft and (mechanically) phase-locked to the operating speed. When the rotational speed is perfectly constant, then, the usual sampling at uniform time increments results very effective in producing spectra in which the machine signature is clear (synchronous spectrum). Indeed, the signal power is concentrated at the frequencies corresponding to the relevant **orders**, namely multiples of the shaft frequency. Unfortunately, common machines are far from being operated at constant speed. In this case, the constant sampling rate leads to **asynchronously** sampled signals (with respect to the shaft frequency) which produces spectra featuring a spread of the signal power in a range dictated by the speed variability (asynchronous spectrum). The machine signature is no longer sharp, and the related diagnostic features becomes meaningless. This can be easily understood with a simple simulated signal featuring a single wave (e.g. the shaft rotational signal) with a variable frequency. A time-frequency analysis is suggested to visualize the problem. In particular, the Short Time Fourier Spectrum (STFT) is used in this case.

The **STFT** algorithm is here summarized:

- A long time-signal is divided into K shorter segments of length N samples. An overlap of O samples is usually considered in the division.
- Each realization k of length N samples is then windowed, and the Fourier transform is computed separately on each short segment. The k-th spectrum is then associated with time instant $t_k = {}^{1}/_{f_s}\left[{}^{N}/_{2} + (k-1)(N-O)\right]$, implying a time discretization $\Delta t = \frac{N-O}{f_s}$.
- The spectrum evolution with time can be obtained by placing the K spectra side by side. The 3-D surface so obtained takes the name of Waterfall (or Cascade) Plot. The same information can be summarized in a 2-D plot of frequency vs time, where the amplitude information is associated to a given color-map. This is the so-called Spectrogram.

The division in shorter chunks is ruled by the latent hypothesis that the speed can be considered quasi-constant in a limited time period $T = Nf_s$, leading to quasi-synchronous spectra. If this holds, the use of the STFT can give good results in picturing the phenomenon of the speed variation, at the expense of a reduced frequency resolution $\Delta f = \frac{f_s}{N}$.

The simple simulation reported in Figure 3 shows two waterfall plots. The first is referred to a signal featuring a speed fluctuation, very common in usual machines. This is modelled as a sinusoidal frequency variation in time. The second is a simulation of a

machine start up, with a linear speed increment up to the stable working condition. Comparing the Fourier spectra of the two overall signals to the ideal, constant speed case, it can be easily noticed how the power spreads over a range of frequencies (Figure 4). The main asynchronous spectra issue is then highlighted with respect to the synchronous spectrum case.

To overcome this issue, traditional analog DAQs were adapted to feature a sampling triggered by a square wave generated by a tachometer sensing angular references on the main shaft, instead of the usual electronic oscillator circuit giving a constant triggering frequency [1]. But some limits remain. First, a variable sampling frequency implies a non-optimized anti-aliasing filter, which can lead to a poorer acquisition. Moreover, the high cost and complexity of the equipment make it less desirable. Nowadays it is far more common to rely on usual (constant f_s) DAQs, adding to the acquisition a channel reserved for recording the signal from a tachometer mounted on the main shaft (keyphasor). The vibration signals are only later resampled at constant angular increments through a procedure called Computed Order Tracking.

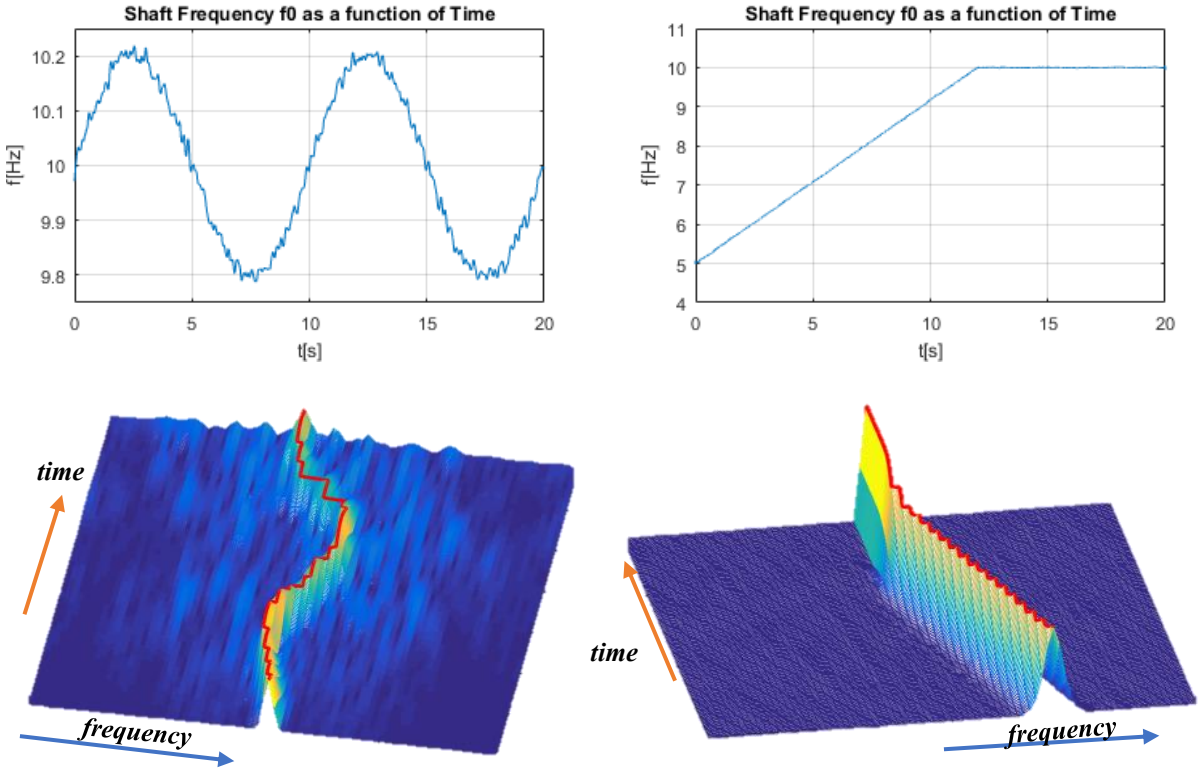


Figure 3: Waterfall plots for two frequency modulated harmonic waves. In the first case the frequency oscillates in time, while in the second there a linear dependence is considered.

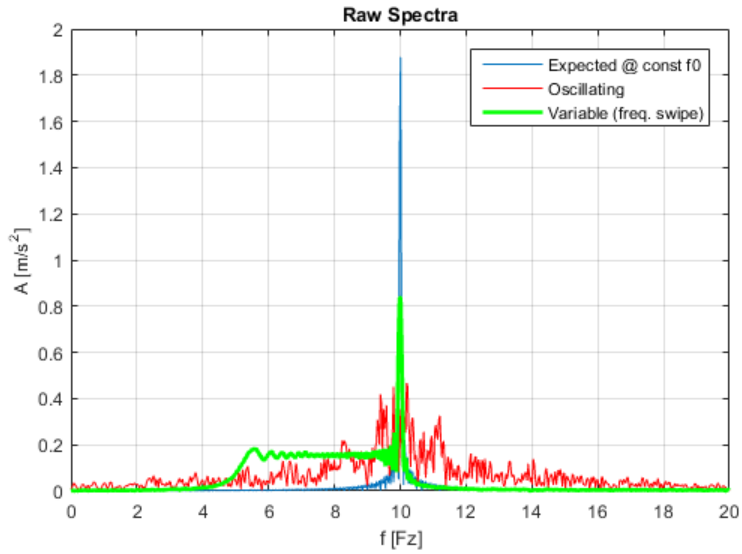


Figure 4: Amplitude spectrum for the two frequency-modulated harmonic waves (Figure 3) compared to the expected spectrum in absence of modulation (in blue).

2.2. Computed Order Tracking (COT) and Synchronous Average (SA)

As established, the COT is an algorithm meant to resample the asynchronous acceleration signal, originally sampled at a constant frequency f_s, to obtain a new discrete representation of the underlying continuous signal given at constant angular increments of the main shaft $\Delta\vartheta$. The COT algorithm, proposed by Fyfe & Munck in 1996 [1], does this in two steps.

- The signal from the tachometer is processed to extract the exact keyphase information: a vector t_{kp} (*keyphasor*) containing all the times at which a reference on the shaft (also *key*, or *trigger*) passes in front of the sensor is produced. The pulses induced in the tachometer signal must be then detected. At this purpose, a rising edge detector can be used to locate the passage times (e.g. the zero crossings), as shown in Figure 5.

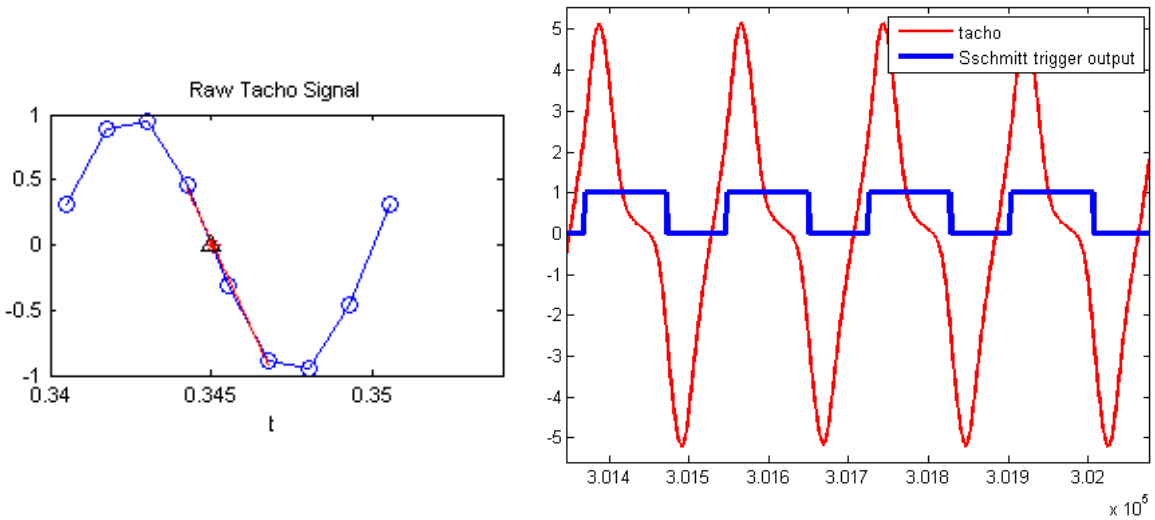


Figure 5: Zero crossing detection on the left side; rising edge detector (thresholds at $\pm 2{,}5$) on the right side.

- The new time axis for the angular resampling is built via interpolation. At first, the signal $\vartheta(t_{kp}) = n_{kp}\,\Delta\emptyset$ is reconstructed on the basis of the physical angle among two following keys $\Delta\emptyset$. Then, assuming a quadratic law of motion $\vartheta(t) = b_0 + b_1 t + b_2 t^2$, corresponding to a constant angular acceleration, the 3-points parabolic interpolating function can be inverted. Given a desired number of samples per rotation of the reference shaft (Samples Per Cycle, SPC), the sampling times corresponding to a uniform angular sampling $\theta(k) = k\frac{2\pi}{SPC} = k\,d\theta$ can be found, as better explained visually in Figure 6.

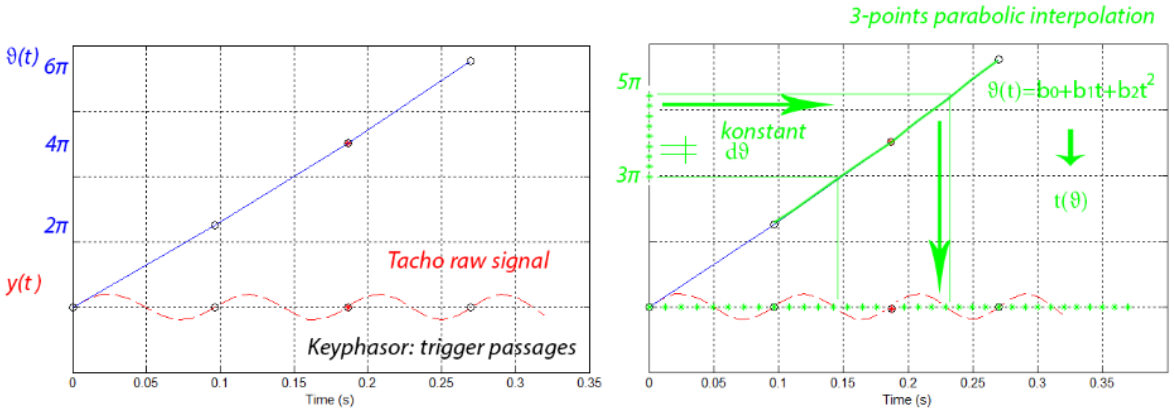


Figure 6: COT visually explained.

In particular, the coefficients b_i can be found solving the system of 3 equations:

$$\begin{Bmatrix} 0 \\ \Delta\emptyset \\ 2\Delta\emptyset \end{Bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \end{bmatrix} \begin{Bmatrix} b_0 \\ b_1 \\ b_2 \end{Bmatrix} \qquad (1$$

Then, by inverting the law of motion, the new time axis can be expressed by

$$t(\theta) = \frac{1}{2b_2}\left[\sqrt{4b_2(\theta - b_0) + b_1^2} - b_1\right] \qquad (2$$

- Finally, the new time axis can be used for resampling via interpolation of the original vibration signal, as shown in Figure 7 (spline interpolation). It is important to remember that the parameter **SPC** is the number of samples per order of the shaft speed and is then the analogous of the sampling frequency in the orders domain.

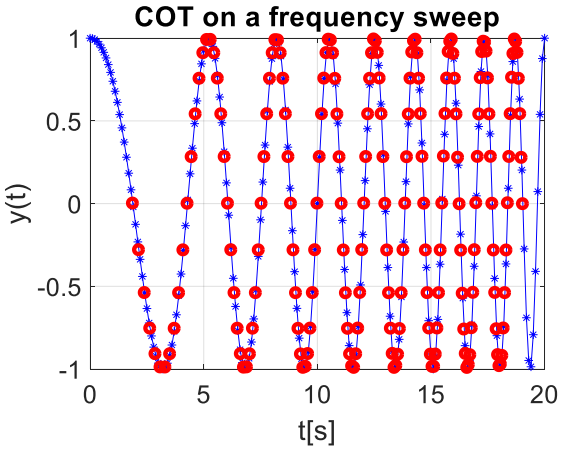


Figure 7: COT of a frequency sweep: in blue the original sweep sampled at constant time increments, in red the spline resampling at constant angular increments

The COT allows then to switch from an asynchronous spectrum in the frequency domain, to a synchronous spectrum in the domain of the orders of the main reference shaft. The power related to a given order is then concentrated again on a single order line, and not spread over wide ranges of frequency, as shown in Figure 8, where the order spectrogram for both the simulations introduced in subsection 2.1 of this chapter is reported.

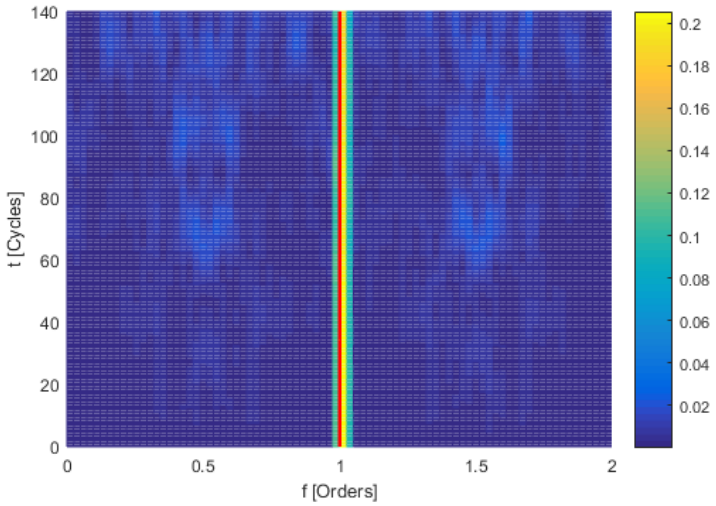


Figure 8: The effect of COT on the two frequency-modulated harmonic waves (Figure 3).

The advantage of having a signal sampled at constant angular increments can be also exploited to separate the deterministic component from the non-deterministic contribution. In particular, this can be done with the so-called **Synchronous Average** (SA) algorithm. SA means averaging the vibration signal acquired at the same angular position. This corresponds to a statistical aggregation allowing to discard all the contributions which are not periodic with the reference shaft, as it is possible to prove that they feature a zero-mean distribution [2]. The periodic component is then generated by replicating the base period for all the length of the original dataset, while the residual corresponds to the subtraction of the periodic part from the original signal. The procedural scheme of SA is graphically explained in Figure 9, where a simple mathematical formulation for aggregating the samples at angle θ_k (in red in the image) over the N cycles is also reported.

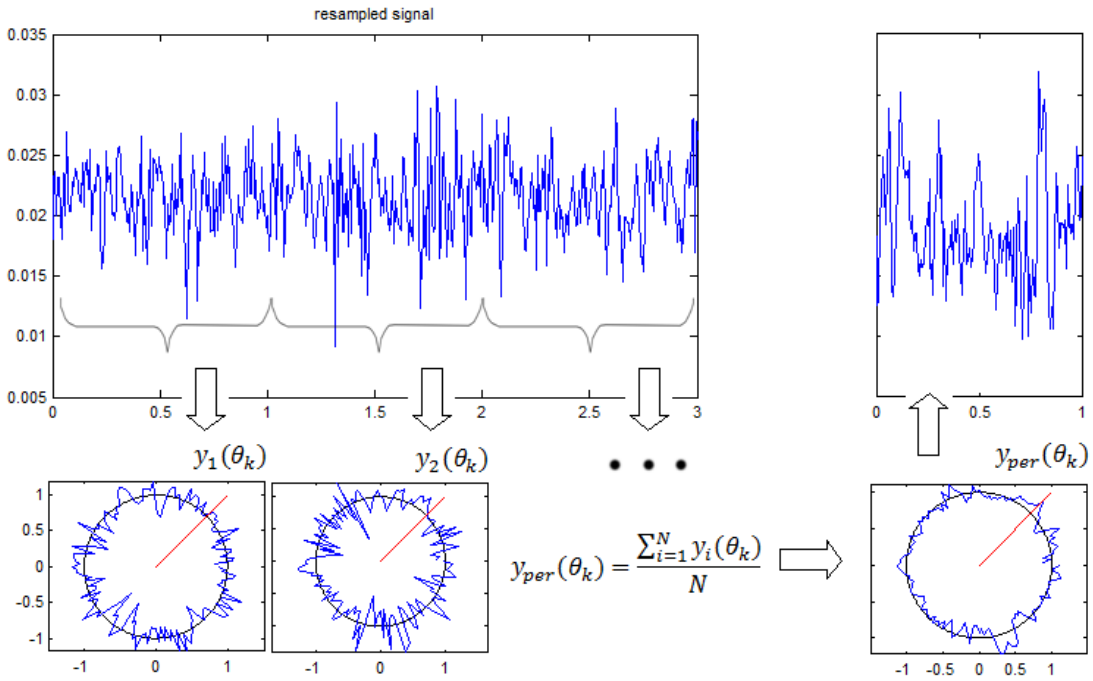


Figure 9: Synchronous Average explained visually.

In the frequency domain SA can be proved to be a comb filter [1] leaving unaffected the signal synchronous with the shaft and its multiples, while annihilating all other contributions (even the submultiples of the reference shaft speed). The orders representation of the corresponding filter is given by the equation

$$H(f) = \frac{Y_{per}(f)}{Y(f)} = \frac{1}{N}\frac{\sin(N\pi Tf)}{\sin(\pi Tf)} \qquad (3$$

where N is the number of averaged cycles, and a graphical representation can be found in Figure 10.

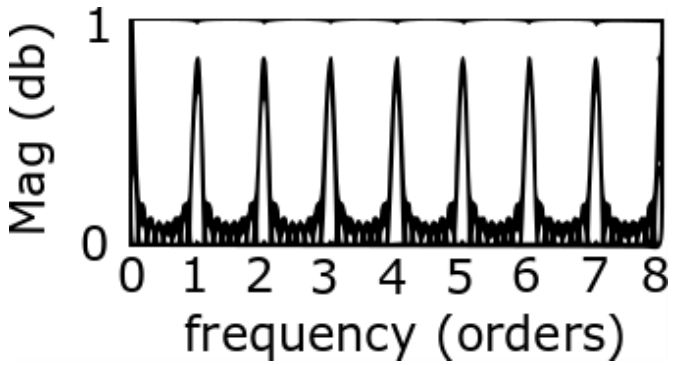


Figure 10: Order tracking filter frequency response (Randall & Antoni, 2011 [1])

The combination of COT and SA is a very powerful tool in diagnostics. Focusing on the periodic part for example, all the gears GMF harmonics and sidebands can be extracted at once, being entire multiples of the shaft speed. In the residual term, on the contrary, the bearing signals is left with minimum disruption. Therefore, its effectiveness in the separation is impressive, but its efficiency is limited by the fact that the whole procedure must be repeated for each shaft present in a machine, so that the computational burden may become huge in case of complex gearboxes.

2.3. Prediction based separation

The issue of decomposing a signal into a deterministic and a non-deterministic component is recurrent in signal processing. This task can be accomplished exploiting the signal properties introduced in Appendix 3. In particular, the definition of deterministic signals implies a complete specification, with no uncertainty at any time instant. On the contrary, random signals cannot be defined unless a probabilistic approach is used, as it is not possible to precisely define their value at a specific instant of time. Namely, periodic signals are predictable from any past values, arbitrarily far back in the past with zero prediction error, while predictability vanishes for random or chaotic signals. Non-deterministic signals are usually in between, and prediction from past values becomes less and less accurate as data become older (their correlation decays towards zero with increasing time-lags). Furthermore, focusing on stationary time series, Wold's work [4] demonstrates that it is always possible to write them as a unique sum of a deterministic part (singular) and a stochastic part (regular). This enables to separate the gear signal, which is periodic with respect to the shaft angle, and then deterministic in a synchronously sampled signal, from all other contributions (bearings plus noise).

2.3.1. Prediction theory: one step ahead prediction

The scope of prediction is to forecast the signal $y(n)$ from N past values, usually collected in a regressor vector $\overline{w_n} = [y(n-1), y(n-2), \dots, y(n-i), \dots, y(n-N)\ ]^t$. It is possible to prove that this problem can be optimally solved considering a *linear autoregressive form* of such as

$$y_p(n) = \bar{h}^t \cdot \overline{w_n} \qquad (4$$

In this framework, the prediction coefficients vector $\bar{h} = [h_1, h_2, \dots, h_i, \dots, h_N]^t$ minimizing the mean square prediction error (the cost function of the optimization) can be proved to be the Wiener filter.

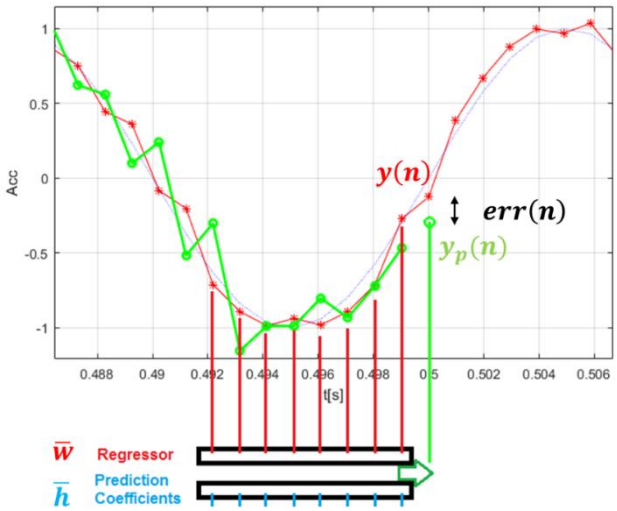


Figure 11: One step ahead prediction – example of AR(8) Wiener filter

Considering the cost function:

$$MSE = \sigma_{err}^2 = E[SE(n)] = E[|err(n)|^2] = E\left[\left|y(n) - y_p(n)\right|^2\right] \qquad (5$$

the optimal FIR filter $h = \text{argmin}\,\sigma_{err}^2$ (exactly the Wiener filter) has then a closed form solution resulting from the so-called Wiener-Hopf system of equations

$$\sum_{i=1}^{N} R_{yy}(i-m)h_i = R_{yy}(m) \quad for\ m = 1, \dots, N \qquad (6$$

Where $Ryy(k)$ is the autocorrelation function $E[y(n)y(n-k)]$ of the sequence $y(n)$ for a lag k, estimated as $Ryy(k) = \frac{1}{N-k}\sum_{l=1}^{N} y(l)y^*(l-k)$. Solving the least squares problem via the normal equations, the Yule-Walker system of equations is obtained, so that the vector of the coefficients $\bar{h}$ can be computed from $\bar{\bar{R}}\bar{h} = \bar{r}$.

$$\begin{bmatrix} Ryy(0) & Ryy(-1) & \dots & Ryy(1-N) \\ Ryy(1) & Ryy(0) & \dots & Ryy(2-N) \\ \vdots & \vdots & \ddots & \vdots \\ Ryy(N-1) & Ryy(N-2) & \dots & Ryy(0) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} = \begin{bmatrix} Ryy(1) \\ Ryy(2) \\ \vdots \\ Ryy(N) \end{bmatrix} \qquad (7$$

The algorithm summarizing these operations is named Linear Prediction (LP), often called **Linear Predictive Coding** (LPC). It is important to note that to improve the deterministic/non-deterministic separation, the algorithm can be modified for performing a Δ steps ahead prediction, so as to exploit the difference in the autocorrelation of the two components.

In practice, however, the filter length N can be very large (hundreds or thousands of samples) so that the Yule-Walker system cannot be solved directly. In this case, the optimization problem can be tackled considering an adaptive process in which the coefficients of the FIR filter are adjusted from iteration to iteration using a gradient descent algorithm such as the Least Mean Square. The simple LMS adaptation rule turns out to be

$$\begin{aligned} h_i^{n+1} &= h_i^n - \mu\, \nabla SE(n) \\ h_i^{n+1} &= h_i^n - \mu\ err(n)\, y(n-i) \end{aligned} \qquad (8$$

with a strictly positive convergence coefficient μ, also called forgetting factor. To overcome the sensitivity to the input signal scale, the Normalized LMS (NLMS) is often used

$$\overline{h^{n+1}} = \overline{h^n} + \frac{\mu\ err(n)\, \overline{w_n}}{\epsilon + \overline{w_n}\,\overline{w_n}^t} \qquad (9$$

adding a scale parameter ϵ preventing the ratio from diverging.
The LMS-based recursive algorithm takes the name of **Self-Adaptive Noise Cancellation** (SANC) in the acoustic community, while in the signal processing community the name **Adaptive Line Enhancer** (ALE) prevails [4].

It can be useful to face the problem of finding the optimal Wiener filter also from a frequency domain point of view. In this case, the convolution of the FIR to the signal simplifies to the product

$$y_p(n) = h(n) \otimes y(n) \quad \rightarrow \mathcal{F} \rightarrow \quad Y_p(f) = H(f)^* Y(f) \qquad (10$$

The scope is then to estimate the transfer function among a delayed version of the original signal $Y^d(f)$ and the only deterministic part of the signal $D(f)$.

$$Y_p(f) = H(f)^* Y(f) \quad \rightarrow \quad D(f) = H(f)^* Y^d(f) \qquad (11$$

This is performed by the **Discrete-Random Separation** (DRS) algorithm [5], which exploits an estimator of the transfer function $H(f)$ based on the average power spectral density of chunks of the signal.

$$H(f) = \frac{\sum_{k=1}^{K} Y_k^d(f)^* Y_k(f)}{\sum_{k=1}^{K} Y_k^d(f)^* Y_k^d(f)^*} \qquad (12$$

The scheme in Figure 12 better describes the procedure, identifying the main steps, and highlighting the N-long sliding windows (e.g. Parzen window in this case), producing the M-long filter, with $M \geq N$.

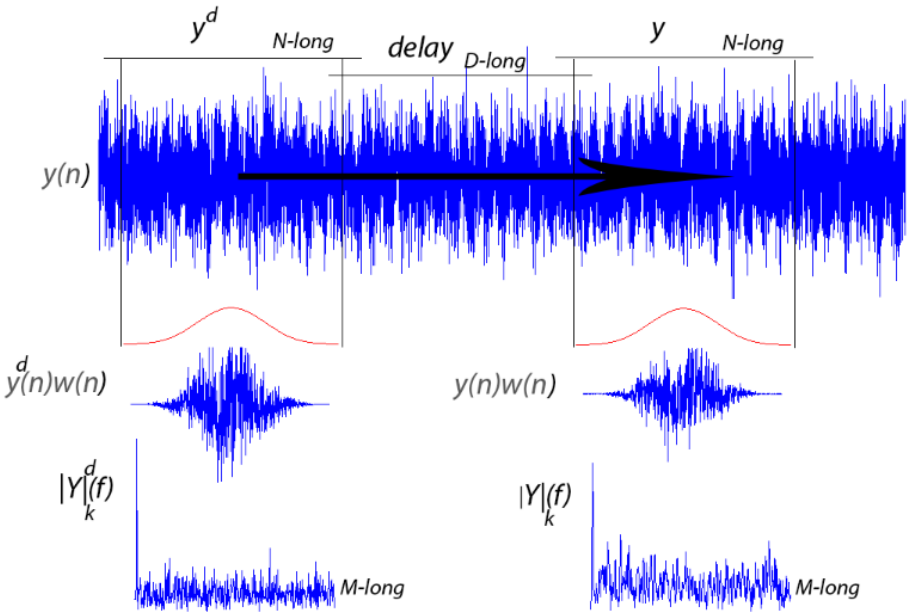


Figure 12: DRS explained visually.

2.4. Envelope Analysis and Spectral Kurtosis

The diagnostics of rolling element bearings is founded on Envelope Analysis. Localized faults in a bearing, in fact, generate impacts which excite the structural resonant frequencies and induce a modulation phenomenon in the acquired acceleration signals. In particular, faulty bearings are believed to cause an amplitude modulation to the high frequency noise, which becomes a carrier for the diagnostic information. Demodulation of the original signal via analysis of the envelope is then the natural processing (Appendix 2) to recover the bearing-characteristic spectral lines, which are usually so weak with respect to the background noise, they can rarely be detected in the row signals spectra. To enhance the signal-to-noise ratio of a bearing signal, a band-pass filter is usually set around the desired resonance frequency band before the demodulation. However, the

selection of the most appropriate band for demodulation raised a heated debate over the years. In accordance with the believed modulation process, in the past, a hammer tap testing was used to find beating housing resonances [7]. In any case, the band-selection problem has now largely been solved using the Spectral Kurtosis (SK) and in particular the Kurtogram to find the most impulsive band of the residual signal (i.e. after the removal of the deterministic part of the signal).

The use of kurtosis is justified by the fact that the CS2 impulse response train signal typical of a damaged bearing shows a distribution which is not normal and with heavy tails, as highlighted by the simple simulation with a synthetic signal (refer to chapter 4) reported in Figure 13. It is important to point out that for such a simulated signal, the value of the excess kurtosis shoots up to a value of 72 and is then the ideal feature highlighting the impacts produced by a damaged bearing. In fact, Kurtosis corresponds to the expected value of the standardized data raised to the fourth power. All the values within the range ± 1 (the 68% in case of a normal distribution) contribute almost nothing to kurtosis (raising a number smaller than 1 to the fourth power makes it closer to zero). On the contrary, the farther values show larger and larger contributions as their distance from the mean increases. Hence, the only interpretation of kurtosis is in terms of tail extremity i.e. outliers and must not be mistaken for a measure of the "peakedness" of the distribution itself [22,23]. In this regard, Figure 13 helps in visualizing the heavy tails which are related to such a high value of kurtosis.

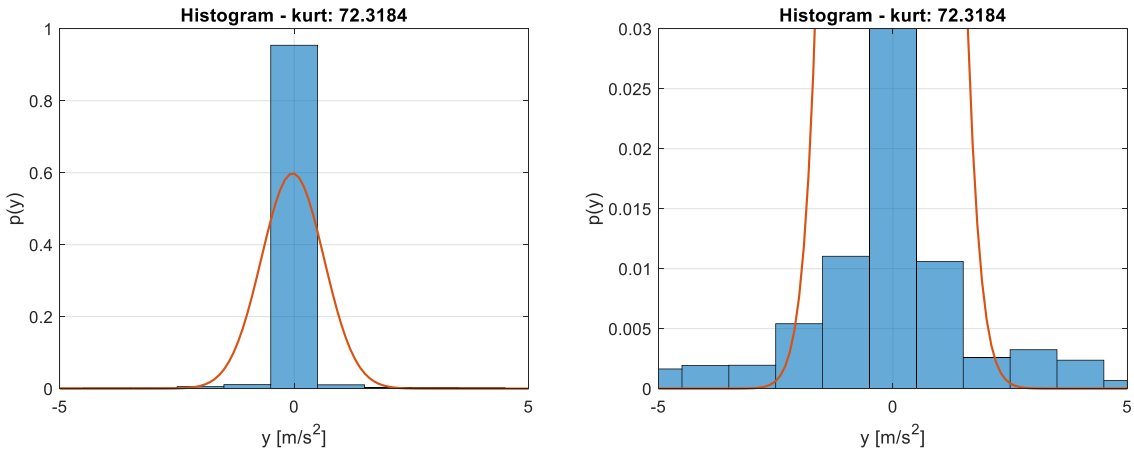


Figure 13: CS2 impulse train histogram vs normal distribution (in red)

2.4.1. The Spectral Kurtosis

The Spectral Kurtosis (SK) was first defined by Dwyer [9] as the normalized fourth order moment of the real part of the STFT of a signal. Later Pagan and Ottonello [10] proposed to move to the fourth order moment of the magnitude of the STFT. Traditionally, SK is based on the STFT, introduced in subsection 2.1 of this chapter. To summarize, the standard Fourier analysis, which gives no time information, is improved by transforming windowed sections of the data. Time resolution is then obtained by centring the window function on the epoch of interest and by sliding the window along the time axis. The drawback is that STFT is not able to give simultaneously high time and frequency resolutions. In any case, starting from a spectrogram, the time frequency representation $X(t, f)$ can be seen as the complex envelope of the signal band-pass filtered around the frequency f (ideal, infinitely narrow band filter), and its squared magnitude then indicates

the way energy is flowing in that frequency with respect to time [2]. If the frequency band happens to carry pulses, bursts of energy will appear, and this can be detected by computing the excess, normalized fourth order moment of the complex envelope:

$$SK(f) = \frac{\langle |X(t,f)|^4 \rangle}{\langle |X(t,f)|^2 \rangle^2} - 2 \qquad (13$$

where the -2 is used to enforce $SK(f) = 0$ in case of a complex Gaussian $X(t,f)$.

From a practical point of view, STFT gives a finite frequency discretization Δf which is function of the time window length. Then, the complex envelope becomes a function of this additional parameter: $X(t,f,\Delta f)$. Computing the spectral kurtosis for different frequency discretizations ($SK(f,\Delta f)$) and summarizing it as a colormap in the $[f,\Delta f]$ plane, the Kurtogram is built. As already introduced, the main weakness of this procedure is the difficulty of STFT in obtaining a good discretization in both frequency and time domain, so that a different approach was developed by Antoni [8]. In order to find the kurtosis in all the required frequency bands, the signal is processed by a quasi-analytic FIR filter bank producing a division of the $[f,\Delta f]$ plane (paving). This paving was originally dyadic but was later improved to a 1/3 binary tree division, which can better cover the frequency axis. The procedure finally obtained takes the name of **Fast Kurtogram (FK)**.

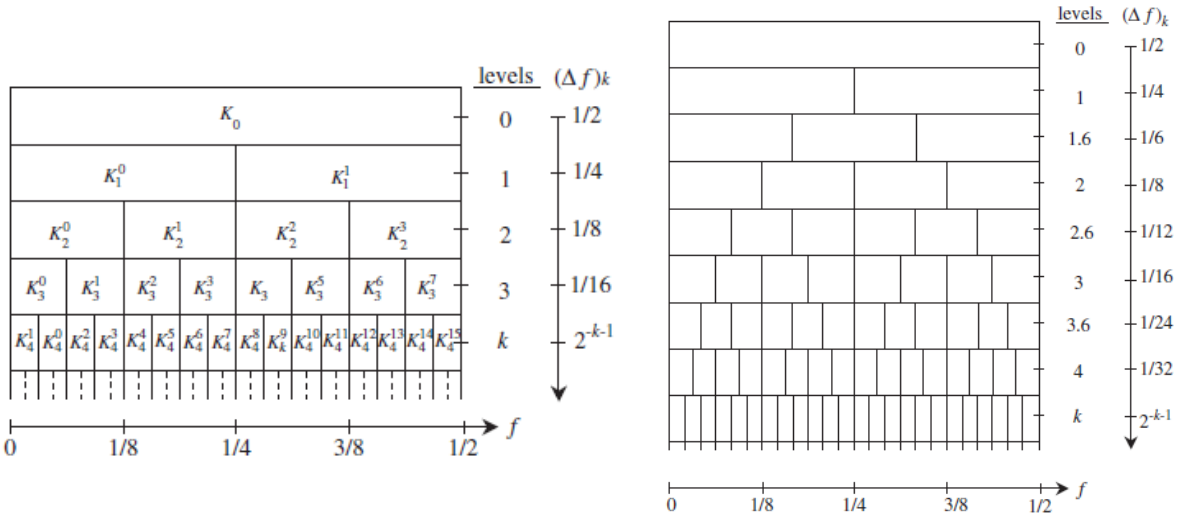


Figure 14: binary vs 1/3 binary tree paving for FK [8].

Several improvements have been proposed over the years, but no one proved to be as reliable and computationally efficient as the FK.

A simple but effective novel implementation is presented in Chapter 6.

2.5. Empirical Mode Decomposition (EMD)

In order to broaden the perspective on bearing diagnostics, a different approach to Envelope Analysis is reported. It is based on the Empirical Mode Decomposition (EMD), a relatively novel technique proposed by Huang in the mid 90's [12] and later successfully applied in many fields, from acoustics to biology (e.g. oceans, earthquakes, climate, etc.), and then also extended to fault diagnostics. EMD is a powerful self-adaptive and data driven method to decompose a multi-component signal into a sum of possibly non-linear, non-stationary empirical modes. In contrast with Fourier transform, which decomposes a signal into a sum of simple harmonic waves, EMD is able to produce a sum of amplitude and frequency modulated components called Intrinsic Mode Functions (IMFs) in an empirical way. Despite EMD is often referred to as Hilbert-Huang's Transform (improperly, as EMD is just a part of HHT), it is not a true transform, but an empirical approach to produce an unpredictable but finite number of elementary components which admit a Hilbert transform. Hence, a signal $s(t)$ can be modelled as multimodal entity and can be summarized as

$$s(t) = \sum_{j=1}^{M} a_j(t) \cos\Big(\varphi_j(t) \Big) \qquad (14$$

In simple words, EMD considers the signal $s(t)$ as a fast mode on top of slower oscillations. Hence, with a procedure called "sifting", it identifies locally the fastest oscillation (first IMF), subtracts it from the signal and iterates on the residual to find in the same way the remaining IMFs.

An IMF must feature two main properties:

- the number of local maxima (typically positive) and minima (negative) and the number of zeros of the function must be equal or differ at most by one
- at any instant, the mean value calculated averaging the maximum and the minimum envelopes must be null.

In light of these requirements, the sifting algorithm can be derived. It consists of few fundamental iterative steps:

1. Identification of local maxima and minima and interpolation (e.g. cubic splines) to generate the lower and the upper envelopes $e_l(t),\ e_u(t)$.
2. Compute the mean envelope $m_1 = 0{,}5\ (e_l(t) + e_u(t))$ and subtract it from the signal to obtain the presumed IMF $h_1(t) = s(t) - m_1(t)$
3. Check whether $h_1(t)$ shows the main IMF properties. If this is false, iterate the procedure on $h_1(t)$ until the first IMF is found.
4. Compute the residual $r_1(t) = s(t) - IMF_1$ and start again the procedure on this residual.
5. The procedure is stopped when $r_n(t)$ becomes a monotonic function or a constant.

The procedure leads then to an automatic split of the original signal in a non-predictable number of IMFs, characterized by progressively lower frequencies.

The great advantage is that EMD can automatically adapt to any signal, even with wide frequency contents, and does not require a priori knowledge on the signal nor control parameters.

Unfortunately, some weaknesses cannot be underestimated. First of all, the algorithm uses interpolation, and can eventually introduce aliasing. Then, the separation, even for a

dummy sum of two harmonic modes, is possible only under some restrictions: amplitude ratio $a = a_2/a_1 < 0.5$ and frequency ratio $f = f_2/f_1 < 1/a$ [13], as shown in Figure 15.

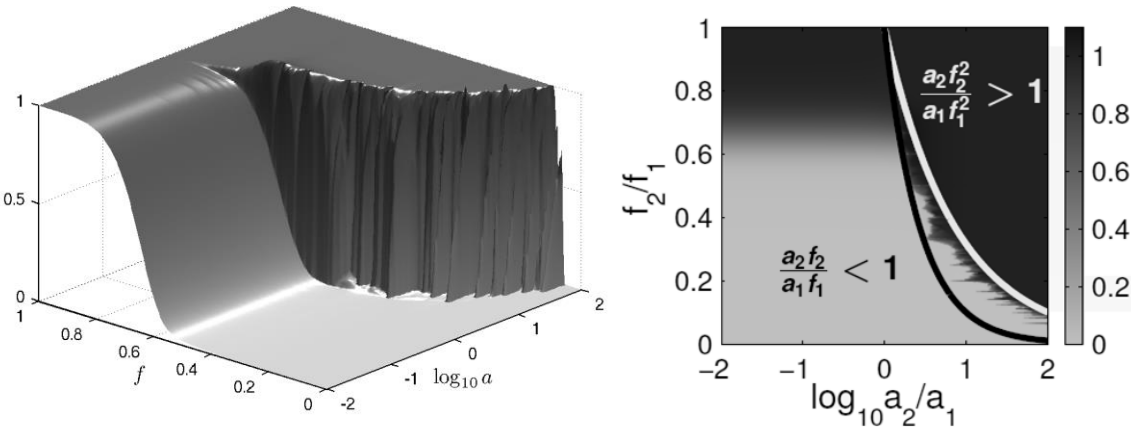


Figure 15: Separation index: a value of 0 indicates a perfect separation. [13]

In case of more complex amplitude and frequency modulated components, the problem of mode-mixing arises, so that EMD is no more able to distinguish the components and the resulting IMFs jump intermittently over time from one mode to another. This phenomenon is highlighted in Figure 16, where two frequency modulated modes are synthesized according to

$$c(t) = a \sin\left(\omega_c t + \int \omega_m(t) dt\right) \qquad (15$$

In the first case, the two carriers are featuring respectively 200 and 1600 Hz, while in the second case they are brought to 300 and 1000 Hz. The modulation is generated to increase the resulting frequency up to a maximum and then get back to the original value. The overall frequency as a function of time is shaped then as a raised cosine, so that at $t = 10s$ both the modes are at the maximum frequency. In this condition, the low carrier frequency (mode 2) is incremented of 300 Hz while the high frequency (mode 1) is incremented of 150 Hz. As it is easy to notice in Figure 16, aliases are present in both cases, but in the second one mode-mixing occurs. In practice, focusing on the frequency domain, EMD behaves as a series of overlapping auto-tuned band pass filters characterized by progressively lower centre frequencies. This is also confirmed by Flandrin [15] in which EMD is recognized to be equivalent to a dyadic filter bank (a wavelet-like filter bank) when it was applied to fractional Gaussian noise.

In general, in the presence of noise, the performance of EMD improves. The addition of Gaussian noise for example is at the base of the Ensemble Empirical Mode Decomposition (EEMD), a variation of EMD able to mitigate the mode mixing problem. Indeed, EEMD defines the true IMF components as the mean of an ensemble of trials, each consisting of the signal plus a white noise of finite amplitude [14].

Hence, in actual vibration signals, where the noise level is usually very high, EMD approaches the behaviour of a filter bank, and acts in almost the same way as the kurtogram. That is why it can be used to assess the most suitable band for Envelope Analysis [16, 17]. Comparing the IMFs on the basis of their kurtosis, for example, the IMF with the highest impulsiveness content can be found. This IMF can be considered as a filtered version of the original signal, at a given band and centre frequency which are

automatically selected by EMD in an adaptive way. Envelope analysis can be then conducted on such a signal in the very same way as seen in paragraph 5.

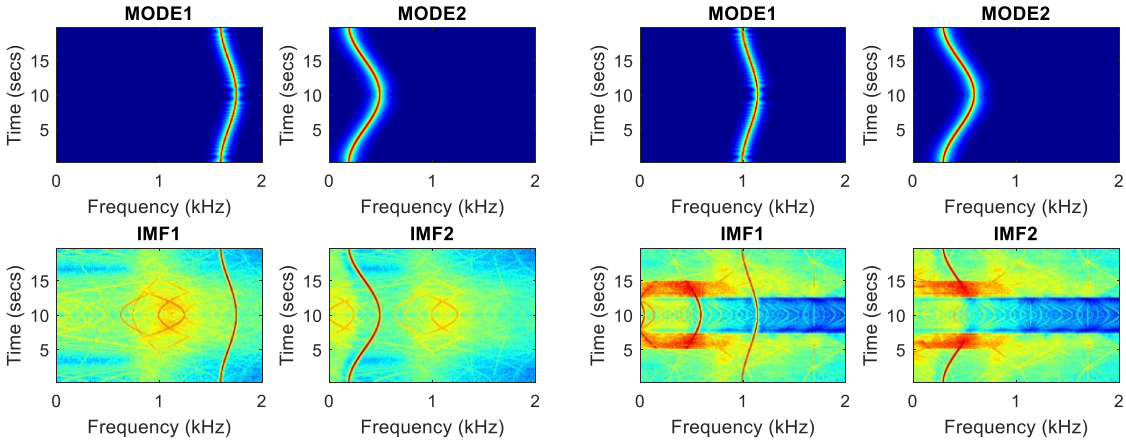


Figure 16: EMD applied on two different synthetic signals composed by a sum of 2 frequency modulated modes. On the left: $\omega_{c,1} = 1600Hz$ *and* $\omega_{c,2} = 200Hz$*. On the right:* $\omega_{c,1} = 1000Hz$ *and* $\omega_{c,2} = 300Hz$*. In both cases* $\Delta\omega_{max,1} = 150Hz, \Delta\omega_{max,2} = 300Hz$,

2.6. Stochastic resonance (SR)

The scope of Envelope Analysis, as clarified in previous chapters, is to highlight the bearing-characteristic signature, which is in practical cases always very weak with respect to the gears signal and the background noise. In any case, Envelope Analysis is not the only technique capable of increasing the signal to noise ratio selectively for the bearing spectral features. Among others, the Stochastic Resonance stands out [24,25,26,27,28].

Stochastic Resonance was first introduced in 1981 by Benzi et al. in a study about the evolution of the Earth's climate to explain the phenomenon of glacial cycles, substantial variations of the average Earth's temperature of about 10K every 100.000 years [18]. In electronics, SR found many applications, in particular as a sort of random amplifier for the faint signals, far below the resolution of the available acquisition board [19]. In simple words, a bi-stable system such as the Schmidt trigger can be used to detect a weak deterministic signal to which a white noise is added.

The Schmidt trigger is a comparator circuit with hysteresis, implemented by applying a positive feedback to the noninverting input of a differential amplifier (Figure 17-up). It can be proved that adding a white noise to a weak input (smaller than the thresholds of the bi-stable device) can produce as output a square wave whose frequency content reproduces the spectrum of the undetectable original signal. The critical issue is the selection of the noise intensity. In fact, in both cases when the noise level is too low or too high, no detection is possible. This form of "resonance" only occurs at the optimal noise level. Hence the name Stochastic Resonance. An example of the stochastic resonance phenomenon is highlighted in Figure 17, where one can appreciate how the output square wave (blue) can roughly reproduce the frequency content of the deterministic but faint signal (magenta) thanks to the addition of Gaussian noise.

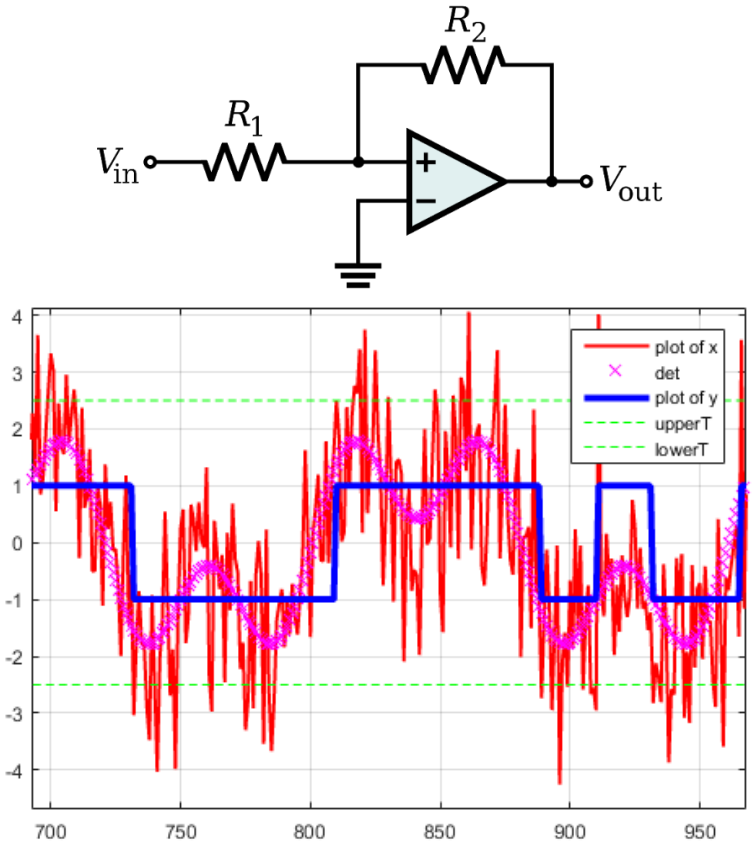


Figure 17: Schmidt trigger op-amp circuit (up) and detection of a signal (red) composed by the sum of a deterministic part (violet) and noise via symmetric thresholds (green). The output in blue proves to roughly follow the deterministic part.

SR also found many applications in the field of physics, in particular to study the Brownian motion of particles suspended in a fluid, resulting from their collision with the fast-moving atoms or molecules of the gas or liquid. This is ruled by the Langevin equation:

$$m\ddot{x}(t) + \lambda\dot{x}(t) = -U'(x) + \sqrt{D}\,\xi(t) \qquad (16$$

where m is the particle mass, λ is a friction coefficient, $\sqrt{D}\,\xi$ is the stochastic force of intensity D and $U(x)$ is the potential function. A bi-stable potential featuring two wells can be given by the equation:

$$U(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4 \qquad (17$$

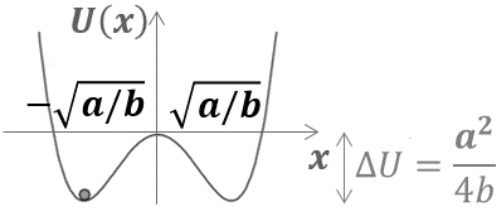


Figure 18: The double-well potential.

This differential equation is used as a reference also for the machine diagnostics field, which in most of cases uses a simplification for the overdamped motion (negligible inertia) of a Brownian particle in the bi-stable potential.

The differential reduces to

$$\dot{x}(t) = a\,x(t) - b\,x^3(t) + \ sig\ (t) \qquad (18$$

where $sig\ (t)$ is the input signal, which for diagnostics applications is a sum of a deterministic, a non-deterministic and a noise term. It is obvious that in order to apply SR, the parameters characterizing the potential a and b must be optimized in some way, and a numerical integration scheme is needed to solve the differential equation.

The simpler explicit method for the integration of ordinary differential equations is the Forward Euler method. Based on truncated Taylor series expansion, it approximates the derivative of a signal as the difference of two following samples divided by the time step:

$$\begin{gathered} y(t+h) = y(t) + h\dot{y}(t) + \ {}^{1}/_{2}\ h^2\,y''(t) + O(h^3) \\ y(t+h) = y(t) + h\dot{y}(t) \\ \dot{y}(t) = \frac{y(t+h) - y(t)}{h} \end{gathered} \qquad (19$$

This formulation accepts then an estimation error of order $O(h^2)$. A generalization for improving the accuracy of the estimation of the derivative is possible using a weighted average of K increments k_i of the function inside a single interval. This takes than name of Runge-Kutta order K:

$$y(t_0 + h) = y(t_0) + h \sum_{i \in [1:K]} b_i k_i \qquad (20$$

where b_i are the non-negative weights which must sum up to one.
Runge-Kutta 4 is probably the most widespread, usually implemented as the Runge–Kutta–Fehlberg 4(5) method, which is of order 4 with an error estimator of order $O(h^5)$. Quantifying this error at each step and using it to control the step-size adaptively, finally brings to the ODE45 Matlab® algorithm. Focusing on bearing diagnostics, it is common to have acquisitions at very high sampling rates, which aims to represent also the high frequency resonances useful in the demodulation process. When the sampling frequency is very high then the error is restrained to acceptably low values, so that also the simple Euler solver, corresponding to a Runge-Kutta 1, can lead to good results. A comparison of the main solvers on a chunk of experimental signal is shown in Figure 19.
Using RK1, the differential equation can be brought to a finite difference form

$$y(t+h) = \big(1 - h(b\,y^2(t) - a)\big)\,y(t) + h\,sig\ (t) \qquad (21$$

which can be compared to the one for a first order low pass filter with time constant $\tau = 1/\omega_{cutoff}$:

$$y(t+h) = \left(1 - \frac{h}{\tau + h}\right) y(t) + \frac{h}{\tau + h}\,x(t) \qquad (22$$

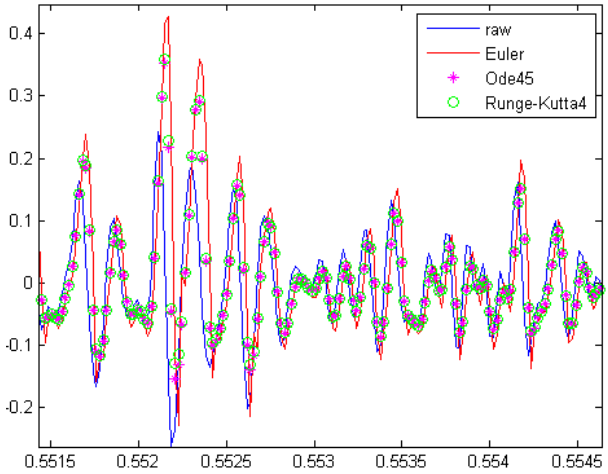


Figure 19: Different integrators applied to the same input (raw, in blue).

Relevant considerations arise from the SR interpretation as a sort of non-linear low pass filter. First of all, the non-linearity term $b\,y^2(t) - a$ does not only depend on the squared signal, but also on the potential parameters a and b. This explains why the signal magnitude should match the double well potential size (given by a and b) and is often rescaled to a given rms value $\boldsymbol{s}$.

$$\tilde{y}(t) = \boldsymbol{s}\frac{y(t)}{rms\big(y(t)\big)} \qquad (23$$

Furthermore, in case of $b\,y^2(t) - a \approx 1$ SR reduces to a perfect low pass filter with a given, nearly unitary cut-off ($\omega_{cutoff} \approx 1/(1-h)$). This clarifies the need of an additional parameter R deforming the time axis to tune the band of action of the SR.

$$\tilde{h} = \boldsymbol{R}h \qquad (24$$

These last considerations are obviously very qualitative and cannot be used to select the best values of the SR parameters. Summarizing, four parameters must be finally selected: a, b, R and s. A performance evaluation criterion is then fundamental for such optimization. As introduced at the beginning of paragraph 5, the maximization of the kurtosis is a very good criterion for detecting a train of impulse responses. In practical cases, however, when multiple impact signals are present with different amplitudes and frequencies, the reliability of kurtosis may be affected i.e. gear impacts and strong noise may mask the damaged bearing signal.

A more robust parameter proposed in [20] involves the Parson correlation coefficient $\left|\rho_{y,sig}\right| \leq 1$

$$\rho_{y,sig} = \frac{cov\big(y(t), sig(t)\big)}{\sigma_y \sigma_{sig}} \qquad (25$$

to produce a weighted kurtosis indicator such as

$$wk = \left|\rho_{y,sig}\right| kurtosis(y) \qquad (26$$

The motivation is related to the fact that the correlation coefficient between y and the bearing signal to be detected is maximum when y is exactly that signal. Obviously, the true bearing signal is never known, so that it is substituted with the input signal $sig(t)$ which contains it.

When a particular spectral line is sought, a signal to noise ratio computed in the spectrum can be used as an optimization parameter for enhancing that particular spectral line or a group of them with respect to the other spectral lines [21].

$$snr = \frac{A(n_b)}{\sum_i A(i)} \qquad (27$$

where $A(i)$ is the amplitude of the i-th spectral line and $i \in [n_b - m, n_b) \cup (n_b, n_b + m]$ is the range of frequencies in the vicinity of the frequency n_b of interest.

Finally, the four parameters can be optimized to find the best combination of values which produces an enhancement of the bearing characteristic spectral lines with respect to the background noise. The Genetic Algorithm proposed for example in [21] is just one of the many optimization algorithms available but thanks to its simplicity and proper balance between exploitation and exploration of the solution space (Appendix 5) it is a very convenient alternative.

3. Selected Machine Learning techniques for Continuous monitoring

As already introduced in Chapter 2, whenever the risk of catastrophic failure is very high, and a failure can cause costly damages to the machine, a permanent condition monitoring is desirable. In order to continuously performing such a monitoring, sensors should be integrated in the machinery and an autonomous, quick, on-line data processing is needed. Because of this, continuous monitoring is normally based on relatively simple parameters and its scope is limited to diagnose impending failure to give a warning in a short advance. Obviously, before integrating expensive sensors in an industrial machine, the system should be accurately designed and tested. In such design or research stage, short acquisitions are usually available. Nevertheless, if the training acquisition is long enough to represent the whole variability of the machine (both operational and environmental) the algorithms can be accurately calibrated and tested to generalize beyond the examples in the training set. In general, two kind of analysis are possible. The features in fact can be treated independently, with multiple univariate analyses, or altogether, considering the complex correlation structure with multivariate analyses. In any case, the analysis will be based on pattern recognition: an intelligence should be used to univocally put in relation the statistically-significant changes in the features to the presence of a damage, excluding possible confounding effects induced by operational or environmental variations.

Coherently with the idea of starting from simple but easily interpretable models, statistical learning is first tackled, to increase the complexity as going through the chapter and cross the frontier of machine leaning.

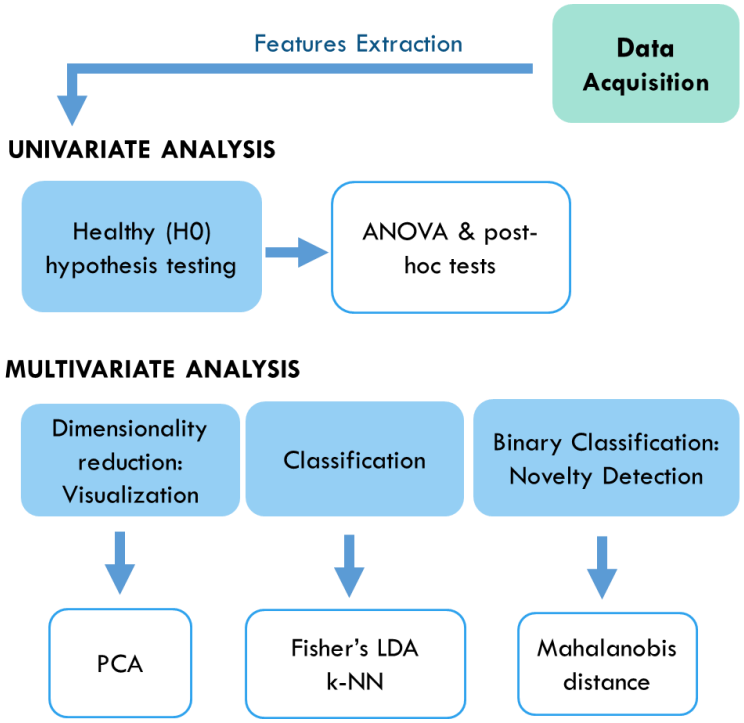


Figure 20: The proposed continuous monitoring methodology.

The statistical pattern recognition is tackled starting from the hypothesis testing philosophy, which can be interpreted from a binary (i.e. two-classes) unsupervised classification, healthy vs damaged. Hence, the univariate Analysis Of Variance (ANOVA)

together with the usual post hoc, multi-comparison tests is proposed to infer from the data the omnibus (i.e. *variance based*) null hypothesis of a detectable effect distinguishing the different groups, that is, for diagnostics purposes, the presence of damage (level 1 diagnostics).

Furthermore, in order to "condensate" the information contained in the different features, enhancing then the effect of damage, a multivariate analysis is preferable. In any case, this analysis is generally based on the selection of a subspace (often a 1-D projection) of the high dimensional dataset able to foster the damage detection. In particular, extending the idea of effect size to measure the distance of two distributions, a projection maximizing this measure can be found. This is the Fisher's Linear Discriminant Analysis (LDA) principle. On this 1-D projection, the univariate considerations about damage detection via hypothesis testing and classification can be applied.

Another way of analysing a multivariate dataset, typically for visualization purposes is the Principal Component Analysis. In this case, the original reference frame is rotated to match the directions explaining most of the variance. Then, the first can be selected to visualize the dataset on 2-D or 3-D plots. The dataset projected on the first component (explaining most of the variance) can also be used for classification. In any case, the use of a single component from PCA or LDA can be regarded as a lossy dimensionality reduction, as some portion of the dataset variability is neglected. In general, nothing ensures that the damage information is contained in the first component. On the contrary, in many cases, the first components can be proved to represent strong latent effects such as operational (e.g. speed, load, …) and environmental (e.g. temperature, humidity, …) variations. A lossless non-linear 1-D dimensionality reduction can be anyway found through Novelty Detection via Mahalanobis distance. In this regard, the information about novelty (i.e. deviation from normality) is additive on the feature space dimensions, so that the whole variability is preserved. Furthermore, it automatically accounts for a compensation of linear or quasi-linear hidden confounding effects.

Finally, the problems related to high dimensionality d (i.e. the curse of dimensionality) and to the selection of the sample size n are not secondary and must be tackled as well. In this respect, relevant considerations are proposed in Chapter 7.

3.1. The selected features

In Chapter 2, it was stated that the acquired data-set made of raw acceleration signals may contain relevant diagnostics information which is unfortunately ''hidden". In order to explore the data and foster the learning, some derived quantities called features are usually extracted, as they summarize the signal, allowing for a simplified and better interpretation. In general, many different features are commonly used in diagnostics, but these can be classified in two main categories: time-domain and frequency domain features. The frequency-domain features are known to be more stable with respect to changes in the overall machine configuration (e.g. speed, load, . . .), but have the drawback of being less general, requiring a deeper prior knowledge of the machine under analysis (geometry, dimensions, effective running speed. . .).

In the present methodology then, the most common time-series features are selected, coping with the need of speed and automation of the analysis. Root mean square, skewness, kurtosis, peak value and crest factor (peak/RMS) are then computed on chunks of the original acquisitions generating statistically significant samples of n observations for each of the channel-feature combination defining the whole feature

space. This numerousness n should be chosen in accordance to the considerations about the curse of dimensionality found in Chapter 7.

3.2. Statistics and probability: an introduction to hypothesis testing

Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation. The word, first introduced in English by Sir John Sinclair in 1829 and meaning "numerical data collected and classified", comes from the German *Statistik* introduced by Gottfried Achenwall (1749) to originally designate the analysis of data about the state (from the Italian *statista*, "statesman, politician" descending from the Latin *status* "position, place; condition" or figuratively "public order").

In particular, the word statistics can be interpreted as the "investigation of large numbers" or "theory of frequencies" [29]. This links statistics to **probability theory** according to von Mises's definition of the term. Indeed, despite in common language the word probability refers to the measure of the likelihood that an event will occur, from 0 i.e. impossible to 1 i.e. certain, the frequentist definition is much stricter:
"The probability is the limiting value of the relative frequency of a given attribute within a considered collective. The probabilities of all the attributes within the collective form its distribution."

The starting point of the probability theory is the concept of a collective (or population), an infinite sequence of observations, each consisting in the recording of a certain attribute. Then, the fundamental frequentist axiom follows. Selecting just n recordings (N.B. a new finite collective is formed by selection of a sample from the population), the relative frequency of an attribute, n_1/n, will approach a constant limiting value when n is increasing indefinitely. From this axiom, the Law of Large Numbers (LLN) can be derived. Actually, the LLN can be approached by alternative point of views, such as the Bernoulli-Poisson or the Bayes's. In any case, the idea of the strong law of large number is that the sample average $\overline{x_n} = \frac{1}{n}\sum x_i$ converges to a constant limiting value, i.e. the expected value μ, for an increasing $n \to \infty$. This can be further generalized to any statistical function (e.g. the median, the variance, etc), namely a function depending on the true frequency distribution but not on the order of the observations or their total number.

Focusing on the LLN applied to the mean, it is easy to get a proof involving the known, true statistical functions $E[x_i] = \mu$, $var[x_i] = \sigma^2$ and simple properties of expectation and variance:

$$E[\overline{x_n}] = \frac{1}{n}E\left[\sum x_i\right] = \frac{1}{n}\sum E[x_i] = \frac{n\mu}{n} = \mu$$
$$var[\overline{x_n}] = \frac{1}{n^2}var\left[\sum x_i\right] = \frac{1}{n^2}\sum var[x_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (28$$

So that, for $n \to \infty$, $var[\overline{x_n}] \to 0$, implying $\overline{x_n} \to \mu$.

And this is not all. It can be also proved that, for any generic sample distribution featuring finite statistical functions $E[x_i] = \mu$ and $var[x_i] = \sigma^2$, as n approaches infinity, the variable $\overline{x_n}$ asymptotically converges in distribution to a normal distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$. This corresponds the so-called Central Limit Theorem (CLT), name given in 1920 by the mathematician Polya to the Gauss's theory of errors derived by Laplace's exponential law.

That is, the overall error induced by the sum of many small elementary errors follows an exponential distribution, which takes the name of Gaussian "bell" curve.

Because of the nature of frequentist probability as a limiting value for $n \to \infty$, in practical cases, the true probability distribution and related statistical functions are not known a-priori but can be **inferred** from a sample i.e. extrapolated from the sample to the population, that is, considering for example the mean value, $\overline{x_n}$ can be used as an estimator of $\mu = \overline{x_\infty}$ at a given confidence or significance (i.e. the result is convincing up to some degree of trust).

Consider CLT. Taking $\overline{x_n}$ as estimator, the dispersion of its normally distributed values is given by $var[\overline{x_n}] = \frac{\sigma^2}{n}$. Then, the true expected value μ falls in an interval $\left(\overline{x_n} - \frac{\sigma}{\sqrt{n}}, \overline{x_n} + \frac{\sigma}{\sqrt{n}}\right)$ at a confidence $1 - \alpha = 68\%$ or significance $\alpha = 32\%$. Standardizing the estimator distribution to get $z_n = \frac{\overline{x_n} - \mu}{\sigma/\sqrt{n}}$, it is possible to derive the results in Table 1, which holds for any Gaussian distribution [30].

Table 1: Confidence level and corresponding interval [30]

Standard interval	Inside to outside ratio	Confidence $1 - \alpha$
± 0.6745	$1\ to\ 1$	50%
± 1	$2{,}15\ to\ 1$	$68{,}3\%$
± 2	$21\ to\ 1$	$95{,}5\%$
± 3	$369\ to\ 1$	$99{,}7\%$

Generalizing, a **critical value** for the given confidence can always be found to form a **confidence interval** such that:

$$-N_{(0,1),\frac{\alpha}{2}} \leq z_n \leq N_{(0,1),\frac{\alpha}{2}}$$
$$\overline{x_n} - \frac{\sigma}{\sqrt{n}} N_{(0,1),\frac{\alpha}{2}} \leq \mu \leq \overline{x_n} + \frac{\sigma}{\sqrt{n}} N_{(0,1),\frac{\alpha}{2}} \qquad (29$$

Some definitions are needed. A confidence interval (CI) is a type of interval estimate giving a range of values in which the true, unknown population parameter will fall at chosen probability rate (the confidence, $1 - \alpha$). The critical values which limit the interval are the values exceeded only $100\frac{\alpha}{2}$ times in a hundred and are given by $N_{(0,1),\frac{\alpha}{2}}$.

Unfortunately, also the population variance σ^2 in most of cases is unknown. When n is large enough anyway, the variance can be estimated from the sample as well as for the mean.

The maximum-likelihood (ML) estimate of the sample variance is given by

$$s^2_{n,ML} = \frac{1}{n}\sum (x_i - \overline{x_n})^2 = \frac{1}{n} S^2_n \qquad (30$$

Focusing on the LLN applied to $s^2_{n,ML}$, it is easy to get a proof of

$$E\left[s^2_{n,ML}\right] = \frac{n-1}{n}\sigma^2 \qquad (31$$

The $s^2_{n,ML}$ is then a biased estimator of the population variance. Hence, it is usually corrected to its unbiased form:

$$s_n^2 = \frac{1}{n-1} S_n^2 \qquad (32$$

Furthermore, under the hypothesis of Gaussianity ($x_i \sim N_{(0,1)}$), Cochran's theorem shows that

$$\frac{S_n^2}{\sigma^2} = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1} \qquad (33$$

where χ^2_{n-1} is a Chi-square distribution with $n-1$ degrees of freedom. This leads to

$$E[s_n^2] = E\left[\frac{\sigma^2}{n-1}\chi^2_{n-1}\right] = \frac{\sigma^2}{n-1} E[\chi^2_{n-1}] = \sigma^2$$
$$var[s_n^2] = var\left[\frac{\sigma^2}{n-1}\chi^2_{n-1}\right] = \frac{\sigma^4}{(n-1)^2} var[\chi^2_{n-1}] = \frac{2\sigma^4}{n-1} \qquad (34$$

Hence, asymmetric confidence intervals can be built

$$\frac{(n-1)s_n^2}{\chi^2_{(n-1),\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi^2_{(n-1),1-\frac{\alpha}{2}}} \qquad (35$$

If the sample variance is used in place of the true variance, the formula given for the confidence interval of the mean holds just for large n.

$$\overline{x_n} - \frac{s_n}{\sqrt{n}} N_{(0,1),\frac{\alpha}{2}} \leq \mu \leq \overline{x_n} + \frac{s_n}{\sqrt{n}} N_{(0,1),\frac{\alpha}{2}} \qquad (36$$

Otherwise, if the sample size is small, the estimates cannot be considered independent from the observations and it can be proved that

$$z_n^* = \frac{\overline{x_n} - \mu}{s_n/\sqrt{n}} \sim t_{(n-1)} \qquad (37$$

where t_{n-1} is a Student's t distribution with $n-1$ degrees of freedom. Hence, a correction of the confidence interval for the mean follows:

$$\overline{x_n} - \frac{s_n}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}} \leq \mu \leq \overline{x_n} + \frac{s_n}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}} \qquad (38$$

Notice that, as shown in Figure 21, for $n \to \infty$, $t_{n-1} \to N_{(0,1)}$ and $s_n \to \sigma$, so that this CI tends to the formula for σ known as n increases. N.B. The critical values are found inverting the CDF, that is reading the abscissa for a given probability (ordinate). Hence, even if the curves seem very near in Figure 21, at the tails the difference becomes much more important, leading to very different critical values, in particular for small *dof*-s.

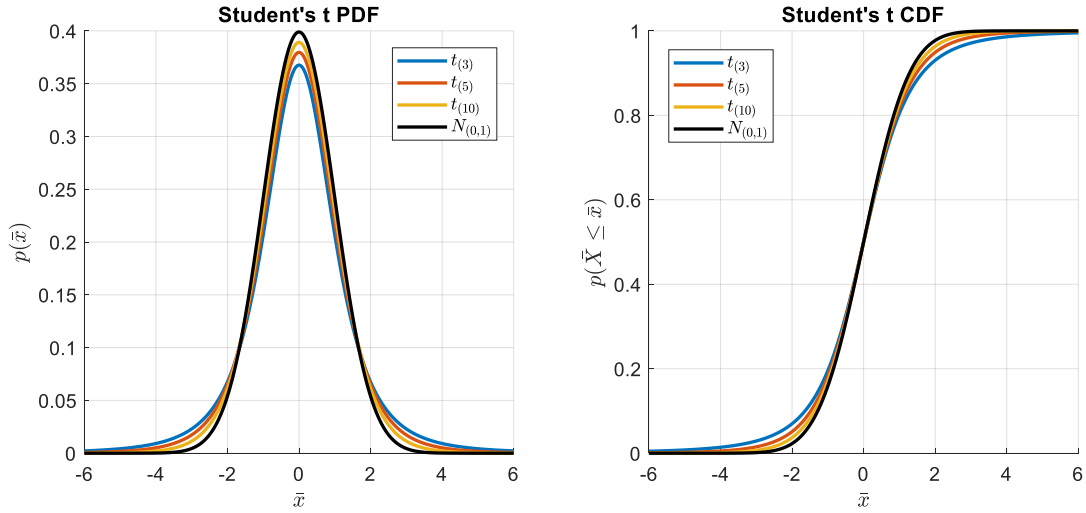


Figure 21: Student's t distribution for increasing degrees of freedom. Notice the heavier tails than a normal, which is the limiting distribution for $dof \to \infty$.

This demonstrates how **inferential statistics** can extrapolate information of the population from a sample at a confidence depending on the size of the sample. Anyway, the **estimation theory** is not the only subject of inferential statistics. **Hypothesis testing** is also a fundamental.

A statistical hypothesis test is a method of statistical inference that is meant to compare two statistical samples, or a sample against a model. A hypothesis is proposed for the statistical relationship among the two and this is compared to an alternative hypothesis. Originally, the hypothesis to be tested was that of no relationship, taking then the name of null hypothesis. The comparison is deemed statistically significant if the observed relationship can be proved to be an unlikely realization of the null hypothesis according to a threshold probability (i.e. the confidence). This is strictly related to the idea of confidence interval. The formula derived in this section, for example, can be used to test a single population mean. Also, outlier detection can be performed following the same logic.

3.2.1. Hypothesis testing of a single population mean

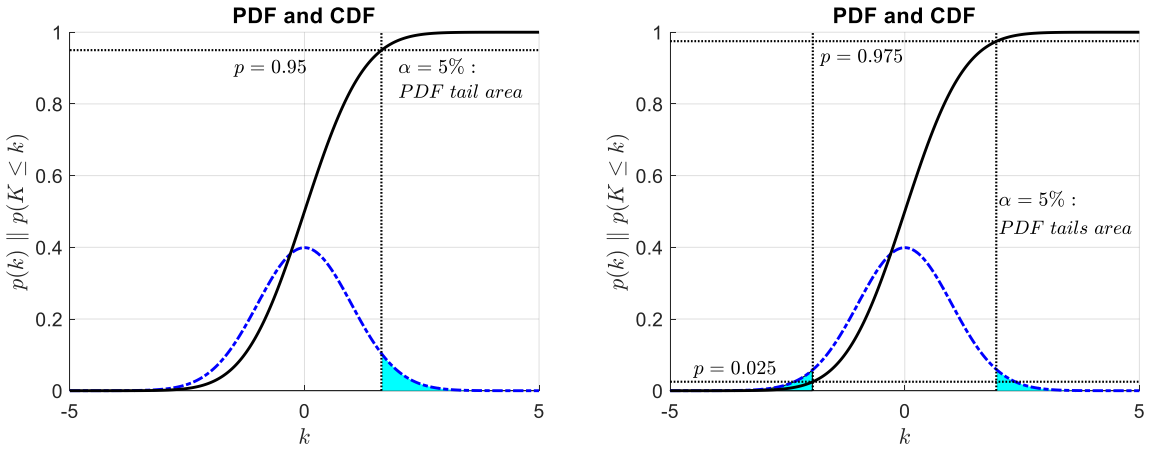
A null hypothesis about a population mean such as $H_0: \mu = \mu_0$ can be tested against an alternative hypothesis $H_a: \mu \neq \mu_0$. Statistical summaries can be found when the sample comes from given distributions:

Distribution of the population:	*Statistical summary k from the sample:*	
• A normal distribution with a given variance or a generic distribution (also non-normal) assuming $n > 30$, thanks to CLT	$k = z = \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}} \sim N_{(0,1)}$	*(39*
• A normal distribution with unknown variance	$k = t = \frac{\overline{x_n} - \mu_0}{s_n/\sqrt{n}} \sim t_{(n-1)}$	*(40*

Despite the hypothesis testing can be tackled through confidence intervals via the computation of a critical value, just as performed in the previous section, it is far more common to compute the so-called **p-value**. The p-value (i.e. probability value or

asymptotic significance) is the probability that, given H_0, the statistical summary would be more extreme than the actual observed results. Hence, if this p-value is less than or equal to a selected significance level α, the hypothesis is rejected in favour of the H_a implying no relationship. Depending on the point of view, "more extreme than" can take different meanings:

Tails:	*p-value:*	
• For a right tail event, it can be stated as	$\Pr(K \geq k \mid H_0)$	*(41*
• For a left tail event, it is	$\Pr(K \leq k \mid H_0)$	*(42*
• For a double tail event (on a symmetric distribution), it becomes	$2\min(\Pr(K \geq k \mid H_0), \Pr(K \leq k \mid H_0))$	*(43*



One tail Two tails

Figure 22: One or two tail hypothesis testing principle: α critical values and confidence intervals.

3.2.2. Hypothesis testing of outliers

In experimental statistics, samples from a population are usually drawn through measurements. While performing such an operation, discordant values can often be recorded. These inconsistent values are usually called **outliers** and they may indicate either measurement error or that the population has a distribution different from the believed one (e.g. heavier tails). The judgement of discordancy can be seen then as a hypothesis test on the single suspected of being an outlier.

Consider for example the Chauvenet's criterion. The idea behind this method of assessing outliers is to find a confidence interval that should reasonably contain all n values of a sample. Hence, under the assumption of a normal population $x \sim N(\mu, \sigma)$, it follows that a believed outlier x_o shows a statistical summary $z = \frac{x_o - \mu}{\sigma}$ which can be compared to a corresponding critical value $N_{(0,1),\frac{\alpha}{2}}$ for a significance $\alpha = \frac{1}{n}$. In other words, the value which is exceeded just once every n values is used as critical, so that:

$$\mu - \sigma\, N_{(0,1),\frac{1}{2n}} \leq x_o \leq \mu + \sigma\, N_{(0,1),\frac{1}{2n}} \qquad (44$$

The other way around, comparing $2\min(\Pr(Z \geq z|H_0), \Pr(Z \leq z|H_0))$, namely the p-value of the summary $z = \frac{x_o - \mu}{\sigma}$, to the significance $\alpha = \frac{1}{n}$, an equivalent test for the hypothesis $H_0: x_o \; is \; not \; an \; outlier$ can be found. If the p-value is less than or equal to the significance, H_0 is rejected. Obviously, substituting μ and σ with their sample estimates, the same formula can be used under the assumption of large n. Unfortunately, in many cases, this assumption does not hold, so that a compensation is needed.

Consider the squared Mahalanobis distance $(x-\mu)'\Sigma^{-1}(x-\mu)$ (see chapter 6 for further considerations). For a dimension 1 it degenerates to $\frac{(x-\mu)^2}{\sigma^2} = Z^2$, which is exactly the square of the original statistical summary. In the ideal case of μ and σ known, in can be proved that this distance is distributed as a χ^2 [33], so that $Z^2 \sim \chi_1^2$. The probability of exceeding $|z|$ or z^2 in any case must be equal. Hence, $\Pr(Z^2 \geq z^2) = \Pr(|Z| \geq |z|)$ must hold, giving an equivalent single tail criterion for hypothesis testing. The p-value

$$2\min(\Pr(Z \geq z|H_0), \Pr(Z \leq z|H_0)) \equiv \Pr(Z^2 \geq z^2|H_0) \qquad (45$$

can be compared to the significance $\alpha = \frac{1}{n}$. Analogously, a bounding critical value for the statistical summary can be found:

$$z^2 \leq \chi^2_{(1),\frac{1}{n}} \qquad (46$$

Thanks to [33], it is easy to correct this formulation for the use of the sample estimates $\overline{x_n}$ and s_n which are not independent from the observations for n small. Hence, the so-called Wilks's critical value is given

$$z^2 = \frac{(x-\overline{x_n})^2}{s_n^2} \leq \frac{(n-1)^2 F_{(1,n-2),\frac{1}{n^2}}}{n\left(n-2+F_{(1,n-2),\frac{1}{n^2}}\right)} \qquad (47$$

where $F_{(1,n-2)}$ is the Fisher–Snedecor distribution with degrees of freedom 1 and $n-2$. Note that $F_{(1,n-2),\frac{1}{n^2}} = \left(t_{(n-2),\sqrt{\frac{1}{n^2}}}\right)^2 \approx \left(N_{(0,1),\sqrt{\frac{1}{n^2}}}\right)^2$ for n large.

The relevant consideration is that, given the multivariate nature of the Mahalanobis distance, this can be used for testing the presence of one outlier also in a multivariate normal distribution of dimension d. The corresponding bounding critical value for the squared Mahalanobis distance is given by:

$$(x_o - \bar{x})'S^{-1}(x_o - \bar{x}) \leq \frac{d(n-1)^2 F_{(d,n-d-1),\frac{\alpha}{n}}}{n\left(n-d-1+d\,F_{(d,n-d-1),\frac{\alpha}{n}}\right)} \qquad (48$$

3.2.3. Hypothesis testing of the difference between two population means

A two-sample location test of the null hypothesis such that the means of two populations are equal, namely $H_0: \mu_1 = \mu_2$, can be tested against an alternative hypothesis $H_a: \mu_1 \neq \mu_2$ when it can be assumed that the two distributions have the same variance (i.e. homoscedasticity) and the samples come from:

Distribution of the population:	*Statistical summary k from the sample:*	
• Normal distributions with given variance or Generic distributions (also non-normal) assuming $n > 30$, thanks to CLT	$k = z = \frac{E[x_1] - E[x_2]}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N_{(0,1)}$	*(49*
• Normal distributions with unknown variance	$k = t = \frac{E[x_1] - E[x_2]}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \sim t_{(n_1+n_2-2)}$	*(50*

where s_p is a pooled estimate of the unknown variance of the two samples ($i = 1,2$):

Biased estimate	*Unbiased estimate*	
$s_{p,B}^2 = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i n_i}$	$s_p^2 = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i (n_i - 1)}$	*(51*

The analysis of the differences among multiple group means take the name of ANalysis Of VAriance (ANOVA), and is the subject of the following Section 3.3.

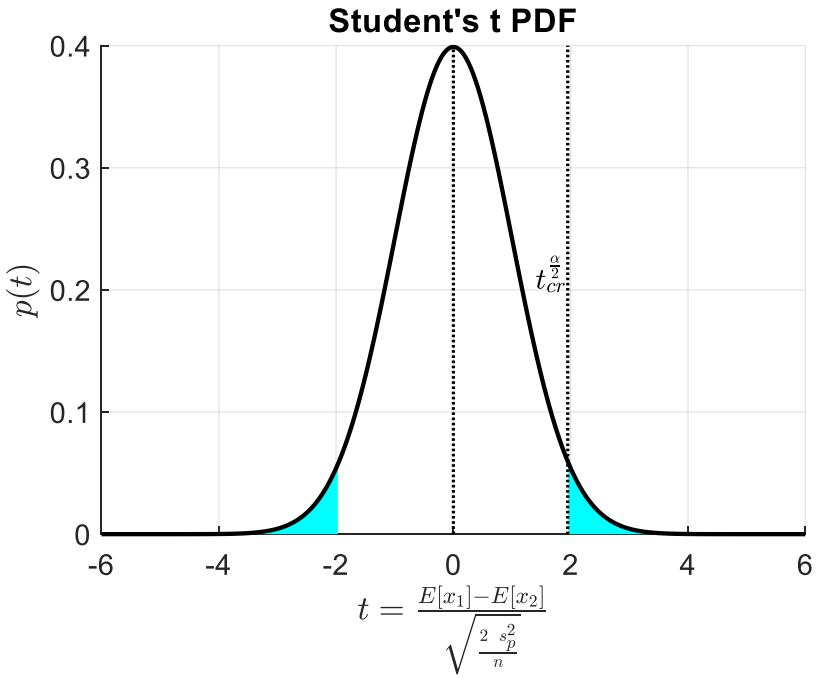


Figure 23: Hypothesis testing of the difference between two population means – graphical representation and critical values highlighted.

3.2.4. Diagnostics, hypothesis testing and errors

So far, hypothesis testing was tackled just from an abstract mathematical perspective. Two kind of null hypotheses were formulated. The first regards the comparison of two population means. The second compares a single data point to a reference population. Both these general points of views can be used in the field of diagnostics. Adopting the null hypothesis H_0: $the\ machine\ is\ healthy$ both the tests can be considered. In the first case, a new sample from the machine under investigation is compared to a reference healthy sample acquired in a calibration stage (known as healthy), implying then H_0: $\mu_{new} = \mu_{ref}$. In the second, H_0: $x_o\ is\ not\ an\ outlier$ is used to test if a discordant measure can be considered "novel", implying that it is believed to be generated by an alternate mechanism such as a damage (when all other possible influences are excluded). These two approaches can be both linked to statistical classification and are then fundamental to perform Level 1 diagnostics.

Classification refers to the problem of identifying to which category (in the considered case just two options, healthy vs non-healthy) a new observation (i.e. a point in the feature space) belongs, based on a training data set taken as reference. This set is "labelled", as the data points are known (or at least believed) to come from a given health condition. Classification is always a two-step procedure:

a) In the **training** phase, the labelled samples are used to build a classifier, namely a function which divides the feature (variable) space in groups. This separation is then found in terms of distributions. When a single feature is used to investigate the machine, the classifiers correspond to the selection of a threshold. It is relevant to point out that this feature-space partitioning can also be obtained in an unsupervised way (i.e. without exploiting the labels). This takes the name of clustering.
b) Just in a second phase the new observations are assigned to the corresponding class (i.e. **classified**) according to the classifier function. Each new unlabelled data point is now treated individually.

Typically, a **validation** phase is added among these two to assess the performances of the classifier function "out of sample", namely on data points different from the ones used for the training.

According to these considerations, hypothesis testing is closely linked to classification. Nevertheless, classification implies the knowledge (or at least the belief) that the different samples are NOT coming from the same distribution, so that the alternative hypothesis takes much more relevance.

Furthermore, an additional step is needed to fully understand hypothesis testing. Imagine the case in which a difference among the means is present (i.e. H_a is true). If H_0 is rejected, it means that the two population averages are discriminable. Obviously, if the difference is small, a huge sample size n is needed to detect the difference at a significance α. In fact, for an increasing n, the "resolution" of the test (i.e. the minimum significant distance between two means according to which H_0 is rejected) can be reduced at will. Consider the confidence interval for testing H_0 in case of equal sample sizes:

$$|t - t_0^*| = \left|\frac{E[x_1] - E[x_2] - (\mu_1 - \mu_2)_0}{\sqrt{2s_p^2/n}}\right| \leq t_{(2n-2),\frac{\alpha}{2}} = t_{cr}^{\frac{\alpha}{2}} \qquad (52$$

$$|d - d_0^*| = |t - t_0^*|\sqrt{\frac{2}{n}} = \left|\frac{E[x_1] - E[x_2] - (\mu_1 - \mu_2)_0}{s_p}\right| \leq \sqrt{\frac{2}{n}} t_{(2n-2),\frac{\alpha}{2}} = d_{cr}^{\frac{\alpha}{2},n} \quad (53$$

$$|D - D_0^*| = |E[x_1] - E[x_2] - (\mu_1 - \mu_2)_0| \leq s_p\sqrt{\frac{2}{n}} t_{(2n-2),\frac{\alpha}{2}} = D_{cr}^{\frac{\alpha}{2},n,s_p} \quad (54$$

where the distance $(\mu_1 - \mu_2)_0$ is null under the null hypothesis.

Testing H_0 when H_a is actually true then can lead to:

$H_0\ accepted$	$\left\|\frac{E[x_1] - E[x_2]}{\sqrt{2s_p^2/n}}\right\| = \|t\|_{H_a}\| \leq t_{cr}^{\frac{\alpha}{2}}$	Probability: β
$H_0\ rejected$	$\left\|\frac{E[x_1] - E[x_2]}{\sqrt{2s_p^2/n}}\right\| = \|t\|_{H_a}\| > t_{cr}^{\frac{\alpha}{2}}$	Probability: $1 - \beta$

(55

Focusing on the distribution of $t|_{H_a}$, this will be centred on $t^* = \frac{\mu_1 - \mu_2}{\sqrt{2\sigma^2/n}}$, from which it is easy to get the value of $d^* = \frac{\mu_1 - \mu_2}{\sigma} = \frac{D^*}{\sigma} = t^*\sqrt{2/n}$, the so-called **effect size**, while $1 - \beta$ is commonly identified as the **power** of the test.

In any case, the test does not consider at all whether $D_{cr}^{\frac{\alpha}{2},n,\sigma}$, namely the minimum resolved distance, is physically meaningful or not, as this consideration also depends on the original populations' variance and on the numerousness of the sample size. Furthermore, no information about the probability of resolving a given d^* (i.e. the power) is taken into account by the test itself. Such considerations should come prior to the test, at a Design of Experiment (DOE) stage.

The power of a two population means test is visualized in Figure 24-above for a particular t^*, and generalized for any t^* in Figure 24-below. This second curve can be obtained by shifting the Cumulative Distribution Function (CDF) of the $t_{(2n-2)}$. For $n > 30$ the Student's t distribution is practically equal to a standard normal, whose CDF is used in this case to obtain the graph of Figure 24-below (which holds even for smaller n if the known variance σ substitutes s_p).

Considering that $\alpha = 5\%$ is probably the most common value and is rarely changed, this graph can be used at a stage of design of the experiment to evaluate the optimal n able to resolve an expected effect size at a probability $1 - \beta$ larger than a selected value (e.g. 0.8). This d^* can be approximated from prior research as $d = \left|\frac{E[x_{dam}] - E[x_{ref}]}{s_p}\right|$, or through conventions such as the one proposed by Cohen [31,32] and here reported:

Effect size	d
Small	0,2
Medium	0,5
Large	0,8

For example, for a power $1 - \beta = 0{,}8$, the graph in Figure 24-below gives $t^* = \left|\frac{(\mu_1 - \mu_2)_a}{\sqrt{2\sigma^2/n}}\right| = 2{,}8$ which implies at least $n = 2\left(\frac{2{,}8}{d}\right)^2$. For detecting a large effect size then,

$n \cong 25$ is enough, but the required n increases to 63 for a medium and to 392 for a small effect size. As n is obviously limited by physical constraints, a trade-off between confidence $1-\alpha$ and power $1-\beta$ is then necessary to control both the $type\ I$ and II error rates.

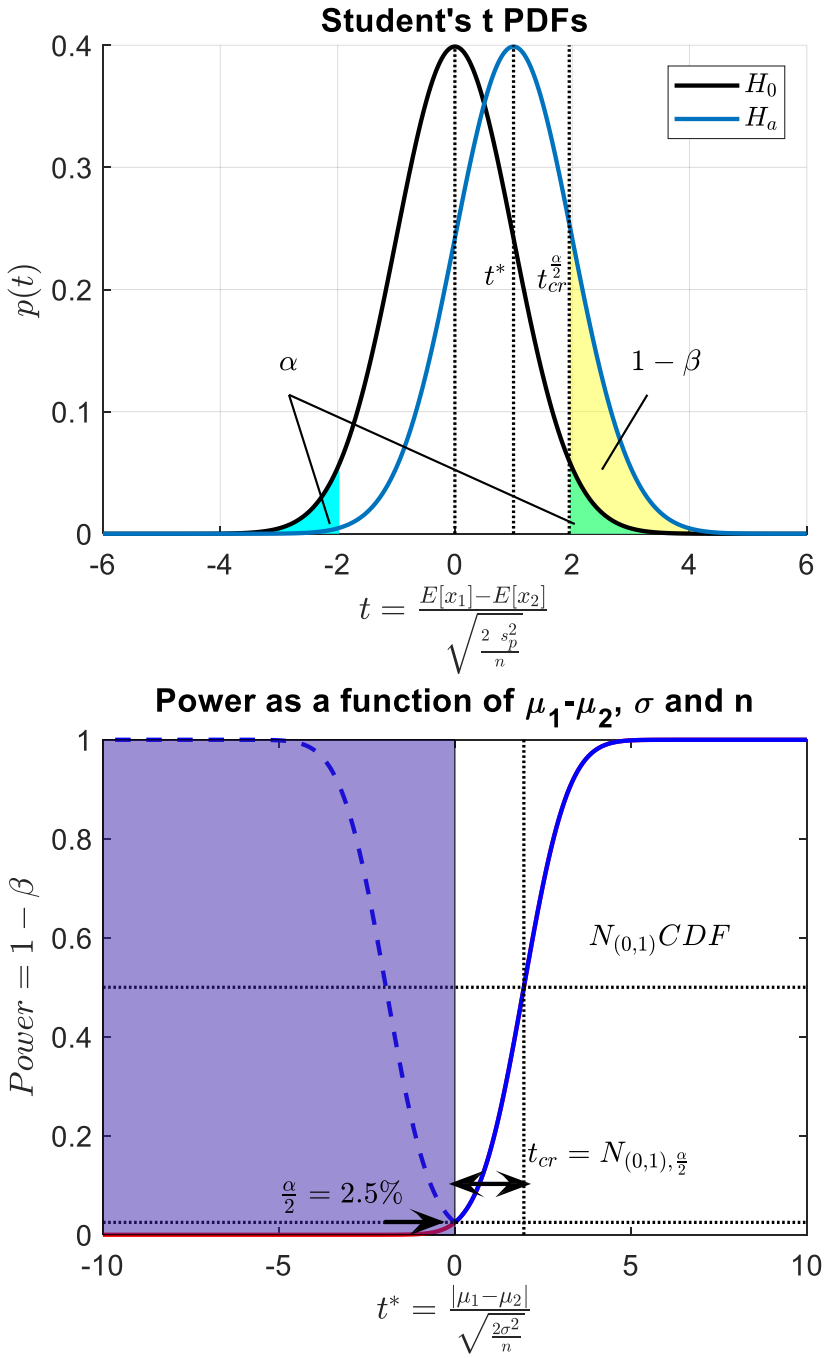


Figure 24: The power of a two population means test – above, a visualization of the significance α and of the power $1-\beta$ for a particular case – below, the power (under assumption of normality) as a function of t^* which depends on the effect size $\frac{\mu_1-\mu_2}{\sigma} = d$ and the sample size n.

From a diagnostics point of view, in fact, the confidence associated to the test implies a $type\ I\ error\ rate$ equivalent to the significance α, which corresponds to the probability of rejecting a true H_0. This must be as small as possible as a too high number of triggered False Alarms (FA) can erode the confidence of the damage detection. At the same time, also the $type\ II\ error\ rate$ should be kept under control. This is the

probability of failing to reject a false H_0, usually referred to as β, the complementary of the power of the test. This value corresponds to a missed indication of damage although present (Missed Alarm MA) and is very detrimental as can bring serious economic and life-safety implications. These error rates are usually collected in "confusion matrices", tables such as:

		True Health Condition:	
		Healthy (H_0)	*Damaged*
CBM Actions	*accept H_0: Healthy*	No Alarm	**Missed Alarm *type II error***
	reject H_0: Damaged	**False Alarm *type I error***	Alarm

Figure 25: Type I and II errors in hypothesis testing for CBM: confusion matrix

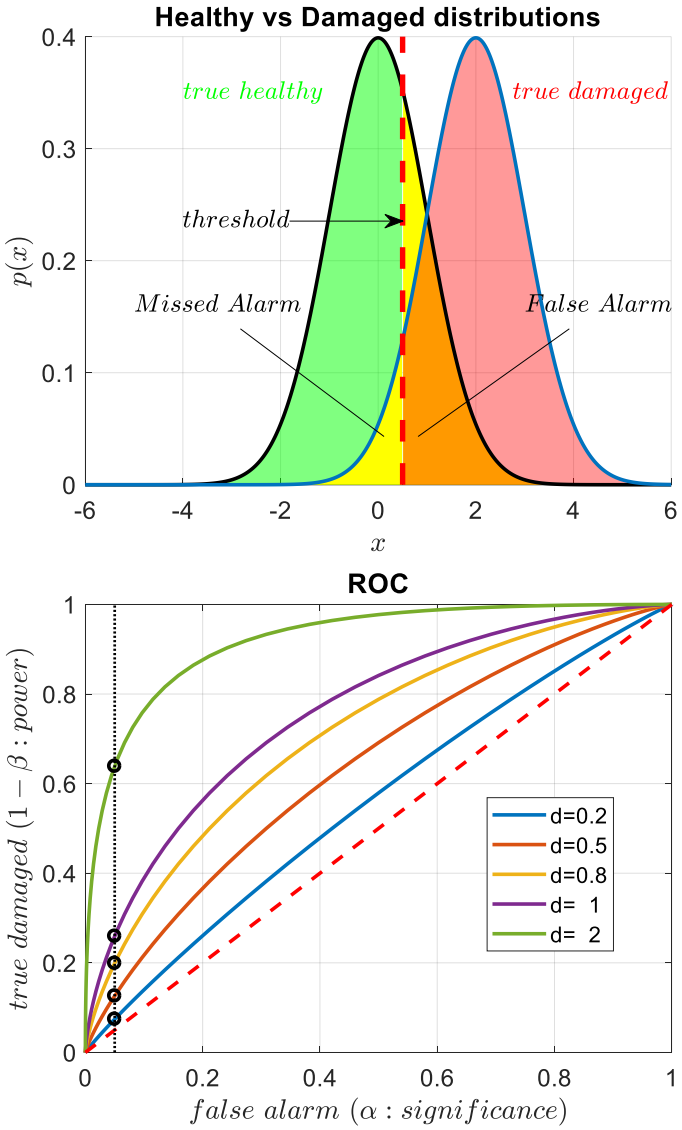


Figure 26: Receiver Operating Characteristic (ROC) function of the threshold (Gaussian distributions).Above- graphical summary of the table of type I and type II errors given in Figure 25. Below- ROC for binary classification with different effect sizes d^* and the position of the 95% critical value (black dotted). For $d = 0{,}2$ the performance is very poor as the ROC is near the 1st-3rd quadrant bisector (random classifier).

On the contrary, in the field of Operational Research (OR), a discipline that deals with the application of analytical methods for making better decisions, the Receiver Operating Characteristic (ROC) is usually preferred for assessing the diagnostic ability of a binary classifier as its discrimination threshold is varied.

Interpreting the critical value as a threshold allows to understand how this can be varied to find the best compromise between α and β and to assess the overall performance of the test or of the classification. Figure 6(b) summarizes the true damaged rate (the power $1-\beta$) as a function of the false alarm rate (the significance α) for some relevant effect sizes while the threshold takes all the possible values. The threshold corresponding to the $\alpha = 5\%$ critical value is highlighted. In general, anyway, the farthest is the ROC curve from the 1st – 3rd quadrant bisector, the better the classification, which obviously improves as the effect size is enlarged.

Summarizing, in the light of these considerations, classification comes from hypothesis testing, so that classifiers "inspired" by corresponding tests of hypothesis can be found. It is the case of the Fisher's LDA, which comes from the test for two population means and from its extension to multiple populations, the Analysis Of Variance (ANOVA). These will be treated in the next sections, together with the Novelty Detection, a form of two-groups classification trained only on the reference condition data (i.e. healthy) but not on the damaged, which can be put in relation to the hypothesis testing of outliers (paragraph 2.2). Indeed, remembering that a univariate classifier function degenerates to a constant (i.e. a threshold), this can be again derived as a critical value.

In conclusion, the classification which is widely used in diagnostics is basically a form of hypothesis testing, so that it is subject to the same general rules:

- A classifier (i.e. a threshold in the 1-D case) must be trained so as to find the best compromise between type I and type II errors. In the design of the diagnostic system, a study for selecting the optimal numerousness of the samples is fundamental.
- In any case, the performances are not only depending on the classifier itself, but also on the populations' distributions under investigation. The effect size, that is the normalized distance of the two populations, changes when different variables are considered, so that a wise feature selection is fundamental.

3.3. The univariate Analysis of Variance: one-way ANOVA

The one-way (i.e. univariate) ANOVA is a very common hypothesis testing method in experimental statistics. In its typical formulation, ANOVA is meant to test the omnibus (i.e. variance based) null hypothesis that all the G groups are random samples from the same population. As it is easy to notice, $H_0: \mu_j = \mu \; \forall j \in 1:G$ is then a generalization of the test for two population means introduced in chapter 2.3.

Mathematically, ANOVA assumes a linear model:

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N_j(0, \sigma) \qquad (56$$

according to which, an observation of the j-th group is a random draw from a normal distribution $N(\mu_j, \sigma)$, where the variance σ^2 is assumed the same for all the groups. Three hypotheses are made:

- Normality
- Homoscedasticity (homogeneity of variances)
- Independence of observations

If these assumptions hold, the overall sum of squares SS_t can be decomposed into a *within groups part* SS_{wg} , and a *between groups parts* SS_{bg}. These sums are related to two different estimates of the common population variance σ^2:

- the *within groups variance* s_{wg}^2, which is an average of the estimated s_j^2 weighted according to the numerousness n_j of each group,
- the *between groups variance* s_{bg}^2, which comes from the average squared deviation of the groups means μ_j from the overall mean $\overline{y}$. The variance of the means, as introduced in previous chapters, actually corresponds to an estimate of σ^2/n.

In any case, the sums of squares are known to be distributed as χ^2. The ratio of the two sums of squares over their degrees of freedom is then distributed as a Fisher–Snedecor F and can be used as a statistical summary for verifying the hypothesis. Hence, for an equal numerousness of the groups $n_j = n\ \forall j = 1{:}\,G$ and $N = nG$:

$$\begin{array}{|c|c|c|}\hline SS_{bg} = n\sum_{j=1}^{G}(\mu_j - \mu)^2 & SS_{bg} \sim \chi^2_{(G-1)} & f = \dfrac{SS_{bg}/_{G-1}}{SS_{wg}/_{N-G}} = \dfrac{MS_{bg}}{MS_{wg}} \\ \hline SS_{wg} = \sum_{j=1}^{G}\sum_{i=1}^{n}(y_{ij} - \mu_j)^2 & SS_{wg} \sim \chi^2_{(N-G)} & = \dfrac{n\,s_{bg}^2}{s_{wg}^2} \sim F_{(G-1,N-G)} \\ \hline \end{array} \quad (57$$

where μ_j is the average of the elements in each group j while μ is the total mean.
The summary f can be finally compared to the critical value $F_{(G-1,N-G),\alpha}$. If f is more extreme, the null hypothesis is rejected, highlighting a difference in the group means.
This can be tackled also in terms of p-values, so that, if $\Pr(F \geq f|H_0)$ is lower than the significance α (typically 5%), H_0 is rejected. This is summarized in Figure 27.

Unfortunately, in case of H_0 rejection, ANOVA is not able to provide additional information about which population average differs the most and from which one of the others. Multiple two-sample tests (ANOVA reduces to the Student's t-test of paragraph 2.3 in this case) could be performed. In this regard, the Fisher's LSD multi-comparison and the Bonferroni correction are proposed in the next paragraph.

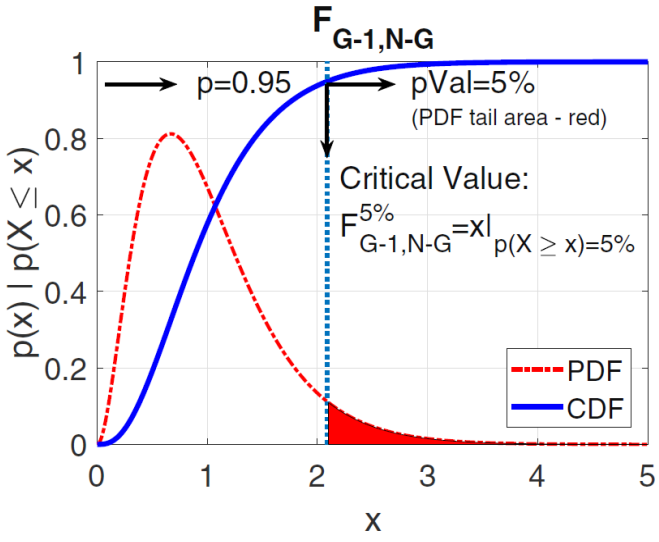


Figure 27: Fisher-Snedecor's $F_{(G-1,N-G)}$ critical value (≈ 2.1) for a p-value of 5% - one sided test.

3.4. Multi-comparison post-hoc test: Fisher's Least Significant Difference (LSD) and Bonferroni correction

The Fisher's Least Significant Difference (LSD) is basically a generalization of the Student's t-test of Section 3.2.1. The two are also related from the fact that the Fisher's F CDF reduces to the Student's t CDF when only two groups at a time are compared.
Multiple two means tests can be then conducted following the procedure in Section 2.3. According to the formula, considering equal samples size n, it is easy to get:

$$t = \frac{E[x_1] - E[x_2]}{\sqrt{2\ s_p^2/n}} \sim t_{(2n-2)} \qquad (58$$

from which a confidence interval based on a significance α can be found. The so-called LSD is defined:

$$LSD = |E[x_1] - E[x_2]| \leq t_{(2n-2),\frac{\alpha}{2}} \sqrt{\frac{2\ s_{wg}^2}{n}} \qquad (59$$

When multiple hypotheses are tested anyway, the chance of a rare event increases and the type-I error rate rises as well. To compensate for this, Bonferroni proposed to correct the formula by decreasing the significance from α to α/m, where $m = \frac{G(G-1)}{2}$ is the number of possible pairwise combinations of G groups. In any case, a confidence interval $\pm {LSD}/{2}$ can be built around the healthy distribution mean: intersecating groups will be considered not significantly distant, meaning that it will be hard to recognize them with enough confidence.

3.5. Multivariate data classification: Fisher's Linear Discriminant Analysis

As introduced in Section 3.2.4, the parameter which characterizes the distance between two distributions is the effect size $d^* = \frac{\mu_1 - \mu_2}{\sigma}$, which can be estimated from the samples as $d = \frac{E[x_{dam}] - E[x_{ref}]}{s_p}$. Fisher found a simple way to use this distance squared as a measure of separation also in case of multivariate problems, creating the Liner Discriminant Analysis (LDA). In short, collecting the multivariate features in the rows of a matrix X, LDA searches for optimal linear dimensionality reduction $y = w'X$, namely the direction w which maximizes the difference between the projected class-means distance, normalized by a measure of the within-class variance (also called "scatter") along the same direction. The measure of separation is then the squared effect size, also resulting as the ratio s_{bg}^2/s_{wg}^2. The formulation for the separation measure $J(w)$ in a multivariate feature space for 2 groups under homoscedasticity assumption is given by:

Between class scatter matrix:	$J(w) = \frac{w' S_b w}{w' S_w w}$
$S_b = (\mu_2 - \mu_1)'(\mu_2 - \mu_1)$	
Within class scatter matrix:	$\arg\max_w J(w)$:
$S_w = \sum_{h \in C_1} (y^h - \mu_1)'(y^h - \mu_1) + \sum_{k \in C_1} (y^k - \mu_2)'(y^k - \mu_2)$	$w \propto S_w^{-1}(\mu_2 - \mu_1)'$

(60

Extending it to multiple classes, it's possible to prove that, when w is an eigenvector of $S_w^{-1}S_b$, the corresponding eigenvalue λ can be interpreted as a measure of the separation of the classes. The eigenvalue w corresponding to the largest λ, usually referred to as the first principal component of the matrix $S_w^{-1}S_b$, is then the direction of maximum separation [34].

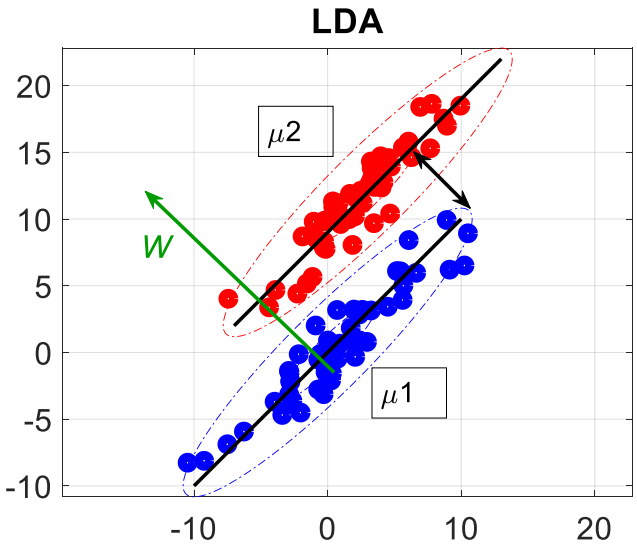


Figure 28: Visualization of the LDA idea for a 2D, 2 groups case.

Once this maximum separation direction w is found, a projection of the observations on this direction (i.e. a linear combination of the features) is performed, and the classification decision can be taken using as a threshold the average position of the projection of the two means on this single dimension. This is equivalent of finding a hyperplane able to separate the different groups in the multivariate feature space.

This algorithm, although very interesting from a theoretical point of view, is not commonly used in most of practical cases, as it expects a linear separation among the classes, and this does not happen often. More refined algorithms based on complex separation surfaces can be found in the literature. A very simple non-parametric one which deserves to be considered is the k-nearest-neighbours algorithm (k-NN). In this case, a new observation is classified taking the k closest observations (usually according to Eulerian distance) in its neighbourhood and searching for the class appearing most frequently. Obviously, the value of the free parameter k must be optimized on the particular application. With a similar reasoning, also unsupervised classification (i.e. clustering) can be performed. Randomly selecting k centroids and separating the dataset according to the distance of each point from these mean values, clusters can be found. This grouping can be refined by repeating the classification using as centroids the means of the clusters found in previous iteration. This is the so-called k-means clustering.

Anyway, when condition monitoring is chosen in real-life industrial applications, a training phase can be easily programmed, so that it is very unlikely that clustering will be needed. On the contrary, it is never advisable to let expensive machines run in damaged conditions, not even in a training phase. A training based on healthy acquisitions alone is then the most likely in industrial condition monitoring. Because of this, Novelty Detection will be then introduced in next Section 3.7.

3.6. Multivariate dimensionality reduction for data visualization: Principal Component Analysis

The PCA is a technique that uses an orthogonal space transformation to convert a set of correlated quantities into uncorrelated variables called principal components. This transformation is basically a rotation of the feature space in such a way that the first principal component will explain the largest possible variance, while each succeeding component will show the highest possible variance under the constraint of orthogonality with the preceding ones. This is usually accomplished by eigenvalue decomposition of the data covariance matrix or singular value decomposition of the data matrix after mean centring [35]. PCA is sensitive to the relative scaling of the original variables, so a data normalization is often advisable. Alternatively, the data correlation matrix can be used.

In general, the main application of PCA is for reducing a complex data set to a lower dimension using the first few components that explain the majority of the variation. The dimensionality reduction is then commonly used to obtain easily visualizable 2D or 3D projections of multivariate datasets. This can also reveal sometimes hidden, simplified dynamics.

Anyway, although very useful for data visualization, the application of PCA to diagnostics is usually risky as some condition-information can be neglected together with the lower variance components, making the detection more challenging.

Mathematically, given a d-dimensional centred dataset of n observations $X \in R^{d \times n}$, an unbiased estimator for the covariance can be used to obtain

$$S = \frac{1}{n-1} XX' \qquad (61$$

PCA corresponds to the solution of the eigenproblem

$$S\,V = V\Lambda \qquad (62$$

where V is the orthogonal matrix whose columns are the d eigenvectors v_j while Λ is the diagonal matrix of the d eigenvalues λ_j (usually sorted in descending magnitude) of the matrix S.

The matrix V can be used then to decorrelate the dataset X, that is, to rotate the reference frame to the one identified by the eigenvectors (i.e. the principal components, PCs) of matrix S:

$$Z = V'X \qquad (63$$

If the eigenvectors in V are normalized to have unit length ($v_j' v_j = 1$), the transform is a pure rotation, and it can be proved that $\sigma_j^2 = var(z_j) = \lambda_j$. Namely, the diagonal Λ is the covariance matrix of Z. Different normalizations are obviously possible. Another quite common one consists in normalizing for $v_j' v_j = \lambda_j$. In this case $var(z_j) = 1$ so that the covariance matrix of Z is the identity matrix I. In this case, on top of the rotation, a rescaling on the principal component occurs. V is then commonly called a “whitening matrix”.

It is important to point out that through PCA the overall dataset variability is decomposed into a sum of decreasing contributions, while also the whole covariance matrix S can undergo a so-called spectral decomposition: $S = V\Lambda V' = \sum_j \lambda_j v_j v_j' = \sum_j S_j$.

The geometric interpretation of PCA is related to the fact that an ellipsoid centred in the origin can be associated to any positive definite matrix such as the covariance S. Its equation can be proved to be:

$$X'S^{-1}X = 1 \qquad (64$$

The eigenvectors of S^{-1} define then the principal axes of the ellipsoid while the eigenvalues of S^{-1} are the reciprocals of the squares of the semi-axes. This can be verified remembering that the eigenvectors of S^{-1} are the same as the eigenvectors of S and the eigenvalues of S^{-1} are the reciprocal of those of S. Indeed, using the inverse transformation $X = VZ$, one can get:

$$X'S^{-1}X = Z'V'S^{-1}VZ = Z'\Lambda^{-1}Z = \sum_j \frac{z_j^2}{\lambda_j} = 1 \qquad (65$$

which is clearly the equation of an ellipsoid whose half principal axes are $\sqrt{\lambda_j} = \sigma_j$ long.
If the whitening matrix is used, a spheroid is obtained in place of the ellipsoid. To visualize the whole geometrical interpretation, Figure 29 is added.

After these considerations, a dimensionality reduction can be easily obtained considering the projection of the original X on the first PC explaining most of the dataset variability. In fact,

$$z_1 = v_1'X = v_{11}x_1 + v_{12}x_2 + \cdots + v_{1d}x_d = \sum_{k=1}^{d} v_{1k}x_k \qquad (66$$

is basically a linear combination of the d features according to the weights given by the first eigenvector and features the greatest variance $\sigma_1^2 = var(z_1) = \lambda_1$.
In general, the set of eigenvalues λ, commonly defined as the spectrum $\sigma(S)$ of the matrix S, are collected in a graph which highlights their magnitude as a function of their index j. From such graphs, the subset of eigenvalues which explains most of the variability of the dataset can be easily selected, even though, for data visualization, just the first 2 or 3 can be used.

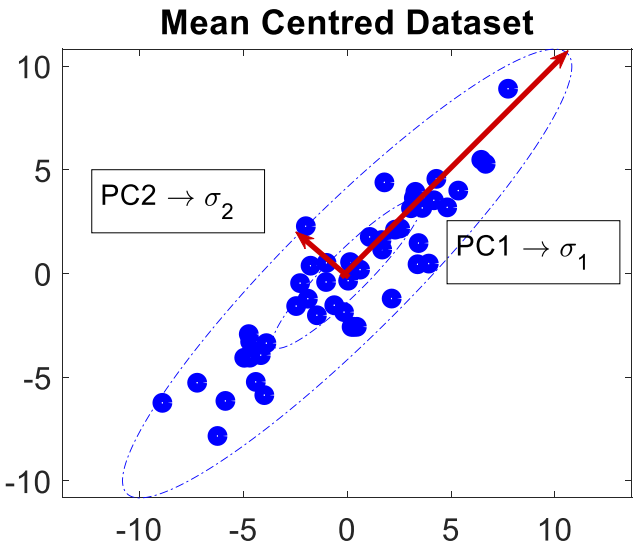


Figure 29: Visualization of the PCA principle – geometric interpretation.

3.7. Multivariate dimensionality reduction for diagnostics: Novelty Detection

In this chapter, the hypothesis testing of outliers was introduced. An outlier is a measure discordant from all the others and is then believed to be generated by an alternate mechanism. When it is possible to exclude all other possible influences (e.g. errors, latent factors like load, speed, temperature...) this inconsistency can be attributed to the presence of damage. In this respect, the detection of novelty can be successfully used to perform Level 1 diagnostics. The judgment on discordancy usually depends on a measure of distance from a reference distribution, which takes the name of Novelty Index (NI).

The Mahalanobis Distance (MD) is the optimal candidate for evaluating discordancy in a multi-dimensional space, because it is unitless and scale-invariant, and takes into account the correlations of the data set. For a mean centred dataset X the Mahalanobis distance is defined as

$$MD(X) = \sqrt{X'S^{-1}X} = \sqrt{Z'V'S^{-1}VZ} = \sqrt{Z'\Lambda^{-1}Z} = \sqrt{\sum_j \frac{z_j^2}{\lambda_j}} \equiv NI \qquad (67$$

Remembering the geometrical interpretation of PCA (derived in previous section), it is easy to understand that the Mahalanobis distance is equivalent to a Euclidean distance on the whitened space (i.e. undergoing a rotation to PCs and a standardization). This is visualized in Figure 30 where the Mahalanobis distance is decomposed into a series of simple steps.

The Novelty Indices computed with Mahalanobis distance are then a 1-D lossless compression of the multivariate dataset and can be compared against some objective criterion (i.e. a threshold) to judge whether the corresponding data comes from the healthy distribution; furthermore, even for graphical purposes, these NI are the optimal univariate dimensionality reduction tool to display possible outliers of a multivariate dataset.

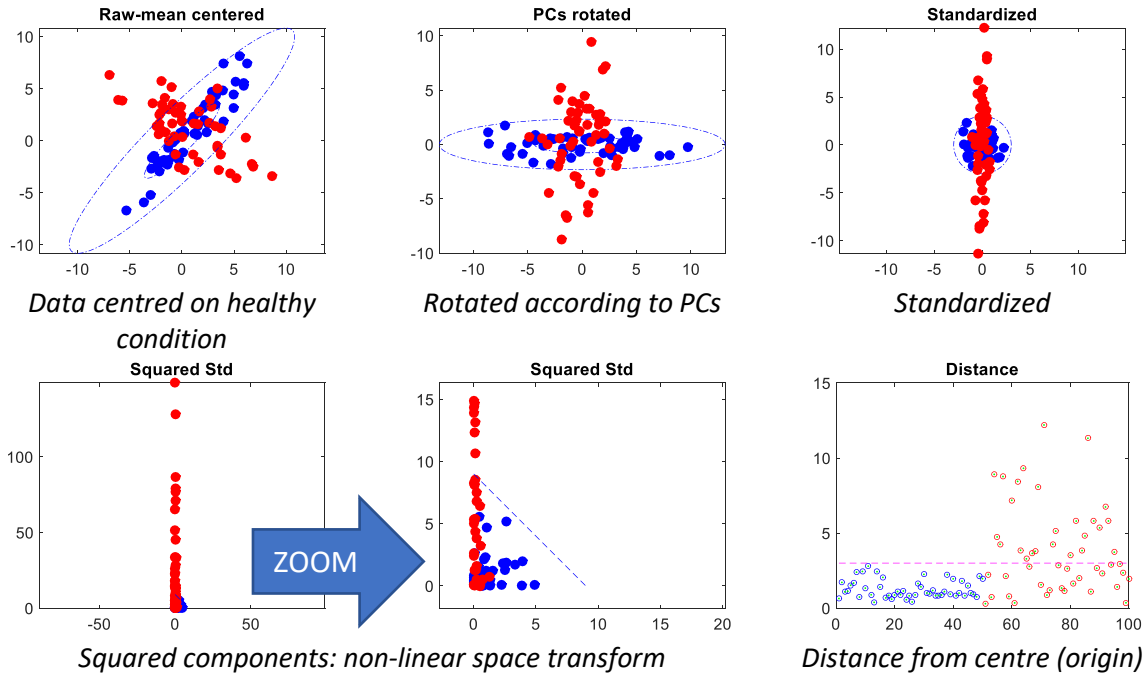


Figure 30: Mahalanobis equivalent procedure on a 2D simplified plane, for 2 simulated normal classes (blue: healthy, red: damaged). Notice that the Centring and Standardization of the space is unique and based on the healthy reference. All the damaged acquisitions will be projected in this plane.

Unfortunately, the procedure to generate a suitable threshold is not trivial, as the distribution of the healthy data may be in general non-normal. In this respect, probability theory and hypothesis testing, can offer some good hints.

3.8. Thresholding

In order to find a significant threshold for the NI, several considerations should be kept in mind. Considering that the Mahalanobis distance was proved to be equal to an Euclidean distance from the spheroid centre on the standardized principal components, a simplification can be helpful. If a d-dimensional Gaussian distribution is assumed, in fact, after centring, the squared Eulerian distance corresponds to the sum of the squared components. It is then asymptotically distributed (i.e. n large) as a χ^2_d distribution. A correction for small samples n with estimated mean and covariance matrix was also given in Section 3.2.2. In any case, this means that if the assumption of normality holds, the distribution of the NI^2 is known and can be used to derive a threshold. Obviously, this assumption can result quite restrictive in the majority of the real applications, so that alternatives and improvements are needed.

Chebyshev's inequality

Chebyshev's inequality guarantees that, for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean. In particular, it can be stated as "nearly all values are close to the mean — no more than a fraction of $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean" [61]. This hypothesis in many cases

overestimates the tails of a distribution (for example for an ideal gaussian, the tails decay more rapidly than that), but it may be very helpful to obtain first-guess values. [19]

Mathematically, it is easy to understand that, if the second order moment of a distribution (the variance) can be written as the equality

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x) dx \qquad (68$$

then, also the inequality obtained by removing a small area in the vicinity of the average holds:

$$\sigma^2 \geq \int_{|x| \geq \epsilon} x^2 p(x) dx \qquad (69$$

This inequality can only be strengthened if the variable x^2 is substituted with its smaller considered value ϵ^2.

$$\sigma^2 \geq \epsilon^2 \int_{|x| \geq \epsilon} p(x) dx \qquad (70$$

Hence, it is proved that the probability of exceeding a given ϵ is bounded by the ratio $\frac{\sigma^2}{\epsilon^2}$

$$\sigma^2 \geq \epsilon^2 \Pr(|X| \geq \epsilon) \qquad \Pr(|X| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \qquad (71$$

If the Chauvenet's criterion introduced in Section 2.2 is recalled, a rule of thumb for the thresholding can be found. Indeed, the critical value for assessing outliers in a sample of n observations can be obtained as the value which is exceeded just once every n values, so that

$$\Pr(|X| \geq \epsilon) = \frac{1}{n} \qquad (72$$

Then, it is clear that

$$\frac{1}{n} \leq \frac{\sigma^2}{\epsilon^2} \qquad \epsilon > \sigma\sqrt{n} \qquad (73$$

Or in other words, no matter what the NI distribution is, in general, no more than 1 every n values can exceed $\sqrt{n}$ standard deviations from the mean.

This rule of thumb can be practically quite useful, but it strongly overestimates the threshold, in particular for large n. Too many missed alarms (type II errors) would be then produced. A more refined threshold focusing with a higher accuracy on the extrema is needed and can be found again from Chauvenet's criterion.

Consider again the probability $1/n$ of being more extreme than a given threshold. This means that repeating m times a draw of a random sample of size n and selecting the

maximum absolute value in each repetition, the value occurring more frequently among the m maxima turns out to be exactly the given threshold. On this intuition the so-called Extreme Value Theory is built [36]. It will be the subject of Chapter 7, Section 1.3.

Monte carlo thresholding

A robust threshold is fundamental when large dimensional spaces are considered. This can be found through several repeated Monte Carlo (MC) simulations of a p-dimensional Gaussian distribution. Drawing n observations in p variables and computing the NIs, the maximum operator could be used to generate a robust threshold, for example taking the 99th percentile of the maxima distribution [37]. It will be the subject of Chapter 7, Section 1.1.

3.9. Mahalanobis distance and confounding influences

Hereinbefore, it was stated that the distance from a population centroid can be used as a measure of discordancy and can be used to discover the presence damage. This very simple idea can be exploited for Damage Detection when the healthy vibration signal can be modelled as a stationary stochastic process, meaning that the joint probability distribution function is invariant under time translation, so that damage is left as the only possible cause of discordancy (Figure 31).

Figure 31: Stationary stochastic process and the biconditional relationship of novelty and damage.

Unfortunately, hidden latent (non-measured) factors like measurement errors, operational conditions (e.g. speed, load, …) and environmental conditions (e.g. temperature, humidity, …) will always affect the measurements. When their effect is important, non-stationarities will arise, leading to misinterpretations of the novelty (and then damage), so that they are often referred to as confounders (Figure 32).

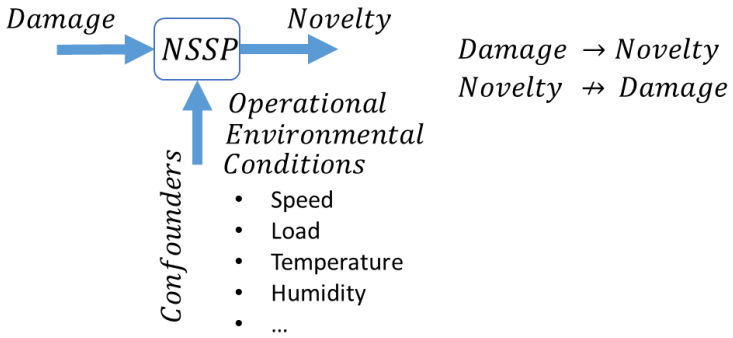


Figure 32: Non-stationary stochastic process and the effect of confounders

When measurement errors enter the game, robust statistics should be used to remove these events from the training dataset. Examples can be found in the literature, such as in [42,43,44], where the main methods for robust covariance estimation are compared. In particular, it is clear how the Minimum Covariance Determinant (MCD) is

more performing than Minimum Volume Enclosing Ellipsoid (MVEE), but computationally less efficient. So that the FAST-MCD is probably the most advisable and most widely applied.

On the contrary, when the operational and environmental influence is present and a wise feature selection is not enough for reducing the effect of the confounders, algorithms for compensating such effects become essential.

Generally speaking, when the confounding influence is strong, it can be proved to be a main source of variability in the dataset, so that it will be pictured by PCA in one of the first (or most probably the first) principal component [38, 39]. Assuming any generic influence on the selected features that can lead to a strong linear or at least a quasi-linear relationship among the features, this will be captured by a principal component, the removal of which can result helpful in highlighting the damage influence.

Actually, the Mahalanobis distance Novelty Detection already accounts for a compensation which reduces the effect of the first components. Remembering the geometrical interpretation of PCA and Mahalanobis distance seen in Section 3.7, it results:

$$NI = \sqrt{\sum_j \frac{z_j^2}{\lambda_j}} \qquad (74$$

it is obvious, then, how the influence of the components with larger variance (i.e. the first components and then also the confounding factors) is mitigated by their normalization on the corresponding eigenvalue. This means that the Mahalanobis Distance-based Novelty Indices implicitly compensate for quasi-linear confounding effects [40, 41].

Nevertheless, even if a lossless analysis which does not neglect any information is more conservative (and safer), in some cases a pre-processing to remove the confounding influences (e.g. the removal of first PCs) is fundamental to improve the performance of the otherwise not very effective MD ND.

When non-linear confounding influences occurs, on the contrary this kind of pre-processing can result ineffective, and MD ND is usually unable to face the problem. In any case, more complex methods can be used to generalize the Novelty detection.

In the next sections these ideas will be briefly developed.

3.10. Confounding influences compensation via pre-processing

To clarify the effect of confounding influences, the second part of the DIRG dataset which will be introduced in next Chapter is used for a concrete example. In particular, acquisitions at controlled temperature and constant load but variable speed are available. The measurement involves an uncontrolled braking of the machine from full speed (470 Hz) to a stop. The features from the first channel of the first accelerometer with null load are reported in Figure 33, where the strong influence of the variable speed on the RMS and peak value of the acceleration signal is highlighted.

Stationarity is noticeably violated as a trend is clearly visible in Figure 33. In the literature [45] two simple models for such violations can be found. These are tested in Figure 34.

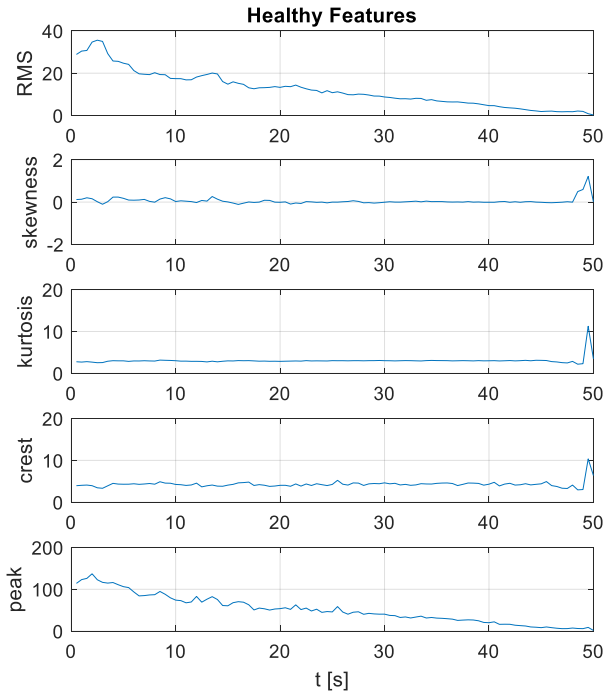


Figure 33: The five selected features from the first channel of the first accelerometer – 0N. Stationarity is noticeably violated as a trend is clearly visible in Figure 5. In the literature [45] two simple models for such violations can be found. These are tested in Figure 6.

The first model involves a deterministic trend, so that the resulting signal takes the name of **trend stationary**. A polynomial fitting can be used in this case to find and subtract the trend, leading to a stationary residual which is said to be "white" as the resulting frequency spectrum turns out to be flat (i.e. the residual is a white noise) or "decorrelated" as its autocorrelation is null for any lag different from 0.

$$y_t = \beta t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2) \qquad (75$$

A second model on the contrary involves a stochastic trend. In the simplest case, this means that the increment in the signal from time to time (innovation) is defined as a stochastic process ε such that

$$y_t - y_{t-1} = \varepsilon_t \qquad (76$$

In this case the signal y is the result of the integration of the considered stochastic process ε and is then called integrated of order 1 or $I(1)$. This process which corresponds to a random walk is **difference stationary** as its first difference is stationary. Again, it is possible to get a stationary signal which can be considered white.

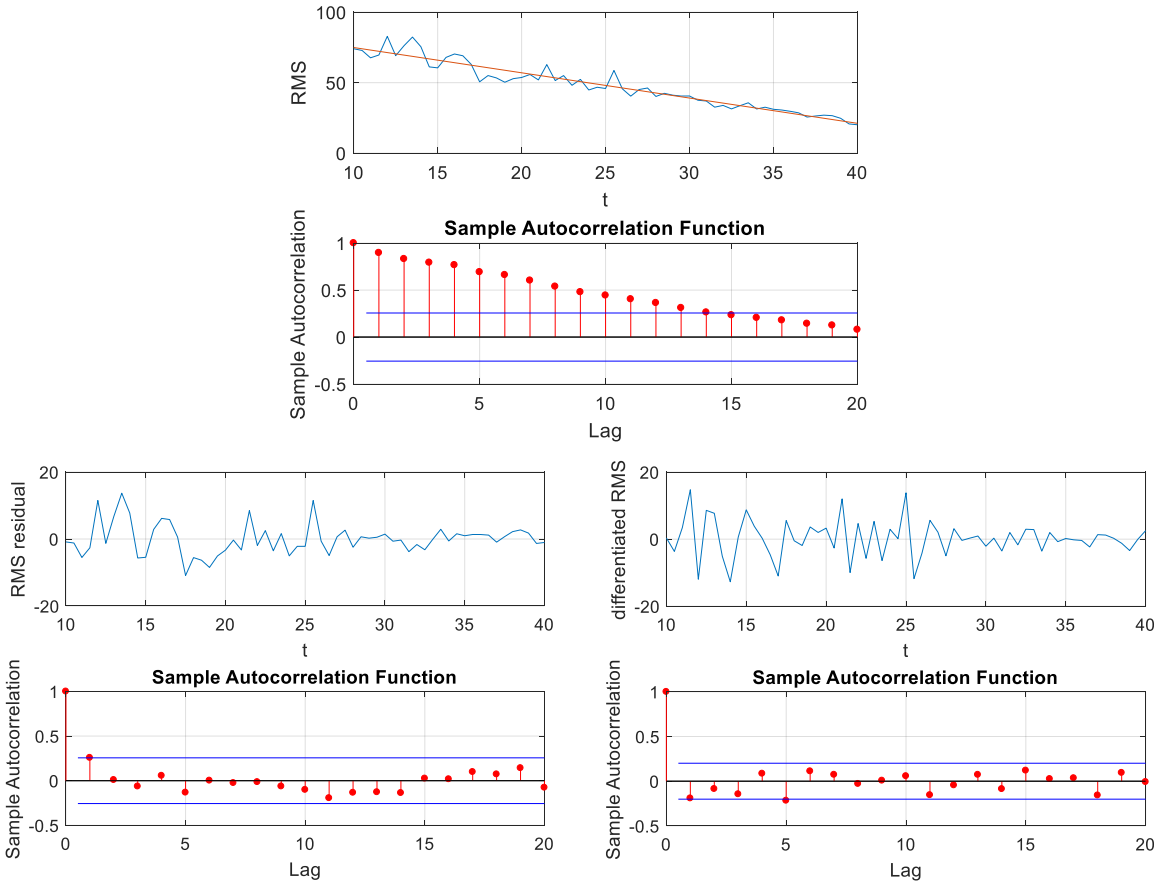


Figure 34: The raw RMS from 10 to 40s and its autocorrelation function (ACF). Below, on the left the residual from removing the linear regression and its ACF, on the right, the differencing (equivalent to the residual after an AR(1) fit) and its ACF.

In general, the random walk can be considered as a particular case of an autoregressive AR(1) model with a unitary coefficient. That explains why it is very common in the literature to whiten data by fitting an AR(1) to the series and focusing on the residual, as done in [46] to highlight the damaged bearing signature.

These concepts can be extended to multivariate spaces. In fact, when the features are affected by the same confounder, they turn out to be strongly correlated (in simple terms, they vary in sympathy).

Under the first assumption (trend stationarity) then, a multivariate regression can be used. In this case, considering that both the variables are affected by measurement errors and that it is not easy to find a dependent and an independent quantity, the orthogonal regression [47] based on PCA is proposed.

(1)

(2)

(3)

(4)

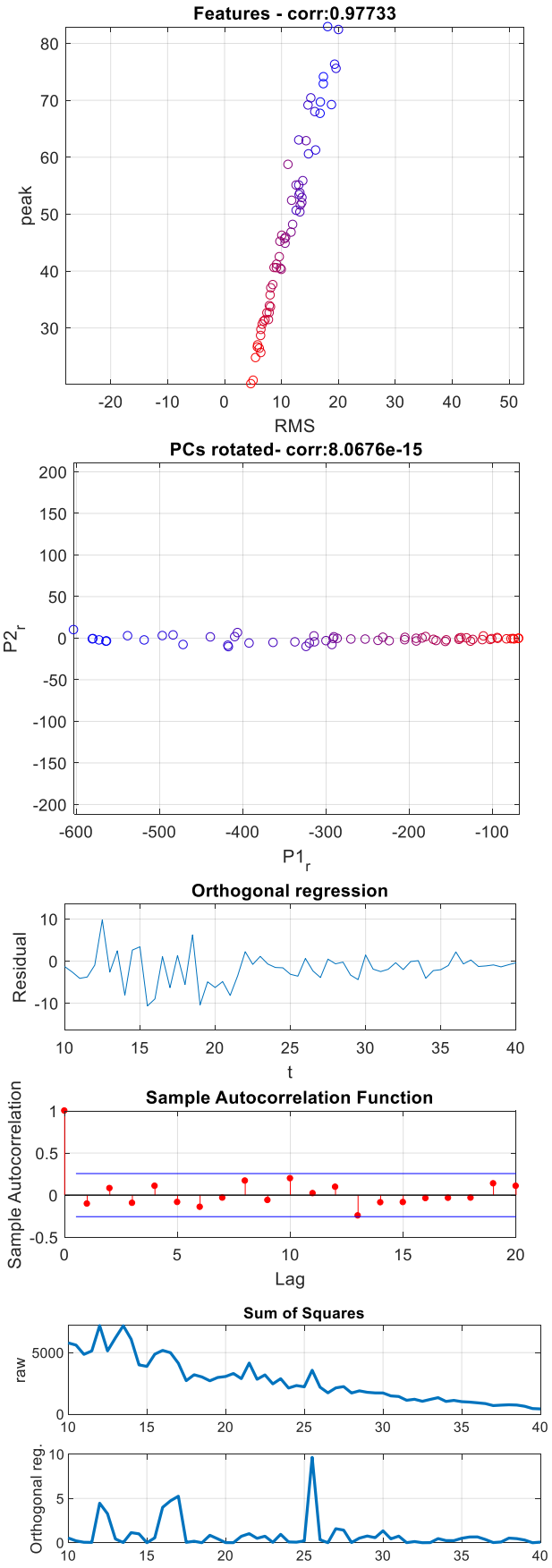


Figure 35: (1) The bivariate scatterplot with time evolution (from blue – 10s to red – 40s) (2) PCA rotation (3) Rotated PC2 corresponding to the OR residual and its ACF (4) sum of squares of the OR residual compared to the raw Euclidean squared distance

Orthogonal regression is fundamentally a reconstruction of the dataset in a subspace obtained by neglecting the first principal component. This methodology can be merged to the PCA whitening, to directly obtain a white, unitary covariance residual. PCA orthogonal regression and whitening are mathematically tackled in next subsection 3.10.1.

The results of a simple bivariate analysis on the RMS and the Peak value of the first channel are reported in Figure 35 to show the ability of the two methods on a real acquisition. In case of real measurements in fact, as confirmed by this simple analysis, it may be difficult to confidently identify the underlying model as both may work in a quite proper way.

Nevertheless, focusing on the final scope of detecting novelty (and damage), a relevant consideration can be made about novelty indices (NI). Novelty detection in fact, is commonly based on the Mahalanobis distance [37,50] which is known to be equivalent to a Euclidean distance on a features space rotated to match the principal components and normalized to obtain unitary variance PCs.

Hence, the squared Mahalanobis distance equals the sum of the squared whitened principal components. Obviously, it involves also the first PC which pictures the confounding factor, so that it is not stationary, as shown in Figure 36. A good idea then is to use as novelty index the sum of the squared principal components rejecting the first or, at most, a few of them.

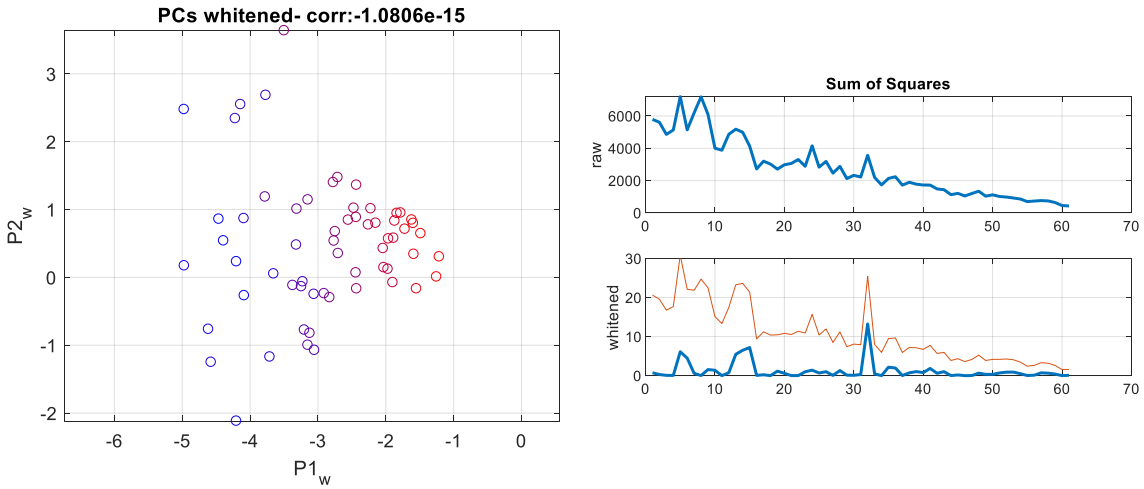


Figure 36: Standardized principal components (whitening) and sum of squares of the two PCs (in red) and of the second alone (blue, mid graph) compared to the squared Euclidean distance (raw)

3.10.1. PCA orthogonal regression and whitening

Orthogonal regression is an extension of traditional regression for datasets in which the independent variable is not assumed to be perfectly known but admits errors. Indeed, in statistical literature this is known as "errors-in-variables" model, or also, "total least squares". A simple but effective way to perform this task is based on Principal component analysis PCA [35,51].

Mathematically, given a d-dimensional centred dataset of n observations $X \in R^{d \times n}$, an unbiased estimator for the covariance can be used to obtain:

$$S = \frac{1}{n-1} XX' \qquad (77$$

PCA corresponds to the solution of the eigenproblem:

$$S\,V = V\Lambda \qquad (78$$

where V is the orthogonal matrix whose columns are the d eigenvectors v_j while Λ is the diagonal matrix of the d eigenvalues λ_j (usually sorted in descending magnitude) of the matrix S.

The matrix V can be used then to decorrelate the dataset X, that is, to rotate the reference frame to the one identified by the eigenvectors (i.e. the principal components, PCs) of matrix S:

$$Z = V'X \qquad (79$$

If the eigenvectors in V are normalized to have unit length ($v_j' v_j = 1$), the transform is a pure rotation, and it can be proved that $\sigma_j^2 = var(z_j) = \lambda_j$. Namely, the diagonal matrix Λ is the covariance of Z.

Focusing on linear orthogonal regression, the direction given by the first eigenvalue corresponds to the regression line, so that the residuals X_L can be simply found as a projection on the subspace generated by the $L = d - 1$ components other than the first:

$$z_{\mathrm{j}} = v_{\mathrm{j}}' X = v_{\mathrm{j}1} x_1 + v_{\mathrm{j}2} x_2 + \cdots + v_{\mathrm{j}d} x_d = \sum_{k=1}^{d} v_{\mathrm{j}k} x_k$$

$$Z_L = V_L' X \qquad (80$$

$$X_L = V_L Z_L = V_L V_L' X$$

The orthogonal regression is visualized in Figure 37.

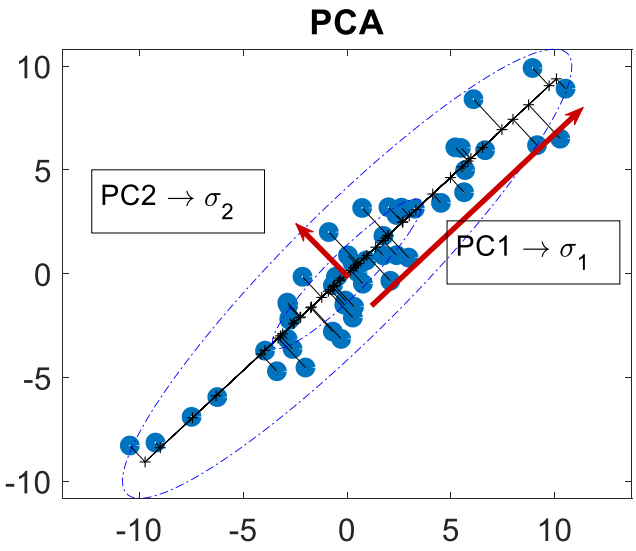


Figure 37: Visualization of the PCA orthogonal regression – the residuals corresponding to PC2 are highlighted.

Different normalizations for the eigenvectors are obviously possible. Another quite common one consists in normalizing for $v_j' v_j = \lambda_j$. In this case $var(z_j) = 1$ so that the covariance matrix of Z is the identity matrix I. In this case, on top of the rotation, a rescaling on the principal component occurs. V is then commonly called a "whitening matrix" W or also sphering matrix as it transforms the data covariance ellipsoid to a spheroid [47].

$$Z_W = W'X = \Lambda^{-1/2} V'X = \Lambda^{-1/2} Z \qquad (81$$

Finally, the squared Mahalanobis distance can be then written as

$$SMD = X'S^{-1}X = Z'V'S^{-1}VZ = Z'\Lambda^{-1}Z = \sum_j \frac{z_j^2}{\lambda_j} = Z_W' Z_W = \sum_j z_{Wj}^2 \qquad (82$$

This proves that the squared Mahalanobis distance corresponds to the sum of squares of the whitened features. Hence, removing the first whitened component(s) from the sum corresponds to merging orthogonal regression and PCA-whitening: the so found distance is therefore robust to confounders. This makes it a good candidate to substitute the Mahalanobis distance as NI in the presence of non-stationary operational or environmental conditions.

3.11. Confounding influences compensation via improved Novelty Detection

Mahalanobis distance Novelty Detection is suitable when the distribution of the dataset in the feature space is normal or quasi-normal. Hence, when strong confounder influences enter the game, MD ND performance starts to decrease. Nevertheless, the same approach can be used by switching from distance to probability.

Given the assumption of normality in fact, it is easy to derive a bijective relation among distance and probability density. In polar coordinates, when the radius r corresponds to the Euclidean distance from the centroid (the mean), the probability density function of a d-dimensional Gaussian (null mean vector and identity covariance) can be written as

$$f(r|d) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{r^2}{2\sigma^2}} \qquad r^2 = \sum_i x_i^2 \qquad (83$$

A visualization of the bijective relation of distance r and probability density f is reported in Figure 38 for a d-dimensional Gaussian with an increasing space dimensionality.

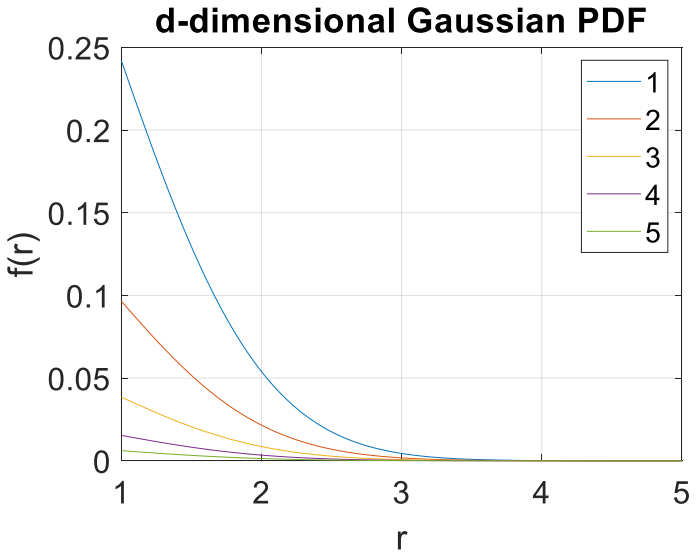


Figure 38: Probability density function in polar coordinates as a function of the distance r for an increasing space dimension d.

Then, it is possible to univocally translate the Mahalanobis NIs to probability density NIs on which an equivalent threshold can be found. Changing the perspective in this way, the problem of Novelty Detection can be extended to any generic distribution, as far as it is possible to estimate its probability density function. Two methods are introduced hereinafter at this purpose.

3.11.1. Kernel Density Estimation

Kernel Density Estimation (KDE), also known as Parzen–Rosenblatt window method, is a non-parametric way to estimate the pdf of a variable. In its 1D original framework it corresponds to a data smoothing problem where inferences about the population are made, based on a finite data sample [52,53]. The idea comes directly from

that of Histograms [54], that is, chosen a bin width h, to count the number of samples n_i out of the total n that falls inside the i-th bin. Mathematically

$$\hat{f}(x) = \frac{1}{n}\ \frac{n_i}{h_i} \qquad (84$$

This discrete estimate can be smoothed just by defining a range of influence for the samples (a bandwidth h). A rectangular window of width h and height $1/nh$ is then placed on each observation and the pdf is retrieved by summing up the contributions on a discretization which can go beyond the window size h. Furthermore, just by changing the window shape, which takes the name of kernel function, different smoothing can be obtained. Common kernel functions are the uniform (or rectangular), the triangular, the Epanechnikov, the normal, and others. Nevertheless, given its convenient mathematical formulation and good performance, the normal kernel is often used [55].

For a generic kernel function $K(x)$, the KDE can be formulated as:

$$\int_{-\infty}^{\infty} K(x)dx = 1 \qquad \hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (85$$

This can be turned into a multidimensional KDE by

$$\hat{f}(x|d,h) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{h}\right) \qquad (86$$

which, using a Multivariate Gaussian Kernel (in polar coordinates), corresponds to:

$$\hat{f}(r|d,h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}h^d} e^{-\frac{(\|r-R_i\|)^2}{2h^2}} \qquad (87$$

As it is easy to notice, KDE is a non-parametric estimator (it does not make assumptions about the parameters and their distributions), but it relies on the external variable h, which should be optimized. Handpicking of bandwidth h can lead to either too much bias or variance in the estimate, so that cross-validation is usually applied.

Nevertheless, when the space dimension d is high, problems related to the curse arise. In particular, it is known that point to point distance tends to get uniform increasing d, so that the estimated kernel density tends to flatness.

In these cases, then, it is not really advisable to use KDE. Furthermore, it must be considered that all the training point information must be stored to compute the density of new points. This can become a big drawback in case of continuous acquisitions.

However, treating the generic multimodal distribution as a mixture of a proper number of unimodal models, it can be enough to improve the novelty detection. In the next section, Gaussian Mixture Models will be then considered.

3.11.2. Finite Gaussian Mixture Models

A Gaussian Mixture Model (GMM) assumes that the underlying distribution can be modelled as a weighted sum of simple Gaussians [63,66], such as, given m mixture components:

$$\hat{f}(x|d,m) = \sum_{i=1}^{m} \phi_i \, N(\mu_i, \Sigma_i) = \sum_{i=1}^{m} \phi_i \frac{\exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right)}{\sqrt{(2\pi)^d|\Sigma_i|}} \qquad (88$$

Where weights ϕ_i and the component parameters μ_i Σ_i should be estimated. This estimation is simplified if the variance matrices can be assumed to be equal (the homoscedastic case), and further simplified in case of $\Sigma_i = \sigma^2 I$ (with I the identity matrix). In any case, the number of mixtures should always be much lower than the number of samples ($m \ll n$) for getting decent estimates.

The common estimation process for GMM is the Expectation maximization (EM). Given the number of mixtures m, EM is a particular way of implementing maximum likelihood estimation via iterative optimization [56,57]. EM is usually initialized by a clustering algorithm (e.g. k-means) [58].

On the other hand, the number of components m can be optimized by focusing on the Negative Log Likelihood (NLL). The likelihood expresses how probable the given set of observations is for different values of the statistical parameter m

$$L = \prod_i L(x_i|m) = \prod_i f_m(x_i) \qquad NLL = -\ln(L) \qquad (89$$

In information theory, other similar criteria can be found, such:

- Akaike information criterion (AIC) [65]:
  Let k be the number of estimated parameters in the model. Let L be the maximum value of the likelihood function for the model. Then the AIC value of the model can be computed as

$$AIC = 2k - 2\ln(L) \qquad (90$$

- Bayesian information criterion (BIC) [64]:
  BIC also accounts for the sample size n and can be formulated as

$$BIC = \ln(n)\, k - 2\ln(L) \qquad (91$$

Alternative methods for parameter estimation, such as the Bayesian approach based on a Markov-chain Monte Carlo algorithm can be found [62].

The GMM formulation, can be interpreted also in terms of Neural Networks. A brief introduction to Radial Basis Functions Neural Networks is given in the following section.

3.11.3. Gaussian Radial Basis Functions Neural Network

In general, a radial basis function (RBF) is a real-valued function whose value depends only on the distance between the input and some fixed point, either the origin, or a centre c_i. Hence, any function $\varphi(x)$ that satisfies the property $\varphi(x) = \varphi(|x|)$ is a radial function. The distance metric is usually the Euclidean, although other metrics can be used. Anyway, one of the most common RBF is the Gaussian RBF, thanks to its compact formulation:

$$\varphi(x) = e^{-(\epsilon|x-c_i|)^2} \quad (92$$

where the shape parameter ϵ is obviously related to the variance.

In general, RBF can be interpreted as a simple kind of neural network, which was the context in which they were originally applied to machine learning [59]. Radial basis functions are meant to build up an approximation of a function $y(x)$ as a sum of N RBFs, each associated with a different center c_i and weighted by an appropriate coefficient w_i

$$y(x) = \sum_{i=1}^{N} w_i \varphi(|x - c_i|) \quad (93$$

This sum anyway corresponds to a simple single-layer type of artificial neural network called a radial basis function network, with the RBFs used as activation functions of the network. It can be proved that this formulation is able to interpolate with arbitrary accuracy any continuous function on a compact interval if a sufficiently large number of RBFs is used. This approximating function is linear in the weights, so that they can be estimated with the matrix methods of linear least squares or using any of the standard iterative methods for neural networks [60]. Anyway, in contrast to training an MLP network, learning in an RBF network is usually done in two stages:

- Adjustment of the parameters of the RBF (i.e. number of RBFs N, centres c_i and scaling parameter ϵ) using unsupervised procedures such as clustering,
- Training of the output weights.

Bibliography

[1] Fyfe, K. R., & Munck, E., "Analysis of computed order tracking", Mechanical Systems and Signal Processing, Volume 11, Issue 2, Pages 187-205, 1997, DOI:10.1006/mssp.1996.0056

[2] R.B. Randall, J. Antoni, "Rolling Element Bearing Diagnostics - A Tutorial", Mechanical Systems and Signal Processing 25(2):485-520, 2011, DOI: 10.1016/j.ymssp.2010.07.017

[3] J. Antoni et al., "Feedback on the Surveillance 8 challenge: Vibration-based diagnosis of a Safran aircraft engine", Mechanical Systems and Signal Processing, Volume 97, Pages 112-144, 2017, DOI: 10.1016/j.ymssp.2017.01.037

[4] J. Antoni, R.B. Randall, "Unsupervised noise cancellation for vibration signals: part I—evaluation of adaptive algorithms", Mechanical Systems and Signal Processing, 18, pp89–101, 2004, DOI:10.1016/S0888-3270(03)00012-8

[5] J. Antoni, R.B. Randall, "Unsupervised noise cancellation for vibration signals: part II—a novel frequency-domain algorithm", Mechanical Systems and Signal Processing, 18, pp103–117, 2004, DOI:10.1016/S0888-3270(03)00013-X

[6] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals",
Volume 20, Issue 2, Pages 282-307, 2006, DOI: 10.1016/j.ymssp.2004.09.001

[7] J. Antoni, R.B. Randall, "The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines", Mech. Syst. Signal Process. 20 308–331, 2006. DOI: 10.1016/j.ymssp.2004.09.002.

[8] J. Antoni, "Fast computation of the kurtogram for the detection of transient faults", Mech. Syst. Signal Process. 21,108–124, 2007. DOI: 10.1016/j.ymssp.2005.12.002.

[9] R.F. Dwyer, "Detection of non-Gaussian signals by frequency domain kurtosis estimation", International Conference on Acoustic, Speech, and Signal Processing, Boston, pp. 607-610, 1983

[10] C. Otonnello, S. Pagnan, "Modified frequency domain kurtosis for signal processing", Electronics Letters, 30 (14), pp. 1117-1118, 1994

[11] W. A. Smith, R. B. Randall, "Rolling Element Bearing Diagnostics Using the Case Western Reserve University Data: A Benchmark Study", Mechanical Systems and Signal Processing, 64-65, 2015. DOI: 10.1016/j.ymssp.2015.04.021

[12] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", 1998, DOI: 10.1098/rspa.1998.0193

[13] Rilling G., & Flandrin P., “One or Two Frequencies? The Empirical Mode Decomposition Answers”, IEEE Transactions on Signal Processing, 56, 2008. DOI:10.1109/TSP.2007.906771

[14] Wu Z., Huang N.E., “Ensemble empirical mode decomposition: a noise-assisted data analysis method”, Advances in Adaptive Data Analysis, Vol. 1, No. 1, pp. 1–41, 2009.

[15] P. Flandrin, P. Goncalves and G. Rilling, “EMD equivalent filter banks, from interpretation to applications, in Hilbert–Huang Transform: Introduction and Applications”, eds. N. E. Huang and S. S. P. Shen, World Scientific, Singapore, 2005. DOI: 10.1142/9789814508247_0005

[16] A. A. Tabrizi, L. Garibaldi, A. Fasana, S. Marchesiello, “Automatic damage identification of roller bearings and effects of sifting stop criterion of IMFs”, Measurement, Vol. 93, 2016, DOI: 10.1016/j.measurement.2016.07.047.

[17] Q. Du, S. Yang, “Application of the EMD method in the vibration analysis of ball bearings”, Mechanical Systems and Signal Processing 21, 2007. DOI: 10.1016/j.ymssp.2007.01.006

[18] R. Benzi, A. Sutera, A. Vulpiani A, “The mechanism of stochastic resonance”, Journal of Physics A: mathematical and general, 1981

[19] G. P. Harmer, B. R. Davis, D. Abbott, “A Review of Stochastic Resonance: Circuits and Measurement”, IEEE transactions on instrumentation and measurement, VOL. 51, 2002.

[20] J. Li, X. Chen, Z. He, “Adaptive stochastic resonance method for impact signal detection based on sliding window”, Mechanical Systems and Signal Processing, Volume 36, Pages 240-255, 2013. DOI: 10.1016/j.ymssp.2012.12.004.

[21] H. Li, C. He, R. Malekian, Z. L, “Weak Defect Identification for Centrifugal Compressor Blade Crack Based on Pressure Sensors and Genetic Algorithm”. Sensors, 18, 2018. DOI: 10.3390/s18041264.

[22] Westfall P. H. “Kurtosis as Peakedness, 1905 - 2014. R.I.P.” The American statistician, 68(3), 191-195, 2014. DOI: 10.1080/00031305.2014.917055

[23] S. Missiakoulis Comments to: Westfall, P. H. (2014), "Kurtosis as Peakedness, 1905-2014. RIP," The American Statistician, 68, 191-195. The American Statistician 69(1):62-62, 2015. DOI: 10.1080/00031305.2014.984816

[24] Mba C. U., Makis V., Marchesiello S., Fasana A., Garibaldi L., Condition monitoring and state classification of gearboxes using stochastic resonance and hidden Markov models, MEASUREMENT, 2018, ISSN: 0263-2241

[25] Mba C. U., Marchesiello S., Fasana A., Garibaldi L., Fault Detection in Gears Using Stochastic Resonance, Advances in Condition Monitoring of Machinery in Non-Stationary Operations, 2018, ISBN: 978-3-319-61926-2

[26] Mba C. U., Marchesiello S., Fasana A., Garibaldi L., Gearbox damage identification and quantification using stochastic resonance, MECHANICS & INDUSTRY, 2017 ISSN: 2257-7777

[27] Mba C. U., Marchesiello S., Fasana A., Garibaldi L., On the use of stochastic resonance for fault detection in spur gearboxes, DIAGNOSTYKA, 2001, ISSN: 1641-6414

[28] Worden K., Antoniadou I., Marchesiello S., Mba C. U., Garibaldi L., An illustration of new methods in machine condition monitoring, Part I: Stochastic resonance, In: JOURNAL OF PHYSICS. CONFERENCE SERIES, 2017. ISSN: 1742-6588

[29] R. Von Mises, "Probability, Statistics, and Truth". Oxford, England: Macmillan, 1939. Dover, 1957 reprint. ISBN 0-486-24214-5

[30] J. P. Holman, "Experimental Methods for Engineers" 8th edition, Mcgraw-hill Series in Mechanical Engineering, 2011. ISBN 10: 0073529303

[31] W. W. Daniel, C. L Cross "Biostatistics: A Foundation for Analysis in the Health Sciences", Wiley Series in Probability and Statistics, 2012. ISBN-13: 978-1118302798

[32] D. C. Howell, "Fundamental Statistics for the Behavioral Sciences" - 8th Edition, Cengage Learning, 2013. ISBN-13: 978-1285076911

[33] Penny, K. I. "Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance". Journal of the Royal Statistical Society: Series C (Applied Statistics), 45(1), 73-81, 1996. DOI: 10.2307/2986224

[34] C. Bishop, "Pattern Recognition and Machine Learning", Springer-Verlag New York, 2006. ISBN: 978-0-387-31073-2

[35] I.T. Jolliffe, "Principal Component Analysis", Springer, 2002. DOI: 10.2307/1270093.

[36] K. Worden, D. W. Allen, H. Sohn, C. R. Farrar, "Damage detection in mechanical structures using extreme value statistics", Proceedings of SPIE - The International Society for Optical Engineering, 4693, 289–299 (2002). DOI: 10.1117/12.475226

[37] K. Worden, G. Manson, N. R. J. Fieller, "Damage detection using outlier analysis", Journal of Sound and Vibration (2000). DOI: 10.1006/jsvi.1999.2514

[38] A.M. Yan, G. Kerschen, P. De Boe, J.C. Golinval, "Structural damage diagnosis under varying environmental conditions – Part I: a linear analysis", Mech. Syst. Signal Process. 19 (2005) 847–864, DOI: 10.1016/j.ymssp.2004.12.002.

[39] A.M. Yan, G. Kerschen, P. De Boe, J.C. Golinval, "Structural damage diagnosis under varying environmental conditions – Part II: local PCA for non-linear cases", Mech. Syst. Signal Process. 19 (2005) 865–880, DOI: 10.1016/j.ymssp.2004.12.003.

[40] A. Deraemaeker, K. Worden, “A comparison of linear approaches to filter out environmental effects in structural health monitoring”, Mechanical Systems and Signal Processing, Volume 105, 2018. DOI: 10.1016/j.ymssp.2017.11.045.

[41] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., “The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data”, Mechanical Systems and Signal Processing, Volume 120, 2019. DOI: 10.1016/j.ymssp.2018.10.010.

[42] Hubert M., Debruyne M., “Minimum covariance determinant” Wiley Interdiscip. Rev. Comput. Stat., 2010. DOI: 10.1002/wics.61

[43] N. Dervilis, E.J. Cross, R.J. Barthorpe, K. Worden, Robust methods of inclusive outlier analysis for structural health monitoring, Journal of Sound and Vibration, 2014. DOI: 10.1016/j.jsv.2014.05.012.

[44] L.A. Bull, K. Worden, R. Fuentes, G. Manson, E.J. Cross, N. Dervilis, Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data, Journal of Sound and Vibration, 2019. DOI: 10.1016/j.jsv.2019.03.025.

[45] Stadnitski T., Deterministic or Stochastic Trend: Decision on the Basis of the Augmented Dickey-Fuller Test. Methodology European Journal of Research Methods for the Behavioral and Social Sciences. 6. 83-92. 2010. DOI: 10.1027/1614-2241/a000009.

[46] Sawalhi N., Randall R.B., Vibration response of spalled rolling element bearings: Observations, simulations and signal processing techniques to track the spall size, Mechanical Systems and Signal Processing, 2011, DOI: 10.1016/j.ymssp.2010.09.009.

[47] Kessy A., Lewin A., Strimmer K., Optimal Whitening and Decorrelation, The American Statistician. 2015.DOI: 10.1080/00031305.2016.1277159.

[48] Cross E. J., Worden K., Chen Q., Cointegration: a novel approach for the removal of environmental trends in structural health monitoring data, Proc. R. Soc. A, 2011. DOI: 10.1098/rspa.2011.0023

[49] K. Worden, T. Baldacchino, J. Rowson and E. J. Cross, Some Recent Developments in SHM Based on Nonstationary Time Series Analysis, Proceedings of the IEEE, vol. 104, no. 8, pp. 1589-1603, Aug. 2016. DOI: 10.1109/JPROC.2016.2573596

[50] A. Deraemaeker, K. Worden, A comparison of linear approaches to filter out environmental effects in structural health monitoring, Mechanical Systems and Signal Processing, 2018, DOI: 10.1016/j.ymssp.2017.11.045.

[51] Van Huffel S., Vandewalle J., The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, 1991, ISBN: 978-0-898712-75-9

[52] Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". The Annals of Mathematical Statistics. doi: 10.1214/aoms/1177728190.

[53] Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode". The Annals of Mathematical Statistics. doi: 10.1214/aoms/1177704472.

[54] Silverman B. E., “Density estimation for statistics and data analysis”, Chapman & Hall - CRC Monographs on Statistics & Applied Probability, 1986. ISBN 0412246201, 780412246203

[55] Wand, M.P; Jones, M.C. (1995). Kernel Smoothing. London: Chapman & Hall/CRC. ISBN 978-0-412-55270-0.

[56] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society.

[57] Yu, Guoshen (2012). "Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity". IEEE Transactions on Image Processing. doi:10.1109/tip.2011.2176743.

[58] McLachlan, G., and D. Peel. Finite Mixture Models. Hoboken, NJ: John Wiley & Sons, Inc., 2000.

[59] Broomhead D. S., Lowe D., "Multivariable functional interpolation and adaptive networks", 1988. Complex Systems.

[60] Schwenker F., Kestler H. A., Palm G., "Three learning phases for radial-basis-function networks", 2001. Neural Networks. doi: 10.1016/s0893-6080(01)00027-2.

[61] Kvanli A. H., Pavur R. J., Keeling K. B., “Concise Managerial Statistics”, cEngage Learning, 2006. ISBN 9780324223880.

[62] Figueiredo E, Radu L, Worden K, Farrar CR, “A Bayesian approach based on a Markov-chain Monte Carlo method for damage detection under unknown sources of variability”, Engineering Structures, 2014. DOI: 10.1016/j.engstruct.2014.08.042.

[63] Figueiredo, E. & Cross, E. J, “Linear approaches to modeling nonlinearities in long-term monitoring of bridges”, Civil Struct Health Monit, 2013. DOI: 10.1007/s13349-013-0038-3

[64] Schwarz G. et al, “Estimating the dimension of a model”, Ann. Stat, 1978.

[65] Akaike H., “A new look at the statistical model identification”, IEEE Trans. Autom. Control, 1974.

[66] Rogers T. J., Worden K., Fuentes R., Dervilis N., Tygesen U. T., Cross E. J., “A Bayesian non-parametric clustering approach for semi-supervised Structural Health Monitoring”, Mechanical Systems and Signal Processing, 2019. DOI: 10.1016/j.ymssp.2018.09.013.

[67] Miller T., “Explanation in artificial intelligence: Insights from the social sciences.” arXiv:1706.07269. (2017)

Chapter 5:

The signals of interest: simulations and experimental datasets

1. Introduction

This thesis is devoted to the individuation of the most promising techniques for the detection, identification and quantification of faults in gearboxes starting from raw acceleration signals, in accordance with the vibration monitoring philosophy. In order to develop and optimize the proposed diagnostic algorithms, theoretical signals were synthesized in agreement with the considerations found in the literature review in Chapter 3. Experimental signals acquired on a test-rig at the Dynamics and Identification Group (DIRG) laboratory were also used, together with real-life signals from aeronautic and windmills gearboxes. Such acquisitions are described hereinafter.

2. Simulated signal

As first, using the knowledge collected in the state of the art of Chapter 3, a theoretical signal simulating a typical gearbox signature is synthesized. It is a sum of a deterministic signal coming from the gear, a cyclostationary train of impulses generated by the bearings and a random noise, all convolved by a transfer function corresponding to the transmission path. An additional measurement noise is also added. The procedure is described in the following paragraphs, to produce a simplified simulation of the finally measured signal.

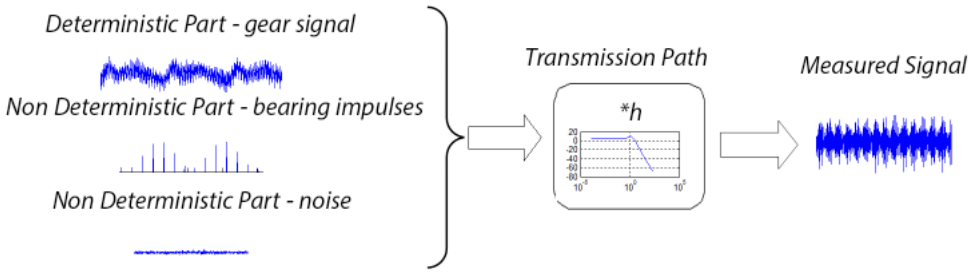


Figure 1: Acquisition modelled as a MISO system (see Chapter 3).

2.1. General information and acquisition settings

The generated signal is meant to simulate an acquisition at $fs = 22528\ Hz$ for a time-duration of $1\ s$. The measurement regards a shaft rotating at a speed $fr = 53\ Hz$ on which a mating gear-wheel with $z = 23$ teeth is mounted. The shaft is supported by common rolling element bearings. The SKF 6006-Z is selected as a reference, so that 11 balls with diameter $6\ mm$ are considered to run between an inner ring with diameter $30\ mm$ and an outer ring of $55\ mm$. An inner-race fault is considered. The bearing characteristic frequencies relative to the shaft frequency are reported in Table 1.

Table 1: SKF 6006-Z Characteristic frequencies normalized on the shaft frequency

FTF	***BSF***	***BPFI***	***BPFO***
0,43	3,47	4,72	6,28

2.2. Shaft and Gear deterministic signal

The shaft contribution f_r is modelled as a sinewave at the rotational speed. Its first harmonics are also added with a linearly decreasing amplitude. In addition, the gear-mesh frequency $f_{GM} = zf_r$ for the 23 teeth gear-wheel and its harmonics are also considered, amplitude modulated by the shaft frequency and its first harmonic.

$$y_{det}(t) = \left(1 + \sum_k \mathrm{A}_k \sin(2\pi\, k f_r\, t)\right) \cdot \sum_k \mathrm{A}_{GM,k} \sin(2\pi\, k f_{GM}\, t) + \sum_k \mathrm{A}_{S,k} \sin(2\pi\, k f_r\, t) \quad (1$$

2.3. Non-deterministic component: Bearing cyclostationary signal and noise

The bearing signal is modelled as a Pseudo-Cyclostationarity (CS2) impulse train, as introduced in Chapter 3, Section 6. The generic period $\Delta T_i = T_{i+1} - T_i$ is then generated as random draw from a normal distribution. In case of rolling element faults or inner ring faults, an additional amplitude modulation is introduced at the cage frequency (FTF) or at the shaft frequency respectively. Amplitude modulation references are given in Appendix 2. The case of an inner race defect is depicted in Figure 2. In red, the additive white gaussian noise is also highlighted.

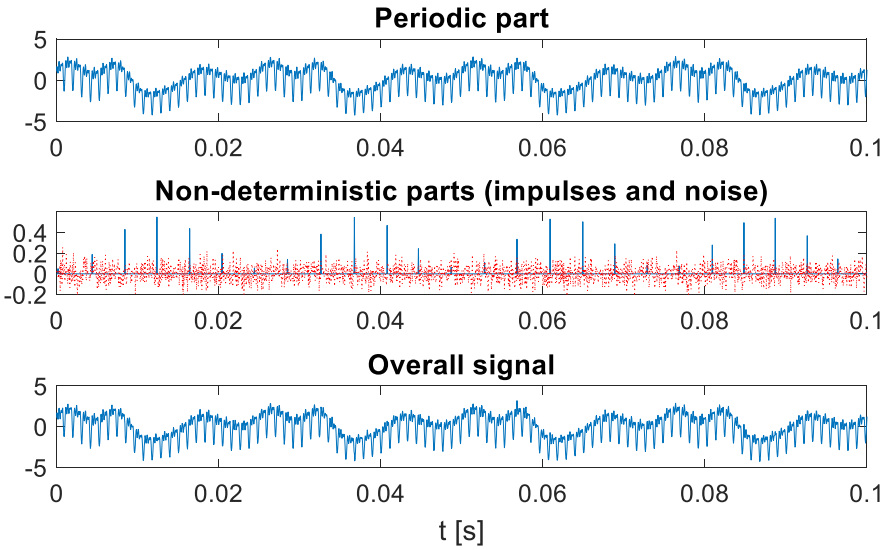


Figure 2: Sum of the deterministic and non-deterministic components due to gears and bearings + noise.

2.4. Transmission path

In order to simulate the transmission path to the sensor, a simple LTI system modelling a Transmissibility transfer function (TF) is implemented. Supposing a given resonance frequency for the whole structure ($5600\ Hz$) and a damping factor (5%), a simple 1-DOF TF is determined in accordance to Appendix 8 to simulate the time response of the modelled dynamic systems to the input signal, which corresponds to the sum of the previously generated deterministic and non-deterministic parts.

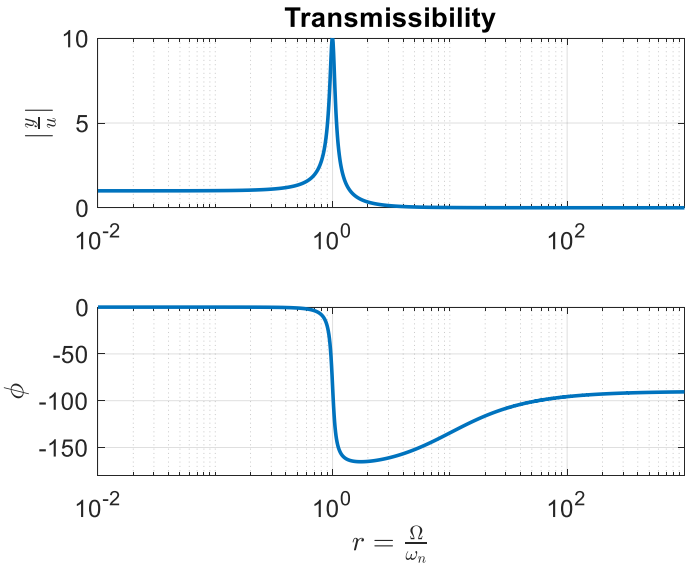


Figure 3: 1-DOF Transmissibility as introduced in Appendix 8.

2.5. Measured signal

The measured signal, further corrupted with some measurement noise (additive white gaussian noise), is finally obtained, as shown in Figure 4. It can be noticed that, in accordance with the given transmissibility, the gear-mesh harmonics undergo some amplitude deformation. This is particularly strong in the resonance region, while at higher frequencies the signal is filtered out.

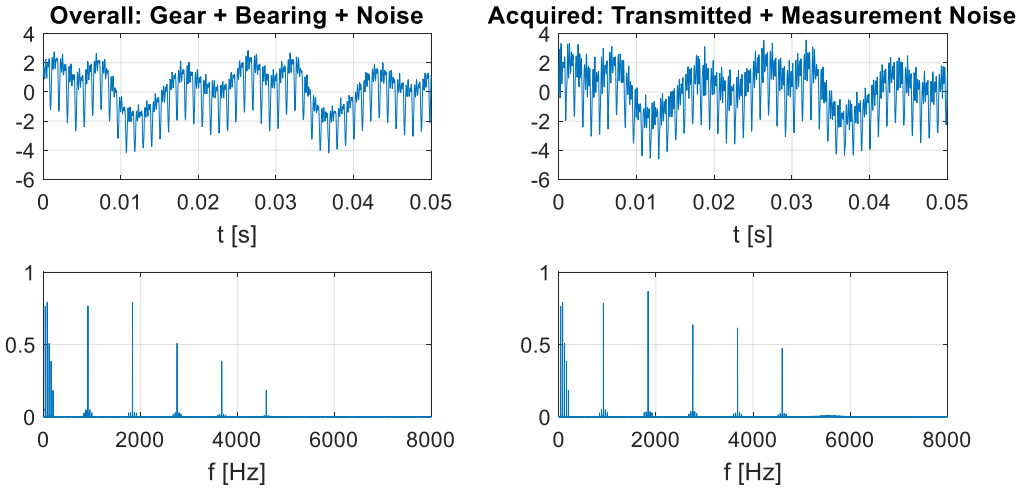


Figure 4: Effect of the simple transmission path and of the measurement noise.

3. Experimental acquisitions

The diagnostic algorithms proposed in this work were validated also on laboratory and real-life machines. The measurement setups for the 3 different acquisitions are reported hereinafter, to describe the obtained signals. In particular, the focus will be on

- Dynamic & Identification Research Group (DIRG) test rig for high speed bearings,
- SAFRAN aeronautical engine from the SAFRAN Contest, Conference Surveillance 8, October 20-21, 2015, Roanne, France,
- Italian windfarm composed by six multi-megawatt wind turbines located near the Adriatic Sea in Molise region.

3.1. DIRG test rig for high speed bearings – part 1

At the DIRG laboratory, acquisitions were performed over a rig specifically conceived to test high speed aeronautical bearings [1]. The considered test rig consists of a direct drive rotating shaft supported by two bearings, one of which (the farthest from the motor) exhibits different damage levels, as reported in Table 2. A third central bearing is used to load the shaft with an increasing force of 0, 1000, 1400 and 1800 N, while the speed is set at four different values of about 90, 180, 280, 370, 470 Hz for a total number of 17 combinations of load and speed (Table 3). The structure is equipped with four tri-axial accelerometers (positioned as reported in Figure 5) sampled at a frequency *fs* = 51200 Hz for a duration of *T* = 10 s.

The test rig is instrumented to acquire the accelerations of the two most significant points of the structure, A1 and A2 in Fig. 1b, located respectively on the support of the damaged bearing under test B1 and the support of the larger bearing dedicated to the application of the external load B2. The accelerometers are of the tri-axial IEPE type, with frequency range 1-12000 Hz (amplitude ±5%, phase ±10°), nominal resonant frequency 55 kHz and nominal sensitivity 1 $\frac{mV}{m/s^2}$. The radial force on the central bearing is measured by a static load cell whose sensitivity (0.499 $\frac{mV}{N}$) was measured by repeated load cycles, showing almost null hysteresis. A K-type thermocouple and a dedicated digital thermometer with 0.1 °C resolution are also placed as near as possible to the external ring of the damaged bearing, mainly to check that the different acquisitions are comparable in terms of temperature.

The digital data acquisition is achieved by an OR38 signal analyser, produced by OROS, whose accuracy on the input channels is: phase ± 0.02°, amplitude ± 0.02 dB, frequency ± 0.005%.

The analogue-to-digital transformation is performed by a 24 bits delta-sigma converter and is synchronous on all channels (no multiplexing); the range of every channel can be independently set between a minimum (±17mV) and a maximum (±40V) so as to avoid saturation of the channels while reaching the optimal amplitude resolution.

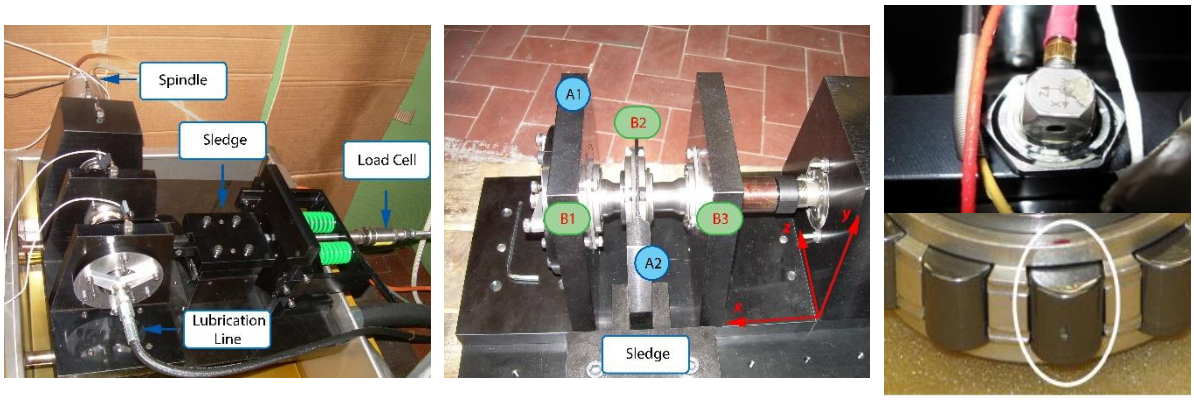


Figure 5: The test rig, the orientation of the triaxial accelerometers and the 4A damaged roller.

Table 2: Bearing codification according to damage type (Inner Ring or Rolling Element) and size.

Code	*0A*	*1A*	*2A*	*3A*	*4A*	*5A*	*6A*
Damage type	none	I.R.	I.R.	I.R.	R.E.	R.E.	R.E.
Damage size [µm]	-	450	250	150	450	250	150

Table 3: The operational conditions

Label	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
f dHz	9	9	9	9	18	18	18	18	28	28	28	28	37	37	37	47	47
F kN	0	1	1,4	1,8	0	1	1,4	1,8	0	1	1,4	1,8	0	1	1,4	0	1

Five features per each acquisition channel (six, given the 2 triaxial accelerometers) were then computed. Channels 1 and 4 measure x acceleration, channels 2 and 5 are recordings of y acceleration while channels 3 and 6 acquire acceleration along z direction.

Root mean square, skewness, kurtosis, peak value and crest factor (peak/RMS) have extracted on 0,1s chunks of the original 10s acquisitions generating 100 data points for each of the 17 acquisitions. The data-set is then composed by 7 differently damaged conditions, from 0A (healthy), to 6A, containing 1700 measurements in a 30-dimensional space (6 channels, 5 features). Focusing on condition 12 (Table 3), the 7 health conditions are pictured per each feature and channel combination, to help visualize the dataset.

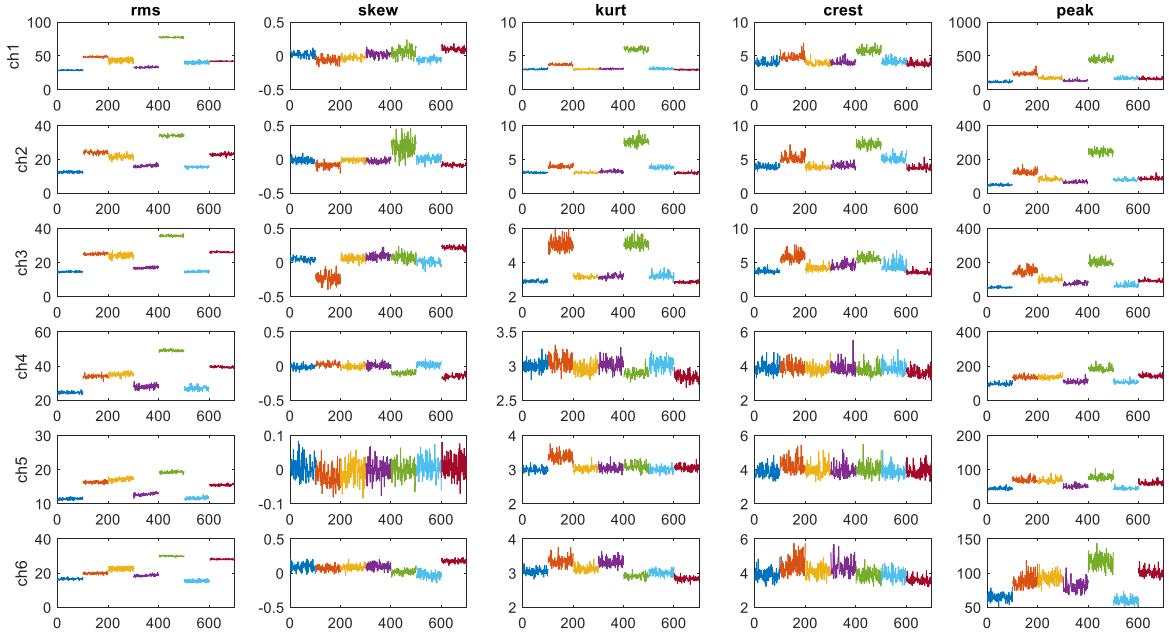


Figure 6: Condition 12 pictured by all the feature/channel combinations in all the health conditions, from 0A to 6A (100 samples each).

The described dataset was made available as an open access resource at the link that can be found in the journal paper [1], *The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data.*

3.2. DIRG test rig for high speed bearings – part 2

The second dataset obtained from the DIRG test rig regards run-down acquisitions from the maximum speed (470 Hz) to a full stop (0 Hz) while B2 is withstanding a discrete increasing force of 0, 1000, 1400 and 1800 N, which characterizes 4 different operational conditions (as highlighted in Table 4). The same two tri-axial accelerometers located respectively on the B1 bearing support (accelerometer A1, as reported in Figure 5) and on the loading sledge (accelerometer A2). The acquisitions last about T = 50 s at a sampling frequency fs = 102400 Hz. In order to perform a significant analysis, the five selected

features root mean square, skewness, kurtosis, peak value and crest factor are extracted on one hundred independent chunks (about 0,5 s each) for each of the 6 channels of the 4 original acquisitions in all the 7 health conditions (from 0A, healthy, to 6A).

Finally, 100 observations in a 30-dimensional space (6 channels, 5 features) per each operational condition are obtained. A part of the dataset is visually summarized in Figure 7.

Table 4: The operational conditions: the different loads while the speed is decreasing from 470 to 0 Hz (run-down acquisitions).

Label	1	2	3	4
F [kN]	0	1	1,4	1,8

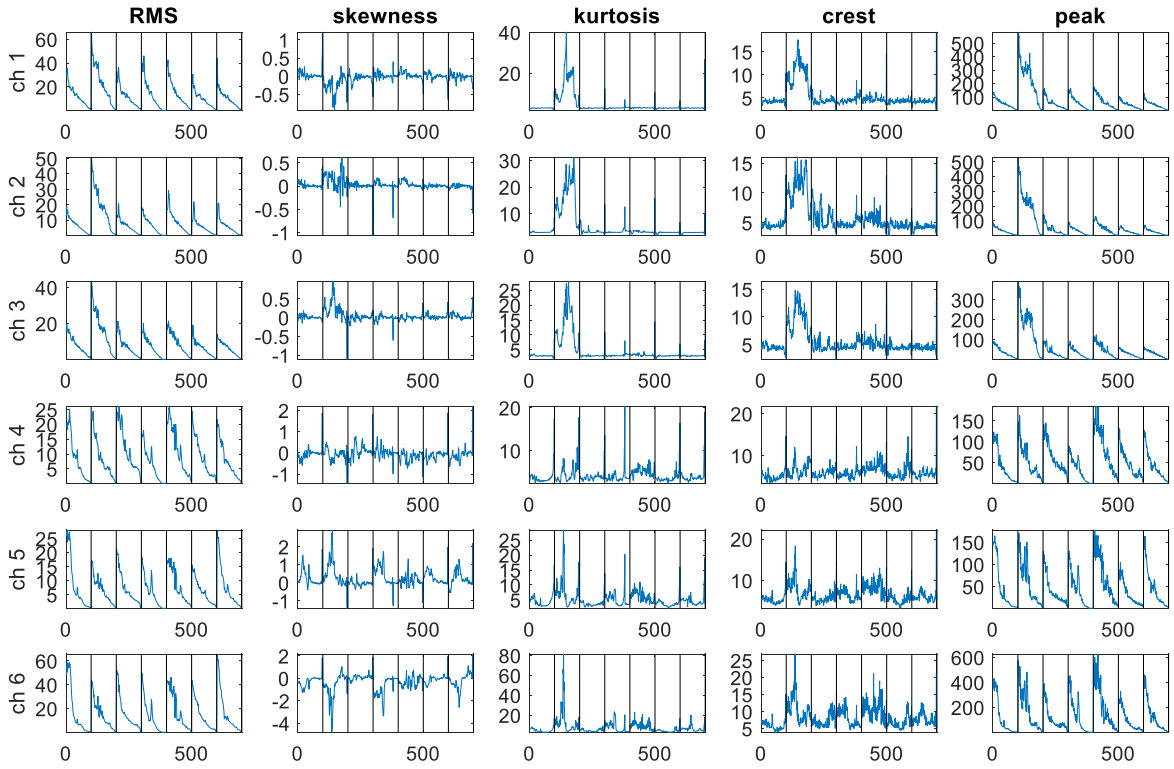


Figure 7: The considered dataset after features extraction for load condition 1 (0 N) while the speed is decreasing until a stop starting from 470 Hz. The black dotted lines divide the different damage conditions (0A to 6A). For each, 100 observations are plotted sequentially.

3.3. SAFRAN civil aircraft engine with two damaged bearings

The Safran contest data from Surveillance 8 conference were also used to test some of the algorithms. The data provided for the contest are vibration and tachometer signals acquired during a ground test campaign on a civil aircraft engine with two damaged bearings.

The engine has two main shafts and an accessory gearbox for the equipment such as pumps, filters, alternators and starter. The accessory gearbox is linked to the high-pressure shaft HP by a radial drive shaft RDS and a horizontal drive shaft HDS. The general layout of the engine is given in Figure 8, where the damaged bearings and the sensors locations are also highlighted.

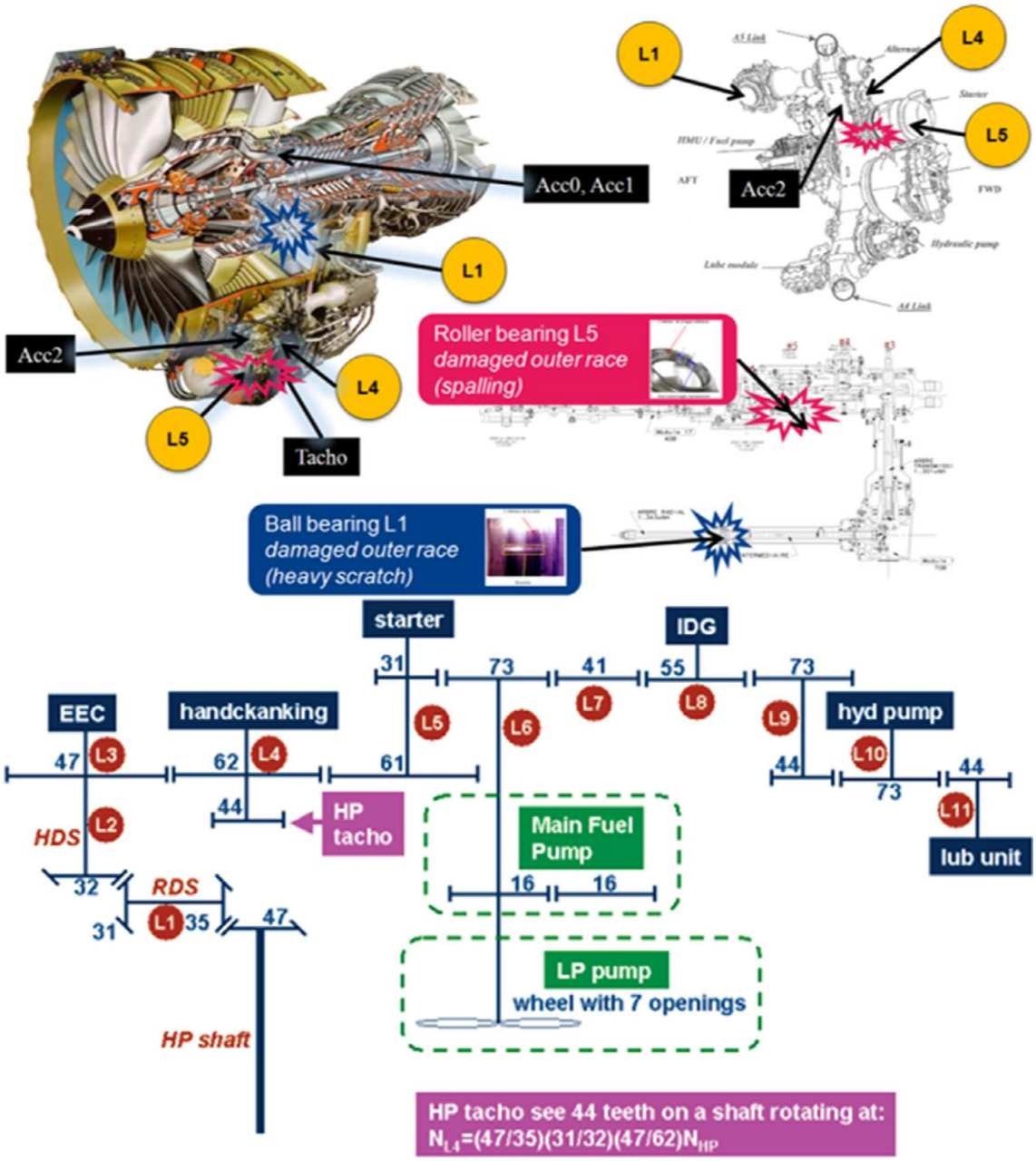


Figure 8: SAFRAN aeronautical engine and gearbox scheme. Courtesy of SAFRAN company, SAFRAN Contest, Conference Surveillance 8, October 20-21, 2015, Roanne, France.

The two damaged bearings are the BL1, supporting the radial drive shaft L1 and the BL5 on shaft L5. The first (BL1) is a ball bearing located at an intermediate position on the radial drive shaft. The damage is a heavy scratch on the outer race with a depth of about 0.3 mm and a width of about 1 mm. An unbalance was also installed on shaft L1 in order to load the bearing and therefore to be representative of a what a nearly-failed bearing would develop. The damage of the roller bearing on shaft L5 is a wide spalled area on the outer race, in a sector of 32° along all the functional line of the race, with a depth of approximately 0.1 mm. It is purposely located at the position where the static force between the outer race and the rolling elements is at maximum, thus where it has the highest probability to develop. All equipments of the accessory gearbox were operating in normal conditions, hence torque applied on the shafts and resulting static load applied on the bearing outer race were representative of usual operating conditions.

Three sensors are mounted. Two accelerometers $acc1$ and $acc2$ are located respectively on the intermediate case near the radial drive shaft and on the flange of the accessory gearbox in the vicinity of shaft L5. Additionally, a tachometer featuring a resolution of 44 pulses per revolution is mounted on shaft L4.

Two acquisitions are performed. The first one, involving only $acc1$ corresponds to a 200 s waveform sampled at 50 kHz during a steady state at full power followed by a slow deceleration down to idle. In the second acquisition all the three available sensors are recorded for the same duration and at the same rate (200 s @ 50 kHz) during a slow acceleration from idle to full power.

The raw acceleration data can be downloaded as supplementary material on the online version of [2], *Feedback on the Surveillance 8 challenge: Vibration-based diagnosis of a Safran aircraft engine*.

3.4. Italian windfarm

The final experimental setup considered in this work [3] regards on-site accelerometric measurements from the Italian windfarm in Molise region composed by six multi-megawatt wind turbines installed as depicted in Figure 9.

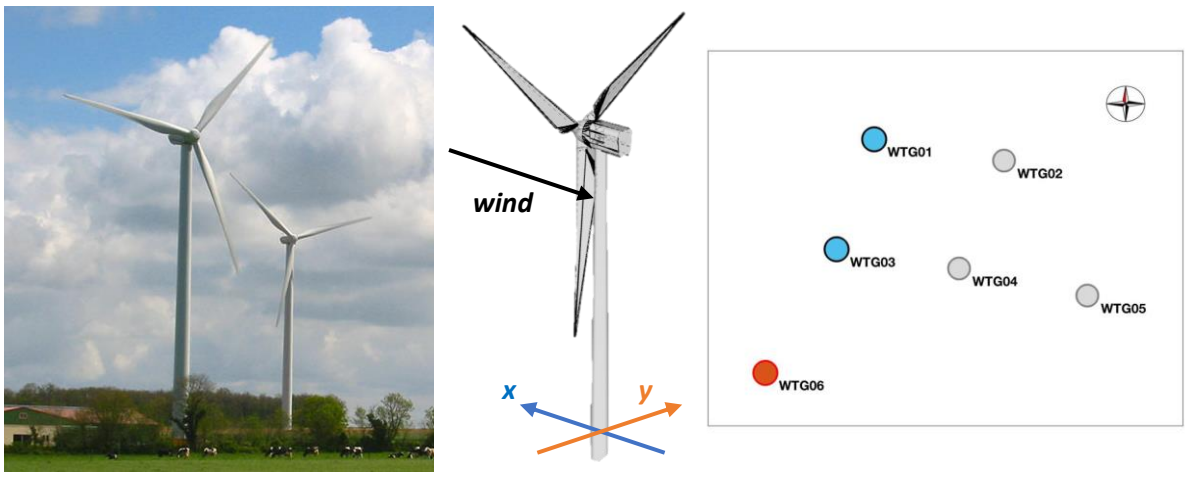


Figure 9: Italian windfarm plant and accelerometers sensing directions according to the wind direction.

The wind-mills gearboxes in this plant are monitored with an oil particle counting system which identified a damage on WTG06 turbine. An additional vibration monitoring system is then added to assess whether it is able to produce concordant diagnosis. Windmills WTG06, WTG03 and WTG01 are sensored with four mono-axial accelerometers, two on the superior level 7 m above ground and two at the inferior level 2 m above ground. Each couple is aligned so as to measure respectively the longitudinal (x-axis) and transversal (y-axis) vibrations.

Two acquisitions in different operational conditions are considered, both performed at a sampling rate of 12.8 kHz for 2 minutes. In the first case, identified as acquisition 17.20, only the healthy WTG03 and WTG01 are recorded. Acquisition 15.00 on the contrary, is involving the damaged windmill WTG06 and the healthy WTG03 (see Table 5). During the experimental campaign all the turbines were operating at rated power, in reasonably similar operating conditions. This was cross-checked thanks to the acquisition of the operational data provided by the wind turbine manufacturer in real time at a sampling rate of about 1 Hz, as shown in Figure 10.

Again, to ensure the statistical significance, each acquisition is divided in 100 sub-parts on which the five selected time features are computed. The result of this operation is graphically summarized in Figure 11.

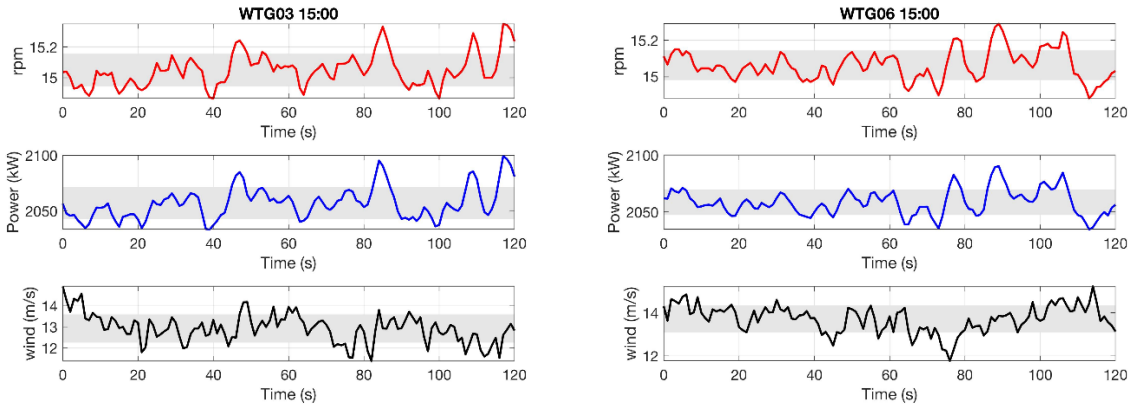


Figure 10: Wind and Operational data at a sampling rate of about 1 Hz.

Table 5: Summary of the four acquisitions and identification of the training and validation sets.

1	**WTG01 @ 17.20**	HEALTHY	Reference →Calibration: Training
2	**WTG03 @ 17.20**		
3	**WTG03 @ 15.00**		Validation
4	**WTG06 @ 15.00**	DAMAGED	

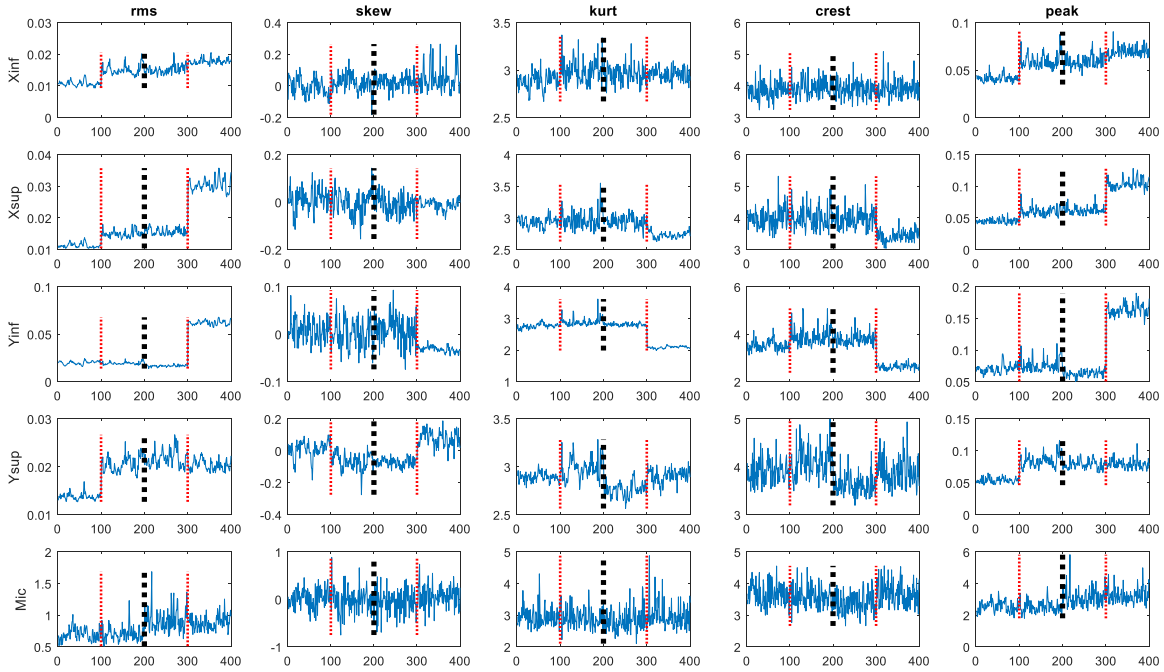


Figure 11: The extracted features. The samples 0-100 are referred to the machine WTG01 @17.20, 101-200 to WTG03 @ at 17.20, 201-300 to WTG03 @ 15.00 and samples 301-400 are from WTG06 (the damaged wind turbine) @ 15.00. The first 2 sets are used for calibration and are separated from the last 2, left for validation, by the black dotted line.

Bibliography

[1] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., "*The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data*", Mechanical Systems and Signal Processing, Volume 120, 2018. DOI: 10.1016/j.ymssp.2018.10.010.
ftp://ftp.polito.it/people/DIRG_BearingData/

[2] Antoni J., et al., "*Feedback on the Surveillance 8 challenge: Vibration-based diagnosis of a Safran aircraft engine*", Mechanical Systems and Signal Processing, Volume 97, 2017. DOI: 10.1016/j.ymssp.2017.01.037.
https://www.sciencedirect.com/science/article/pii/S0888327017300584#s0385

[3] Castellani F., Garibaldi L., Astolfi D., Daga A. P., Becchetti M., Fasana A., Marchesiello S. "*Fault diagnosis of wind turbine gearboxes through on-site measurements and vibrational signal processing*", International Conference on Structural Engineering Dynamics ICEDyn 2019, Viana do Castelo, Portugal.

Chapter 6:

Signal Processing for Intermittent Monitoring: Spectral Kurtosis, a novel estimate

1. Time-frequency representation of signals: from Parseval to Wigner-Ville

Time-frequency analysis is one of the most important areas of signal processing. The standard Fourier analysis, although very useful in identifying individual frequency components of a signal, has no time resolution.

The simplest solution is the short-time Fourier transform (i.e. the spectrogram), in which one transforms windowed section of the data to the frequency domain. Time resolution is obtained by centring the window function on the time range of interest and then sliding the window along the time axis.

The drawback is that it fails to provide high time and frequency resolution simultaneously. To localize some frequency component in time one must choose a very short window, which will inevitably lead to poor frequency resolution upon Fourier transformation. Conversely, to increase the frequency resolution one must Fourier-transform long sections of the data, which adversely affects the localization in time.

Time-varying spectra were studied in the classical works of Gabor, Ville, Page, and Wigner [1,2]. Their work was not devoted to make improvements on the spectrogram, but to construct a joint time and frequency distribution of the energy of a waveform based on general mathematical principles.

Consider an infinite duration waveform $s(t)$. The instantaneous energy of the signal per unit time (**power**) at time t is given by $|s(t)|^2$. The intensity per unit frequency (**energy spectral density**) is given by $|S(f)|^2$, where

$$S(f) = \mathcal{F}[s(t)] = \int_{-\infty}^{\infty} s(t) e^{-i2\pi f t} dt \qquad (1$$

is the standard Fourier transform relationship. Parseval's theorem then states that the **total energy** E can be computed in the time or the frequency domain; thus,

$$E = \int_{-\infty}^{\infty} E(t)\, dt = \int_{-\infty}^{\infty} |s(t)|^2 dt \equiv \int_{-\infty}^{\infty} |S(f)|^2 df = \int_{-\infty}^{\infty} E(f)\, df \qquad (2$$

The fundamental goal is then to find a joint function of time and frequency that represents the energy or intensity of a waveform per unit time and per unit frequency. This **joint** function, dented as $E(t, f)$, is then a distribution of energy (**energy density**) satisfying the **marginals**:

Total energy spectral density	$E(f) = \lvert S(f)\rvert^2 = \int_{-\infty}^{\infty} E(t,f)\, dt$	(3
Total energy density in time (total power envelope)	$E(t) = \lvert s(t)\rvert^2 = \int_{-\infty}^{\infty} E(t,f)\, df$	(4

Focusing on "slices" of the $E(t,f)$, the **conditional** distributions can be found:

Power spectral density at time t	$tPSD(f) = E(t,f\lvert t)$	(5
Power for a given frequency f (power envelope)	$fPE(t) = E(t,f\lvert f)$	(6

So, exploiting the conditionals, two equivalent point of views can be adopted. Starting from the expression of energy density at a given time $\hat{t}$ and frequency $\hat{f}$:

$$E(\hat{t},\hat{f}) = \hat{t}\, PSD(\hat{f}) = \hat{f} PE(\hat{t}) \qquad (7$$

Two relations hold:

$$E(\hat{f}) = \sum_{t=-\infty}^{\infty} tPSD(\hat{f}) = \int_{-\infty}^{\infty} \hat{f} PE(t)\, dt$$
$$E(\hat{t}) = \sum_{f=-\infty}^{\infty} fPE(\hat{t}) = \int_{-\infty}^{\infty} \hat{t}\, PSD(f)\, df \qquad (8$$

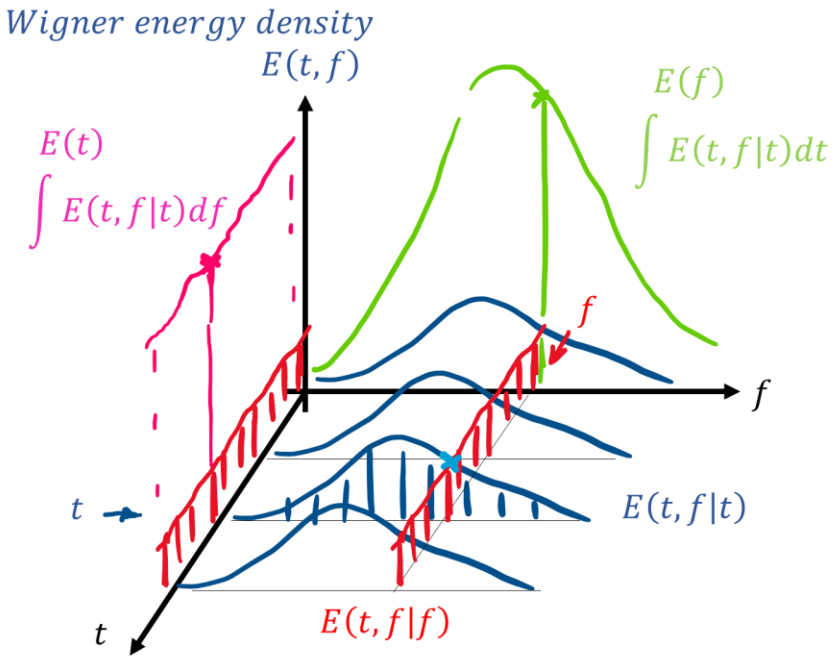


Figure 1: Wigner Energy Density function with indication of its marginals and conditional distributions

This means that the energy spectral density $E(f)$ is the sum of all the power spectral densities $tPSD(f)$ over all the time instants t (in rigorous terms, an integration with respect

to time). But pointwise, per each frequency f, it is the integral over the time axis of the power envelope $fPE(t)$ for the corresponding frequency.

$$E(\hat{f}) = \sum_{t=-\infty}^{\infty} tPSD(\hat{f}) \equiv \int_{-\infty}^{\infty} tPSD(\hat{f})\, dt = \int_{-\infty}^{\infty} E(t, f|\hat{f})\, dt = \int_{-\infty}^{\infty} \hat{f}PE(t)\, dt \qquad (9$$

The energy density in time $E(t)$ (the total power envelope) is the sum of all the power envelopes $fPE(t)$ along all the frequencies f. But pointwise, per each time t, it is the integral over the frequency axis of the power spectral density $tPSD(f)$ for the corresponding time instant.

$$E(\hat{t}) = \sum_{f=-\infty}^{\infty} fPE(\hat{t}) \equiv \int_{-\infty}^{\infty} fPE(\hat{t})\, df = \int_{-\infty}^{\infty} E(t, f|\hat{t})\, df = \int_{-\infty}^{\infty} \hat{t}\, PSD(f)\, df \quad (10$$

Important **convergence considerations** can be deduced. For any continuous-time finite-valued signal, the total power envelope $E(t) = |s(t)|^2$ is a finite function. All the single $tPSDs$ must feature a finite power as well:

$$E(\hat{t}) = \int_{-\infty}^{\infty} \hat{t}\, PSD(f)\, df = finite, \forall\, \hat{t} \qquad (11$$

Each power spectral density, which is positive by definition, must then converge to 0 at the limits for $f \to \pm\infty$.

On the contrary, the power envelope for a given frequency does not have, in general, a bounded integral.

The **energy spectrum**, then, exists finite only for transient or finite duration signals

$$E(\hat{f}) = \int_{-\infty}^{\infty} \hat{f}PE(t)\, dt\,, \forall \hat{f} \qquad (12$$

Nevertheless, the **power spectrum** (average PSD) exists finite as the $tPSD(f)$ are finite:

$$PSD(f) = \langle tPSD(f) \rangle_t = \lim_{T\to\infty} \frac{1}{T} \int_0^T E(t, f)\, dt \qquad (13$$

Then, we can get back to Parseval's theorem in a circular way:

$$\int_{-\infty}^{\infty} PSD(f)\, df = \lim_{T\to\infty} \frac{1}{T} \int_0^T E(t)\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T |s(t)|^2\, dt = \langle |s(t)|^2 \rangle_t = rms^2(s(t)) \qquad (14$$

1.1. Averages of the conditionals: PSD estimation

Once the conditionals are established, the first order moments of these conditionals can be found as:

Estimate via average of following functions	*Pointwise est. via integration*	
$PSD(f) = \langle tPSD(f) \rangle_t = \lim_{T\to\infty} \frac{1}{T} \int_0^T E(t,f)\, dt = \lim_{T\to\infty} \frac{E(f)}{T}$	$\equiv \lim_{T\to\infty} \frac{1}{T} \int_0^T fPE(t)\, dt$	(average) PSD *(15*
$APE(t) = \langle fPE(t) \rangle_f = \lim_{F\to\infty} \frac{1}{F} \int_0^F E(t,f)\, df = \lim_{F\to\infty} \frac{E(t)}{F}$	$\equiv \lim_{F\to\infty} \frac{1}{F} \int_0^F tPSD(f)\, df$	(inst. avg.) APE *(16*

The PSD is defined as the integral average along the time axis of all the following $tPSD(f)$ functions, but it can also be interpreted pointwise, per each frequency f, as the integral average over time of the power envelope $fPE(t)$ for the corresponding frequency.

Indeed, pointwise, per each frequency $\hat{f}$:

$$PSD(\hat{f}) = \langle tPSD(\hat{f}) \rangle_t = \langle \hat{f}PE(t) \rangle_t \qquad (17$$

This has no practical application as the true time-frequency energy distribution $E(t,f)$ is never known a priori. But it can be used to analyse the PSD estimation process.

In fact, the first part $\langle tPSD(\hat{f}) \rangle_t$ can be seen as the ideal Welch's estimate of the PSD. Considering an infinitely short sliding window in time (window centre-time t), the corresponding power spectra are averaged. But exploiting the here introduced considerations, the PSD can also be derived pointwise as $\langle \hat{f}PE(t) \rangle_t$. Considering an infinitely sharp band-pass filter at centre-frequency $\hat{f}$, the corresponding power envelope can be averaged to find the same power density.

Note that the ideal filtering operation, would allow to find a time signal containing a single harmonic contribution $s_f(t)$ whose power envelope corresponds to $fPE(t) = |s_f(t)|^2$.

Then

$$PSD(\hat{f}) = \langle \hat{f}PE(t) \rangle_t = \lim_{T\to\infty} \frac{1}{T} \int_0^T |s_{\hat{f}}(t)|^2\, dt = \langle |s_{\hat{f}}(t)|^2 \rangle_t = ms\left(s_{\hat{f}}(t)\right) \qquad (18$$

where ms stands for mean square.

So, regardless of the true $E(t,f)$, the ideal filtering opens to an alternative PSD estimation process.

From a practical point of view, a finite length T sliding window in time is necessary to perform a Fourier transform. The power spectrum can be computed (squaring the Fourier coefficients) at a finite frequency resolution $\Delta f = 1/T$. The other way around, an equivalent result can be obtained using an ideal filter-bank with a band $B = 2\Delta f$, as found in the literature [3].

This approach leads to a histogram-like approximation such as:

$$PSD(f \mid \Delta f) \cong \frac{1}{B}\langle \left| s_{[f-\Delta f \;\; f+\Delta f]}(t) \right|^2 \rangle_t \qquad (19$$

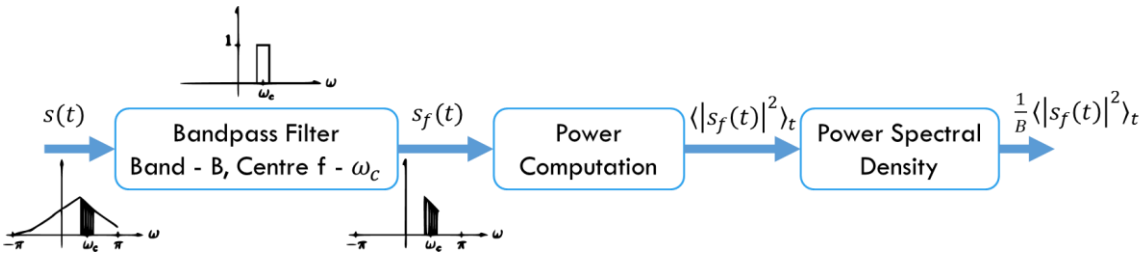


Figure 2: Procedure for filter bank power spectral density estimation [3]

1.2. Hilbert envelope and energy considerations

As introduced, the instantaneous energy of the signal per unit time (the **power**) as a function of time is given by $|s(t)|^2$, and at every time instant must be equal to the integral of the corresponding instantaneous power spectrum $tPSD(f)$.

Because of the nature of the Fourier spectrum, this definition of the power leads to a variable marginal $fPE(t)$ even when the amplitude of the $s_f(t)$ is constant.
For example, given a constant amplitude cosine

$$s_f(t) = A\cos(\omega t) \qquad (20$$

the related power results variable in time

$$fPE(t) = \left| s_f(t) \right|^2 = A^2 \cos^2(\omega t) \qquad (21$$

This can be tedious for amplitude modulate signals, so that, to enhance the modulation function, it can be worth to use a smoothing involving mean square values, able to keep constant the overall energy content:

$$ms\left(s_f(t)\right) = \langle \left| s_f(t) \right|^2 \rangle_t = \frac{A^2}{2} \qquad (22$$

Hilbert transform and analytic signals turn out to be very useful for this purpose.

In fact, considering the corollary of the Euler's formula, it can be written that

$$s_f(t) = A\frac{1}{2}\left(e^{i\omega t} + e^{-i\omega t}\right) \qquad (23$$

From which the analytic signal can be retrieved by discarding the negative frequency component, and doubling the positive frequency component:

$$s_{f,A}(t) = 2\,A\frac{1}{2}e^{i\omega t} = Ae^{i\omega t} \qquad (24$$

This analytic signal turns out to be very useful to smooth the power envelope $fPE(t)$ so as to highlight the possible amplitude modulation. Using half of the squared absolute value of the analytic signal in fact, or the squared Hilbert envelope, a smoothed power envelope is found:

$$fPE_s(t) = smooth\big(fPE(t)\big) = \frac{1}{2}\, abs^2\left(s_{f,A}(t)\right) = \frac{1}{2}s_{f,env}^2(t) \qquad (25$$

This smoothing is particularly effective as it keeps the energy balance:

$$\langle \hat{f}PE(t)\rangle_t = \langle \left|s_{\hat{f}}(t)\right|^2\rangle_t \equiv \frac{A^2}{2} \equiv \langle \frac{1}{2}s_{f,env}^2(t)\rangle_t = \langle \hat{f}PE_s(t)\rangle_t \qquad (26$$

In fact, energy is also conserved under Hilbert transformation [4]. Furthermore, also for non-ideally filtered signals (i.e. when $s_f(t)$ is composed by a band of frequencies around f) these considerations still hold, as the power of a sum of harmonics is the sum of the powers:

$$s(t) = \sum A_i \cos\left(\omega_i t\right)$$

$$\langle fPE(t)\rangle_t = \langle |s(t)|^2\rangle_t = \lim_{T\to\infty}\frac{1}{T}\int \sum A_i^2 \cos^2(\omega_i t) \; + \sum crossterms\; dt \; = \frac{1}{2}\sum A_i^2 \qquad (27$$

$$s_A(t) = \sum A_i \mathrm{e}^{i\omega_i t}$$

$$\langle fPE_s(t)\rangle_t = \langle \frac{1}{2}\, abs^2\big(s_A(t)\big)\rangle_t = \frac{1}{2}\left(\lim_{T\to\infty}\frac{1}{T}\int \sum A_i^2 + \sum crossterms\; dt\right) = \frac{1}{2}\sum A_i^2 \qquad (28$$

So

$$PSD\big(\hat{f}\big) = \langle \hat{f}PE(t)\rangle_t = \langle \hat{f}PE_s(t)\rangle_t \qquad (29$$

1.3. Variance of the conditionals: Spectral kurtosis estimation

Second order moment (variance) of the conditionals can also be found. For example, $PSDvar$, the dispersion of $\hat{f}PE(t) = \left|s_{\hat{f}}(t)\right|^2$ around its expectation, can be seen as

$$\begin{aligned} PSDvar\big(\hat{f}\big) = var\left(\hat{f}PE(t)\right) &= \langle \left|\hat{f}PE(t) - \langle \hat{f}PE(t)\rangle\right|^2\rangle_t = \langle \left|\hat{f}PE(t) - PSD\big(\hat{f}\big)\right| \\ &= \langle \left|\hat{f}PE(t)\right|^2\rangle_t - \left|\langle \hat{f}PE(t)\rangle_t\right|^2 = \langle \left|\hat{f}PE(t)\right|^2\rangle_t - \left|PSD\big(\hat{f}\big)\right|^2 \end{aligned} \qquad (30$$

or

$$\begin{aligned} var\left(\left|s_{\hat{f}}(t)\right|^2\right) &= \langle \left|\left|s_{\hat{f}}(t)\right|^2 - \langle \left|s_{\hat{f}}(t)\right|^2\rangle_t\right|^2\rangle_t = \langle \left|\left|s_{\hat{f}}(t)\right|^2\right|^2\rangle_t - \left|\langle \left|s_{\hat{f}}(t)\right|^2\rangle_t\right|^2 \\ &= \langle \left|s_{\hat{f}}(t)\right|^4\rangle_t - \left|\langle \left|s_{\hat{f}}(t)\right|^2\rangle_t\right|^2 = \langle \left|s_{\hat{f}}(t)\right|^4\rangle_t - \left|PSD\big(\hat{f}\big)\right|^2 \end{aligned} \qquad (31$$

Normalizing on the squared PSD, Moors' interpretation of the kurtosis as a measure of the dispersion of the squared filtered signal (squared envelope) around its expectation can be found:

$$\frac{PSDvar(\hat{f})}{\left|PSD(\hat{f})\right|^2} = \frac{\langle\left|s_{\hat{f}}(t)\right|^4\rangle}{\left|PSD(\hat{f})\right|^2} - 1 = \frac{\langle\left|s_{\hat{f}}(t)\right|^4\rangle}{\left|\langle\left|s_{\hat{f}}(t)\right|^2\rangle\right|^2} - 1 = K(\hat{f}) \qquad (32$$

If $s_{\hat{f}}(t)$ is substituted with its envelope $s_{f,env}$, then, the normalized variance of the squared envelope $var(s^2_{f,env})$ at a given frequency corresponds to the kurtosis of the signal envelope around that frequency, that leads to the common definition of spectral kurtosis via time-frequency representation (e.g. spectrogram).

In brief, the spectrogram of a signal can be seen as the complex envelope of the signal band-pass filtered around the frequency f (ideal, infinitely narrow band filter), and its squared magnitude then indicates the way energy is flowing in that frequency with respect to time. If the frequency band happens to carry pulses, bursts of energy will appear, and this can be detected by computing the excess, normalized fourth order moment of the complex envelope:

$$S(t,f) = STFT(s(t))$$

$$K(f) = \frac{\langle|S(t,f)|^4\rangle}{\langle|S(t,f)|^2\rangle^2} - 2 \qquad (33$$

where the -2 is used to enforce $K(f) = 0$ in case of a complex Gaussian $X(t,f)$.

From a practical point of view, STFT gives a finite frequency discretization Δf which is function of the time window length. Then, the complex envelope becomes a function of this additional parameter: $S(t,f,\Delta f)$. Computing the spectral kurtosis for different frequency discretizations ($K(f,\Delta f)$) and summarizing it as a colormap in the $[f,\Delta f]$ plane, the Kurtogram is built. As already introduced, the main weakness of this procedure is the difficulty of STFT in obtaining a good discretization in both frequency and time domain, so that a different approach was developed by Antoni [8]. In order to find the kurtosis in all the required frequency bands, the signal is processed by a quasi-analytic FIR filter bank producing a division of the $[f,\Delta f]$ plane (paving). This paving was originally dyadic but was later improved to a 1/3 binary tree division, which can better cover the frequency axis. The procedure finally obtained takes the name of **Fast Kurtogram (FK)**.

Several improvements have been proposed over the years, but no one proved to be as reliable and computationally efficient as the FK.
An alternative is proposed in next section, derived from the considerations here introduced.

2. Spectral kurtosis as a sliding filter

Despite its efficient implementation, the Fast Kurtogram has the limit of computing the spectral kurtosis only on the discrete grid generated by the dyadic or 1/3 binary tree paving.

Nevertheless, the considerations here introduced allows to make a parallel of this methodology with the one introduced for filter-bank estimation of the PSD. In each region of the frequency domain defined as a frequency band $B_f = [f - \Delta f \;\; f + \Delta f]$, the PSD and the Spectral Kurtosis can be estimated as

$$PSD(B_f) \cong \frac{1}{B_f}\langle \left| s_{B_f}(t) \right|^2 \rangle_t \qquad (34$$

$$K(B_f) \cong \frac{\langle \left| s_{B_f,env}(t) \right|^4 \rangle}{\left| \langle \left| s_{B_f,env}(t) \right|^2 \rangle \right|^2} \qquad (35$$

The main issue with this methodology is that, as Δf increases, the variance in the PSD estimation is lowered, but frequency resolution decreases, leading to a trade-off between statistical accuracy and resolution.

Furthermore, it is necessary a complex optimization for designing the best filter which passes the desired frequency (band) the more undistorted as possible in order not to compromise the estimate. For the PSD estimation, the Slepian baseband filters and the Capon method can be found in the literature [3].

To overcome these limits, breaking the accuracy resolution trade-off, a sliding filter can be used.

Substituting the discrete paving with a sliding filter in fact, smoothed versions of the true PSD and true SK can be found, as the result of the application of a symmetric kernel average smoother (e.g. a symmetric moving average filter is a kernel smoothers that use a rectangular kernel) defined by the shape of the used filter in the frequency domain.

Selecting a FIR band-pass filter of bandwidth B_f and center frequency f and sliding this filter along the frequency axis, in fact, an approximation of the PSD and SK trends can be found. Additionally, decreasing the bandwidth B_f, a series of more detailed approximations of SK can be produced. Finally, $\arg\max SK(f, B_f)$ defines in the same way the best filter, but more freedom is left to the centre-frequency, at a reasonable additional computational cost. The order of the FIR filter used in FK is not extremely high [5], so that in the proposed SK* order 16 was used.

3. Comparison over synthetic data

Synthetic data was used at first to assess the performance of the proposed estimator against a selection of opponent algorithms introduced in Chapter 4 for the enhancement of bearing signature and reported in Figure 3.

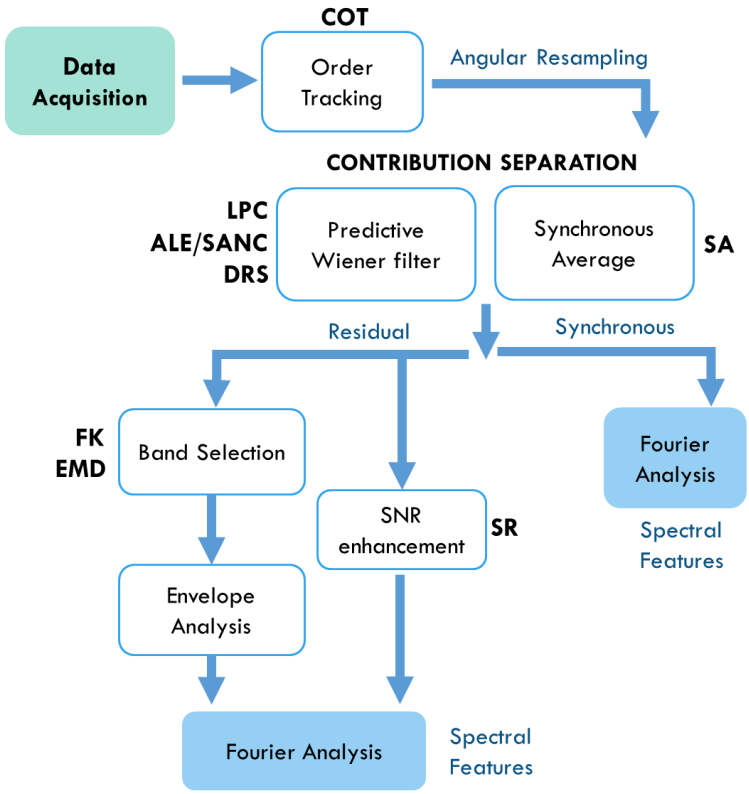


Figure 3: Signal Processing procedure

The comparison was performed on 4 different synthetic signals generated using the model described in Chapter 5 for increasing noise levels, as shown in the following Table 1.

In particular, both the input noise (upstream of the transfer path) and the measurement noise were varied, to simulate different masking of the bearing signature.

In this particular case, the simulated bearing damage regards the inner race of a SKF 6006-Z as previously described, rotating at a constant speed of $fr = 40\ Hz$, which leads to a theoretical damage frequency (BPFI) of $f_{BPFI} = 251{,}4\ Hz$.

Table 1: Noise levels for the simulated signals from the model described in Chapter 5.

Signal id #	Input Noise	Measurement noise
1	5	1
2	10	10
3	5	5
4	5	3

In order to characterize and compare the performance of the different algorithms in enhancing the bearing signature, some evaluation parameter is needed. At this scope, two performance indicators in the frequency domain are introduced.

The first is the amplitude of the normalized power spectrum P corresponding to the damage frequency (BPFI), or

$$NA_{BPFI} = P_{BPFI}/\max(P) \quad (36$$

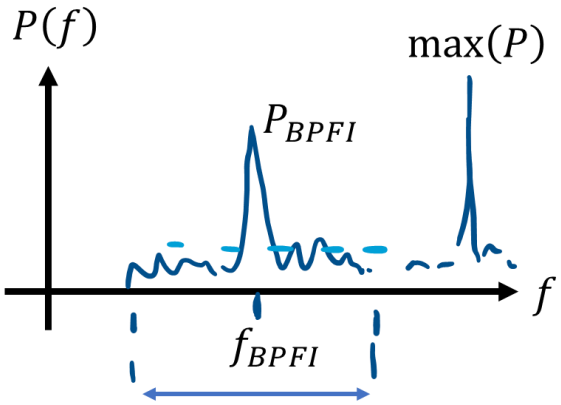


Figure 4: Performance parameter – Normalized power spectrum

The second is the signal to Noise Ratio of the damaged spectral line with respect to the surrounding noise level calculated as the 97th percentile of the distribution of surrounding spectral lines in a range of $\pm 5\ Hz$:

$$SNR_{BPFI} = P_{BPFI}/P_{97\%} \quad (37$$

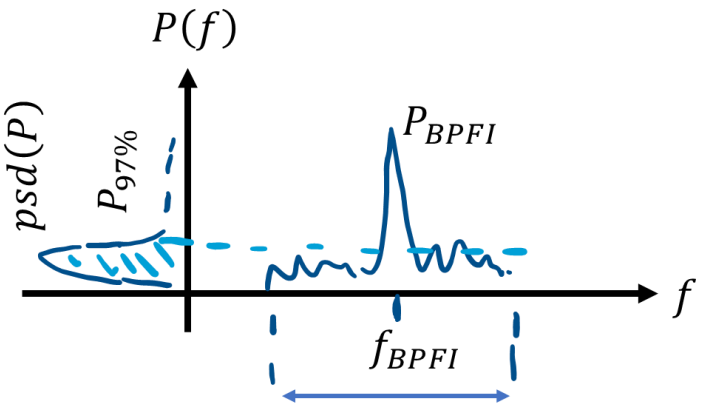


Figure 5: Performance parameter – Signal to Noise Ratio from the power spectrum

The selected algorithms have been tested on the overall raw signal, as well as on the residual signal after LPC and after TSA.

The optimal order selected for LPC was 20, while angular resampling was also performed to obtain an exact number of samples per cycle. The original sampling frequency was in fact $fs = 22528\ Hz$. Since the rotational speed was set to fr, this leads to a non-entire number of samples of

$$\frac{fs}{fr} = 563{,}2 \quad (38$$

An entire number of samples per cycle of $SPC = 550$ was then selected, and the original signal was resampled in the angular domain (i.e. at constant angular increments). In this particular case featuring a constant speed fr anyway, this corresponds to a resampling at a given, reduced, sampling frequency $fs_1 = SPC * fr = 22000\ Hz$.

The result of this Computed Order Tracking pre-processing followed by the Synchronous Average (OT+SA) is a systematic improvement of the bearing signature as both the performance indicators (SNR and NA) always increase. This means that the isolation of the non-deterministic components by removal of the periodic contribution obtained via SA is already enhancing the bearing damage signature. This is not the same for LPC which cannot ensure the same improvement with respect to the raw signal (i.e. no OT). This result is highlighted in Table 2.

Table 2:

Signal # and parameter		**RAW**	**LPC**	**OT+SA**
1	SNR_{BPFI}	3,4	3,2	4
	NA_{BPFI}	0,0033	0,0029	0,15
4	SNR_{BPFI}	2,5	2,8	4,7
	NA_{BPFI}	0,0035	0,0028	0,16
3	SNR_{BPFI}	2,7	3	3,5
	NA_{BPFI}	0,0028	0,0028	0,17
2	SNR_{BPFI}	2	2	2
	NA_{BPFI}	0,004	0,0022	0,18

The scope of this analysis was anyway to compare the performance of the selected algorithms (the Fast Kurtogram FK, the Empirical Mode Decomposition EMD, the Stochastic Resonance SR and the Improved Spectral Kurtosis estimator SK*) on the synthetic dataset.

The results of the application of the different selected algorithms on the raw signal and downstream of LPC and of OT+SA are reported in the following Table 3 in terms of the declared performance indicators SNR and NA.

It can be seen from Table 3 that when the raw signal is used, the proposed SR* algorithm is the only able to enhance the bearing signature. In particular, the NA is always improved, while the SNR is in general improved omitting the Signal 2 for which the original SNR of 2 is slightly decreased.

Considering the LPC pre-processing, no algorithm is able to output acceptable results from Signal 2, while FK outperforms the others for Signal 1 (which however features the lowest noise level), SR is the best for Signal 4, but only SK* ensures stable results on all the three Signals 1, 4 and 3. In particular, Signal 3 shows stronger noise contamination, and in this case SK* turns out to be the best.

Table 3:

	RAW		LPC		OT+SA	
	SNR_{BPFI}	NA_{BPFI}	SNR_{BPFI}	NA_{BPFI}	SNR_{BPFI}	NA_{BPFI}
Signal 1:						
FK	1	-	6	1	8,2	0,95
SK*	5,4	1	4,85	0,98	3,7	0,75
EMD	1	-	4,5	0,93	4,5	1
SR	2,5	0,0007	4,2	0,16	3,75	0,97
Signal 4:						
FK	1,2	-	1	-	6	0,6
SK*	3,2	0,7	2,9	0,62	3,6	0,75
EMD	1,4	-	3,2	0,69	2,9	0,6
SR	1,2	-	3,3	0,067	4,4	0,6
Signal 3:						
FK	2	0,006	3,1	0,8	1,5	-
SK*	3,75	0,87	3,8	0,81	4,4	0,76
EMD	1,2	-	2,4	0,8	2,2	0,7
SR	1,6	-	3	0,04	3,2	0,6
Signal 2:						
FK	1	-	1	-	1	-
SK*	1,65	0,8	1	-	1	-
EMD	1	-	1	-	1,7	0,7
SR	1,7	0,001	1,4	0,018	1,9	1

When OT+SA pre-processing is taken into account, similar considerations can be made. In particular, neglecting Signal 2 which is considered separately, SK* proves to produce acceptable results on all the three Signals 1, 4 and 3, even if on Signal 1 it is outperformed by FK and on Signal 2 both FK and SK lead to better results. When OT+SA is performed on Signal 2 on the contrary, only EMD and SR give good results.

To conclude, FK is proved to be more effective and efficient than SR and EMD for finding the best demodulation band for envelope analysis to highlight the bearing damage signature. Furthermore, the FK improvement via sliding filter ensures

SK* ensures consistent performance improvements of the bearing signal and can be considered more effective than FK in finding the best demodulation band as the discretization of the filter centre frequency can be arbitrarily small in the frequency domain.

To further prove the success of SK*, in the next section a further comparison is reported on a real-life data from an aeronautical gearbox. Both the algorithms for deterministic/random separation and the ones for signature sharpening will be tested.

4. Application and comparison of the algorithms on real-life data: SAFRAN civil aircraft engine gearbox from Surveillance 8 contest

The Safran contest data from Surveillance 8 conference was used to test the algorithms. The provided data consists of vibration and tachometer signals acquired during a ground test campaign on a civil aircraft engine with two damaged bearings.

As described in Chapter 5, the engine has two main shafts and an accessory gearbox for the equipment. The diagnostic problem of the contest was to assess the presence of possible damages on the bearings given accelerometric acquisitions at variable speed. The procedure introduced in this work and reported in Figure 3 is then followed, so that a comparison between EMD, SR, FK and SK* is possible.

4.1. Computed Order Tracking and Synchronous Average

A practical implementation adopted for the Surveillance 8 Contest can be found in [7] where the acquisition relative to the SAFRAN engine (Chapter 5) is the object of bearing diagnostics. Five shafts (HP and L1 to L4 in Figure 8) need to be taken into account in this case. At first, STFT is applied on the signal to verify that the speed is not constant in time.

A quite accurate speed profile can be recovered by tracking the harmonics visible in the spectrogram. Obviously, at low frequency (e.g. shaft frequency, about 250 Hz) the relative estimation error due to the frequency discretization can be huge. It is then preferable to track the sharpest high frequency peaks. For example, the 38th and the 75th harmonics of the shaft, highlighted in Figure 6 can be proved to produce very accurate reconstructions of the instantaneous angular speed (IAS) for the machine shut down (decreasing speed). In any case a tachometer is available, then, despite the possible geometric errors, the IAS can be recovered from the time difference between two following trigger passage times. The recovered speed for a machine start-up (increasing speed) is shown in Figure 7.

When the speed is variable, as in this case, the issues of an asynchronously sampled vibration signal should be faced. A computed order tracking is then required and implemented thanks to the tachometer signal. A first deterministic/non-deterministic separation is obtained by Synchronous Average. COT+SA is then applied in cascade for each shaft to remove all the periodic components, isolating the very last residual containing the bearing contribution which, as in most of cases, cannot be seen directly in the residual spectrum, but requires Envelope Analysis.

The cascade of COT+SA for each of the 5 shafts taken into account is reported in Figure 8. Focusing on the residuals, it is easy to see how most of the main spectral lines related to gear characteristic orders are gradually removed.

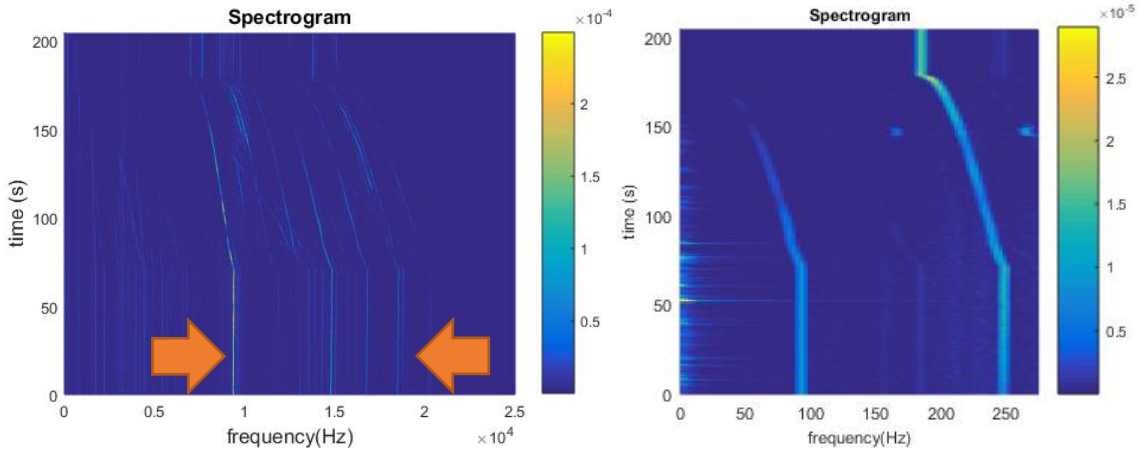


Figure 6: Spectrogram of the given vibration signal and zoom in the low frequency region. The main shaft frequency is pointed by the yellow arrow, while the 38th and the 75th harmonics of the main shaft are in red. – Window length: 49152, Step: 8192.

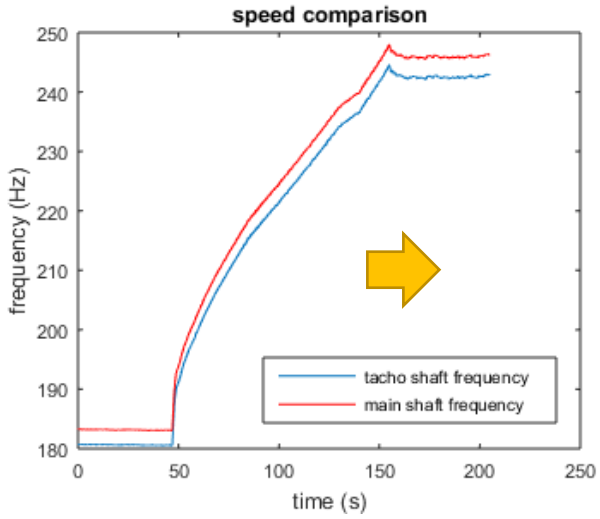


Figure 7: IAS recovered from the tacho trigger passage times. Knowing the transmission ratio, the speed of the main shaft (HP) can be easily computed from the tacho shaft speed.

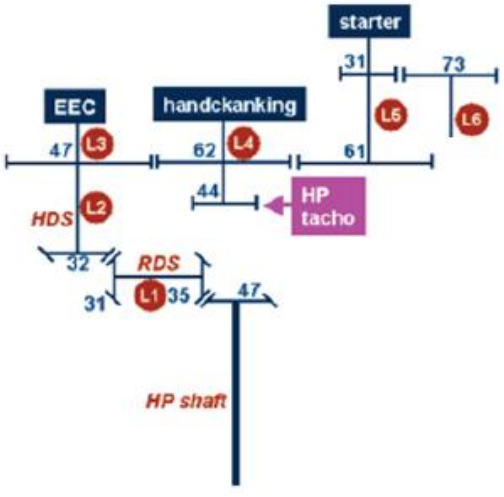


Figure 8: Part of the scheme of the Surveillance 8 SAFRAN contest gearbox.

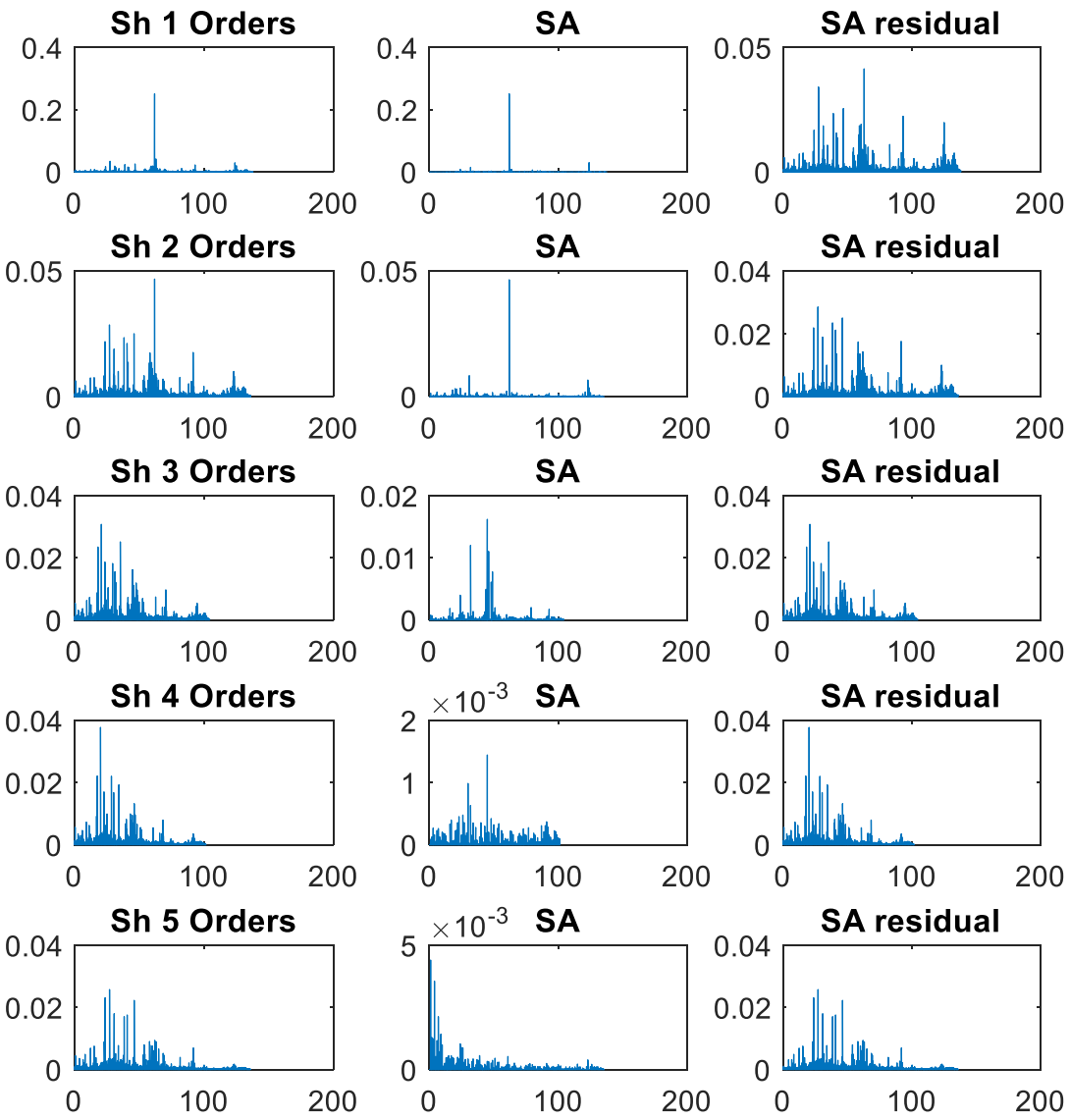


Figure 9: Cascade COT and SA for the 5 considered shafts. Shafts 1 to 5 corresponds to shafts L4, L5, L2-3 , L1, HP.

4.2. Deterministic/non-deterministic separation

The same deterministic/non-deterministic separation performed via OT+SA can also be performed using the predictive algorithms presented in Chapter 4. A comparison of the effect of the different filters on the time domain signal is shown in Figure 10. Taking the SA signal as a reference, the Mean Square Difference from SA $MSD^{SA} = E[|d^{SA}(n) - d^{P}(n)|^2]$ is computed for the LPC with order 1000, for a SANC of analogous order and forgetting factor $\mu = 10^{-3}$, and for a DRS using a Parzen window of length $N = 1024$ and equivalent delay, step and filter length M. The computational times are also recorded.

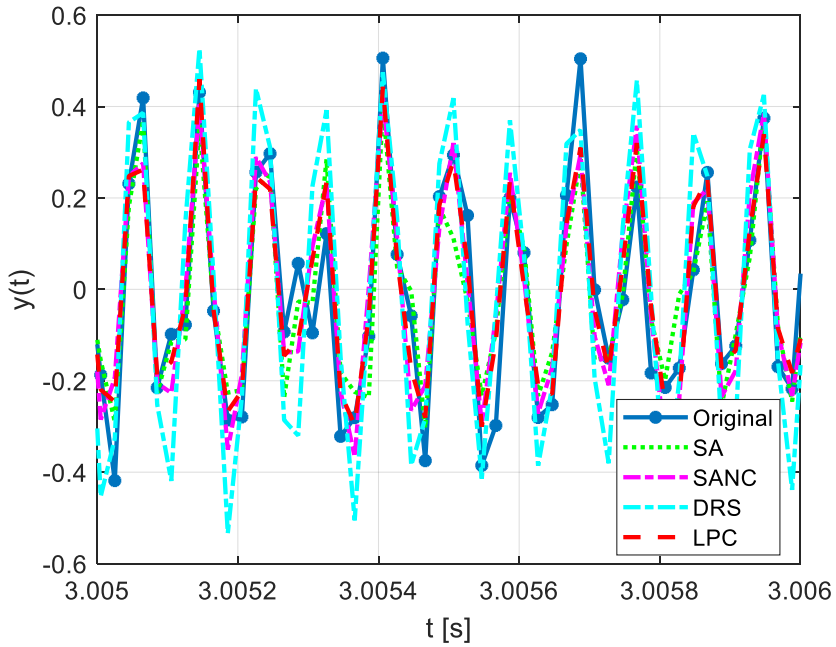


Figure 10: Comparison of the different predictive algorithms and SA. In particular LPC refers to an AR(1000), SANC has order 1000 and forgetting factor $\mu = 10^{-3}$, DRS uses N=M=D=Step=1024.

An accurate parameters optimization for the different algorithms is out of the scope of this work. Conversely, similar parameters were set, so as to ensure the comparability of the results in terms of MSD and relative computational time. According to these two assessment variables summarized in Table 4, in this particular application, the LPC results the best both in terms of prediction error and computational effort. When the filter order is not too large in fact, LPC provides a closed form solution corresponding to the optimal Wiener filter. Furthermore, no additional parameters are needed, unlike the other predictive algorithms.

Table 4: Mean Squared Difference from SA and relative computational times for the different predictive algorithms

	LPC	***SANC***	***DRS***
$\boldsymbol{MSD^{SA}}$	0,009	0,013	0,022
$\boldsymbol{t/t^{OT+SA}}$	0,25	0,5	0,25

Switching to the frequency domain, further comments can be made about LPC and SA. SA, in fact, ensures the minimum disruption of the residual signal but requires a repetition for each shaft in cascade. In this particular example, only the main shafts

contributions were removed (the accelerometers were known to be near to shafts 1 to 5) but the whole gearbox consisted of more than 12 shafts (chapter 4, paragraph 3.2, Figure 7), so that only the major deterministic contributions are identified, while the minor ones are likely to be missed. This is reflected by the amplitude spectrum reported in Figure 11. Indeed, as it is easy to notice, LP recognizes as deterministic additional spectral lines with respect to SA, despite not extracting all the deterministic frequencies identified by SA. This highlights why predictive algorithms are preferable in case of complex gearboxes.

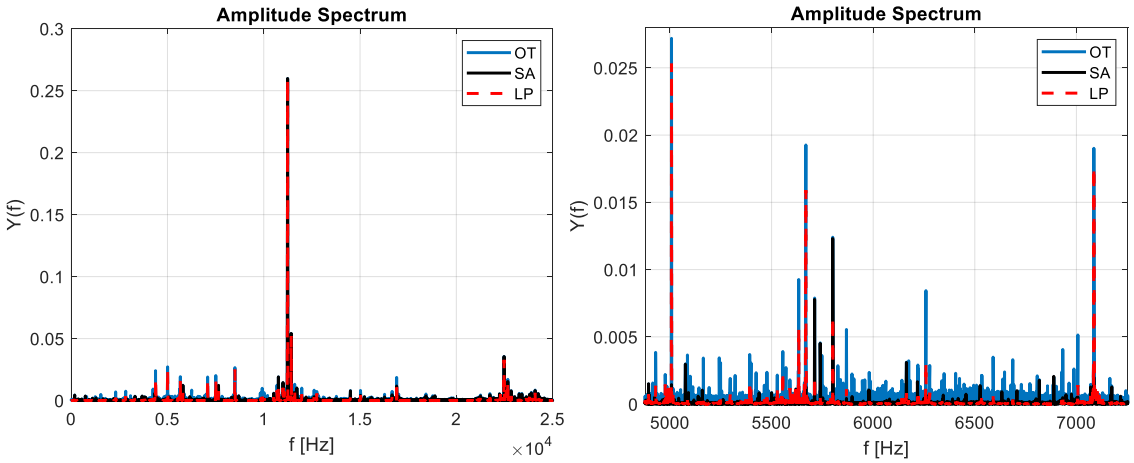


Figure 11: Amplitude spectrum of the predictive algorithms applied to the first 10 seconds of Acc2 signal (at constant speed - see Figure 7).

4.3. Kurtogram, sliding window spectral kurtosis SK* and a novel envelope analysis visualization via Envelope Spectrogram

The SAFRAN contest core requirement was to check whether one or more of the bearings supporting shafts L1, L4 and L5 were damaged. The difficulty was related to the complexity of the gearbox, featuring many shafts and supports, and running at a variable speed. To highlight the bearing signal then, COT and deterministic part removal (e.g with LPC) are fundamental to produce a residual signal. The residual contains only the bearing vibration covered with noise and resampled at constant angular increments of a reference shaft (e.g. the L5 shaft in this case). For diagnostic purposes, then, the bearing defect frequencies should be computed taking as a reference the L5 shaft frequency. Table 5 summarizes the so obtained features in the order domain (adimensionalized frequencies).

Table 5: Bearing characteristic frequencies for 3 bearings. The identifying colour code is also defined.

Bearings:	***FTF***	***BSF***	***BPFO***	***BPFI***	*colour*
L1 (*main*)	0.407	2.584	4.066	5.934	Green
L4 (*tacho*)	0.409	2.515	4.087	6.054	Red
L5	0.430	3.548	7.741	10.22	Blue

Thereafter, envelope analysis can take place. It is always worth to try EA on the raw residual signal, as in many cases the damage features are already emphasized. Unfortunately, this is not the case. Focusing on Figure 14 both the raw residual spectrum and the envelope spectrum of the raw residual (first 10 seconds at constant speed – see

Figure 7) are reported. As it is easy to notice, none of the possible defect frequencies are rising from the background noise.

The STFT-based kurtogram and the Fast Kurtogram should be then implemented, to find the most suitable band for envelope demodulation. Figure 12 shows the kurtogram of the signal and the selected band in bright yellow (about 0-11 orders). In order to better visualize the kurtosis as a function of frequency, the spectral kurtosis can be also computed with both the STFT algorithm and the sliding filter algorithm previously introduced. Despite the different levels, the two trends are comparable, proving the effectiveness of the proposed simplification. In terms of computational times, the Fast Kurtogram results 3 times faster than the traditional STFT-based kurtogram. The sliding filter SK* computational burden is also reduced with respect to STFT-based spectral kurtosis, as highlighted in Table 6.

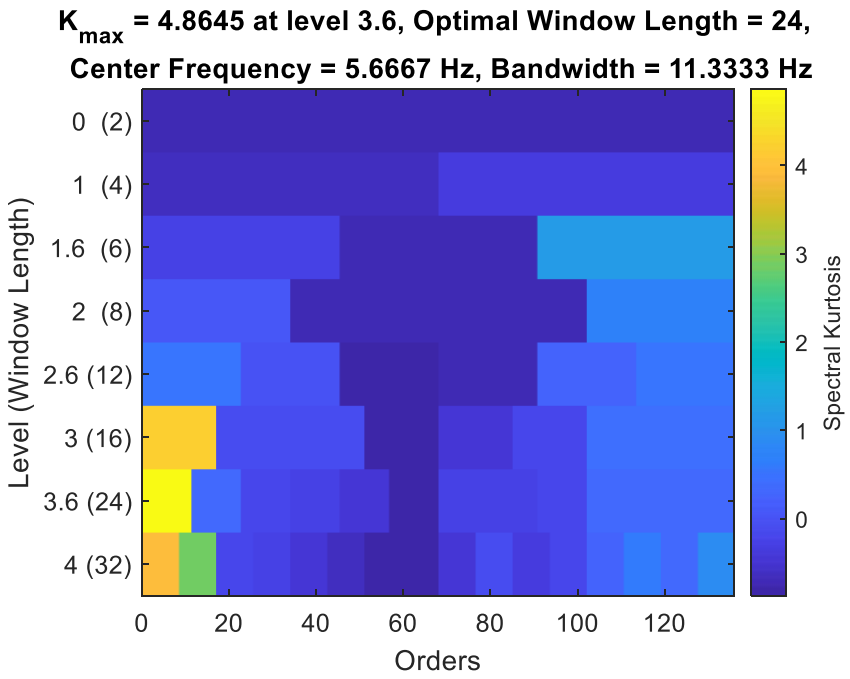


Figure 12: Kurtogram of the LPC residual.

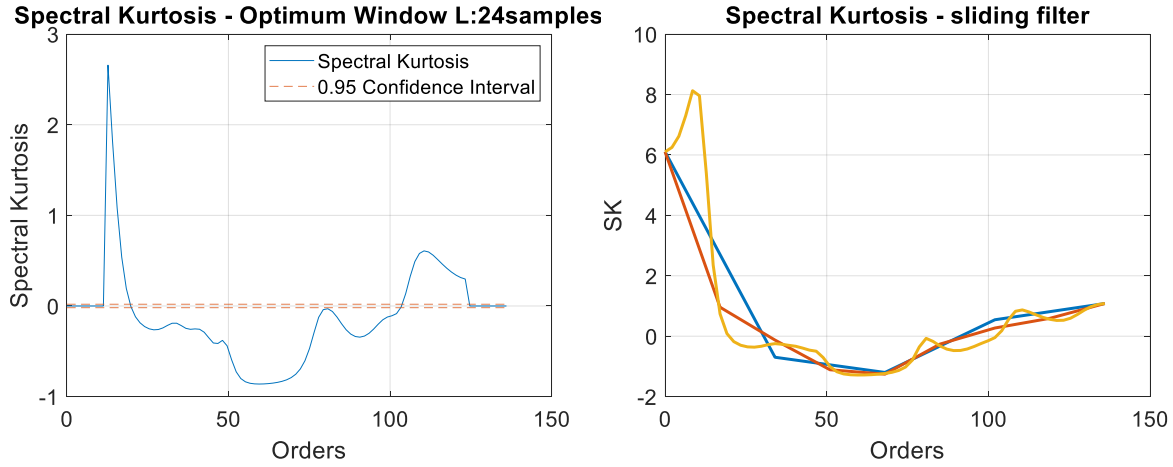


Figure 13: Spectral kurtosis – results from STFT-based algorithm (left -L=24) vs sliding-filter-based algorithms (right – $\frac{1}{2}, \frac{1}{4}, \frac{1}{32}$ *of Nyquist frequency).*

Table 6: Comparison of STFT-SK, STFT-FK, FK and SK in terms of computational times.*

	STFT based kurtogram	*Fast Kurtogram*
Computational time [s]	0.97	0.30
	STFT Spectral Kurtosis	*Sliding filter SK**
Computational time [s]	4.0	3.2

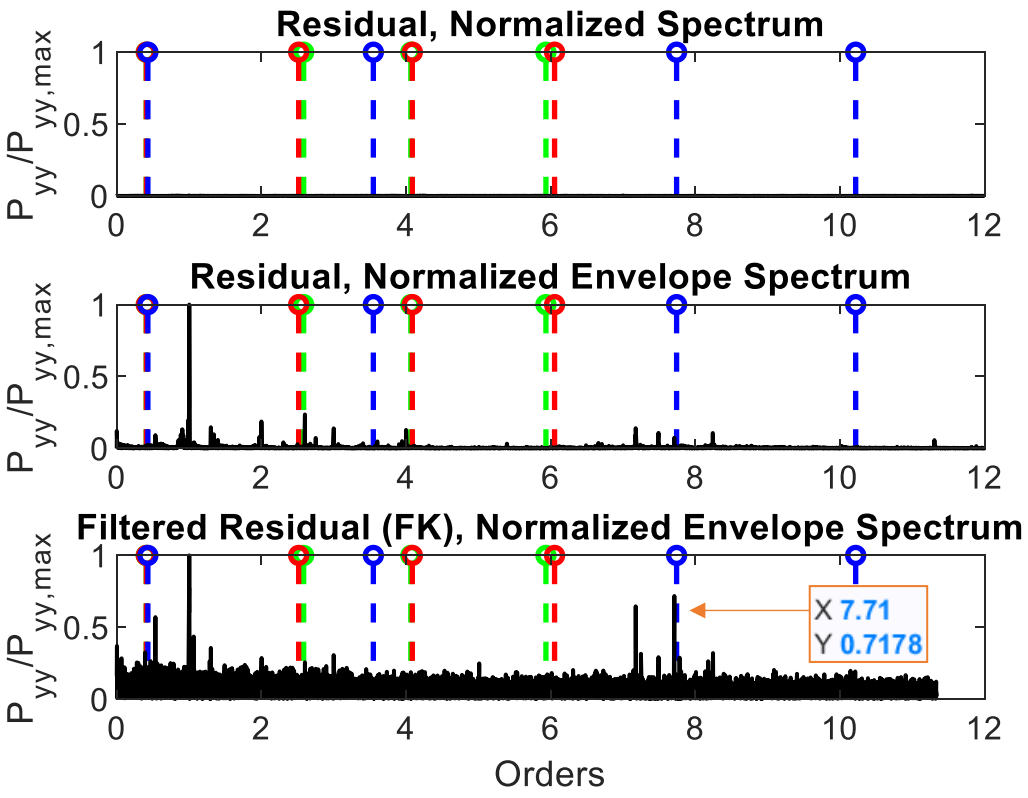


Figure 14: Normalized spectra of the LPC residual and of EA in orders of main shaft, for a window of Acc2 from 0s to 10s. In green L1, in red L4, in blue L5 bearing frequencies (Table 6). The L5 BPFO is highlighted. The result of EA on the signal band filtered according to the FK-selected band is reported in the third graph.

4.3.1. Novel envelope analysis visualization: Envelope Spectrogram

Considering that the overall signal is very long but the speed is changing over time (see Figure 7) it is wise to divide the signal in several shorter chunks and perform the envelope demodulation on each, fixing a constant band of interest B on the basis of prior knowledge (e.g. possible resonances) or on tests with FK or SK* on some chunks of the signal. The envelope spectra computed on all the chunks generated by a sliding time window can later be "packed" one after the other to form a surface in a spectrogram like manner.

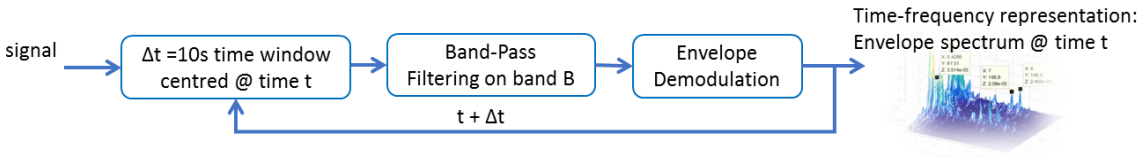


Figure 15: Envelope spectrogram block scheme

In SAFRAN dataset two accelerometers are available, so that two envelope spectrograms can be obtained. In any case, as it is possible to notice in Figure 16, Acc1 signal mainly contains the harmonics of the shaft. Furthermore, its overall amplitude is significantly lower than Acc2 channel, so that it is better to focus the analysis on this second signal, whose sensor is proved to be located nearer to the damage position. In Acc2 envelope spectrogram, despite a slight difference attributable to the common phenomenon of rolling elements slip, two bearings features are highlighted: 0.43 and 7.7, and the second is compatible with the L5 BPFO order (7.741 - Table 6). Hence, this feature denotes the presence of a damage on the outer race of shaft L5 support bearing.

Additionally, these *envelope spectrograms* prove to be very useful to put in relation the damage detectability with the operational speed of the engine. Indeed, Figure 16 shows that the L5-BPFO amplitude increases in time, while the speed is increasing.

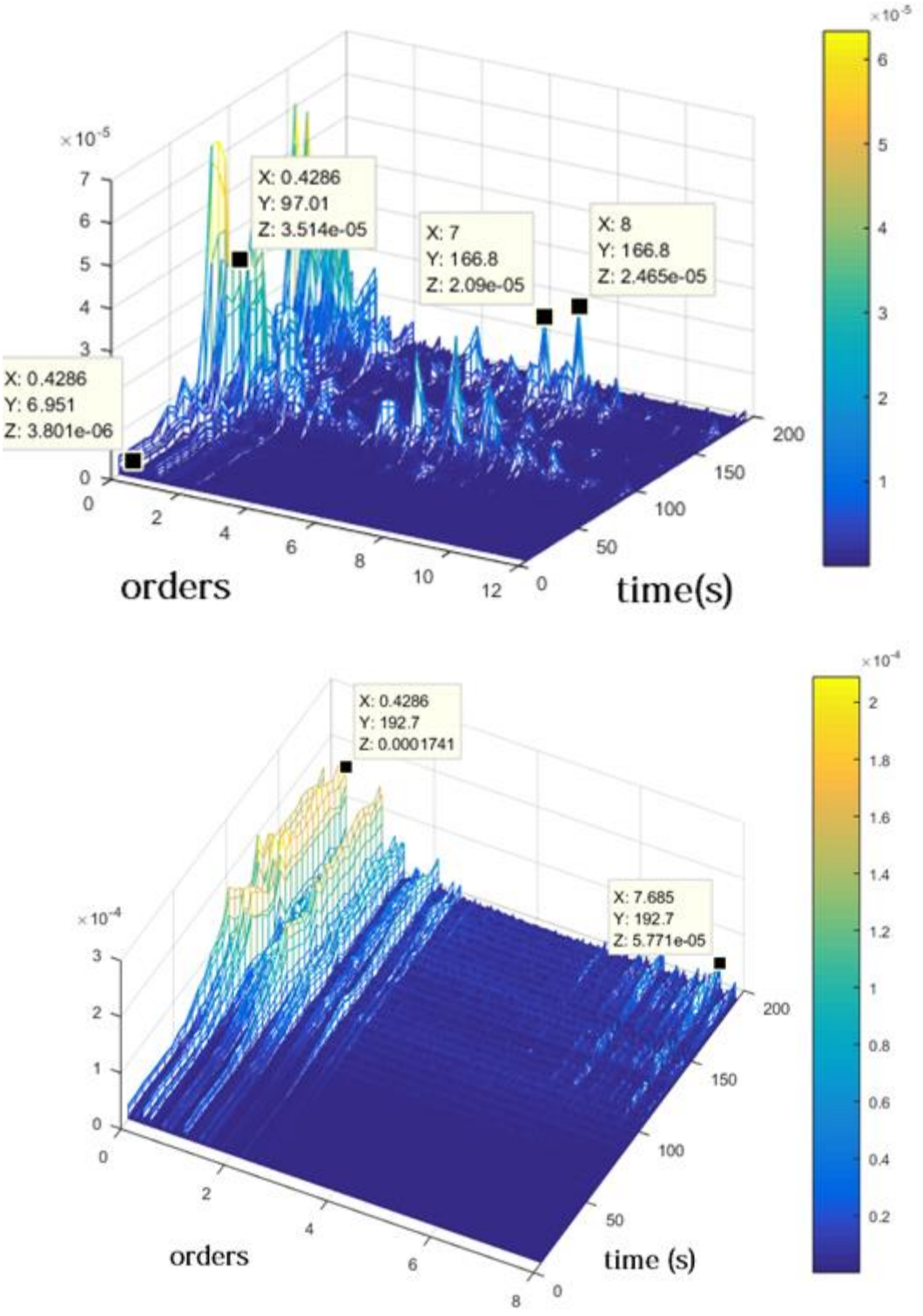


Figure 16: Envelope spectrograms of Acc1 (up) and Acc2 (down) signals – Window length: 512000, Step: 256000. [7]

4.4. EMD and envelope analysis

The envelope demodulation can be repeated using EMD for selecting the more appropriate band. The residual signal from OT and deterministic/random separation, can be then decomposed into several IMFs featuring different kurtosis levels. The first six IMFs obtained by treating with traditional EMD the OT+LPC residual signal are reported in Figure 17. As it can be easily noticed, the second IMF is the one featuring the highest excess kurtosis (around 1,83). Focusing on its spectrum, one can notice IMF-2 is almost equivalent to the signal band-pass filtered in the range 25-50 orders, far away from the band highlighted by the kurtogram (about 0-11 orders). In any case, the Envelope Analysis performed on IMF-2 proves to be very effective in highlighting the L5 BPFO order (7.71). Figure 18, in fact, shows that the SNR of such a bearing feature is much higher than that produced by EA in the kurtogram-selected band (Figure 14).

Table 7: EMD computational time in seconds

	EMD
Computational time [s]	0.30

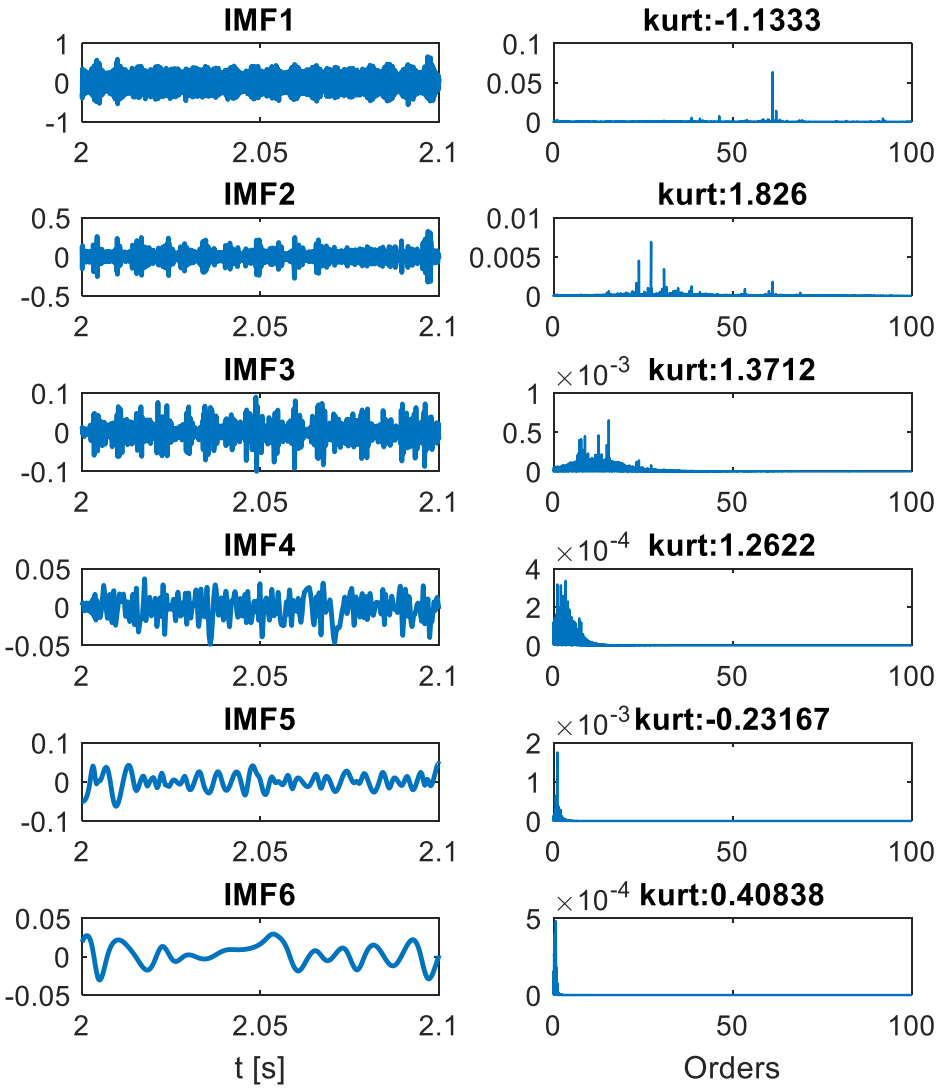


Figure 17: IMFs from EMD of the LPC residual of Acc2 from 0s to 10s. The kurtosis of each IMF is also reported.

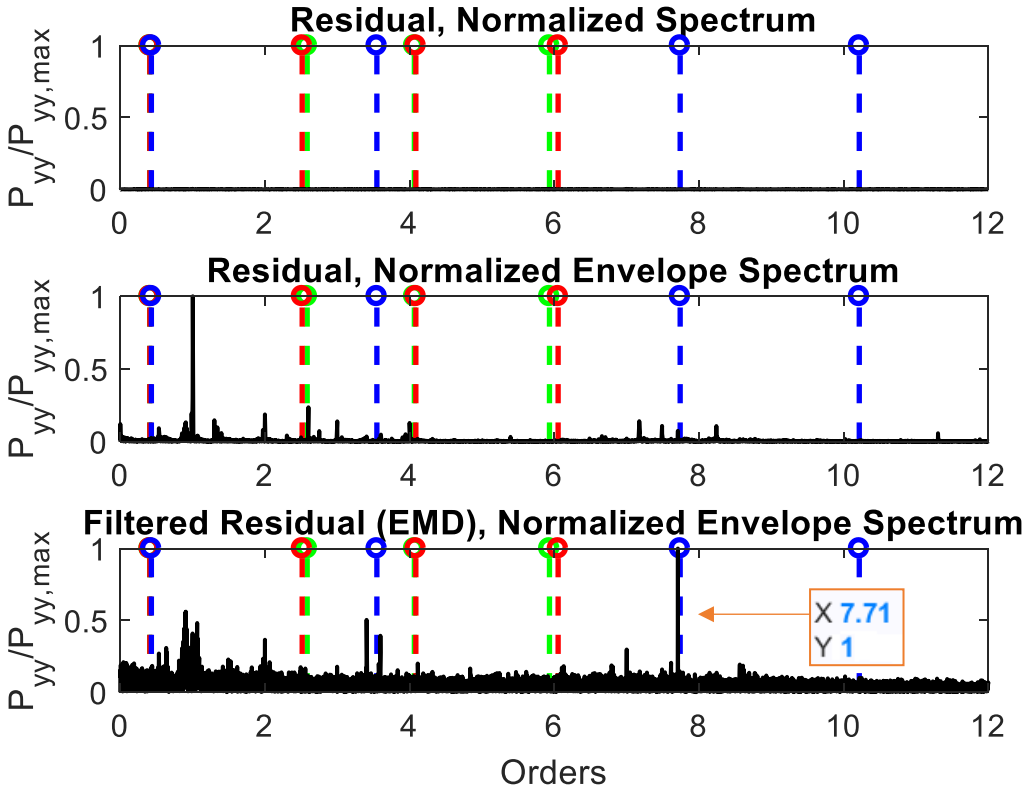


Figure 18: Normalized spectra of the LPC residual and of EA in orders of main shaft, for a window of Acc2 from 0s to 10s. In green L1, in red L4, in blue L5 bearing frequencies (Table 6). The L5 BPFO is highlighted. The result of EA on the EMD computed IMF2 is reported in the third graph.

4.5. Stochastic Resonance

The analysis diagnostic analysis is repeated switching from EA-based algorithms to stochastic resonance. The residual signal from OT and deterministic/random separation (OT+LPC) is then used as input to the SR non-linear differential equation, whose parameters a, b, R and s are selected via optimization with a Genetic Algorithm (Appendix 5) using the snr optimization criterion. The result is reported in Figure 20. With the parameters $a = 89$, $b = 235$, $R = 575$ and $s = 5$ the SR can really highlight the L5 BPFO order (7.71) with respect to the background level, even if its amplitude is just the 5% of the highest spectral line which is not depicted as occurring at order 61. It is relevant to consider that the GA, despite being very effective, has a limited computational efficiency as it must recall the SR function for a number of times determined by the population size by the number of generations needed to carry out the optimization.

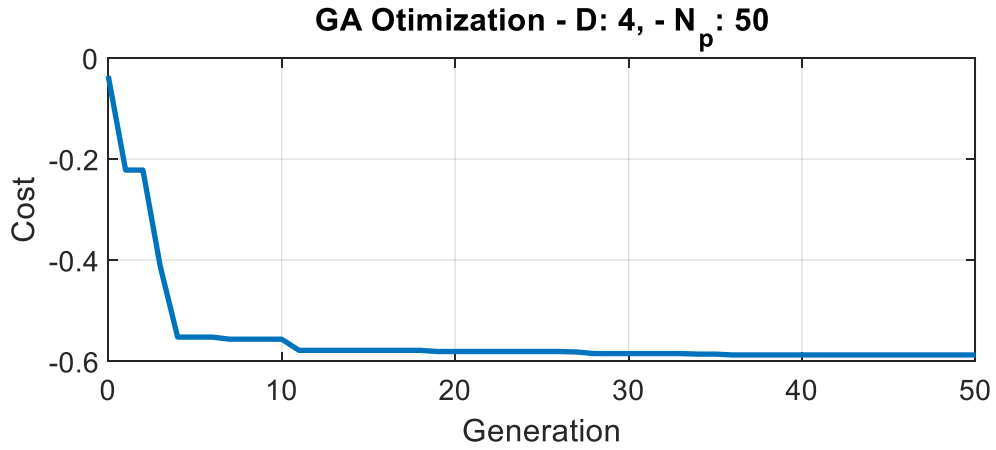


Figure 19: GA generations summary for the optimization of the four SR parameters on the snr criterion.

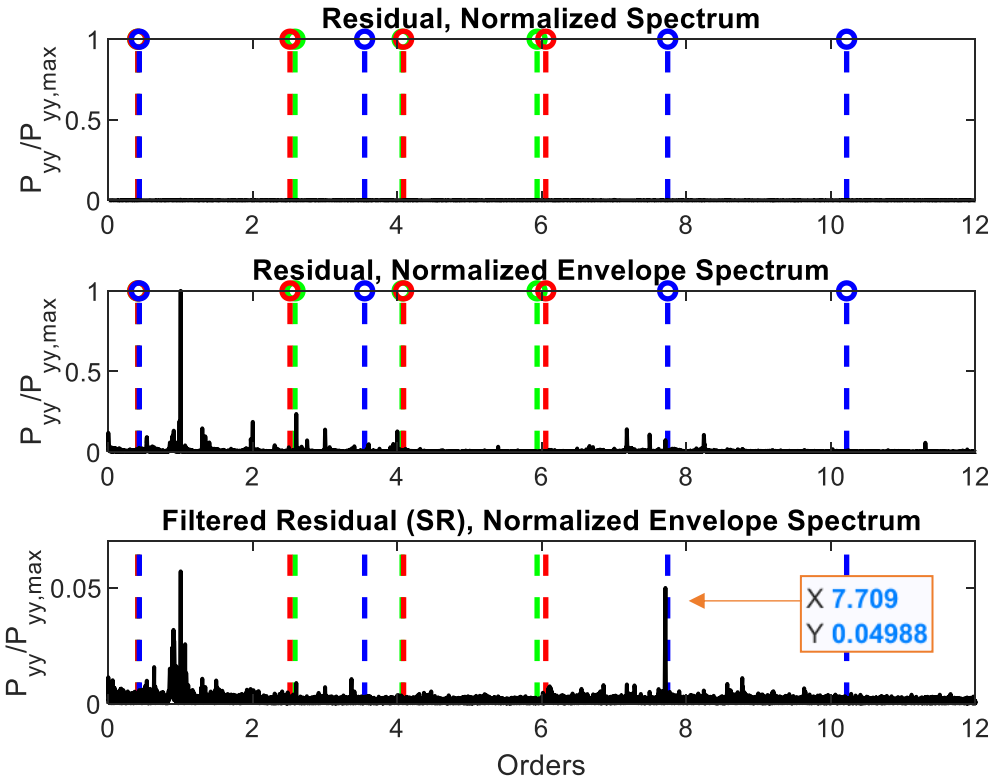


Figure 20: Normalized spectra of the LPC residual and of EA in orders of main shaft, for a window of Acc2 from 0s to 10s. In green L1, in red L4, in blue L5 bearing frequencies (Table 6). The L5 BPFO is highlighted. The result from SR given R=575, a=89, b=235, s=5 is reported in the third graph. Notice the different y-axis scale.

Table 8: SR+GA computational times in s.

	SR	*GA*
Computational time [s]	0.15	380

5. Conclusions

In this chapter the proposed intermittent monitoring methodology was first tested on a synthetic dataset generated with different noise contamination levels. The novel SK* algorithm was then compared to the reference algorithms individuated in the state of the art. Then, a second test of the methodology was conducted on the SAFRAN Contest data from Conference Surveillance 8 held in Roanne, France on October 20&21 2015.

In both cases, two evaluation parameters were used to assess and compare the performance of the bearing signature sharpening: the normalized amplitude of the spectral line corresponding to the bearing damage frequency and the signal to noise ratio of the same spectral line.

According to these parameters, the SK* proved to combine the advantages of the Fast Kurtogram in terms of time efficiency and effectiveness in the damage identification, removing the drawback of a poor spectral discretization (the paving of the fast-kurtogram). The analysis showed that the results found using the SK* were more stable, as in almost all the noise condition, the algorithm was able to find good results. It was not the same for FK, whose performances were sometimes outmatched by SR or EMD.

The SAFRAN data about an aeronautical engine gearbox opened to the possibility of testing the methodology on a real-life application of interest. In accordance with the contest requirement, the focus was primarily on bearings, so that after the first stage of COT and contribution separation, the analysis prosecuted on the residual signal.

In the first stage, the traditional synchronous average algorithm was compared to the most common prediction-based algorithms. The analysis confirmed the ability of SA in extracting the deterministic component with minimum disruption of the residual signal but underlined at the same time the deeper amount of geometric information needed to perform the analysis and the longer computational times required in case of complex gearboxes (just as the SAFRAN gearbox).

In the second stage, envelope analysis was performed on selected bands of the residual obtained via linear predictive coding (LPC). Both SK*, FK and EMD bands selected on the basis of the excess kurtosis index proved to highlight the damage, even if in this particular case, despite the lack of physical motivation, EMD demonstrated to be an effective alternative. Finally, SR was also implemented to denoise the low frequency region, highlighting the bearing signature. Exploiting the knowledge of the characteristic frequencies of interest, the snr-based cost function was selected to optimize the SR parameters via genetic algorithm. The result is again the same. In all the three cases in fact the algorithms are able to highlight the presence of the outer race damage (BPFO frequency) of the bearing supporting the L5 shaft.

In general, despite the objective comparison of the different algorithms on the synthetic dataset, it is impossible to be sure that an algorithm will outperform the other in the different conditions and applications. SK*, for example, proved to be the most consistent in producing a good sharpening of the bearing signature, but in some cases other algorithms outperformed it. In any case this is not necessarily an issue. Using all these algorithms in parallel in fact, can enhance the robustness of the damage detection, increasing the reliability of the results.

A final remark regards the here considered spectral features (i.e. the characteristic spectral lines). For the simple detection of damage, in fact, the assessment of the presence of such characteristic spectral lines in the signal's PSD (even by eye) is enough to perform level 1 diagnostics. Nevertheless, such features can also be processed by the more refined

pattern recognition algorithms that will be taken into account in Chapter 8, enabling eventually also level 2, 3 and 4 diagnostics.

For the sake of computational speed and quickness of the detection anyway, next chapter is devoted to the application of the pattern recognition algorithms on way simpler features which better adapt to the needs of continuous monitoring.

Bibliography

[1] Najmi, A. Y., The Wigner Distribution: A Time-Frequency analysis, 2015

[2] Ville, J., Theory and applications of the notion of the analytic signal. 1948

[3] Stoica, Petre, and Randolph L. Moses, Spectral Analysis of Signals. Englewood Cliffs, NJ: Prentice Hall, 2005. ISBN-10: 0131139568

[4] K. Worden, G.R. Tomlinson, Nonlinearity in Structural Dynamics: Detection, Identification and Modelling, CRC Press, 2001. ISBN 13:9780750303569

[5] J. Antoni, "*Fast computation of the kurtogram for the detection of transient faults*", Mech. Syst. Signal Process. 21,108–124, 2007. DOI: 10.1016/j.ymssp.2005.12.002.

[6] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., "Bearing damage detection techniques and their enhancement: comparison over real data", International Conference Surveillance 8, 2015.

[7] Antoni J. et Al, "Feedback on the Surveillance 8 challenge: Vibration-based diagnosis of a Safran aircraft engine", Mechanical Systems and Signal Processing, 2017. DOI: 10.1016/j.ymssp.2017.01.037

Chapter 7:

Novelty Detection: statistical considerations via Monte Carlo simulations

1. Novelty Detection as Outlier Detection

In a data set, a discordant measure is usually defined "outlier", when, being inconsistent with the others, it is believed to be generated by an alternate mechanism. The judgment on discordancy will depend on a measure of distance from the reference distribution, often called Novelty Index (NI), on which a threshold can be defined [1].

Outliers may occur by chance in any distribution, but they often indicate either that the population has a heavy-tailed distribution or the presence of an anomaly such as a measurement error or a damaged condition. The identification of a damaged condition is the scope of data-based diagnostic systems, so that a deep knowledge of the outliers is needed to foster their robustness and reliability.

Unfortunately, there is no rigid mathematical definition of what constitutes an outlier, therefore, in many cases, determining whether an observation is an outlier or not is ultimately a subjective exercise: according to some algorithms an expert could detect doubtful measurements and take some decisions on them.

In order to improve the ability of distinguishing outliers, a study of the behaviour of extreme values coming from a reference normal distribution is proposed in this chapter, based on Monte Carlo repetitions. The essence of the Monte Carlo method (MC) is the invention of games of chance whose behaviour and outcome can be used to study some phenomenon of interest [10, 11] (outliers in this case). Carrying out games of chance or random sampling in fact, can not only result in sharp estimates of numerical quantities (for a high enough number of repetitions), but, given the statistical nature of the MC estimation, it enables also the determination of the degree of accuracy of the obtained estimates.

In this chapter, the MC simulations are based on the **Normal Distribution** whose mathematical formulation is reported hereinafter.

The Normal Probability Density Function (PDF) for a normally distributed variable x is:

$$x \sim N(\mu, \sigma^2)$$
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1$$

which can be standardized to get what is usually called z-score:

$$z = \frac{x - \mu}{\sigma}$$
$$z \sim N(0,1) \equiv Z(0,1)$$
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (2$$

A corresponding Cumulative Distribution Function (CDF) can be computed as:

$$\phi(\hat{z}) = P(z \leq \hat{z}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{z}} e^{-\frac{t^2}{2}} dt \qquad (3$$

Often the Complementary CDF (CCDF), or *survival function*, is used, as it corresponds to the tail area.

$$\overline{\phi(\hat{z})} = P(z > \hat{z}) = \frac{1}{\sqrt{2\pi}} \int_{\hat{z}}^{\infty} e^{-\frac{t^2}{2}} dt = 1 - \phi(\hat{z}) \qquad (4$$

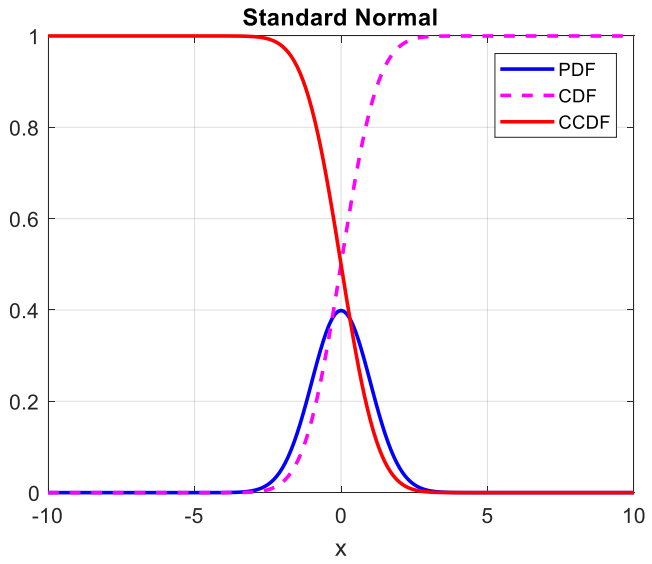


Figure 1: Standard Normal PDF, CDF, CCDF

1.1. Extrema empirical distribution via Monte Carlo

The analysis of outliers is related to the knowledge of the tails of the distribution, which experimentally are typically difficult to study because of the limited length of the commonly available samples. In this regard, MC offers a powerful advantage as it is possible to perform also vary long samplings, form which the extrema can be selected. This is the approach described in [1], whose algorithm is briefly summarized:

- Construct a vector of n observations randomly generated from a standard normal distribution,
- Compute the deviation of each observation (its absolute value),
- Save the maximum deviation and repeat the draw (back to point 1) for d times.
- A statistic on the maxima distribution could be performed.

In [1] this is used to find a threshold against which compare possible outliers.

The result of such approach for a univariate standard normal is reported in Figure 2, where the distribution of the extrema (in this case the maximum absolute values) from $d = 1000$ MC samples of size $n = 100$ is reported. The critical value for a confidence of 95% (the 5% significance critical value) is also highlighted. This implies that, given a sample of size $n = 100$, a value can exceed the ±3,4 confidence interval only 5 times over 100. This consideration can be then built to find a robust threshold for outliers' detection.

Similar considerations, resulting from the fundamental principles of the Calculus of Probabilities, could be seen in Peirce [2] and Gould [3]. A brief analysis of their proposed method and conclusions is reported in the following chapter.

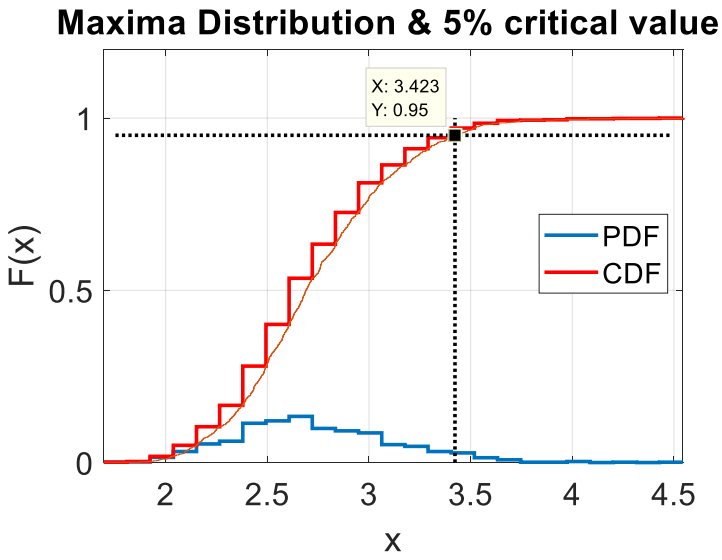


Figure 2: Maxima Distribution Histogram and 5% critical value (one tailed) for a sample of $n = 100$ MC observations repeated $d = 1000$ times.

1.2. Peirce's criterion

According to Peirce [2], in order to determine the limit of error in a series of n observations, it is possible to rely on a comparison between the probability of the system of errors obtained by retaining the abnormal observations and the probability of the system of errors obtained by their rejection multiplied by the probability of making so many and no more abnormal observations.

This may seem complex to compute, but the considerations in [2] and [3] have been translated by Ross [4] in a simpler table of R values, representing the maximum allowable deviation of a measured value from the mean μ, normalized on the standard deviation σ, for different numbers of doubtful observations (m).

The procedure simplifies to choosing a value of R (n, m), computing the threshold value for deviation:

$$|x - \mu|_{max} = R * \sigma \qquad (5$$

and finally comparing it to the deviation of the abnormal observations.

Table 1: R values, Ross [4]

Total #:	*Number of doubtful observations*					
n	1	2	3	4	5	6
3	1.196	-				
4	1.383	1.078	-			
5	1.509	1.200	-			
6	1.610	1.299	1.099	-		
7	1.693	1.382	1.187	1.022	-	
8	1.763	1.453	1.261	1.109	-	
9	1.824	1.515	1.324	1.178	1.045	-
10	1.878	1.570	1.380	1.237	1.114	-
11	1.925	1.619	1.430	1.289	1.172	1.059
…						

These results are anyway related to the Maxima distribution previously introduced, deeply analyzed in the Extreme Value Theory.

1.3. Extreme Value Theory

The Extreme Value Theory (EVT) is a branch of statistics dealing with the extreme deviations from the mean. It seeks to assess the probability of events that are more extreme than a selected reference, from a sample of size n of a given random variable. In this section, we will focus on the distribution of the maxima drawn from a standard normal.
According to EVT, in the limit for the number m of sampled vectors tending to infinity, the induced distribution on the maxima of the samples (e.g. see Figure 2) can only take 3 forms: Gumbel, Weibull or Frechet (those last two can be easily transformed into a Gumbel [5]) which can be combined into a single family of Generalized Extreme Value CDFs:

$$G(z) = \begin{cases} exp\left\{-\left[1+\xi\left(\frac{z-\mu_g}{\sigma_g}\right)^{-1/\xi}\right]\right\} & \xi \neq 0 \\ exp\left\{-exp\left[-\left(\frac{z-\mu_g}{\sigma_g}\right)\right]\right\} & \xi = 0 \end{cases} \quad (6$$

where μ_g is the location parameter, σ_g is the scale (or dispersion) parameter, while ξ is the shape parameter (0 for Gumbel- type I distribution) [6].

Analysing the behaviour of the maxima from a standard normal via Monte Carlo repetitions (see the next section of this chapter), in accordance to EVT, they can be proved to follow a Gumbel distribution, whose peak (the mode) is (asymptotically) placed in a point where the probability of exceedance for the original, generating distribution is $1/n$ (1 maximum in each sample of size n). The location parameter of the maxima distribution then tends to the inverse of the standard normal CDF for a probability of $1 - 1/n$ [7], [8].

Focusing on the **Gumbel** distribution for the maxima, its CDF and PDF can be mathematically expressed as:

$$G(z|\mu_g, \sigma_g) = e^{-e^{-\frac{z-\mu_g}{\sigma_g}}}$$

$$g\big(z|\mu_g, \sigma_g\big) = \frac{1}{\sigma_g} e^{-\left(\frac{(z-\mu_g)}{\sigma_g} + e^{-\frac{(z-\mu_g)}{\sigma_g}}\right)} \quad (7$$

$$G(p)^{-1} = \mu_g - \sigma_g\, ln\big(-ln(p)\big)$$

Figure 3 shows the shape of the PDF and CDF of the Standard Gumbel distribution together with the formulation of some relevant distribution features given as a function of the location and scale parameters μ_g and σ_g.

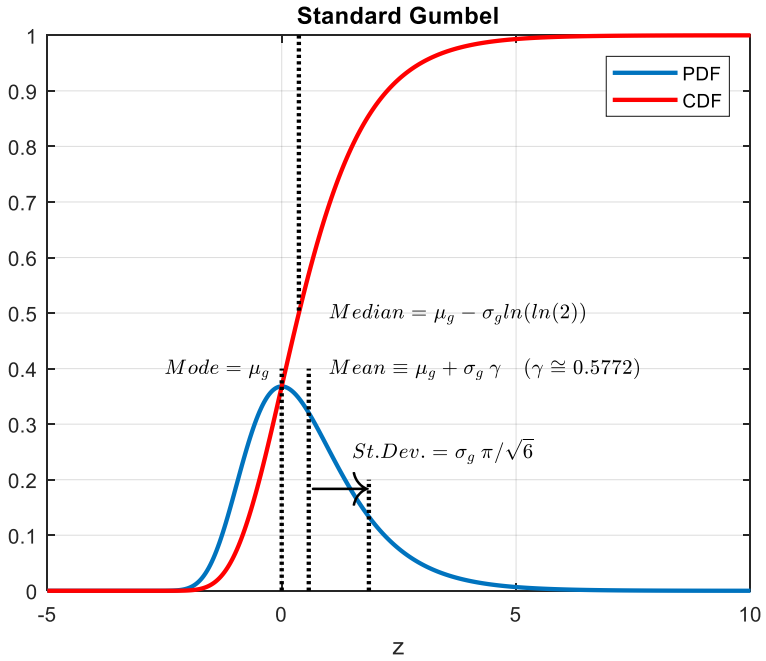


Figure 3: Standard Gumbel $G(z|0,1)$ PDF and CDF for the maxima ($\mu_g = 0, \sigma_g = 1$)

As can be even deduced from the MC simulations shown in following chapter, the standard normal belongs to the domain of attraction of the Gumbel distribution (the maxima distribution will tend to a type-1), and the characteristic parameters μ_g, σ_g for which $\frac{(z-\mu_g)}{\sigma_g}$ shows the standard Gumbel distribution $G(z|0,1) = \exp(-e^{-z})$ can be related to the μ, σ parameters of the original normal $N(x|\mu, \sigma)$.

This could be performed starting from von Mises's theorem [8]:

Theorem. Let F be a distribution CDF. Suppose:

- $F(x) < 1 \ \forall \ finite \ x \ \ (i.e. \ F \ has \ an \ infinite \ endpoint)$
- $F(x) \ is \ twice \ differentiable, at \ least \ for \ x > x_{lim} \ (tail)$
- $\lim_{x\to\infty} \frac{d}{dx}\left[\frac{1-F(x)}{F\prime(x)}\right] = 0$

Then

$$\lim_{n\to\infty} F^n\big(\sigma_g(n)\, x + \mu_g(n)\big) = G(z|0,1)$$

holds uniformly $\forall x \in R$, where

$$\mu_g(n) = \inf\{x: 1 - F(x) \geq 1/n\}$$
$$\sigma_g(n) = 1/n\, F'(\mu_g) \qquad \blacksquare$$

So, the location parameter μ_g of the Gumbel can be computed from the inverse normal cdf of $1/n$:

$$\mu_g(n) = mode = \underset{z}{\text{argmax}}(g(z)) = z_n \equiv x_n$$
$$\overline{\phi(x_n)} = \frac{1}{n} \qquad (8$$

And it can be proved that, asimptotically, for n large enough:

$$\mu_g(n) = \mu + \sigma\left[(2\ln n)^{1/2} - \frac{(\ln\ln n + \ln 4\pi)}{2(2\ln n)^{1/2}}\right] \qquad (9$$

While the dispersion parameter σ_g will correrspond to

$$\sigma_g(n) = \frac{\overline{\phi\left(\mu_g(n)\right)}}{f\left(\mu_g(n)\right)} = \frac{1}{n}\frac{1}{f(\mu_g)} = \frac{1}{n}\frac{1}{F'(\mu_g)} \qquad (10$$

And asimptotically:

$$\sigma_g(n) = \sigma(2\ln n)^{-1/2} \qquad (11$$

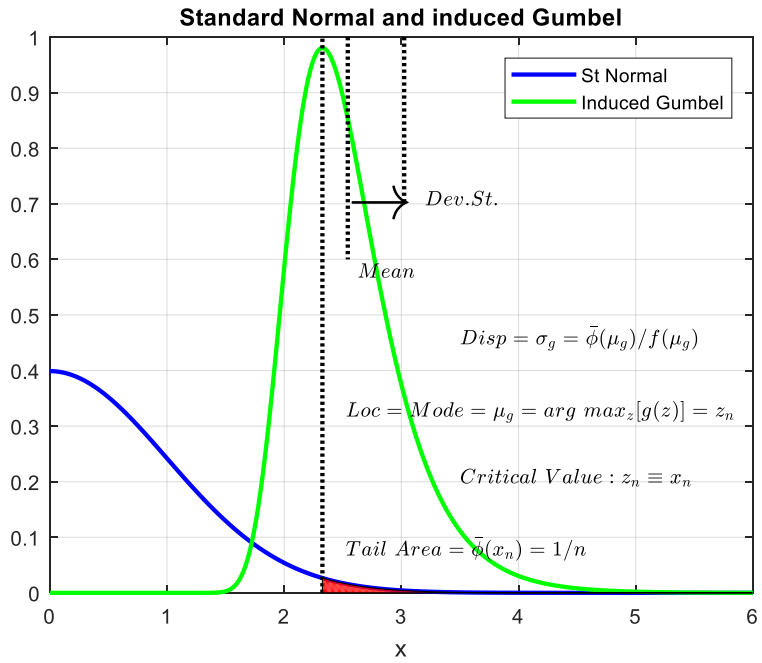


Figure 4: Standard Normal and Induced Gumbel ($n = 100$)

It is interesting to note that

$$\sigma_g = \frac{\overline{\phi(\mu_g)}}{f(\mu_g)} = \frac{1}{h(\mu_g)} \qquad (12$$

which corresponds to the inverse of what is often called *intensity* or *hazard* function h . This basically comes from the probability of observing a particular value x, conditional on the fact that the observation will at least exceed x.

$$A: x\ is\ observed \quad B: X \geq x\ is\ observed$$
$$\Pr(A|B) = \frac{\Pr(A\ \cap B)}{\Pr\ (B)} = \frac{\int_x^{x+dx} f(t)dt}{\int_x^{+\infty} f(t)dt} = \frac{f(x)dx}{1-\phi(x)} \qquad (13$$
$$h(x) = \lim_{dx\to 0} \frac{\int_x^{x+dx} f(t)dt}{\int_x^{+\infty} f(t)dt} = \frac{f(x)}{1-\phi(x)} = \frac{f(x)}{\overline{\phi(x)}}$$

Note that the hazard function is not a conditional probability, but rahter a sort of "conditional probability density", corresponding to the expected number of events per unit of x.

Another asymptotical interpretation can be achieved remembering the de l'Hôpital's rule:

$$\lim_{x\to\infty} \frac{f(x)}{1-\phi(x)} = -\frac{f'(x)}{f(x)} = -\frac{d\ln f(x)}{dx} \qquad (14$$

where the resulting ratio computed by *Logarithmic derivative* [$f'/f = (\ln(f))'$] is usually identified as *semi-elasticity*: the percentage change in the function corresponding to an absolute change in x.

Notice that, being $f(x)$ the standard normal PDF, $h(x)$ will asymptotically tend to the first-third quadrant bisector.

$$h(x) \approx -\frac{\Delta f(x)/f(x)}{\Delta x} \approx -\frac{d\ln f(x)}{dx} = x \qquad (15$$

Therefore, the inverse ratio of $h(x)$ is an absolute change in x; this is obviously related to the dispersion of the induced maxima distribution and approaches $1/x$ for x sufficiently large (namely for n large enough, as $x \equiv \mu_g(n)$).

1.3.1. Monte Carlo Simulation

Similar to what described in section 1.1 of this chapter, $d = 10^3$ Monte Carlo Repetitions for different sample sizes (n) are performed to study the evolution of the maxima distribution as a function of n. The result is graphically summarized in Figure 5, where each empirical distribution is fitted with a corresponding Gumbel, whose paramters are also reported.

As expected, the Normal distribution shows a type-1 (Gumbel) domain of attraction of class E1: the extreme distributions become more and more peaked as the numerousness n increases, while at the same time, the spacing between them decreases. These two trends are almost linear $\ln(n)$.

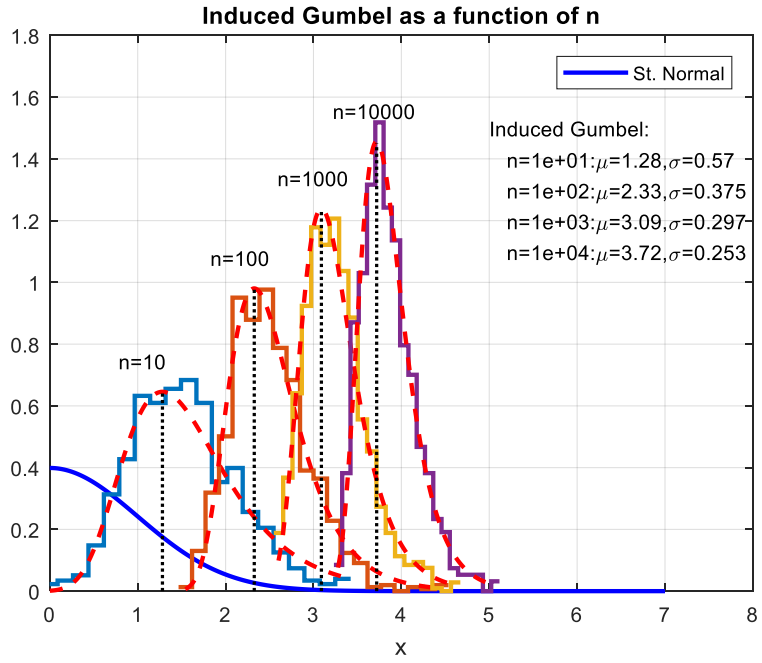


Figure 5: Several different Induced Gumbel distributions starting from the standard normal, as a function of n, for $d = 10^3$ Monte Carlo Repetitions.

Anyway, the dispersion and location parameters (reported in Table 2) could be quite easily foreseen using the formulas reported in previous section.

From Table 2, which summarizes the theoretical and estimated location and dispersion parameters, it is easy to understand that the error made by using the asymptotical formulation is sufficiently low for n large enough (e.g. $n > 10^2$).

Table 2: Comparison among theoretical, asymptotical and estimated location and dispersion parameters μ_g and σ_g.

n	μ_g	μ_g as.	MC est. μ_g	σ_g	σ_g as.	MC est. σ_g
10	1.28	1.36	1.23	0.57	0.47	0.53
10^2	2.33	2.37	2.31	0.38	0.33	0.38
10^3	3.09	3.12	3.08	0.30	0.27	0.29
10^4	3.72	3.74	3.70	0.25	0.23	0.25
10^5	4.26	4.28	4.26	0.22	0.21	0.21
10^6	4.75	4.77	4.75	0.20	0.19	0.19

1.3.2. EVT vs Peirce's criterion

In section 1.2, in accordance to Ross work [4], the quantity R was introduced as the ratio of the maximum allowable deviation from the data mean to the standard deviation.

$$R = \frac{|x - \mu|_{max}}{\sigma} \qquad (16$$

Referring to a standard normal, R can be then considered as the threshold that limits the one outlier region for the absolute value of x.

According to this consideration, a threshold similar to R could be computed with EVT starting from the absolute value of a draw from a standard normal. For a more rigorous formulation we will try to recover the first column of Table 1 (1 doubtfoul observation → 1 outlier) applying EVT to the square of a standard normal draw, corresponding to a draw from a $\chi^2(y|1)$ (chi-square, one degree of freedom) distribution.
In fact, it holds:

$$x \sim N(\mu, \sigma^2)$$
$$y = x^2 \;\rightarrow\; y \sim \chi^2(y|k) \;\; with\; k = 1\; dof$$
$$\chi^2 pdf(y|1) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$
$$\chi^2 cdf(\hat{y}|1) = P(y \le \hat{y}) = \int_{-\infty}^{\hat{y}} \frac{1}{\sqrt{2\pi t}} e^{-\frac{t}{2}} dt \qquad (17$$
$$\equiv 2 \int_{-\infty}^{\sqrt{\hat{y}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt - 1 = 2 * \phi(\hat{x}) - 1$$

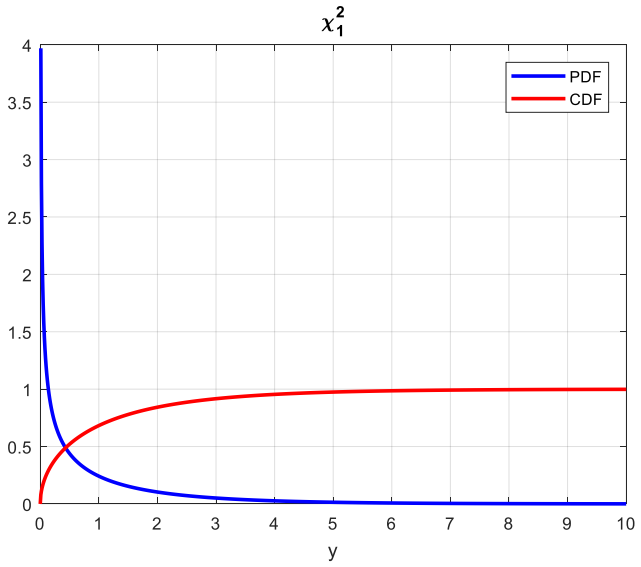


Figure 6: $\chi^2(1)$ PDF and CDF.

A Monte Carlo simulation, similar to the one proposed in the previous section, was then repeated, and the results are reported in Figure 7.
The picture shows a type-1 (Gumbel) domain of attraction similar to class E2: the extreme distributions has a "stable behaviour" preserving the dispersion (actually it shows a slight change, but at a very low rate), while the PDF is basically just shifted as the number n increases.

It is interesting to note that the square root of the Gumbel location parameter computed starting from the $\chi^2(y|1)$ distribution, $\sqrt{\mu_g}$, which can be used as a threshold for absolute deviation outlier detection, can be also calculated starting from the standard normal distribution in the following way (it is sufficient to focus on the relation between the $\chi^2 cdf$ and the standard normal CDF reported above):

$$\sqrt{\mu_g} = \left[\chi^2 cdf^{-1}\left(1-\frac{1}{n}\right)\right]^{1/2} = \sqrt{\hat{y}} \equiv \hat{x} = \phi^{-1}\left(1-\frac{1}{2n}\right) \quad (18$$

The MC results proposed in Figure 7, are summarized, in terms of thresholds, in Figure 8, where a comparison against Pierce's R values was also performed.

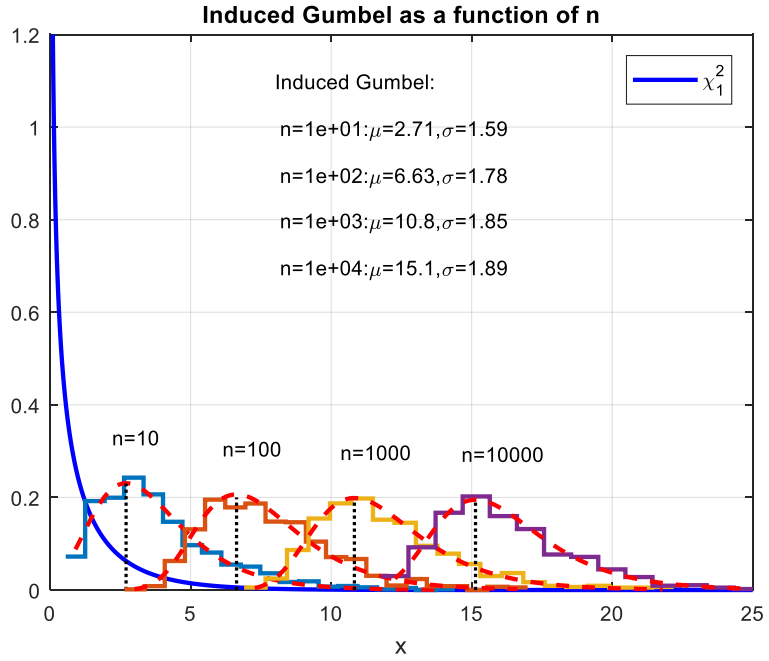


Figure 7: Several different Induced Gumbel distributions starting from the $\chi^2(1)$, as a function of n, for $m = 10^3$ Monte Carlo Repetitions.

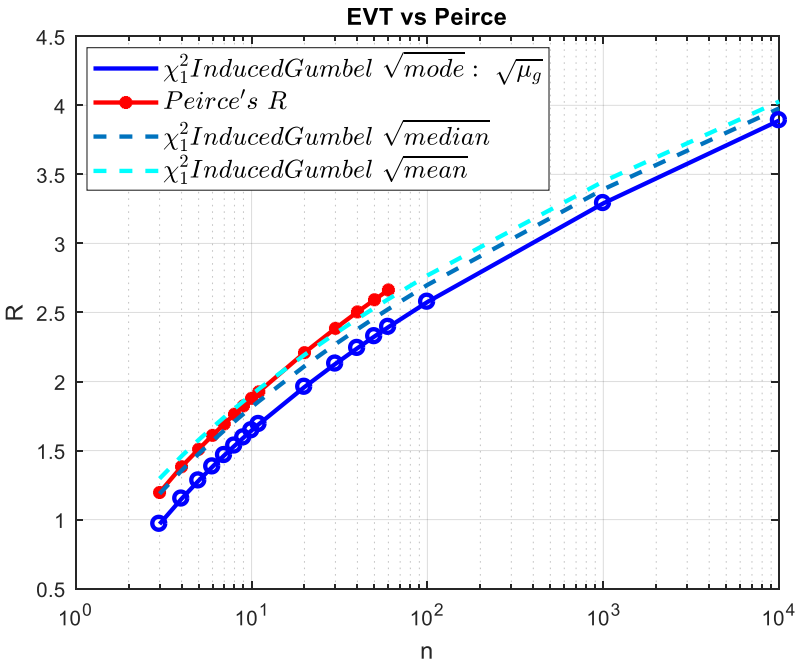


Figure 8: Peirce's R ($m = 1$) vs EVT threshold

In this image, it's possible to notice that the square root of the location parameter computed from the $\chi^2(y|1)$ ($\sqrt{\mu_g}$ – which corresponds to the square root of the mode) underestimates R, although the trend is quite similar.
A better approximation of the $R = f(n)$ function can be obtained focusing on the square root of the expected value of the EV distribution (namely the mean: $\bar{y} = \int_{-\infty}^{+\infty} y\, f(y) dy$).

2. Multivariate Extension

If the assumption of multivariate Normality holds for the original multivariate distribution, the sum of squares NI^2 is distributed as a perfect χ^2_d. That is

$$NI^2 = \sum_j \frac{z_j^2}{\lambda_j} \sim \chi^2_{(d)} \qquad (19$$

Then the ideal location and dispersion asymptotical parameters for NI can be obtained as a function of the dimension d of the feature space and the size n of the samples. For example, it can be written that:

$$\mu_g(n) = \sqrt{\chi^2_{(d),\frac{1}{n}}} \qquad (20$$

As introduced in Chapter 4, when the Mahalanobis distance is computed with sample estimates of mean and covariance matrix coming from small samples n, Wilks's correction should be used:

$$\mu_g(n) = \sqrt{\frac{d(n-1)^2 F_{(d,n-d-1),\frac{1}{n^2}}}{n\left(n-d-1+d\, F_{(d,n-d-1),\frac{1}{n^2}}\right)}} \qquad (21$$

These considerations can be easily proved through Monte Carlo repetitions on a multivariate Gaussian distribution, as suggested by Worden [1]:

1. Draw a sample of n observations randomly generated from a d-dimensional standard normal distribution,
2. Compute the deviation of each observation in terms of distance from the centroid i.e. the NI,
3. Save the maximum deviation and repeat the draw for m times.

The result of such operation is a collective from which several considerations can be obtained. In particular, in Figure 9, the NI probability distribution (histogram) from $m = 100$ repetitions with increasingly larger n drawn from a bivariate Gaussian is shown, together with a fitting of a Gumbel per each n. It is easy to appreciate the goodness of such fit, in accordance with the EVT theory. Furthermore, the so estimated $\hat{\mu}_g$ are stored and reported in Table 3 together with the values found for $d = 5$ and 10. There, the squared $\hat{\mu}_g^2$ are compared to their expected asymptotical value $\chi^2_{(d),\frac{1}{n}}$. As inferable from the EVT theory, the two values are matching very well. Finally, these asymptotical values are graphed as a function of the dimension d in Figure 11.

In order to verify Wilks's intuition, a second Monte Carlo repetition is proposed, where the observations come from a multivariate normal with randomly generated covariance and mean. The results of the simulation for $d = 5$ is summarized in Figure 11, where it is compared to the Wilks's μ_g^2 and to the $\chi^2_{(d),\frac{1}{n}}$.

To account both for small and large n a final summarizing formula can be given:

$$\mu_g(n) = \min\left(\sqrt{\chi^2_{(d),\frac{1}{n}}} \quad , \sqrt{\frac{d(n-1)^2 F_{(d,n-d-1),\frac{1}{n^2}}}{n\left(n-d-1+d\,F_{(d,n-d-1),\frac{1}{n^2}}\right)}} \right) \qquad (22$$

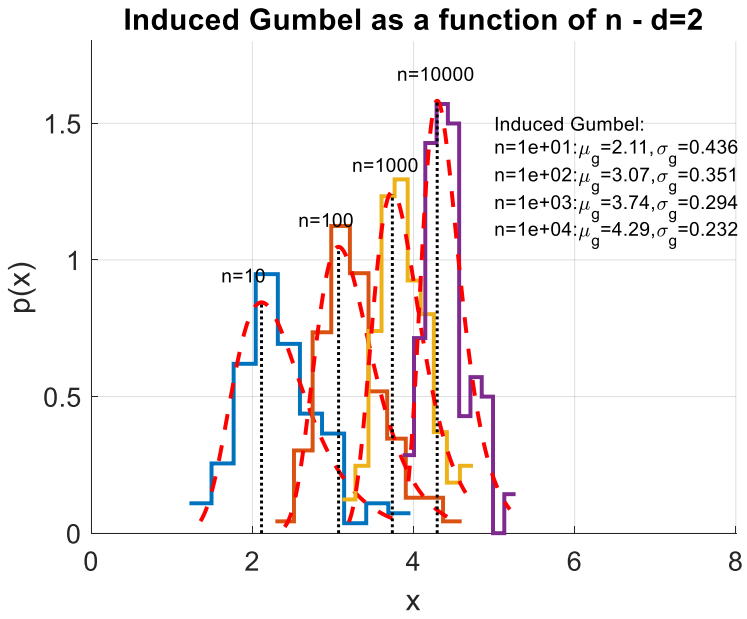


Figure 9: Several different Induced Gumbel distributions for the maxima arising from a bivariate standard normal, as a function of n, for $m = 100$ Monte Carlo Repetitions.

Table 3: Monte Carlo sampling of increasing n observations from Multivariate Standard Normals (d=2,5,10). The $\hat{\mu}_g$ values obtained from fitting a Gumbel are compared to the theoretical $\chi^2_{(d),1/n}$ critical values.

m=100	**d=2**		**d=5**		**d=10**	
n	$\hat{\mu}_g^2$	$\chi^2_{(d),\frac{1}{n}}$	$\hat{\mu}_g^2$	$\chi^2_{(d),\frac{1}{n}}$	$\hat{\mu}_g^2$	$\chi^2_{(d),\frac{1}{n}}$
10	4,72	4,61	9,12	9,23	16,12	15,99
100	8,85	9,21	14,37	15,01	22,57	23,21
1000	13,73	13,82	20,73	20,51	29,42	29,59
10000	18,29	18,42	25,90	25,75	35,83	35,56

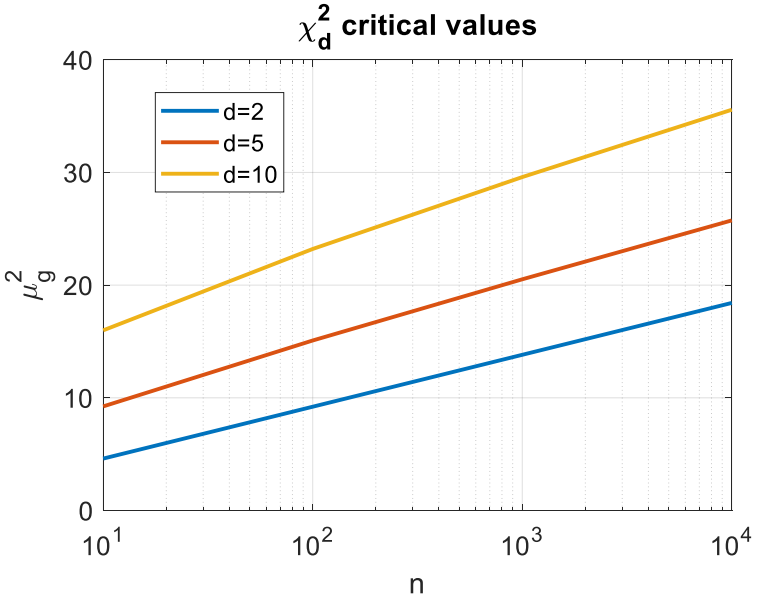


Figure 10: Expected asymptotical value $\chi^2_{(d),\frac{1}{n}}$ for MC sampling (d=2,5,10) as a function of n.

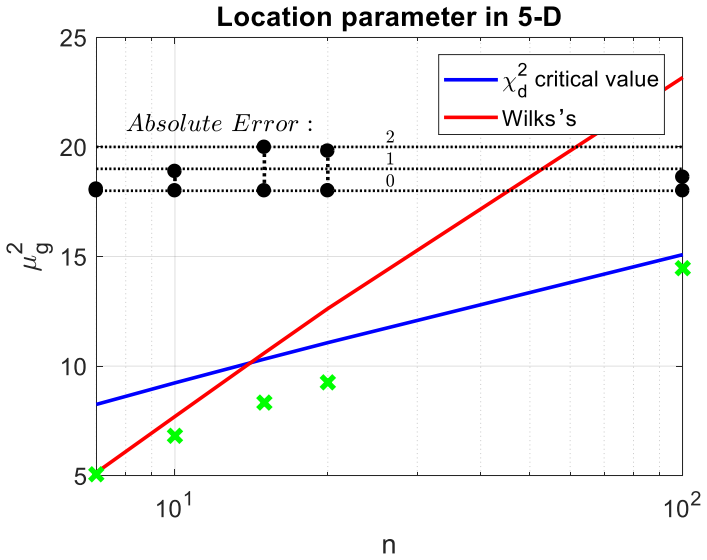


Figure 11: Monte Carlo sampling from a 5-dimensional Multivariate Normal with random mean and covariance matrix. The $\hat{\mu}_g^2$ values obtained from fitting a Gumbel (green dots) are compared to the theoretical $\mu_g^2(n)$ critical values form χ^2 and Wilks's criteria.

In any case, the relevant result is that a threshold for the NI can be found using statistical hypothesis testing considerations. In this respect, the theoretical μ_g turns out to be a very good candidate, even if, due to the asymmetric shape of the Gumbel, it can be good to increase it (e.g. of σ_g) to increment the significance α of the test, possibly at the expense of the power (see the considerations about confidence-power trade-off in Section 2.4). Figure 11 shows that a region in which the theoretical μ_g overestimates the location parameter $\hat{\mu}_g$ is always present. This area changes with the dimension d, so that in many cases it can be more conservative to use a MC simulation to find a robust threshold.

Furthermore, considerations about the so-called curse of dimensionality can be derived from such an analysis.

3. The curse of dimensionality

When the space dimensionality increases, the volume of the space becomes larger so fast that the available data are usually not enough to "fill" it and the data cloud turns out to be "sparse", so that the confidence on statistical estimates is eroded. Moreover, comparing the volume of a hypercube to the volume of the inscribed hypersphere it is possible to derive that the ratio $V_{h-sphere}/\,V_{h-cube}$ tends to 0 as $d \to \infty$, while the distance between the centre and the corners increases without any bound with d. The high-dimensional unit hypercube then can be said to consist almost entirely of the "corners" with almost no "middle". This space density deformation is highlighted also by the χ^2 distribution shape. As illustrated in Figure 12, in fact, most of the d-cube volume concentrates near the surface of a sphere of radius $\sqrt{d}$. Indeed, the limiting distribution of the χ^2 for an increasing d (i.e. $d > 50$) can be proved to be the normal $N_{(d,\sqrt{2d})}$.

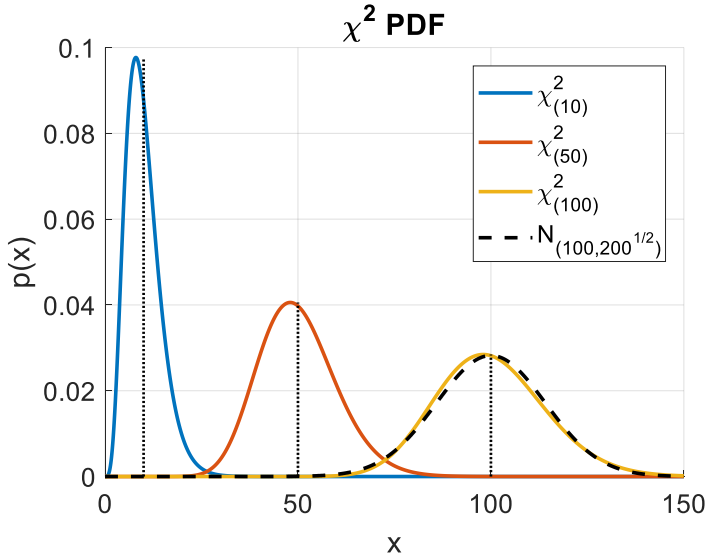


Figure 12: Three χ^2 distributions for an increasing number of dofs. The χ^2_{100} is compared to the asymptotical tendency distribution $N_{(100,\sqrt{200})}$. The asymptotical means are highlighted as black dotted lines. Notice that increasing the dofs, the distribution concentrates around the asymptotical mean.

4. Multivariate issues: Robust covariance estimation

The main issue brought by the so-called curse of dimensionality is related to the sparsity of the high dimensional datasets. In fact, the amount of data needed to "fill the space" grows exponentially with the dimensionality, so that it is usually practically impossible to have n large enough. Furthermore, in these sparse datasets, possible outliers can truly kill the reliability of statistical estimates such as the mean value and the covariance matrix. These estimates are fundamental for hypothesis testing, classification and novelty quantification, so that the problem of robustness to outliers and samples numerousness cannot be neglected.

Focusing on the estimation of the covariance matrix Σ, similar considerations to the one introduced in Chapter 4 can be derived. In particular, the maximum likelihood estimate result biased

$$S_{n,ML} = \frac{1}{n} X^t X \qquad E\left[S_{n,ML}\right] \neq \Sigma \qquad (23$$

so that, the unbiased estimator should be preferred.

$$S_n = \frac{n}{n-1} S^{ML} \qquad (24$$

Notice that, for a dimension d, the estimation of S_n involves the computation of $\frac{d^2-d}{2}$ covariances and d variances, or overall $\frac{d^2+d}{2}$ parameters. In this regard, the overall mean square estimation error can be decomposed as

$$MSE(S) = bias(S)^2 + var(S)$$

$$E\left[||\Sigma - S||_{fro}^2\right] = \sum_i E^2[\sigma_i - s_i] + var(s_i) \qquad (25$$

where $||\cdot||_{fro}^2$, the so-called Frobenius norm accounting for elementwise errors, can be written as the sum of a bias and a variance error term. The problem of **bias – variance trade-off** is a common one in estimation theory. In particular, when $n \sim \frac{d^2+d}{2}$, the variance model used is too complex for the available data, so that the risk of overfitting is very high.

That is, the estimation is good on the particular sample, but out of sample, the variability can be very high as the model is not properly generalizing the hidden patterns but is picturing also the noise in the training sample. On the contrary, using the unbiased estimator, the problem of underfit (n too large) will not arise.

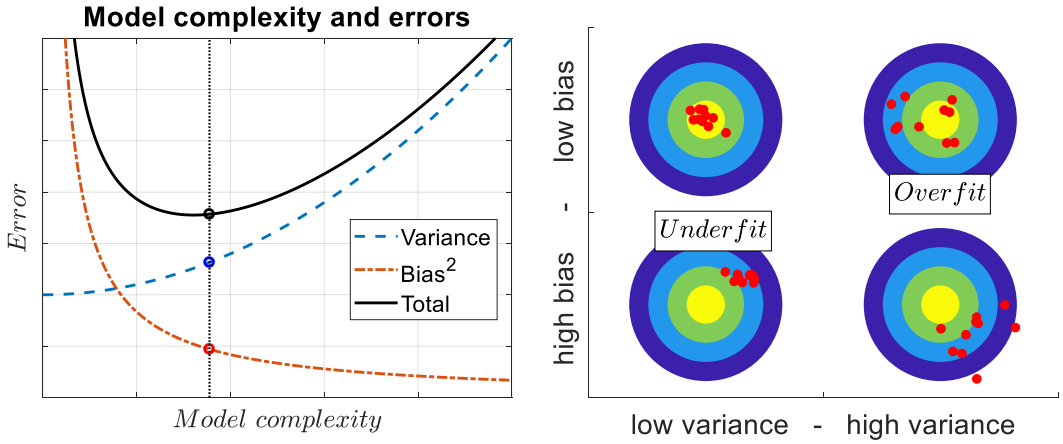


Figure 13: Total error vs model complexity and the corresponding conditions of underfit and overfit as a function of accuracy (low bias) and precision (low variance).

In order to improve the reliability of the covariance estimation, statistics offers a number of options:

- Cross validation and leave one out cross validation
- Bootstrap aggregation (“bagging”)
- Covariance concentration
- Shrinkage

The first three are mainly useful to produce a covariance estimate robust to invclusive outliers, while the last one faces the problem of small n.

4.1. Robust covariance estimators

In this section, the main approaches to the subject of robust covariance estimates is faced. The algorithms are first introduced using the simple dataset of reference [12]. This dataset collects the values of 13 constituents found in three types of Italian wines. For the sake of simplicity, as in [12], just the first group of wines is considered, and the number of features is limited to two (i.e. "malic acid" and "proline"). As it is easy to note in Figure 14, the dataset features many outliers which affect the estimate of the covariance matrix, whose corresponding ellipse (for confidence $97{,}5\%$) is highlighted. In general, the measure of the volume of an ellipsoid is proportional to the determinant of the covariance matrix, so that this measure can be used to quantify the ellipsoid dimension. (Precisely, it is the determinant times the volume of the unit spheroid). Being this a bi-dimensional case, the volume shrinks down to an area, whose value is reported in figure.

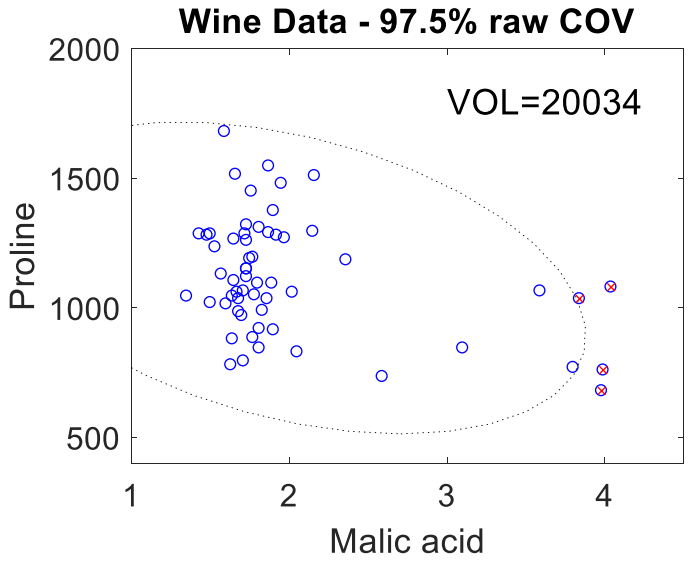
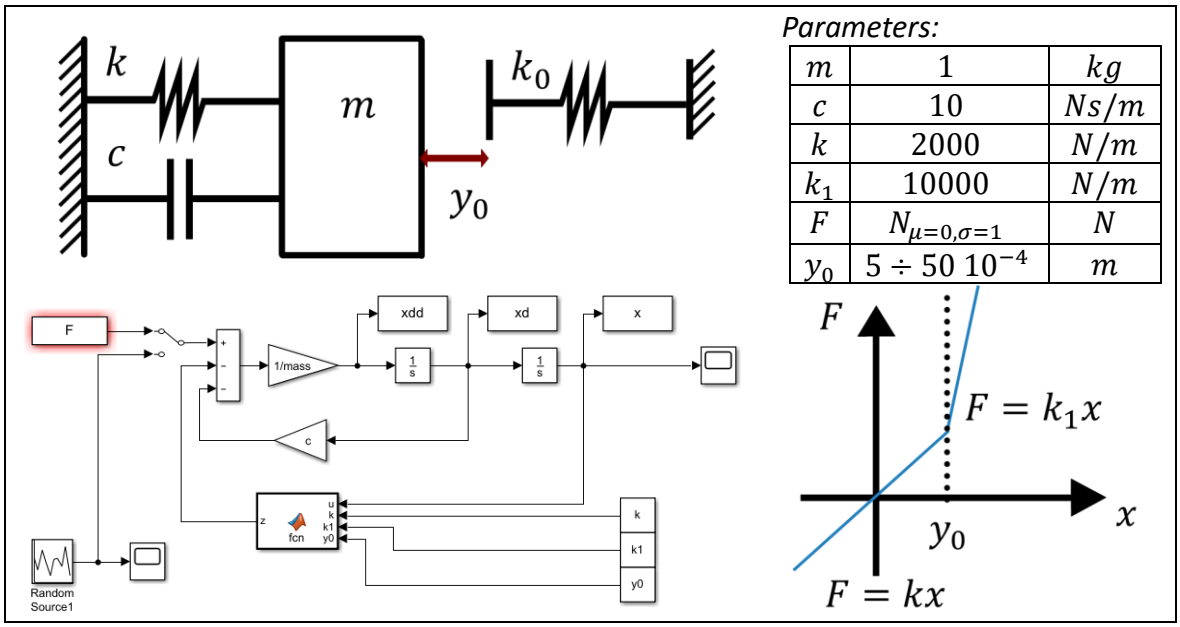


Figure 14: Dataset of Italian wines in terms of their constituents [14]. A family of wines is plotted as a function of two features: "malic acid" and "proline". The ellipsoid corresponding to the estimated covariance delimiting the 97,5% confidence region results affected by the presence of many outliers. In red the only detected outliers at a confidence 97,5%.

Then, a non-linear mechanical system was simulated, and the so generated dataset was used for comparing the different algorithms.

The system selected for this analysis is the simple 1-DOF damped harmonic oscillator, with an asymmetric nonlinearity introduced by a bilinear stiffness which simulates the presence of clearance. When the positive displacement exceeds the clearance, the mass comes into contact with a hard stop that restricts the motion of the body. This is simulated by an instantaneous increase in the stiffness from k to k_1, which accounts for the parallel of the two stiffnesses (i.e. the spring k and the hard stop k_0 stiffnesses), as shown in Figure 15. Outputs of the model for different values of the clearance y_0 were then computed. An example of the standard normal input and the output in terms of position and acceleration are reported in Figure 16 for $y_0 = 1\ \mu m$. The signals consisting of 10000 samples at a frequency of $100\ Hz$ were then brought to the frequency domain (i.e. Welch Periodogram) and a limited number of spectral lines around the natural frequency of the system ($f_n = \frac{1}{2\pi}\sqrt{k/m} = 7{,}2\ Hz$) were selected as features.



m	1	kg
c	10	Ns/m
k	2000	N/m
k_1	10000	N/m
F	$N_{\mu=0,\sigma=1}$	N
y_0	$5 \div 50\ 10^{-4}$	m

Figure 15: The modelled system and the block scheme of the corresponding differential equation

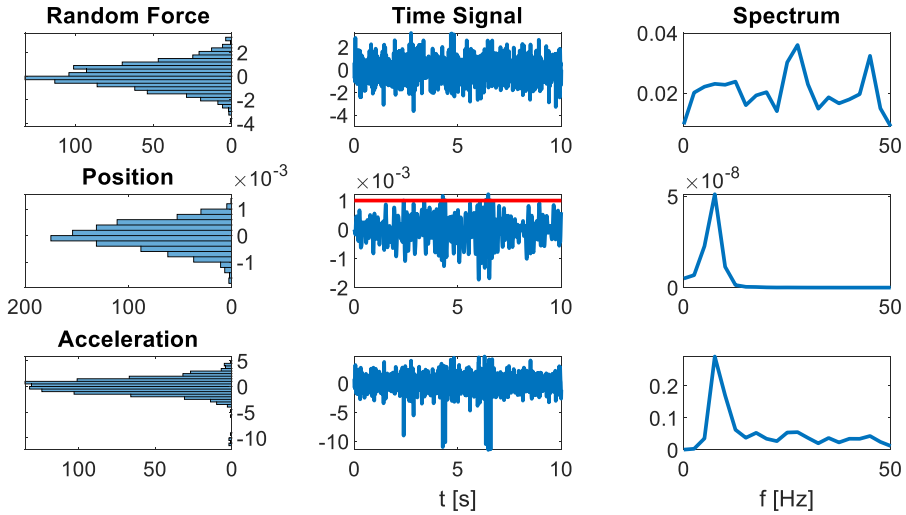


Figure 16: An example of the standard normal input and the output in terms of position and acceleration for $y_0 = 1\ \mu m$.

4.1.1. Cross validation and leave one out cross validation

Cross validation is a common procedure in model validation for assessing how the results of a statistical analysis will generalize to an independent data set. In particular, the leave-p-out cross-validation involves using p observations as the validation set and the remaining observations as the training set. This procedure for $p = 1$ (i.e. leave one out) is adopted by [13] to make the mean value and covariance estimates independent from the dataset. The so called Jackknifed estimates turn out to be:

$$\overline{x_i} = E[x]_{j \neq i}$$

$$S_i = E\left[\left(x - E[x]_{j \neq i}\right)\left(x - E[x]_{j \neq i}\right)^t\right]_{j \neq i} \quad (26$$

This way it is possible to obtain an independent measure of the Mahalanobis distance for the i-th observation

$$D_i^2 = (x_i - \overline{x_i})' S_i^{-1} (x_i - \overline{x_i}) \qquad (27$$

In the first instance, a threshold $\chi^2_{(2),2.5\%}$ can be easily built at a confidence $1 - \alpha = 97{,}5\%$ to detect and remove outliers. The result is shown in Figure 17, where a substantial reduction in the volume is obtained. In any case this is an approximation, as the correct critical value according to [13] should be $\frac{nd(n-2)}{(n-1)(n-d-1)} F_{(d,n-d-1),\alpha/n}$.

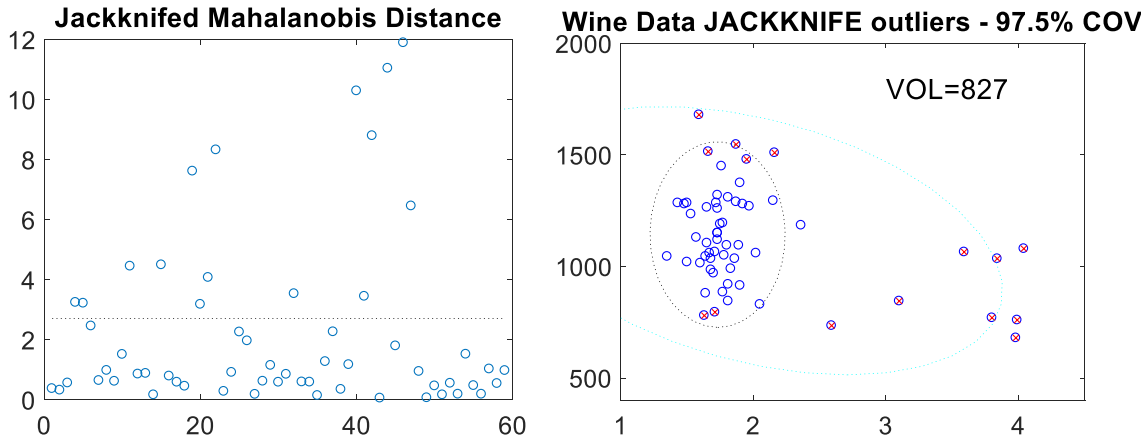


Figure 17: Jackknifed Mahalanobis distance and the 97.5% confidence threshold on the left. On the right the farther excluded points are highlighted in red, while the covariance ellipsoid of the remaining points is drawn.

4.1.2. Bootstrap aggregation ("bagging")

The bootstrap aggregation (or bagging) is a procedure which relies on random sampling with replacement. A way of using this procedure for making the covariance consists of

- Estimating the multivariate pdf from the data to assign importance weight to the observations (i.e. weights proportional to the probability). In the example a gaussian kernel is used (this is suitable for low space dimensions). As it easy to notice in Figure 18, despite the bias in the PDF, the importance weights can be correctly estimated. In the simplest case, the importance weight can be set all equal, so that this step can be simplified (random bootstrap),
- Resampling with replacement according to importance weights and computing the determinant,
- Repeating and selecting the minimum determinant until a stopping criterion is met (e.g. max iterations or minimum error).

The result of two iterations is shown in Figure 18, where the corresponding sampled values are highlighted in green.

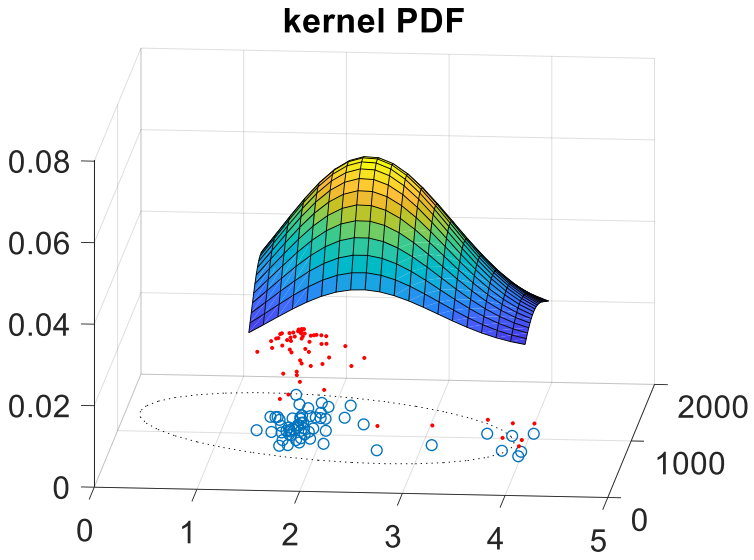


Figure 18: Bivariate Gaussian kernel PDF and corresponding importance weights in red.

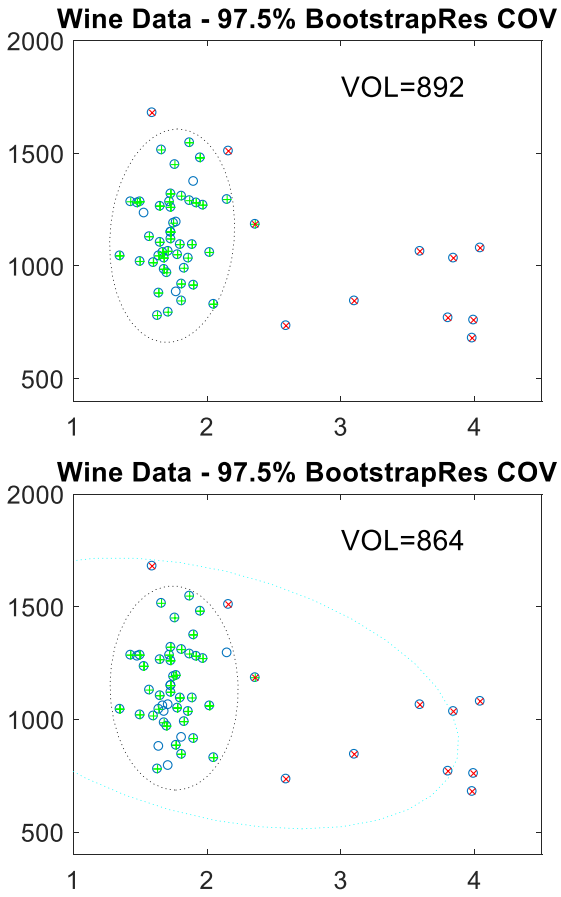


Figure 19: Two different importance samplings (the points drawn are highlighted in green) and the corresponding covariance ellipsoids. In red the outliers.

4.1.3. Minimum Covariance Determinant (MCD)

A very established technique for producing robust covariance estimates is the so-called Minimum Covariance Determinant (MCD) [12] also used for Structural Health Monitoring [23,24]. It is based on concentration steps which are exploiting both the bootstrap and the cross-validation idea. The procedure is based on:

- Random selection of a sub-sample of n_s observation, from which a covariance S_j is estimated. N.B. $\frac{n}{2} \leq n_s \leq n$ and j identifies the iteration.
- Use S_j to compute the Mahalanobis distance of the entire sample and select the smallest n_s observations.
- Compute a new covariance S_{j+1} with these values and compare its determinant to S_j determinant: if smaller iterate the procedure, otherwise if the determinant is the same or 0 stop and repeat for a different initial random sub-sample.

The result of this algorithm to the wine data, as produced in [12], is shown in Figure 20.

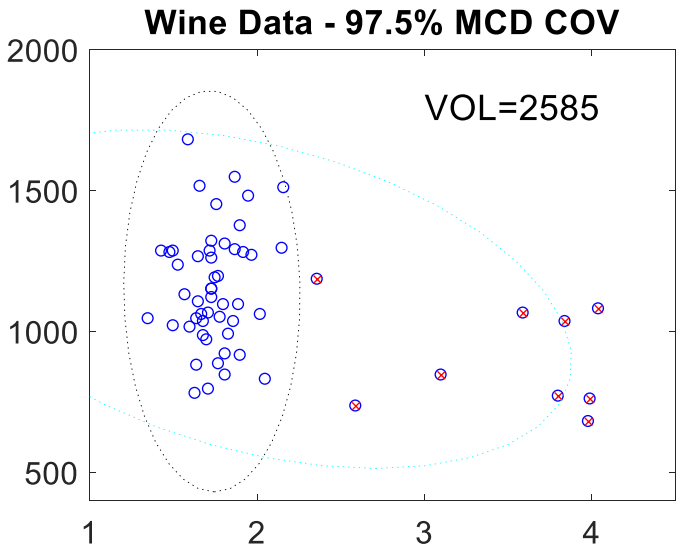


Figure 20: MCD 97,5% covariance ellipsoid and the detected outliers in red.

4.1.4. Comparison of the robust covariance estimators over the simulated signals from the non-linear mechanical system for Novelty Detection

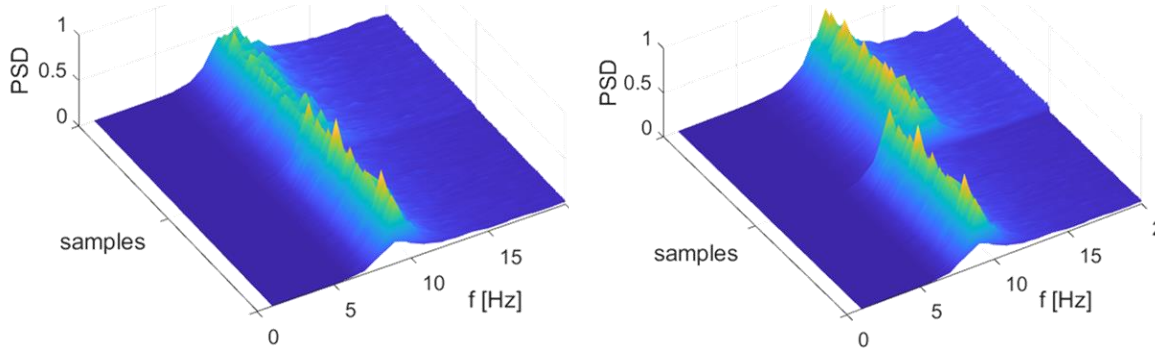
As introduced at the beginning of the chapter, outliers are of great interest in statistics. In machine diagnostics they are fundamental for the damage detection. Mahalanobis Distance Novelty Detection for example, assumes that the healthy data from a machine can be modelled as a multivariate normal cloud, so that outliers can be used to assess the presence of novelty, which denotes damage in absence of confounding influences. Nevertheless, if the healthy training dataset is corrupted by outliers (in this case extraordinary events which may have affected the measurements), the estimate of the normal distribution parameters (i.e. mean vector and covariance matrix) can be biased by such outliers.

The performance of ND in case of outliers' inclusive trainings can be strongly reduced because of biased parameters estimation. In these cases, then, it is fundamental to remove such outliers. Robust ND can be implemented with the three methods introduced above, so that it is relevant to test their performance. In order to make an objective comparison, the simulated data from the non-linear 1DOF oscillator are used, and the performance in terms of robustness is assessed via the Kullback-Leibler divergence of the estimated distribution (in terms of mean vector and covariance matrix) with respect to the theoretical distribution in the feature space. In particular, up to 20 spectral lines around

the natural frequency of the system in the linear range are used as features for a multivariate analysis. The smoothed spectral density was estimated with the Welch's Periodogram methodology with a Hamming window of length 1000 samples, 90% overlap and 110 spectral lines of frequency discretization.

The system was simulated changing the values of the clearance to produce different responses to the Gaussian random input. In particular, the theoretical distribution parameters are estimated from a simulation with a very large value of $y_0 = 5\ \mu m$.

This dataset was later corrupted with observations from the same system when the clearance is first reduced to $y_0 = 1\ \mu m$ (soft non-linearity effect) and then decreased again up to $y_0 = 0{,}5\ \mu m$ (strong non-linearity effect). From a visual inspection of the spectra, in the first case (Figure 21-a) the natural frequency is starting a slight shift to higher frequencies, which is evident in Figure 21-b (compared to the spectra obtained in case of $y_0 = 5\ \mu m$, for which the system is linear, which are reported in the first half of the pictures).



a) $y_0 = 1\ \mu m$ b) $y_0 = 0{,}5\ \mu m$

Figure 21: Cascade of spectra from several MC repetitions of the model compared to the same reference obtained with $y_0 = 5\ \mu m$ (linear system), reported in the first half of the pictures; a) comparison with results from $y_0 = 1\ \mu m$ (soft non-linearity effect); b) comparison with results from $y_0 = 1\ \mu m$ (strong non-linearity effect).

Different levels of contamination were accounted for an overall samples numerousness of 200 in Figure 22 and 23 and a reduced numerousness of 60 in Figure 24. In all the pictures, the Kullback-Leibler Divergence under the assumption of normality was computed. The aggregated results for all the MC repetitions are shown.

Note that the KL divergence was computed under the assumption of normality of the estimated d-dimensional pdf $q \sim N(\mu_2, \Sigma_2)$ from a normal reference pdf $p \sim N(\mu_1, \Sigma_1)$ as [22]:

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx = \\ = \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)' \Sigma_2^{-1} (\mu_2 - \mu_1)\right] \quad (28$$

Several considerations can be deduced by looking at the MC aggregated results (mean value and $\pm$ standard deviation region) in Figures 22, 23 and 24, where the traditional inclusive (non-robust) estimators for mean and covariance are compared to exclusive estimates, where the removed outliers are spotted by Mahalanobis Distance (MD), Jacknife Mahalanobis Distance (JKN) and Minimum Entropy Deconvolution (MCD) or via Random Bootstrap (RB). The algorithms are implemented assuming that no prior

knowledge about the true number of outliers is known, so that no parameters optimization is performed.

The RB was selected in place of the importance bootstrap as the analysis was interested in estimating the performance of the methods for a feature space dimensionality up to 20. In this case the use of KDE for the multivariate pdf is not advisable, so that the importance bootstrap was dropped. The concentration effect of 50 repeated draws of samples of 75% the size of the original is not substantial. Just when the number of samples is big enough and the space dimensionality is large, RB proves to slightly reduce the estimation error. When a simple χ^2 critical value with significance $\alpha = 1/n$ is used to detect outliers from MD, the estimation error is reduced only when the contamination is below 10%. In this case, the sample size is also relevant, as for small samples (e.g. Figure 24) the so estimated threshold results inappropriate. In this case, if the number of included outliers is not big, the performance can be improved using a Jacknifed estimate of the Mahalanobis distance with an appropriate threshold, as the one based on Fisher's F introduced in previous section. Finally, if the number of included outliers is large, the FAST MCD with a $n_s = 0{,}5\ n$ proves to be the best in reducing the estimation error. This robust estimator is in fact an effective mix of Bootstrap ad MD outlier detection, which can outperform all the other algorithms if the outlier fraction is optimized according to the expected inclusiveness value. For example, reducing the fraction to 25% ($n_s = 0{,}25\ n$), the first two graphs in Figure 23 can be improved as seen in Figure 24.

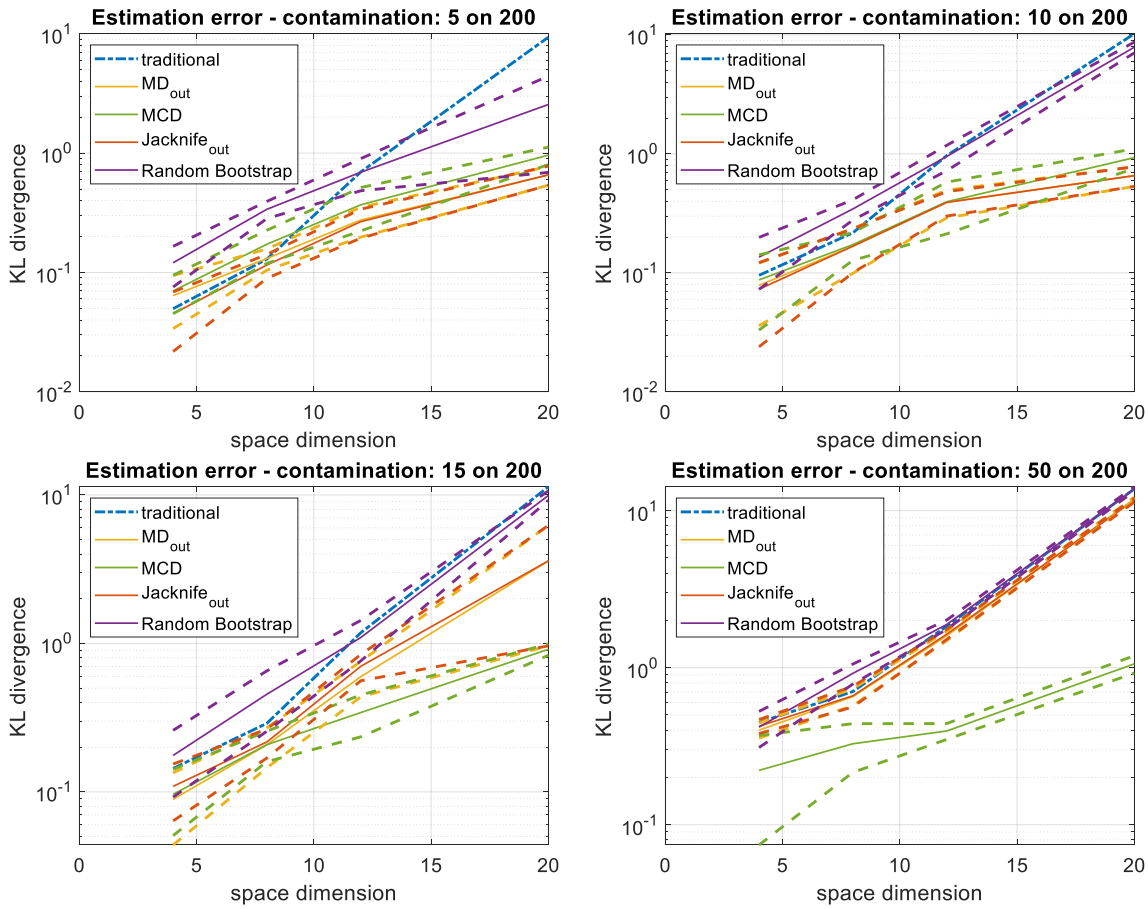


Figure 22: Robustness assessed in terms of KL divergence of the estimated multivariate psd (model with $y_0 = 1\ \mu m$) from the theoretical psd (with $y_0 = 5\ \mu m$) when an outlier inclusive set with declared contamination proportion is used.

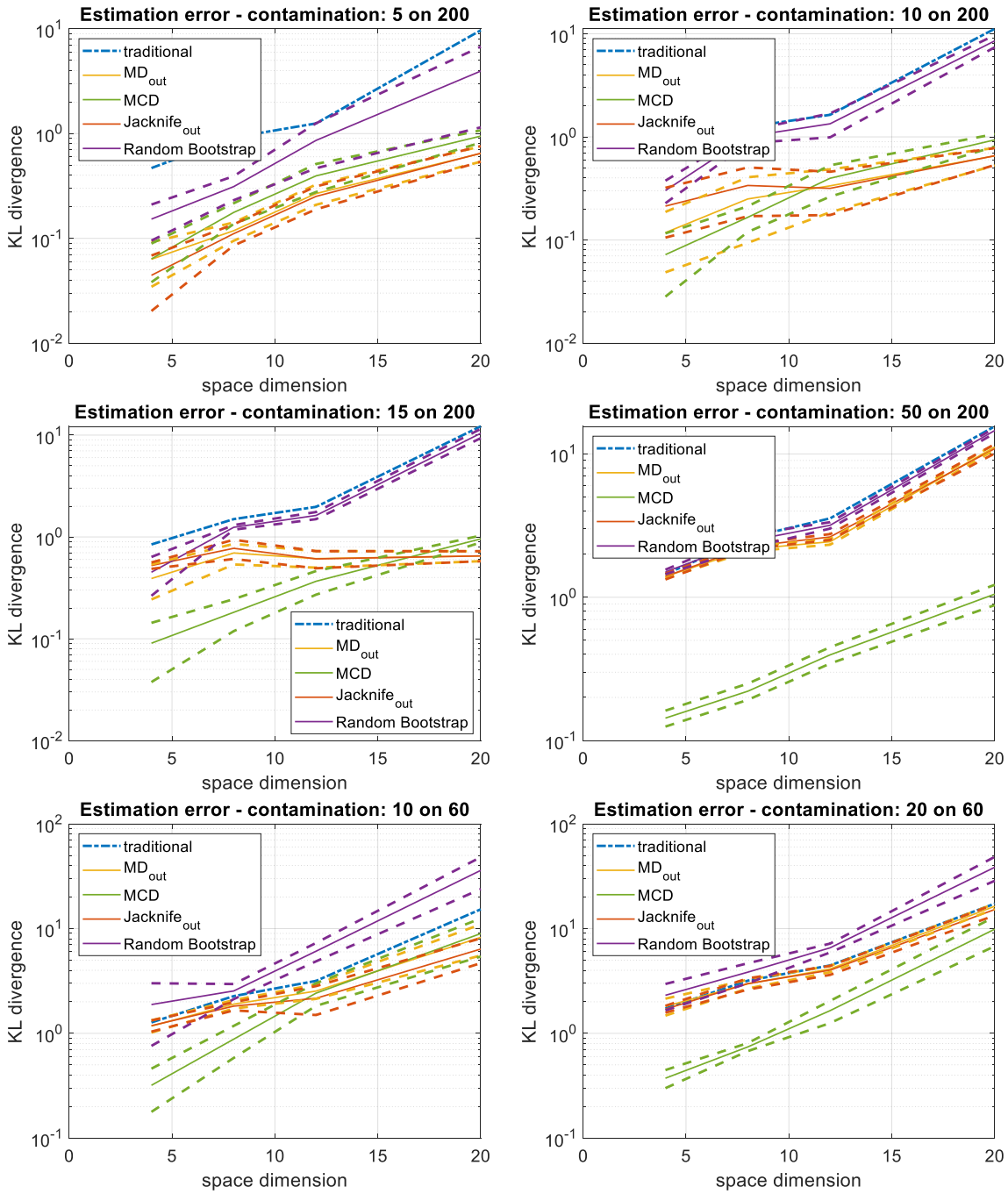


Figure 23: Robustness assessed in terms of KL divergence of the estimated multivariate psd (model with $y_0 = 0{,}5\ \mu m$) from the theoretical psd (with $y_0 = 5\ \mu m$) when an outlier inclusive set with declared contamination proportion is used.

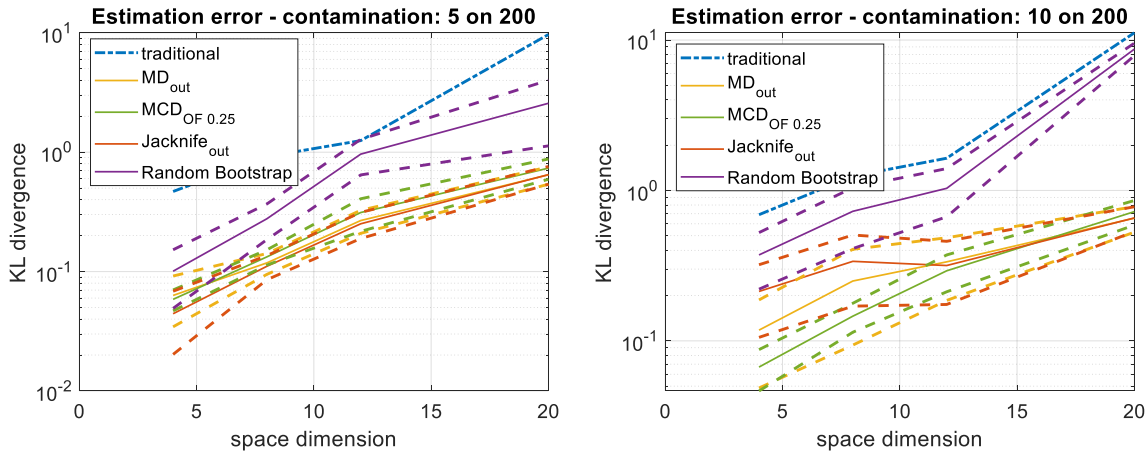


Figure 24: Robustness assessed in terms of KL divergence of the estimated multivariate psd (model with $y_0 = 0{,}5\ \mu m$) from the theoretical psd (with $y_0 = 5\ \mu m$) when an outlier inclusive set with declared contamination proportion is used. MCD outliers fraction reduced to 25%.

4.2. Shrinkage (regularization)

As repeated at the beginning of Section 4, the unbiased covariance estimator virtually reaches by definition no bias when $n \to \infty$. Actually, in practical cases n is always finite, and often small compared to the dimension d, implying a very large variance term (overfit). An idea appeared in [14,15,16] to face very small sample sizes is then to accept a limited amount of introduced bias provided that the variance will reduce more. This procedure is commonly called shrinkage:

$$R_\alpha = (1-\gamma)\, S + \gamma\, T \qquad (29$$

the overfitted estimate S is combined to a predefined underfit target T through a shrinkage coefficient γ. Common options for the underfit target T are covariances with simplified structures such as

diagonal, unit variance	*diagonal, common variance*	*diagonal, unequal variance*
$T \equiv I$	$T = t^2\, I$ $t^2 = tr(S)/d$	$t_{ij} \begin{cases} s_{ij} & i = j \\ 0 & i \neq j \end{cases}$

Obviously, the selection of a particular structure in place of another is up to the experience and expectation of a researcher. A quite neutral selection anyway can be the diagonal, unequal variance matrix T, which basically keeps the marginals of the estimate S but removes the information of the correlation. Whatever the selected T anyway, the shrinkage parameter γ can be optimized by minimizing a loss function. Taken as expected loss the mean square error and recalling the bias variance decomposition, it can be proved that

$$\begin{gathered} MSE(R) = bias(R)^2 + var(R) \\ E\left[||R-\Sigma||_{fro}^2\right] = \sum E^2[r_i - \sigma_i] + var(r_i) \; = \\ = \sum (E[\alpha t_i + (1-\alpha)s_i - \sigma_i])^2 + var(\alpha t_i + (1-\alpha)s_i) \end{gathered} \qquad (30$$

$$\alpha^* = \frac{\sum var(s_i) - cov(t_i, s_i) - bias(s_i)E[t_i - s_i]}{E[(t_i - s_i)^2]} \qquad (31$$

And for the particular case of T diagonal, unequal variance, the optimal value of α can be proved:

$$\alpha^* = \frac{\sum_{i \neq j} var(s_{ij})}{\sum_{i \neq j} s_{ij}^2} \qquad (32$$

where the unbiased empirical variance of the individual entries of S equals

$$var(s_{ij}) = \frac{n}{(n-1)^3} \sum_k \left(w_{kij} - \bar{w}_{ij}\right)^2 \qquad (33$$

With

$$w_{kij} = (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}) \qquad \bar{w}_{ij} = \frac{1}{n} \sum_k w_{kij} \qquad (34$$

A synthetic example can be found in Figure 25, where two random samples of 4 observations in a dimension $d = 2$ are drawn from a bivariate normal having the 97.5% confidence ellipse shown in black. Having just 4 samples, the probability of correctly picture the covariance matrix are very low. In most of cases, the points are either grouped near the sample average (in green) and featuring then a low correlation, or almost aligned, with a high correlation (Parsons's ρ). In the first case the shrinkage is not improving the situation, but is quite conservative, as the sample covariance (i.e. the red ellipse) is much affected. In the second case on the contrary, the shrinkage improves the covariance estimate (i.e. the green ellipse). As it is easy to notice, the marginals obtained by the sample covariance are left unaffected as the diagonal, unequal variance T only modifies the off-diagonal elements of S. The 97.5% shrunk ellipse (in green) is anyway much nearer to the true ellipse (in black).

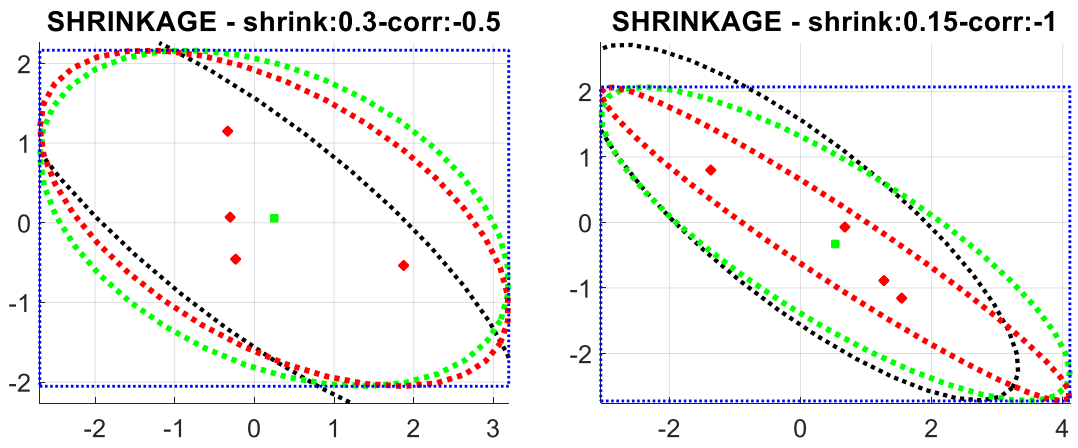


Figure 25: Two samples (size $n = 4$) from the distribution featuring the black 97.5% ellipsoid. The estimated mean (green dot) and covariance (red 97,5% ellipsoid) are shown. In the first case, a small correlation is found, so that the optimal shrinkage $\alpha = 0{,}3$ is not much effective. In the second case, the very high correlation is reduced by the shrinkage ($\alpha = 0{,}15$) to a more probable value.

In any case, when there is the freedom of selecting a proper n, this must always be done in a design of experiment phase (DOE). In the next section, the selection of a proper sample size n is tackled.

5. Proper sample size n: how large is large enough?

In the design of experiment phase a fundamental step is the selection of the sample size n to ensure a proper confidence in the estimation of the covariance matrix. In [17] a nice derivation of the minimum appropriate n based on probability distributions and confidence intervals is given.

The derivation for the 1-D variance is quite straightforward. The confidence interval is given by

$$\Pr\left[\left|\frac{S_n}{\sigma} - 1\right| < \epsilon\right] = 1 - \alpha \qquad (35$$

Given that for large n it holds:

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{(n-1)} \qquad Z = \sqrt{\frac{2(n-1)s_n^2}{\sigma^2}} - \sqrt{2(n-1)-1} \sim N_{(0,1)} \qquad (36$$

it can be derived:

$$n \cong 1 + \frac{1}{2}\left(\frac{Z_{\alpha/2}}{\epsilon}\right)^2 \qquad (37$$

where n can be computed as a function of the significance α and the acceptable relative error on the variance estimate ϵ. The table summarizing relevant values of α and ϵ is given in Table 4.

Table 4: Minimum number of observations n as a function of the significance α for having a maximum relative error on the estimated variance ϵ (1D case).

Significance α	*Variance Relative Error -* ϵ								
	0,01	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,08*	*0,09*	*0,10*
0,01	33180	8295	3588	2075	1329	933	679	411	333
0,05	19210	4804	2136	1202	770	535	394	239	194
0,10	13530	3384	1505	847	543	377	278	169	197

For the multivariate case, similar considerations under the assumption of simplified a diagonal Σ lead to

$$\beta = \frac{1}{2}\left[1 - (1-\alpha)^{1/d}\right] \qquad n = 1 + 2\left(\frac{Z_\beta}{\epsilon}\right)^2 \qquad (38$$

so that a similar table involving also the space dimension d can be found. It is relevant to point out that, in accordance with the curse of dimensionality, a sharp increase in n is found

as d increases, so that with large d it is very likely that n will be too big to be matched. A generalization for any covariance Σ shape can be derived, but the n will depend on the particular covariance, so that it is not really helpful.

Another method based on Monte Carlo repetitions is proposed in this thesis.

5.1. Proper sample size: A novel methodology via Monte Carlo simulations

In general, some rule of thumb can be found in the literature for bounding the minimum number of samples for training a classification algorithm. As a general guideline, references [18,19,20,21] suggest that having at least 5 to 10 times as many training samples per class as the number of features is a good practice to follow in classifier design, and that the minimum acceptable ratio is 2. Hence, the rule is:

$$\frac{n}{d} > 5 \div 10 \qquad (39$$

Anyway, based on the considerations about the geometric interpretation of the Mahalanobis distance highlighted in Chapter 4 (section 3.7), a simple methodology for assessing the appropriateness of a sample size n with a particular focus on Novelty Detection is here derived. In particular, the MD was proved to be equivalent to a Euclidean distance on the standardized principal component space. In this respect then, repeating $m = 1000$ draws of size n from a multivariate Gaussian of dimension d the Euclidean distance can be compared to the Mahalanobis distance obtained by estimating the sample mean (whose expected value is the null vector) and the sample covariance matrix (whose expected value is the identity matrix). Hence, the Mahalanobis distance is expected to be equal to the Euclidean, but because of the mean and covariance estimation errors, the two values will start deviating when n is not large enough to fill the d-dimensional space. Fixing some relevant n values and letting d increase, it is possible to compute some statistics on the m values produced by each MC repetition. In particular, the mean value of the two distances and the $\pm\sigma$ confidence interval are given in Figure 26.

Focusing on $d = 30$ (of interest for example in the DIRG dataset analysis), it can be noticed that the MD and ED confidence intervals for $n = 100$ are still intersecting, pointing out that the MD computation is already reliable from a statistical point of view. A $n = 50$ on the contrary would produce untrustworthy results. This confirm the rule of thumb which suggests for $d = 30$ a numerousness $n = 150 \div 300$ and a minimum $n_{min} = 60$.

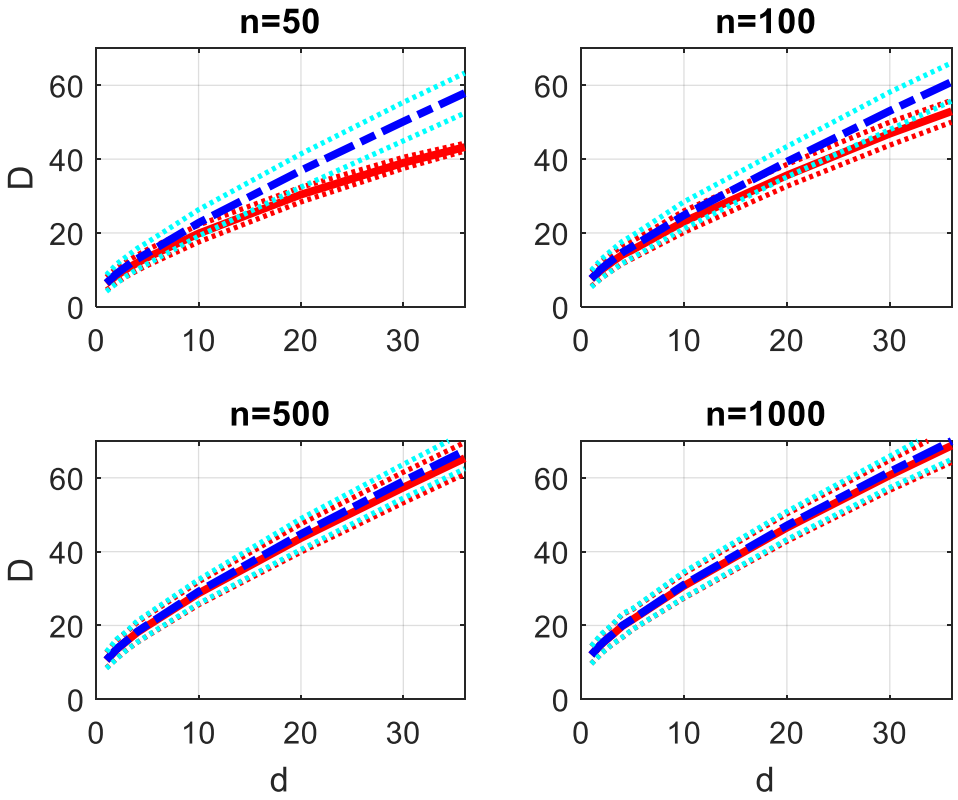


Figure 26: Averages of the Mahalanobis distances (red) and Euclidean distances (blue) for the Maxima with respect of the space dimension d, in 4 sets with growing number of samples n, considering 1000 Monte Carlo repetitions; ±σ confidence intervals of the estimations are also given in cyan (Euclidean) and magenta (Mahalanobis).

6. Conclusions

Throughout all this chapter, Monte Carlo simulations were used to analyse the main issues related to Novelty Detection based on outliers' detection.

First, considerations were made about the optimal threshold to identify outliers from the Mahalanobis distance given a known normal distribution. Extreme Values Theory is also used to confirm the considerations in [13]. A criterion summarizing both the χ^2 and the Wilks's was proposed. Then, the process of estimating the mean and the variance from a training set was accounted and the issues related to finite sample sizes in high-dimensional feature spaces are considered.

The problem of estimating mean and the variance from outliers' inclusive training datasets was also faced introducing and comparing some robust methods for such inclusive outliers' removal. The comparison was done using Monte Carlo Repetitions of a simulated model of a 1DOF oscillator with a clearance, whose response to a white noise excitation was brought to frequency domain in order to find a multidimensional feature space.

The Minimum Covariance Determinant in its FAST implementation proved to be the best method in terms of robustness.

Finally, the evaluation of the minimum number of samples to ensure a meaningful Mahalanobis Distance Novelty Detection was finally performed using an original intuition which exploits the Mahalanobis distance geometric interpretation.

Bibliography

[1] K. Worden, G. Manson, N. R. J. Fieller, "*Damage detection using outlier analysis*", Journal of Sound and Vibration (2000). DOI: 10.1006/jsvi.1999.2514

[2] Benjamin Peirce, "*Criterion for the rejection of doubtful observations*", The Astronomical Journal II, 45, 161- 163 (1852).

[3] B.A. Gould, "*On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application*", The Astronomical Journal IV, 83, 81-87 (1855).

[4] S. Ross, "*Peirce's Criterion for the Elimination of Suspect Experimental Data*", Journal of Engineering Technology, 20(2), 1–12 (2003).

[5] K. Worden, D. W. Allen, H. Sohn, C. R. Farrar, "*Damage detection in mechanical structures using extreme value statistics*", Proceedings of SPIE - The International Society for Optical Engineering, 4693, 289–299 (2002). DOI: 10.1177/1475921704041866

[6] D. Toshkova, N. Lieven, P. Morrish, P. Hutchinson, "*Applying Extreme Value Theory for alarm and warning levels setting under variable operating conditions*", EWSHM (2016). www.ndt.net/search/docs.php3?showForm=off&id=20060

[7] Ian Jordaan, "*Decisions under Uncertainty - Probabilistic Analysis for Engineering Decisions*", Cambridge University Press (2005). ISBN-10: 0521369975

[8] Rinya Takahashi, "*Normalizing constants of a distribution which belongs to the domain of attraction of the Gumbel distribution*", Statistics & Probability Letters Volume 5, Issue 3, Pages 197-200, April 1987. DOI: 10.1016/0167-7152(87)90039-3

[9] J. W. Tukey, "*Exploratory Data Analysis"*, Addison-Wesley (1977).

[10] Rossi F., "*Theory of Semiconductor Quantum Devices*", Springer, 2011. DOI: 10.1007/978-3-642-10556-2

[11] Brandimarte P., "*Handbook in monte carlo simulation. Applications In Financial Engineering, Risk Management, And Economics*", John Wiley And Sons Ltd, 2014.
ISBN: 0470531118

[12] Hubert M., Debruyne M., "*Minimum covariance determinant*" Wiley Interdiscip. Rev. Comput. Stat., 2010. DOI: 10.1002/wics.61

[13] Penny, K. I. "*Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance*". Journal of the Royal Statistical Society: Series C (Applied Statistics), 45(1), 73-81, 1996. DOI: 10.2307/2986224

[14] Ledoit O., Wolf M., "*Honey, I Shrunk the Sample Covariance Matrix*", The Journal of Portfolio Management. 30, 2003. DOI:10.2139/ssrn.433840

[15] Schäfer J., Strimmer K. "*A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics*", Statistical applications in genetics and molecular biology. 4. Article 32, 2005. DOI:10.2202/1544-6115.1175

[16] Theiler J. "*The incredible shrinking covariance estimator*", Proceedings of SPIE - The International Society for Optical Engineering. 8391. 22, 2012. DOI:10.1117/12.918718

[17] Gupta P. L., Gupta R. D., "*Sample size determination in estimating a covariance matrix*", Computational Statistics & Data Analysis, Volume 5, Issue 3, 1987. DOI:10.1016/0167-9473(87)90014-4

[18] Jain AK, Duin RPW, Mao J., Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell. 2000

[19] A.K. Jain and B. Chandrasekaran, Dimensionality and Sample Size Considerations in Pattern Recognition Practice, Handbook of Statistics. P.R. Krishnaiah and L.N. Kanal, eds., vol. 2, pp. 835-855, Amsterdam: North-Holland, 1982

[20] Duin R. P. W., On the accuracy of statistical pattern recognizers, Dutch Efficiency Bureau, 1978. ISBN 90 6231 052 4

[21] Raudys S., Pikelis V., On Dimensionality, Sample Size, Classification Error, and Complexity of Classification Algorithm in Pattern Recognition, IEEE transactions on pattern analysis and machine intelligence, VOL. PAMI-2, No. 3, May 1980

[22] Duchi J., "Derivations for Linear Algebra and Optimization", 2014, University of California, Berkeley; URL: http://www.eecs.berkeley.edu/∼jduchi/projects/general notes.pdf.

[23] Dervilis N., Cross, E.J., Barthorpe R., Worden K., "Robust methods of inclusive outlier analysis for structural health monitoring", Journal of Sound and Vibration, 2014. 10.1016/j.jsv.2014.05.012.

[24] L. A. Bull, K. Worden, E. J. Cross, N. Dervilis, "Outlier Ensembles: an Alternative Robust Method for Inclusive Outlier Analysis with Structural Health Monitoring Data", 9th European Workshop on Structural Health Monitoring July 10-13, 2018, Manchester, United Kingdom.

Chapter 8:

Machine Learning for Continuous Monitoring: Comparison of the selected algorithms over real life applications

1. Introduction

The Machine Learning algorithms selected in Chapter 4 were first experimented on the dataset collected from the DIRG test rig, conceived to test high speed aeronautical bearings. In second instance, the real-life acquisitions from the Italian windfarm were used to assess the suitability of the considered methodology, summarized in the block scheme introduced in chapter 4 and here reported in Figure 1.

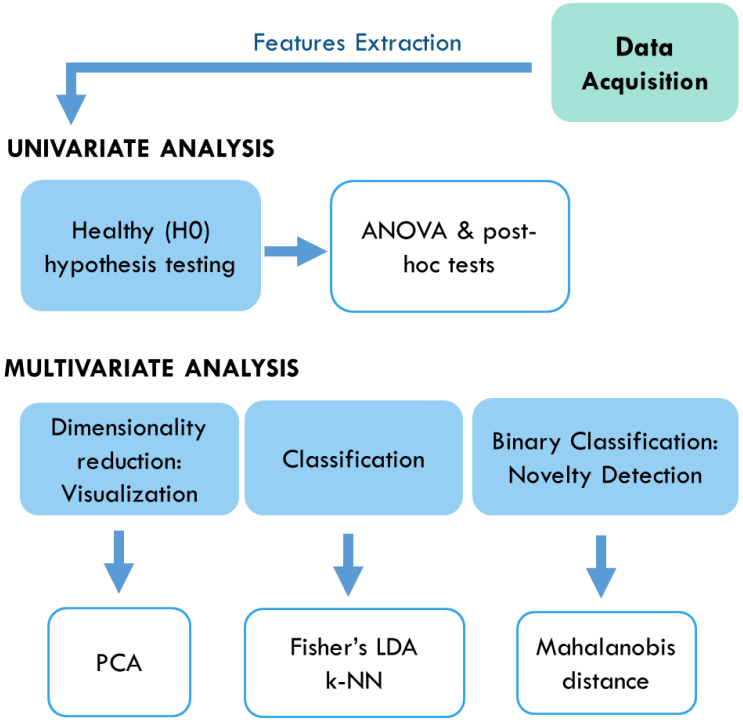


Figure 1: The proposed continuous monitoring methodology.

The analysis started with an explorative univariate analysis of variance (ANOVA), to verify the presence of diagnostic information in the dataset. Then, in order to "condensate" the information contained in the different features enhancing the effect of damage, multivariate analyses were considered. In particular, Fisher's Linear Discriminant Analysis (LDA), k-Nearest Neighbours (k-NN) classification, Principal Component Analysis (PCA) and Novelty Detection (ND).

2. DIRG test rig data analysis – part 1

Root mean square, skewness, kurtosis, peak value and crest factor (peak/RMS) were computed on 0,1 s chunks of the original 10 s acquisitions generating 100 data points for each of the 17 acquisitions described in Chapter 4. Figure 2 from the same chapter is then reported for completeness, to show the dataset, composed by 7 differently damaged conditions, from 0A (healthy), to 6A, containing 1700 measurements in a 30-dimensional space made by the 5 features extracted from the 6 channels (2 sensor positions and 3 possible directions in the space). The sample size $n = 100$ is chosen in accordance to the reflections in Chapter 7 about the curse of dimensionality.

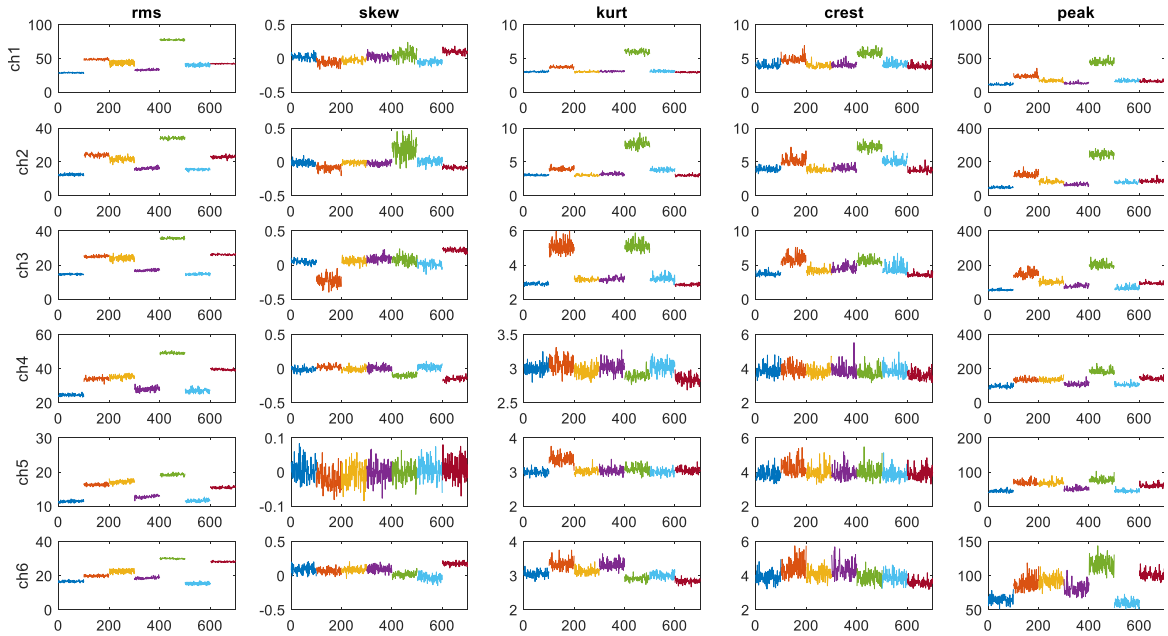


Figure 2: Condition 12 pictured by all the feature/channel combinations in all the health conditions, from 0A to 6A (100 samples each).

2.1. ANOVA and post-hoc for diagnostics of high-speed bearings

In order to assess the diagnostic-ability of each channel (corresponding to a sensor measurement direction and positioning) and feature combination, the ANOVA is performed for each channel and feature to assess the diagnostics-ability of each combination. Although the assumptions of normality and homoscedasticity (Appendix 4) are not completely met, ANOVA is generally considered robust to these violations, in particular for the case in which all the groups under analysis show equal numerousness, so that the method can be applied quite confidently.

In all the 30 tests, the ANOVA p-values result almost negligible, meaning that the omnibus null hypothesis H_0 is rejected at a very high confidence. A significant difference is detectable among the groups mean values. Obviously, this does not tell much about the real effect size, furthermore, ANOVA is not able to add anything about which groups are the farthest and with respect to which other group, so that a more informative multi-comparison post-hoc test can be very helpful. LSD limits can be then computed. A graph summarizing the confidence interval around the healthy sample is given in Figure 3. There, it is easy to notice that *kurtosis* and *crest* are probably the best features, able to discriminate the most damaged 1A and 4A conditions in all the channels. They also seem to be quite

consistent with the damage level. For example, focusing on channel 4, it's easy to notice a linear relationship between the distance from the healthy reference and the damage level.

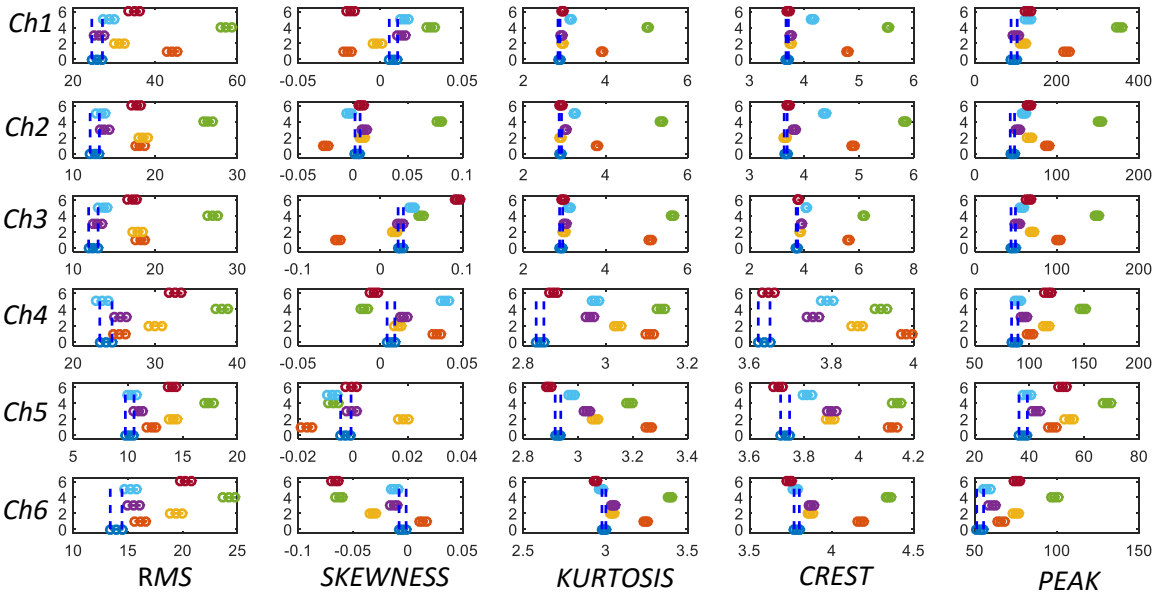


Figure 3: ANOVA post-hoc, Multicomparison result. For different channels and features, all the 6 damage conditions (from 0A to 6A, respectively: blue, orange, yellow, violet, green, light blue, red) are compared to the healthy reference (0A) through LSD limits.

In any case, ANOVA and the post-hoc test prove that the damage effect is detectable on all the considered features-channels combinations with the given sample size. It is wise then to condensate all this information into a single, unique analysis via multivariate statistics.

This could be performed through the multivariate analysis of variance (MANOVA). Unfortunately, this hypothesis test shows the same limitations of ANOVA. Hence, instead of focusing on p-values of non-immediate interpretation, a multivariate classification is preferred. As already introduced in fact, classification is just another point of view on the same subject. Furthermore, extending the hypothesis testing considerations about effect size, the Fisher's Linear Discriminant Analysis (LDA) can be easily derived.

2.2. LDA and k-NN classification

The Fisher's LDA classifier here introduced is tested on the DIRG test rig data and compared to the k-NN classifier. The dataset features 1700 points (i.e. 100 measurements per each of the 17 different combinations of load and speed) in a 30-dimensional space (6 channels, 5 features). In agreement with the classification philosophy, the samples of size $n = 100$ are divided in two subsamples $n_{tr} = 60$ and $n_{val} = 40$ for training (i.e. in sample) and out of sample validation respectively. Unequal sample sizes are selected because with a feature space dimensionality $d = 30$ it would be meaningless to train algorithms using less than $n_{tr} = 60$ (refer to Chapter 6 for considerations about the minimum sample size).

Seven differently damaged conditions, from 0A (healthy), to 6A are present in the dataset, so that the error table (meant for maximum two classes) becomes more complex and takes the name of confusion matrix. In order to assess the classifier performance both in and out of sample, two of such confusion matrices are produced, collecting the (rounded) percentages of the classified samples for the different expected classes.

Table 1 highlights that LDA, despite being very sensitive to 1A and 4A damages, which are correctly classified in more than 90% of cases, shows some troubles in distinguishing the other damages and in particular the healthy condition, which is correctly classified only in less than 40% of cases.

Table 1: LDA confusion matrices in rounded percentages by columns, computed on the training set (in sample) and on the validation set (out of sample)

LDA		*in sample* Target Class							*out of sample* Target Class						
		0A	1A	2A	3A	4A	5A	6A	0A	1A	2A	3A	4A	5A	6A
Output class	0A	38	0	11	13	0	9	17	37	0	11	15	0	10	16
	1A	9	93	1	5	2	6	0	10	92	1	4	2	5	0
	2A	11	0	60	15	0	4	8	13	0	57	17	0	5	9
	3A	18	0	12	47	0	10	8	17	0	13	43	0	12	8
	4A	1	5	1	0	97	9	0	1	6	0	1	97	10	0
	5A	13	1	7	6	0	55	6	12	2	9	9	1	51	5
	6A	10	0	9	14	0	7	61	11	0	9	12	0	6	63

As expected, the linear classifier limits arise. In this respect, k-NN classifier is tested to understand whether the difficulties in the classification are related to the separability of the dataset or to the LDA algorithm. Table 2 highlights that with k-NN the misclassifications are very limited both in sample than out of sample. Only 0A and 3A (the weakest damage on the inner race) are sometimes confounded, so that the validation of the classifier confirms the separability of the different damaged conditions in the multidimensional feature space, given the selected features.

Table 2: k-NN confusion matrices in rounded percentages by columns

k-NN		*in sample* Target Class							*out of sample* Target Class						
		0A	1A	2A	3A	4A	5A	6A	0A	1A	2A	3A	4A	5A	6A
Output class	0A	88	0	1	8	0	2	1	83	0	1	8	0	2	2
	1A	0	96	0	0	1	0	2	0	93	0	0	2	0	2
	2A	1	0	95	2	0	1	1	1	0	94	2	0	1	1
	3A	12	0	2	83	0	2	2	11	0	2	83	0	2	2
	4A	0	4	0	1	94	0	0	0	4	0	1	98	0	0
	5A	3	2	1	3	0	90	2	2	2	1	3	0	92	2
	6A	3	2	2	3	0	2	89	2	2	2	3	0	2	92

In conclusion, a very good classification is possible as the groups are parted very well in the 30-dimensional feature space, even if a linear separation is not the optimal. All the different damages are far enough not to be confounded, but they are also quite far from

the healthy condition. Indeed, focusing on a binary classification without distinguishing among the different damage levels (i.e. level 1 diagnostics, healthy vs damaged), the confusion matrix simplifies to the reported error table (Table 3). In this case it is easy to find the corresponding type I error rate (the significance α) of about 20-30% for LDA and 10-15% for k-NN and the type II error rate (β) of 13% for LDA and 2% for k-NN.

Table 3: LDA and k-NN confusion matrices in rounded percentages by columns for a two-class case

		in		*out*	
		Target		Target	
LDA		H	D	H	D
	H	77	13	71	13
	D	23	87	29	87

		in		*out*	
		Target		Target	
k-NN		H	D	H	D
	H	89	2	85	2
	D	11	98	15	98

Despite a separation is with no doubt present between the healthy and the damaged conditions, such distance cannot be visualized because of the feature space being multivariate. At an exploratory stage the visualization can be very useful. At this scope, the Principal Component Analysis is introduced in next chapter.

2.3. PCA visualization of DIRG test rig data

In order to visualize the 30-dimensional dataset (6 channels, 5 features) corresponding to the DIRG acquisition, it was summarized by a 2D representation through PCA. In particular, the 17 healthy acquisitions (different combinations of load and speed) of 100 samples each were used (after mean centring) to produce a reference covariance matrix. Its decomposition via PCA produces, after selection of the first two principal components, the graph reported in Figure 4-above.

From the picture, neglecting the zero-load condition (1,5,9,13 labels), which is anyway not very meaningful, the data could be clustered in equal speed subgroups, almost regardless from the load (2-3-4, 6-7-8, 10-11-12 and 14-15 clusters). Just the highest load acquisitions (16 and 17) proves to remain out of this scheme.

In order to visualize the variability related to the damage, an analogous procedure can be followed. Focusing on acquisition 12 (280 Hz, 1800 N) for example, its corresponding healthy sample (0A) can be used to "train" PCA, that is, to find the rotation matrix to PCs. Applying the same transform to all the other samples of acquisition 12, Figure 4-below can be produced.

In this case the diagnostic information is very effectively pictured, as the most damaged conditions (1A and 4A) result to be the furthest from the healthy cluster. Remembering that this dimensional reduction is just a simplified projection which is neglecting a lot of information, this underlines once more the separability of the difference health conditions. Moreover, the influence of the work condition on the data distribution can be appreciated, highlighting the necessity of some strategy to compensate for such confounding factor.

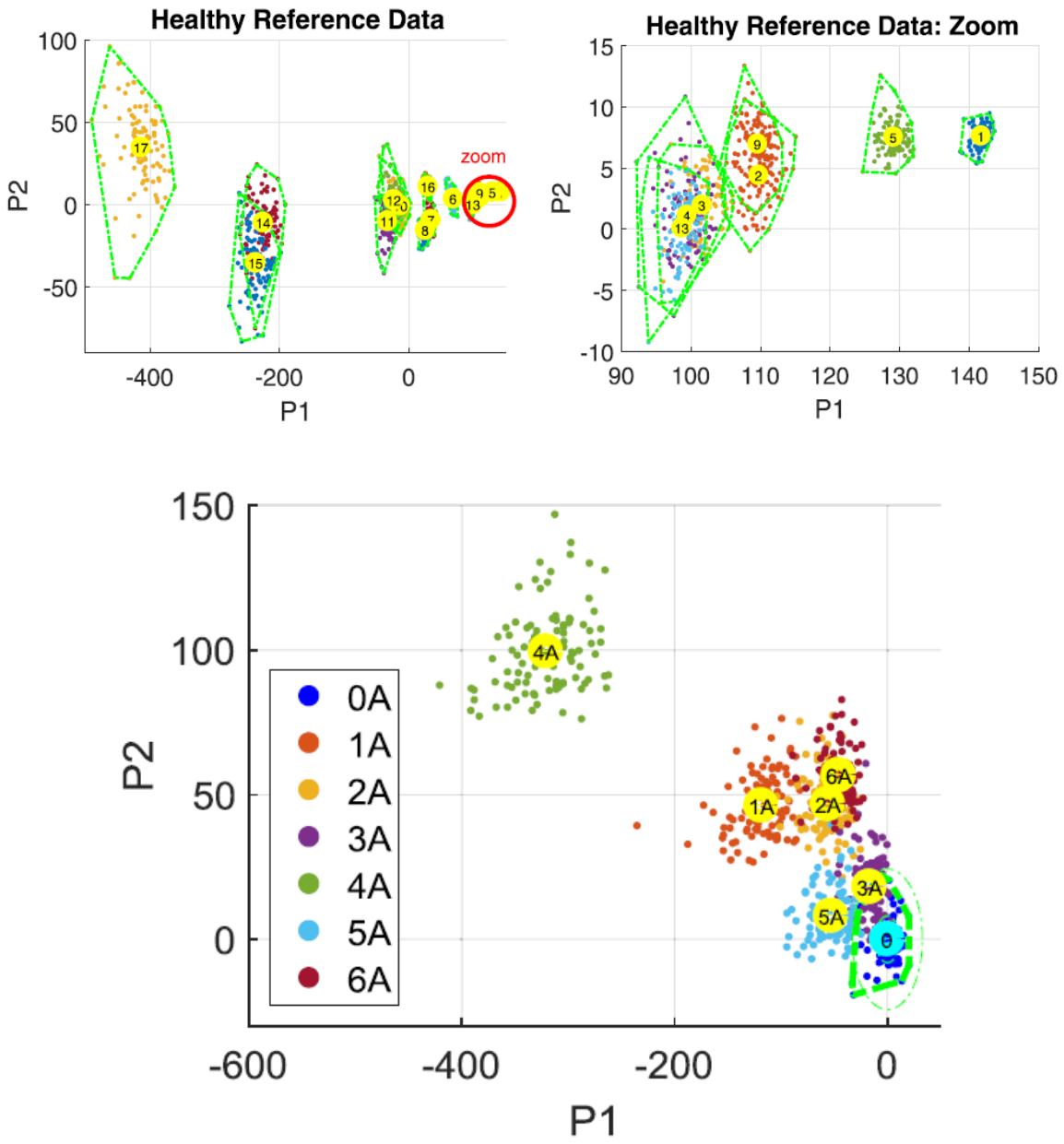


Figure 4: Above- Healthy data for the 17 speed and load combinations. Below- Healthy data compared to damaged acquisitions, centred on the same reference (work condition 12)

2.4. Multivariate Novelty Detection

The novelty detection procedure was first implemented and tested on the DIRG test rig data. Differently from the classification proposed in section 2.2, the healthy condition (0A) alone is taken as reference for computing the training covariance matrix. In a first trial stage, all the 17 different speed and load conditions are considered as a whole in the training phase, where a threshold is also produced using the 99th percentile (significance $\alpha = 1\%$) of the maxima distribution from 1000 Monte Carlo repetitions. The NIs computed for the entire dataset, are reported in Figure 5.

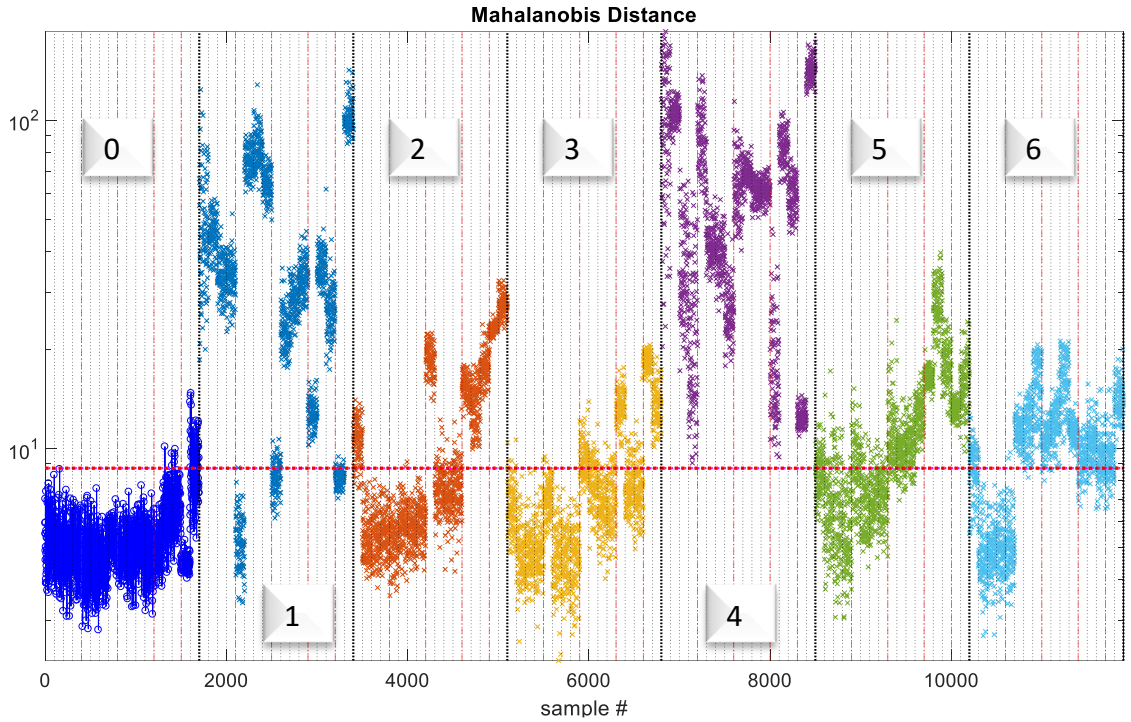


Figure 5: Mahalanobis distance referred to the entire availablet training dataset, without discerning the different operational conditions; in red the threshold: the 99th percentile of the maxima distribution from 1000 Monte Carlo repetitions.

Table 4: Missed and False Alarms of MD-ND

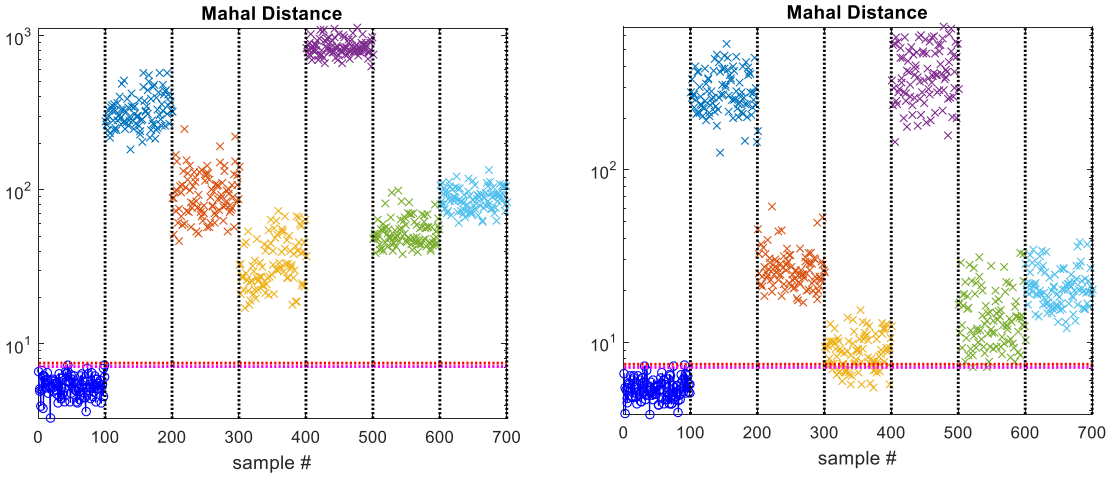
	0A	1A	2A	3A	4A	5A	6A
Alarms	2,9%	15,3%	56,9%	73,4%	0,0%	40,0%	33,7%

In the picture, the effect of the strong variations of the speed and load is already evident in the healthy (training) data. Although all the conditions are used in the training, the too large range of variation of speed and load causes a high rate of False Alarms (FA). Furthermore, only the biggest damages can be recognized effectively, while many Missed Alarms (MA) are present in all the other damaged conditions (Table 4).

An independent analysis for each of the 17 measured operational conditions can be then performed, reporting the results in terms of FA and MA in Table 5. These results are graphically summarized in Figure 6 for just a couple of interesting conditions (work conditions 3 and 12). In this case the results are really improved, as the alarm rates are very low for almost all the conditions, while just at low speed, for conditions 3 and 4 it seems difficult to detect the 3A damage. Furthermore, a consistency between NIs and the damage is evident, as they vary almost monotonically with the defect severity. This could be used to extract further information about the size of the damage. Additionally, in most of cases (as in Figure 6 - condition 12) all the damaged conditions show a wide distance from the healthy state, so that the threshold may be increased to reduce the FA rate without worsening the MA rate (i.e. to increase the power leaving unaffected the confidence).

The ability of the Mahalanobis distance of compensating for linear or quasi-linear variations of the operational conditions, in accordance to what introduced in the previous section, can be easily tested on this experimental data. For this purpose, a test at a constant speed of 300 Hz while the load changes from 0 to its maximum value is performed,

agglomerating acquisitions 9 to 12. Indeed, training the algorithm on a variable-load healthy set, it is possible to filter out quite well the effect of load variation. This is highlighted in Figure 7, where a table summarizing the False and the Missed Alarms is also reported.

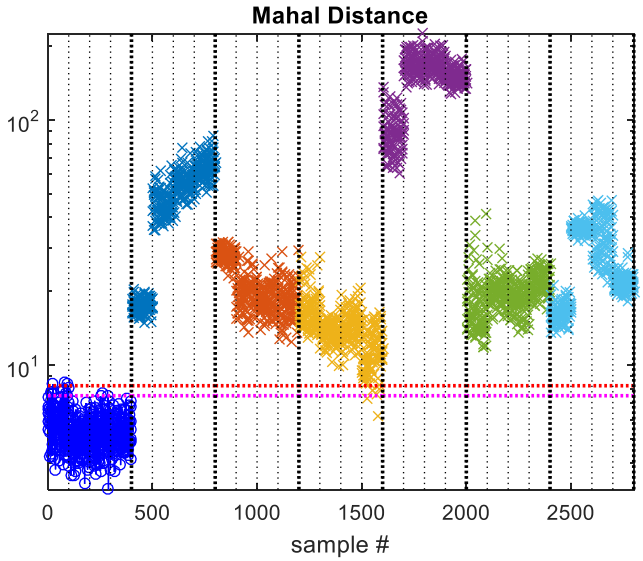


Work condition 12 *Work condition 3*

Figure 6: Mahalanobis distance for two different operational conditions.

Table 5: False and Missed Alarms in % for the 17 operational conditions, considered independently and compared to their own reference healthy acquisitions (see Figure 6); the 99th percentile of the maxima distribution from 1000 Monte Carlo repetitions, used as a threshold, is reported as well.

	FA	MA						MC 99% threshold
	0A	**1A**	**2A**	**3A**	**4A**	**5A**	**6A**	
1	2	0	0	0	0	0	0	7,42
2	2	0	0	0	0	0	0	7,37
3	3	0	0	21	0	2	0	7,38
4	4	0	0	7	0	0	0	7,46
5	3	0	0	0	0	0	0	7,42
6	1	0	0	0	0	0	0	7,43
7	1	0	0	0	0	0	0	7,37
8	3	0	0	0	0	0	0	7,41
9	3	0	0	0	0	0	0	7,44
10	1	0	0	0	0	0	0	7,37
11	4	0	0	0	0	0	0	7,41
12	2	0	0	0	0	0	0	7,39
13	4	0	0	0	0	0	0	7,43
14	2	0	0	0	0	0	0	7,49
15	4	0	0	0	0	0	0	7,42
16	1	0	0	0	0	0	0	7,40
17	0	0	0	0	0	0	0	7,38
							average threshold:	7,41



Alarms:	
0A	*3 %*
1A	*0 %*
2A	*0 %*
3A	*0,5%*
4A	*0 %*
5A	*0 %*
6A	*0 %*
MC 99% threshold: 8,2	

Figure 7: Mahalanobis distance for operational conditions 9-12 (constant speed 300 Hz).

2.5. Kernel Density Novelty Detection

In order to improve the novelty detection, it is possible to switch from distance to probability density so that a non-parametric density estimator can be used to infer a generic multivariate pdf on which a threshold can be fixed. Given the space dimensionality ($d = 30$) KDE is not really advisable, but it is anyway tested so as to prove the validity of the change of paradigm in Novelty Detection. If a 30-dimensional Gaussian Kernel function with a bandwidth $h = 1{,}2$ is used to estimate the pdf from the healthy reference dataset, the estimation can be easily extended to the damaged acquisitions, whose probability density values are reported in the graph in Figure 8. Table 6 summarizes the performance in terms of missed alarms when a threshold ensuring no false alarms is used.

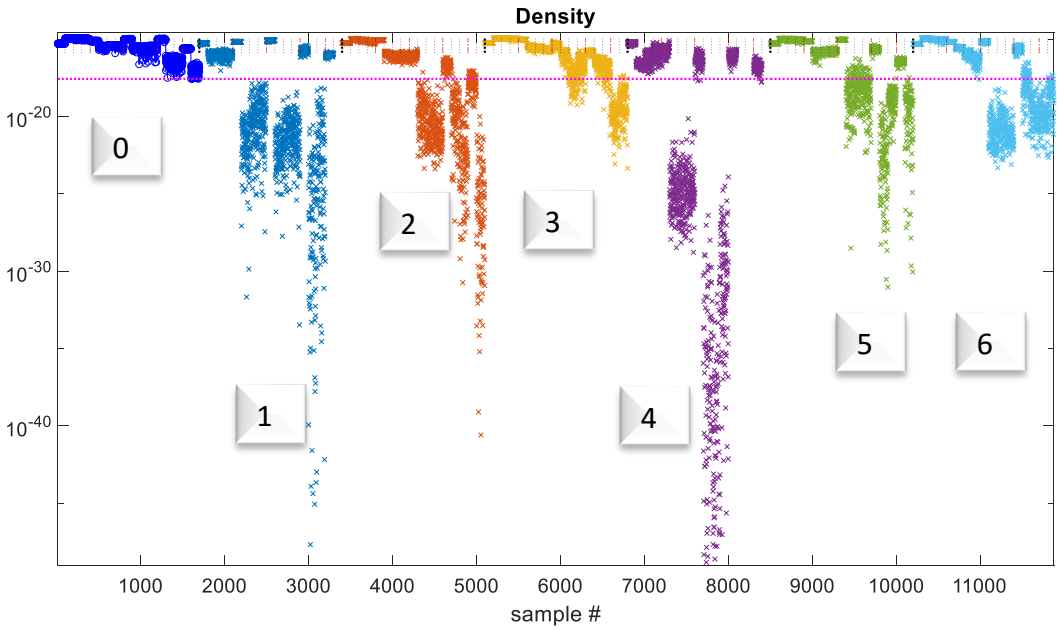


Figure 8: Probability density Novelty Indices from KDE trained without discerning the different operational conditions; in magenta the best separation threshold.

Table 6: Missed and False Alarms of KDE-ND

	0A	1A	2A	3A	4A	5A	6A
Alarms	0,0%	47,1%	60,8%	84,1%	46,8%	68,2%	60,9%

Comparing Table 6 to Table 4, it is easy to notice that the classification performance is decreased. In general, it would be possible to improve KDE-NE by the selection of an optimal bandwidth, but given the drawbacks of KDE related to the curse of dimensionality (i.e. the pdf tends to flatness as the features space dimension increases) it is not worth to spend time in this optimization. Furthermore, to perform the classification, all the training points must be kept stored, leading to possible storage issues. It is much more advisable to switch back to parametric models but of a more complex nature, such as the Gaussian Mixture Models, as proposed in the next section.

2.6. Gaussian Mixture Novelty Detection

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. The weights and the Gaussians parameters can be estimated from training data using the iterative Expectation-Maximization (EM) algorithm, but in this case the number of mixtures m must be selected by the user on the basis of the particular dataset.

In this regard, a prior analysis is run on the training data with several values of m (in the range $1 \div 20$). The likelihood of the data to be drawn by such 20 GMMs is then evaluated so that the Negative Log Likelihood (NLL), the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) can be compared to assess the relative quality of the 20 statistical models for the given training set. The best number of mixtures, according to the result reported in Figure 9, is $m = 8$, so that this value was used for the following Novelty Detection.

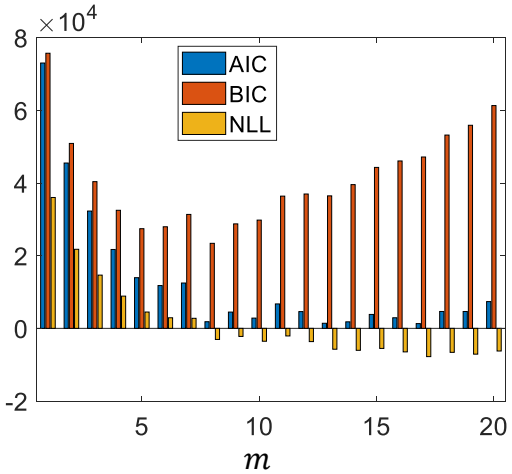


Figure 9: Relative quality of the 20 GMMs on the basis of Negative Log Likelihood, AIC and BIC criteria.

The novelty detection in terms of probability density is summarized by Figure 10, while in Table 7 is reported the result in terms of false and missed alarms using a threshold which ensures no false alarms.

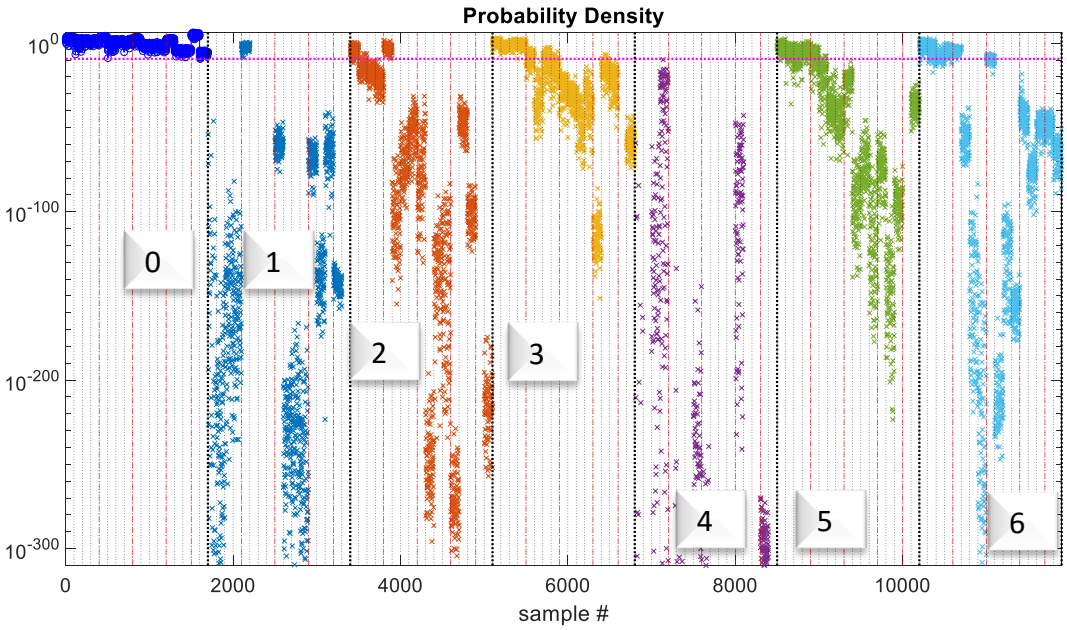


Figure 10: Probability density Novelty Indices from GMM trained without discerning the different operational conditions; in magenta the best separation threshold.

Table 7: Missed and False Alarms of GMM-ND

	0A	1A	2A	3A	4A	5A	6A
Alarms	0,0%	5,7%	16,1%	29,4%	0,0%	21,5%	26,3%

Comparing Table 7 to Table 4 it is easy to notice that when a training which does not discern the different operational conditions is used, GMM-ND outperforms the traditional MD-ND. The increase in the model complexity is then justified in this case.

2.7. Note about probability density values in Matlab®

The estimated psd values are very small, as wall as the ANOVA p-values obtained by Matlab® environment, so that one could ask if these values are "logic" and the computation makes sense.

In general, the so called "machine precision" is considered the smallest number (usually called eps) such that the difference between 1 and 1 + eps is nonzero, representing then the smallest difference between two numbers that the computer recognizes. A 32-bit computer with IEEE® double precision has then an eps of 2^{-52} (or $2{,}22\ 10^{-16}$).

Nevertheless, that precision does not represent the smallest number that can be stored in a computer and is not the same for the whole range of values.

The smallest positive normalized floating-point number in IEEE® double precision is equal to 2^{-1022} (or $2{,}22\ 10^{-308}$), while the maximum is $1{,}8\ 10^{+308}$ (namely realmin and realmax in Matlab®). In this range, the machine precision varies in a deterministic way so as to be always orders of magnitude smaller than the represented value. For example, $eps(realmin) = 4.9^{-324}$ and $eps(realmax) = 2\ 10^{+292}$. In between the trend can be visualized with a graph, such as the one reported in Figure 11.

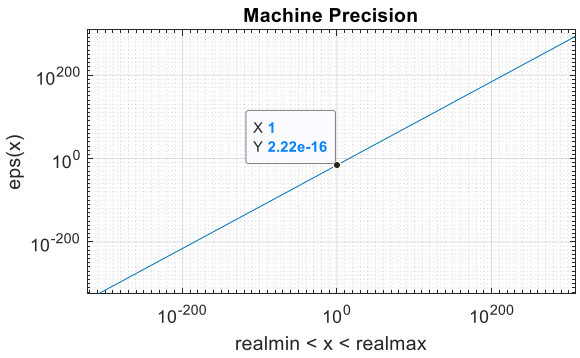


Figure 11: Machine precision eps in the representable range with IEEE® double precision

3. DIRG test rig data analysis – part 2: confounders compensation

The second DIRG dataset involves run down acquisitions while the speed is reducing from 470 to 0 Hz and the load is set to 0, 1000, 1400 and 1800 N (respectively condition 1,2,3 and 4). The acquisitions last about T = 50 s at a sampling frequency fs = 102400 Hz, so that the features extracted on independent chunks of 0,5s lead to 100 samples per each of the 6 signals (6 channels from 2 tri-axial accelerometers) in the 7 health conditions (from 0A to 6A) at the 4 different loading conditions. The dataset for null load is shown in Figure 12.

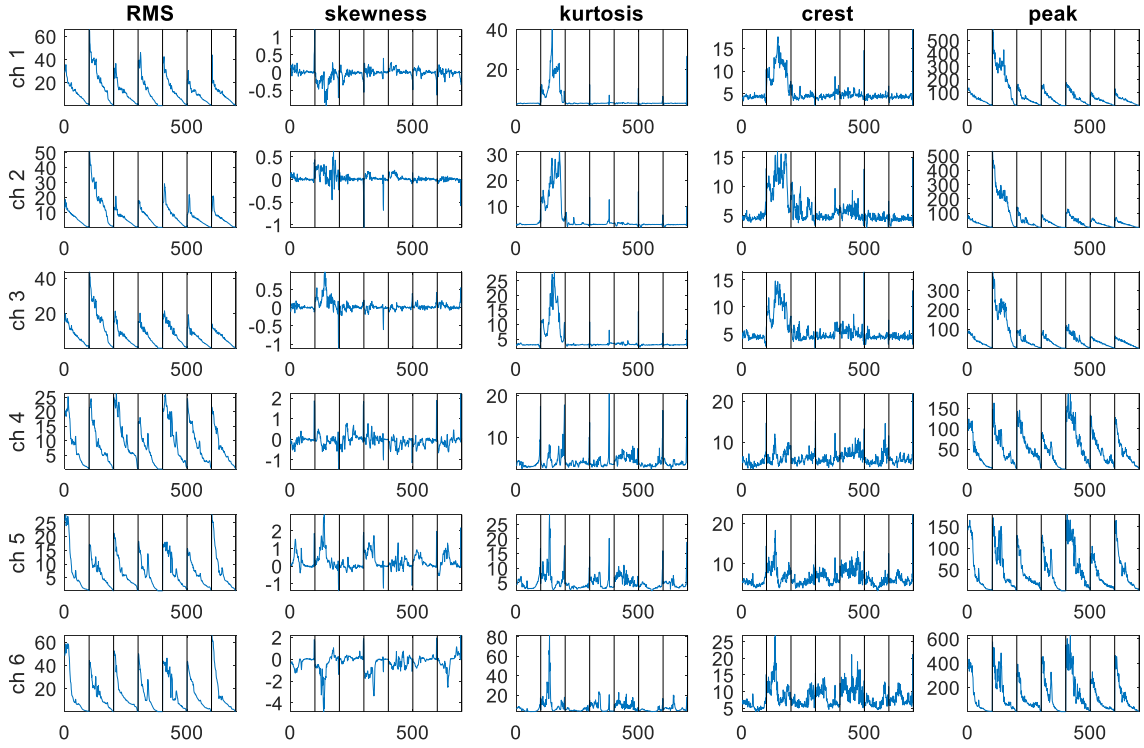


Figure 12: The considered dataset after features extraction for 1st load condition (0 N) while the speed is decreasing until a stop starting from 470 Hz. The black dotted lines divide the different damage conditions (0A to 6A). For each, 100 observations are plotted sequentially.

The dataset collecting 5 simple time features per each of the 6 channels (i.e., 30 features) is analysed separately for the 4 different load conditions. The confounding effect of the reducing speed (from 470 Hz to 0 Hz) will be compensated during the healthy training with 3 different approaches which will be compared:

- Plain Euclidean distance (raw)
- Mahalanobis distance,
- PCA orthogonal regression and whitening,

The novelty indices for the healthy reference and for the damaged conditions are reported in Figure 13 for condition 4 (load 1800N, with decreasing speed). This graph highlights relevant considerations previously introduced.

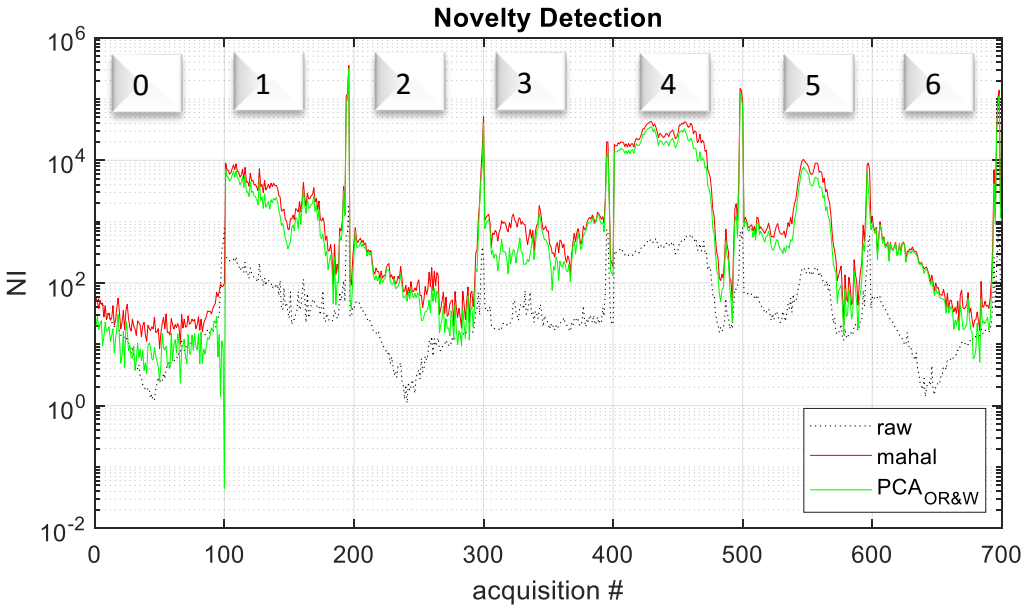


Figure 13: The Novelty Detection with the different NIs at maximum load. 0-100 samples are the healthy reference, 100-200 corresponds to 1A damage, and so on until 600-700 coming from 6A damage.

In particular, Raw Euclidean NIs (samples 0-100) are strongly non-stationary as a trend is clearly visible by eye. The Mahalanobis NIs are non-stationary too, but a slight improvement is obtained by neglecting the first 20 whitened principal components and focusing on the subspace individuated by the last 10 (PCA orthogonal regression). An additional note should be added to explain that the behaviour of the NIs at the end of all the run-down is ascribable to the fact that the record is not stopped exactly when the machine stops, so that the last points are practically acquiring just noise as the machine is already at rest.

Despite the NIs curve already gives a qualitative impression of good detectability of almost all the different damages (from 1A to 6A), a quantitative comparison of the performances of the different methods is necessary. At this scope, Figure 14 reports the ROC curves for all the 4 different load conditions (overall for all the damaged conditions).

All the 4 graphs lead to a similar result: in this particular application of damage detection in case of non-stationary rotational speed, Mahalanobis novelty detection proves to perform well. Nevertheless, removing the first 20 principal whitened components from the NIs computation (PCA-Ortogonal Regression & whitening), the ND results are further improved.

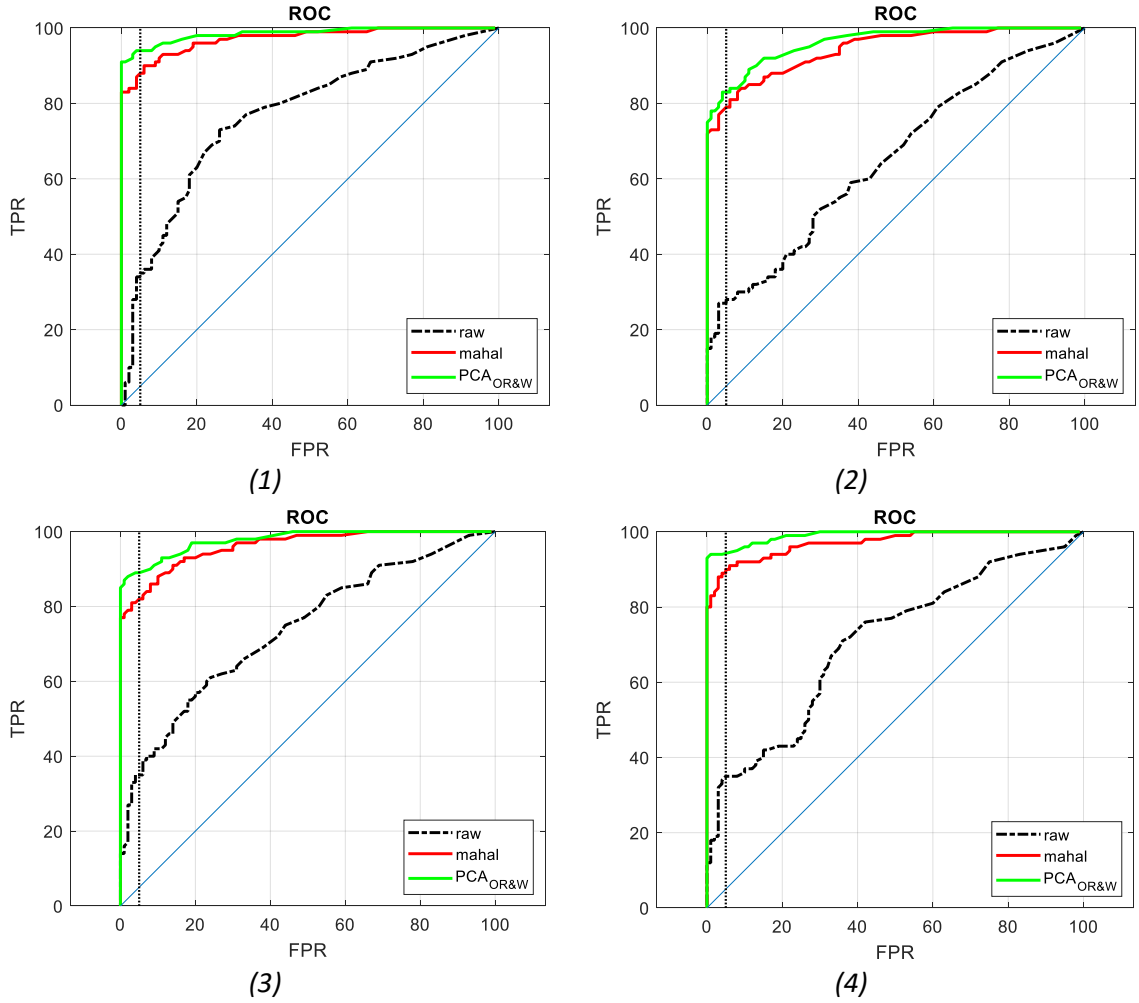


Figure 14: The ROC for Novelty Detection in the 4 loading conditions, respectively 1800, 1400, 1000, 0N. The 5% FPR point is highlighted.

4. Italian windfarm data analysis

The same methodology was tested also on the real-life application concerning multimegawatt windmill gearboxes from the Italian windfarm in Molise region. 4 acquisitions from 3 gearboxes were considered. The first two were acquired from WTG01 and WTG03 at the same time and were used as calibration (training). The second two area from WTG03 and WTG06 (the damaged one) and were recorded in a second moment but in similar environmental and operational conditions, so that confounding influences can be neglected. The dataset is reported in the following Figure 15.

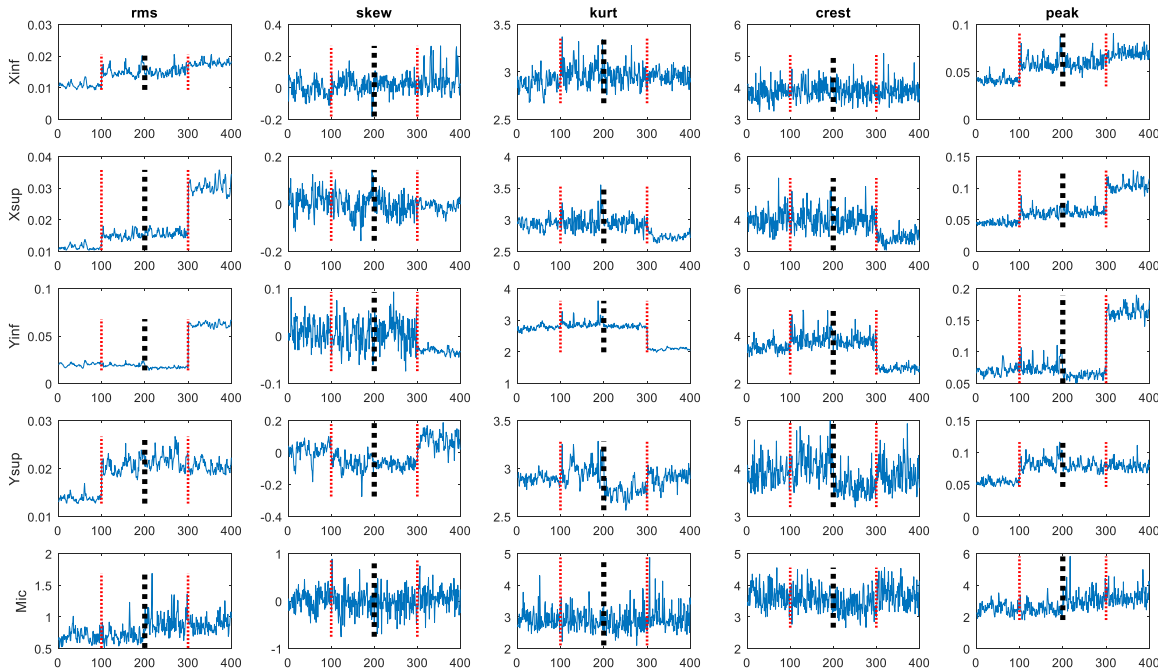


Figure 15: The Italian windfarm dataset. Samples 0-100 are from WTG01 @17.20, 101-200 from WTG03 @ at 17.20, 201-300 from WTG03 @ 15.00 and 301-400 from WTG06 (the damaged wind turbine) @ 15.00. The first 2 sets are used for calibration and are separated by the black dotted line from the last 2, left for validation. X is the wind direction; Y is the orthogonal. Inf is a position on the tower 2m from the ground, Sup is 7m from the ground (N.B. the Nacelle is located about 100m from the ground).

4.1. Hypothesis testing of two means

The hypothesis testing of two population means, which is a particular case of ANOVA for 2 groups only, is performed on the here considered data regarding the Italian windfarm. The dataset is divided in 2 groups: the healthy one contains the first 300 observations from the healthy windmills (WTG01 and WTG03), while the last 100 observations, coming from the damaged WTG06, are labelled as damaged. The assumption of normality can be considered verified with enough confidence. The same does not hold for the homoscedasticity, but the ANOVA is commonly considered robust to such violations, so that the trustworthiness of the results will not be affected. It is relevant to remember that the ANOVA is a univariate technique, so it will be repeated per each channel and feature combination (25 times). This enables to make some considerations about the more relevant channels and features for diagnosing a damage. The results are reported in Table 8.

Focusing on this table, it is easy to notice that the p-values are in general very small, implying the rejection of H_0. The damage is then proved to be detectable also using the simple time-domain features proposed. This is true in particular for channel Yinf, which shows the smallest p-values. On the contrary, channel Mic, channel Ysup and channel Xinf are less performing in detecting the damage using Skewness, Kurtosis and Crest factor.

Because of this, considering also the different nature of the Mic acquisition, the fifth channel will not be considered in further analyses, which will try to aggregate the diagnostic information of all the 5 features from the 4 accelerometers, using a multivariate approach.

Table 8: ANOVA p-values for the different channel-feature combinations – Italian windfarm dataset.

Feature \ Channel	**Xinf**	**Xsup**	**Yinf**	**Ysup**	**Mic**
RMS	2.33 e-48	5.24 e-220	0	2.71 e-05	1.15 e-08
Skewness	2.74 e-06	0.033	1.51 e-34	1.42 e-54	6.14 e-03
Kurtosis	0.330	4.02 e-62	2.66 e-222	0.019	0.023
Crest factor	0.661	8.26 e-50	1.81 e-117	0.646	1.53 e-4
Peak	1.81 e-40	3.54 e-160	9.22 e-260	2.06 e-05	2.92 e-18

4.2. PCA visualization

The result of the PCA applied to the centred, healthy reference set (WTG01 and 03 at 17.20) in the 20-dimensional space (4 channels, 5 features) is reported in Figure 16, where the validation set is also projected according to the same mapping.

In Figure 16 one can easily notice that 2 clusters arise. The damaged acquisitions (in red) can be clearly separated by all the other healthy points (both from the calibration and the validation sets). The first component is then enough to perform the damage detection. In order to compare the weights of the features involved in the linear combination producing the first principal component, a PCA is repeated on the standardized features (centred and normalized on their standard deviation). The absolute value of the weights for the first principal component are reported in Table 9. As it is easy to notice, the features kurtosis and crest shows the highest absolute weights, proving to be the most influent in the computation. Furthermore, the higher weights are used with Yinf, which confirms to be the most informative channel (N.B. note from Figure 10 of Chapter 4 that the selected features do not vary in the same range of values, so that the PCA on the standardized features is needed to meaningfully compare the weights involved in PC1).

Table 9: PC1 absolute weights for the standardized features (centred and normalized on their standard deviation)

	Xinf	***Xsup***	***Yinf***	***Ysup***
rms	0,0008	0,0006	0,0001	0,0011
skew	0,0035	0,0009	0,0014	0,0115
kurt	0,0237	0,0153	0,0338	0,0240
crest	0,0295	0,0212	0,0963	0,0516
peak	0,0035	0,0028	0,0022	0,0055

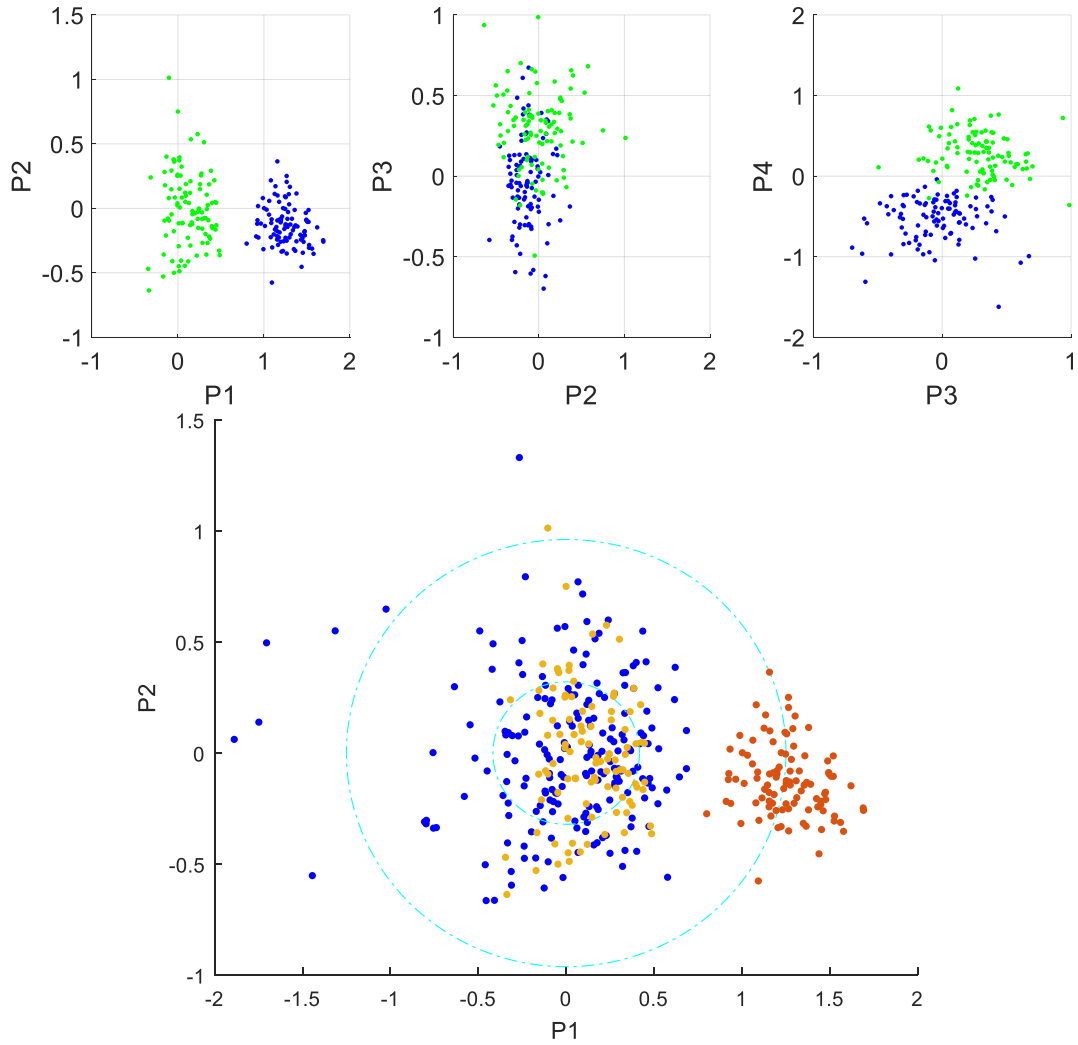


Figure 16: Above- PCA of the referenced set - the validation set is added later (green), projected on the same space. Below- Zoom of the space generated by the first 2 principal components (above), highlighting the healthy reference set (BLUE), the healthy validation set (yellow) and the damaged set (red).

For the sake of inquisitiveness, LDA is used after PCA in the reduced-dimensionality space created by P1 and P2 but also on the one generated by P2 and P3.

As it is easy to notice in Figure 17, on the P1-2 plane the separation is almost perfect. This is also highlighted by the confusion matrix and by the ROC curve, which is far from the bisector which represents the performance of a random classifier. The ROC curve summarizes an entire set of confusion matrices generated moving the threshold from a minimum to a maximum. The farthest point from the bisector gives the optimal threshold. On the contrary, on the P2-3 plane (Figure 18) the two clouds are almost indistinguishable.

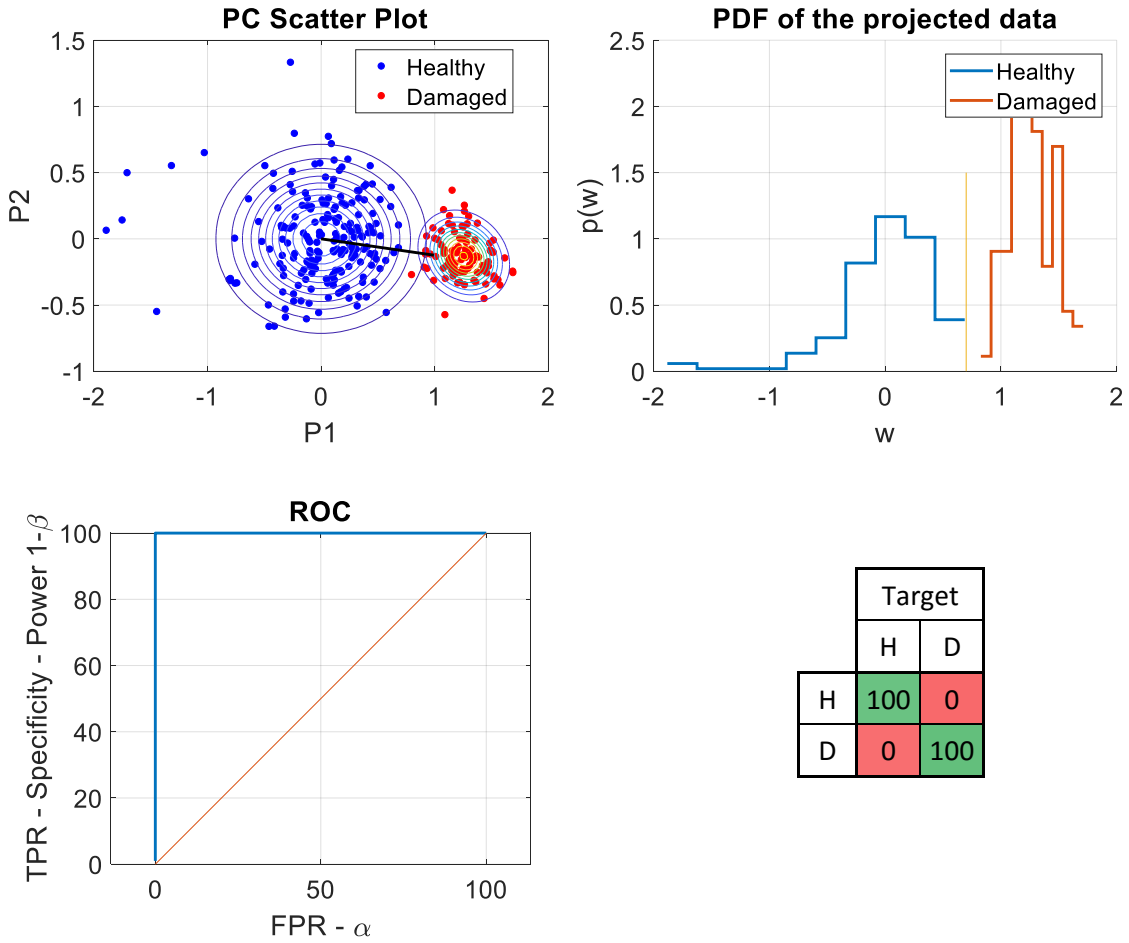


Figure 17: P1-P2 visualization of the multivariate dataset. The LDA direction is showed (black) and on this 1D projection the empirical PDFs are shown. The error table and the corresponding ROC curve are also reported.

The relevant consideration is that the dimensionality reduction found via PCA is basically a "lossy compression" of the original dataset. Then, there is no guarantee that the diagnostic information will be highlighted instead of being rejected. In this respect, a "lossless" algorithm reducing the dimensionality while concentrating the information would be much more appreciated. The Mahalanobis-Distance-based Novelty Detection is then applied in next section.

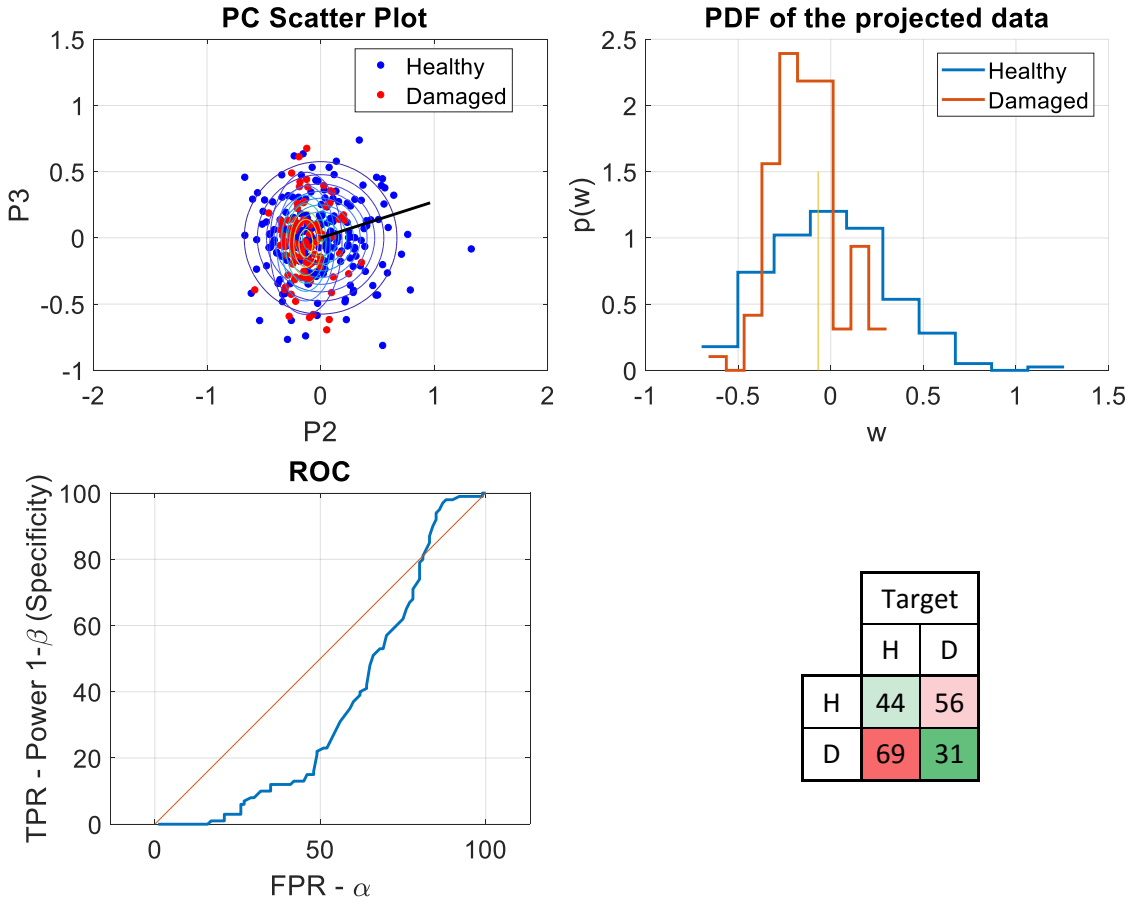


Figure 18: P2-P3 visualization of the multivariate dataset. The LDA direction is showed (black) and on this 1D projection the empirical PDFs are shown. The error table and the corresponding ROC curve are also reported.

4.3. Multivariate Novelty Detection

In order to assess whether the here proposed novelty detection on simple time domain features, already tested on experimental data, can be generalized for industrial machineries, the procedure is repeated on real windmills instrumented with a multichannel vibration monitoring system. The result is reported in Figure 19. There, the NIs of the damaged set are all very large and can be easily distinguished from the healthy NIs, allowing a perfect damage detection with no missed alarms. Unfortunately, the calibration set (observations 1-200) is not very big and is then non-representative of the entire variability in the different operational and environmental conditions. This explains why the proposed MC threshold is crossed many times in the healthy validation set (observations 201-300), implying a way too high false alarm rate. The considerable distance of the damaged NIs (observations 301-400) anyway, provides a wide margin to improve the threshold without increasing the missed alarms, that is, increasing the confidence without affecting the power, obtaining then a perfect detection.

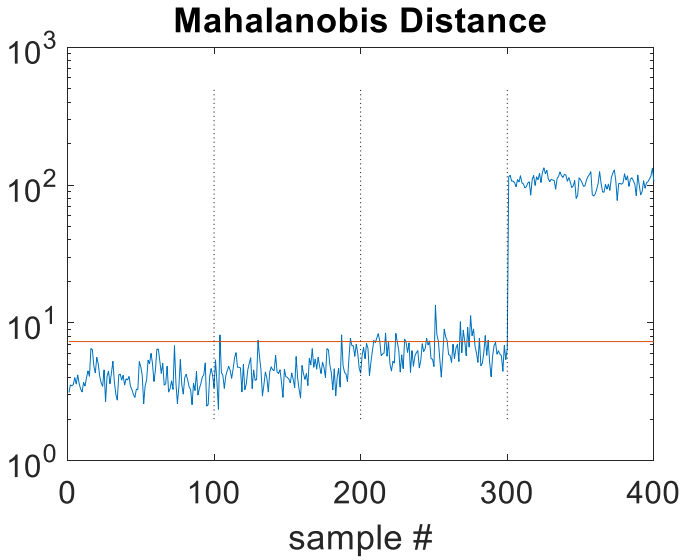


Figure 19: Mahalanobis Distance from the Calibration set (observations 1-200) – observations 201-300 form the healthy validation set, while observations 301-400 correspond to the damaged condition.

5. Conclusions

This chapter is devoted to the implementation and comparison of the selected techniques to perform machine diagnostics on three different applications using features extracted from vibration accelerometric measurements.

The selected techniques were organized to form a methodology. First, this methodology is meant to assess the goodness of the selected time features in performing diagnostics through ANOVA.

Subsequently, multivariate methods were applied. These are ultimately based on a dimensionality reduction of high-dimensional datasets to a 1-D case, on which a threshold can be set to separate (or at least try to separate) a healthy condition from a damaged one.

At this scope, PCA, the most famous dimensionality reduction algorithm was implemented. PCA is commonly used to produce 2-D or 3-D projections which enable the visualization of multivariate datasets. The scope is that of summarizing most of the dataset variability in the first components. Obviously, discarding components means pieces of information, which, despite being usually small, might be detrimental for diagnostic. This procedure is defined a “lossy compression”.

In this regard, a particular algorithm meant to find the 1-D projection that best separates healthy acquisitions from damaged exists. Fisher’s LDA in fact is able to maximize the distance of the two distributions (i.e. to find the direction with maximum “effect size”). LDA can be easily extended to multiple groups becoming a classifier algorithm. This possibly enables also level 2 and 3 detection. Unfortunately, the LDA is again lossy, as it is blind to variability in directions orthogonal to that of the 1-D reduction.

To face this issue, the k-NN classifier was first taken into account. The main advantage of k-NN is that it potentially involves all training data into the decision-making procedure. This is also its disadvantage, as it becomes computationally costly in case of long and heavy training sets (which must be stored entirely). A good lossless alternative is then the Novelty Detection. In fact, the use of the Mahalanobis distance as a measure of novelty, is a lossless (i.e. the distance from the centroid is based on the whole features set and

normalized according to the direction), non-linear (i.e. it is a sum of squares), dimensionality reduction (1-D Novelty Indices) which allows to substitute the training dataset with a centre and a covariance matrix. This is admissible when the training set distribution is normal or nearly normal, which is the optimal case for MD-ND.

MD-ND proved to be robust to quasilinear confounding influences such as varying operational conditions (e.g. speed, load, …), as it performs a PCA whitening of the data: eigenvectors of the data covariance matrix are normalized so as to have unitary eigenvalues (sphering transform), so that the contribution of the first PCs to the NIs is very limited. When the quasilinear confounding influence is very strong then, a residual found by removing the first principal component(s) (i.e. the Orthogonal Regression Residual) can be used to compute robust NIs.

The issue of confounders should be treated also for a more general case of strongly non-linear influences. In this case, the hypothesis of a normal healthy distribution does not hold at all, so that the novelty detection should be improved.

Switching the novelty information from distance to probability density, a generic multivariate pdf can be estimated.

In this regard, two different algorithms were tested. First a kernel density estimation was used for the pdf. Unfortunately, this shows the same limits as the k-NN classifier, as the whole training set must be stored to compute the probability density. To get a reliable estimate, the bandwidth parameter optimization becomes fundamental to find the best estimate in terms of bias-variance trade-off. Nevertheless, issues arise when high-dimensional feature spaces are involved. In general, then, it is wiser to switch again to parametric models, and find estimates of the non-normal pdf as a mixture of known pdf models. Gaussian Mixture Models were then taken into account.

The successful application of the here-proposed methodology to the laboratory research about the high-speed aeronautical bearings of the DIRG test rig was confirmed by the impressive results in the industrial application to the windmill gearboxes of the Italian windfarm. These results are very interesting also in terms of quickness, simplicity and full independence from human interaction, making the methodology suitable for real time implementation.

Bibliography

[1] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., "*The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data*", Mechanical Systems and Signal Processing, Volume 120, 2019. DOI: 10.1016/j.ymssp.2018.10.010.

[2] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., "Bearing damage detection techniques and their enhancement: comparison over real data", International Conference Surveillance 8, 2015.

[3] Daga A. P., Fasana A., Marchesiello S., Garibaldi L., "ANOVA and other statistical tools for bearing damage detection", International Conference Surveillance 9, 2017.

[4] Castellani F., Garibaldi L., Astolfi D., Daga A. P., Becchetti M., Fasana A., Marchesiello S. "*Fault diagnosis of wind turbine gearboxes through on-site measurements and vibrational signal processing*", International Conference on Structural Engineering Dynamics ICEDyn 2019, Viana do Castelo, Portugal.

Chapter 9:

Summary, Conclusions and further work

1. Summary and Conclusions

The main focus of this thesis was to test a selection of promising and reliable methods on laboratory and real-life applications evaluating and comparing their performances. In order to foster the potential benefits of Vibration Monitoring in industrial maintenance regimes the algorithms were summarized into two methodologies able to meet the needs of both continuous and intermittent monitoring of Gearboxes. In this regard, the state of the art was explored. The selection was primarily made according to model interpretability to select practical and representative algorithms able to diagnose the presence of incipient damage (damage detection, lev. 1 diagnostics) and possibly identify, locate and quantify the damage. This is primarily a matter of pattern recognition performed on some selected features in both time and frequency domain. The raw data in fact is "mined" so as to monitor the health condition of the gearbox. This information is then used to take judgements about the eventuality of triggering red alarms and stop the machine in case of impending failure or to neatly schedule maintenance interventions, in a data-to-decision framework.

The selected algorithms were tested not only on synthetic simulated signals, but also on three experimental acquisitions. The first laboratory measurement regards a test rig built at the Dynamic & Identification Research Group (DIRG) laboratory and specifically conceived to test high-speed aeronautical bearings. The other two on the contrary refers to real life applications such as a SAFRAN aeronautical engine (SAFRAN Contest data from Conference Surveillance 8 held in Roanne, France on October 20&21 2015) and an Italian windfarm composed by six multi-megawatt wind turbines installed in Molise region.

In Chapter 4 and 5 the proposed methodology and the datasets were illustrated in detail.

In Chapter 6 the Signal Processing algorithms for Intermittent Monitoring were first compared over synthetic data with different noise contamination, and later tested on the SAFRAN dataset. In this practical application, the Short Time Fourier Transform (STFT) was used to highlight the non-stationarity induced by the variable speed. The signal was then "re-phased" using Computed Order Tracking (COT) so that the signal was resampled on the basis of the tachometer key-phase signal. Once a synchronously sampled signal was obtained, algorithms for deterministic/non-deterministic contribution separation were tested. The Synchronous Average (SA) was compared to prediction-based algorithms such Linear Prediction (LP) and Discrete/Random Separation (DRS). The analysis highlighted the ability of SA in extracting the deterministic component with minimum disruption of the residual signal but underlined at the same time the deeper amount of geometric information needed to perform the analysis and the longer computational times required in case of complex gearboxes. Finally, to highlight the bearing signature, the Fast Kurtogram (FK) was compared to the Empirical Mode Decomposition (EMD), to the Stochastic Resonance (SR) but also on the improved algorithm here proposed for Spectral Kurtosis estimation (SK*). Both SK*, FK and EMD bands, selected on the basis of the excess kurtosis index, demonstrated to highlight the damage, even if in this particular case, despite the lack

of physical motivation, EMD demonstrated to be at least as effective as FK and SK*. SR was also effective in highlighting the bearing signature through denoising of the low frequency region, even if a long optimization of the parameters was needed, exploiting the prior knowledge of the characteristic frequencies of interest. These frequencies can be considered spectral features and are commonly used to assess the presence of damage. Despite this is usually performed by eye by a trained operator, more refined pattern recognition algorithms can be used.

This Machine Learning perspective is tested in Chapter 8. For the sake of computational speed and quickness, the analysis focused on the simpler, time-domain features which better adapt to the needs of continuous monitoring. In this case, the damage detection was tackled through statistical modelling. The analysis started introducing the philosophy of hypothesis testing, which was used to establish the problem and to get important considerations. At first, ANOVA was used to assess the goodness of the selected time features for diagnostic purposes. Then, in order to "condensate" the information contained in the different features enhancing the effect of damage, multivariate analyses were considered. In particular, most of the proposed algorithms (apart from k-NN classification) were binary classifiers based on dimensionality reduction of the multi-dimensional dataset. In this regard, the Principal Component Analysis, the most famous dimensionality reduction algorithm, was described. PCA summarizes the majority of a dataset variability in the first components, so that the last can be neglected. This is very useful for visualization (2D or 3D projections can be easily obtained), but not for diagnostics, in which the discard of components corresponds to a possible loss of information, leading to a "lossy" compression. Even the Fisher's Linear Discriminant Analysis, which maximizes the distance of the healthy and damaged distributions, was considered, but it shares the same problem of not exploiting all the available information. A "lossless" classification was obtained through k-NN, potentially involving all the training data into the decision-making strategy, but this leads to long and heavy calibration stages which can become costly. A good lossless solution is the Novelty Detection based on the Mahalanobis distance, which was finally considered. In this case the Novelty Index is based on the whole features set, and only the direction information is neglected. The whole methodology was applied to the laboratory research about high speed aeronautical bearings from the DIRG test rig, and then extended to the industrial application about windmill gearboxes damage detection from the Italian windfarm. The goodness of such results both in terms of Missed and False alarms than in terms of quickness, simplicity and full independence from human interaction, encouraged the further development of these algorithms also for the case of confounding influences (e.g. variable environmental conditions or operational conditions such as load or rotational speed producing non-stationary accelerometric signals). Even if the Mahalanobis Distance Novelty Detection is robust to weak *quasi-linear* confounding influences, some methods for facing stronger influences were tested. In particular, as first, the underlying model of MDND was sophisticated using Kernel Density Estimation and Gaussian Mixture Models. Then, also the removal of the confounding influences through PCA orthogonal regression and whitening was tested.

To complete the picture of the multivariate analyses, the curse of dimensionality was covered in Chapter 7 through Monte Carlo simulations. In particular, the selection of an optimal threshold, the minimum sample size ensuring statistical confidence, and the robust estimation of the covariance matrix against inclusive outliers were analysed.

2. Further work

This thesis presented two methodologies based on reliable algorithms which can be used for diagnostics of gearboxes. From a practical point of view, the part related to intermittent monitoring is usually performed off-line, but the continuous monitoring is meant for on-line applications. In this regard, the implementation of the algorithms in a true real-time environment is lacking and can then be the subject of further work.

Concerning signal processing algorithms, the novel estimator for the Spectral Kurtosis (SK*) introduced in Chapter 6 of this thesis deserves further attention. Despite it proved to work in a very effective way also on real life cases, additional tests are worth. A stronger mathematical analysis about the estimation error in terms of bias and variance could, for example, strengthen the motivations of the methodology. Furthermore, an optimization of the computational times could be performed. In its first implementation in fact, a large time is lost in recomputing the filter coefficients for each bandpass filtering operation (as the fir filter is meant to slide over the frequency axis). Nevertheless, it could be possible to exploit the shift theorem of the Discrete Fourier Transform to shift the signal spectrum while keeping the filter constant.

Focusing on the machine learning approach, the features and algorithms here considered proved to work very well in case of stationary conditions, and possibly with quasi-linear confounding influences, but in general, the effect of variable operational conditions (e.g. speed, load, …) or environmental conditions (e.g. temperature, humidity, …) should be tackled with a much deeper analysis, also considering more complex correlations and possibly cointegration.

Cointegration for example, is able to produce stationary signals from two or more non-stationary time series, compensating then for latent, unmeasured factors inducing the non-stationarities. Latent factors can be also accounted with Independent Component Analysis, Factor Analysis or also with Canonical Correlation Analysis. Non-linear dimensionality reduction methods like kernel-PCA, Locally-Linear Embedding or other Manifold algorithms should also be considered. Non-linear classification on the contrary can be addressed by Support Vector Machines through the kernel trick.

Differently, another way to address the problem of non-stationarities can be that of selecting features robust to the latent factors. Feature selection algorithms based on regression are already present in the literature such as Lasso or Ridge regression, so that they could possibly be adapted to objectively automate the choice.

Finally, in order to optimize the decision stage of the condition-based maintenance, the evaluation of the remaining useful life is fundamental. This is commonly addressed by prognostics, which requires deeper studies about the failure models and massive amounts of endurance (run-to-failure) acquisitions on which to tune them. These are not always easy to be obtained, especially for applications for which safety issues arise, but are fundamental to understand the failure modes, the early signs of wear and ageing and the system parameters that are to be monitored. The discipline that links the failure mechanisms to the maintenance management is often referred to as Prognostics and Health Management (PHM), and is based on two kinds of approaches [1,2]:

- Data-driven prognostics, which uses pattern recognition and machine learning techniques to detect changes in system states.
- Model-based prognostics, which tries to use physical models for estimating the remaining useful life (RUL)

Model-based approaches are usually preferred, but when the system is complex, the development of accurate models is often too expensive. Hence, data driven approaches are considerably more common. The RUL in this case can be learnt directly from data or by using models of cumulative damage on which a threshold is extrapolated [1].

This will give a strong push to the effectiveness of the data-to-decision process, fostering the reliability and then the spread of preventive maintenance regimes based on vibration monitoring.

Bibliography

[1] Mosallam A., Medjaher K., Zerhouni N., "Component based data-driven prognostics for complex systems: Methodology and applications," First International Conference on Reliability Systems Engineering (ICRSE), 2015. DOI: 10.1109/ICRSE.2015.7366504

[2] Atamuradov V., Medjaher K., Dersin P., Lamoureux B., Zerhouni N., “Prognostics and Health Management for Maintenance Practitioners-Review, Implementation and Tools Evaluation”, International Journal of Prognostics and Health Management, Special Issue on Railways & Mass Transportation, 2017.

Appendix 1:

Failure rate and reliability

1. Introduction

Nowadays machines are increasingly complex, so that the problem of reliability becomes more and more important. The reliability of a system is defined as the probability of proper operation under specified operational and environmental conditions for a defined period of time. It is then clear that increased levels of complexity imply increased risk of systems becoming prone to failure.

Reliability obviously involves economic considerations. In particular, a failure implies an external action such as repair which is usually very expensive also in terms of down times and production losses. Hence, external maintenance actions are commonly planned to avoid failure. In case of low reliability, maintenance costs may be very high. On the other side, increasing the reliability means improve the overall product quality, so that the product price becomes higher. A trade-off between the two aspects usually lead to the optimal reliability level which minimizes the overall costs (i.e. purchase price + repair and maintenance costs). Obviously, in case of risks for the human health and safety, maximum reliability is sought regardless of the costs.

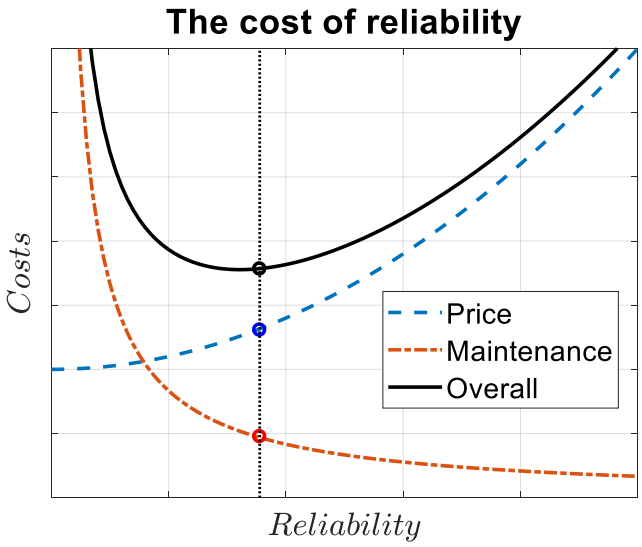


Figure 1: The cost of reliability: trade-off of purchase price and repair and maintenance costs.

2. Reliability measures

In order to quantify the reliability of a system, parameters describing the time between failures (or TBF i.e. the lifetime) distributions can be derived. In particular, the distribution is commonly specified by either one of the three possible functions here reported:

- A failure probability density function (PDF) $f(t)$, specifying the probability of failing within the time interval $[t, t + \Delta t]$ in the limit for $\Delta t \to 0$.

- A failure cumulative distribution function (CDF) $F(t)$, giving the probability of failing before time t.
- A survival cumulative distribution function, often identified as **reliability function** $R(t)$, which is the probability of surviving up to time t and corresponds then to the complementary of the CDF ($R = 1 - CDF$).

In general, the probability of death is commonly studied using Weibull family of distributions which features

$$f(t) = \frac{\beta}{\mu}\left(\frac{t}{\mu}\right)^{\beta-1} e^{-\left(\frac{t}{\mu}\right)^{\beta}} \qquad F(t) = 1 - e^{-\left(\frac{t}{\mu}\right)^{\beta}} \qquad R(t) = e^{-\left(\frac{t}{\mu}\right)^{\beta}} \qquad (1$$

Where μ is the "scale parameter" and β is the "slope parameter" or Weibull gradient defining the shape. Indeed, the Weibull distribution interpolates between the exponential distribution ($\beta = 1$) and the Rayleigh distribution ($\beta = 2$) while for $\beta \cong 3{,}57$ it approaches the Normal.
The shape of the Weibull β strongly influences the final parameter of interest of reliability engineering: the **failure rate** $\lambda(t)$. The failure rate can be looked from two perspectives.

- Focusing on a population of M equal machines it gives the relative number of units (n) failing on average in a unit of time.

$$\lambda(t) = \frac{1}{\Delta t}\frac{n}{M} \qquad (2$$

- Focusing on a single machine, it corresponds to the probability (density) of failure in the interval $[t, t + \Delta t]$ given that no failure is yet occurred.

$$\lambda(t) = \frac{1}{\Delta t}\frac{R(t) - R(t + \Delta t)}{R(t)} \qquad (3$$

which, brought to the limit for $\Delta t \to 0$, gives

$$\lambda(t) = \frac{f(t)}{R(t)} = -\frac{R'(t)}{R(t)} = h(t) \qquad (4$$

often called hazard function $h(t)$ which is NOT a conditional probability, but rahter a "conditional probability density" and therefore possibly greater than 1.

The influence of the shape parameter β on the failure rate can be easily obtained:

- A value of $\beta < 1$ indicates that the failure rate $\lambda(t)$ decreases over time. This condition usually holds at the beginning of the life of a product, when a significant "infant mortality" is due to design and manufacturing defects which are late recognized. Anyway, such issues are solved quickly, so as to find the ultimate optimal design.
- A value of $\beta > 1$ indicates that the failure rate $\lambda(t)$ increases with time. This happens if there is an "aging" process due to wear and is typical at the end of the service life of a product.

- A value of $\beta = 1$ indicates that the failure rate is constant over time. This suggests that failure (or mortality) is caused by a superimposition of random independent factors. The Weibull distribution in this case reduces to an exponential.

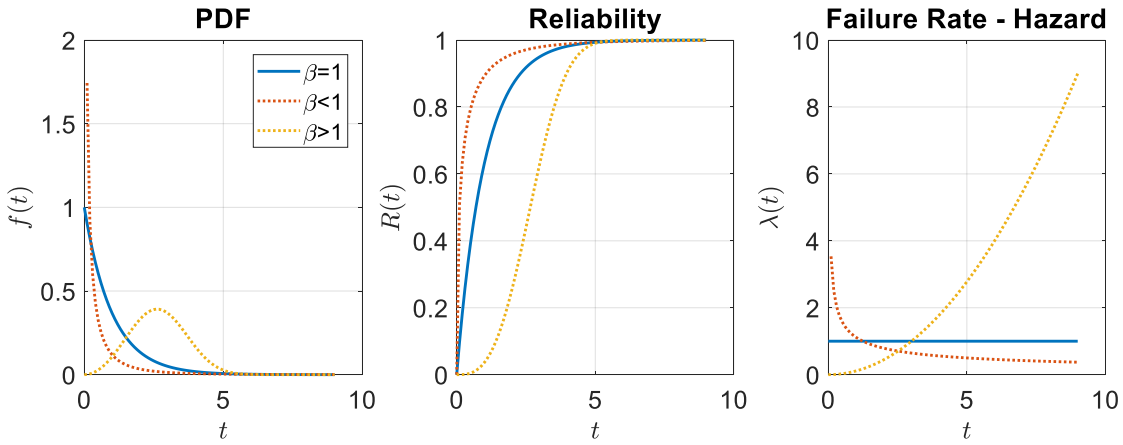


Figure 2: Weibull distribution of the time between failures – PDF, Reliability and Failure Rate are shown for different values of β.

Considering the evolution of the failure rate through the whole life of a product then, three regions can be distinguished, producing the well-known bathtub curve.

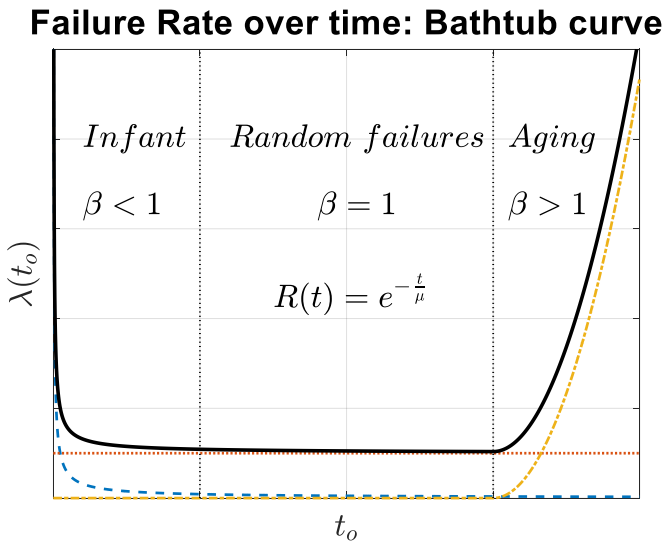


Figure 3: Failure rate evolution over the overall life of a product – The bathtub curve.

In any case, the most interesting region for scheduling maintenance interventions (i.e. Programmed Maintenance Regime) is the central region in which the statistical superimposition of a number of independent factors leads to random failures featuring an exponential distribution:

$$f(t) = \frac{1}{\mu} e^{-\frac{t}{\mu}} \qquad F(t) = 1 - e^{-\frac{t}{\mu}} \qquad R(t) = e^{-\frac{t}{\mu}} \qquad (5$$

In this particular case, an important parameter characterizing the distribution becomes the average time between failures (or "mean lifetime") which can be computed via expected value. Exploiting the property:

$$E[t] = \int_{-\infty}^{+\infty} t\, f(t)dt = \int_{-0}^{+\infty} (1 - F(t))dt - \int_{-\infty}^{0} F(t)dt \qquad (6$$

and remembering that the time axis is defined only for positive values of t, the formula simplifies to:

$$E[t] = \int_{-0}^{+\infty} R(t)dt = \int_{-0}^{+\infty} e^{-\frac{t}{\mu}}dt = \mu \qquad (7$$

The scale parameter can be then directly obtained as the average lifetime, usually called Mean Time To Failure (MTTF) for non-repairable products or Mean Time Between Failures (MTBF) in case of repairable machines. Having available a record of the TBFs t_i with $i = 1{:}\, r$ where r is the total number of failures, the MTBF can be simply estimated as:

$$MTBF = \frac{1}{r}\sum_{i=1}^{r} t_i = \mu \qquad (8$$

Furthermore, the reciprocal of this time period corresponds to the constant failure rate λ, as:

$$\lambda = \frac{f(t)}{R(t)} = \frac{\frac{1}{\mu}e^{-\frac{t}{\mu}}}{e^{-\frac{t}{\mu}}} = \frac{1}{\mu} \qquad (9$$

It is relevant to point out that, despite being the fundamental parameter on which basing reliability predictions for maintenance scheduling, the MTBF corresponds to the critical value for a confidence (i.e. reliability) of about 37% only, meaning that just the 37% of the machines will survive up to the MTBF.

3. Practical example

For the sake of exemplification, the dataset collecting the operation and maintenance data described in [1] is considered. This consists of field tracing records for $N = 24$ machining centres over a period of one year, with a particular focus on machine No. 1, which failed for $r = 27$ times in a period $T_o = 3744{,}4\ h$.

The analysis of the TBFs is conducted by fitting an exponential distribution, whose characteristic parameter μ can be estimated as:

$$\mu = \frac{T_o}{r} = 138{,}7\ h \qquad\qquad \lambda = \frac{1}{\mu} = 7{,}2\ 10^{-3}\ \ failures/h$$

In Figure 4 the so fitted exponential PDF is superimposed to the empirical PDF (i.e. the probability histogram). The goodness of the fit is remarkable. The corresponding CDF is also pictured together with the reliability function. As it is easy to notice, the MTBF features

a reliability of just 37%. It is obviously possible to look for a stricter critical value. For example, the time for which the 95% of the machines are surviving $t_{R=95\%}$ is computed and proved to corresponds to about 10 h, as highlighted in the picture.

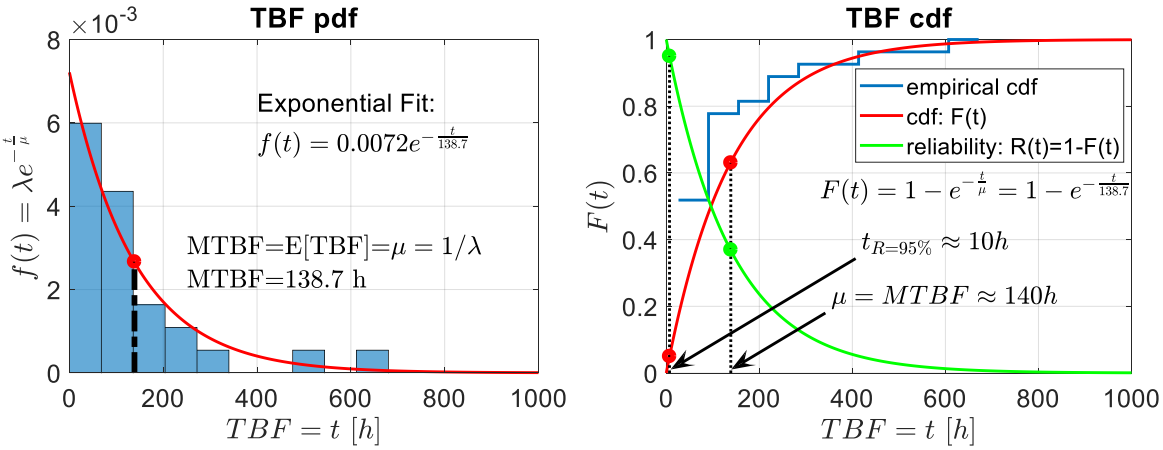


Figure 4: On the left, the fitted exponential PDF is compared to the empirical probability histogram obtained from the experimental data of machine No. 1 from [1]. On the right, the fitted CDF and the empirical CDF are shown, together with the reliability function. The MTBF and the 95% survival time $t_{R=95\%}$ are reported.

The goodness of the fit is also assessed comparing the theoretical constant failure rate coming after the assumption of exponential distribution, to the actual empirical failure rate obtained from the ratio of the empirical PDF and the reciprocal of the empirical CDF (the reliability R). The comparison using 8 bins is shown in Figure 5, computed as

$$\lambda(bin) = \frac{f(bin)}{R(bin)} \qquad (10$$

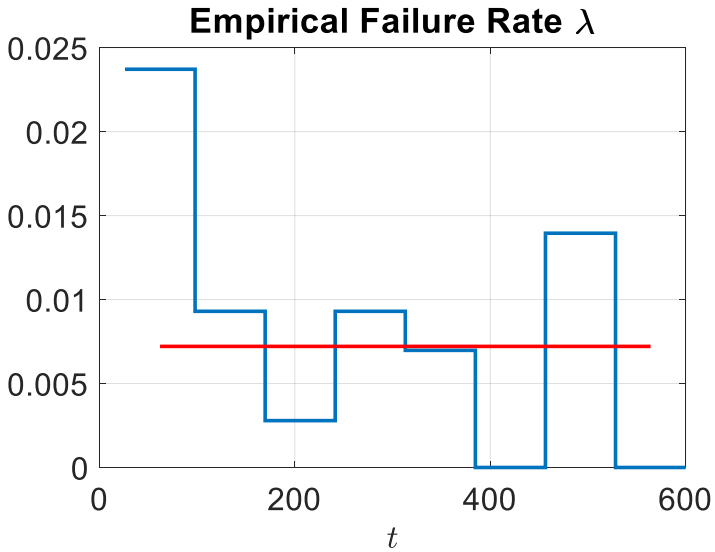


Figure 5: 8 bins empirical failure rate vs theoretical failure rate from the fitted exponential distribution.

Bibliography

[1] Yazhou J., Molin W., Zhixin J. "*Probability distribution of machining center failures".* Reliability Engineering & System Safety – Elsevier, 1995. DOI: 10.1016/0951-8320(95)00070-I

[2] Lienig J., Bruemmer H., "*Fundamentals of Electronic Systems"*, Springer, 1st ed. 2017, DOI: 10.1007/978-3-319-55840-0

Appendix 2:

Hilbert Transform and the analytic signal

1. The Hilbert Transform

The so-called Hilbert transform is a linear operator produced by the convolution of a generic function of time $f(t)$ by the function $-1/(\pi t)$. This produces the usual definition

$$\tilde{f}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{f(\tau)}{t-\tau} d\tau \qquad (1$$

which can be translated in the frequency domain by Fourier Transform:

$$\tilde{F}(\omega) = F(\omega)\left(-i\, sgn(\omega)\right) \qquad (2$$

The Hilbert transform imparts then a phase shift of $\pi/2$, positively or negatively depending on the sign of ω while leaving unaltered the amplitudes of the spectral components. Obviously, applying two Hilbert transforms in succession reverses the phases of all components, while applying four will restore the original function.

For example, the Hilbert transform of $cos(\omega t)$, where $\omega > 0$, is $cos\left(\omega t - \frac{\pi}{2}\right) = sin(\omega t)$, as it can be easily understandable from Figure 1.

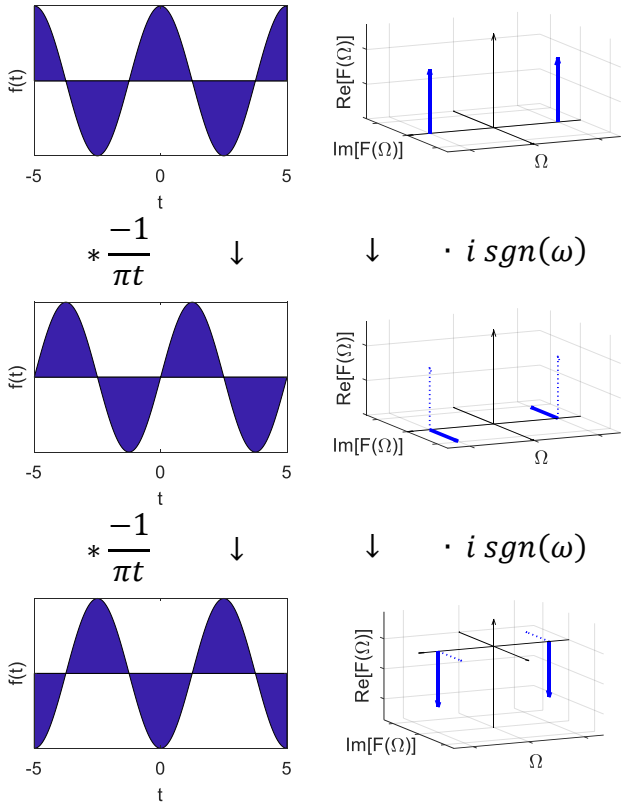


Figure 1: Two successive Hilbert transforms of a cosine wave.

It is relevant to point out that the Hilbert transform can be said to be the relationship between the real and imaginary parts of the FT of a one-sided function of time. In fact, given a causal function $f(t)$ ($f(t) = 0\ for\ t < 0$), finite at $t = 0$, its Fourier transform $F(\omega)$ will be a complex quantity $F(\omega) = X(\omega) + iY(\omega)$ for which the real and the imaginary parts are related by the Hilbert transform, so that $Y(\omega) = \tilde{X}(\omega)$. This can be highlighted by considering that a causal function is made up of even and odd components which are identical for positive time, and thus cancel for negative time. Since the even part of a time function transforms to the real part of its FT, and the odd part to the imaginary part, it easy to notice that:

$$f(t) = f_o(t) + f_e(t) = f_o(t) + f_o(t)\ sgn(t)$$
$$F(\omega) = \mathcal{F}[f_o(t)] + \mathcal{F}[f_e(t)] = Re[F(\omega)] + Im[F(\omega)] = F_o(\omega) + F_0(\omega) * \frac{1}{i\pi\omega} \quad (3$$

where $F_0(\omega) * \frac{1}{\pi\omega} = \widetilde{F_0}(\omega)$ defines the Hilbert transform, so that $F_o(\omega) = X(\omega)$ and $Y(\omega) = \widetilde{F_0}(\omega) = \tilde{X}(\omega)$.

In the same way, a one-sided frequency spectrum, which is causal in the frequency domain and has no negative frequency components, undergoing IFT will generate a complex time signal whose real and imaginary parts will be related by the Hilbert transform. This is usually known as analytic signal.

2. The analytic signal

In mathematics and signal processing, an analytic signal $f_A(t)$ is a complex-valued time function built not to have negative frequency components, namely to be causal in the frequency domain.

Therefore, in continuous time, every analytic signal can be generated via IFT of a one-sided spectrum:

$$f_A(t) = \frac{1}{2\pi}\int_0^\infty F(\Omega)e^{i\Omega t}d\Omega \quad (4$$

This analytic signal $f_A(t)$ can be proved to be:

$$f_A(t) = f(t) - i\,\mathcal{H}\big(f(t)\big) = f(t) - i\tilde{f}(t) \quad (5$$

where $f(t)$ is the original real-valued function of time and $\tilde{f}(t)$ is its Hilbert transform, often called quadrature signal. This is basically an expansion of the idea of phasors coming from Euler's equation

$$e^{i(\phi)} = cos(\phi) + i\ sin(\phi) \quad (6$$

according to which, any real harmonic $cos(\phi)$ can be converted into a positive-frequency complex harmonic $e^{i(\phi)}$ by adding as imaginary part the phase-quadrature component $sin(\phi)$ obtained by a simple quarter-cycle time shift.

Generalizing for an Amplitude and Frequency Modulated signal (AM-FM), its corresponding analytic form can be written as:

$$f_A(t) = A(t)\, e^{i(\phi(t))} \qquad (7$$

$A(t)$ represents the amplitude modulating function, and $\phi(t)$ represents the phase modulating function in radians. From this analytic signal then, it is possible to recover both the instantaneous amplitude $|A(t)|$, usually referred to as envelope, and the instantaneous angular frequency, corresponding to the time rate of change of the phase of the analytic signal $d\phi(t)/dt$.

3. AM signals and Amplitude Demodulation

The term Amplitude Modulation (AM), is traditionally used in connection with radio transmission. An AM signal encodes information into a carrier wave by modulating its amplitude in accordance with the signal to be sent (i.e. the baseband). A radio receiver must then demodulate the transmitted signal to recover the information. This can be done with a simple circuit called "envelope detector", able to produce the envelope of the modulates signal, which can be proved to correspond to be equivalent to the baseband signal. In common AM radio broadcasts, the carrier frequency ranges from 550kHz to 1600kHz while the audio signal to be transmitted (the modulating) covers the human hearing range from 20Hz to 20kHz.

One can consider, for the sake of exemplification, a $f_c = 600\ kHz$ carrier modulated by a $f_m = 98\ Hz$ sound (i.e. a pure G2 musical note frequency) with some additional Gaussian noise:

$$\begin{gathered} mod(t) = cos(2\pi f_m t) \\ carr(t) = A\, cos(2\pi f_c t) \\ n(t) \sim N(0,1) \end{gathered} \qquad s(t) = \left(1 + \frac{mod(t)}{A}\right) carr(t) + n(t); \qquad (8$$

In Figure 2, both the time and frequency representation of the synthesized signal for $A = 1$ can be found. The expected first order sidebands $f_c \pm f_m$ arise. A demodulation via Hilbert transform can be implemented directly on the raw signal. To improve the effectiveness of the demodulation in terms of signal to noise ratio (SNR), the raw signal can be band filtered around the carrier frequency, so as to remove most of the background noise. Figure 2 clearly highlight these considerations.

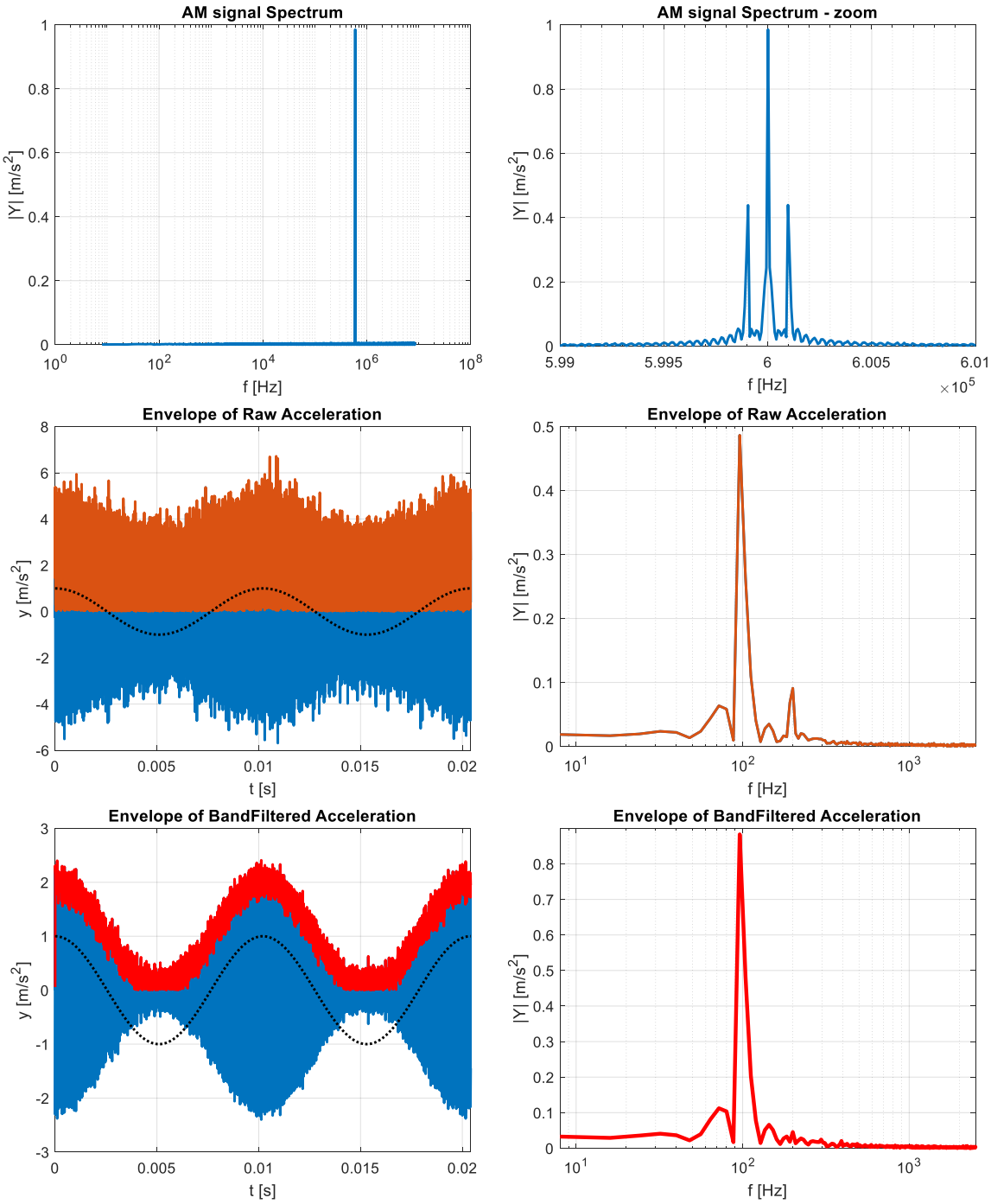


Figure 2: Synthesized AM signal in the typical radio transmission framework ($600kHz$ carrier). In the first row the spectrum of the AM signal is given. In the second row the envelope of the overall signal is reproduced to highlight the presence of the $98Hz$ modulating. In time domain, the envelope (orange) is compared to the AM signal (blue) and the modulating signal (black dotted). In the third row the effect of filtering around the carrier is highlighted.

3.1. Bearings Envelope Analysis

The rolling element bearing, is a fundamental component of most of mechanical systems. Bearing faults are commonly believed to cause an amplitude modulation to the high frequency noise [2], which becomes then a carrier for the diagnostic information. Hence, envelope demodulation is the natural signal processing to recover the bearing-characteristic spectral lines. In the example, a synthetic signal generated in accordance to

Chapter 4, Section 2 is given. In particular, the sampling frequency is set to $fs = 22528\ Hz$, the shaft to $fr = 53\ Hz$, the structural resonance to $fn = 5600\ Hz$. Five harmonics of shaft and five of the gear-mesh are considered for gear-wheel with $z = 23$. The SKF 6006-Z bearing is modelled, featuring an inner ring fault ($BPFI = 251\ Hz$).

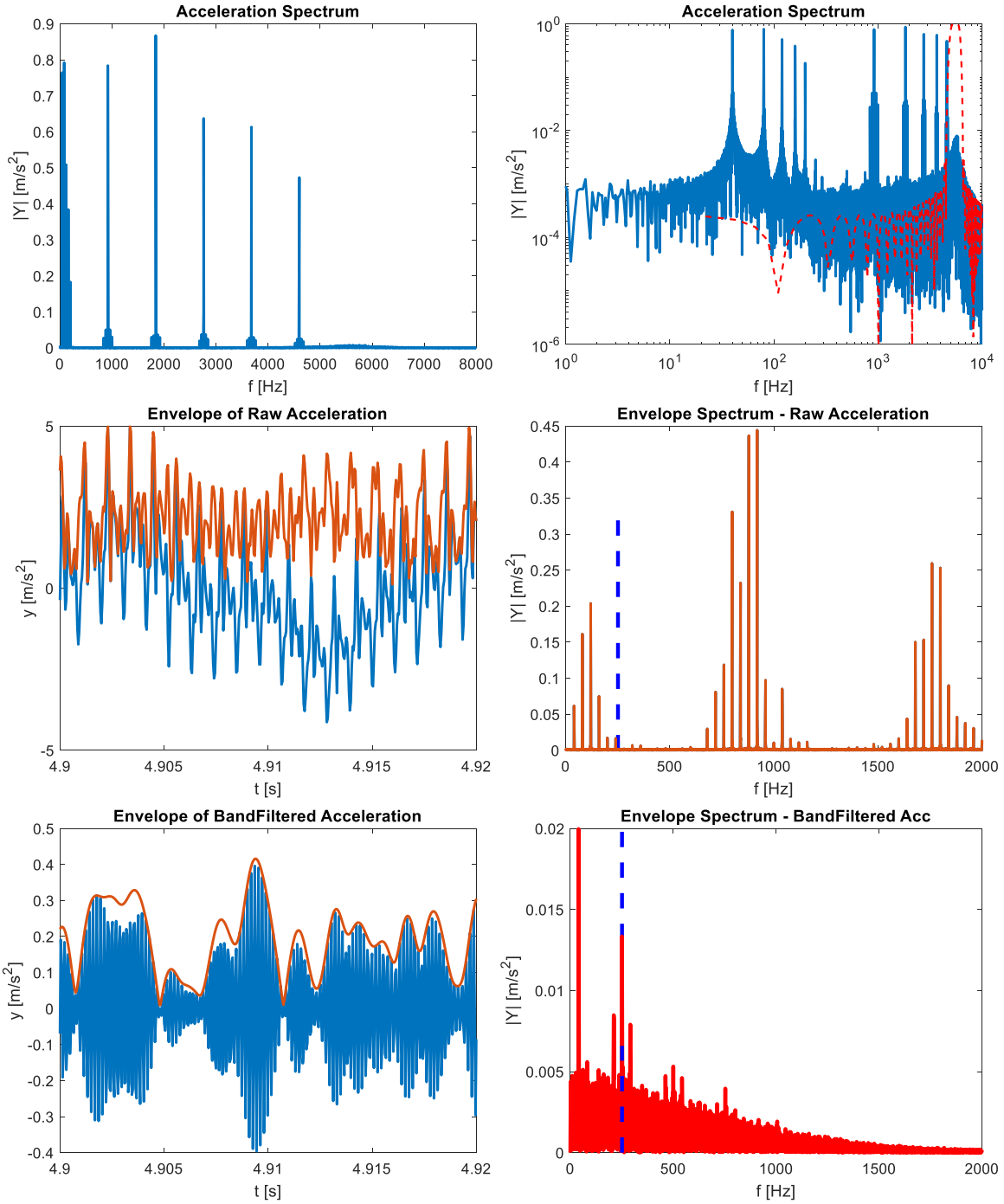


Figure 3: Synthesized signal from a bearing with an inner race damage (BPFI). In the first row the spectrum of the bearing signal is given, also in log-scale. The transfer function of a band-pass filter selecting the resonance band is reported in red-dotted line. In the second row the envelope of the overall signal is reproduced. The expected BPFI spectral line (blue-dashed) is not present in the spectrum. No detection occurs. In the third row, the band-filtered signal envelope is analyzed. In the spectrum the BPFI frequency is clearly highlighted.

Bibliography

[1] Ronald Newbold Bracewell, "*The Fourier Transform and Its Applications"*. McGraw Hill, 2000. ISBN: 0070070156

[2] Robert Bond Randall, "*Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications"*. John Wiley & Sons, 2011. ISBN: 9780470747858

[3] Bendat, J.S., "*The Hilbert Transform and Applications to Correlation Measurements"*. Report prepared for Brüel & Kjær, Copenhagen No. BT0008-11.

[4] Michael Feldman, "*Hilbert Transform Applications in Mechanical Vibration*", Wiley, 2011. ISBN: 978-0-470-97827-6

Appendix 3:

Stochastic processes, probability theory and spectra

1. Introduction

A stochastic process is the ensemble of all the possible realizations of a random variable over time. The properties of such a process can be then derived from the properties of the realizations (acquired signals). Random basically stands for unpredictable, as the opposite of deterministic. A deterministic signal is a completely specified function of time whose behaviour can be determined with certainty at every time instant. An example of a deterministic signal can be a periodic signal (e.g. a harmonic $x(t) = A\cos(\omega t)$), but also aperiodic signals like transients and quasi-periodic signals (sum of more sinusoids whose frequencies are not entire multiple of the fundamental) belong to this category. Unfortunately, in the real world, no measured signal will be perfectly deterministic, because different realizations will always show some differences, albeit small. Although it is not possible to make perfect predictions, a probabilistic approach can still be used to deal with such an uncertainty. At each time t a stochastic process $y(t)$ can be then defined in terms of a probability distribution.

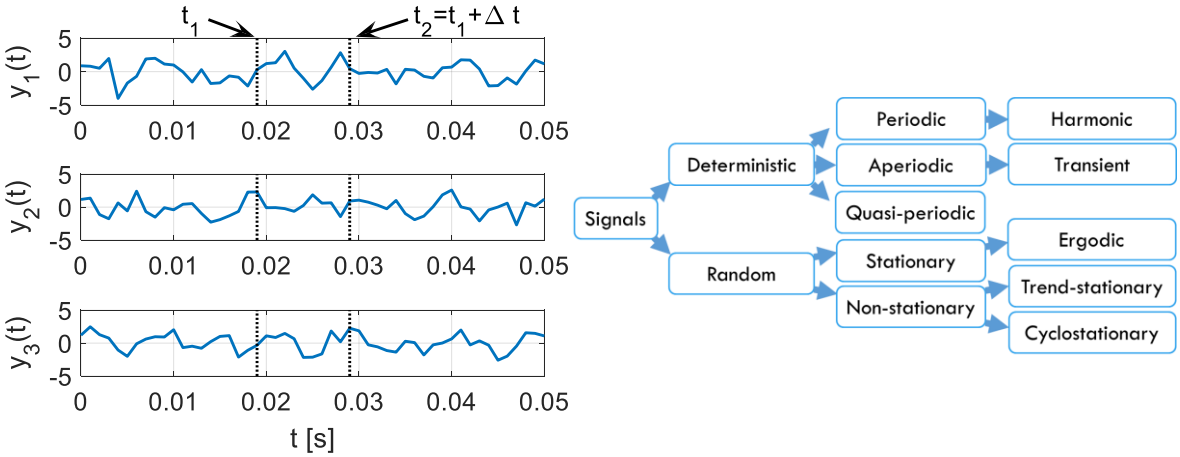


Figure 1: Three realizations of a stochastic process on the left. The classification of signals from chapter 3 on the right.

2. The probabilistic approach

Probability is a measure of the likelihood that an event will occur. Consider the discrete variable $y(t)$, whose discrete realization $y(kT_s) = y(k)$ has been measured. This measure can be easily compared to a threshold y: the likelihood of $y(k)$ being less or equal than the threshold, takes the name of Cumulative Distribution Function (cdf):

$$P(y) = prob[y(k) < y] \qquad (1$$

Assuming a continuous cdf, the Probability Density Function (pdf) of such a variable is defined as:

$$p(y) = \lim_{\Delta y \to 0}\left(\frac{prob[y < y(k) \leq y + \Delta y]}{\Delta y}\right) = \lim_{\Delta y \to 0}\left(\frac{P(y+\Delta y) - P(y)}{\Delta y}\right) = \frac{dP}{dy} \qquad (2$$

with the following properties:

$$p(y) \geq 0 \qquad P(y) = \int_{-\infty}^{y} p(\gamma)d\gamma \qquad \int_{-\infty}^{+\infty} p(y)dy = 1 \qquad (3$$

In mathematics, in particular in statistics, specific quantitative measures of the shape of a function can be computed. These statistical functions are said to summarize the distribution. In statistics, a relevant summary is the measure of the centre, or location of the distribution, which can be given by:

- the mode: the most frequent value, corresponding to a peak in the pdf (N.B. if the distribution is not unimodal, more local peaks can be found)
- the median: the value separating the higher half from the lower half of the data (the value which divides the pdf in two parts of equal areas)
- the mean: the expected value of the process $E[y]$, the value at which the arithmetic average tends at the limit for large number of repetitions.

In general, the main statistical functions which describe the shape of the pdf are called moments, and are defined as

$$\int_{-\infty}^{+\infty} (x-c)^n f(x)dx \qquad (4$$

where n is the moment order, while c is a constant equal to 0 for the raw moments and to the mean value for centred moments. When also a normalization is performed, the moment is said to be standardized. The most relevant moments are:

Order 1 – raw moment: *Location* (5	**mean**	$\mu_y = E[y(k)] = \int_{-\infty}^{+\infty} y\, p(y)dy$
Order 2 – central moment: *Dispersion* (6	**variance**	$\sigma_y^2 = E\left[\left(y(k) - \mu_y\right)^2\right] = \int_{-\infty}^{+\infty} \left(y - \mu_y\right)^2 p(y)dy$
Order 3 – standardized moment *Shape:* symmetry (7	**skewness**	$\frac{\mu_3}{\sigma_y^3} = E\left[\left(\frac{y(k) - \mu_y}{\sigma_y}\right)^3\right]$
Order 4 – standardized moment *Shape: tailedness* (8	**kurtosis**	$\frac{\mu_4}{\sigma_y^4} = E\left[\left(\frac{y(k) - \mu_y}{\sigma_y}\right)^4\right]$

It is important to highlight that the idea of expected value of a generic variable q intrinsically implies that the average value of repeated realizations of q will converge, at least on the long run ($N \to \infty$) to a finite value $E(q)$. This is basically the Law of Large Numbers (LLN), which is a fundamental principle as it guarantees stable long-term results for the averages of random processes. Furthermore, in probability theory, a second fundamental principle is the Central Limit Theorem (CLT). It establishes that when independent random variables are added, their sum tends to the well-known Gauss bell

distribution (which takes the name of Normal or Gaussian) independently from the distributions of the original variables.

3. Gaussian distribution and Gaussian processes

One of the most relevant distributions is then the Normal (or Gaussian) distribution, which is the limiting probability distribution of complicated sums, according to the CLT. A Gaussian process $y(t) \sim N\big(\mu(t), \sigma^2(t)\big)$, is then a stochastic process described by the bell-shaped Gaussian probability density function (pdf)

$$N(\mu, \sigma^2): \;\; p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|y-\mu|^2}{2\sigma^2}} \qquad (9$$

This pdf, which holds for the time instant t, is completely defined by the first two moments:

$$\mu(t) = E[y(t)] \qquad \text{The mean (expected value)} \qquad (10$$

$$\sigma^2(t) = Var(y(t)) = E\left[\big(y(t) - \mu(t)\big)^2\right] \qquad \text{The variance} \qquad (11$$

But this is not enough to describe the Gaussian process over time. Let us consider two time-instants t_1 and t_2. The two distributions $y(t_1) \sim N(\mu(t_1), \sigma^2(t_1))$ and $y(t_2) \sim N(\mu(t_2), \sigma^2(t_2))$ can be considered as marginal distributions, but in the general case, considering that $y(t_1)$ and $y(t_2)$ may not be independent, a joint pdf is necessary to completely depict the process. Indeed, a correlation among two following instants is possible, related to a memory effect which characterizes the dynamic of every system. In this case a bivariate normal should be considered. This can be defined as:

$$p(y(t_1), y(t_2)) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{z}{2(1-\rho^2)}} \qquad (12$$

where:

$$z = \frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} \qquad \text{Normalized radius} \qquad (13$$

$$\gamma = E[(y(t_1)-\mu_1)(y(t_2)-\mu_2)] = cov\big(y(t_1), y(t_2)\big) \qquad \text{Auto-Covariance} \qquad (14$$

$$R = E[y(t_1)y(t_2)] = corr\big(y(t_1), y(t_2)\big) \qquad \text{Auto-Correlation} \qquad (15$$

$$\rho = \frac{\gamma}{\sigma_1\sigma_2} = corr(y(t_1), y(t_2)) \qquad \text{Correlation coefficient} \qquad (16$$

It is relevant to notice that $-1 \le \rho \le 1$ and $\rho = 0$ implies uncorrelation between the two time-instants (N.B. independence implies uncorrelation, but the opposite does not hold). Finally, in order to completely define a Gaussian process, at least the mean and the auto-covariance function γ (ACF) are needed. Conversely, in the case of null mean value, which is common for accelerometric measures, the auto-covariance simplifies to the auto-correlation $\gamma = R = E[y(t_1)y(t_2)]$, which becomes the only characterizing parameter.

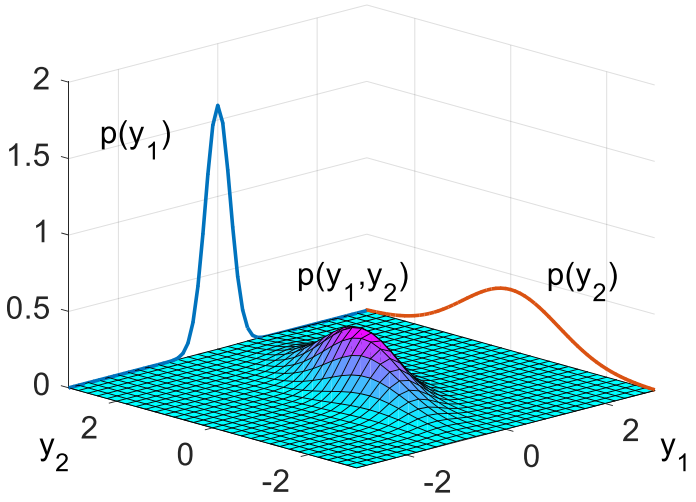


Figure 2: Bivariate Normal joint distribution with the two marginal distributions highlighted.

4. Correlation and Spectral densities

It is important to remember that, when the properties of a process are not known a priori, one will have to rely on estimates, always affected by uncertainty. According to the LLN, in fact, given a finite number of realizations, the estimate is computed at a given confidence level.

The main estimates characterizing a stochastic process of N realizations i of duration T are:

Through the different realizations:	Through the different time-instants:	
$\mu(t) = \lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} y_i(t)\right)$	$\mu(i) = \lim_{T\to\infty}\left(\frac{1}{T}\int_0^T y_i(t)dt\right)$	*(17*
$R_{yy}(t,\tau) = \lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} y_i(t)y_i(t+\tau)\right)$	$R_{yy}(\tau,i) = \lim_{T\to\infty}\frac{1}{T}\int_0^T y_i(t)y_i(t+\tau)\,dt$	*(18*

If the mean $\mu(t)$ and the auto-correlation $R_{yy}(t,\tau)$ are variable with time, the process is said to be non-stationary. On the contrary, if at least such quantities are constant in time, the process is at least weakly stationary: $\mu(t) = \mu$ and $R_{yy}(t,\tau) = R_{yy}(\tau)$. Furthermore, a stationary process is said ergodic if these quantities does not change with the realizations.

Focusing on stationary processes, the characterization can be moved to the frequency domain and, exploiting the Wiener-Khinchine theorem, it is possible to define the power spectral density (PSD) as

$$S_{yy}(\Omega) = \mathcal{F}\left[R_{yy}(\tau)\right] = \int_{-\infty}^{+\infty} R_{yy}(\tau)e^{-i\Omega\tau}d\tau \qquad (19$$

This real and even quantity, periodic with period $2\pi/T_s$, is often called two-sided auto-spectral density function, but in the engineering field, the one-sided is usually preferred, so that:

$$G_{yy}(\Omega) = 2S_{yy}(\Omega), \quad \Omega \geq 0 \qquad (20$$

It is wise to notice that integrating the spectral density, the average power of the signal (the second order moment) is obtained:

$$R_{yy}(0) = E[y(t)^2] = \int_{-\infty}^{+\infty} S_{yy}(\Omega)d\Omega = \int_{0}^{+\infty} G_{yy}(\Omega)d\Omega \qquad (21$$

These considerations are fundamental from the theoretical point of view, but in practice, the estimate of the spectral density of a stationary stochastic process is usually obtained through a procedure based on Fourier transform called **Periodogram**.
Basically, following the Welch method, a long acquisition is divided into K shorter overlapping parts which corresponds to different realizations. Each realization k of length T is then windowed and brought to the frequency domain. The k-th auto spectral density can be written as:

$$\hat{S}_{yy}(\Omega, T, k) = \frac{1}{T} Y_k^*(\Omega, T) Y_k(\Omega, T) \qquad (22$$

where $\cdot^*$ means the complex conjugate.
The results will be finally aggregated to find the estimate:

$$\hat{S}_{yy}(\Omega, T) = E[\hat{S}_{yy}(\Omega, T, k)] \qquad (23$$

Obviously, this procedure reduces the variability in the estimated spectral density (thanks to aggregation of the results) in exchange for reducing the frequency resolution (the chunks are shorter than the overall signal), but it can be proved that the estimator is unbiased, so that:

$$S_{yy}(\Omega) = \lim_{T \to \infty} E[\hat{S}_{yy}(\Omega, T, k)] \qquad (24$$

In the discrete domain, considering K chunks, the following results can be proved:

$$\begin{gathered} \hat{S}_{yy}(\Omega) = \frac{1}{K} \sum_{k=1}^{K} Y_k^*(\Omega) Y_k(\Omega) \\ \frac{2K\hat{S}_{yy}(\Omega)}{S_{yy}(\Omega)} \sim \chi^2(2K) \\ \frac{2K\hat{S}_{yy}(\Omega)}{\chi^2_{1-\alpha/2}} \leq S_{yy}(\Omega) \leq \frac{2K\hat{S}_{yy}(\Omega)}{\chi^2_{\alpha/2}} \end{gathered} \qquad (25$$

The χ^2 distribution with $2K$ degrees of freedom can be then used in order to find the range within which the true psd is supposed to fall at a confidence α. A simple bi-harmonic signal (sum of two harmonics at 100 and 150 Hz) sampled at $1\,kHz$ and corrupted by random white noise is used to show the idea of range of confidence. The synthesized time signal and the corresponding PSD estimated via Welch method are reported in Figure 3.

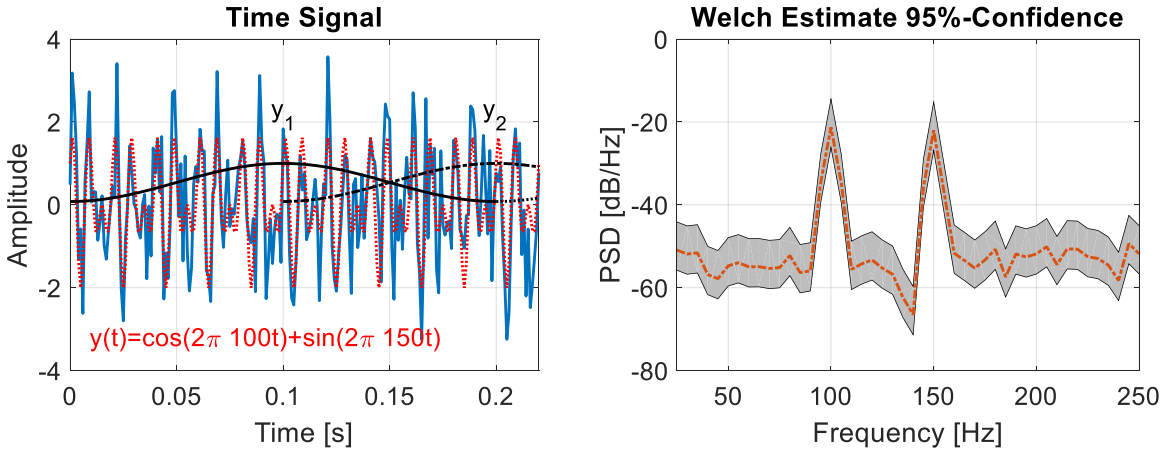


Figure 3: Modelled time signal (red) corrupted by noise on the left and the corresponding Welch periodogram with 95% confidence interval on the right. The 50% overlap used is highlighted by the reported Hamming windows on the left plot (in black).

5. LTI systems and stochastic signals

According to the Wold theorem, every covariance-stationary time series can be decomposed into a unique deterministic and stochastic part. Then, focusing on a generic stationary random signal $y(t)$, this can be obtained as the response of a linear system characterized by the impulse response $h(t)$ when excited by a white noise $w(t) \sim N(0, \sigma_w^2)$.

In the time domain this can be written as the convolution:

$$y(t) = h(t) * w(t) = \sum_{\tau=-\infty}^{t} h(t-\tau) \cdot w(\tau) = \sum_{\tau=0}^{+\infty} h(t) \cdot w(t-\tau) \qquad (26$$

The ACF of the output signal can be then found to convey, to a scale factor, the impulse response $h(t)$, so that:

$$R(\tau) = h(\tau) * h(-\tau) \cdot \sigma_w^2 = \sum_{t=0}^{\infty} h(t) \cdot h(t+\tau) \cdot \sigma_w^2 \qquad (27$$

And in the frequency domain:

$$S(\Omega) \ = \mathcal{F}[\gamma(\tau)] = |\mathcal{F}[h(\tau)]|^2 \cdot \sigma_w^2 = |H(\Omega)|^2 \cdot \sigma_w^2 \qquad (28$$

The usefulness of the Wold Theorem is inherent to the description of the dynamic evolution of a system through a causal linear model. Two considerations are particularly relevant:

- Causality implies predictability from past values. The linear model allows then to specify any generic stationary stochastic signal as it represents the relation of an observed value to its past values. In discrete time domain, this means that the moving average model is the only possible representation of such a relationship:

$$y(n\Delta t) = \sum_{k=0}^{\infty} b_k \varepsilon_{n-k} \qquad (29$$

  where ε is a white noise signal usually called innovation, while b are the (possibly infinite) weights of a linear moving average filter.

- Furthermore, even if the innovation is random, the MA parameters b are determined. The ACF of a generic stochastic signal is then a periodic function and therefore deterministic. The PSD of such a generic stochastic signal is deterministic as well.

Bibliography

[1] Fasana A., & Marchesiello S., "*Meccanica delle vibrazioni".* Torino: Clut, 2006. SBN: 88-7992-217-3

[2] Richard W. Hamming, "*Digital Filters - Dover Civil and Mechanical Engineering - third edition".* Courier Corporation, 2013. ISBN-13: 978-0-486-65088-3

[3] Fassois S.D., Sakellariou J.S., "*Time-series methods for fault detection and identification in vibrating structures"*. Phil. Trans. R. Soc. A (2007) 365, 411–448, 2007. Doi: 10.1098/rsta.2006.1929

[4] J. Antoni, R.B. Randall, "*Unsupervised noise cancellation for vibration signals: part I—evaluation of adaptive algorithms*", Mechanical Systems and Signal Processing, 18, pp89–101, 2004, DOI:10.1016/S0888-3270(03)00012-8

Appendix 4:

Hypothesis tests for Normality and Homoscedasticity

1. Normality tests

In statistics, normality tests are used to find if a dataset can be fitted well by a normal distribution.
Normality tests can be classified in three categories:

- Visual tests,
- Moment tests,
- Empirical distribution tests.

Among the visual, the most used are the *Q-Q plots*, which are plots of the sorted values from the data set against the expected values of the corresponding quantiles from the standard normal distribution. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the I-III quadrant bisector (y = x).

This test is easily extended to multivariate analysis with the name of *Mahalanobis Q-Q plot*. It can be proved in fact that Mahalanobis distance of the P-dimensional data set from the mean (d_M^2) has a χ_P^2 distribution.

$$d_M^2 = (x - \bar{x})^t \, CS_x (x - \bar{x}) \sim \chi^2_{(P)} \qquad (1$$

where S_x is the covariance matrix composed by:

$$s_{ij} = \frac{1}{K} \sum_{k=1}^{K} (x_{ki} - \bar{x}_i)\left(x_{kj} - \bar{x}_j\right) \qquad (2$$

This means that, for example, considering a 2-dimensional data set, it's possible to write with a confidence level of *99,7%* that

$$d_M^2 \leq \chi^2_{(2),0,3\%} \cong 11{,}6 \qquad (3$$

where $\chi^2_{(P),\alpha}$ is called *critical value* for the corresponding *p-value* of 0.997.
This is shown graphically in Figure 1, where the $\chi^2_{(2)}$ Cumulative Distribution Function is reported, together with the $99{,}7\%$ confidence critical value, featuring a significance $\alpha = 0{,}3\%$.

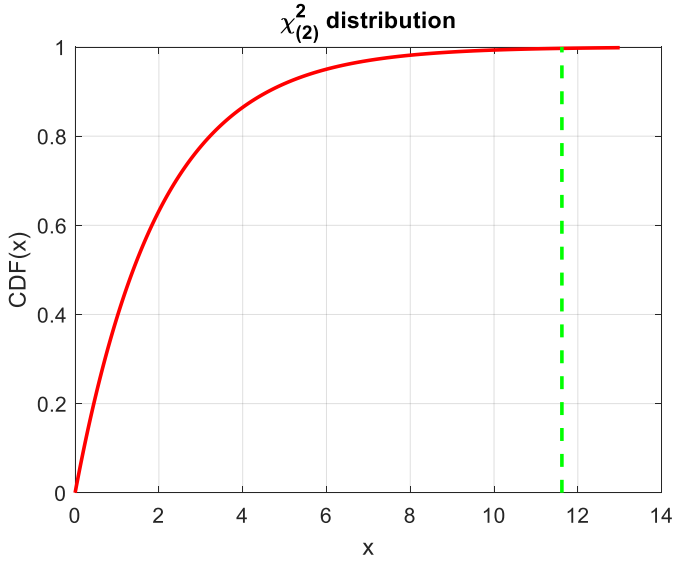


Figure 1: Cumulative probability density function (CDF) of $\chi^2_{(2)}$; The critical value $\chi^2_{(2),0.003}$ is shown in green.

Considering moment test, one of the most used results to be the *Jarque-Bera* test, which is based on descriptors of the shape of a probability density function (PDF) such as *skewness* (s), a measure of the *asymmetry*, and *kurtosis* (k), that measures the "*tailedness*" of the PDF.
The method is based on the fact that the statistical test JB will be asymptotically distributed as a χ^2_2.
So, the H_0 hypothesis of normality will be verified if

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right) \leq \chi^2_{(2),\alpha} \qquad (4$$

where *n* is the number of observations in the data set and $\chi^2_{(2),\alpha}$ is the upper-tail critical value for the χ^2_2 distribution.

Empirical distribution tests like *Lilliefors* can even be carried out, comparing the empirical cumulative distribution function (CDF) of the sample data, with the CDF of the normal distribution with estimated parameters equal to the sample parameters.

2. Homoscedasticity tests

In order to test if univariate data sets have homogeneous variances, a lot of tests are available, starting from *Fisher-Snedecor F-test*, limited to comparison among 2 groups, and switching to *Bartlett's test* to compare multiple variances at the same time.

2.1. Fisher-Snedecor F-test

Fisher-Snedecor F-test, is simply based on the fact that, given two normally distributed sets *x* ~ *N* and y ~ *N,* their squared deviations from the mean will be distributed as $\chi^2_x(nx)$ and $\chi^2_y(ny)$, where *nx* and *ny* are the dimensions of the 2 sets, while their ratio will show a particular distribution called Fisher-Snedecor *F*.

So, the H_0 hypothesis of homoscedasticity will be verified if

$$F = \frac{S_x^2}{S_y^2} \leq F_{(nx-1,ny-1),\alpha} \qquad (5$$

Where $S_x^2 = \frac{1}{n_x - 1}\sum_i (x_i - \bar{x})$ and similarly for S_y^2, are the non-distorted estimators for sample variance.

2.2. Bartlett's test

Bartlett's test on the contrary uses Bratlett's χ_B^2 statistic to be compared with a χ_{k-1}^2 distribution. If there are *G* samples (groups) of size n_i and sample variances S_i^2:

$$\chi_B^2 = \frac{(N-G)ln\left(S_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\, ln(S_i^2)}{1 + \frac{1}{3(G-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-G}\right)} \leq \chi_{(G-1),\alpha}^2 \qquad (6$$

where $N = \sum_{i=1}^{G} n_i$ and $S_p^2 = \frac{1}{N-G}\sum_i (n_i - 1)S_i^2$ is the pooled estimate for the variance.

Bartlett's test can be extended to multivariate cases, taking the name of *Box's M test*.

Bibliography

[1] Ghasemi, A., & Zahediasl, S. (2012). "*Normality tests for statistical analysis: a guide for non-statisticians*". International journal of endocrinology and metabolism, 10(2), 486–489. DOI: 10.5812/ijem.3505

[2] Goldfeld, S. M., & Quandt, R. E. (1965). "*Some Tests for Homoscedasticity*". Journal of the American Statistical Association, 60(310), 539. DOI: 10.2307/2282689

[3] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Appendix 5:

The Genetic Algorithm: evolutionary optimization

1. Evolutionary optimization and the Genetic Algorithm

Optimization is the selection of a best element from a set of available alternatives according to some criteria. In a more formal way, given an objective function $f: D \to \mathbb{R}$ which links the search space of feasible solutions to the corresponding utility or cost, the optimization process seeks to find the element $x_o \in D$ such that $f(x_o) \leq f(x)\ \forall\ x \in D$ (minimization) or such that $f(x_o) \geq f(x)\ \forall\ x \in D$ (maximization). Fixing a target for convenience, in the simplest case, an optimization problem consists of a minimization of a cost function over a search space obtained by constraining the overall Euclidean space. Or, $\underset{x \in D}{\operatorname{argmin}} f(x)$. From a mathematical point of view, the minimization of a function typically involves derivatives. Then, the more a function is complex (e.g. defined on a wide multidimensional support, non continuous, or with non continuous derivatives, featuring many lcoal minima etc.), the harder is the computation of such derivatives, so that the optimization may become very tricky in practical cases. Furthermore, the optimization is very likely to get stuck into local minima in the vicinity of an initial guess value for the optimum location (local optimization), with no guarantees (unless particular properties of the cost function i. e. convexity) that the result corresponds to the actual global minimum (global minimization).

In general, the assessment of the performance of an optimizer can be expressed in terms of

- *Exploration*: the optimizer discovers a wide region of the search space,
- *Exploitation*: the optimizer "pounds the pavement" on a limited but promizing region,
- *Reliability*: repeatability of the fund solution.

It is important to highlihgt that exploration and exploitation are competing properties. Local optimizers show very good exploitation at the expense of a very poor exploration. On the contrary, a good global optimizer should sacrifice exploitation to gain in exploration and speed. This is usualy obtained taking advantage of heuristic or meta-heuristic techniques implementing some form of stochastic optimization.

An important category of global population-based metaheuristic optimization algorithms is the Evolutionary. An evolutionary algorithm (EA) uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the cost function determines the quality of a solution. A "direct search" is performed to find the best individuals within the population according to their quality. These best individuals are then selected to determine the offspring, namely the new trial solutions, which will substitute lower quality individuals.

The most famous EA is the Genetic Algorithm, developed by John Holland introduced genetic algorithms in 1960 based on the concept of Darwin's theory of evolution. The GA evolutionary cycle starts initiating a population randomly and evaluating the quality of each individual on the basis of his genotype. The best individuals are then

selected to produce via modification the new offspring, while the worst are discarded. Modifications are stochastically triggered operators such as the crossover (the offspring is a random mix of the genotypes of their parents) or the mutation (the offspring features new genes which were not present in the parents). The first is important to ensure exploitation, while the second guarantees exploration of the search pace of all possible genotypes. Finally, a new population is ready for starting again the cycle until some stopping criteria is met. The cycle is outlined in Figure 1.

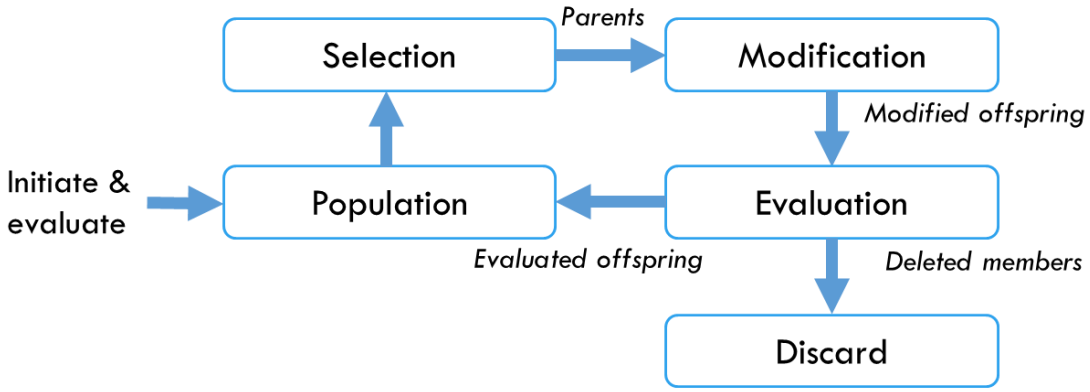


Figure 1: The evolutionary cycle

GA usual parameters in Matlab® environment are:

- Population Size: $N_p = 200$ or reduced to 50 if the problem dimension D (number of variables) is lower than 5.
- Elite Count: 5%. It defines the number of best individuals selected as a percentage of N_p.
- Crossover Fraction: 80%. It defines the offspring quantity at next generation as a percentage of N_p.
- Default mutation: Shrinking Gaussian. Each newborn features a degree of random mutation which decreases in time according to the linear law: $\sigma_g = \sigma_{g-1}\left(1 - s\,\frac{g}{G}\right)$. $\sigma_0 = 1$, s$= 1$.
- The maximum number of generations G is by default 100 times the problem dimension D.
- As the total Np is fixed, the percentage of discarded individuals equals the crossover fraction.
- Additional to G another stopping criterion is the maximum number of stall generations. The algorithm stops if the average relative change in the best objective function value is less than a function tolerance of 10^{-6} for 50 generations.

2. GA examples

In order to assess the performance of GA algorithm, benchmark functions are commonly used. These functions are typically generalized on a D-dimensional domain and are ruled by some parameters. Furthermore, they are built to have a unique and well-known global minimum in the origin, featuring a null cost. The performances of the algorithm are assessed in terms of

- Effectiveness: The optimization ensures to reach the right global minimum. This involves all the three properties introduced in previous chapter.

- The exploration, to find the area in which the global minimum is without getting stuck into local minima.
- The exploitation, to descend at the very bottom of the minimum and reach the value of null cost.
- The repeatability, so that even running the algorithm multiple time, the first two properties holds for each run.

- Efficiency: The time or number of generations in which the minimization is carried out.

while the number of generations G is fixed to 200 and is selected as the only stopping criterion.

To highlight the repeatability, each test is run 5 times. In the following examples, for convenience, the search space is bounded to the region $-5 \leq x_i \leq 5, i = 1{:}D$.

2.1. The Sphere Function

The first and simplest function tested is the Sphere Function, basically the Eulerian distance from the origin of a point in a D dimensional space.

$$f(x) = \sum_{i=1}^{D} x_i^2 \qquad (1$$

This function is convex, and has then a single minimum in the origin, which is obviously global. Its representation for $D = 2$ can be found in Figure 2.

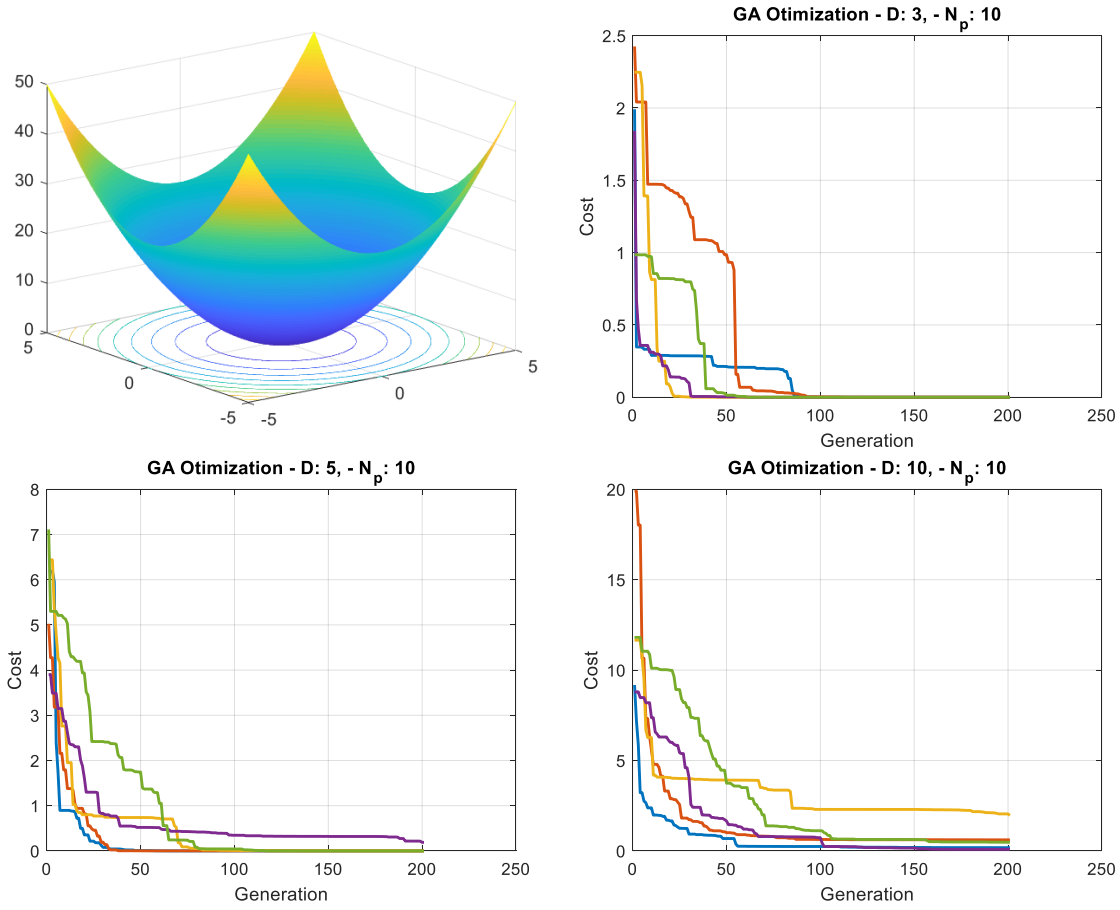


Figure 2: Sphere Function convex GA optimization performance in term of costs over generations (iterations) for different space dimensions D given a population size $N_p = 10$.

As it is easy to notice in Figure 2, even with a small population of only 10 individuals GA is very effective in minimizing the sphere function for a not too large dimensionality. When D increases larger N_p are needed to ensure a reliable minimization in acceptable numbers of iterations. Indeed, the search space volume increases as a power of D.

2.2. The Rastrigin Function

The second tested function is the Rastrigin, which features several local minima. It is non-convex and highly multimodal, but locations of the minima are regularly distributed. Its formulation, as a function of parameter a is given by:

$$f(x) = a\,D + \sum_{i=1}^{D} [x_i^2 - a\cos(2\pi x_i)] \qquad (2$$

The recommended value of a is 10 and its two-dimensional form is shown in Figure 3.
In the same figure the GA minimum cost at each generation is given for a constant dimension $D = 5$ while the population number N_p is increasing from 10 to 40. Despite the performances are obviously improving as N_p gets larger, it is also evident that an exploration issue may arise. GA often get stuck into local minima and even for $N_p = 40$ just 3 times out of 5 the algorithm is able to really find the global minimum. Larger population sizes and possibly stronger mutations would surely increase the reliability.

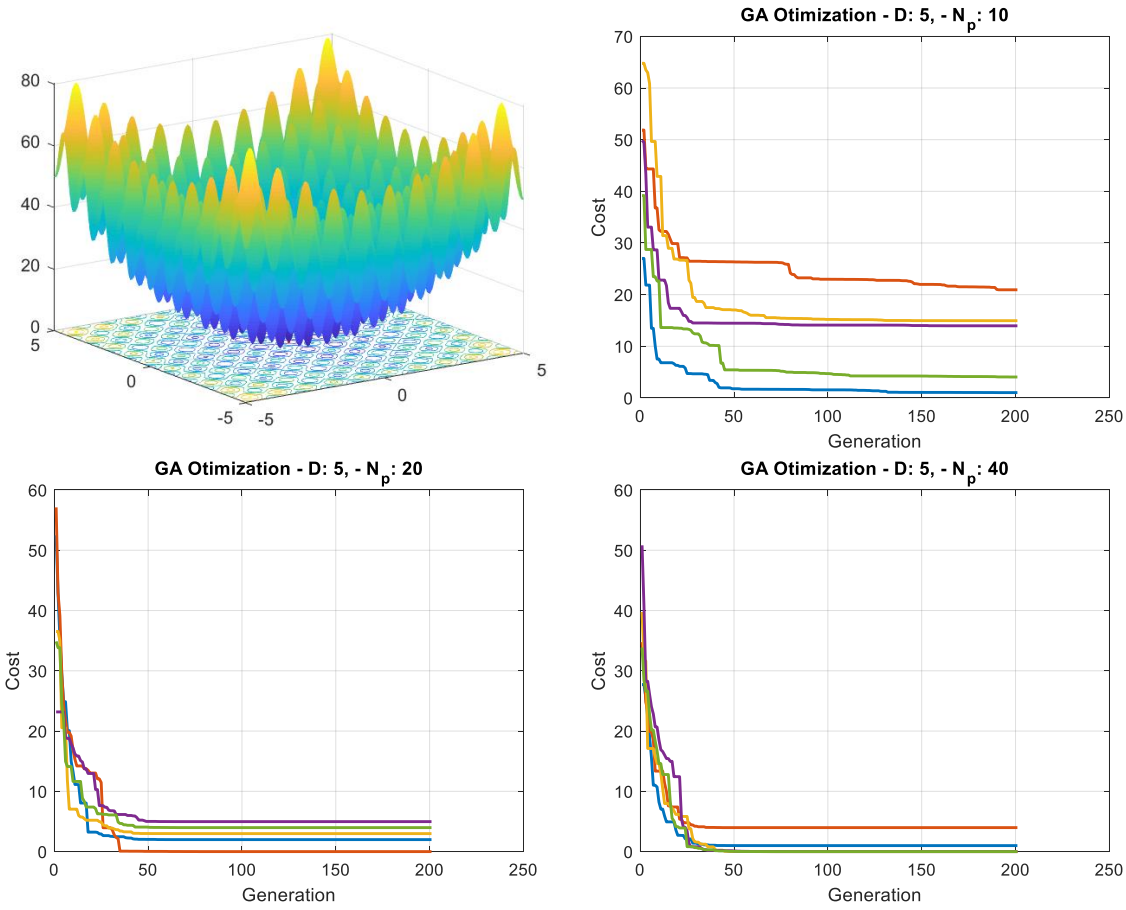


Figure 3: Rastrigin Function non-convex GA optimization performance in term of costs over generations (iterations) for a given space dimensions $D = 5$ increasing the population size N_p.

2.3. The Ackley Function

The last function proposed in this performance assessment is the Ackley function, shown in its two-dimensional form in Figure 4. As evident in the plot, it is characterized by a nearly flat outer region, and a large hole at the centre, while the surface features many local minima. Its general formulation is given by:

$$f(x) = a + \exp(1) - a\exp\left(-b\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - \exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(cx_i)\right) \qquad (3$$

Recommended variable values are: $a = 20$, $b = 0{,}2$ and $c = 2\pi$.

The Ackley function is again non-convex and shares the same criticalities of the Rastrigin function. The optimizer in fact, can get stuck in one of the many local minima and never reach the global optimum. In any case, with the same parameters, GA seems to perform much better than with the Rastrigin as already with $N_p = 20$ just in 1 case out of 5 it fails to find the global optimum, while for the other 4 it is reached in less than 100 generations.

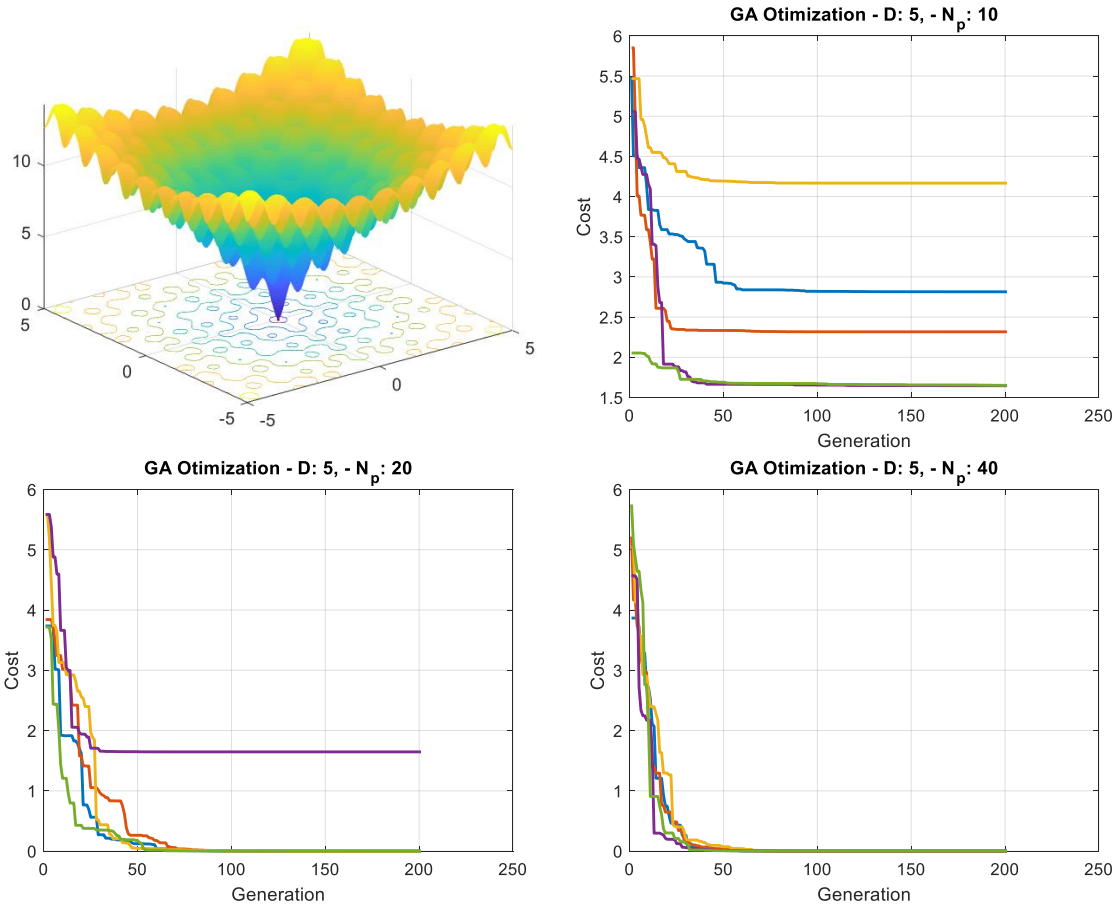


Figure 4: Ackley Function non-convex GA optimization performance in term of costs over generations (iterations) for a given space dimensions $D = 5$ increasing the population size N_p.

Bibliography

[1] Goldberg, David E., "*Genetic Algorithms in Search, Optimization & Machine Learning*", Addison-Wesley, 1989. ISBN:0201157675

[2] Simon D., "*Evolutionary Optimization Algorithms: Biologically-Inspired and Population-Based Approaches to Computer Intelligence*", John Wiley & Sons, 2013. ISBN-10: 0470937416